Oracle® Database

Machine Learning for SQL Use Cases

Release 23ai
F74720-05
April 2025

Oracle Database Machine Learning for SQL Use Cases, Release 23ai

F74720-05

Primary Author: Sarika Surampudi

Contents

1 Overview

2 Get Started

3 Develop

5 Examples

6 Reference

7 Concepts

1 Overview

Machine Learning Overview

Machine learning is a subset of Artificial Intelligence (AI) that focuses on building systems that learn or improve performance based on the data they consume.

What Is Machine Learning?

Machine learning is a technique that discovers previously unknown relationships in data.

Machine learning and AI are often discussed together. An important distinction is that although all machine learning is AI, not all AI is machine learning. Machine learning automatically searches potentially large stores of data to discover patterns and trends that go beyond simple statistical analysis. Machine learning uses sophisticated algorithms that identify patterns in data creating models. Those models can be used to make predictions and forecasts, and categorize data.

The key features of machine learning are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Ability to analyze potentially large volumes of data

Machine learning can answer questions that cannot be addressed through traditional deductive query and reporting techniques.

Benefits of Machine Learning

Machine learning is a powerful technology that can help you find patterns and relationships within your data.

Find trends and patterns - Machine learning discovers hidden information in your data. You might already be aware of important patterns as a result of working with your data over time. Machine learning can confirm or qualify such empirical observations in addition to finding new patterns that are not immediately distinguishable through simple observation. Machine learning can discover predictive relationships that are not causal relationships. For example, machine learning might determine that males with incomes between $50,000 and $65,000 who subscribe to certain magazines are likely to buy a given product. You can use this information to help you develop a marketing strategy. Machine learning can handle large volume of data and can be used in financial analysis. Some of the benefits include stock price predictions (algorithmic trading) and portfolio management.

Make data driven decisions - Many companies have big data and extracting meaningful information from that data is important in making data driven business decisions. By leveraging machine learning algorithms, organizations are able to transform data into knowledge and actionable intelligence. With the changing demands, companies are able to make better decisions faster by using machine learning techniques.

Recommend products - Machine learning results can also be used to influence customer decisions by promoting or recommending relevant and useful products based on behavior patterns of customers online or their response to a marketing campaign.

Detect fraud, anomalies, and security risks - The Financial Services sector has benefited from machine learning algorithms and techniques by discovering unusual patterns or fraud and responding to new fraud behaviors much more quickly. Today companies and governments are conducting business and sharing information online. In such cases, network security is a concern. Machine learning can help in detecting anomalous behavior and automatically take corrective actions.

Retail analysis - Machine learning helps to analyze customer purchase patterns to provide promotional offers for target customers. This service ensures superior customer experience and improves customer loyalty.

Healthcare - Machine learning in medical use is becoming common, helping patients and doctors. Advanced machine learning techniques are used in radiology to make an intelligent decision by reviewing images such as radiographs, CT, MRI, PET images, and radiology reports. It is reported that machine learning-based automatic detection and diagnosis are at par or better than the diagnosis of an actual radiologist. Some of the machine learning applications are trained to detect breast cancer. Another common use of machine learning in the medical field is that of automated billing. Some applications using machine learning can also point out patient's risk for various conditions such as stroke, diabetes, coronary artery diseases, and kidney failures and recommend medication or procedure that may be necessary.

To summarize, machine learning can:

- easily identify trends and patterns
- simplify product marketing and sales forecast
- facilitate early anomaly detection
- minimize manual intervention by "learning"
- handle multidimensional data

Define Your Business Problem

Enterprises face problems such as classifying documents, predicting the financial outcomes, detecting hidden patterns and anomalies, and so on. Machine learning can help solve such problems provided that you have clear understanding of the business problem with enough data and learn to ask the right questions to obtain meaningful results.

You require skills in preparing data, applying ML techniques, and evaluating results. The patterns you find through machine learning may be very different depending on how you formulate the problem. For example, rather than trying to learn how to "improve the response to a direct mail campaign," you might try to find the characteristics of people who have responded to your campaigns in the past. You can then classify if a given profile of a prospect would respond to a direct email campaign.

Many forms of machine learning are predictive. For example, a model can predict income level based on education and other demographic factors. Predictions have an associated probability (How likely is this prediction to be true?). Prediction probabilities are also known as confidence (How confident can I be of this prediction?). Some forms of predictive machine learning generate rules, which are conditions that imply a given outcome. For example, a rule can specify that a person who has a bachelor's degree and lives in a certain neighborhood is likely to have an income greater than the regional average. Rules have an associated support (What percentage of the population satisfies the rule?).

Other forms of machine learning identify groupings in the data. For example, a model might identify the segment of the population that has an income within a specified range, that has a good driving record, and that leases a new car on a yearly basis.

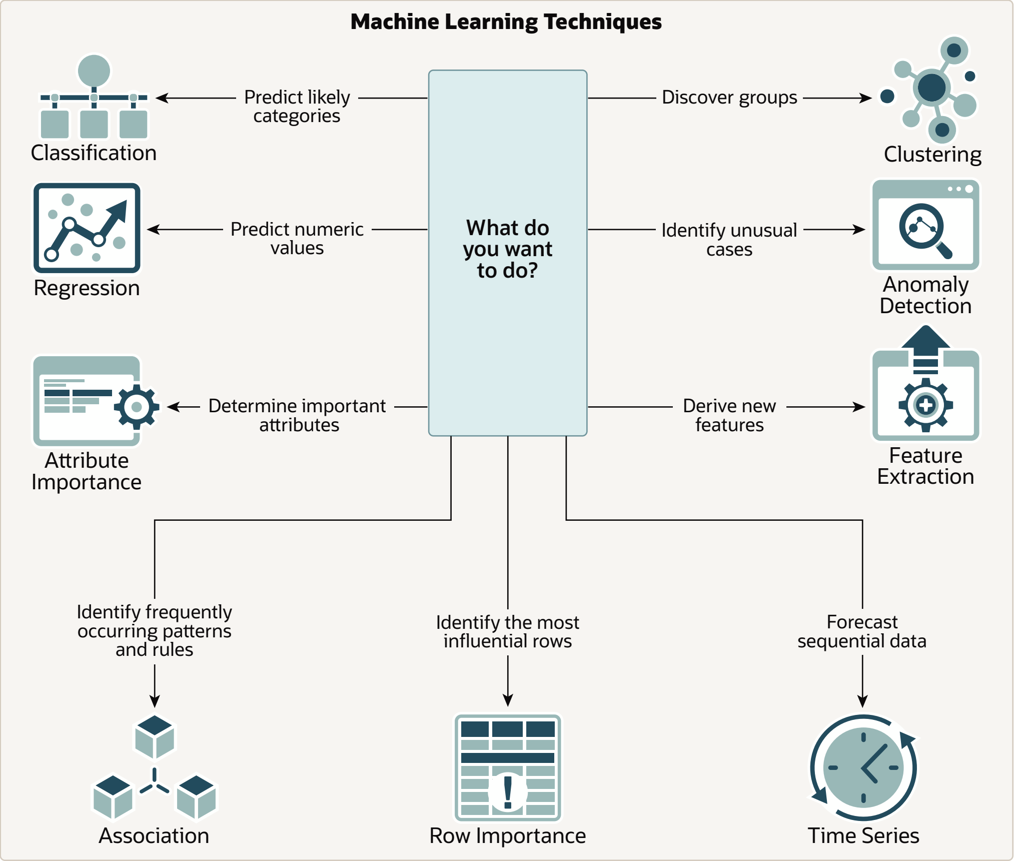
What Do You Want to Do?

Multiple machine learning techniques, also referred to as "mining function", are available through Oracle Database and Oracle Autonomous Database. Depending on your business problem, you can identify the appropriate mining function, or combination of mining functions, and select the algorithm or algorithms that may best support the solution.

For some mining functions, you can choose from among multiple algorithms. For specific problems, one technique or algorithm may be a better fit than the other or more than one algorithm can be used to solve the problem.

The following diagram provides a basic idea on how to select machine learning techniques that are available across Oracle Database and Oracle Autonomous Database.

Figure 1-1 Machine Learning Techniques



OML provides machine learning capabilities within Oracle Database by offering a broad set of in-database algorithms to perform a variety of machine learning techniques such as Classification, Regression, Clustering, Feature Extraction, Anomaly Detection, Association (Market Basket Analysis), and Time Series. Others include Attribute Importance, Row Importance, and Ranking. OML uses built-in features of Oracle Database to maximize scalability, improved memory, and performance. OML is also integrated with open source languages such as Python and R. Through the use of open source packages from R and Python, users can extend this set of techniques and algorithms in combination with embedded execution from OML4Py and OML4R.

Discover More Through Interfaces

Oracle supports programming language interfaces for SQL, R, and Python, and no-code user interfaces such as OML AutoML UI and Oracle Data Miner, and REST model management and deployment through OML Services.

Oracle Machine Learning Notebooks (OML Notebooks) is based on Apache Zeppelin technology enabling you to perform machine learning in Oracle Autonomous Database (Autonomous Data Warehouse (ADW), Autonomous Transactional Database (ATP), and Autonomous JSON Database (AJD)). OML Notebooks helps users explore, visualize, and prepare data, and develop and document analytical methodologies.

AutoML User Interface (AutoML UI) is an Oracle Machine Learning interface that provides you no-code automated machine learning. When you create and run an experiment in AutoML UI, it automatically performs algorithm and feature selection, as well as model tuning and selection, thereby enhancing productivity as well as model accuracy and performance. Business users without extensive data science background can use AutoML UI to create and deploy machine learning models.

Oracle Machine Learning Services (OML Services) extends OML functionality to support model deployment and model lifecycle management for both in-database OML models and third-party Open Neural Networks Exchange (ONNX) format machine learning models through REST APIs. The REST API for Oracle Machine Learning Services provides REST API endpoints hosted on Oracle Autonomous Database. These endpoints enable you to store machine learning models along with its metadata, and create scoring endpoints for the model.

Oracle Machine Learning for Python (OML4Py) enables you to run Python commands and scripts for data transformations and for statistical, machine learning, and graphical analysis on data stored in or accessible through Oracle Autonomous Database service using a Python API. OML4Py is a Python module that enables Python users to manipulate data in database tables and views using Python syntax. OML4Py functions and methods transparently translate a select set of Python functions into SQL for in-database execution. OML4Py users can use Automated Machine Learning (AutoML) to enhance user productivity and machine learning results through automated algorithm and feature selection, as well as model tuning and selection. Users can use Embedded Python Execution to run user-defined Python functions in Python engines spawned by the Autonomous Database environment.

Oracle Machine Learning for R (OML4R) provides a database-centric environment for end-to-end analytical processes in R, with immediate deployment of user-defined R functions to production environments. OML4R is a set of R packages and Oracle Database features that enable an R user to operate on database-resident data without using SQL and to run user-defined R functions, also referred to as "scripts",in one or more database-controlled R engines. OML4R is included with Oracle Database and Oracle Database Cloud Service.

Oracle Machine Learning for SQL (OML4SQL) provides SQL access to powerful, in-database machine learning algorithms. You can use OML4SQL to build and deploy predictive and descriptive machine learning models that can add intelligent capabilities to applications and dashboards. OML4SQL is included with Oracle Database, Oracle Database Cloud Service, and Oracle Autonomous Database.

Oracle Data Miner (ODMr) is an extension to Oracle SQL Developer. Oracle Data Miner is a graphical user interface to discover hidden patterns, relationships, and insights in data. ODMr provides a drag-and-drop workflow editor to define and capture the steps that users take to explore and prepare data and apply machine learning technology.

Machine Learning Techniques and Algorithms

Machine learning problems are categorized into mining techniques. Each machine learning function specifies a class of problems that can be modeled and solved. An algorithm is a mathematical procedure for solving a specific kind of problem.

Machine Learning Techniques

Each machine learning **technique** specifies a class of problems that can be modeled and solved.

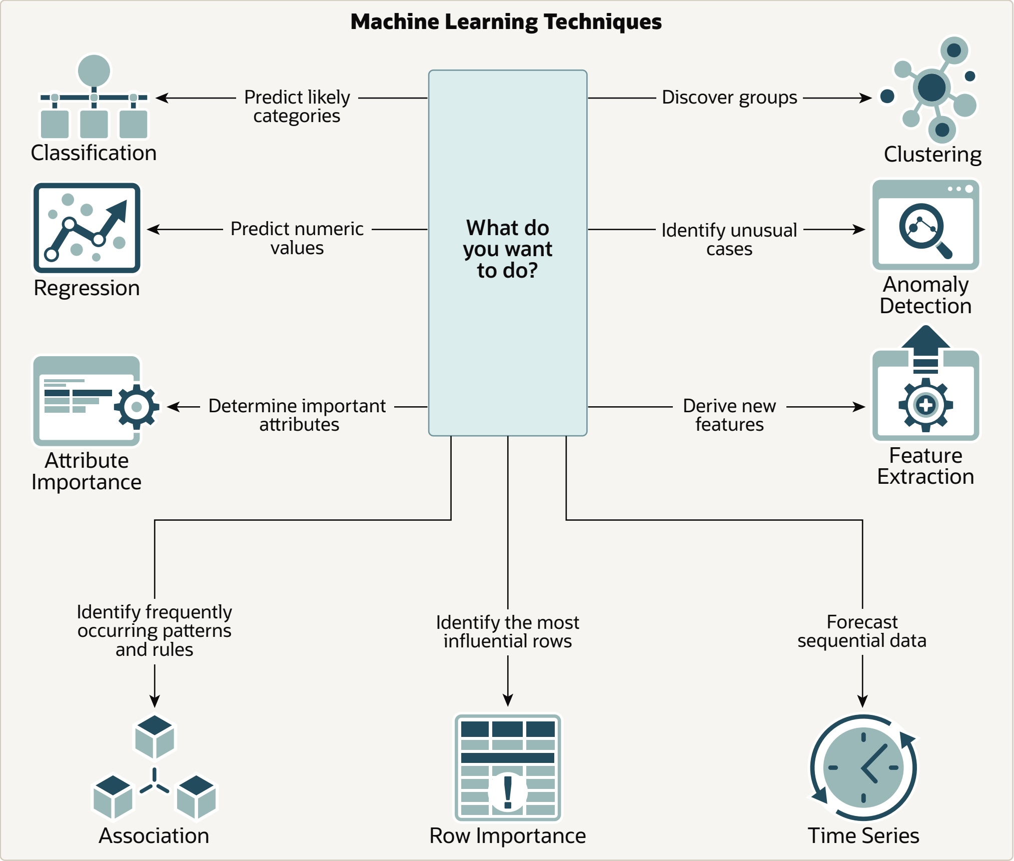
A basic understanding of machine learning techniques and algorithms is required for using Oracle Machine Learning.

Machine learning techniques fall generally into two categories: **supervised** and **unsupervised**. Notions of supervised and unsupervised learning are derived from the science of machine learning, which has been called a sub-area of artificial intelligence.

Artificial intelligence refers to the implementation and study of systems that exhibit autonomous intelligence or behavior of their own. Machine learning deals with techniques that enable devices to learn from their own performance and modify their own functioning.

The following illustration provides an idea of how to use Oracle machine learning techniques.

Figure 1-2 How to Use Machine Learning techniques



Related Topics

- What is a Machine Learning Algorithm
  An algorithm is a mathematical procedure for solving a specific kind of problem. For some machine learning techniques, you can choose among several algorithms.

Supervised Learning

Supervised learning is also known as directed learning. The learning process is directed by a previously known dependent attribute or target.

Supervised machine learning attempts to explain the behavior of the target as a function of a set of independent attributes or predictors. Supervised learning generally results in predictive models.

The building of a supervised model involves training, a process whereby the software analyzes many cases where the target value is already known. In the training process, the model "learns" the patterns in the data that enable making predictions. For example, a model that seeks to identify the customers who are likely to respond to a promotion must be trained by analyzing the characteristics of many customers who are known to have responded or not responded to a promotion in the past.

Oracle Machine Learning supports the following supervised machine learning functions:

Table 1-1 Supervised Machine Learning Functions

Function	Description	Sample Problem	Supported Algorithms
Feature Selection or Attribute Importance	Identifies the attributes that are most important in predicting a target attribute	Given customer response to an affinity card program, find the most significant predictors	• cur Matrix Decomposition • Expectation Maximization • Minimum Description Length
Classification	Assigns items to discrete classes and predicts the class to which an item belongs	Given demographic data about a set of customers, predict customer response to an affinity card program	• Decision Tree • Explicit Semantic Analysis • XGBoost • Generalized Linear Model • Naive Bayes • Neural Network • Random Forest • Support Vector Machine
Regression	Approximates and forecasts continuous values	Given demographic and purchasing data about a set of customers, predict customers' age	• XGBoost • Generalized Linear Model • Neural Network • Support Vector Machine
Ranking	Predicts the probability of one item over other items	Recommend products to online customers based on their browsing history	XGBoost
Time Series	Forecasts target value based on known history of target values taken at equally spaced points in time	Predict the length of the ocean waves, address tactical issues such as projecting costs, inventory requirements and customer satisfaction, and so on.	Exponential Smoothing

Unsupervised Learning

Unsupervised learning is non-directed. There is no distinction between dependent and independent attributes. There is no previously-known result to guide the algorithm in building the model.

Unsupervised learning can be used for descriptive purposes. In unsupervised learning, the goal is pattern detection. It can also be used to make predictions.

Oracle Machine Learning supports the following unsupervised machine learning functions:

Table 1-2 Unsupervised Machine Learning Functions

Function	Description	Sample Problem	Supported Algorithms
Anomaly Detection	Identifies rows (cases, examples) that do not satisfy the characteristics of "normal" data	Given demographic data about a set of customers, identify which customer purchasing behaviors are unusual in the dataset, which may be indicative of fraud.	• One-Class SVM • Multivariate State Estimation Technique - Sequential Probability Ratio Test
Association	Finds items that tend to co-occur in the data and specifies the rules that govern their co-occurrence	Find the items that tend to be purchased together and specify their relationship	Apriori
Clustering	Finds natural groupings in the data	Segment demographic data into clusters and rank the probability that an individual belongs to a given cluster	• Expectation Maximization • k-Means • O-Cluster
Feature Extraction	Creates new attributes (features) using linear combinations of the original attributes	Given demographic data about a set of customers, transform the original attributes into fewer new attributes.	• Explicit Semantic Analysis • Non-Negative Matrix Factorization • PCA scoring • Singular Value Decomposition
Row Importance	Row importance technique is used in dimensionality reduction of large data sets. Row importance identifies the most influential rows of the data set.	Given a data set, select rows that meet a minimum importance value prior to model building.	cur Matrix Decomposition

What is a Machine Learning Algorithm

An algorithm is a mathematical procedure for solving a specific kind of problem. For some machine learning techniques, you can choose among several algorithms.

Each algorithm produces a specific type of model, with different characteristics. Some machine learning problems can best be solved by using more than one algorithm in combination. For example, you might first use a feature extraction model to create an optimized set of predictors, then a classification model to make a prediction on the results.

About Unstructured Text

Unstructured text may contain important information that is critical to the success of a business.

Machine learning algorithms act on data that is numerical or categorical. Numerical data is ordered. It is stored in columns that have a numeric data type, such as `NUMBER` or `FLOAT`. Categorical data is identified by category or classification. It is stored in columns that have a character data type, such as `VARCHAR2` or `CHAR`.

Unstructured text data is neither numerical nor categorical. Unstructured text includes items such as web pages, document libraries, Power Point presentations, product specifications, emails, comment fields in reports, and call center notes. It has been said that unstructured text accounts for more than three quarters of all enterprise data. Extracting meaningful information from unstructured text can be critical to the success of a business.

About Machine Learning and Oracle Text

Understand machine learning operations on text and Oracle Text.

Machine learning operations on text is the process of applying machine learning techniques to text terms, also called text features or tokens. Text terms are words or groups of words that have been extracted from text documents and assigned numeric weights. Text terms are the fundamental unit of text that can be manipulated and analyzed.

Oracle Text is an Oracle Database technology that provides term extraction, word and theme searching, and other utilities for querying text. When columns of text are present in the training data, Oracle Machine Learning for SQL uses Oracle Text utilities and term weighting strategies to transform the text for machine learning operations. OML4SQL passes configuration information supplied by you to Oracle Text and uses the results in the model creation process.

Related Topics

- *Oracle Text Application Developer's Guide*

About Partitioned Models

Introduces partitioned models to organize and represent multiple models.

When you build a model on your data set and apply it to new data, sometimes the prediction may be generic that performs badly when run on new and evolving data. To overcome this, the data set can be divided into different parts based on some characteristics. Oracle Machine Learning for SQL supports partitioned model. Partitioned models allow users to build a type of ensemble model for each data partition. The top-level model has sub models that are automatically produced. The sub models are based on the attribute options. For example, if your data set has an attribute called `REGION` with four values and you have defined it as the partitioned attribute. Then, four sub models are created for this attribute. The sub models are automatically managed and used as a single model. The partitioned model automates a typical machine learning task and can potentially achieve better accuracy through multiple targeted models.

The partitioned model and its sub models reside as first class, persistent database objects. Persistent means that the partitioned model has an on-disk representation. In a partition model, the performance of partitioned models with a large number of partitions is enhanced, and dropping a single model within a partition model is also improved.

To create a partitioned model, include the `ODMS_PARTITION_COLUMNS` setting. To define the number of partitions, include the `ODMS_MAX_PARTITIONS` setting. When you are making predictions, you must use the top-level model. The correct sub model is selected automatically based on the attribute, the attribute options, and the partition setting. You must include the partition columns as part of the `USING` clause when scoring. The `GROUPING` hint is an optional hint that applies to machine learning scoring functions when scoring partitioned models.

The partition names, key values, and the structure of the partitioned model are available in the `ALL_MINING_MODEL_PARTITIONS` view.

Related Topics

- *Oracle Database Reference*

> **See Also:**
>
> *Oracle Database SQL Language Reference* on how to use `GROUPING` hint.
> *Oracle Machine Learning for SQL User’s Guide* to understand more about partitioned models.

Partitioned Model Build Process

To build a partitioned model, Oracle Machine Learning for SQL requires a partitioning key specified in a settings table.

The partitioning key is a comma-separated list of one or more columns (up to 16) from the input data set. The partitioning key horizontally slices the input data based on discrete values of the partitioning key. That is, partitioning is performed as list values as opposed to range partitioning against a continuous value. The partitioning key supports only columns of the data type `NUMBER` and `VARCHAR2`.

During the build process the input data set is partitioned based on the distinct values of the specified key. Each data slice (unique key value) results in its own model partition. The resultant model partition is not separate and is not visible to you as a standalone model. The default value of the maximum number of partitions for partitioned models is 1000 partitions. You can also set a different maximum partitions value. If the number of partitions in the input data set exceeds the defined maximum, Oracle Machine Learning for SQL throws an exception.

The partitioned model organizes features common to all partitions and the partition specific features. The common features consist of the following metadata:

- The model name
- The machine learning function
- The machine learning algorithm
- A super set of all machine learning model attributes referenced by all partitions (signature)
- A common set of user-defined column transformations
- Any user-specified or default build settings that are interpreted as global; for example, the Auto Data Preparation (ADP) setting

DDL in Partitioned model

Learn about maintenance of partitioned models thorough DDL operations.

Partitioned models are maintained through the following DDL operations:

- Drop model or drop partition
- Add partition

Drop Model or Drop Partition

Oracle Machine Learning for SQL supports dropping a single model partition for a given partition name.

If only a single partition remains, you cannot explicitly drop that partition. Instead, you must either add additional partitions prior to dropping the partition or you may choose to drop the model itself. When dropping a partitioned model, all partitions are dropped in a single atomic operation. From a performance perspective, Oracle recommends DROP_PARTITION followed by an ADD_PARTITION instead of leveraging the REPLACE option due to the efficient behavior of the DROP_PARTITION option.

Add Partition

Oracle Machine Learning for SQL supports adding a single partition or multiple partitions to an existing partitioned model.

The addition occurs based on the input data set and the name of the existing partitioned model. The operation takes the input data set and the existing partitioned model as parameters. The partition keys are extracted from the input data set and the model partitions are built against the input data set. These partitions are added to the partitioned model. In the case where partition keys for new partitions conflict with the existing partitions in the model, you can select from the following three approaches to resolve the conflicts:

- ERROR: Terminates the ADD operation without adding any partitions.
- REPLACE: Replaces the existing partition for which the conflicting keys are found.
- IGNORE: Eliminates the rows having the conflicting keys.

If the input data set contains multiple keys, then the operation creates multiple partitions. If the total number of partitions in the model increases to more than the user-defined maximum specified when the model was created, then you get an error. The default threshold value for the number of partitions is 1000.

Partitioned Model Scoring

The scoring of the partitioned model is the same as that of the non-partitioned model.

The syntax of the machine learning function remains the same but is extended to provide an optional hint. The optional hint can impact the performance of a query which involves scoring a partitioned model.

For scoring a partitioned model, the signature columns used during the build for the partitioning key must be present in the scoring data set. These columns are combined to form a unique partition key. The unique key is then mapped to a specific underlying model partition, and the identified model partition is used to score that row.

The partitioned objects that are necessary for scoring are loaded on demand during the query execution and are aged out depending on the System Global Area (SGA) memory.

In this example an SVM model is used to predict the number of years a customer resides at their residence but partitioned on customer gender. The model is then used to predict the target. This example highlights the model settings that you can define when you create a partitioned model. The following example is using a view created from the SH schema tables.

The CREATE_MODEL2 procedure is used for creating the model. The partition attribute is CUST_GENDER. This attribute has two options *M* and *F*.

```
%script
BEGIN DBMS_DATA_MINING.DROP_MODEL('SVM_MOD_PARTITIONED');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME'):= 'ALGO_SUPPORT_VECTOR_MACHINES';
    v_setlst('SVMS_KERNEL_FUNCTION')  :='SVMS_LINEAR';
    v_setlst('ODMS_PARTITION_COLUMNS'):='CUST_GENDER';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'SVM_MOD_PARTITIONED',
        MINING_FUNCTION     => 'REGRESSION',
        DATA_QUERY          => 'SELECT * FROM CUSTOMERS_DEMO',
        SET_LIST            => v_setlst,
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        TARGET_COLUMN_NAME  => 'YRS_RESIDENCE');
END;
```

The output is as follows:

```
PL/SQL procedure successfully completed.


---------------------------

PL/SQL procedure successfully completed.
```

The following code sample shows the prediction.

```
%script

SELECT cust_id, YRS_RESIDENCE,
       ROUND(PREDICTION(SVM_MOD_PARTITIONED USING *),2) pred_YRS_RESIDENCE
FROM CUSTOMERS_DEMO;
```

```
CUST_ID   YRS_RESIDENCE   PRED_YRS_RESIDENCE
   100100               4                 4.71
   100200               2                 1.62
   100300               4                 4.66
   100400               6                  5.9
   100500               2                 2.07
   100600               3                 2.74
   100700               6                 5.78
   100800               5                 7.22
   100900               4                 4.88
```

```
101000               7                 6.49
101100               4                 3.54
101200               1                 1.46
101300               4                 4.34
101400               4                 4.34 ...
```

Related Topics

- *Oracle Database SQL Language Reference*

Automatic Data Preparation

Most algorithms require some form of data transformation. During the model build process, Oracle Machine Learning for SQL can automatically perform the transformations required by the algorithm.

You can choose to supplement the automatic transformations with additional transformations of your own, or you can choose to manage all the transformations yourself.

In calculating automatic transformations, Oracle Machine Learning for SQL uses heuristics that address the common requirements of a given algorithm. This process results in reasonable model quality in most cases.

Binning and normalization are transformations that are commonly needed by machine learning algorithms.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Binning

Binning, also called discretization, is a technique for reducing the cardinality of continuous and discrete data. Binning groups related values together in bins to reduce the number of distinct values.

Binning can improve resource utilization and model build response time dramatically without significant loss in model quality. Binning can improve model quality by strengthening the relationship between attributes.

Supervised binning is a form of intelligent binning in which important characteristics of the data are used to determine the bin boundaries. In supervised binning, the bin boundaries are identified by a single-predictor decision tree that takes into account the joint distribution with the target. Supervised binning can be used for both numerical and categorical attributes.

Normalization

Learn about normalization.

Normalization is the most common technique for reducing the range of numerical data. Most normalization methods map the range of a single variable to another range (often 0,1).

How ADP Transforms the Data

The following table shows how ADP prepares the data for each algorithm.

Table 1-3 Oracle Machine Learning Algorithms With ADP

Algorithm	Machine Learning Function	Treatment by ADP
Apriori	Association rules	ADP has no effect on association rules.
CUR Matrix Decomposition	Feature selection	ADP has no effect on CUR Matrix Decomposition
Decision Tree	Classification	ADP has no effect on Decision Tree. Data preparation is handled by the algorithm.
Expectation Maximization	Clustering	Single-column (not nested) numerical columns that are modeled with Gaussian distributions are normalized. ADP has no effect on the other types of columns.
GLM	Classification and regression	Numerical attributes are normalized.
k-Means	Clustering	Numerical attributes are normalized.
MDL	Attribute importance	All attributes are binned with supervised binning.
MSET-SPRT	Classification (for anomaly detection)	Z-score normalization is performed.
Naive Bayes	Classification	All attributes are binned with supervised binning.
Neural Network	Classification and regression	Numerical attributes are normalized.
NMF	Feature extraction	Numerical attributes are normalized.
O-Cluster	Clustering	Numerical attributes are binned with a specialized form of equi-width binning, which computes the number of bins per attribute automatically. Numerical columns with all nulls or a single value are removed.
Random Forest	Classification	ADP has no effect on Random Forest. Data preparation is handled by the algorithm.
SVD	Feature extraction	Numeric attributes are centered if PCA is selected.
SVM	Classification, anomaly detection, and regression	Numerical attributes are normalized.
XG Boost	Classification and regression	ADP has no effect on XG Boost.

See Also:

- *Oracle Database PL/SQL Packages and Types Reference*
- Part III, Algorithms, in *Oracle Machine Learning for SQL Concepts* for more information about algorithm-specific data preparation

Missing Value Treatment in Oracle Machine Learning for SQL

Summarizes the treatment of missing values in Oracle Machine Learning for SQL.

Missing value treatment depends on the algorithm and on the nature of the data (categorical or numerical, sparse or missing at random). Missing value treatment is summarized in the following table.

Note:

Oracle Machine Learning for SQL performs the same missing value treatment whether or not you are using Automatic Data Preparation (ADP).

Table 1-4 Missing Value Treatment by Algorithm

Missing Data	EM, GLM, NMF, k-Means, SVD, SVM	DT, MDL, NB, OC	Apriori
NUMERICAL missing at random	The algorithm replaces missing numerical values with the mean. For Expectation Maximization (EM), the replacement only occurs in columns that are modeled with Gaussian distributions.	The algorithm handles missing values naturally as missing at random.	The algorithm interprets all missing data as sparse.
CATEGORICAL missing at random	Generalized Linear Model (GLM), Non-Negative Matrix Factorization (NMF), *k*-Means, and Support Vector Machine (SVM) replaces missing categorical values with the mode. Singular Value Decomposition (SVD) does not support categorical data. EM does not replace missing categorical values. EM treats NULLs as a distinct value with its own frequency count.	The algorithm handles missing values naturally as missing random.	The algorithm interprets all missing data as sparse.
NUMERICAL sparse	The algorithm replaces sparse numerical data with zeros.	O-Cluster does not support nested data and therefore does not support sparse data. Decision Tree (DT), Minimum Description Length (MDL), and Naive Bayes (NB) replace sparse numerical data with zeros.	The algorithm handles sparse data.
CATEGORICAL sparse	All algorithms except SVD replace sparse categorical data with zero vectors. SVD does not support categorical data.	O-Cluster does not support nested data and therefore does not support sparse data. DT, MDL, and NB replace sparse categorical data with the special value `DM$SPARSE`.	The algorithm handles sparse data.

Data Preparation

Data preparation involves cleaning, transforming, and organizing data for building effective machine learning models. Quality data is essential for accurate model predictions.

The quality of a model depends to a large extent on the quality of the data used to build (train) it. Much of the time spent in any given machine learning project is devoted to data preparation.

The data must be carefully inspected, cleansed, and transformed, and algorithm-appropriate data preparation methods must be applied.

The process of data preparation is further complicated by the fact that any data to which a model is applied, whether for testing or for scoring, must undergo the same transformations as the data used to train the model.

Simplify Data Preparation with Oracle Machine Learning for SQL

Oracle Machine Learning for SQL (OML4SQL) provides inbuilt data preparation, automatic data preparation, custom data preparation through the `DBMS_DATA_MINING_TRANSFORM` PL/SQL package, model details, and employs consistent approach across machine learning algorithms to manage missing and sparse data.

OML4SQL offers several features that significantly simplify the process of data preparation:

- Embedded data preparation: The transformations used in training the model are embedded in the model and automatically run whenever the model is applied to new data. If you specify transformations for the model, you only have to specify them once.
- Automatic Data Preparation (ADP): Oracle Machine Learning for SQL supports an automated data preparation mode. When ADP is active, Oracle Machine Learning for SQL automatically performs the data transformations required by the algorithm. The transformation instructions are embedded in the model along with any user-specified transformation instructions.
- Automatic management of missing values and sparse data: Oracle Machine Learning for SQL uses consistent methodology across machine learning algorithms to handle sparsity and missing values.
- Transparency: Oracle Machine Learning for SQL provides model details, which are a view of the attributes that are internal to the model. This insight into the inner details of the model is possible because of reverse transformations, which map the transformed attribute values to a form that can be interpreted by a user. Where possible, attribute values are reversed to the original column values. Reverse transformations are also applied to the target of a supervised model, thus the results of scoring are in the same units as the units of the original target.
- Tools for custom data preparation: Oracle Machine Learning for SQL provides many common transformation routines in the `DBMS_DATA_MINING_TRANSFORM` PL/SQL package. You can use these routines, or develop your own routines in SQL, or both. The SQL language is well suited for implementing transformations in the database. You can use custom transformation instructions along with ADP or instead of ADP.

Case Data

Case data organizes information in single-record rows for each case, essential for most machine learning algorithms in Oracle Machine Learning for SQL.

Most machine learning algorithms act on single-record case data, where the information for each case is stored in a separate row. The data attributes for the cases are stored in the columns.

When the data is organized in transactions, the data for one case (one transaction) is stored in many rows. An example of transactional data is market basket data. With the single exception of Association Rules, which can operate on native transactional data, Oracle Machine Learning for SQL algorithms require single-record case organization.

Nested Data

Nested data supports attributes in nested columns, enabling effective mining of complex data structures and multiple sources.

Oracle Machine Learning for SQL supports attributes in nested columns. A transactional table can be cast as a nested column and included in a table of single-record case data. Similarly, star schemas can be cast as nested columns. With nested data transformations, Oracle Machine Learning for SQL can effectively mine data originating from multiple sources and configurations.

Text Data

Text data involves transforming unstructured text into numeric values for analysis, utilizing Oracle Text utilities and configurable transformations.

Oracle Machine Learning for SQL interprets `CLOB` columns and long `VARCHAR2` columns automatically as unstructured text. Additionally, you can specify columns of short `VARCHAR2`, `CHAR`, `BLOB`, and `BFILE` as unstructured text. Unstructured text includes data items such as web pages, document libraries, Power Point presentations, product specifications, emails, comment fields in reports, and call center notes.

Oracle Machine Learning for SQL uses Oracle Text utilities and term weighting strategies to transform unstructured text for analysis. In text transformation, text terms are extracted and given numeric values in a text index. The text transformation process is configurable for the model and for individual attributes. Once transformed, the text can by mined with a Oracle Machine Learning for SQL algorithm.

Related Topics

- Prepare the Data
- Machine Learning Operations on Unstructured Text

Data Requirements

Understand how data is stored and viewed for Oracle Machine Learning.

Machine learning activities require data that is defined within a single table or view. The information for each record must be stored in a separate row. The data records are commonly called **cases**. Each case can optionally be identified by a unique **case ID**. The table or view itself can be referred to as a **case table**.

The `CUSTOMERS` table in the `SH` schema is an example of a table that could be used for machine learning. All the information for each customer is contained in a single row. The case ID is the `CUST_ID` column. The rows listed in the following example are selected from `SH.CUSTOMERS`.

Note:

Oracle Machine Learning requires single-record case data for all types of models except association models, which can be built on native transactional data.

Example 1-1 Sample Case Table

```
select cust_id, cust_gender, cust_year_of_birth,
          cust_main_phone_number from sh.customers where cust_id < 11;
```

The output is as follows:

```
CUST_ID CUST_GENDER CUST_YEAR_OF_BIRTH CUST_MAIN_PHONE_NUMBER
------- ----------- ---- ------------- -------------------------
1        M               1946          127-379-8954
2        F               1957          680-327-1419
3        M               1939          115-509-3391
4        M               1934          577-104-2792
5        M               1969          563-667-7731
6        F               1925          682-732-7260
7        F               1986          648-272-6181
8        F               1964          234-693-8728
9        F               1936          697-702-2618
10       F               1947          601-207-4099
```

Related Topics

- Use Market Basket Data
  Understand the use of association and Apriori for market basket analysis.

Column Data Types

Understand the different types of column data in a case table.

The columns of the case table hold the attributes that describe each case. In Example 1-1, the attributes are: `CUST_GENDER`, `CUST_YEAR_OF_BIRTH`, and `CUST_MAIN_PHONE_NUMBER`. The attributes are the predictors in a supervised model or the descriptors in an unsupervised model. The case ID, `CUST_ID`, can be viewed as a special attribute; it is not a predictor or a descriptor.

Oracle Machine Learning for SQL supports standard Oracle data types except `DATE`, `TIMESTAMP`, `RAW`, and `LONG`. Oracle Machine Learning supports date type (datetime, date, timestamp) for case_id, `CLOB`/`BLOB`/`FILE` that are interpreted as text columns, and the following collection types as well:

```
DM_NESTED_CATEGORICALS
DM_NESTED_NUMERICALS
DM_NESTED_BINARY_DOUBLES
DM_NESTED_BINARY_FLOATS
```

Note:

The attributes with the data type `BOOLEAN` are treated as numeric with the following values: `TRUE` means `1`, `FALSE` means `0`, and `NULL` is interpreted as an unknown value. The `CASE_ID_COLUMN_NAME` attribute does not support `BOOLEAN` data type.

Related Topics

- Use Nested Data
  A join between the tables for one-to-many relationship is represented through nested columns.
- #unique_47
- *Oracle Database SQL Language Reference*

Vector Data Type

You can provide `VECTOR` data as input to Oracle Machine Learning in-database algorithms to complement other structured data or be used alone. The vector data type is supported for clustering, classification, anomaly detection, and feature extraction.

While dense vectors with arbitrary precision and dimensions are supported, in a flex vector column, precision may differ and dimensions within a single vector column must remain consistent. Errors are raised for mismatched dimensions.

Partitioned models track vector dimensions alongside partition statistics. Different partitions can have different vector dimensions, however, dimensions must remain consistent within a single partition. Errors are raised for mismatched dimensions within a single partition.

The system supports `FLOAT32`, `FLOAT64`, and `INT8` as datatypes. Vectors with `FLEX` dimension and precision are supported. These features can be used in combination with the other data types supported by OML (numerical, categorical, nested, and text).

Scoring with Vectors

The system treats each vector dimension as an individual predictor and provides model details at the vector component level, labeled as `DM$$VECxxx`, where `xxx` represents the component's position. For example, `DM$$VEC1`. During scoring, the system matches vector dimensions between the model and input data at compile time or runtime, raising errors if mismatches occur. A vector cannot be a target or a `case_id` column, errors are raised if you set vector as a target or `case_id`.

The system does not support:

- analytic scoring with vectors, analytic scoring operators skip the vector inputs without displaying any error.
- sparse vectors, raises an error that the format is not supported if sparse vectors are identified. To learn more about sparse vectors, see Create Tables Using the VECTOR Data Type.
- binary vector precision, raises an error that the format is not supported

OML supports the vector data type for the following algorithms and the scoring operators:

Technique	Algorithms	Scoring Operator
Classification or Regression	SVM, Neural Network, GLM	`PREDICTION`, `PREDICTION_PROBABILITY`, `PREDICTION_SET`, `PREDICTION_BOUNDS`
Anomaly Detection	One-class SVM, Expectation Maximization	`PREDICTION`, `PREDICTION_PROBABILITY`, `PREDICTION_SET`

Technique	Algorithms	Scoring Operator
Clustering	*k*-Means, Expectation Maximization	CLUSTER_ID, CLUSTER_PROBABILITY, CLUSTER_SET, CLUSTER_DISTANCE
Feature Extraction	SVD, PCA	FEATURE_ID, FEATURE_VALUE, FEATURE_SET, VECTOR_EMBEDDING

See Example: Using Vector Data for Dimensionality Reduction and Clustering for more details.

Scoring Requirements

Learn how scoring is done in Oracle Machine Learning for SQL.

Most machine learning models can be applied to separate data in a process known as **scoring**. Oracle Machine Learning for SQL supports the scoring operation for classification, regression, anomaly detection, clustering, and feature extraction.

The scoring process matches column names in the scoring data with the names of the columns that were used to build the model. The scoring process does not require all the columns to be present in the scoring data. If the data types do not match, Oracle Machine Learning for SQL attempts to perform type coercion. For example, if a column called PRODUCT_RATING is VARCHAR2 in the training data but NUMBER in the scoring data, Oracle Machine Learning for SQL effectively applies a TO_CHAR() function to convert it.

The column in the test or scoring data must undergo the same transformations as the corresponding column in the build data. For example, if the AGE column in the build data was transformed from numbers to the values CHILD, ADULT, and SENIOR, then the AGE column in the scoring data must undergo the same transformation so that the model can properly evaluate it.

Note:

Oracle Machine Learning for SQL can embed user-specified transformation instructions in the model and reapply them whenever the model is applied. When the transformation instructions are embedded in the model, you do not need to specify them for the test or scoring data sets.

Oracle Machine Learning for SQL also supports Automatic Data Preparation (ADP). When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model along with any user-specified transformations.

See Also:

Automatic Data Preparation and Embed Transformations in a Model for more information on automatic and embedded data transformations

About Attributes

Attributes are the items of data that are used in machine learning. Attributes are also referred as variables, fields, or predictors.

In predictive models, attributes are the predictors that affect a given outcome. In descriptive models, attributes are the items of information being analyzed for natural groupings or associations. For example, a table of employee data that contains attributes such as job title, date of hire, salary, age, gender, and so on.

Data Attributes and Model Attributes

Data attributes are columns in the data set used to build, test, or score a model. **Model attributes** are the data representations used internally by the model.

Data attributes and model attributes can be the same. For example, a column called `SIZE`, with values `S`, `M`, and `L`, are attributes used by an algorithm to build a model. Internally, the model attribute `SIZE` is most likely be the same as the data attribute from which it was derived.

On the other hand, a nested column `SALES_PROD`, containing the sales figures for a group of products, does not correspond to a model attribute. The data attribute can be `SALES_PROD`, but each product with its corresponding sales figure (each row in the nested column) is a model attribute.

Transformations also cause a discrepancy between data attributes and model attributes. For example, a transformation can apply a calculation to two data attributes and store the result in a new attribute. The new attribute is a model attribute that has no corresponding data attribute. Other transformations such as binning, normalization, and outlier treatment, cause the model's representation of an attribute to be different from the data attribute in the case table.

Related Topics

- Use Nested Data
  A join between the tables for one-to-many relationship is represented through nested columns.
- Embed Transformations in a Model
  You can specify your own transformations and embed them in a model by creating a transformation list and passing it to `DBMS_DATA_MINING.CREATE_MODEL2` or `DBMS_DATA_MINING.CREATE_MODEL`.

Target Attribute

Understand what a **target** means in machine learning and understand the different target data types.

The **target** of a supervised model is a special kind of attribute. The target column in the training data contains the historical values used to train the model. The target column in the test data contains the historical values to which the predictions are compared. The act of scoring produces a prediction for the target.

Clustering, feature extraction, association, and anomaly detection models do not use a target.

Nested columns and columns of unstructured data (such as `BFILE`, `CLOB`, or `BLOB`) cannot be used as targets.

Table 1-5 Target Data Types

Machine Learning Function	Target Data Types
Classification	VARCHAR2, CHAR NUMBER, FLOAT BINARY_DOUBLE, BINARY_FLOAT, ORA_MINING_VARCHAR2_NT BOOLEAN
Regression	NUMBER, FLOAT BINARY_DOUBLE, BINARY_FLOAT

You can query the *_MINING_MODEL_ATTRIBUTES view to find the target for a given model.

Related Topics

- ALL_MINING_MODEL_ATTRIBUTES
  Describes an example of ALL_MINING_MODEL_ATTRIBUTES and shows a sample query.
- *Oracle Database PL/SQL Packages and Types Reference*

Numericals, Categoricals, and Unstructured Text

Explains numeric, categorical, and unstructured text attributes.

Model attributes are numerical, categorical, or unstructured (text). Data attributes, which are columns in a case table, have Oracle data types, as described in "Column Data Types".

Numerical attributes can theoretically have an infinite number of values. The values have an implicit order, and the differences between them are also ordered. Oracle Machine Learning for SQL interprets NUMBER, FLOAT, BINARY_DOUBLE, BINARY_FLOAT, BOOLEAN, DM_NESTED_NUMERICALS, DM_NESTED_BINARY_DOUBLES, and DM_NESTED_BINARY_FLOATS as numerical.

Categorical attributes have values that identify a finite number of discrete categories or classes. There is no implicit order associated with the values. Some categoricals are binary: they have only two possible values, such as yes or no, or male or female. Other categoricals are multi-class: they have more than two values, such as small, medium, and large.

Oracle Machine Learning for SQL interprets CHAR and VARCHAR2 as categorical by default, however these columns may also be identified as columns of unstructured data (text). Oracle Machine Learning for SQL interprets columns of DM_NESTED_CATEGORICALS as categorical. Columns of CLOB, BLOB, and BFILE always contain unstructured data.

The target of a classification model is categorical. (If the target of a classification model is numeric, it is interpreted as categorical.) The target of a regression model is numerical. The target of an attribute importance model is either categorical or numerical.

Related Topics

- Column Data Types
  Understand the different types of column data in a case table.
-

Model Signature

Learn about model signature and the data types that are considered in the build data.

The model signature is the set of data attributes that are used to build a model. Some or all of the attributes in the signature must be present for scoring. The model accounts for any missing columns on a best-effort basis. If columns with the same names but different data types are present, the model attempts to convert the data type. If extra, unused columns are present, they are disregarded.

The model signature does not necessarily include all the columns in the build data. Algorithm-specific criteria can cause the model to ignore certain columns. Other columns can be eliminated by transformations. Only the data attributes actually used to build the model are included in the signature.

The target and case ID columns are not included in the signature.

Scoping of Model Attribute Name

Learn about model attribute name.

The model attribute name consists of two parts: a column name, and a subcolumn name.

```
column_name[.subcolumn_name]
```

The `column_name` component is the name of the data attribute. It is present in all model attribute names. Nested attributes and text attributes also have a `subcolumn_name` component as shown in the following example.

Example 1-2   Model Attributes Derived from a Nested Column

The nested column `SALESPROD` has three rows.

```
SALESPROD(ATTRIBUTE_NAME, VALUE)
--------------------------------
((PROD1, 300),
 (PROD2, 245),
 (PROD3, 679))
```

The name of the data attribute is `SALESPROD`. Its associated model attributes are:

```
SALESPROD.PROD1
SALESPROD.PROD2
SALESPROD.PROD3
```

Model Details

Model details reveal information about model attributes and their treatment by the algorithm. Oracle recommends that users leverage the model detail views for the respective algorithm.

Transformation and reverse transformation expressions are associated with model attributes. Transformations are applied to the data attributes before the algorithmic processing that creates the model. Reverse transformations are applied to the model attributes after the model has been built, so that the model details are expressed in the form of the original data attributes, or as close to it as possible.

Reverse transformations support model transparency. They provide a view of the data that the algorithm is working with internally but in a format that is meaningful to a user.

Deprecated `GET_MODEL_DETAILS`

There is a separate `GET_MODEL_DETAILS` routine for each algorithm. Starting from Oracle Database 12*c* Release 2, the `GET_MODEL_DETAILS` are deprecated. Oracle recommends to use Model Detail Views for the respective algorithms.

Related Topics

- Model Detail Views

Use Nested Data

A join between the tables for one-to-many relationship is represented through nested columns.

Oracle Machine Learning for SQL requires a case table in single-record case format, with each record in a separate row. What if some or all of your data is in multi-record case format, with each record in several rows? What if you want one attribute to represent a series or collection of values, such as a student's test scores or the products purchased by a customer?

This kind of one-to-many relationship is usually implemented as a join between tables. For example, you can join your customer table to a sales table and thus associate a list of products purchased with each customer.

Oracle Machine Learning for SQL supports dimensioned data through nested columns. To include dimensioned data in your case table, create a view and cast the joined data to one of the machine learning nested table types. Each row in the nested column consists of an attribute name/value pair. Oracle Machine Learning for SQL internally processes each nested row as a separate attribute.

> **Note:**
>
> O-Cluster is the only algorithm that does not support nested data.

Related Topics

- Example: Creating a Nested Column for Market Basket Analysis
  The example shows how to define a nested column for market basket analysis.

Nested Object Types

Nested tables are object data types that can be used in place of other data types.

Oracle Database supports user-defined data types that make it possible to model real-world entities as objects in the database. **Collection types** are object data types for modeling multi-valued attributes. Nested tables are collection types. Nested tables can be used anywhere that other data types can be used.

OML4SQL supports the following nested object types:

```
DM_NESTED_BINARY_DOUBLES
DM_NESTED_BINARY_FLOATS
DM_NESTED_NUMERICALS
DM_NESTED_CATEGORICALS
```

Descriptions of the nested types are provided in this example.

Example 1-3 OML4SQL Nested Data Types

```
describe dm_nested_binary_double
 Name                                      Null?    Type
 ----------------------------------------- --------
```

```
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              BINARY_DOUBLE

describe dm_nested_binary_doubles
 DM_NESTED_BINARY_DOUBLES TABLE OF SYS.DM_NESTED_BINARY_DOUBLE
 Name                                       Null?    Type
 ------------------------------------------ --------
---------------------------
 ATTRIBUTE_NAME                                      VARCHAR2(4000)
 VALUE                                               BINARY_DOUBLE

describe dm_nested_binary_float
 Name                                      Null?    Type
 ----------------------------------------- --------
---------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              BINARY_FLOAT

describe dm_nested_binary_floats
 DM_NESTED_BINARY_FLOATS TABLE OF SYS.DM_NESTED_BINARY_FLOAT
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              BINARY_FLOAT

describe dm_nested_numerical
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              NUMBER

describe dm_nested_numericals
 DM_NESTED_NUMERICALS TABLE OF SYS.DM_NESTED_NUMERICAL
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              NUMBER

describe dm_nested_categorical
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              VARCHAR2(4000)

describe dm_nested_categoricals
 DM_NESTED_CATEGORICALS TABLE OF SYS.DM_NESTED_CATEGORICAL
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 ATTRIBUTE_NAME                                     VARCHAR2(4000)
 VALUE                                              VARCHAR2(4000)
```

Related Topics

- *Oracle Database Object-Relational Developer's Guide*

Example: Transforming Transactional Data for Machine Learning

In this example, a comparison is shown for sale of products in four regions with data before transformation and then after transformation.

Example 1-4 shows data from a view of a sales table. It includes sales for three of the many products sold in four regions. This data is not suitable for machine learning at the product level because sales for each case (product), is stored in several rows.

Example 1-5 shows how this data can be transformed for machine learning. The case ID column is `PRODUCT`. `SALES_PER_REGION`, a nested column of type `DM_NESTED_NUMERICALS`, is a data attribute. This table is suitable for machine learning at the product case level, because the information for each case is stored in a single row.

Oracle Machine Learning for SQL treats each nested row as a separate model attribute, as shown in Example 1-6.

> **Note:**
>
> The presentation in this example is conceptual only. The data is not actually pivoted before being processed.

Example 1-4 Product Sales per Region in Multi-Record Case Format

```
PRODUCT   REGION         SALES
-------   --------    ----------
Prod1       NE            556432
Prod2       NE            670155
Prod3       NE              3111
.
.
Prod1       NW             90887
Prod2       NW            100999
Prod3       NW            750437
.
.
Prod1       SE             82153
Prod2       SE             57322
Prod3       SE             28938
.
.
Prod1       SW           3297551
Prod2       SW           4972019
Prod3       SW            884923
.
.
```

Example 1-5   Product Sales per Region in Single-Record Case Format

```
PRODUCT      SALES_PER_REGION
          (ATTRIBUTE_NAME, VALUE)
------    --------------------------
Prod1       ('NE' ,     556432)
            ('NW' ,      90887)
            ('SE' ,      82153)
            ('SW' ,    3297551)
Prod2       ('NE' ,     670155)
            ('NW' ,     100999)
            ('SE' ,      57322)
            ('SW' ,    4972019)
Prod3       ('NE' ,       3111)
            ('NW' ,     750437)
            ('SE' ,      28938)
            ('SW' ,     884923)
.
.
```

Example 1-6   Model Attributes Derived From SALES_PER_REGION

```
PRODUCT    SALES_PER_REGION.NE    SALES_PER_REGION.NW    SALES_PER_REGION.SE
SALES_PER_REGION.SW
-------    ------------------    -------------------    ------------------
-------------------
Prod1                556432                90887                  82153
3297551
Prod2                670155               100999                  57322
4972019
Prod3                  3111                750437                 28938
884923
.
.
```

Handle Missing Values

Understand sparse data and missing values.

Oracle Machine Learning for SQL distinguishes between **sparse data** and data that contains **random missing values**. The latter means that some attribute values are unknown. Sparse data, on the other hand, contains values that are assumed to be known, although they are not represented in the data.

A typical example of sparse data is market basket data. Out of hundreds or thousands of available items, only a few are present in an individual case (the basket or transaction). All the item values are known, but they are not all included in the basket. Present values have a quantity, while the items that are not represented are sparse (with a known quantity of zero).

Oracle Machine Learning for SQL interprets missing data as follows:

- Missing at random: Missing values in columns with a simple data type (not nested) are assumed to be missing at random.
- Sparse: Missing values in nested columns indicate sparsity.

Missing Values or Sparse Data?

Some real life examples are described to interpret missing values and sparse data.

The examples illustrate how Oracle Machine Learning for SQL identifies data as either sparse or missing at random.

Sparsity in a Sales Table

Understand how Oracle Machine Learning for SQL interprets missing data in nested column.

A sales table contains point-of-sale data for a group of products that are sold in several stores to different customers over a period of time. A particular customer buys only a few of the products. The products that the customer does not buy do not appear as rows in the sales table.

If you were to figure out the amount of money a customer has spent for each product, the unpurchased products have an inferred amount of zero. The value is not random or unknown; it is zero, even though no row appears in the table.

Note that the sales data is dimensioned (by product, stores, customers, and time) and are often represented as nested data for machine learning.

Since missing values in a nested column always indicate sparsity, you must ensure that this interpretation is appropriate for the data that you want to mine. For example, when trying to mine a multi-record case data set containing movie ratings from users of a large movie database, the missing ratings are unknown (missing at random), but Oracle Machine Learning for SQL treats the data as sparse and infer a rating of zero for the missing value.

Missing Values in a Table of Customer Data

When the data is not available for some attributes, those missing values are considered to be missing at random.

A table of customer data contains demographic data about customers. The case ID column is the customer ID. The attributes are age, education, profession, gender, house-hold size, and so on. Not all the data is available for each customer. Any missing values are considered to be missing at random. For example, if the age of customer 1 and the profession of customer 2 are not present in the data, that information is unknown. It does not indicate sparsity.

Note that the customer data is not dimensioned. There is a one-to-one mapping between the case and each of its attributes. None of the attributes are nested.

Changing the Missing Value Treatment

Transform the missing data as sparse or missing at random.

If you want Oracle Machine Learning for SQL to treat missing data as sparse instead of missing at random or missing at random instead of sparse, transform it before building the model.

If you want missing values to be treated as sparse, but OML4SQL interprets them as missing at random, you can use a SQL function like `NVL` to replace the nulls with a value such as "NA". OML4SQL does not perform missing value treatment when there is a specified value.

If you want missing nested attributes to be treated as missing at random, you can transform the nested rows into physical attributes in separate columns — as long as the case table stays within the column limitation imposed by the Database. Fill in all of the possible attribute names, and specify them as null. Alternatively, insert rows in the nested column for all the items that are not present and assign a value such as the mean or mode to each one.

Related Topics

- *Oracle Database SQL Language Reference*

Prepare the Case Table

The first step in preparing data for machine learning is the creation of a case table.

If all the data resides in a single table and all the information for each case (record) is included in a single row (single-record case), this process is already taken care of. If the data resides in several tables, creating the data source involves the creation of a view. For the sake of simplicity, the term "case table" is used here to refer to either a table or a view.

Convert Column Data Types

In OML, string columns are treated as categorical, number columns as numerical, and BOOLEAN columns are treated as numerical. If you have a numeric column that you want to be treated as a categorical, you must convert it to a string. For example, the day number of the week.

For example, zip codes identify different postal zones; they do not imply order. If the zip codes are stored in a numeric column, they are interpreted as a numeric attribute. You must convert the data type so that the column data can be used as a categorical attribute by the model. You can do this using the TO_CHAR function to convert the digits 1-9 and the LPAD function to retain the leading 0, if there is one.

```
LPAD(TO_CHAR(ZIPCODE),5,'0')
```

The attributes with the data type BOOLEAN are treated as numeric with the following values: TRUE means 1, FALSE means 0, and NULL is interpreted as an unknown value. The CASE_ID_COLUMN_NAME attribute does not support BOOLEAN data type.

Extract Datetime Column Values

You can extract values from a datatime or interval value using the EXTRACT function.

The EXTRACT function extracts and returns the value of a specified datetime field from a datetime or interval value expression. The values that can be extracted are YEAR, MONTH, DAY, HOUR, MINUTE, SECOND, TIMEZONE_HOUR, TIMEZONE_MINUTE, TIMEZONE_REGION, and TIMEZONE_ABBR.

```
sales_tssales_tsCUST_IDTIME_STAMP

select cust_id, time_stamp,
    extract(year from time_stamp) year,
    extract(month from time_stamp) month,
    extract(day from time_stamp) day_of_month,
    to_char(time_stamp,'ww') week_of_year,
    to_char(time_stamp,'D') day_of_week,
    extract(hour from time_stamp) hour,
    extract(minute from time_stamp) minute,
    extract(second from time_stamp) second
from sales_ts
```

Text Transformation

Learn text processing using Oracle Machine Learning for SQL.

You can use Oracle Machine Learning for SQL to process text. Columns of text in the case table can be processed once they have undergone the proper transformation.

The text column must be in a table, not a view. The transformation process uses several features of Oracle Text; it treats the text in each row of the table as a separate document. Each document is transformed to a set of text tokens known as **terms**, which have a numeric value and a text label. The text column is transformed to a nested column of `DM_NESTED_NUMERICALS`.

About Business and Domain-Sensitive Transformations

Understand why you need to transform data according to business problems.

Some transformations are dictated by the definition of the business problem. For example, you want to build a model to predict high-revenue customers. Since your revenue data for current customers is in dollars you need to define what "high-revenue" means. Using some formula that you have developed from past experience, you can recode the revenue attribute into ranges Low, Medium, and High before building the model.

Another common business transformation is the conversion of date information into elapsed time. For example, date of birth can be converted to age.

Domain knowledge can be very important in deciding how to prepare the data. For example, some algorithms produce unreliable results if the data contains values that fall far outside of the normal range. In some cases, these values represent errors or unusualities. In others, they provide meaningful information.

Related Topics

- Outlier Treatment
  Understand what you must do to treat outliers.

Create Nested Columns

In transactional data, the information for each case is contained in multiple rows. When the data source includes transactional data (multi-record case), the transactions must be aggregated to the case level in nested columns.

An example is sales data in a star schema when machine learning at the product level. Sales is stored in many rows for a single product (the case) because the product is sold in many stores to many customers over a period of time.

> **See Also:**
>
> Using Nested Data for information about converting transactional data to nested columns

Machine Learning Process

The lifecycle of a machine learning project is divided into six phases. The process begins by defining a business problem and restating the business problem in terms of a machine learning objective. The end goal of a machine learning process is to produce accurate results for solving your business problem.

Workflow

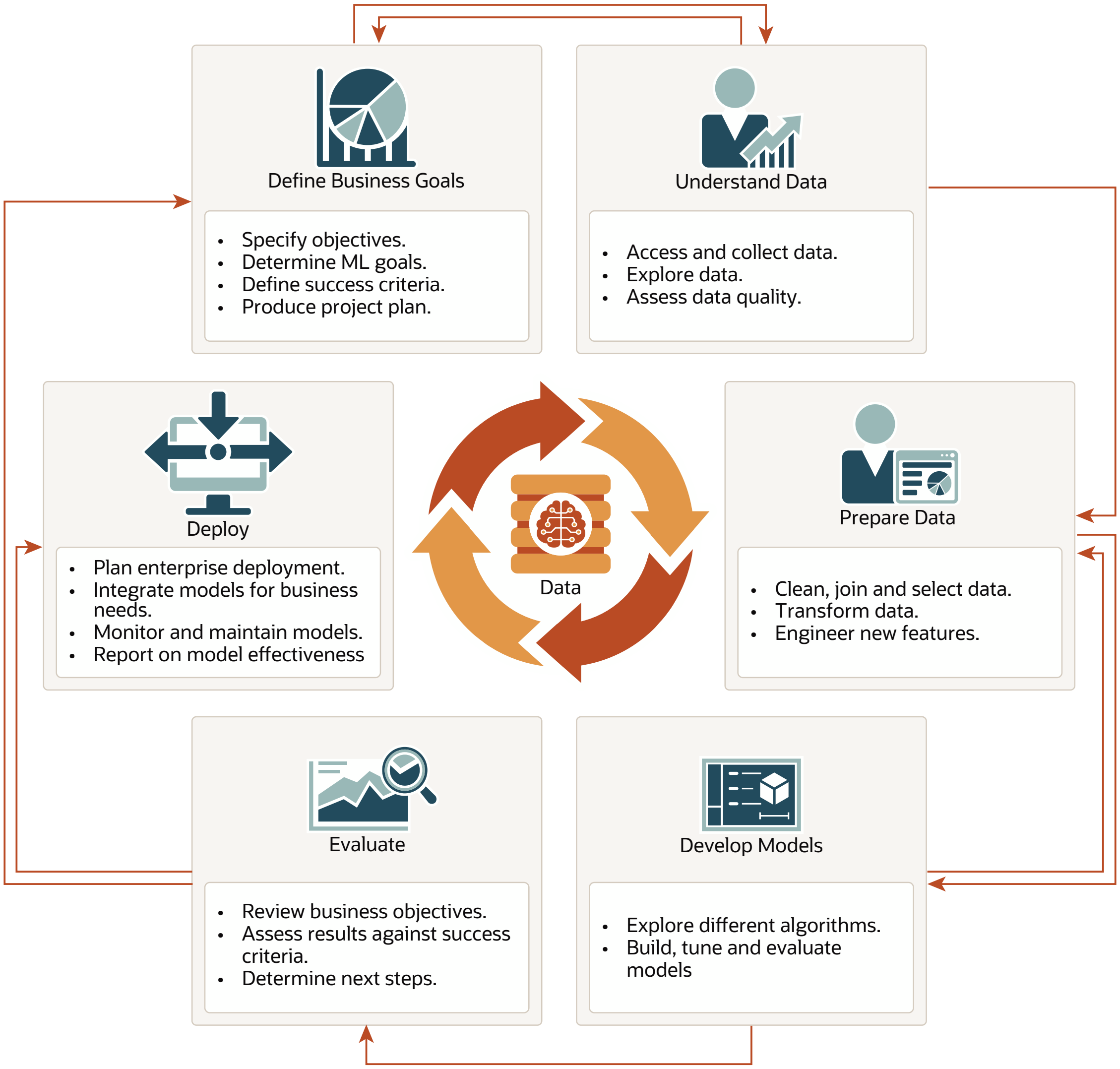
The machine learning process workflow illustration is based on the CRISP-DM methodology. Each stage in the workflow is illustrated with points that summarize the key tasks. The CRISP-DM methodology is the most commonly used methodology for machine learning.

The following are the phases of the machine learning process:

- Define business goals
- Understand data
- Prepare data
- Develop models
- Evaluate
- Deploy

Each of these phases are described separately. The following figure illustrates machine learning process workflow.

Figure 1-3   Machine Learning Process Workflow



Related Topics

- https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome
- https://www.sv-europe.com/crisp-dm-methodology/

Define Business Goals

The first phase of machine learning process is to define business objectives. This initial phase of a project focuses on understanding the project objectives and requirements.

Once you have specified the problem from a business perspective, you can formulate it as a machine learning problem and develop a preliminary implementation plan. Identify success criteria to determine if the machine learning results meet the business goals defined. For example, your business problem might be: "How can I sell more of my product to customers?" You might translate this into a machine learning problem such as: "Which customers are most likely to purchase the product?" A model that predicts who is most likely to purchase the product is typically built on data that describes the customers who have purchased the product in the past.

To summarize, in this phase, you will:

- Specify objectives
- Determine machine learning goals
- Define success criteria
- Produce project plan

Understand Data

The data understanding phase involves data collection and exploration which includes loading the data and analyzing the data for your business problem.

Assess the various data sources and formats. Load data into appropriate data management tools, such as Oracle Database. Explore relationships in data so it can be properly integrated. Query and visualize the data to address specific data mining questions such as distribution of attributes, relationship between pairs or small number of attributes, and perform simple statistical analysis. As you take a closer look at the data, you can determine how well it can be used to addresses the business problem. You can then decide to remove some of the data or add additional data. This is also the time to identify data quality problems such as:

- Is the data complete?
- Are there missing values in the data?
- What types of errors exist in the data and how can they be corrected?

To summarize, in this phase, you will:

- Access and collect data
- Explore data
- Assess data quality

Prepare Data

The preparation phase involves finalizing the data and covers all the tasks involved in making the data in a format that you can use to build the model.

Data preparation tasks are likely to be performed multiple times, iteratively, and not in any prescribed order. Tasks can include column (attributes) selection as well as selection of rows in a table. You may create views to join data or materialize data as required, especially if data is collected from various sources. To cleanse the data, look for invalid values, foreign key values that don't exist in other tables, and missing and outlier values. To refine the data, you can apply transformations such as aggregations, normalization, generalization, and attribute constructions needed to address the machine learning problem. For example, you can transform a `DATE_OF_BIRTH` column to `AGE`; you can insert the median income in cases where the `INCOME` column is null; you can filter out rows representing outliers in the data or filter columns that have too many missing or identical values.

Additionally you can add new computed attributes in an effort to tease information closer to the surface of the data. This process is referred as *Feature Engineering*. For example, rather than using the purchase amount, you can create a new attribute: "Number of Times Purchase Amount Exceeds $500 in a 12 month time period." Customers who frequently make large purchases can also be related to customers who respond or don't respond to an offer.

Thoughtful data preparation and feature engineering that capture domain knowledge can significantly improve the patterns discovered through machine learning. Enabling the data professional to perform data assembly, data preparation, data transformations, and feature engineering inside the Oracle Database is a significant distinction for Oracle.

> **Note:**
>
> Oracle Machine Learning supports Automatic Data Preparation (ADP), which greatly simplifies the process of data preparation.

To summarize, in this phase, you will:

- Clean, join, and select data
- Transform data
- Engineer new features

Related Topics

- *Oracle Machine Learning for SQL User's Guide*

Develop Models

In this phase, you select and apply various modeling techniques and tune the algorithm parameters, called *hyperparameters*, to desired values.

If the algorithm requires specific data transformations, then you need to step back to the previous phase to apply them to the data. For example, some algorithms allow only numeric columns such that string categorical data must be "exploded" using one-hot encoding prior to modeling. In preliminary model building, it often makes sense to start with a sample of the data since the full data set might contain millions or billions of rows. Getting a feel for how a given algorithm performs on a subset of data can help identify data quality issues and algorithm setting issues sooner in the process reducing time-to-initial-results and compute costs. For supervised learning problem, data is typically split into train (build) and test data sets using an 80-20% or 60-40% distribution. After splitting the data, build the model with the desired model settings. Use default settings or customize by changing the model setting values. Settings can be specified through OML's PL/SQL, R and Python APIs. Evaluate model quality through metrics appropriate for the technique. For example, use a confusion matrix, precision, and recall for classification models; RMSE for regression models; cluster similarity metrics for clustering models and so on.

Automated Machine Learning (AutoML) features may also be employed to streamline the iterative modeling process, including algorithm selection, attribute (feature) selection, and model tuning and selection.

To summarize, in this phase, you will:

- Explore different algorithms
- Build, evaluate, and tune models

Related Topics

- *Oracle Machine Learning for SQL User's Guide*

Splitting the Data

Separate data sets are required for building (training) and testing some predictive models. Typically, one large table or view is split into two data sets: one for building the model, and the other for testing the model.

The build data (training data) and test data must have the same column structure. The process of applying the model to test data helps to determine whether the model, built on one chosen sample, is generalizable to other data.

You need two case tables to build and validate supervised (like classification and regression) models. One set of rows is used for training the model, another set of rows is used for testing the model. It is often convenient to derive the build data and test data from the same data set. For example, you could randomly select 60% of the rows for training the model; the remaining 40% could be used for testing the model. Models that implement unsupervised machine learning techniques, such as attribute importance, clustering, association, or feature extraction, do not use separate test data.

Evaluate

At this stage of the project, it is time to evaluate how well the model satisfies the originally-stated business goal.

During this stage, you will determine how well the model meets your business objectives and success criteria. If the model is supposed to predict customers who are likely to purchase a product, then does it sufficiently differentiate between the two classes? Is there sufficient lift? Are the trade-offs shown in the confusion matrix acceptable? Can the model be improved by adding text data? Should transactional data such as purchases (market-basket data) be included? Should costs associated with false positives or false negatives be incorporated into the model?

It is useful to perform a thorough review of the process and determine if important tasks and steps are not overlooked. This step acts as a quality check based on which you can determine the next steps such as deploying the project or initiate further iterations, or test the project in a pre-production environment if the constraints permit.

To summarize, in this phase, you will:

- Review business objectives
- Assess results against success criteria
- Determine next steps

Deploy

Deployment is the use of machine learning within a target environment. In the deployment phase, one can derive data driven insights and actionable information.

Deployment can involve scoring (applying a model to new data), extracting model details (for example the rules of a decision tree), or integrating machine learning models within applications, data warehouse infrastructure, or query and reporting tools.

Because Oracle Machine Learning builds and applies machine learning models inside Oracle Database, the results are immediately available. Reporting tools and dashboards can easily display the results of machine learning. Additionally, machine learning supports scoring single cases or records at a time with dynamic, batch, or real-time scoring. Data can be scored and the results returned within a single database transaction. For example, a sales representative can run a model that predicts the likelihood of fraud within the context of an online sales transaction.

To summarize, in this phase, you will:

- Plan enterprise deployment
- Integrate models with application for business needs
- Monitor, refresh, retire, and archive models
- Report on model effectiveness

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

Supervised Learning: Testing

Testing evaluates the model's generalizability to new data, ensuring it avoids overfitting and accurately predicts outcomes.

The process of applying the model to test data helps to determine whether the model, built on one chosen sample, is generalizable to other data. In other words, test data is used for scoring. In particular, it helps to avoid the phenomenon of overfitting, which can occur when the logic of the model fits the build data too well and therefore has little predictive power.

Supervised Learning: Scoring

Scoring applies the model to new data to make predictions, using techniques like classification and regression for different types of data.

Apply data, also called scoring data, is the actual population to which a model is applied. For example, you might build a model that identifies the characteristics of customers who frequently buy a certain product. To obtain a list of customers who shop at a certain store and are likely to buy a related product, you might apply the model to the customer data for that store. In this case, the store customer data is the scoring data.

Most supervised learning can be applied to a population of interest. The principal supervised machine learning techniques, **classification** and **regression**, can both be used for scoring.

Oracle Machine Learning does not support the scoring operation for **attribute importance**, another supervised technique. Models of this type are built on a population of interest to obtain information about that population; they cannot be applied to separate data. An attribute importance model returns and ranks the attributes that are most important in predicting a target value.

Oracle Machine Learning supports the supervised machine learning techniques described in the following table:

Table 1-6 Oracle Machine Learning Supervised Techniques

Technique	Description	Sample Problem
Attribute Importance	Identifies the attributes that are most important in predicting a target attribute	Given customer response to an affinity card program, find the most significant predictors
Classification	Assigns items to discrete classes and predicts the class to which an item belongs	Given demographic data about a set of customers, predict customer response to an affinity card program

Table 1-6 (Cont.) Oracle Machine Learning Supervised Techniques

Technique	Description	Sample Problem
Regression	Approximates and forecasts continuous values	Given demographic and purchasing data about a set of customers, predict customers' age

Unsupervised Learning: Scoring

Scoring in unsupervised learning applies models to data for clustering and feature extraction, revealing hidden patterns and structures.

Although unsupervised machine learning does not specify a target, most unsupervised learning can be applied to a population of interest. For example, clustering models use descriptive machine learning techniques, but they can be applied to classify cases according to their cluster assignments. **Anomaly Detection**, although unsupervised, is typically used to predict whether a data point is typical among a set of cases.

Oracle Machine Learning supports the scoring operation for **Clustering** and **Feature Extraction**, both unsupervised machine learning techniques. Oracle Machine Learning does not support the scoring operation for **Association Rules**, another unsupervised function. Association models are built on a population of interest to obtain information about that population; they cannot be applied to separate data. An association model returns rules that explain how items or events are associated with each other. The association rules are returned with statistics that can be used to rank them according to their probability.

OML supports the unsupervised techniques described in the following table:

Table 1-7 Oracle Machine Learning Unsupervised Techniques

Function	Description	Sample Problem
Anomaly Detection	Identifies items (outliers) that do not satisfy the characteristics of "normal" data	Given demographic data about a set of customers, identify customer purchasing behavior that is significantly different from the norm
Association Rules	Finds items that tend to co-occur in the data and specifies the rules that govern their co-occurrence	Find the items that tend to be purchased together and specify their relationship
Clustering	Finds natural groupings in the data	Segment demographic data into clusters and rank the probability that an individual belongs to a given cluster
Feature Extraction	Creates new attributes (features) using linear combinations of the original attributes	Given demographic data about a set of customers, group the attributes into general characteristics of the customers

Related Topics

- Machine Learning Techniques
  Provides an overview of concepts related to Oracle Machine Learning techniques.
- In-Database Scoring
  In-database scoring applies machine learning models to new data within the database, ensuring security, efficiency, and ease of integration with applications.

About Oracle Machine Learning for SQL

Oracle Machine Learning for SQL (OML4SQL) provides scalable in-database machine learning algorithms through PL/SQL and SQL APIs. The algorithms are fast and scalable, support algorithm-specific automatic data preparation, and can score in batch or real-time.

OML4SQL provides a powerful, state-of-the-art machine learning capability within Oracle Database. The parallelized algorithms in the database keep data under database control. There is no need to extract data to separate machine learning engines, which adds latency to data access and raises concerns about data security, storage, and recency. The algorithms are fast and scalable, support algorithm-specific automatic data preparation, and can score in batch or real-time. You can use OML4SQL to build and deploy predictive and descriptive machine learning applications, to add intelligent capabilities to existing applications, and to generate predictive queries for data exploration. OML4SQL provides explanatory prediction details when scoring data, so you can understand why an individual prediction is made.

OML4SQL offers a broad set of in-database algorithms for performing a variety of machine learning tasks, such as classification, regression, anomaly detection, feature extraction, clustering, and market basket analysis. The algorithms can work on standard case data, transactional data, star schemas, and unstructured text data. OML4SQL is uniquely suited to the analysis of very large data sets.

Oracle Machine Learning for SQL, along with Oracle Machine Learning for R and Oracle Machine Learning for Python, is a component of Oracle Machine Learning that provides three powerful APIs for in-database machine learning, among other features.

Oracle Machine Learning for SQL in the Database Kernel

Learn about the implementation of Oracle Machine Learning for SQL (OML4SQL) in Oracle Database kernel and its advantages.

OML4SQL is implemented in the Oracle Database kernel. OML4SQL models are first class database objects. Oracle Machine Learning for SQL processes use built-in features of Oracle Database to maximize scalability and make efficient use of system resources.

OML4SQL within Oracle Database offers many advantages:

- No Data Movement: Some machine learning products require that the data be exported from a corporate database and converted to a specialized format. With OML4SQL, no data movement or conversion is needed. This makes the entire process less complex, time-consuming, and error-prone, and it allows for the analysis of very large data sets.
- Security: Your data is protected by the extensive security mechanisms of Oracle Database. Moreover, specific database privileges are needed for different machine learning activities. Only users with the appropriate privileges can define, manipulate, or apply machine learning model objects.
- Data Preparation and Administration: Most data must be cleansed, filtered, normalized, sampled, and transformed in various ways before it can be mined. Up to 80% of the effort in a machine learning project is often devoted to data preparation. OML4SQL can automatically manage key steps in the data preparation process. Additionally, Oracle Database provides extensive administrative tools for preparing and managing data.
- Ease of Data Refresh: Machine learning processes within Oracle Database have ready access to refreshed data. OML4SQL can easily deliver machine learning results based on current data, thereby maximizing its timeliness and relevance.

- Oracle Database Analytics: Oracle Database offers many features for advanced analytics and business intelligence. You can easily integrate machine learning with other analytical features of the database, such as statistical analysis and analytic views.
- Oracle Technology Stack: You can take advantage of all aspects of Oracle's technology stack to integrate machine learning within a larger framework for business intelligence or scientific inquiry.
- Domain Environment: Machine learning models have to be built, tested, validated, managed, and deployed in their appropriate application domain environments. Machine learning results may need to be post-processed as part of domain specific computations (for example, calculating estimated risks and response probabilities) and then stored into permanent repositories or data warehouses. With OML4SQL, the pre- and post-machine learning activities can all be accomplished within the same environment.
- Application Programming Interfaces: The PL/SQL API and SQL language operators provide direct access to OML4SQL functionality in Oracle Database.

Related Topics

- Overview of Database Analytics

Oracle Machine Learning for SQL in Oracle Exadata

Understand how complex scoring and algorithmic processing is done using Oracle Exadata.

Scoring refers to the process of applying a OML4SQL model to data to generate predictions. The scoring process may require significant system resources. Vast amounts of data may be involved, and algorithmic processing may be very complex.

With OML4SQL, scoring can be off-loaded to intelligent Oracle Exadata Storage Servers where processing is extremely performant.

Oracle Exadata Storage Servers combine Oracle's smart storage software and Oracle's industry-standard hardware to deliver the industry's highest database storage performance. For more information about Oracle Exadata, visit the Oracle Technology Network.

Related Topics

- https://www.oracle.com/engineered-systems/exadata/

Highlights of the Oracle Machine Learning for SQL API

Learn about the advantages of OML4SQL application programming interface (API).

Machine learning is a valuable technology in many application domains. It has become increasingly indispensable in the private sector as a tool for optimizing operations and maintaining a competitive edge. Machine learning also has critical applications in the public sector and in scientific research. However, the complexities of machine learning application development and the complexities inherent in managing and securing large stores of data can limit the adoption of machine learning technology.

OML4SQL is uniquely suited to addressing these challenges. The machine learning engine is implemented in the database kernel, and the robust administrative features of Oracle Database are available for managing and securing the data. While supporting a full range of machine learning algorithms and procedures, the API also has features that simplify the development of machine learning applications.

The OML4SQL API consists of extensions to Oracle SQL, the native language of the database. The API offers the following advantages:

- Scoring in the context of SQL queries. Scoring can be performed dynamically or by applying machine learning models.
- Automatic Data Preparation (ADP) and embedded transformations.
- Model transparency. Algorithm-specific queries return details about the attributes that were used to create the model.
- Scoring transparency. Details about the prediction, clustering, or feature extraction operation can be returned with the score.
- Simple routines for predictive analytics.
- A workflow-based graphical user interface (GUI) within Oracle SQL Developer. You can download SQL Developer free of charge from the following site:

  `Oracle Data Miner`

> **Note:**
>
> The examples in this publication are taken from the OML4SQL examples that are available on GitHub. For information on the examples, see About the OML4SQL Examples.

Related Topics

- *Oracle Machine Learning for SQL Concepts*

2 Get Started

Install Database On-premises

You can download the latest database version on your system and use clients like Oracle SQL Developer to connect to the Oracle database.

About Installation

Oracle Machine Learning components associated with Oracle Database are included with the database license.

To install Oracle Database, follow the installation instructions for your platform. Choose a Data Warehousing configuration during the installation.

Oracle Data Miner, the graphical user interface to Oracle Machine Learning for SQL, is an extension to Oracle SQL Developer. Instructions for downloading SQL Developer and installing the Data Miner repository are available on https://www.oracle.com/database/technologies/odmrinstallation.html.

To perform machine learning activities, you must be able to log on to the Oracle Database, and your user ID must have the database privileges described in Grant Privileges for Oracle Machine Learning for SQL.

Related Topics

- `Oracle Data Miner`

> **See Also:**
>
> **Install and Upgrade** page of the Oracle Database online documentation library for your platform-specific installation instructions: `Oracle Database 23ai Release`

Enable or Disable a Database Option

You can access Oracle Machine Learning components after you complete the Oracle Database installation.

After installation, you can use the command-line utility `chopt` to enable or disable a database option.

Database Tuning Considerations for Oracle Machine Learning for SQL

Standard administrative practices can be followed to manage workload on the system when machine learning activities are running.

DBAs managing production databases that support Oracle Machine Learning for SQL must follow standard administrative practices as described in *Oracle Database Administrator's Guide*.

Building machine learning models and batch scoring of machine learning models tend to put a DSS-like workload on the system. Single-row scoring tends to put an OLTP-like workload on the system.

Database memory management can have a major impact on machine learning. The correct sizing of Program Global Area (PGA) memory is very important for model building, complex queries, and batch scoring. From a machine learning perspective, the System Global Area (SGA) is generally less of a concern. However, the SGA must be sized to accommodate real-time scoring, which loads models into the shared cursor in the SGA. In most cases, you can configure the database to manage memory automatically. To do so, specify the total maximum memory size in the tuning parameter `MEMORY_TARGET`. With automatic memory management, Oracle Database dynamically exchanges memory between the SGA and the instance PGA as needed to meet processing demands.

Most machine learning algorithms can take advantage of parallel execution when it is enabled in the database. Parameters in `INIT.ORA` control the behavior of parallel execution.

Related Topics

- Oracle® Database Administrator's Guide
- Part I Database Performace Fundamentals
- Part III Tuning Database Memory
- Oracle® Database VLDB and Partitioning Guide

Install SQL Developer

Oracle SQL Developer is a free, integrated development environment that simplifies the development and management of Oracle Database in both traditional and Cloud deployments.

About SQL Developer

Oracle SQL Developer is a graphical version of SQL*Plus that gives database developers a convenient way to perform basic tasks. You can browse, create, edit, and delete (drop); run SQL statements and scripts; edit and debug PL/SQL code; manipulate and export (unload) data; and view and create reports.

You can connect to any target Oracle Database schema using standard Oracle Database authentication. Once connected, you can perform operations on objects in the database.

You can connect to schemas for MySQL and selected third-party (non-Oracle) databases, such as Microsoft SQL Server, Sybase Adaptive Server, and IBM DB2, and view metadata and data in these databases; and you can migrate these databases to Oracle Database.

Install and Get Started with SQL Developer

To install and start SQL Developer, download a ZIP file and unzip it into the desired parent directory on your system or folder and then type a command or double-click a file name.

If Oracle Database (Release 11 or later) is also installed, a version of SQL Developer is also included and is accessible through the menu system under Oracle. This version of SQL Developer is separate from any SQL Developer kit that you download and unzip on your own, so do not confuse the two, and do not unzip a kit over the SQL Developer files that are included with Oracle Database.

> **Tip:**
>
> Create a shortcut for the SQL Developer executable file that you install, and use it to start SQL Developer.

1. Unzip the SQL Developer kit into a folder (directory) of your choice, which will be referred to as `<sqldeveloper_install>`. Unzipping the SQL Developer kit causes a folder named `sqldeveloper` to be created under the `<sqldeveloper_install>` folder.
   For example, if you unzip the kit into `C:\`, the folder `C:\sqldeveloper` is created, along with several sub-folders under it.
2. To start SQL Developer, go to the `sqldeveloper` directory under the `<sqldeveloper_install>` directory, and do one of the following:
   On Linux and Mac OS X systems, run `sh sqldeveloper.sh`.

   On Windows systems, double-click `sqldeveloper.exe`.

   If you are asked to enter the full pathname for the JDK, click **Browse** and find it. For example, on a Windows system the path might have a name similar to `C:\Program Files\Java\jdk1.7.0_51`. (If you cannot start SQL Developer, it could be due to an error in specifying or configuring the JDK.)
3. Create at least one database connection (or import some previously exported connections), so that you can view and work with database objects, use the SQL Worksheet, and use other features.
   To create a new database connection:

   a. Right-click the **Connections** node in the **Connections** navigator

   b. Select **New Connection**, and complete the required entries in the Create/Edit/Select Database Connection dialog box. (You may also be able to generate connections automatically by right-clicking the Connections node and selecting Create Local Connections.)

Related Topics

- Database Connections

Access Autonomous Database

Oracle Autonomous Database is a family of self-driving, self-securing, and self-repairing cloud services. You can sign up for an Oracle Cloud Free Tier account and create a database instance.

Provision an Autonomous Database

A LiveLabs workshop (a set of labs) that teaches you to manage and monitor Autonomous Database (ADB) is available. A part of the workshop aims to provision an Autonomous Database instance on Oracle Cloud.

Manage and Monitor Autonomous Database

Create and Update User Accounts for Oracle Machine Learning Components on Autonomous Database

An administrator can add an existing database user account to use with Oracle Machine Learning components or create a new user account and user credentials with the Oracle Machine Learning User Management interface.

Create User

An administrator creates new user accounts and user credentials for Oracle Machine Learning in the User Management interface.

> **Note:**
>
> You must have the administrator role to access the Oracle Machine Learning User Management interface.

To create a user account:

1. On the Autonomous Databases page, under the **Display Name**, select an Autonomous Database.
2. On the Autonomous Database Details page, select **Database Actions** and click **Database users**.

   As an alternative, select Database Actions and click **View all database actions** to access the Database Actions Launchpad. From the Database Actions launchpad, under **Administration** click **Database Users**.
3. Click **+ Create User**.
4. In the **User Name** field, enter a username for the account. Using the username, the user will log in to an Oracle Machine Learning instance.
5. (Optional) Select the option **Password Expired (user must change).** to required the user to change their password when they login for the first time.
6. In the **Password** field, enter a password for the user.
7. In the **Confirm Password** field, enter a password to confirm the value that you entered in the **Password** field.
8. Select **OML** to enable Oracle Machine Learning for the user.
9. Click **Create User**.

This creates a new database user and grants the required privileges to use Oracle Machine Learning.

Note:

With a new database user, an administrator needs to issue grant commands on the database to grant table access to the new user for the tables associated with the user's Oracle Machine Learning notebooks.

Add Existing Database User Account to Oracle Machine Learning Components

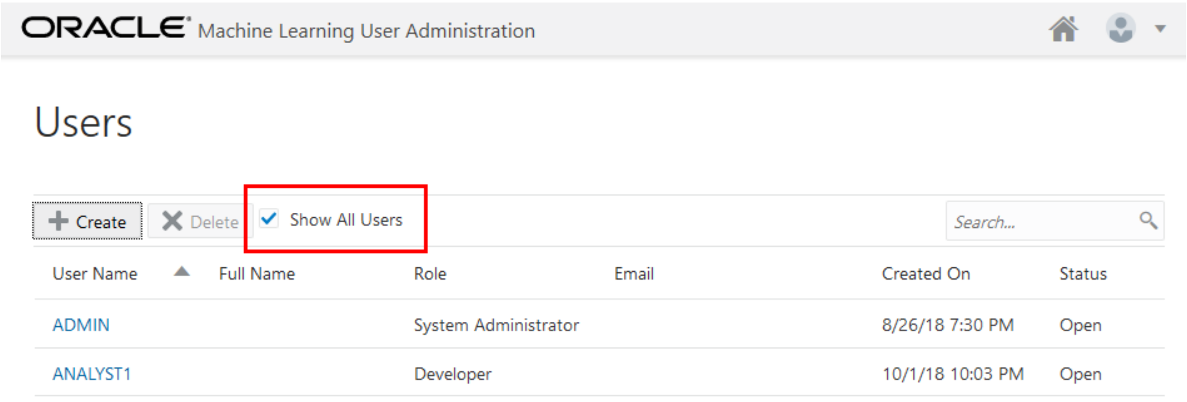
As the ADMIN user you can add an existing database user account for Oracle Machine Learning components.

Note:

You must have the ADMIN role to access the Oracle Machine Learning User Management interface.

To add an existing database user account:

1. On the Autonomous Databases page, under the **Display Name** column, select an Autonomous Database.
2. On the Autonomous Database Details page, select **Database Actions** and click **View all database actions**.
3. On the Database Actions Launchpad, under **Development**, click **Oracle Machine Learning**.
4. Expand the navigator by clicking the ≡ next to Oracle Machine Learning.
5. Under Admin, select **Manage OML Users** to add Oracle Machine Learning Notebooks users.
6. Click **Show All Users** to display the existing database users.



User Name	Full Name	Role	Email	Created On	Status
ADMIN		System Administrator		8/26/18 7:30 PM	Open
ANALYST1		Developer		10/1/18 10:03 PM	Open

Note:

Initially, the **Role** field shows the role **None** for existing database users. After adding a user the role **Developer** is assigned to the user.

7. Select a user. To select a user select a name in the **User Name** column. For example, select **ANALYST1**.

   Selecting the user shows the Oracle Machine Learning **Edit User** page.

8. Enter a name in the **First Name** field. (Optional)

9. Enter the last name of the user in the **Last Name** field. (Optional)

10. In the **Email Address** field, enter the email ID of the user.

    Making any change on this page adds the existing database user with the required privileges as an Oracle Machine Learning component user.

11. Click **Save**.

This grants the required privileges to use the Oracle Machine Learning application. In Oracle Machine Learning this user can then access any tables the user has privileges to access in the database.

Access OML Notebooks

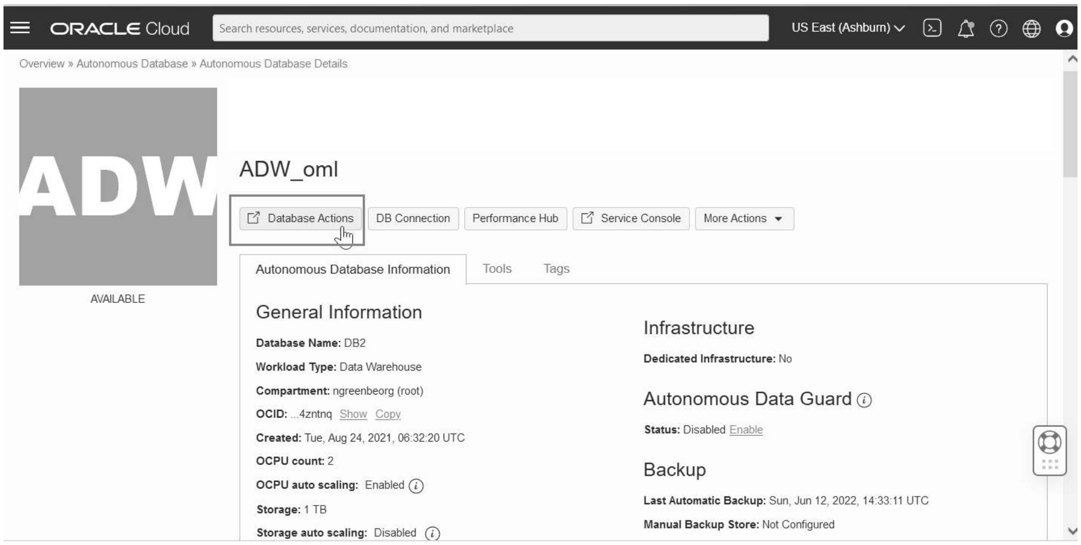
To perform Oracle Machine Learning tasks, you can access Oracle Machine Learning Notebooks from Autonomous Database

Access Oracle Machine Learning User Interface

You can access Oracle Machine Learning **User Interface** from Autonomous Database.

To access Oracle Machine Learning User Interface (UI) from the Autonomous Database:

1. Select your Autonomous Database instance and on the Autonomous Database details page click **Database Actions**.



ORACLE Cloud
Search resources, services, documentation, and marketplace
US East (Ashburn)
Overview » Autonomous Database » Autonomous Database Details
ADW
AVAILABLE
ADW_oml
Database Actions
DB Connection
Performance Hub
Service Console
More Actions
Autonomous Database Information
Tools
Tags

General Information

Database Name: DB2

Workload Type: Data Warehouse

Compartment: ngreenbeorg (root)

OCID: ...4zntnq Show Copy

Created: Tue, Aug 24, 2021, 06:32:20 UTC

OCPU count: 2

OCPU auto scaling: Enabled

Storage: 1 TB

Storage auto scaling: Disabled

Infrastructure

Dedicated Infrastructure: No

Autonomous Data Guard

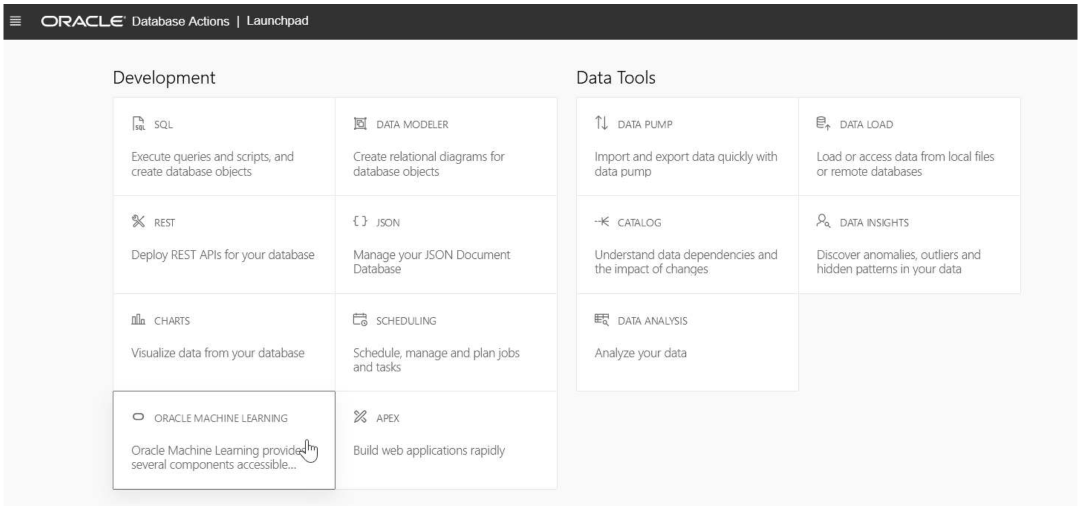
Status: Disabled Enable

Backup

Last Automatic Backup: Sun, Jun 12, 2022, 14:33:11 UTC

Manual Backup Store: Not Configured

2. On the Database Actions page, go to the **Development** section and click **Oracle Machine Learning**. The Oracle Machine Learning sign in page opens.


ORACLE Database Actions | Launchpad
Development
Data Tools
SQL
Execute queries and scripts, and create database objects
DATA MODELER
Create relational diagrams for database objects
DATA PUMP
Import and export data quickly with data pump
DATA LOAD
Load or access data from local files or remote databases
REST
Deploy REST APIs for your database
JSON
Manage your JSON Document Database
CATALOG
Understand data dependencies and the impact of changes
DATA INSIGHTS
Discover anomalies, outliers and hidden patterns in your data
CHARTS
Visualize data from your database
SCHEDULING
Schedule, manage and plan jobs and tasks
DATA ANALYSIS
Analyze your data
ORACLE MACHINE LEARNING
Oracle Machine Learning provides several components accessible...
APEX
Build web applications rapidly

3. On the Oracle Machine Learning sign in page, enter your username and password.
4. Click **Sign In**.

This opens the Oracle Machine Learning user application.

Create a Notebook Classic

A Notebook Classic is a web-based interface for data analysis, data discovery, data visualization and collaboration.

Whenever you create a Notebook Classic, it has an interpreter settings specification. The Notebook Classic contains an internal list of bindings that determines the order of the interpreter bindings. A Notebook Classic comprises paragraphs which is a notebook component where you can write SQL statements, run PL/SQL scripts, and run Python commands. A paragraph has an input section and an output section. In the input section, specify the interpreter to run along with the text. This information is sent to the interpreter to be executed. In the output section, the results of the interpreter are provided.

To create a Notebook Classic:

1. On the Oracle Machine Learning UI home page, click **Notebooks Classic.** The Notebooks Classic page opens.
2. On the Notebooks Classic page, click **Create.**

   The Create Notebook window appears.
3. In the **Name** field, provide a name for the notebook.
4. In the **Comments** field, enter comments, if any.
5. Click **OK.**

Your Notebook Classic is created and it opens in the notebook editor. You can now use it to run SQL statements, run PL/SQL scripts, run Python, R and Conda commands. To do so, specify any one of the following directives in the input section of the paragraph:

- `%sql` — To connect to the SQL interpreter and run SQL statements
- `%script` — To connect to the PL/SQL interpreter and run PL/SQL scripts
- `%md` — To connect to the Markdown interpreter and generate static html from Markdown plain text

- `%python` — To connect to the Python interpreter and run Python scripts
- `%r` — To connect to the R interpreter and run R scripts.
- `%conda` — To connect to the Conda interpreter, and install third-party Python and R libraries inside a notebook session.

Edit Your Notebook Classic

Upon creating an OML Notebook Classic, it opens automatically, presenting you with a single paragraph using the default `%sql` interpreter. You can change the interpreter by explicitly specifying one of `%script, %python, %sql` , `%r` , `%md` or `%conda`.

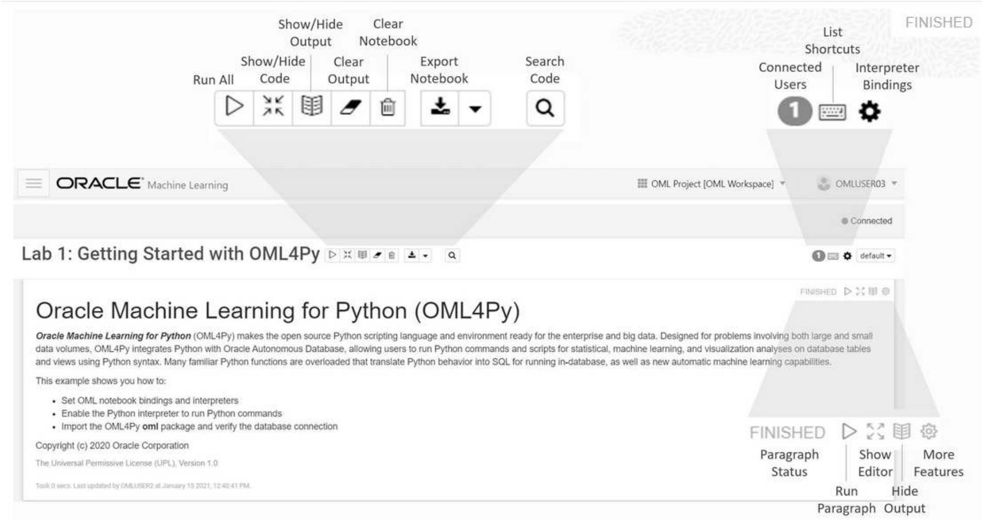
Set the context with a project with which your notebook is associated.

You can edit an existing Notebook Classic in your project. To edit an existing Notebook Classic:

1. On Oracle Machine Learning UI home page, select the project in which your notebook is available.
2. Go to the Oracle Machine Learning UI navigator, and select **Notebooks Classic.** All notebooks that are available in the project are listed.
3. Click the notebook that you want to open and edit.

   The selected notebook opens in edit mode.
4. In the edit mode, you can use the Oracle Machine Learning Notebooks Classic toolbar options to run code in paragraphs, for configuration settings, and display options.

Figure 2-1 Notebook toolbar



Lab 1: Getting Started with OML4Py

Oracle Machine Learning for Python (OML4Py)

Oracle Machine Learning for Python (OML4Py) makes the open source Python scripting language and environment ready for the enterprise and big data. Designed for problems involving both large and small data volumes, OML4Py integrates Python with Oracle Autonomous Database, allowing users to run Python commands and scripts for statistical, machine learning, and visualization analyses on database tables and views using Python syntax. Many familiar Python functions are overloaded that translate Python behavior into SQL for running in-database, as well as new automatic machine learning capabilities.

This example shows you how to:

- Set OML notebook bindings and interpreters
- Enable the Python interpreter to run Python commands
- Import the OML4Py **oml** package and verify the database connection

Copyright (c) 2020 Oracle Corporation

The Universal Permissive License (UPL), Version 1.0

Took 0 secs. Last updated by OMLUSER2 at January 15 2021, 12:40:41 PM.

You can perform the following tasks:

- Write code to fetch data
- Click ▷ to run one or all paragraphs in the notebook.

- Click ↘↙↗↖ to hide all codes from all the paragraphs in the notebook. Click it again to display the codes.
- Click 📖 to hide all outputs from all the paragraphs in the notebook. Click it again to view the outputs.
- Click to remove all outputs from all the paragraphs in the notebook. To view the output, click the run icon again.
- Click 🗑 to delete all the paragraphs in the notebook.
- Click ⬇ to export the notebook.
- Click 🔍 to search any information in the codes present in the notebook.
- Click ⌨ to view the list of keyboard shortcuts.
- Click ⚙ to set the order for interpreter bindings for the notebook.
- Click default ▾ to select one of the three notebook display options.
  - Click **default** to view the codes, output, and metadata in all paragraphs in the notebook.
  - Click **Simple** to view only the code and output in all paragraphs in the notebook. In this view, the notebook toolbar and all edit options are hidden. You must hover your mouse to view the edit options.
  - Click **Report** to view only the output in all paragraphs in the notebook.
- Click ⚙ to access paragraph specific edit options such as clear output, remove paragraph, adjust width, font size, run all paragraphs above or below the selected paragraph and so on.
- Add dynamic forms such as the Text Input form, Select form, Check box form for easy selection of inputs and easy filtering of data in your notebook. Oracle Machine Learning supports the following Apache Zeppelin dynamic forms:
  - Text Input form — Allows you to create a simple form for text input.
  - Select form — Allows you to create a form containing a range of values that the user can select.
  - Check Box form — Allows you to insert check boxes for multiple selection of inputs.

> **Note:**
>
> The Apache Zeppelin dynamic forms are supported only on SQL interpreter notebooks.

5. Once you have finished editing the notebook, click **Back.**

   This takes you back to the Notebooks Classic page.

3 Develop

Interfaces to Oracle Machine Learning for SQL

Introduces supported interfaces for Oracle Machine Learning for SQL.

The programmatic interfaces to Oracle Machine Learning for SQL are PL/SQL for building and maintaining models and a family of SQL functions for scoring. OML4SQL also supports a graphical user interface, which is implemented as an extension to Oracle SQL Developer.

Oracle Predictive Analytics, a set of simplified OML4SQL routines, is built on top of OML4SQL and is implemented as a PL/SQL package.

Oracle Machine Learning Modeling, Transformations, and Convenience Functions

You can access PL/SQL interface to perform data modeling, transformations, and predictive analytics.

The following table displays the PL/SQL packages for Oracle Machine Learning. In Oracle Database releases prior to Release 21c, Oracle Machine Learning was named Oracle Data Mining.

Table 3-1 Oracle Machine Learning PL/SQL Packages

Package Name	Description
`DBMS_DATA_MINING`	Routines for creating and managing machine learning models
`DBMS_DATA_MINING_TRANSFORM`	Routines for transforming the data for machine learning
`DBMS_PREDICTIVE_ANALYTICS`	Routines that perform predictive analytics

Related Topics

- DBMS_DATA_MINING
- DBMS_DATA_MINING_TRANSFORM
- DBMS_PREDICTIVE_ANALYTICS

DBMS_DATA_MINING

The DBMS_DATA_MINING package contains routines for creating machine learning models, for performing operations on the models, and for querying them.

The package includes routines for:

- Creating, dropping, and performing other DDL operations on machine learning models
- Obtaining detailed information about model attributes, rules, and other information internal to the model (model details)
- Computing test metrics for classification models
- Specifying costs for classification models
- Exporting and importing models
- Building models using Oracle Machine Learning native algorithms as well as algorithms written in R

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

About Oracle Machine Learning Models

Machine learning models are database schema objects that perform machine learning techniques.

As with all schema objects, access to machine learning models is controlled by database privileges. Models can be exported and imported. They support comments and they can be tracked in the Oracle Database auditing system.

Machine learning models are created by the CREATE_MODEL2 or the CREATE_MODEL procedures in the DBMS_DATA_MINING PL/SQL package. Models are created for a specific machine learning technique, and they use a specific algorithm to perform that function. **Machine learning technique** is a term that refers to a class of machine learning problems to be solved. Examples of machine learning techniques are: regression, classification, attribute importance, clustering, anomaly detection, and feature selection. OML4SQL supports one or more algorithms for each machine learning technique.

Along with the machine learning technique, in the CREATE_MODEL2 procedure, you can specify an algorithm and other characteristics of a model. In CREATE_MODEL procedure you can specify a settings table to specify an algorithm and other characteristics of a model. Some settings are general, some are specific to a machine learning technique, and some are specific to an algorithm.

> **Note:**
>
> Most types of machine learning models can be used to score data. However, it is possible to score data without applying a model. Dynamic scoring and predictive analytics return scoring results without a user-supplied model. They create and apply transient models that are not visible to you.

Related Topics

- Upgrade and Downgrade
  Explains how to perform administrative tasks related to Oracle Machine Learning for SQL.
- Dynamic Scoring
  You can perform dynamic scoring if, for some reason, you do not want to apply a predefined model.
- DBMS_PREDICTIVE_ANALYTICS
  The `DBMS_PREDICTIVE_ANALYTICS` package contains routines that perform an automated form of machine learning known as predictive analytics. With predictive analytics, you do not need to be aware of model building or scoring. All machine learning activities are handled internally by the procedure.
- Control Access to Oracle Machine Learning for SQL Models and Data
  You can create a Oracle Machine Learning for SQL user and grant necessary privileges by following the steps listed.

DBMS_DATA_MINING_TRANSFORM

The `DBMS_DATA_MINING_TRANSFORM` package contains routines that perform data transformations such as binning, normalization, and outlier treatment.

The package includes routines for:

- Specifying transformations in a format that can be embedded in a machine learning model.
- Specifying transformations as relational views (external to machine learning model objects).
- Specifying distinct properties for columns in the build data. For example, you can specify that the column must be interpreted as unstructured text, or that the column must be excluded from Automatic Data Preparation.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Transformation Methods in DBMS_DATA_MINING_TRANSFORM

Summarizes the methods for transforming data in `DBMS_DATA_MINING_TRANSFORM` package.

Table 3-2 DBMS_DATA_MINING_TRANSFORM Transformation Methods

Transformation Method	Description
`XFORM` interface	`CREATE`, `INSERT`, and `XFORM` routines specify transformations in external views
`STACK` interface	`CREATE`, `INSERT`, and `XFORM` routines specify transformations for embedding in a model
`SET_TRANSFORM`	Specifies transformations for embedding in a model

The statements in the following example create a Support Vector Machine (SVM) classification model called T_SVM_Clas_sample with an embedded transformation that causes the comments attribute to be treated as unstructured text data. The T_SVM_CLAS_SAMPLE model is created by `oml4sql-classification-text-mining-svm.sql` example.

Example 3-1 Sample Embedded Transformation

```
DECLARE
  xformlist dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
  dbms_data_mining_transform.SET_TRANSFORM(
    xformlist, 'comments', null, 'comments', null, 'TEXT');
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'T_SVM_Clas_sample',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_build_text',
    case_id_column_name => 'cust_id',
    target_column_name  => 'affinity_card',
    settings_table_name => 't_svmc_sample_settings',
    xform_list => xformlist);
END;
/
```

Predictive Analytics

Predictive analytics is a technology that captures Oracle Machine Learning for SQL processes in simple routines.

Sometimes called "one-click machine learning," predictive analytics simplifies and automates the machine learning process.

Predictive analytics uses OML4SQL technology, but knowledge of OML4SQL is not needed to use predictive analytics. You can use predictive analytics by specifying an operation to perform on your data. You do not need to create or use OML4SQL models or understand the OML4SQL functions and algorithms summarized in "Oracle Machine Learning for SQL Basics ".

Oracle Machine Learning for SQL predictive analytics operations are described in the following table:

Table 3-3 Oracle Predictive Analytics Operations

Operation	Description
`EXPLAIN`	Explains how individual predictors (columns) affect the variation of values in a target column
`PREDICT`	For each case (row), predicts the values in a target column
`PROFILE`	Creates a set of rules for cases (rows) that imply the same target value

The Oracle predictive analytics operations are implemented in the `DBMS_PREDICTIVE_ANALYTICS` PL/SQL package. They are also available in Oracle Data Miner.

Related Topics

- About Oracle Machine Learning for SQL
  Oracle Machine Learning for SQL (OML4SQL) provides scalable in-database machine learning algorithms through PL/SQL and SQL APIs. The algorithms are fast and scalable, support algorithm-specific automatic data preparation, and can score in batch or real-time.

DBMS_PREDICTIVE_ANALYTICS

The `DBMS_PREDICTIVE_ANALYTICS` package contains routines that perform an automated form of machine learning known as predictive analytics. With predictive analytics, you do not need to be aware of model building or scoring. All machine learning activities are handled internally by the procedure.

The DBMS_PREDICTIVE_ANALYTICS package includes these routines:

- **EXPLAIN** ranks attributes in order of influence in explaining a target column.
- **PREDICT** predicts the value of a target column based on values in the input data.
- **PROFILE** generates rules that describe the cases from the input data.

The EXPLAIN statement in the following example lists attributes in the view mining_data_build_v in order of their importance in predicting affinity_card.

Example 3-2 Sample EXPLAIN Statement

```
BEGIN
    DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
        data_table_name      => 'mining_data_build_v',
        explain_column_name  => 'affinity_card',
        result_table_name    => 'explain_results');
END;
/
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Oracle Machine Learning Data Dictionary Views

Lists Oracle Machine Learning data dictionary views.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

Table 3-4 Data Dictionary Views for Oracle Machine Learning

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_PARTITIONS	Provides information about the partitions of all accessible partitioned machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

SQL Functions

Oracle Machine Learning for SQL supports SQL functions for performing prediction, clustering, and feature extraction.

The functions score data by applying an OML4SQL model object or by running an analytic clause that performs dynamic scoring.

The following example shows a query that applies the classification model `svmc_sh_clas_sample` to the data in the view `mining_data_apply_v`. The query returns the average age of customers who are likely to use an affinity card. The results are broken out by gender.

Example 3-3 The PREDICTION Function

```
SELECT cust_gender,
       COUNT(*) AS cnt,
       ROUND(AVG(age)) AS avg_age
  FROM mining_data_apply_v
 WHERE PREDICTION(svmc_sh_clas_sample USING *) = 1
GROUP BY cust_gender
ORDER BY cust_gender;



C        CNT    AVG_AGE
- ---------- ----------
F         59         41
M        409         45
```

Related Topics

- In-Database Scoring
  In-database scoring applies machine learning models to new data within the database, ensuring security, efficiency, and ease of integration with applications.

Oracle Machine Learning for SQL Scoring Functions

Use OML4SQL functions score data. Functions can apply a machine learning model schema object to data or dynamically mine it with an analytic clause. SQL functions exist for all OML4SQL scoring algorithms.

All OML4SQL functions, as listed in the following table can operate on an R machine learning model with the corresponding OML4SQL function. However, the functions are not limited to the ones listed here.

Table 3-5 OML4SQL Functions

Function	Description
`CLUSTER_ID`	Returns the ID of the predicted cluster
`CLUSTER_DETAILS`	Returns detailed information about the predicted cluster
`CLUSTER_DISTANCE`	Returns the distance from the centroid of the predicted cluster
`CLUSTER_PROBABILITY`	Returns the probability of a case belonging to a given cluster
`CLUSTER_SET`	Returns a list of all possible clusters to which a given case belongs along with the associated probability of inclusion

Table 3-5 (Cont.) OML4SQL Functions

Function	Description
FEATURE_COMPARE	Compares two similar and dissimilar set of texts from two different documents or keyword phrases or a combination of both
FEATURE_ID	Returns the ID of the feature with the highest coefficient value
FEATURE_DETAILS	Returns detailed information about the predicted feature
FEATURE_SET	Returns a list of objects containing all possible features along with the associated coefficients
FEATURE_VALUE	Returns the value of the predicted feature
ORA_DM_PARTITION_NAME	Returns the partition names for a partitioned model
PREDICTION	Returns the best prediction for the target
PREDICTION_BOUNDS	(GLM only) Returns the upper and lower bounds of the interval wherein the predicted values (linear regression) or probabilities (logistic regression) lie.
PREDICTION_COST	Returns a measure of the cost of incorrect predictions
PREDICTION_DETAILS	Returns detailed information about the prediction
PREDICTION_PROBABILITY	Returns the probability of the prediction
PREDICTION_SET	Returns the results of a classification model, including the predictions and associated probabilities for each case
VECTOR_EMBEDDING	Generates a single vector embedding for different data types

The following example shows a query that returns the results of the CLUSTER_ID function. The query applies the model em_sh_clus_sample, which finds groups of customers that share certain characteristics. The query returns the identifiers of the clusters and the number of customers in each cluster. The em_sh_clus_sample model is created by the oml4sql-clustering-expectation-maximization.sql example.

Example 3-4 CLUSTER_ID Function

```
-- -List the clusters into which the customers in this
-- -data set have been grouped.
--
SELECT CLUSTER_ID(em_sh_clus_sample USING *) AS clus, COUNT(*) AS cnt
  FROM mining_data_apply_v
GROUP BY CLUSTER_ID(em_sh_clus_sample USING *)
ORDER BY cnt DESC;
```

```
-- List the clusters into which the customers in this
-- data set have been grouped.
--
SELECT CLUSTER_ID(em_sh_clus_sample USING *) AS clus, COUNT(*) AS cnt
FROM mining_data_apply_v
GROUP BY CLUSTER_ID(em_sh_clus_sample USING *)
ORDER BY cnt DESC;
```

The output is as follows:

```
      CLUS        CNT
---------- ----------
         9        311
         3        294
         7        215
        12        201
        17        123
        16        114
        14         86
        19         64
        15         56
        18         36
```

Related Topics

- Oracle Machine Learning for SQL Functions

Oracle Machine Learning for SQL Statistical Functions

Various SQL statistical functions are available in Oracle Database to explore and analyze data.

A variety of scalable statistical functions are accessible through SQL in Oracle Database. These statistical functions are implemented as SQL functions. The SQL statistical functions can be used to compute standard univariate statistics such as MEAN, MAX, MIN, MEDIAN, MODE, and standard deviation on the data. Users can also perform various other statistical functions such as t-test, f-test, aggregate functions, analytic functions, or ANOVA. The functions listed in the following table are available from SQL.

Table 3-6 SQL Statistical Functions Supported by OML4SQL

Function	Description
APPROX_COUNT	Returns approximate count of an expression
APPROX_SUM	Returns approximate sum of an expression
APPROX_RANK	Returns approximate value in a group of values
CORR	Retuns the coefficient of correlation of a set of number pairs
CORR_S	Calculates the Spearman's rho correlation coefficient
CORR_K	Calculates the Kendall's tau-b correlation coefficient
COVAR_POP	Returns the population covariance of a set of number pairs

Table 3-6 (Cont.) SQL Statistical Functions Supported by OML4SQL

Function	Description
COVAR_SAMP	Returns the sample covariance of a set of number pairs.
LAG	LAG is an analytic function. It provides access to more than one row of a table at the same time without a self join.
LEAD	LEAD is an analytic function. It provides access to more than one row of a table at the same time without a self join.
STATS_BINOMIAL_TEST	STATS_BINOMIAL_TEST is an exact probability test used for dichotomous variables, where only two possible values exist.
STATS_CROSSTAB	STATS_CROSSTAB is a method used to analyze two nominal variables.
STATS_F_TEST	STATS_F_TEST tests whether two variances are significantly different.
STATS_KS_TEST	STATS_KS_TEST is a Kolmogorov-Smirnov function that compares two samples to test whether they are from the same population or from populations that have the same distribution.
STATS_MODE	Takes as its argument a set of values and returns the value that occurs with the greatest frequency
STATS_MW_TEST	A Mann Whitney test compares two independent samples to test the null hypothesis that two populations have the same distribution function against the alternative hypothesis that the two distribution functions are different.
STATS_ONE_WAY_ANOVA	Tests differences in means (for groups or variables) for statistical significance by comparing two different estimates of variance
STATS_T_TEST_*	The t-test measures the significance of a difference of means
STATS_T_TEST_ONE	A one-sample t-test
STATS_T_TEST_PAIRED	A two-sample, paired t-test (also known as a crossed t-test)
STATS_T_TEST_INDEP and STATS_T_TEST_INDEPU	A t-test of two independent groups with the same variance (pooled variances) A t-test of two independent groups with unequal variance (unpooled variances)
STDDEV	returns the sample standard deviation of a set of numbers
STDDEV_POP	Computes the population standard deviation and returns the square root of the population variance
STDDEV_SAMP	Computes the cumulative sample standard deviation and returns the square root of the sample variance
SUM	Returns the sum of values

DBMS_STAT_FUNCS PL/SQL package is also available for users.

Related Topics

- R Operators and Functions Supported by Oracle Machine Learning for R

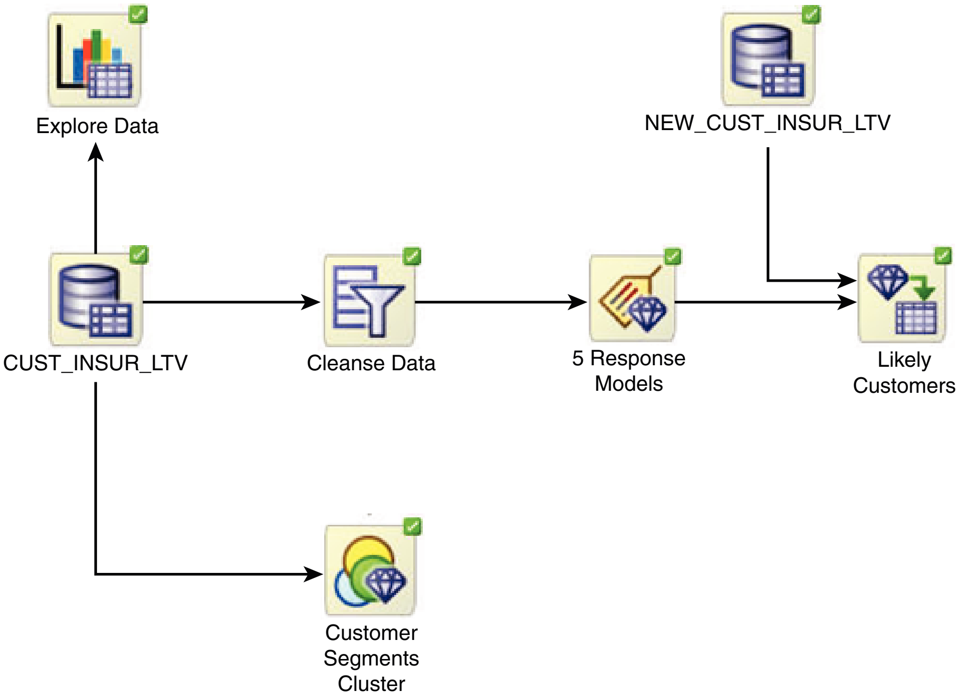
Oracle Data Miner

Oracle Machine Learning for SQL supports a graphical interface called Oracle Data Miner.

Oracle Data Miner is a graphical interface to OML4SQL. Oracle Data Miner is an extension to Oracle SQL Developer, which is available for download free of charge on the Oracle Technology Network.

Oracle Data Miner uses a work flow paradigm to capture, document, and automate the process of building, evaluating, and applying OML4SQL models. Within a work flow, you can specify data transformations, build and evaluate multiple models, and score multiple data sets. You can then save work flows and share them with other users.

Figure 3-1 An Oracle Data Miner Workflow



For information about Oracle Data Miner, including installation instructions, visit Oracle Technology Network.

Related Topics

- `Oracle Data Miner`

About Transformations

Understand how you can transform data by using Automatic Data Preparation (ADP) and embedded data transformation.

A transformation is a SQL expression that modifies the data in one or more columns. Data must typically undergo certain transformations before it can be used to build a model. Many Oracle Machine Learning algorithms have specific transformation requirements. Before data can be scored, it must be transformed in the same way that the training data was transformed.

Oracle Machine Learning for SQL supports ADP, which automatically implements the transformations required by the algorithm. The transformations are embedded in the model and automatically run whenever the model is applied.

If additional transformations are required, you can specify them as SQL expressions and supply them as input when you create the model. These transformations are embedded in the model as they are with ADP.

With automatic and embedded data transformation, most of the work of data preparation is handled for you. You can create a model and score multiple data sets in a few steps:

1. Identify the columns to include in the case table.
2. Create nested columns if you want to include transactional data.
3. Write SQL expressions for any transformations not handled by ADP.
4. Create the model, supplying the SQL expressions (if specified) and identifying any columns that contain text data.
5. Ensure that some or all of the columns in the scoring data have the same name and type as the columns used to train the model.

Related Topics

- Scoring Requirements
  Learn how scoring is done in Oracle Machine Learning for SQL.

> **See Also:**
>
> OML provides algorithm-specific automatic data preparation and other model building-related features

Embed Transformations in a Model

You can specify your own transformations and embed them in a model by creating a transformation list and passing it to `DBMS_DATA_MINING.CREATE_MODEL2` or `DBMS_DATA_MINING.CREATE_MODEL`.

The transformation instructions are embedded in the model and reapplied whenever the model is applied to new data.

The schema of how you can use `xform_list` to embed your transformations is shown here with `CREATE_MODEL` procedure.

```
DBMS_DATA_MINING.CREATE_MODEL2 (
model_name         IN VARCHAR2,
mining_function       IN VARCHAR2,
data_query         IN CLOB,
set_list           IN SETTING_LIST,
case_id_column_name      IN VARCHAR2 DEFAULT NULL,
target_column_name       IN VARCHAR2 DEFAULT NULL,
xform_list         IN TRANSFORM_LIST DEFAULT NULL);
```

```
DBMS_DATA_MINING.CREATE_MODEL(
                  model_name           IN VARCHAR2,
                  mining_function      IN VARCHAR2,
                  data_table_name      IN VARCHAR2,
                  case_id_column_name  IN VARCHAR2,
                  target_column_name   IN VARCHAR2 DEFAULT NULL,
                  settings_table_name  IN VARCHAR2 DEFAULT NULL,
                  data_schema_name     IN VARCHAR2 DEFAULT NULL,
                  settings_schema_name IN VARCHAR2 DEFAULT NULL,
                  xform_list           IN TRANSFORM_LIST DEFAULT NULL);
```

The following examples show how to create an embedded transform list with `CREATE_MODEL` and `CREATE_MODEL2` procedures.

Here is an example with `DBMS_DATA_MINING.CREATE_MODEL` procedure:

```
BEGIN
DBMS_DATA_MINING.DROP_MODEL('model_sample2');
EXCEPTION WHEN OTHERS THEN NULL;
END;
/
CREATE TABLE sett_table (SETTING_NAME  VARCHAR2(30),
                                    SETTING_VALUE VARCHAR2(4000));

BEGIN
   INSERT INTO sett_table (SETTING_NAME, SETTING_VALUE) VALUES
('KMNS_DISTANCE','KMNS_EUCLIDEAN');
   INSERT INTO sett_table (SETTING_NAME, SETTING_VALUE) VALUES
('PREP_AUTO','ON');
   INSERT INTO sett_table (SETTING_NAME, SETTING_VALUE) VALUES
('KMNS_DETAILS', 'KMNS_DETAILS_ALL');
END;
DECLARE
  xformlist dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'N_TRANS_ATM', null,
'TO_CHAR(N_TRANS_ATM)', null);
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'BANK_FUNDS', null,
'BANK_FUNDS+BANK_FUNDS+BANK_FUNDS', null);
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'AGE', null,
'log(10,AGE+1)', 'power(10, AGE)-1');

 DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'model_sample2',
    mining_function     => dbms_data_mining.clustering,
    data_table_name     => 'INSUR_CUST_LTV',
    case_id_column_name => 'customer_id',
    settings_table_name => 'sett_table',
    xform_list          => xformlist);
END;
```

The following example shows how to create an embedded transformation using the `DBMS_DATA_MINING.CREATE_MODEL2` procedure:

```
DECLARE
  xformlist dbms_data_mining_transform.TRANSFORM_LIST;
  v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'N_TRANS_ATM', null, 
'TO_CHAR(N_TRANS_ATM)', null);
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'BANK_FUNDS', null, 
'BANK_FUNDS+BANK_FUNDS+BANK_FUNDS', null);
  dbms_data_mining_transform.SET_TRANSFORM(xformlist, 'AGE', null, 
'log(10,AGE+1)', 'power(10, AGE)-1');

  v_setlst('ALGO_NAME') := 'ALGO_KMEANS';

 DBMS_DATA_MINING.CREATE_MODEL2(
    model_name          => 'model_sample3',
    mining_function     => 'CLUSTERING',
    data_query          => 'select * from INSUR_CUST_LTV',
    set_list            => v_setlst,
    case_id_column_name => 'customer_id',
    xform_list          => xformlist);
END;
```

Build a Transformation List

You can build transformation list by `SET_TRANSFORM`, `STACK`, and `GET_*` methods. These methods are listed here.

A transformation list is a collection of transformation records. When a new transformation record is added, it is appended to the top of the transformation list. You can use any of the following methods to build a transformation list:

- The `SET_TRANFORM` procedure in `DBMS_DATA_MINING_TRANSFORM`
- The `STACK` interface in `DBMS_DATA_MINING_TRANSFORM`
- The `GET_MODEL_TRANSFORMATIONS` and `GET_TRANSFORM_LIST` functions in `DBMS_DATA_MINING`

SET_TRANSFORM

The `SET_TRANSFORM` procedure applies a specified SQL expression to a specified attribute.

The `SET_TRANSFORM` procedure adds a single transformation record to a transformation list.

```
DBMS_DATA_MINING_TRANSFORM.SET_TRANSFORM (
         xform_list               IN OUT NOCOPY TRANSFORM_LIST,
         attribute_name           VARCHAR2,
         attribute_subname        VARCHAR2,
         expression               VARCHAR2,
         reverse_expression       VARCHAR2,
         attribute_spec           VARCHAR2 DEFAULT NULL);
```

SQL expressions that you specify with SET_TRANSFORM must fit within a VARCHAR2. To specify a longer expression, you can use the SET_EXPRESSION procedure, which builds an expression by appending rows to a VARCHAR2 array. For example, the following statement appends a transformation instruction for country_id to a list of transformations called my_xforms. The transformation instruction divides country_id by 10 before algorithmic processing begins. The reverse transformation multiplies country_id by 10.

```
dbms_data_mining_transform.SET_TRANSFORM (my_xforms,
   'country_id', NULL, 'country_id/10', 'country_id*10');
```

The reverse transformation is applied in the model details. If country_id is the target of a supervised model, the reverse transformation is also applied to the scored target.

The STACK Interface

The STACK interface creates transformation records from a table of transformation instructions and adds them to a transformation list.

The STACK interface offers a set of pre-defined transformations that you can apply to an attribute or to a group of attributes. For example, you can specify supervised binning for all categorical attributes.

The STACK interface specifies that all or some of the attributes of a given type must be transformed in the same way. For example, STACK_BIN_CAT appends binning instructions for categorical attributes to a transformation list. The STACK interface consists of three steps:

1. A CREATE procedure creates a transformation definition table. For example, CREATE_BIN_CAT creates a table to hold categorical binning instructions. The table has columns for storing the name of the attribute, the value of the attribute, and the bin assignment for the value.
2. An INSERT procedure computes the bin boundaries for one or more attributes and populates the definition table. For example, INSERT_BIN_CAT_FREQ performs frequency-based binning on some or all of the categorical attributes in the data source and populates a table created by CREATE_BIN_CAT.
3. A STACK procedure creates transformation records from the information in the definition table and appends the transformation records to a transformation list. For example, STACK_BIN_CAT creates transformation records for the information stored in a categorical binning definition table and appends the transformation records to a transformation list.

GET_MODEL_TRANSFORMATIONS and GET_TRANSFORM_LIST

Use the functions to create a new transformation list.

These two functions can be used to create a new transformation list from the transformations embedded in an existing model.

The GET_MODEL_TRANSFORMATIONS function returns a list of embedded transformations.

```
DBMS_DATA_MINING.GET_MODEL_TRANSFORMATIONS (
      model_name      IN VARCHAR2)
RETURN DM_TRANSFORMS PIPELINED;
```

GET_MODEL_TRANSFORMATIONS returns a table of dm_transform objects. Each dm_transform has these fields

```
attribute_name       VARCHAR2(4000)
attribute_subname    VARCHAR2(4000)
```

```
expression          CLOB
reverse_expression  CLOB
```

The components of a transformation list are `transform_rec`, not `dm_transform`. The fields of a `transform_rec` are described in Table 3-7. You can call `GET_MODEL_TRANSFORMATIONS` to convert a list of `dm_transform` objects to `transform_rec` objects and append each `transform_rec` to a transformation list.

```
DBMS_DATA_MINING.GET_TRANSFORM_LIST (
      xform_list           OUT NOCOPY TRANSFORM_LIST,
      model_xforms         IN   DM_TRANSFORMS);
```

> **See Also:**
>
> "DBMS_DATA_MINING_TRANSFORM Operational Notes", "SET_TRANSFORM Procedure", "CREATE_MODEL Procedure", and "GET_MODEL_TRANSFORMATIONS Function" in *Oracle Database PL/SQL Packages and Types Reference*

Transformation List and Automatic Data Preparation

You can provide transformation list and Automatic Data Preparation (ADP) to customize the data transformation.

The transformation list argument to `CREATE_MODEL2` and `CREATE_MODEL` interacts with the `PREP_AUTO` setting, which controls ADP:

- When ADP is on and you specify a transformation list, your transformations are applied with the automatic transformations and embedded in the model. The transformations that you specify are processed before the automatic transformations.
- When ADP is off and you specify a transformation list, your transformations are applied and embedded in the model, but no system-generated transformations are performed.
- When ADP is on and you do not specify a transformation list, the system-generated transformations are applied and embedded in the model.
- When ADP is off and you do not specify a transformation list, no transformations are embedded in the model; you must separately prepare the data sets you use for building, testing, and scoring the model.

Related Topics

- Embed Transformations in a Model
  You can specify your own transformations and embed them in a model by creating a transformation list and passing it to `DBMS_DATA_MINING.CREATE_MODEL2` or `DBMS_DATA_MINING.CREATE_MODEL`.
- *Oracle Database PL/SQL Packages and Types Reference*

Specify Transformation Instructions for an Attribute

You can pass transformation instructions for an attribute by defining a transformation list.

A transformation list is defined as a table of transformation records. Each record (`transform_rec`) specifies the transformation instructions for an attribute.

```
TYPE transform_rec IS RECORD (
    attribute_name      VARCHAR2(30),
    attribute_subname   VARCHAR2(4000),
    expression          EXPRESSION_REC,
    reverse_expression  EXPRESSION_REC,
    attribute_spec      VARCHAR2(4000));
```

The fields in a transformation record are described in this table.

Table 3-7 Fields in a Transformation Record for an Attribute

Field	Description
`attribute_name` and `attribute_subname`	These fields identify the attribute, as described in "Scoping of Model Attribute Name"
`expression`	A SQL expression for transforming the attribute. For example, this expression transforms the age attribute into two categories: child and adult:[0,19) for 'child' and [19,) for adult `CASE WHEN age < 19 THEN 'child' ELSE 'adult'` Expression and reverse expressions are stored in `expression_rec` objects. See "Expression Records" for details.
`reverse_expression`	A SQL expression for reversing the transformation. For example, this expression reverses the transformation of the age attribute: `DECODE(age,'child','(-Inf,19)','[19,Inf)')`
`attribute_spec`	Specifies special treatment for the attribute. The `attribute_spec` field can be null or it can have one or more of these values: • `FORCE_IN` — For GLM, forces the inclusion of the attribute in the model build when the `ftr_selection_enable` setting is enabled. (`ftr_selection_enable` is disabled by default.) If the model is not using GLM, this value has no effect. `FORCE_IN` cannot be specified for nested attributes or text. • `NOPREP` — When ADP is on, prevents automatic transformation of the attribute. If ADP is not on, this value has no effect. You can specify `NOPREP` for a nested attribute, but not for an individual subname (row) in the nested attribute. • `TEXT` — Indicates that the attribute contains unstructured text. ADP has no effect on this setting. `TEXT` may optionally include subsettings `POLICY_NAME`, `TOKEN_TYPE`, and `MAX_FEATURES`. See Example 3-5 and Example 3-6.

Related Topics

- Scoping of Model Attribute Name
  Learn about model attribute name.
- Expression Records
  Example of a transformation record.

Expression Records

Example of a transformation record.

The transformation expressions in a transformation record are `expression_rec` objects.

```
TYPE expression_rec IS RECORD (
     lstmt        DBMS_SQL.VARCHAR2A,
     lb           BINARY_INTEGER DEFAULT 1,
     ub           BINARY_INTEGER DEFAULT 0);

TYPE varchar2a IS TABLE OF VARCHAR2(32767)
INDEX BY BINARY_INTEGER;
```

The `lstmt` field stores a `VARCHAR2A`, which allows transformation expressions to be very long, as they can be broken up across multiple rows of `VARCHAR2`. Use the `DBMS_DATA_MINING_TRANSFORM.SET_EXPRESSION` procedure to create an `expression_rec`.

Attribute Specifications

Learn how to define the characteristics specific to an attribute through attribute specification.

The attribute specification in a transformation record defines characteristics that are specific to this attribute. If not null, the attribute specification can include values `FORCE_IN`, `NOPREP`, or `TEXT`, as described in Table 3-7.

Example 3-5 An Attribute Specification with Multiple Keywords

If more than one attribute specification keyword is applicable, you can provide them in a comma-delimited list. The following expression is the specification for an attribute in a GLM model. Assuming that the `ftr_selection_enable` setting is enabled, this expression forces the attribute to be included in the model. If ADP is on, automatic transformation of the attribute is not performed.

```
"FORCE_IN,NOPREP"
```

Example 3-6 A Text Attribute Specification

For text attributes, you can optionally specify subsettings `POLICY_NAME`, `TOKEN_TYPE`, and `MAX_FEATURES`. The subsettings provide configuration information that is specific to text transformation. In this example, the transformation instructions for the text content are defined in a text policy named `my_policy` with token type is `THEME`. The maximum number of extracted features is 3000.

```
"TEXT(POLICY_NAME:my_policy)(TOKEN_TYPE:THEME)(MAX_FEATURES:3000)"
```

Related Topics

- Configure a Text Attribute
  Provide transformation instructions for text attribute or unstructured text by explicitly identifying the column datatypes.

Oracle Machine Learning for SQL Transformation Routines

Learn about transformation routines.

OML4SQL provides routines that implement various transformation techniques in the `DBMS_DATA_MINING_TRANSFORM` package.

Related Topics

- *Oracle Database SQL Language Reference*

Binning Routines

Explains binning techniques in Oracle Machine Learning for SQL.

A number of factors go into deciding a binning strategy. Having fewer values typically leads to a more compact model and one that builds faster, but it can also lead to some loss in accuracy.

Model quality can improve significantly with well-chosen bin boundaries. For example, an appropriate way to bin ages is to separate them into groups of interest, such as children 0-13, teenagers 13-19, youth 19-24, working adults 24-35, and so on.

The following table lists the binning techniques provided by Oracle Machine Learning for SQL:

Table 3-8 Binning Methods in DBMS_DATA_MINING_TRANSFORM

Binning Method	Description
Top-N Most Frequent Items	You can use this technique to bin categorical attributes. You specify the number of bins. The value that occurs most frequently is labeled as the first bin, the value that appears with the next frequency is labeled as the second bin, and so on. All remaining values are in an additional bin.
Supervised Binning	Supervised binning is a form of intelligent binning, where bin boundaries are derived from important characteristics of the data. Supervised binning builds a single-predictor decision tree to find the interesting bin boundaries with respect to a target. It can be used for numerical or categorical attributes.
Equi-Width Binning	You can use equi-width binning for numerical attributes. The range of values is computed by subtracting the minimum value from the maximum value, then the range of values is divided into equal intervals. You can specify the number of bins or it can be calculated automatically. Equi-width binning must usually be used with outlier treatment.
Quantile Binning	Quantile binning is a numerical binning technique. Quantiles are computed using the SQL analytic function `NTILE`. The bin boundaries are based on the minimum values for each quantile. Bins with equal left and right boundaries are collapsed, possibly resulting in fewer bins than requested.

Related Topics

- Routines for Outlier Treatment
  Understand the transformations used for outlier treatment.

Normalization Routines

Learn about normalization routines in Oracle Machine Learning for SQL.

Most normalization methods map the range of a single attribute to another range, typically 0 to 1 or -1 to +1.

Normalization is very sensitive to outliers. Without outlier treatment, most values are mapped to a tiny range, resulting in a significant loss of information.

Table 3-9 Normalization Methods in DBMS_DATA_MINING_TRANSFORM

Transformation	Description
Min-Max Normalization	This technique computes the normalization of an attribute using the minimum and maximum values. The shift is the minimum value, and the scale is the difference between the maximum and minimum values.
Scale Normalization	This normalization technique also uses the minimum and maximum values. For scale normalization, shift = 0, and scale = max{abs(max), abs(min)}.
Z-Score Normalization	This technique computes the normalization of an attribute using the mean and the standard deviation. Shift is the mean, and scale is the standard deviation.

Related Topics

- Routines for Outlier Treatment
  Understand the transformations used for outlier treatment.

Outlier Treatment

Understand what you must do to treat outliers.

A value is considered an outlier if it deviates significantly from most other values in the column. The presence of outliers can have a skewing effect on the data and can interfere with the effectiveness of transformations such as normalization or binning.

Outlier treatment methods such as trimming or clipping can be implemented to minimize the effect of outliers.

Outliers represent problematic data, for example, a bad reading due to the unusual condition of an instrument. However, in some cases, especially in the business arena, outliers are perfectly valid. For example, in census data, the earnings for some of the richest individuals can vary significantly from the general population. Do not treat this information as an outlier, since it is an important part of the data. You need domain knowledge to determine outlier handling.

Routines for Outlier Treatment

Understand the transformations used for outlier treatment.

Outliers are extreme values, typically several standard deviations from the mean. To minimize the effect of outliers, you can Winsorize or trim the data.

Winsorizing involves setting the tail values of an attribute to some specified value. For example, for a 90% Winsorization, the bottom 5% of values are set equal to the minimum value in the 5th percentile, while the upper 5% of values are set equal to the maximum value in the 95th percentile.

Trimming sets the tail values to NULL. The algorithm treats them as missing values.

Outliers affect the different algorithms in different ways. In general, outliers cause distortion with equi-width binning and min-max normalization.

Table 3-10 Outlier Treatment Methods in DBMS_DATA_MINING_TRANSFORM

Transformation	Description
Trimming	This technique trims the outliers in numeric columns by sorting the non-null values, computing the tail values based on some fraction, and replacing the tail values with nulls.
Windsorizing	This technique trims the outliers in numeric columns by sorting the non-null values, computing the tail values based on some fraction, and replacing the tail values with some specified value.

Understand Reverse Transformations

Reverse transformations ensure that information returned by the model is expressed in a format that is similar to or the same as the format of the data that was used to train the model. Internal transformation are reversed in the model details and in the results of scoring.

Some of the attributes used by the model correspond to columns in the build data. However, because of logic specific to the algorithm, nested data, and transformations, some attributes do not correspond to columns.

For example, a nested column in the training data is not interpreted as an attribute by the model. During the model build,Oracle Machine Learning for SQL explodes nested columns, and each row (an attribute name/value pair) becomes an attribute.

Some algorithms, for example Support Vector Machine (SVM) and Generalized Linear Model (GLM), only operate on numeric attributes. Any non-numeric column in the build data is exploded into binary attributes, one for each distinct value in the column (SVM). GLM does not generate a new attribute for the most frequent value in the original column. These binary attributes are set to one only if the column value for the case is equal to the value associated with the binary attribute.

Algorithms that generate coefficients present challenges in interpreting the results. Examples are SVM and Non-Negative Matrix Factorization (NMF). These algorithms produce coefficients that are used in combination with the transformed attributes. The coefficients are relevant to the data on the transformed scale, not the original data scale.

For all these reasons, the attributes listed in the model details do not resemble the columns of data used to train the model. However, attributes that undergo embedded transformations, whether initiated by Automatic Data Preparation (ADP) or by a user-specified transformation list, appear in the model details in their pre-transformed state, as close as possible to the original column values. Although the attributes are transformed when they are used by the model, they are visible in the model details in a form that can be interpreted by a user.

Before Creating a Model

Explains the preparation steps before creating a model.

Models are database schema objects that perform machine learning. The `DBMS_DATA_MINING` PL/SQL package is the API for creating, configuring, evaluating, and querying machine learning models (model details).

Before you create a model, you must decide what you want the model to do. You must identify the training data and determine if transformations are required. You can specify model settings to influence the behavior of the model behavior. The preparation steps are summarized in the following table.

Table 3-11 Preparation for Creating an Oracle Machine Learning for SQL Model

Preparation Step	Description
Choose the machine learning function	See Choose the Machine Learning Technique
Choose the algorithm	See Choose the Algorithm
Identify the build (training) data	See Data Preparation
For classification and regression models, identify the test data	See Splitting the Data
Determine your data transformation strategy and create and populate a settings tables (if needed)	See Specify Model Settings

Related Topics

- About Oracle Machine Learning Models
  Machine learning models are database schema objects that perform machine learning techniques.
- DBMS_DATA_MINING
  The DBMS_DATA_MINING package contains routines for creating machine learning models, for performing operations on the models, and for querying them.

Choose the Machine Learning Technique

Describes providing an Oracle Machine Learning for SQL machine learning function for the CREATE_MODEL and CREATE_MODEL2 procedure.

An OML4SQL machine learning technique specifies a class of problems that can be modeled and solved. You specify a machine learning with the mining_function argument of the CREATE_MODEL and CREATE_MODEL2 procedure.

OML4SQL machine learning functions implement either **supervised** or **unsupervised** learning. Supervised learning uses a set of independent attributes to predict the value of a dependent attribute or **target**. Unsupervised learning does not distinguish between dependent and independent attributes. Supervised functions are predictive. Unsupervised functions are descriptive.

> **Note:**
>
> In OML4SQL terminology, a **function** is a general type of problem to be solved by a given approach to machine learning. In SQL language terminology, a **function** is an operation that returns a result.
>
> In OML4SQL documentation, the term **function**, or **machine learning function** refers to an OML4SQL machine learning function; the term **SQL function** or **SQL machine learning function** refers to a SQL function for scoring (applying machine learning models).

You can specify any of the values in the following table for the *mining_function* parameter to the CREATE_MODEL and CREATE_MODEL2 procedure.

Table 3-12 Oracle Machine Learning mining_function Values

***mining_function* Value**	**Description**
ASSOCIATION	Association is a descriptive machine learning function. An association model identifies relationships and the probability of their occurrence within a data set (association rules). Association models use the Apriori algorithm.
ATTRIBUTE_IMPORTANCE	Attribute importance is a predictive machine learning function. An attribute importance model identifies the relative importance of attributes in predicting a given outcome. Attribute importance models use the Minimum Description Length algorithm and CUR Matrix Decomposition.
CLASSIFICATION	Classification is a predictive machine learning function. A classification model uses historical data to predict a categorical target. Classification models can use Naive Bayes, Neural Network, Decision Tree, logistic regression, Random Forest, Support Vector Machine, Explicit Semantic Analysis, or XGBoost. The default is Naive Bayes. You can also specify the classification machine learning function for anomaly detection for a One-Class SVM model and a Multivariate State Estimation Technique - Sequential Probability Ratio Test model.
CLUSTERING	Clustering is a descriptive machine learning function. A clustering model identifies natural groupings within a data set. Clustering models can use *k*-Means, O-Cluster, or Expectation Maximization. The default is *k*-Means.
FEATURE_EXTRACTION	Feature extraction is a descriptive machine learning function. A feature extraction model creates a set of optimized attributes. Feature extraction models can use Non-Negative Matrix Factorization, Singular Value Decomposition (which can also be used for Principal Component Analysis) or Explicit Semantic Analysis. The default is Non-Negative Matrix Factorization.
REGRESSION	Regression is a predictive machine learning function. A regression model uses historical data to predict a numerical target. Regression models can use Support Vector Machine, GLM regression, or XGBoost. The default is Support Vector Machine.
TIME_SERIES	Time series is a predictive machine learning function. A time series model forecasts the future values of a time-ordered series of historical numeric data over a user-specified time window. Time series models use the Exponential Smoothing algorithm. The default is Exponential Smoothing.

Choose the Algorithm

Learn about providing the algorithm settings for a model.

The ALGO_NAME setting specifies the algorithm for a model. If you use the default algorithm for the machine learning technique, or if there is only one algorithm available for the machine learning technique, then you do not need to specify the ALGO_NAME setting.

Table 3-13 Oracle Machine Learning Algorithms

ALGO_NAME Value	Algorithm	Default?	Machine Learning Model Function
ALGO_AI_MDL	Minimum Description Length	—	Attribute importance
ALGO_APRIORI_ASSOCIATION_RULES	Apriori	—	Association
ALGO_CUR_DECOMPOSITION	CUR Matrix Decomposition	—	Attribute importance
ALGO_DECISION_TREE	Decision Tree	—	Classification
ALGO_EXPECTATION_MAXIMIZATION	Expectation Maximization	—	Clustering and Anomaly Detection
ALGO_EXPLICIT_SEMANTIC_ANALYS	Explicit Semantic Analysis	—	Feature extraction and classification
ALGO_EXPONENTIAL_SMOOTHING	Exponential Smoothing	—	Time series and time series regression
ALGO_EXTENSIBLE_LANG	Language used for an extensible algorithm	—	All machine learning functions are supported
ALGO_GENERALIZED_LINEAR_MODEL	Generalized Linear Model	—	Classification and regression
ALGO_KMEANS	*k*-Means	yes	Clustering
ALGO_MSET_SPRT	Multivariate State Estimation Technique - Sequential Probability Ratio Test	—	Anomaly detection (classification with no target)
ALGO_NAIVE_BAYES	Naive Bayes	yes	Classification
ALGO_NEURAL_NETWORK	Neural Network	—	Classification
ALGO_NONNEGATIVE_MATRIX_FACTOR	Non-Negative Matrix Factorization	yes	Feature extraction
ALGO_O_CLUSTER	O-Cluster	—	Clustering
ALGO_RANDOM_FOREST	Random Forest	—	Classification
ALGO_SINGULAR_VALUE_DECOMP	Singular Value Decomposition (can also be used for Principal Component Analysis)	—	Feature extraction
ALGO_SUPPORT_VECTOR_MACHINES	Support Vector Machine	yes	Default regression algorithm; regression, classification, and anomaly detection (classification with no target)
ALGO_XGBOOST	XGBoost	—	Classification and regression

The CREATE_MODEL2 Procedure

The CREATE_MODEL2 procedure of the DBMS_DATA_MINING package is a procedure for defining model settings to build a model.

By using the CREATE_MODEL2 procedure, the user does not need to create transient database objects. The model can use configuration settings and user-specified transformations. In the CREATE_MODEL2 procedure, the input is a table or a view and if such an object is not already present, the user must create it.

```
DBMS_DATA_MINING.CREATE_MODEL2 (
model_name            IN VARCHAR2,
mining_function       IN VARCHAR2,
data_query            IN CLOB,
set_list              IN SETTING_LIST,
case_id_column_name   IN VARCHAR2 DEFAULT NULL,
target_column_name    IN VARCHAR2 DEFAULT NULL,
xform_list            IN TRANSFORM_LIST DEFAULT NULL);
```

The `data_query` parameter species a query which provides training data for building the model. The `set_list` parameter specifies the `SETTING_LIST`. `SETTING_LIST` is a table of CLOB index by `VARCHAR2(30)`; Where the index is the setting name and the CLOB is the setting value for that name. The rest of the parameters are covered in the `CREATE_MODEL` procedure.

You can also rename the model using the `RENAME_MODEL` procedure of the `DBMS_DATA_MINING` package. The procedure changes the value of the machine learning model specified against `MODEL_NAME` with another name that you specify.

The following `CREATE_MODEL2` procedure builds a classification model using SVM algorithm. The following example mining_data_build_v data set to arrive at likelihood of customers opting the affinity card program. .

```
DECLARE
    v_setlist DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlist('PREP_AUTO') := 'ON';
    v_setlist('ALGO_NAME') := 'ALGO_SUPPORT_VECTOR_MACHINES';
    v_setlist('SVMS_KERNEL_FUNCTION') := 'SVMS_LINEAR';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'SVM_MODEL',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'select * from mining_data_build_v',
        SET_LIST            => v_setlist,
        CASE_ID_COLUMN_NAME => 'CUST_ID,
    TARGET_COLUMN_NAME  => 'AFFINITY_CARD');
END;
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*
- RENAME_MODEL Procedure

The CREATE_MODEL Procedure

The `CREATE_MODEL` procedure of the `DBMS_DATA_MINING` package uses the specified data to create a machine learning model with the specified name and machine learning function.

The model can be created with configuration settings and user-specified transformations.

```
PROCEDURE CREATE_MODEL(
                  model_name            IN VARCHAR2,
                  mining_function       IN VARCHAR2,
                  data_table_name       IN VARCHAR2,
                  case_id_column_name   IN VARCHAR2,
                  target_column_name    IN VARCHAR2 DEFAULT NULL,
                  settings_table_name   IN VARCHAR2 DEFAULT NULL,
                  data_schema_name      IN VARCHAR2 DEFAULT NULL,
                  settings_schema_name  IN VARCHAR2 DEFAULT NULL,
                  xform_list            IN TRANSFORM_LIST DEFAULT NULL);
```

You can also rename the model using the `RENAME_MODEL` procedure of the `DBMS_DATA_MINING` package. The procedure changes the value of the machine learning model specified against `MODEL_NAME` with another name that you specify.

The following example builds a classification model using the Support Vector Machine algorithm.

```
 Create the settings table
CREATE TABLE svm_model_settings (
  setting_name  VARCHAR2(30),
  setting_value VARCHAR2(30));

-- Populate the settings table
-- Specify SVM. By default, Naive Bayes is used for classification.
-- Specify ADP. By default, ADP is not used.
BEGIN
  INSERT INTO svm_model_settings (setting_name, setting_value) VALUES
     (dbms_data_mining.algo_name, dbms_data_mining.algo_support_vector_machines);
  INSERT INTO svm_model_settings (setting_name, setting_value) VALUES
     (dbms_data_mining.prep_auto,dbms_data_mining.prep_auto_on);
  COMMIT;
END;
/
-- Create the model using the specified settings
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'svm_model',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_data_build_v',
    case_id_column_name => 'cust_id',
    target_column_name  => 'affinity_card',
    settings_table_name => 'svm_model_settings');
END;
/
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*
- RENAME_MODEL Procedure

About Scoring and Deployment

Scoring is the application of models to new data. In Oracle Machine Learning for SQL, scoring is performed by SQL language functions.

Predictive functions perform classification, regression, or anomaly detection. Clustering functions assign rows to clusters. Feature extraction functions transform the input data to a set of higher order predictors. A scoring procedure is also available in the `DBMS_DATA_MINING` PL/SQL package.

Deployment refers to the use of models in a target environment. Once the models have been built, the challenges come in deploying them to obtain the best results, and in maintaining them within a production environment. Deployment can be any of the following:

- Scoring data either for batch or real-time results. Scores can include predictions, probabilities, rules, and other statistics.
- Extracting model details to produce reports. For example: clustering rules, decision tree rules, or attribute rankings from an Attribute Importance model.

- Extending the business intelligence infrastructure of a data warehouse by incorporating machine learning results in applications or operational systems.
- Moving a model from the database where it was built to the database where it used for scoring (export/import)

OML4SQL supports all of these deployment scenarios.

> **Note:**
>
> OML4SQL scoring operations support parallel execution. When parallel execution is enabled, multiple CPU and I/O resources are applied to the execution of a single database operation.
>
> Parallel execution offers significant performance improvements, especially for operations that involve complex queries and large databases typically associated with decision support systems (DSS) and data warehouses.

Related Topics

- *Oracle Database VLDB and Partitioning Guide*
- *Oracle Machine Learning for SQL Concepts*
- Export and Import Oracle Machine Learning for SQL Models
  You can export machine learning models to move models to a different Oracle Database instance, such as from a development database to a production database.

Use the Oracle Machine Learning for SQL Functions

Some of the benefits of using SQL functions for Oracle Machine Learning for SQL are listed.

The OML4SQL functions provide the following benefits:

- Models can be easily deployed within the context of existing SQL applications.
- Scoring operations take advantage of existing query execution functionality. This provides performance benefits.
- Scoring results are pipelined, enabling the rows to be processed without requiring materialization.

The machine learning functions produce a score for each row in the selection. The functions can apply a machine learning model schema object to compute the score, or they can score dynamically without a pre-defined model, as described in "Dynamic Scoring".

Related Topics

- Dynamic Scoring
  You can perform dynamic scoring if, for some reason, you do not want to apply a predefined model.
- Scoring Requirements
  Learn how scoring is done in Oracle Machine Learning for SQL.
- Oracle Machine Learning for SQL Scoring Functions
  Use OML4SQL functions score data. Functions can apply a machine learning model schema object to data or dynamically mine it with an analytic clause. SQL functions exist for all OML4SQL scoring algorithms.

- *Oracle Database SQL Language Reference*

Choose the Predictors

You can select different attributes as predictors in a `PREDICTION` function through a `USING` clause.

The OML4SQL functions support a `USING` clause that specifies which attributes to use for scoring. You can specify some or all of the attributes in the selection and you can specify expressions. The following examples all use the `PREDICTION` function to find the customers who are likely to use an affinity card, but each example uses a different set of predictors.

When predictor values are not in the training data, the models score categorical values that were not in the training data without error. A score is produced using the remaining predictors. This enables batch scoring that does not fail because of a single record with an invalid value. Also, in some algorithms, like k-Means or Gaussian SVM, a new value can change the prediction in a meaningful way, such as resulting in larger distances with the unknown value. Furthermore, additional columns that were not present for building may be present in the table or view provided for scoring, and only the columns matching the model signature are used. Also, scoring may be performed with fewer predictors than are listed in the model signature.

In the case of partitioned models, a `NULL` score is produced if the partition value is invalid. If the partition column value is omitted, an error message is returned.

The query in Example 3-7 uses all the predictors.

The query in Example 3-8 uses only gender, marital status, occupation, and income as predictors.

The query in Example 3-9 uses three attributes and an expression as predictors. The prediction is based on gender, marital status, occupation, and the assumption that all customers are in the highest income bracket.

Example 3-7 Using All Predictors

The dt_sh_clas_sample model is created by the `oml4sql-classification-decision-tree.sql` example.

```
SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
     FROM mining_data_apply_v
     WHERE PREDICTION(dt_sh_clas_sample USING *) = 1
   GROUP BY cust_gender
   ORDER BY cust_gender;
```

The output is follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         25         38
M        213         43
```

Example 3-8 Using Some Predictors

```
 SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
     FROM mining_data_apply_v
     WHERE PREDICTION(dt_sh_clas_sample USING
                     cust_gender,cust_marital_status,
                   occupation, cust_income_level) = 1
   GROUP BY cust_gender
   ORDER BY cust_gender;
```

The output is as follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         30         38
M        186         43
```

Example 3-9 Using Some Predictors and an Expression

```
SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
     FROM mining_data_apply_v
     WHERE PREDICTION(dt_sh_clas_sample USING
                     cust_gender, cust_marital_status, occupation,
                    'L: 300,000 and above' AS cust_income_level) = 1
   GROUP BY cust_gender
   ORDER BY cust_gender;
```

The output is follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         30         38
M        186         43
```

Single-Record Scoring

You can score a single record which produces 0 and 1 to predict customers who are unlikely or likely to use an affinity card.

The Oracle Machine Learning for SQL functions can produce a score for a single record, as shown in Example 3-10 and Example 3-11.

Example 3-10 returns a prediction for customer 102001 by applying the classification model NB_SH_Clas_sample. The resulting score is 0, meaning that this customer is unlikely to use an affinity card. The NB_SH_Clas_Sample model is created by the oml4sql-classification-naive-bayes.sql example.

Example 3-11 returns a prediction for 'Affinity card is great' as the comments attribute by applying the text machine learning model T_SVM_Clas_sample. The resulting score is 1, meaning that this customer is likely to use an affinity card. The T_SVM_Clas_sample model is created by the oml4sql-classification-text-analysis-svm.sql example.

Example 3-10 Scoring a Single Customer or a Single Text Expression

```
SELECT PREDICTION (NB_SH_Clas_Sample USING *)
    FROM sh.customers where cust_id = 102001;
```

The output is as follows:

```
PREDICTION(NB_SH_CLAS_SAMPLEUSING*)
```

```
-----------------------------------
                                  0
```

Example 3-11 Scoring a Single Text Expression

```
SELECT
  PREDICTION(T_SVM_Clas_sample USING 'Affinity card is great' AS comments)
FROM DUAL;
```

The output is as follows:

```
PREDICTION(T_SVM_CLAS_SAMPLEUSING'AFFINITYCARDISGREAT'ASCOMMENTS)
-----------------------------------------------------------------
                                                                1
```

Prediction Details

Prediction details are XML strings that provide information about the score.

Details are available for all types of scoring: clustering, feature extraction, classification, regression, and anomaly detection. Details are available whether scoring is dynamic or the result of model apply.

The details functions, `CLUSTER_DETAILS`, `FEATURE_DETAILS`, and `PREDICTION_DETAILS` return the actual value of attributes used for scoring and the relative importance of the attributes in determining the score. By default, the functions return the five most important attributes in descending order of importance.

Cluster Details

Shows an example of the `CLUSTER_DETAILS` function.

For the most likely cluster assignments of customer 100955 (probability of assignment > 20%), the query in the following example produces the five attributes that have the most impact for each of the likely clusters. The clustering functions apply an Expectation Maximization model named em_sh_clus_sample to the data selected from `mining_data_apply_v`. The "5" specified in `CLUSTER_DETAILS` is not required, because five attributes are returned by default. The em_sh_clus_sample model is created by the `oml4sql-clustering-expectation-maximization.sql` example.

Example 3-12 Cluster Details

```
SELECT S.cluster_id, probability prob,
           CLUSTER_DETAILS(em_sh_clus_sample, S.cluster_id, 5 USING T.*) det
    FROM
      (SELECT v.*, CLUSTER_SET(em_sh_clus_sample, NULL, 0.2 USING *) pset
        FROM mining_data_apply_v v
       WHERE cust_id = 100955) T,
      TABLE(T.pset) S
    ORDER BY 2 DESC;
```

The output is as follows:

```
CLUSTER_ID  PROB DET
```

```
---------- -----
----------------------------------------------------------------------------
        14 .6761 <Details algorithm="Expectation Maximization" cluster="14">
                 <Attribute name="AGE" actualValue="51" weight=".676"
rank="1"/>
                 <Attribute name="HOME_THEATER_PACKAGE" actualValue="1"
weight=".557" rank="2"/>
                 <Attribute name="FLAT_PANEL_MONITOR" actualValue="0"
weight=".412" rank="3"/>
                 <Attribute name="Y_BOX_GAMES" actualValue="0" weight=".171"
rank="4"/>
                 <Attribute name="BOOKKEEPING_APPLICATION"actualValue="1"
weight="-.003"
                  rank="5"/>
                 </Details>

         3 .3227 <Details algorithm="Expectation Maximization" cluster="3">
                 <Attribute name="YRS_RESIDENCE" actualValue="3"
weight=".323" rank="1"/>
                 <Attribute name="BULK_PACK_DISKETTES" actualValue="1"
weight=".265" rank="2"/>
                 <Attribute name="EDUCATION" actualValue="HS-grad"
weight=".172" rank="3"/>
                 <Attribute name="AFFINITY_CARD" actualValue="0"
weight=".125" rank="4"/>
                 <Attribute name="OCCUPATION" actualValue="Crafts"
weight=".055" rank="5"/>
                 </Details>
```

Feature Details

Shows an example of the `FEATURE_DETAILS` function.

The query in the following example returns the three attributes that have the greatest impact on the top Principal Components Analysis (PCA) projection for customer 101501. The `FEATURE_DETAILS` function applies a Singular Value Decomposition (SVD) model named svd_sh_sample to the data selected from the svd_sh_sample_build_num table. The table and model are created by the `oml4sql-singular-value-decomposition.sql` example.

Example 3-13   Feature Details

```
SELECT FEATURE_DETAILS(svd_sh_sample, 1, 3 USING *) proj1det
  FROM svd_sh_sample_build_num
  WHERE CUST_ID = 101501;
```

The output is as follows:

```

PROJ1DET
------------------------------------------------------------------------------
--
<Details algorithm="Singular Value Decomposition" feature="1">
<Attribute name="HOME_THEATER_PACKAGE" actualValue="1" weight=".352"
rank="1"/>
<Attribute name="Y_BOX_GAMES" actualValue="0" weight=".249" rank="2"/>
```

```
<Attribute name="AGE" actualValue="41" weight=".063" rank="3"/>
</Details>
```

Prediction Details

Shows an examples of `PREDICTION_DETAILS` function.

The query in the following example returns the attributes that are most important in predicting the age of customer 100010. The prediction functions apply a Generalized Linear Model regression model named GLMR_SH_Regr_sample to the data selected from `mining_data_apply_v`. The GLMR_SH_Regr_sample model is created by the `oml4sql-regression-glm.sql` example.

Example 3-14 Prediction Details for Regression

```
SELECT cust_id,
      PREDICTION(GLMR_SH_Regr_sample USING *) pr,
      PREDICTION_DETAILS(GLMR_SH_Regr_sample USING *) pd
  FROM mining_data_apply_v
  WHERE CUST_ID = 100010;
```

The output is as follows:

```
CUST_ID    PR PD
-------- ----- -----------
 100010 25.45 <Details algorithm="Generalized Linear Model">
              <Attribute name="FLAT_PANEL_MONITOR" actualValue="1"
weight=".025" rank="1"/>
              <Attribute name="OCCUPATION" actualValue="Crafts" weight=".019"
rank="2"/>
              <Attribute name="AFFINITY_CARD" actualValue="0" weight=".01"
rank="3"/>
              <Attribute name="OS_DOC_SET_KANJI" actualValue="0" weight="0"
rank="4"/>
              <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight="-.004" rank="5"/>
              </Details>
```

The query in the following example returns the customers who work in Tech Support and are likely to use an affinity card (with more than 85% probability). The prediction functions apply an Support Vector Machine (SVM) classification model named svmc_sh_clas_sample. to the data selected from `mining_data_apply_v`. The query includes the prediction details, which show that education is the most important predictor. The svmc_sh_clas_sample model is created by the `oml4sql-classification-svm.sql` example.

Example 3-15 Prediction Details for Classification

```
SELECT cust_id, PREDICTION_DETAILS(svmc_sh_clas_sample, 1 USING *) PD
      FROM mining_data_apply_v
  WHERE PREDICTION_PROBABILITY(svmc_sh_clas_sample, 1 USING *) > 0.85
  AND occupation = 'TechSup'
  ORDER BY cust_id;
```

The output is as follows:

```
CUST_ID PD
-------
------------------------------------------------------------------------------
---------
 100029 <Details algorithm="Support Vector Machines" class="1">
        <Attribute name="EDUCATION" actualValue="Assoc-A" weight=".199"
rank="1"/>
        <Attribute name="CUST_INCOME_LEVEL" actualValue="I: 170\,000 -
189\,999" weight=".044"
         rank="2"/>
        <Attribute name="HOME_THEATER_PACKAGE" actualValue="1" weight=".028"
rank="3"/>
        <Attribute name="BULK_PACK_DISKETTES" actualValue="1" weight=".024"
rank="4"/>
        <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight=".022" rank="5"/>
        </Details>

 100378 <Details algorithm="Support Vector Machines" class="1">
        <Attribute name="EDUCATION" actualValue="Assoc-A" weight=".21"
rank="1"/>
        <Attribute name="CUST_INCOME_LEVEL" actualValue="B: 30\,000 -
49\,999" weight=".047"
         rank="2"/>
        <Attribute name="FLAT_PANEL_MONITOR" actualValue="0" weight=".043"
rank="3"/>
        <Attribute name="HOME_THEATER_PACKAGE" actualValue="1" weight=".03"
rank="4"/>
        <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight=".023" rank="5"/>
        </Details>

 100508 <Details algorithm="Support Vector Machines" class="1">
        <Attribute name="EDUCATION" actualValue="Bach." weight=".19"
rank="1"/>
        <Attribute name="CUST_INCOME_LEVEL" actualValue="L: 300\,000 and
above" weight=".046"
         rank="2"/>
        <Attribute name="HOME_THEATER_PACKAGE" actualValue="1" weight=".031"
rank="3"/>
        <Attribute name="BULK_PACK_DISKETTES" actualValue="1" weight=".026"
rank="4"/>
        <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight=".024" rank="5"/>
        </Details>

 100980 <Details algorithm="Support Vector Machines" class="1">
        <Attribute name="EDUCATION" actualValue="Assoc-A" weight=".19"
rank="1"/>
        <Attribute name="FLAT_PANEL_MONITOR" actualValue="0" weight=".038"
rank="2"/>
        <Attribute name="HOME_THEATER_PACKAGE" actualValue="1" weight=".026"
```

```
rank="3"/>
        <Attribute name="BULK_PACK_DISKETTES" actualValue="1" weight=".022"
rank="4"/>
        <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight=".02" rank="5"/>
        </Details>
```

The query in the following example returns the two customers that differ the most from the rest of the customers. The prediction functions apply an anomaly detection model named SVMO_SH_Clas_sample to the data selected from `mining_data_apply_v`. anomaly detection uses a one-class SVM classifier. The model is created by the `oml4sql-singular-value-decomposition.sql` example.

Example 3-16 Prediction Details for Anomaly Detection

```
SELECT cust_id, pd FROM
  (SELECT cust_id,
        PREDICTION_DETAILS(SVMO_SH_Clas_sample, 0 USING *) pd,
        RANK() OVER (ORDER BY prediction_probability(
              SVMO_SH_Clas_sample, 0 USING *) DESC, cust_id) rnk
  FROM mining_data_one_class_v)
  WHERE rnk <= 2
  ORDER BY rnk;
```

The output is as follows:

```
  CUST_ID PD
----------
------------------------------------------------------------------------------
-----
    102366 <Details algorithm="Support Vector Machines" class="0">
           <Attribute name="COUNTRY_NAME" actualValue="United Kingdom"
weight=".078" rank="1"/>
           <Attribute name="CUST_MARITAL_STATUS" actualValue="Divorc."
weight=".027" rank="2"/>
           <Attribute name="CUST_GENDER" actualValue="F" weight=".01"
rank="3"/>
           <Attribute name="HOUSEHOLD_SIZE" actualValue="9+" weight=".009"
rank="4"/>
           <Attribute name="AGE" actualValue="28" weight=".006" rank="5"/>
           </Details>

    101790 <Details algorithm="Support Vector Machines" class="0">
           <Attribute name="COUNTRY_NAME" actualValue="Canada" weight=".068"
rank="1"/>
           <Attribute name="HOUSEHOLD_SIZE" actualValue="4-5" weight=".018"
rank="2"/>
           <Attribute name="EDUCATION" actualValue="7th-8th" weight=".015"
rank="3"/>
           <Attribute name="CUST_GENDER" actualValue="F" weight=".013"
rank="4"/>
           <Attribute name="AGE" actualValue="38" weight=".001" rank="5"/>
           </Details>
```

GROUPING Hint

OML4SQL functions include PREDICTION*, CLUSTER*, FEATURE*, and ORA_DM_*. The GROUPING hint is an optional hint that applies to machine learning scoring functions when scoring partitioned models.

This hint results in partitioning the input data set into distinct data slices so that each partition is scored in its entirety before advancing to the next partition. However, parallelism by partition is still available. Data slices are determined by the partitioning key columns used when the model was built. This method can be used with any machine learning function against a partitioned model. The hint may yield a query performance gain when scoring large data that is associated with many partitions but may negatively impact performance when scoring large data with few partitions on large systems. Typically, there is no performance gain if you use the hint for single row queries.

Enhanced PREDICTION Function Command Format

```
<prediction function> ::=
    PREDICTION <left paren> /*+ GROUPING */ <prediction model>
        [ <comma> <class value> [ <comma> <top N> ] ]
        USING <machine learning attribute list> <right paren>
```

The syntax for only the PREDICTION function is given but it is applicable to any machine learning function in which PREDICTION, CLUSTERING, and FEATURE_EXTRACTION scoring functions occur.

Example 3-17   Example

```
SELECT PREDICTION(/*+ GROUPING */my_model USING *) pred FROM <input table>;
```

Related Topics

- *Oracle Database SQL Language Reference*

In-Database Scoring

In-database scoring applies machine learning models to new data within the database, ensuring security, efficiency, and ease of integration with applications.

Scoring is the application of a machine learning algorithm to new data. In Oracle Machine Learning for SQL scoring engine and the data both reside within the database. In traditional machine learning, models are built using specialized software on a remote system and deployed to another system for scoring. This is a cumbersome, error-prone process open to security violations and difficulties in data synchronization.

With Oracle Machine Learning for SQL, scoring is simple and secure. The scoring engine and the data both reside within the database. Scoring is an extension to the SQL language, so the results of machine learning can easily be incorporated into applications and reporting systems.

Parallel Execution and Ease of Administration

Parallel execution and in-database scoring provide performance advantages and simplify model deployment, ensuring efficient handling of large data sets.

All Oracle Machine Learning for SQL scoring routines support parallel execution for scoring large data sets.

In-database scoring provides performance advantages. All Oracle Machine Learning for SQL scoring routines support parallel execution, which significantly reduces the time required for executing complex queries and scoring large data sets.

In-database machine learning minimizes the IT effort needed to support Oracle Machine Learning for SQL initiatives. Using standard database techniques, models can easily be refreshed (re-created) on more recent data and redeployed. The deployment is immediate since the scoring query remains the same; only the underlying model is replaced in the database.

Related Topics

- *Oracle Database VLDB and Partitioning Guide*

SQL Functions for Model Apply and Dynamic Scoring

In Oracle Machine Learning for SQL, scoring is performed by SQL language functions. Understand the different ways of scoring using SQL functions.

The functions perform prediction, clustering, and feature extraction. The functions can be loaded in two different ways: By applying a machine learning model object (Example 3-18), or by running an analytic clause that computes the machine learning analysis dynamically and applies it to the data (Example 3-19). Dynamic scoring, which eliminates the need for a model, can supplement, or even replace, the more traditional methodology described in "The Machine Learning Process".

In Example 3-18, the `PREDICTION_PROBABILITY` function applies the model svmc_sh_clas_sample, created in Example 5-1, to score the data in `mining_data_apply_v`. The function returns the ten customers in Italy who are most likely to use an affinity card.

In Example 3-19, the functions `PREDICTION` and `PREDICTION_PROBABILITY` use the analytic syntax (the `OVER` () clause) to dynamically score the data in `mining_data_apply_v`. The query returns the customers who currently do not have an affinity card with the probability that they are likely to use.

Example 3-18 Applying a Oracle Machine Learning for SQL Model to Score Data

```
SELECT cust_id FROM
  (SELECT cust_id,
        rank() over (order by PREDICTION_PROBABILITY(svmc_sh_clas_sample, 1
                     USING *) DESC, cust_id) rnk
   FROM mining_data_apply_v
   WHERE country_name = 'Italy')
WHERE rnk <= 10
ORDER BY rnk;
```

```
 CUST_ID
----------
    101445
    100179
    100662
    100733
    100554
    100081
```

```
     100344
     100324
     100185
     101345
```

Example 3-19   Executing an Analytic Function to Score Data

```
SELECT cust_id, pred_prob FROM
  (SELECT cust_id, affinity_card,
    PREDICTION(FOR TO_CHAR(affinity_card) USING *) OVER () pred_card,
    PREDICTION_PROBABILITY(FOR TO_CHAR(affinity_card),1 USING *) OVER () pred_prob
   FROM mining_data_build_v)
WHERE affinity_card = 0
AND pred_card = 1
ORDER BY pred_prob DESC;
```

```
 CUST_ID PRED_PROB
----------- ---------
     102434       .96
     102365       .96
     102330       .96
     101733       .95
     102615       .94
     102686       .94
     102749       .93
     .
     .
     .
     101656       .51
```

Dynamic Scoring

You can perform dynamic scoring if, for some reason, you do not want to apply a predefined model.

The Oracle Machine Learning for SQL functions operate in two modes: by applying a predefined model, or by executing an analytic clause. If you supply an analytic clause instead of a model name, the function builds one or more transient models and uses them to score the data.

The ability to score data dynamically without a predefined model extends the application of basic embedded machine learning techniques into environments where models are not available. Dynamic scoring, however, has limitations. The transient models created during dynamic scoring are not available for inspection or fine tuning. Applications that require model inspection, the correlation of scoring results with the model, special algorithm settings, or multiple scoring queries that use the same model, require a predefined model.

The following example shows a dynamic scoring query. The example identifies the rows in the input data that contain unusual customer age values.

Example 3-20   Dynamic Prediction

```
SELECT cust_id, age, pred_age, age-pred_age age_diff, pred_det FROM
 (SELECT cust_id, age, pred_age, pred_det,
    RANK() OVER (ORDER BY ABS(age-pred_age) DESC) rnk FROM
    (SELECT cust_id, age,
         PREDICTION(FOR age USING *) OVER () pred_age,
         PREDICTION_DETAILS(FOR age ABS USING *) OVER () pred_det
  FROM mining_data_apply_v))
WHERE rnk <= 5;
```

The output is follows:

```
CUST_ID  AGE   PRED_AGE AGE_DIFF PRED_DET
-------- ---- ----------- --------
--------------------------------------------------------------
 100910   80 40.6686505    39.33 <Details algorithm="Support Vector Machines">
                                 <Attribute name="HOME_THEATER_PACKAGE"
actualValue="1"
                                  weight=".059" rank="1"/>
                                 <Attribute name="Y_BOX_GAMES" actualValue="0"
                                  weight=".059" rank="2"/>
                                 <Attribute name="AFFINITY_CARD"
actualValue="0"
                                  weight=".059" rank="3"/>
                                 <Attribute name="FLAT_PANEL_MONITOR"
actualValue="1"
                                  weight=".059" rank="4"/>
                                 <Attribute name="YRS_RESIDENCE"
actualValue="4"
                                  weight=".059" rank="5"/>
                                  </Details>

 101285   79 42.1753571    36.82 <Details algorithm="Support Vector Machines">
                                 <Attribute name="HOME_THEATER_PACKAGE"
actualValue="1"
                                  weight=".059" rank="1"/>
                                 <Attribute name="HOUSEHOLD_SIZE"
actualValue="2" weight=".059"
                                  rank="2"/>
                                 <Attribute name="CUST_MARITAL_STATUS"
actualValue="Mabsent"
                                  weight=".059" rank="3"/>
                                 <Attribute name="Y_BOX_GAMES"
actualValue="0" weight=".059"
                                  rank="4"/>
                                 <Attribute name="OCCUPATION"
actualValue="Prof." weight=".059"
                                  rank="5"/>
                                 </Details>

 100694   77 41.0396722    35.96 <Details algorithm="Support Vector Machines">
                                 <Attribute name="HOME_THEATER_PACKAGE"
actualValue="1"
                                  weight=".059" rank="1"/>
                                 <Attribute name="EDUCATION"
actualValue="&lt; Bach."
                                  weight=".059" rank="2"/>
                                 <Attribute name="Y_BOX_GAMES"
actualValue="0" weight=".059"
```

```
                                  rank="3"/>
                                 <Attribute name="CUST_ID"
actualValue="100694" weight=".059"
                                  rank="4"/>
                                 <Attribute name="COUNTRY_NAME"
actualValue="United States of
                                  America" weight=".059" rank="5"/>
                                 </Details>

 100308   81 45.3252491    35.67 <Details algorithm="Support Vector Machines">
                                 <Attribute name="HOME_THEATER_PACKAGE"
actualValue="1"
                                  weight=".059" rank="1"/>
                                 <Attribute name="Y_BOX_GAMES"
actualValue="0" weight=".059"
                                  rank="2"/>
                                 <Attribute name="HOUSEHOLD_SIZE"
actualValue="2" weight=".059"
                                  rank="3"/>
                                 <Attribute name="FLAT_PANEL_MONITOR"
actualValue="1"
                                  weight=".059" rank="4"/>
                                 <Attribute name="CUST_GENDER"
actualValue="F" weight=".059"
                                  rank="5"/>
                                 </Details>

 101256   90 54.3862214    35.61 <Details algorithm="Support Vector Machines">
                                 <Attribute name="YRS_RESIDENCE"
actualValue="9" weight=".059"
                                  rank="1"/>
                                 <Attribute name="HOME_THEATER_PACKAGE"
actualValue="1"
                                  weight=".059" rank="2"/>
                                 <Attribute name="EDUCATION"
actualValue="&lt; Bach."
                                  weight=".059" rank="3"/>
                                 <Attribute name="Y_BOX_GAMES"
actualValue="0" weight=".059"
                                  rank="4"/>
                                 <Attribute name="COUNTRY_NAME"
actualValue="United States of
                                  America" weight=".059" rank="5"/>
                                 </Details>
```

Real-Time Scoring

You can perform real-time scoring by running a SQL query. An example shows a real-time query using `PREDICTION_PROBABILITY` function. Based on the result, a customer representative can offer a value card to the customer.

Oracle Machine Learning for SQL functions enable prediction, clustering, and feature extraction analysis to be easily integrated into live production and operational systems. Because machine learning results are returned within SQL queries, machine learning can occur in real time.

With real-time scoring, point-of-sales database transactions can be mined. Predictions and rule sets can be generated to help front-line workers make better analytical decisions. Real-time scoring enables fraud detection, identification of potential liabilities, and recognition of better marketing and selling opportunities.

The query in the following example uses a Decision Tree model named `dt_sh_clas_sample` to predict the probability that customer 101488 uses an affinity card. A customer representative can retrieve this information in real time when talking to this customer on the phone. Based on the query result, the representative can offer an extra-value card, since there is a 73% chance that the customer uses a card. The model is created by the `oml4sql-classification-decision-tree.sql` example.

Example 3-21 Real-Time Query with Prediction Probability

```
SELECT PREDICTION_PROBABILITY(dt_sh_clas_sample, 1 USING *) cust_card_prob
       FROM mining_data_apply_v
       WHERE cust_id = 101488;
```

The output is as follows:

```
CUST_CARD_PROB
--------------
        .72764
```

DBMS_DATA_MINING.APPLY

The `APPLY` procedure in `DBMS_DATA_MINING` is a batch apply operation that writes the results of scoring directly to a table.

The columns in the table are machine learning function-dependent.

Scoring with `APPLY` generates the same results as scoring with the SQL scoring functions. Classification produces a prediction and a probability for each case; clustering produces a cluster ID and a probability for each case, and so on. The difference lies in the way that scoring results are captured and the mechanisms that can be used for retrieving them.

`APPLY` creates an output table with the columns shown in the following table:

Table 3-14 APPLY Output Table

Machine Learning Technique	Output Columns

classification	`CASE_ID`
	`PREDICTION`
	`PROBABILITY`

regression	`CASE_ID` `PREDICTION`

anomaly detection	`CASE_ID`
	`PREDICTION`
	`PROBABILITY`

clustering	`CASE_ID`
	`CLUSTER_ID`
	`PROBABILITY`

Table 3-14 (Cont.) APPLY Output Table

Machine Learning Technique	Output Columns
feature extraction	CASE_ID FEATURE_ID MATCH_QUALITY

Since APPLY output is stored separately from the scoring data, it must be joined to the scoring data to support queries that include the scored rows. Thus any model that is used with APPLY must have a case ID.

A case ID is not required for models that is applied with SQL scoring functions. Likewise, storage and joins are not required, since scoring results are generated and consumed in real time within a SQL query.

The following example illustrates anomaly detection with APPLY. The query of the APPLY output table returns the ten first customers in the table. Each has a a probability for being typical (1) and a probability for being anomalous (0). The SVMO_SH_Clas_sample model is created by the oml4sql-anomaly-detection-1class-svm.sql example.

Example 3-22 Anomaly Detection with DBMS_DATA_MINING.APPLY

```
EXEC dbms_data_mining.apply
        ('SVMO_SH_Clas_sample','svmo_sh_sample_prepared',
         'cust_id', 'one_class_output');

SELECT * from one_class_output where rownum < 11;
```

The output is as follows:

```
   CUST_ID PREDICTION PROBABILITY
---------- ---------- -----------
    101798          1  .567389309
    101798          0  .432610691
    102276          1  .564922469
    102276          0  .435077531
    102404          1   .51213544
    102404          0   .48786456
    101891          1  .563474346
    101891          0  .436525654
    102815          0  .500663683
    102815          1  .499336317
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Create a Model that Includes Machine Learning Operations on Text

Create a model and specify the settings to perform machine learning operations on text.

Oracle Machine Learning for SQL supports unstructured text within columns of VARCHAR2, CHAR, CLOB, BLOB, and BFILE, as described in the following table:

Table 3-15 Column Data Types That May Contain Unstructured Text

Data Type	Description
BFILE and BLOB	Oracle Machine Learning for SQL interprets BLOB and BFILE as text *only if* you identify the columns as text when you create the model. If you do not identify the columns as text, then CREATE_MODEL returns an error.
CLOB	OML4SQL interprets CLOB as text.
CHAR	OML4SQL interprets CHAR as categorical by default. You can identify columns of CHAR as text when you create the model.
VARCHAR2	OML4SQL interprets VARCHAR2 with data length > 4000 as text. OML4SQL interprets VARCHAR2 with data length <= 4000 as categorical by default. You can identify these columns as text when you create the model.

Note:

Text is not supported in nested columns or as a target in supervised machine learning.

The settings described in the following table control the term extraction process for text attributes in a model. Instructions for specifying model settings are in "Specifying Model Settings".

Table 3-16 Model Settings for Text

Setting Name	Data Type	Setting Value	Description
ODMS_TEXT_POLICY_NAME	VARCHAR2(4000)	Name of an Oracle Text policy object created with CTX_DDL.CREATE_POLICY	Affects how individual tokens are extracted from unstructured text.
ODMS_TEXT_MAX_FEATURES	INTEGER	1 <= *value* <= 100000	Maximum number of features to use from the document set (across all documents of each text column) passed to CREATE_MODEL. Default is 3000.

A model can include one or more text attributes. A model with text attributes can also include categorical and numerical attributes.

To create a model that includes text attributes:

1. Create an Oracle Text policy object.
2. Specify the model configuration settings that are described in "Table 3-16".
3. Specify which columns must be treated as text and, optionally, provide text transformation instructions for individual attributes.
4. Pass the model settings and text transformation instructions to DBMS_DATA_MINING.CREATE_MODEL2 or DBMS_DATA_MINING.CREATE_MODEL.

> **Note:**
>
> All algorithms except O-Cluster can support columns of unstructured text.
>
> The use of unstructured text is not recommended for association rules (Apriori).

In the following example, an SVM model is used to predict customers that are most likely to be positive responders to an Affinity Card loyalty program. The data comes with a text column that contains user generated comments. By creating an Oracle Text policy and specifying model settings, the algorithm automatically uses the text column and builds the model on both the structured data and unstructured text.

This example uses a view called `mining_data` which is created from `SH.SALES` table. A training data set called `mining_train_text` is also created.

The following queries show you how to create an Oracle Text policy followed by building a model using `CREATE_MODEL2` procedure.

```
%script

BEGIN

EXECUTE ctx_ddl.create_policy('dmdemo_svm_policy');
```

The output is:

```


PL/SQL procedure successfully completed.

---------------------------

PL/SQL procedure successfully completed.
```

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('T_SVM_Clas_sample');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
    xformlist dbms_data_mining_transform.TRANSFORM_LIST;

BEGIN

    v_setlst(dbms_data_mining.algo_name) :=
dbms_data_mining.algo_support_vector_machines;
    v_setlst(dbms_data_mining.prep_auto) :=  dbms_data_mining.prep_auto_on;
    v_setlst(dbms_data_mining.svms_kernel_function) := dbms_data_mining.svms_linear;
    v_setlst(dbms_data_mining.svms_complexity_factor) := '100';
    v_setlst(dbms_data_mining.odms_text_policy_name) := 'DMDEMO_SVM_POLICY';

    v_setlst(dbms_data_mining.svms_solver) :=  dbms_data_mining.svms_solver_sgd;
    dbms_data_mining_transform.SET_TRANSFORM(
        xformlist, 'comments', null, 'comments', null, 'TEXT');
    DBMS_DATA_MINING.CREATE_MODEL2(
        model_name          => 'T_SVM_Clas_sample',
        mining_function     => dbms_data_mining.classification,
        data_query          => 'select * from mining_train_text',
        set_list            => v_setlst,
        case_id_column_name => 'cust_id',
        target_column_name  => 'affinity_card',
        xform_list => xformlist);
END;
/
```

The output is:

```


PL/SQL procedure successfully completed.

------------------------------

PL/SQL procedure successfully completed.

------------------------------
```

Related Topics

- Specify Model Settings
  You can configure your model by specifying model settings.
- Create a Text Policy
  An Oracle Text policy specifies how text content must be interpreted. You can provide a text policy to govern a model, an attribute, or both the model and individual attributes.
- Configure a Text Attribute
  Provide transformation instructions for text attribute or unstructured text by explicitly identifying the column datatypes.
- Embed Transformations in a Model
  You can specify your own transformations and embed them in a model by creating a transformation list and passing it to DBMS_DATA_MINING.CREATE_MODEL2 or DBMS_DATA_MINING.CREATE_MODEL.

Create a Text Policy

An Oracle Text policy specifies how text content must be interpreted. You can provide a text policy to govern a model, an attribute, or both the model and individual attributes.

If a model-specific policy is present and one or more attributes have their own policies, Oracle Machine Learning for SQL uses the attribute policies for the specified attributes and the model-specific policy for the other attributes.

The CTX_DDL.CREATE_POLICY procedure creates a text policy.

```
CTX_DDL.CREATE_POLICY(
          policy_name    IN VARCHAR2,
                          filter         IN VARCHAR2 DEFAULT NULL,
                          section_group  IN VARCHAR2 DEFAULT NULL,
                          lexer          IN VARCHAR2 DEFAULT NULL,
                          stoplist       IN VARCHAR2 DEFAULT NULL,
                          wordlist       IN VARCHAR2 DEFAULT NULL);
```

The parameters of CTX_DDL.CREATE_POLICY are described in the following table.

Table 3-17 CTX_DDL.CREATE_POLICY Procedure Parameters

Parameter Name	Description
policy_name	Name of the new policy object. Oracle Text policies and text indexes share the same namespace.
filter	Specifies how the documents must be converted to plain text for indexing. Examples are: CHARSET_FILTER for character sets and NULL_FILTER for plain text, HTML and XML. For filter values, see "Filter Types" in *Oracle Text Reference*.
section_group	Identifies sections within the documents. For example, HTML_SECTION_GROUP defines sections in HTML documents. For section_group values, see "Section Group Types" in *Oracle Text Reference*. Note: You can specify any section group that is supported by CONTEXT indexes.
lexer	Identifies the language that is being indexed. For example, BASIC_LEXER is the lexer for extracting terms from text in languages that use white space delimited words (such as English and most western European languages). For lexer values, see "Lexer Types" in *Oracle Text Reference*.
stoplist	Specifies words and themes to exclude from term extraction. For example, the word "the" is typically in the stoplist for English language documents. The system-supplied stoplist is used by default. See "Stoplists" in *Oracle Text Reference*.
wordlist	Specifies how stems and fuzzy queries must be expanded. A stem defines a root form of a word so that different grammatical forms have a single representation. A fuzzy query includes common misspellings in the representation of a word. See "BASIC_WORDLIST" in *Oracle Text Reference*.

Related Topics

- *Oracle Text Reference*

Configure a Text Attribute

Provide transformation instructions for text attribute or unstructured text by explicitly identifying the column datatypes.

As shown in Table 3-15, you can identify columns of CHAR, shorter VARCHAR2 (<=4000), BFILE, and BLOB as text attributes. If CHAR and shorter VARCHAR2 columns are not explicitly identified as unstructured text, then CREATE_MODEL processes them as categorical attributes. If BFILE and BLOB columns are not explicitly identified as unstructured text, then CREATE_MODEL returns an error.

To identify a column as a text attribute, supply the keyword TEXT in an **Attribute specification**. The attribute specification is a field (attribute_spec) in a transformation record (transform_rec). Transformation records are components of transformation lists (xform_list) that can be passed to CREATE_MODELor CREATE_MODEL2.

Note:

An attribute specification can also include information that is not related to text. Instructions for constructing an attribute specification are in "Embedding Transformations in a Model".

You can provide transformation instructions for any text attribute by qualifying the `TEXT` keyword in the attribute specification with the subsettings described in the following table.

Table 3-18 Attribute-Specific Text Transformation Instructions

Subsetting Name	Description	Example
`BIGRAM`	A sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words. Here, `NORMAL` tokens are mixed with their bigrams.	(`TOKEN_TYPE:`*`BIGRAM`*)
`POLICY_NAME`	Name of an Oracle Text policy object created with `CTX_DDL.CREATE_POLICY`	(`POLICY_NAME:`*`my_policy`*)
`STEM_BIGRAM`	Here, `STEM` tokens are extracted first and then stem bigrams are formed.	(`TOKEN_TYPE:`*`STEM_BIGRAM`*)
`SYNONYM`	Oracle Machine Learning for SQL supports synonyms. The following is an optional parameter: <`thesaurus`> where <`thesaurus`> is the name of the thesaurus defining synonyms. If `SYNONYM` is used without this parameter, then the default thesaurus is used.	(`TOKEN_TYPE:`*`SYNONYM`*) (`TOKEN_TYPE:`*`SYNONYM[NAMES]`*)

`TOKEN_TYPE`	The following values are supported:
	`NORMAL` (the default)
	`STEM`
	`THEME`
	See "Token Types in an Attribute Specification"
		`(TOKEN_TYPE:THEME)`

`MAX_FEATURES`	Maximum number of features to use from the attribute.	`(MAX_FEATURES:3000)`

Note:

The `TEXT` keyword is only required for `CLOB` and longer `VARCHAR2` (>4000) when you specify transformation instructions. The `TEXT` keyword is *always* required for `CHAR`, shorter `VARCHAR2`, `BFILE`, and `BLOB` — whether or not you specify transformation instructions.

Tip:

You can view attribute specifications in the data dictionary view `ALL_MINING_MODEL_ATTRIBUTES`, as shown in *Oracle Database Reference*.

Token Types in an Attribute Specification

When stems or themes are specified as the token type, the lexer preference for the text policy must support these types of tokens.

The following example adds themes and English stems to `BASIC_LEXER`.

```
BEGIN
  CTX_DDL.CREATE_PREFERENCE('my_lexer', 'BASIC_LEXER');
  CTX_DDL.SET_ATTRIBUTE('my_lexer', 'index_stems', 'ENGLISH');
  CTX_DDL.SET_ATTRIBUTE('my_lexer', 'index_themes', 'YES');
END;
```

Example 3-23 A Sample Attribute Specification for Text

This expression specifies that text transformation for the attribute must use the text policy named `my_policy`. The token type is `THEME`, and the maximum number of features is 3000.

```
"TEXT(POLICY_NAME:my_policy)(TOKEN_TYPE:THEME)(MAX_FEATURES:3000)"
```

Related Topics

- Embed Transformations in a Model
  You can specify your own transformations and embed them in a model by creating a transformation list and passing it to `DBMS_DATA_MINING.CREATE_MODEL2` or `DBMS_DATA_MINING.CREATE_MODEL`.
- Specify Transformation Instructions for an Attribute
  You can pass transformation instructions for an attribute by defining a transformation list.
- *Oracle Database PL/SQL Packages and Types Reference*
- ALL_MINING_MODEL_ATTRIBUTES

4
Use cases

Regression Use Case Scenario

A real estate agent approaches you, a data scientist, to provide assistance in evaluating house prices in Boston. The agent requires this information on a daily basis to provide targeted services to clients. Using the Generalized Linear Model algorithm for Regression, you estimate the median value of owner-occupied homes in the Boston area.

Related Content

Topic	Link
OML4SQL GitHub Example	Regression - GLM
`CREATE_MODEL2` Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING - Model Settings
Generalized Linear Model Settings	DBMS_DATA_MINING - Algorithm Settings: Generalized Linear Models
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views
Generalized Linear Model - Model Detail Views	Model Detail Views for Generalized Linear Model
About Regression	About Regression
About Generalized Linear Model (GLM)	About Generalized Linear Models

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data set
  Download the data set from https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/datasets/data/boston_house_prices.csv.

  > **Note:**
  >
  > This data set is used for illustrative purpose only.

- Database
  Select or create database out of the following options:

  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements
  Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Load Data

Examine the data set and its attributes. Load the data in your database.

In this use case, you will modify the data set to add a column and upload the data set to your database. If you are using the Oracle Autonomous Database, you will upload files to the Oracle Cloud Infrastructure (OCI) Object Storage, create a sample table, load data into the sample table from files on the OCI Object Storage, and explore the data. If you are using the on-premises database, you will use Oracle SQL developer to import the data set and explore the data.

Examine Data

There are 13 attributes in the data set. This is a customized data set that excludes one attribute from the original data set. The following table displays information about the data attributes:

Attribute Name	Information
CRIM	Per capita crime rate by town
ZN	The proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	The proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	The average number of rooms per dwelling
AGE	The proportion of owner-occupied units built before 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	The pupil-teacher ratio by town
LSTAT	% lower status of the population
MEDV	The median value of owner-occupied homes in $1000’s

Related Topics

- How ADP Transforms the Data

Add a Column

In this data set, no row identifier uniquely identifies each record in the data set. Add a new `case_id` column. The `case_id` assists with reproducible results, joining scores for individual customers with other data in, example, scoring data table.

Add a column called House ID (HID). The HID value is added as a primary key to the table so that identifying and retrieving each record is simple. Each record in the database is called a case and each case is identified by a `case_id`. Here, *HID* is the `case_id`.

To add the HID column:

1. Open the .csv file in a spreadsheet.
2. Delete the first row with 506 and 13. Now, the row with the column names becomes the first row.
3. To the left of the data set, add a column.
4. Enter *HID* as the column name.
5. In the *HID* column enter 1 as the first value identifying the first row.
6. You will see a + icon in the spreadsheet cell. Drag the + icon right to the bottom till the end of the records.
7. Right-click and select **Fill Series**.
8. To remove the column "B" from the data set, select the entire column with the title **B** by right clicking on the top of the column, and then select **Delete**.

Import Data

There are various methods to import data into the database. Two methods are explained here. One using SQL Developer (for on-premises) and the other using Object Storage (for Cloud).

Import Data into the Database (On premises)

To access the data set, import the modified data set into the database using SQL Developer.

The following steps help you to import the data set into an on premises database.

(Optional) Enter task prerequisites here.

1. Launch SQL Developer on your system.
2. Import the modified .csv file. See Tables.
3. Set House ID (HID) as a primary key. This column identifies each record and helps in retrieving information about a specific record. The HID column helps when you join tables or views. See Primary Key Constraint.

You are now ready to query the table in SQL Developer.

Import Data to the Cloud

If you are using a cloud account, one of the methods of importing the data is through Object Storage. Upload the data set to an Object Storage. The Object Storage URI will be used in another procedure.

You can load data into your Oracle Autonomous Database (Autonomous Data Warehouse [ADW] or Autonomous Transaction Processing [ATP]) using Oracle Database tools, and Oracle and 3rd party data integration tools. You can load data:

- from local files in your client computer, or
- from files stored in a cloud-based object store

Follow the steps to upload your data file to the Object Storage bucket.

1. Login to your cloud account.
2. Click the left-side hamburger menu and select **Storage** from the menu.
3. Select **Buckets** from the Object Storage & Archive Storage option.
4. Select the compartment in which you want to upload the data.
5. Click **Create Bucket**.
6. Enter a name for your bucket. For example, Bucket1. Leave the rest of the fields as default.
7. Click **Create**.
8. Click on the bucket that you created. Scroll down and click **Upload** under Objects.
9. Leave the Object Name Prefix field black. Click **select files** to navigate to the data file that you want to upload or drag and drop the data file. In this use case, select the modified .csv file.
10. Click **Upload**. The data file appears under Objects.
11. Click the ellipses on the right side of the data file to view the menu. Click **View Object Details**.
12. Copy the URL PATH (URI) to a text file. This URI is used in the `DBMS_CLOUD.COPY_DATA` procedure.

This procedure creates an object storage containing the data file in your cloud account.

Create Auth Token

The Auth Token is required in the `DBMS_CLOUD.CREATE_CREDENTIAL` procedure. You can generate the Auth Token in your cloud account.

1. Login into your ADW Cloud account.
2. Hover your mouse cursor over the human figure icon at the top right of the console and click **User Settings** from the drop-down menu.
3. Click **Auth Tokens** under Resources on the left of the console.
4. Click **Generate Token**. A pop-up dialog appears.
5. Enter a description (optional).
6. Click **Generate Token**.
7. Copy the generated token to a text file. The token does not appear again.

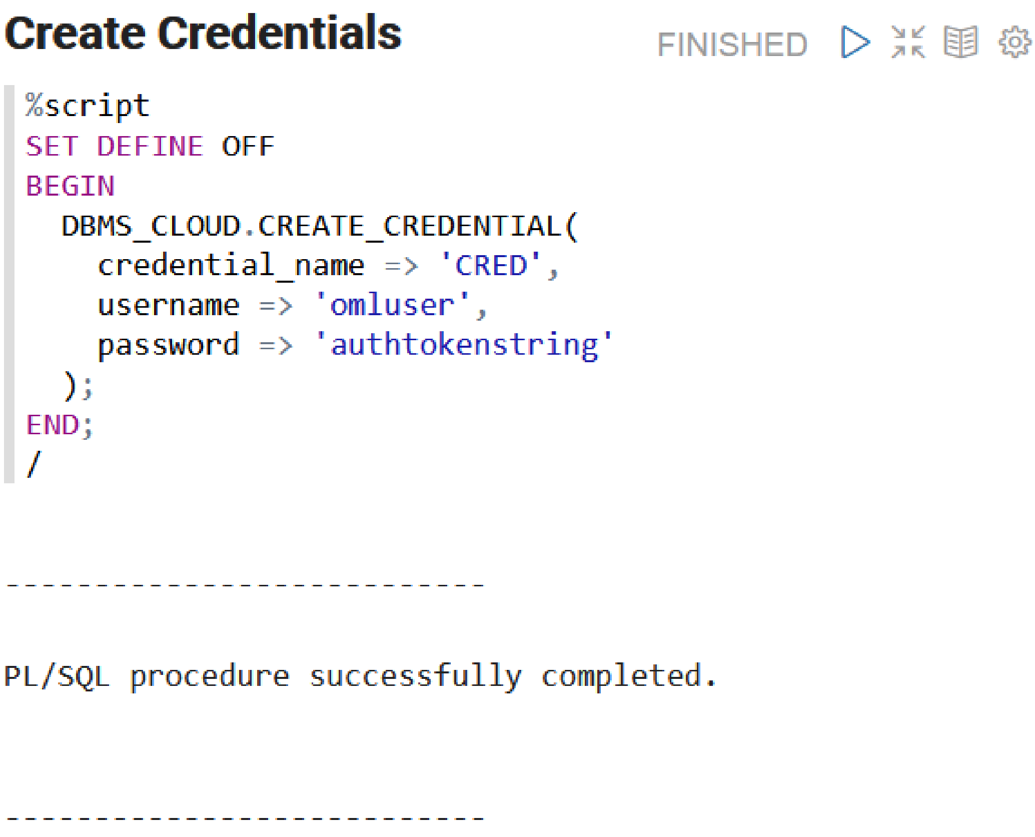
8. Click **Close**.

Create Object Storage Credential

The object storage credential is used in the `DBMS_CLOUD.COPY_DATA` procedure.

1. Login to the OML Notebooks page and create a notebook. See Create a Notebook Classic.
2. Open the notebook that you just created.
3. Enter the following query to create an object storage credentials:

```
%script
begin
  DBMS_CLOUD.create_credential (
    credential_name => 'CRED',
    username => '<your cloud account username>',
    password => '<your Auth Token>'
  );
end;
/



---------------------------  PL/SQL procedure successfully completed.
    ---------------------------
```



```
%script
SET DEFINE OFF
BEGIN
  DBMS_CLOUD.CREATE_CREDENTIAL(
    credential_name => 'CRED',
    username => 'omluser',
    password => 'authtokenstring'
  );
END;
/



---------------------------


PL/SQL procedure successfully completed.


---------------------------
```

Examine the query:

- `credential_name`: The name of the credential to be stored. Provide any name. Here, *CRED* is the name given.
- `username`: This is your cloud account username.
- `password`: Enter your Auth Token password that you copied after generating the Auth Token.

4. Click the play icon to run the query in your notebook. Your credentials are stored in the ADW user schema.
5. In another para, run the following query to check the user credentials:

   ```
   SELECT* FROM USER_CREDENTIALS;
   ```

Create a Table

Create a table called `BOSTON_HOUSING`. This table is used in `DBMS_CLOUD.COPY_DATA` procedure to access the data set.

Enter the following code in a new pare of the notebook that you created and run the notebook.

```
%sql
CREATE table boston_housing
(
 HID NUMBER NOT NULL,
 CRIM NUMBER,
 ZN NUMBER,
 INDUS NUMBER,
 CHAS VARCHAR2(32),
 NOX NUMBER,
 RM NUMBER,
 AGE NUMBER,
 DIS NUMBER,
 RAD NUMBER,
 TAX NUMBER,
 PTRATIO NUMBER,
 LSTAT NUMBER,
 MEDV NUMBER
);
```

Load Data in the Table

Load the data set stored in object storage to the `BOSTON_HOUSING` table.

Add a new para in the OML Notebooks and enter the following statement:

```
%script
BEGIN
 DBMS_CLOUD.COPY_DATA(
    table_name =>'BOSTON_HOUSING',
    credential_name =>'CRED',
    file_uri_list =>'https://objectstorage.us-phoenix-1.oraclecloud.com/n/
namespace-string/b/bucketname/o/filename.csv',
    format => json_object('type' value 'CSV', 'skipheaders' value 1)
```

```
 );
END;
```

Examine the statement:

- `table_name`: is the target table's name.
- `credential_name`: is the name of the credential created earlier.
- `file_uri_list`: is a comma delimited list of the source files you want to load.
- `format`: defines the options you can specify to describe the format of the source file, including whether the file is of type text, ORC, Parquet, or Avro.

In this example, *namespace-string* is the Oracle Cloud Infrastructure object storage namespace and *bucketname* is the storage bucket name that you created earlier (for example, Bucket1), and *filename.csv* is the modified .csv file name that you uploaded to the storage bucket.

Related Topics

- DBMS_CLOUD.COPY_DATA Procedure

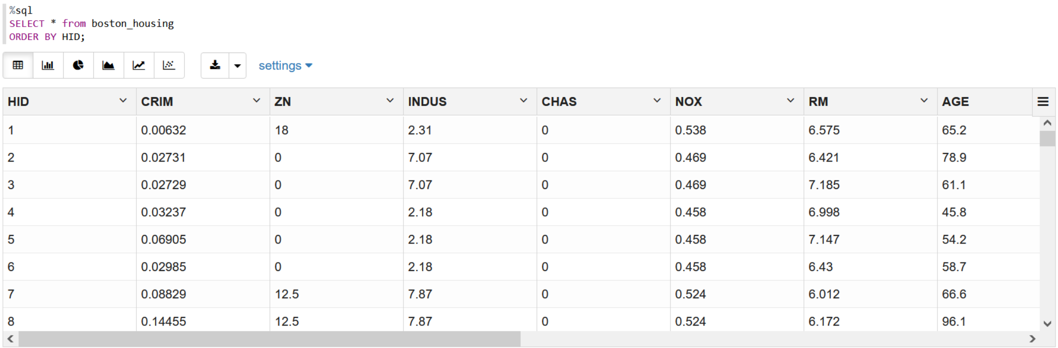
Explore Data

Explore the data to understand and assess the quality of the data. At this stage assess the data to identify data types and noise in the data. Look for missing values and numeric outlier values.

The following steps help you with the exploratory analysis of the data:

1. View the data in the `BOSTON_HOUSING` table by running the following query:

```
SELECT * FROM BOSTON_HOUSING
ORDER BY HID;
```



```
%sql
SELECT * from boston_housing
ORDER BY HID;
```

HID	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
1	0.00632	18	2.31	0	0.538	6.575	65.2
2	0.02731	0	7.07	0	0.469	6.421	78.9
3	0.02729	0	7.07	0	0.469	7.185	61.1
4	0.03237	0	2.18	0	0.458	6.998	45.8
5	0.06905	0	2.18	0	0.458	7.147	54.2
6	0.02985	0	2.18	0	0.458	6.43	58.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6
8	0.14455	12.5	7.87	0	0.524	6.172	96.1

2. Since you created the table specifying each column's datatype, you already know the datatype. However, to view the datatype of the columns, run the following script:

```
%script
DESCRIBE BOSTON_HOUSING;


Name    Null?    Type
------- -------- ------------
HID   NOT NULL NUMBER
```

```
CRIM           NUMBER
ZN             NUMBER
INDUS          NUMBER
CHAS           VARCHAR2(32)
NOX            NUMBER
RM             NUMBER
AGE            NUMBER
DIS            NUMBER
RAD            NUMBER(38)
TAX            NUMBER
PTRATIO        NUMBER
LSTAT          NUMBER
MEDV           NUMBER

---------------------------
```

3. Find the `COUNT` of the dataset to know how many rows are present.

    ```
    SELECT COUNT (*) from BOSTON_HOUSING;



    COUNT(*)
          506
    ---------------------------
    ```

4. To check if there are any missing values (NULL values), run the following query:

    ```
    SELECT COUNT(*) FROM BOSTON_HOUSING WHERE PTRATIO=NULL OR CHAS=NULL OR
     LSTAT=NULL OR TAX=NULL OR CRIM=NULL OR MEDV=NULL OR ZN=NULL OR NOX=NULL
     OR AGE=NULL OR INDUS=NULL OR DIS=NULL OR RAD=NULL OR PTRATIO=NULL OR RM=NULL;



    COUNT(*)
          0
    ---------------------------
    ```

    NULLs, if found, are automatically handled by the OML algorithms. Alternately, you can manually replace NULLs with `NVL` SQL function.

5. To list the distinct values for the categorical column CHAS and the number of records for each distinct value of CHAS, run the following query:

    ```
    %sql
    SELECT CHAS, COUNT(1)
    FROM BOSTON_HOUSING
    GROUP BY CHAS;



    CHAS    COUNT(1)
    0            471
    1             35
    ---------------------------
    ```

6. To calculate mean, median, min, max, and interquartile range (IQR) create a view called `unpivoted`.

The IQR describes the middle 50% of values (also called the mid spread or the H spread) when ordered from lowest to highest. To find the IQR, first, find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1. Sometimes, this assessment is helpful to find outliers in the data.

```
%sql
create or replace view unpivoted as
select *
  from (

SELECT 'CRIM' COL, ROUND(MIN(CRIM),2) MIN_VAL, PERCENTILE_CONT(0.25) WITHIN GROUP
(ORDER BY CRIM) FIRST_QUANTILE, ROUND(AVG(CRIM),2) MEAN_VAL, ROUND(MEDIAN(CRIM),2)
MEDIAN_VAL, PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY CRIM) THIRD_QUANTILE,
ROUND(MAX(CRIM),2) MAX_VAL
FROM BOSTON_HOUSING
UNION
SELECT 'AGE' COL, ROUND(MIN(AGE),2) MIN_VAL,  PERCENTILE_CONT(0.25) WITHIN GROUP
(ORDER BY AGE) FIRST_QUANTILE, ROUND(AVG(AGE),2) MEAN_VAL, ROUND(MEDIAN(AGE),2)
MEDIAN_VAL, PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY AGE) THIRD_QUANTILE,
ROUND(MAX(AGE),2) MAX_VAL
FROM BOSTON_HOUSING
UNION
SELECT 'DIS' COL, ROUND(MIN(DIS),2) MIN_VAL,  PERCENTILE_CONT(0.25) WITHIN GROUP
(ORDER BY DIS) FIRST_QUANTILE, ROUND(AVG(DIS),2) MEAN_VAL, ROUND(MEDIAN(DIS),2)
MEDIAN_VAL, PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY DIS) THIRD_QUANTILE,
ROUND(MAX(DIS),2) MAX_VAL
FROM BOSTON_HOUSING
  ) a
unpivot
(
  VALUE
    for stat in ("MIN_VAL", "FIRST_QUANTILE", "MEAN_VAL","MEDIAN_VAL",
"THIRD_QUANTILE", "MAX_VAL")
);
```

7. To view the values, pivot the table by running the following query:

```
%sql
select *
  from unpivoted
pivot(
  SUM(VALUE)
    for COL in ('CRIM', 'AGE','DIS')
);



STAT              'CRIM'      'AGE'     'DIS'
MEAN_VAL                3.61     68.57         3.8
THIRD_QUARTILE     3.6770825    94.075    5.188425
MAX_VAL                88.98       100       12.13
FIRST_QUARTILE      0.082045    45.025    2.100175
MEDIAN_VAL              0.26      77.5        3.21
MIN_VAL                 0.01       2.9        1.13


6 rows selected.

---------------------------
```

This completes the data understanding and data preparation stage. OML supports Automatic Data Preparation (ADP). ADP is enabled through the model settings. When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model. This step is done during the Build Model stage. The commonly used methods of data preparation are binning, normalization, and missing value treatment.

Related Topics

- How ADP Transforms the Data

Build Model

Build your model using the training data set. Use the DBMS_DATA_MINING.CREATE_MODEL2 procedure to build your model and specify model settings.

For a supervised learning, like Regression, before creating the model, split the data in to training and test data. Although you can use the entire data set to build a model, it is difficult to validate the model unless there are new data sets available. Therefore, to evaluate the model and to accurately assess the performance of the model on the same data, you generally split or separate the data into training and test data. You use the training data set to train the model and then use the test data set to test the accuracy of the model by running prediction queries. The testing data set already contains known values for the attribute that you want to predict. It is thus easy to determine whether the model's predictions are correct.

Algorithm Selection

Before you build a model, choose the suitable algorithm. You can choose one of the following algorithms to solve a regression problem:

- Extreme Gradient Boosting
- Generalized Linear Model
- Neural Network
- Support Vector Machine

When you want to understand the data set, you always start from a simple and easy baseline model. The Generalized Linear Model algorithm is the right choice because it is simple and easy to interpret since it fits a linear relationship between the feature and the target. You can get an initial understanding of a new data set from the result of the linear model.

The following steps guide you to split your data and build your model with the selected algorithm.

1. Split the data into 80/20 as training and test data. Run the following statement:

   ```
   BEGIN
       EXECUTE IMMEDIATE 'CREATE OR REPLACE VIEW TRAINING_DATA AS SELECT * FROM 
   BOSTON_HOUSING SAMPLE (80) SEED (1)';
       DBMS_OUTPUT.PUT_LINE ('Created TRAINING_DATA');
       EXECUTE IMMEDIATE 'CREATE OR REPLACE VIEW TEST_DATA AS SELECT * FROM 
   BOSTON_HOUSING MINUS SELECT * FROM TRAINING_DATA';
       DBMS_OUTPUT.PUT_LINE ('Created TEST_DATA');

   END;
   ```

   After splitting the data, view the count of rows in TRAINING_DATA and TEST_DATA. You can verify the ratio of the training and test data by checking the number of rows of the training and test set.

2. To find the count of rows in TRAINING_DATA, run the following statement:

```
select count(*) from TRAINING_DATA;

COUNT(*)
400
------------------------------
```

3. To find the count of rows from `TEST_DATA`, run the following statement:

   ```
   select COUNT(*) from TEST_DATA;

   COUNT(*)
   106
   ------------------------------
   ```

4. To find if any rows are not sampled (left out) in both `TRAINING_DATA` and `TEST_DATA`, run the following query:

   ```
   SELECT COUNT(1)
   FROM TRAINING_DATA train
   JOIN TEST_DATA test
   ON train.HID = test.HID

   COUNT(*)
   0
   ------------------------------
   ```

5. Build your model using the `CREATE_MODEL2` procedure. First, declare a variable to store model settings or hyperparameters. Run the following script:

   ```
   %script
   DECLARE
       v_setlst DBMS_DATA_MINING.SETTING_LIST;
       BEGIN
       v_setlst('PREP_AUTO') := 'ON';
       v_setlst('ALGO_NAME') := 'ALGO_GENERALIZED_LINEAR_MODEL';
       v_setlst('GLMS_DIAGNOSTICS_TABLE_NAME') := 'GLMR_DIAG';
       v_setlst('GLMS_FTR_SELECTION') := 'GLMS_FTR_SELECTION_ENABLE';
       v_setlst('GLMS_FTR_GENERATION') := 'GLMS_FTR_GENERATION_ENABLE';

       DBMS_DATA_MINING.CREATE_MODEL2(
         MODEL_NAME          =>  'GLMR_REGR',
         MINING_FUNCTION     => 'REGRESSION'
         DATA_QUERY          =>  'SELECT * FROM TRAINING_DATA',
         SET_LIST            =>  v_setlst,
         CASE_ID_COLUMN_NAME =>  'HID',
         TARGET_COLUMN_NAME  =>  'MEDV');
   END;
   ```

   Examine the script:

   - `v_setlst` is a variable to store `SETTING_LIST`.
   - `SETTING_LIST` defines model settings or hyperparameters for your model.
   - `DBMS_DATA_MINING` is the PL/SQL package used for Oracle Machine Learning. These settings are described in DBMS_DATA_MINING - Model Settings.
   - `ALGO_NAME` specifies the algorithm name. Since you are using the Generalized Linear Model as your algorithm, set `ALGO_GENERALIZED_LINEAR_MODEL`.

- PREP_AUTO is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is ON.
- GLMS_DIAGNOSTICS_TABLE_NAME generates per-row statistics if you specify the name of a diagnostics table in the setting. The value of the setting is GLMR_DIAG.
- GLMS_FTR_SELECTION indicates feature selection. The value GLMS_FTR_SELECTION_ENABLE indicates that feature selection is enabled. Feature selection selects columns that are most important in predicting a target attribute. If feature selection is not selected, then all the columns are considered for analysis which may not give accurate results.
- GLMS_FTR_GENERATION indicates feature generation. The value GLMS_FTR_GENERATION_ENABLE indicates that the feature generation is enabled. Feature generation generates new features from existing features which might be useful in our analysis.

The CREATE_MODEL2 procedure has the following parameters:

- MODEL_NAME: A unique model name that you want to give to your model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is GLMR_REGR
- MINING_FUNCTION: Specifies the machine learning function. Since you are solving a linear regression problem, in this use case, select REGRESSION.
- DATA_QUERY: A query that provides training data for building the model. Here, the query is SELECT * FROM TRAINING_DATA.
- SET_LIST: Specifies SETTING_LIST.
- CASE_ID_COLUMN_NAME: A unique case identifier column in the training data. In this use case, case_id is HID. If there is a composite key, you must create a new attribute before creating the model. The CASE_ID assists with reproducible results, joining scores for individual customers with other data in, example, scoring data table.
- TARGET_COLUMN_NAME: Specifies the column that needs to be predicted. Also referred to as the target variable of the model. In this use case, you are predicting MEDV value.

> **Note:**
>
> Any parameters or settings not specified are either system-determined or default values are used.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

Sometimes querying dictionary views and model detail views is sufficient to measure your model's performance. However, you can evaluate your model by computing test metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), confusion matrix, lift statistics, cost matrix, and so on. For Association Rules, you can inspect various rules to see if they reveal new insights for item dependencies (antecedent itemset implying consequent) or for unexpected relationships among items.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

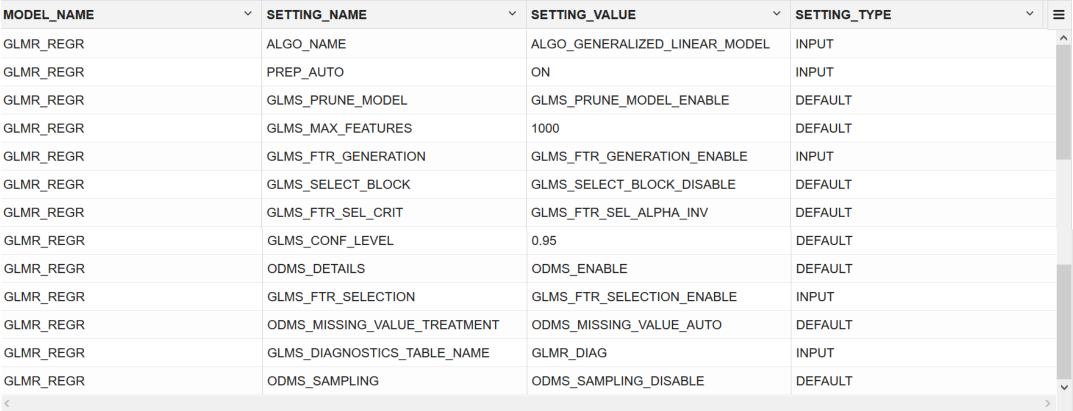
Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views.

The following steps help you to view different dictionary views and model detail views.

1. Run the following statement to view the settings in `USER_MINING_MODEL_SETTINGS`:

   ```
   SELECT * FROM USER_MINING_MODEL_SETTINGS WHERE MODEL_NAME='GLMR_REGR';
   ```

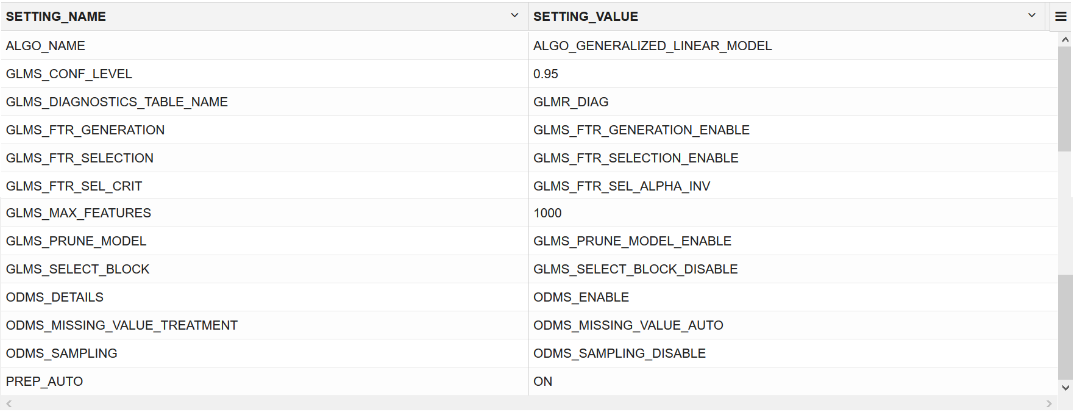
   In this statement, you are selecting all the columns available in the `USER_MINING_MODEL_SETTINGS` view where the model name is `GLMR_REGR`.



MODEL_NAME	SETTING_NAME	SETTING_VALUE	SETTING_TYPE
GLMR_REGR	ALGO_NAME	ALGO_GENERALIZED_LINEAR_MODEL	INPUT
GLMR_REGR	PREP_AUTO	ON	INPUT
GLMR_REGR	GLMS_PRUNE_MODEL	GLMS_PRUNE_MODEL_ENABLE	DEFAULT
GLMR_REGR	GLMS_MAX_FEATURES	1000	DEFAULT
GLMR_REGR	GLMS_FTR_GENERATION	GLMS_FTR_GENERATION_ENABLE	INPUT
GLMR_REGR	GLMS_SELECT_BLOCK	GLMS_SELECT_BLOCK_DISABLE	DEFAULT
GLMR_REGR	GLMS_FTR_SEL_CRIT	GLMS_FTR_SEL_ALPHA_INV	DEFAULT
GLMR_REGR	GLMS_CONF_LEVEL	0.95	DEFAULT
GLMR_REGR	ODMS_DETAILS	ODMS_ENABLE	DEFAULT
GLMR_REGR	GLMS_FTR_SELECTION	GLMS_FTR_SELECTION_ENABLE	INPUT
GLMR_REGR	ODMS_MISSING_VALUE_TREATMENT	ODMS_MISSING_VALUE_AUTO	DEFAULT
GLMR_REGR	GLMS_DIAGNOSTICS_TABLE_NAME	GLMR_DIAG	INPUT
GLMR_REGR	ODMS_SAMPLING	ODMS_SAMPLING_DISABLE	DEFAULT

2. Run the following statement to view only the `SETTING_NAME` and `SETTING_VALUE` column from the above table:

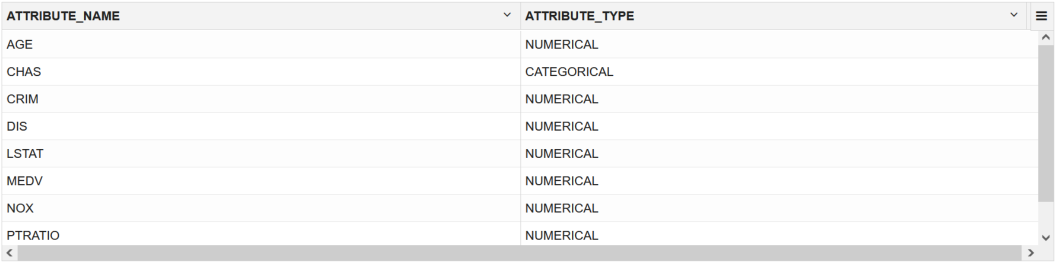
   ```
   SELECT SETTING_NAME, SETTING_VALUE FROM USER_MINING_MODEL_SETTINGS WHERE
   MODEL_NAME = 'GLMR_REGR' ORDER BY SETTING_NAME;
   ```



SETTING_NAME	SETTING_VALUE
ALGO_NAME	ALGO_GENERALIZED_LINEAR_MODEL
GLMS_CONF_LEVEL	0.95
GLMS_DIAGNOSTICS_TABLE_NAME	GLMR_DIAG
GLMS_FTR_GENERATION	GLMS_FTR_GENERATION_ENABLE
GLMS_FTR_SELECTION	GLMS_FTR_SELECTION_ENABLE
GLMS_FTR_SEL_CRIT	GLMS_FTR_SEL_ALPHA_INV
GLMS_MAX_FEATURES	1000
GLMS_PRUNE_MODEL	GLMS_PRUNE_MODEL_ENABLE
GLMS_SELECT_BLOCK	GLMS_SELECT_BLOCK_DISABLE
ODMS_DETAILS	ODMS_ENABLE
ODMS_MISSING_VALUE_TREATMENT	ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING	ODMS_SAMPLING_DISABLE
PREP_AUTO	ON

3. Run the following statement to see attribute information in `USER_MINING_MODEL_ATTRIBUTES` view:

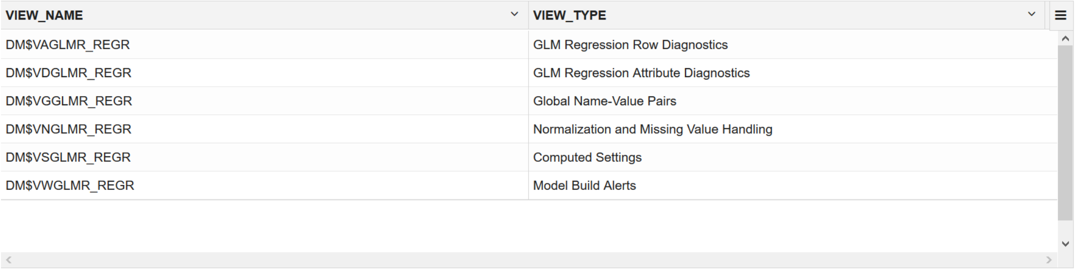
```
SELECT ATTRIBUTE_NAME, ATTRIBUTE_TYPE FROM USER_MINING_MODEL_ATTRIBUTES
WHERE MODEL_NAME = 'GLMR_REGR' ORDER BY ATTRIBUTE_NAME;
```



ATTRIBUTE_NAME	ATTRIBUTE_TYPE
AGE	NUMERICAL
CHAS	CATEGORICAL
CRIM	NUMERICAL
DIS	NUMERICAL
LSTAT	NUMERICAL
MEDV	NUMERICAL
NOX	NUMERICAL
PTRATIO	NUMERICAL

4. Run the following statement to see information on various views in `USER_MINING_MODEL_VIEWS`:

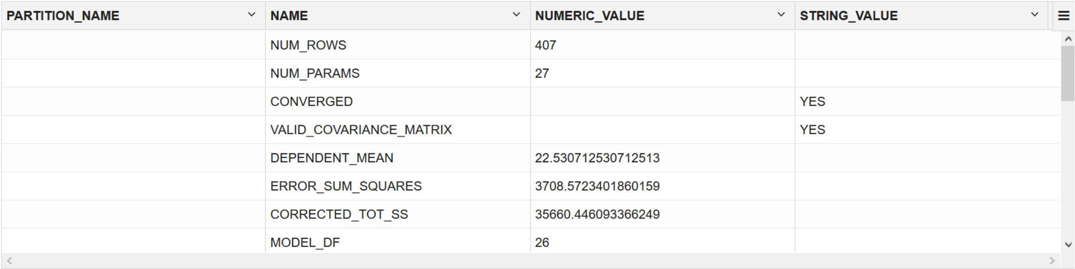
```
SELECT VIEW_NAME, VIEW_TYPE FROM USER_MINING_MODEL_VIEWS WHERE
MODEL_NAME='GLMR_REGR' ORDER BY VIEW_NAME;
```



VIEW_NAME	VIEW_TYPE
DM$VAGLMR_REGR	GLM Regression Row Diagnostics
DM$VDGLMR_REGR	GLM Regression Attribute Diagnostics
DM$VGGLMR_REGR	Global Name-Value Pairs
DM$VNGLMR_REGR	Normalization and Missing Value Handling
DM$VSGLMR_REGR	Computed Settings
DM$VWGLMR_REGR	Model Build Alerts

5. From the table above, query the Global details for linear regression. See Model Detail Views for Generalized Linear Model. Run the following query to see all the columns of the view:

```
SELECT * FROM DM$VGGLMR_REGR;
```

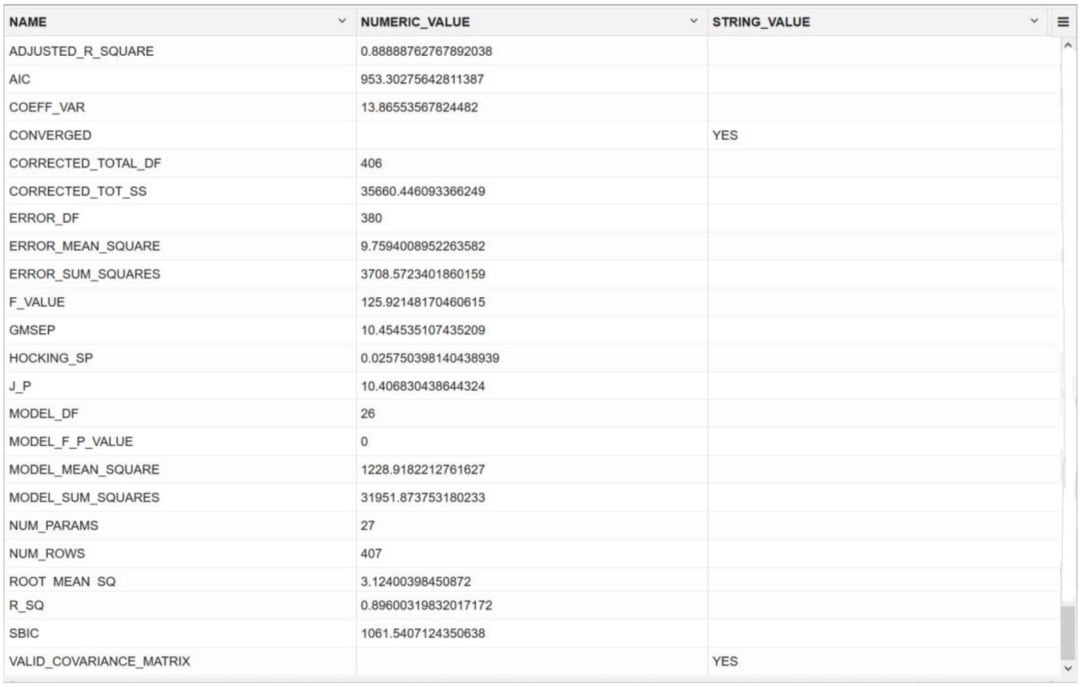


PARTITION_NAME	NAME	NUMERIC_VALUE	STRING_VALUE
	NUM_ROWS	407	
	NUM_PARAMS	27	
	CONVERGED		YES
	VALID_COVARIANCE_MATRIX		YES
	DEPENDENT_MEAN	22.530712530712513	
	ERROR_SUM_SQUARES	3708.5723401860159	
	CORRECTED_TOT_SS	35660.446093366249	
	MODEL_DF	26	

6. From the above table, you can ignore the first column `PARTITION_NAME` and refine the query to display the rest of the columns ordered by name. Run the following statement:

```
SELECT NAME, NUMERIC_VALUE, STRING_VALUE FROM DM$VGGLMR_REGR ORDER BY NAME;
```

When comparing models, a model with a lower Root Mean Square Error (RMSE) value is better. RMSE, which squares the errors, gives more weight to large errors. When we have a low RMSE value, we can say that our model is good at predicting the target.



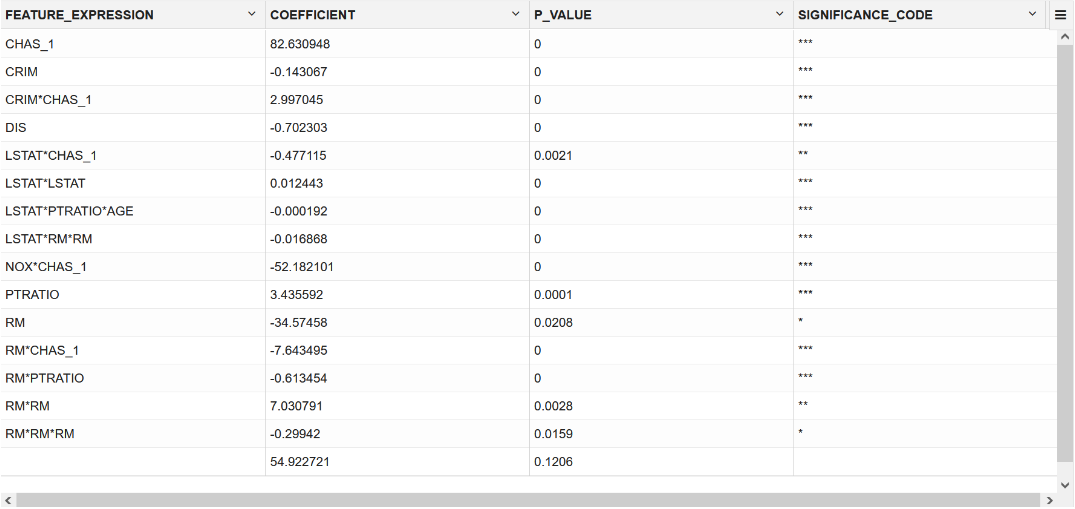
NAME	NUMERIC_VALUE	STRING_VALUE
ADJUSTED_R_SQUARE	0.88888762767892038	
AIC	953.30275642811387	
COEFF_VAR	13.86553567824482	
CONVERGED		YES
CORRECTED_TOTAL_DF	406	
CORRECTED_TOT_SS	35660.446093366249	
ERROR_DF	380	
ERROR_MEAN_SQUARE	9.7594008952263582	
ERROR_SUM_SQUARES	3708.5723401860159	
F_VALUE	125.92148170460615	
GMSEP	10.454535107435209	
HOCKING_SP	0.025750398140438939	
J_P	10.406830438644324	
MODEL_DF	26	
MODEL_F_P_VALUE	0	
MODEL_MEAN_SQUARE	1228.9182212761627	
MODEL_SUM_SQUARES	31951.873753180233	
NUM_PARAMS	27	
NUM_ROWS	407	
ROOT_MEAN_SQ	3.12400398450872	
R_SQ	0.89600319832017172	
SBIC	1061.5407124350638	
VALID_COVARIANCE_MATRIX		YES

7. Query the GLM Regression Attributes Diagnostics view.

```
SELECT FEATURE_EXPRESSION, round(COEFFICIENT,6) COEFFICIENT, round(P_VALUE,4)
P_VALUE,
CASE
    when p_value < 0.001 THEN '***'
    when p_value < 0.01 THEN '**'
    when p_value < 0.05 THEN '*'
    when p_value < 0.1 THEN '.'
    else ' '
END AS significance_statement
FROM DM$VDGLMR_REGR ORDER BY FEATURE_EXPRESSION;
```

The columns of the view are described in Model Detail Views for Generalized Linear Model.
Let us examine the statement:

- `round(COEFFICIENT,6) COEFFICIENT`: returns the coefficient rounded to six places to the right of the decimal point.
- `p_value`: provides information about the relationship between a dependent variable and independent variable such that you could decide to accept or reject the null hypothesis. Generally, p_value less than 0.05 means that you can reject the null hypothesis and accept that there is a correlation between the dependent and independent variables with a significant coefficient value.



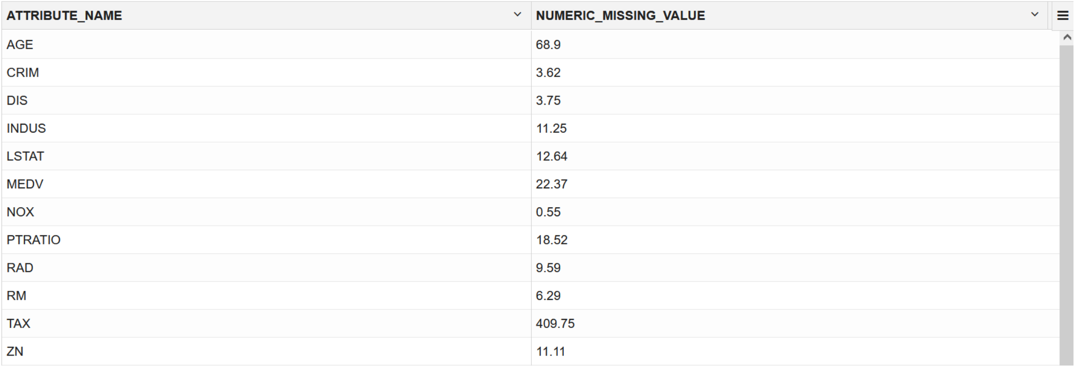
FEATURE_EXPRESSION	COEFFICIENT	P_VALUE	SIGNIFICANCE_CODE
CHAS_1	82.630948	0	***
CRIM	-0.143067	0	***
CRIM*CHAS_1	2.997045	0	***
DIS	-0.702303	0	***
LSTAT*CHAS_1	-0.477115	0.0021	**
LSTAT*LSTAT	0.012443	0	***
LSTAT*PTRATIO*AGE	-0.000192	0	***
LSTAT*RM*RM	-0.016868	0	***
NOX*CHAS_1	-52.182101	0	***
PTRATIO	3.435592	0.0001	***
RM	-34.57458	0.0208	*
RM*CHAS_1	-7.643495	0	***
RM*PTRATIO	-0.613454	0	***
RM*RM	7.030791	0.0028	**
RM*RM*RM	-0.29942	0.0159	*
	54.922721	0.1206	

8. Now, run the following statement to query Normalization and Missing Value Handling view. The columns of the view are described in Model Detail Views for Normalization and Missing Value Handling.

```
SELECT ATTRIBUTE_NAME, round(NUMERIC_MISSING_VALUE,2)
NUMERIC_MISSING_VALUE FROM DM$VNGLMR_REGR
ORDER BY ATTRIBUTE_NAME;
```

Examine the query:

- `ATTRIBUTE_NAME`: Provides the column names in the data set.
- `round(NUMERIC_MISSING_VALUE,2)NUMERIC_MISSING_VALUE`: Provides numeric replacements for the missing values (NULLs) in the data set. The `ROUND (n,integer)` returns results of `NUMERIC_MISSING_VALUE` rounded to integer places to the right.



ATTRIBUTE_NAME	NUMERIC_MISSING_VALUE
AGE	68.9
CRIM	3.62
DIS	3.75
INDUS	11.25
LSTAT	12.64
MEDV	22.37
NOX	0.55
PTRATIO	18.52
RAD	9.59
RM	6.29
TAX	409.75
ZN	11.11

Since there are no missing values (NULLs) in your data, you can ignore the result.

Test Your Model

In this use case, you are evaluating a regression model by computing Root Mean Square Error (RMSE) and Mean Absolute Error Mean (MAE) on the test data with known target values and comparing the predicted values with the known values.

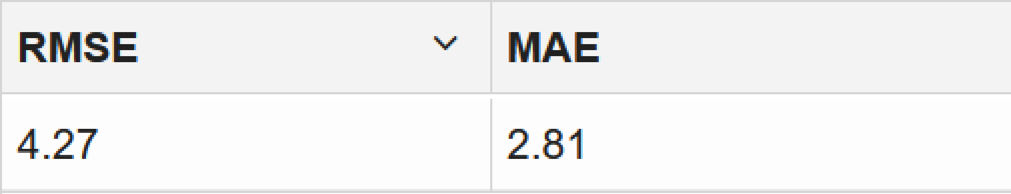
Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets your business requirements, it can then be applied to new data to predict the future. These matrices can help you to compare models to arrive at one model that satisfies your evaluation criteria.
For this use case, you compute Root Mean Square Error (RMSE) and Mean Absolute Error Mean (MAE) values. The RMSE and MAE are popular regression statistics. RMSE is an estimator for predictive models. The score averages the residuals for each case to yield a single indicator of model error. Mean absolute error is useful for understanding how close overall the predictions were to actual values. A smaller score means predictions were more accurate.

The following steps computes the error metrics for your model.

- To compute RMSE and MAE, run the following statement:

```
%sql
SELECT round(SQRT(AVG((A.PRED_MEDV - B.MEDV) * (A.PRED_MEDV - B.MEDV))),2) RMSE,
       round(AVG(ABS(A.PRED_MEDV - B.MEDV)),2) MAE
  FROM (SELECT HID, PREDICTION(GLMR_REGR using *) PRED_MEDV
          FROM TEST_DATA) A,
       TEST_DATA B
  WHERE A.HID = B.HID;
```

  This statement is using the prediction query to score the median value from the test data. The predicted value and the actual value from the test data is used to compute RMSE and MAE .



RMSE	MAE
4.27	2.81

RMSE and MAE convey average model prediction errors in units consistent with the target variable. When comparing models, a model with lower values is better. RMSE, which squares the errors, gives more weight to large errors, while MAE error scales linearly. Therefore, the predictions look fair and the model is a good fit for prediction.

Score

Scoring involves applying the model to the target data. Use `PREDICTION` query to predict the `MEDV` value on the test data.

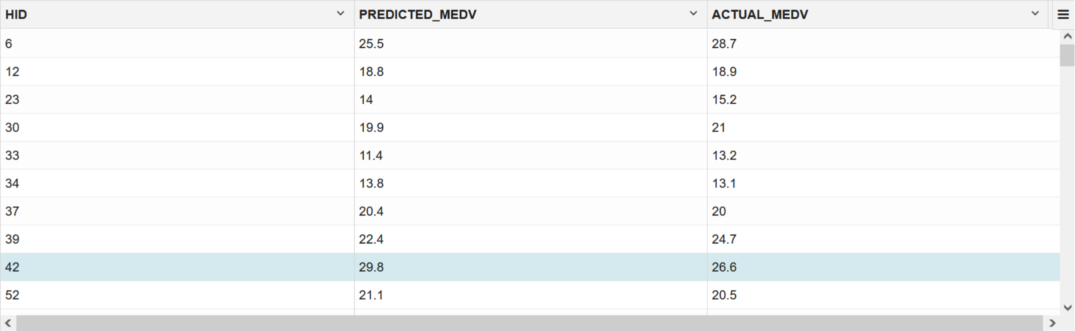
The following step scores the test data comparing with the original data.

- Predict the median value of owner-occupied homes in the Boston area from the `TEST_DATA` and compare the predicted `MEDV` value with the actual `MEDV` value in your result.

```
SELECT HID, ROUND(PREDICTION(GLMR_REGR USING *), 1) AS
PREDICTED_MEDV, MEDV AS ACTUAL_MEDV FROM TEST_DATA ORDER BY HID;
```

Examine the query:

- `HID:` is the House ID.
- `ROUND (n,integer):` in this case, is `ROUND (PREDICTION(GLMR_REGR USING *), 1)` returns results of `PREDICTION(GLMR_REGR USING *)` rounded to integer places to the right. Here, rounded to 1 place to the right.
- `PREDICTED_MEDV:` is the predicted `MEDV` value.
- `ACTUAL_MEDV:` is the `MEDV` value in the test data.



HID	PREDICTED_MEDV	ACTUAL_MEDV
6	25.5	28.7
12	18.8	18.9
23	14	15.2
30	19.9	21
33	11.4	13.2
34	13.8	13.1
37	20.4	20
39	22.4	24.7
42	29.8	26.6
52	21.1	20.5

To conclude, you have successfully predicted the median house prices in Boston using Generalized Linear Model algorithm.

Classification Use Case Scenario

You are working in a retail chain company that sells some products. To better target their marketing materials, they need to identify customers who are likely to purchase a home theater package. To resolve this, you are using the Random Forest algorithm to identify the customers.

Related Content

Topic	**Link**
OML4SQL GitHub Example	Classification - Random Forest
`CREATE_MODEL2` Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING - Model Settings
Random Forest Settings	DBMS_DATA_MINING - Algorithm Settings: Random Forest
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views
Random Forest - Model Detail Views	Model Detail Views for Random Forest
About Classification	About Classification
About Random Forest (RF)	About Random Forest

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data Set
  The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.
- Database
  Select or create database out of the following options:

  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements
  Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Related Topics

- Create a Notebook
- Edit your Notebook
- Uninstalling HR Schema

Load Data

Access the data set from the SH Schema and explore the data to understand the attributes.

> **Remember:**
>
> The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.

To understand the data, you will perform the following:

- Access the data.
- Examine the various attributes or columns of the data set.
- Assess data quality (by exploring the data).

Access Data

You will use `CUSTOMERS` and `SUPPLEMENTARY_DEMOGRAPHICS` table data from the SH schema.

Examine Data

The following table displays information about the attributes from `SUPPLEMENTARY_DEMOGRAPHICS`:

Attribute Name	Information
`CUST_ID`	The ID of the customer
`EDUCATION`	Educational information of the customer
`OCCUPATION`	Occupation of the customer

Attribute Name	Information
HOUSEHOLD_SIZE	People per house
YRS_RESIDENCE	Number of years of residence
AFFINITY_CARD	Whether the customer holds an affinity card
BULK_PACK_DISKETTES	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
FLAT_PANEL_MONITOR	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
HOME_THEATER_PACKAGE	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
BOOKKEEPING_APPLICATION	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
PRINTER_SUPPLIES	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
Y_BOX_GAMES	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
OS_DOC_SET_KANJI	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
COMMENTS	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No

Explore Data

Explore the data to understand and assess the quality of the data. At this stage assess the data to identify data types and noise in the data. Look for missing values and numeric outlier values.

Assess Data Quality

To assess the data, first, you must be able to view the data in your database. For this reason, you will use SQL statements to query the SH.CUSTOMERS and the SH.SUPPLEMENTARY_DEMOGRAPHICS table.

If you are working with Oracle Autonomous Database, you can use the Oracle Machine Learning (OML) Notebooks for your data science project, including assessing data quality. If you are using on-premise Oracle Database, you can use the Oracle SQL Developer to assess data quality. Query the SH schema as described.

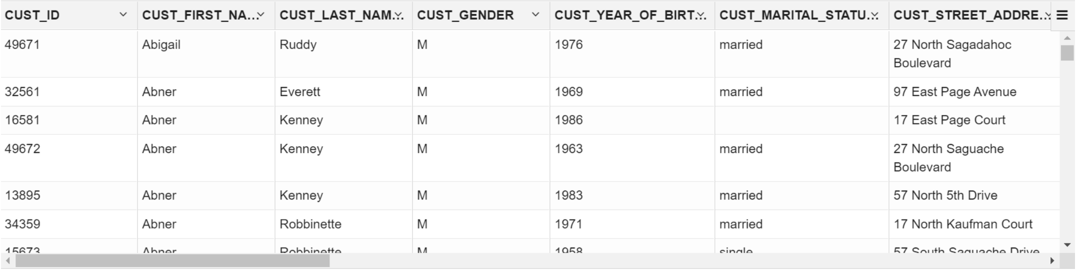
> **Note:**
>
> Each record in the database is called a case and each case is identified by a `case_id`. In this use case, `CUST_ID` is the `case_id`.

1. View the data in the `SH.CUSTOMERS` table by running the following statement:

```
SELECT * FROM SH.CUSTOMERS;
```

2. To see distinct data from the table, run the following statement:

```
SELECT DISTINCT * FROM SH.CUSTOMERS;
```



CUST_ID	CUST_FIRST_NA...	CUST_LAST_NAM...	CUST_GENDER	CUST_YEAR_OF_BIRT...	CUST_MARITAL_STATU...	CUST_STREET_ADDRE...
49671	Abigail	Ruddy	M	1976	married	27 North Sagadahoc Boulevard
32561	Abner	Everett	M	1969	married	97 East Page Avenue
16581	Abner	Kenney	M	1986		17 East Page Court
49672	Abner	Kenney	M	1963	married	27 North Saguache Boulevard
13895	Abner	Kenney	M	1983	married	57 North 5th Drive
34359	Abner	Robbinette	M	1971	married	17 North Kaufman Court
15673	Abner	Robbinette	M	1958	single	57 South Saguache Drive

3. Find the `COUNT` of rows in the data set by running the following statement:

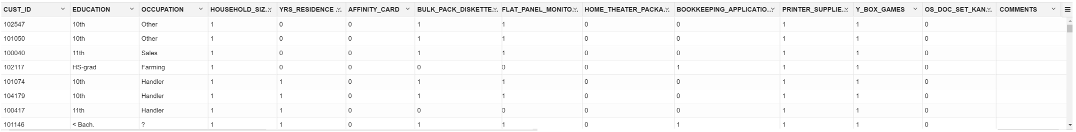
```
SELECT COUNT(*) from SH.CUSTOMERS;



COUNT(*)
     55500
---------------------------
```

4. To identify distinct or unique customers in the table, run the following statement:

```

%script
SELECT COUNT (DISTINCT CUST_ID) FROM SH.CUSTOMERS;



COUNT(DISTINCTCUST_ID)
                   55500
---------------------------
```

5. Similarly, query the `SH.SUPPLEMENTARY_DEMOGRAPHICS` table.

```
SELECT * FROM SH.SUPPLEMENTARY_DEMOGRAPHICS;
```



CUST_ID	EDUCATION	OCCUPATION	HOUSEHOLD_SIZ...	YRS_RESIDENCE ...	AFFINITY_CARD	BULK_PACK_DISKETTE...	FLAT_PANEL_MONITO...	HOME_THEATER_PACKA...	BOOKKEEPING_APPLICATIO...	PRINTER_SUPPLIE...	Y_BOX_GAMES	OS_DOC_SET_KAN...	COMMENTS
102547	10th	Other	1	0	0	1	1	0	0	1	1	0	
101050	10th	Other	1	0	0	1	1	0	0	1	1	0	
100040	11th	Sales	1	0	0	1	1	0	0	1	1	0	
102117	HS-grad	Farming	1	0	0	0	0	0	1	1	1	0	
101074	10th	Handler	1	1	0	1	1	0	0	1	1	0	
104179	10th	Handler	1	1	0	1	1	0	0	1	1	0	
100417	11th	Handler	1	1	0	0	0	0	0	1	1	0	
101146	< Bach.	?	1	1	0	1	1	0	1	1	1	0	

6. To view the count of `SH.SUPPLEMENTARY_DEMOGRAPHICS`, run the following statement:

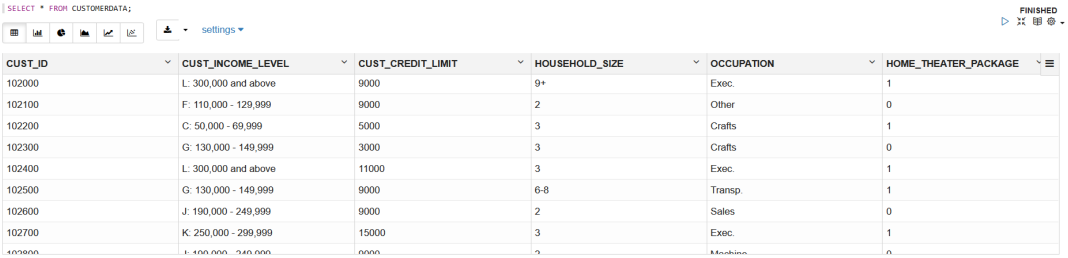
```
SELECT COUNT(*) from SH.SUPPLEMENTARY_DEMOGRAPHICS;



COUNT(*)
      4500
---------------------------
```

7. Create a table called `CUSTOMERDATA` by selecting the required columns from the `SH.CUSTOMERS` and the `SH.SUPPLIMENTARY_DEMOGRAPHICS` tables.

```
%script
CREATE TABLE CUSTOMERDATA AS
   SELECT a.CUST_ID,
         a.CUST_INCOME_LEVEL, a.CUST_CREDIT_LIMIT,
          b.HOUSEHOLD_SIZE, b.OCCUPATION, b.HOME_THEATER_PACKAGE
   FROM SH.CUSTOMERS a, SH.SUPPLEMENTARY_DEMOGRAPHICS b
   WHERE a.CUST_ID = b.CUST_ID;



Table CUSTOMERDATA created.
```

8. View the `CUSTOMERDATA` table.

```
SELECT * FROM CUSTOMERDATA;
```



```
SELECT * FROM CUSTOMERDATA;
```

CUST_ID	CUST_INCOME_LEVEL	CUST_CREDIT_LIMIT	HOUSEHOLD_SIZE	OCCUPATION	HOME_THEATER_PACKAGE
102000	L: 300,000 and above	9000	9+	Exec.	1
102100	F: 110,000 - 129,999	9000	2	Other	0
102200	C: 50,000 - 69,999	5000	3	Crafts	1
102300	G: 130,000 - 149,999	3000	3	Crafts	0
102400	L: 300,000 and above	11000	3	Exec.	1
102500	G: 130,000 - 149,999	9000	6-8	Transp.	1
102600	J: 190,000 - 249,999	9000	2	Sales	0
102700	K: 250,000 - 299,999	15000	3	Exec.	1
102800	J: 190,000 - 249,999	9000	2	Machine	0

9. Find the count of rows in the new table `CUSTOMERDATA`:

```
SELECT COUNT(*) FROM CUSTOMERDATA;



COUNT(*)
      4500
---------------------------
```

10. To view the data type of the columns, run the following script:

```
%script
DESCRIBE CUSTOMERDATA;



Name                 Null?    Type
-------------------- -------- ------------
```

```
CUST_ID        NOT NULL NUMBER
CUST_INCOME_LEVEL          VARCHAR2(30)
CUST_CREDIT_LIMIT          NUMBER
HOUSEHOLD_SIZE          VARCHAR2(21)
OCCUPATION          VARCHAR2(21)
HOME_THEATER_PACKAGE          NUMBER(10)

---------------------------
```

11. To check if there are any missing values (NULL values), run the following statement:

```
SELECT COUNT(*) FROM CUSTOMERDATA WHERE CUST_ID=NULL OR CUST_GENDER=NULL
 OR CUST_MARITAL_STATUS=NULL OR CUST_YEAR_OF_BIRTH=NULL OR CUST_INCOME_LEVEL=NULL
 OR CUST_CREDIT_LIMIT=NULL OR HOUSEHOLD_SIZE=NULL OR YRS_RESIDENCE=NULL OR
Y_BOX_GAMES=NULL;



COUNT(*)
         0
---------------------------
```

NULLs, if found, are automatically handled by the OML algorithms. Alternately, you can manually replace NULLs with `NVL` SQL function.

12. To know the income level of customers who responded to `HOME_THEATER_PACKAGE`, run the following statement:

```
SELECT COUNT(CUST_ID) AS NUM_CUSTOMERS, CUST_INCOME_LEVEL, HOME_THEATER_PACKAGE
FROM   CUSTOMERDATA
GROUP BY CUST_INCOME_LEVEL, HOME_THEATER_PACKAGE;


NUM_CUSTOMERS   CUST_INCOME_LEVEL      HOME_THEATER_PACKAGE
            214 K: 250,000 - 299,999                        0
            315 L: 300,000 and above                        1
            114 E: 90,000 - 109,999                         0
             27 A: Below 30,000                             0
             61 A: Below 30,000                             1
            206 F: 110,000 - 129,999                        1
            446 J: 190,000 - 249,999                        0
            196 E: 90,000 - 109,999                         1
             90 B: 30,000 - 49,999                          0
             99 C: 50,000 - 69,999                          1
            319 I: 170,000 - 189,999                        1
            165 I: 170,000 - 189,999                        0
            179 K: 250,000 - 299,999                        1
            142 H: 150,000 - 169,999                        0

NUM_CUSTOMERS   CUST_INCOME_LEVEL      HOME_THEATER_PACKAGE
            163 F: 110,000 - 129,999                        0
             83 D: 70,000 - 89,999                          1
             50 D: 70,000 - 89,999                          0
            328 L: 300,000 and above                        0
            519 J: 190,000 - 249,999                        1
            189 G: 130,000 - 149,999                        1
            150 G: 130,000 - 149,999                        0
            132 B: 30,000 - 49,999                          1
             72 C: 50,000 - 69,999                          0
            241 H: 150,000 - 169,999                        1


24 rows selected.
---------------------------
```

This completes the data exploration stage. OML supports Automatic Data Preparation (ADP). ADP is enabled through the model settings. When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model. This step is done during the Build Model stage. The commonly used methods of data preparation are binning, normalization, and missing value treatment.

Related Topics

- How ADP Transforms the Data

Build Model

Build your model using the training data set. Use the `DBMS_DATA_MINING.CREATE_MODEL2` procedure to build your model and specify the model settings.

For a supervised learning, like Classification, before creating the model, split the data into training and test data. Although you can use the entire data set to build a model, it is difficult to validate the model unless there are new data sets available. Therefore, to evaluate the model and to accurately assess the performance of the model on the same data, you generally split or separate the data into training and test data. You use the training data set to train the model and then use the test data set to test the accuracy of the model by running prediction queries. The testing data set already contains known values for the attribute that you want to predict. It is thus easy to determine whether the predictions of the model are correct.

Algorithm Selection

Before you build a model, choose the suitable algorithm. You can choose one of the following algorithms to solve a classification problem:

- Decision Tree
- Explicit Semantic Analysis (ESM)
- Generalized Linear Model (GLM)
- Naive Bayes
- Random Forest
- Support Vector Machine (SVM)
- XGBoost

From the above algorithms, ESM is more about Natural Language Processing (NLP) and text mining. ESM does not apply to this use case and data. If you were to select a relatively simple linear model like GLM, the prediction accuracy can be further improved by the Random Forest algorithm. Random Forest is an ensemble method that builds multiple decision trees on subsets of the data re-sampled at each time (bagging). This avoids the overfitting for a single decision tree. The random forest model is a widely used ensemble method that is known to have higher accuracy than linear models. Thus, Random Forest is selected for this use case.

For this use case, split the data into 60/40 as training and test data. You build the model using the training data and once the model is built, score the test data using the model.

The following steps guide you to split your data and build your model with the selected algorithm.

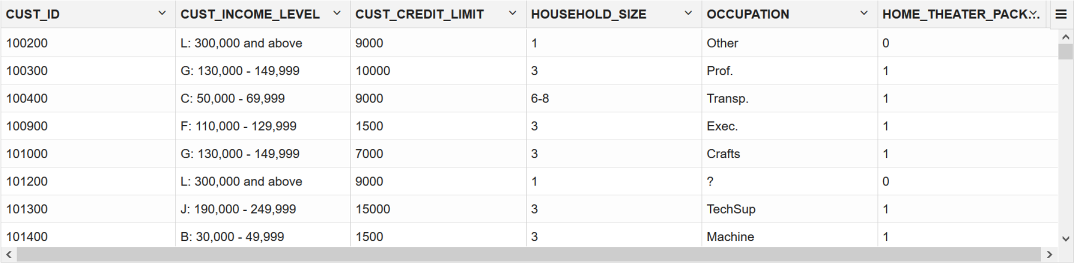
1. To create the training and test data with 60/40 split, run the following statement:

```
CREATE OR REPLACE VIEW TRAINING_DATA AS SELECT * FROM CUSTOMERDATA SAMPLE (60) SEED
(1);
--DBMS_OUTPUT.PUT_LINE ('Created TRAINING_DATA');
CREATE OR REPLACE VIEW TEST_DATA AS SELECT * FROM CUSTOMERDATA MINUS SELECT * FROM
TRAINING_DATA;
--DBMS_OUTPUT.PUT_LINE ('Created TEST_DATA');
```

```
View TRAINING_DATA created.
---------------------------
View TEST_DATA created.
```

2. To view the data in the `training_data` view, run the following statement:

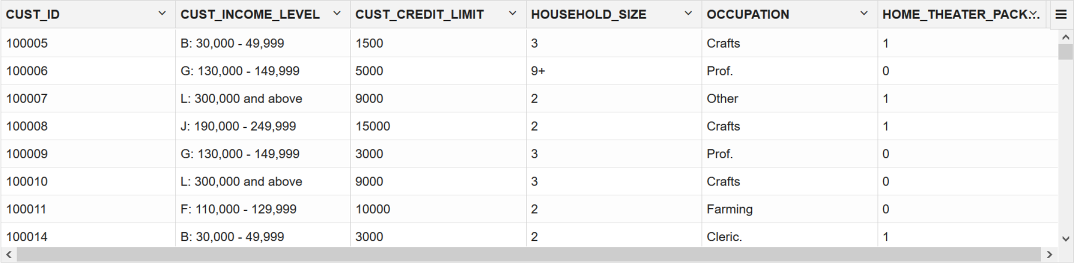
```
SELECT * FROM TRAINING_DATA;
```



CUST_ID	CUST_INCOME_LEVEL	CUST_CREDIT_LIMIT	HOUSEHOLD_SIZE	OCCUPATION	HOME_THEATER_PACK...
100200	L: 300,000 and above	9000	1	Other	0
100300	G: 130,000 - 149,999	10000	3	Prof.	1
100400	C: 50,000 - 69,999	9000	6-8	Transp.	1
100900	F: 110,000 - 129,999	1500	3	Exec.	1
101000	G: 130,000 - 149,999	7000	3	Crafts	1
101200	L: 300,000 and above	9000	1	?	0
101300	J: 190,000 - 249,999	15000	3	TechSup	1
101400	B: 30,000 - 49,999	1500	3	Machine	1

3. To view the data in the `test_data` view, run the following statement:

```
SELECT* FROM TEST_DATA;
```



CUST_ID	CUST_INCOME_LEVEL	CUST_CREDIT_LIMIT	HOUSEHOLD_SIZE	OCCUPATION	HOME_THEATER_PACK...
100005	B: 30,000 - 49,999	1500	3	Crafts	1
100006	G: 130,000 - 149,999	5000	9+	Prof.	0
100007	L: 300,000 and above	9000	2	Other	1
100008	J: 190,000 - 249,999	15000	2	Crafts	1
100009	G: 130,000 - 149,999	3000	3	Prof.	0
100010	L: 300,000 and above	9000	3	Crafts	0
100011	F: 110,000 - 129,999	10000	2	Farming	0
100014	B: 30,000 - 49,999	3000	2	Cleric.	1

4. To view the distribution of `HOME_THEATER_PACKAGE` (target) owners, run the following script:

```
%script
select HOME_THEATER_PACKAGE, count(1)
```

```
from training_data
group by HOME_THEATER_PACKAGE;



HOME_THEATER_PACKAGE   COUNT(1)
                     1       1506
                     0       1208

---------------------------
```

5. Build your model using the `CREATE_MODEL2` procedure. First, declare a variable to store model settings or hyperparameters. Run the following script:

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('MODEL_RF');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlist DBMS_DATA_MINING.SETTING_LIST;

BEGIN
    v_setlist('PREP_AUTO') := 'ON';
    v_setlist('ALGO_NAME') := 'ALGO_RANDOM_FOREST';
    v_setlist('RFOR_NUM_TREES') := '25';

    DBMS_DATA_MINING.CREATE_MODEL2(
      MODEL_NAME          =>  'MODEL_RF',
      MINING_FUNCTION     => 'CLASSIFICATION',
      DATA_QUERY          =>  'SELECT * FROM TRAINING_DATA',
      SET_LIST            =>  v_setlist,
      CASE_ID_COLUMN_NAME =>  'CUST_ID',
      TARGET_COLUMN_NAME  =>  'HOME_THEATER_PACKAGE');
END;
```

```
PL/SQL procedure successfully completed.

---------------------------

PL/SQL procedure successfully completed.
```

Examine the script:

- `v_setlist` is a variable to store `SETTING_LIST`.
- `SETTING_LIST` defines model settings or hyperparameters for your model.
- `DBMS_DATA_MINING` is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- `ALGO_NAME` specifies the algorithm name. Since you are using Random Forest as the algorithm, set `ALGO_RANDOM_FOREST`.

- PREP_AUTO is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is ON.
- RFOR_NUM_TREES is the number of trees in the forest. The value here is 25. Random forest resolves the overfitting problem by training multiple trees on distinct sampled subsets of the data instead of on the same, entire training set. The more trees you select, the more accuracy it can obtain. However, keep in mind that more trees mean more computation load and longer model building time. You need to do a trade-off between the time cost and model accuracy here. Choosing the number of trees equal to 25 allows you to build the model in a reasonably short time and obtain an accurate enough model.

The CREATE_MODEL2 procedure takes the following parameters:

- MODEL_NAME: A unique model name that you will give to the model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is MODEL_RF
- MINING_FUNCTION: Specifies the machine learning function. Since it is a classification problem in this use case, select CLASSIFICATION.
- DATA_QUERY: A query that provides training data for building the model. Here, the query is SELECT * FROM TRAINING_DATA.
- SET_LIST: Specifies SETTING_LIST.
- CASE_ID_COLUMN_NAME: A unique case identifier column in the build data. In this use case, case_id is CUST_ID. If there is a composite key, you must create a new attribute before creating the model. The CASE_ID assists with reproducible results, joining scores for individual customers with other data in, example, scoring data table.

> **Note:**
>
> OML uses either system-determined or default values for any parameters or settings not specified.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

Sometimes querying dictionary views and model detail views is sufficient to measure your model's performance. However, you can evaluate your model by computing test metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), confusion matrix, lift statistics, cost matrix, and so on. For Association Rules, you can inspect various rules to see if they reveal new insights for item dependencies (antecedent itemset implying consequent) or for unexpected relationships among items.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views.

The following steps help you to view different dictionary views and model detail views.

1. Run the following statement to view the settings in `USER_MINING_MODEL_SETTINGS`:

   ```
   %script

   SELECT SETTING_NAME, SETTING_VALUE
     FROM USER_MINING_MODEL_SETTINGS
     WHERE MODEL_NAME='MODEL_RF'
     ORDER BY SETTING_NAME;



   SETTING_NAME                   SETTING_VALUE
   ALGO_NAME                      ALGO_RANDOM_FOREST
   CLAS_MAX_SUP_BINS              32
   CLAS_WEIGHTS_BALANCED          OFF
   ODMS_DETAILS                   ODMS_ENABLE
   ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
   ODMS_RANDOM_SEED               0
   ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE
   PREP_AUTO                      ON
   RFOR_NUM_TREES                 25
   RFOR_SAMPLING_RATIO            .5
   TREE_IMPURITY_METRIC           TREE_IMPURITY_GINI
   TREE_TERM_MAX_DEPTH            16
   TREE_TERM_MINPCT_NODE          .05
   TREE_TERM_MINPCT_SPLIT         .1

   SETTING_NAME             SETTING_VALUE
   TREE_TERM_MINREC_NODE    10
   TREE_TERM_MINREC_SPLIT   20


   16 rows selected.
   ---------------------------
   ```

2. Run the following statement to see attribute information in `USER_MINING_MODEL_ATTRIBUTES` view:

```
%script
SELECT ATTRIBUTE_NAME, ATTRIBUTE_TYPE
FROM USER_MINING_MODEL_ATTRIBUTES
WHERE MODEL_NAME = 'MODEL_RF'
ORDER BY ATTRIBUTE_NAME;



ATTRIBUTE_NAME         ATTRIBUTE_TYPE
CUST_CREDIT_LIMIT      NUMERICAL
HOME_THEATER_PACKAGE   CATEGORICAL
HOUSEHOLD_SIZE         CATEGORICAL
OCCUPATION             CATEGORICAL

---------------------------
```

3. Run the following statement to view various model detail views from `USER_MINING_MODEL_VIEWS`:

```
%script
SELECT VIEW_NAME, VIEW_TYPE
  FROM USER_MINING_MODEL_VIEWS
  WHERE MODEL_NAME='MODEL_RF'
  ORDER BY VIEW_NAME;



VIEW_NAME       VIEW_TYPE
DM$VAMODEL_RF   Variable Importance
DM$VCMODEL_RF   Scoring Cost Matrix
DM$VGMODEL_RF   Global Name-Value Pairs
DM$VSMODEL_RF   Computed Settings
DM$VTMODEL_RF   Classification Targets
DM$VWMODEL_RF   Model Build Alerts


6 rows selected.
---------------------------
```

4. Now, view the Classification targets view. This view describes the target (`HOME_THEATER_PACKAGE`) distribution for classification models.

```
%script
SELECT* from DM$VTMODEL_RF;



PARTITION_NAME   TARGET_VALUE   TARGET_COUNT   TARGET_WEIGHT
                              0           1178
                              1           1549

---------------------------
```

The distribution value from this view validates the earlier target distribution that was obtained from the training data. The difference in the values is minimal.

Related Topics

- PREDICTION_SET

Test Your Model

In this use case, you are evaluating a classification model by computing Lift and Confusion Matrix on the test data with known target values and comparing the predicted values with the known values.

Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets your business requirements, it can then be applied to new data to predict the future. These matrices can help you to compare models to arrive at one model that satisfies your evaluation criteria.
Lift measures the degree to which the predictions of a classification model are better than randomly-generated predictions. Lift can be understood as a ratio of two percentages: the percentage of correct positive classifications made by the model to the percentage of actual positive classifications in the test data.

A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is n-by-n, where n is the number of classes.

1. Create a result table to store the predictions for each row with likely and unlikely probabilities. Run the following script:

   ```
   %script

   BEGIN EXECUTE IMMEDIATE 'DROP TABLE APPLY_RESULT PURGE';
   EXCEPTION WHEN OTHERS THEN NULL; END;
   /

   CREATE TABLE APPLY_RESULT AS
       SELECT cust_id, t.prediction, t.probability
       FROM TEST_DATA, TABLE(PREDICTION_SET(MODEL_RF USING *)) t;
   ```

   ```
   PL/SQL procedure successfully completed.
   ---------------------------
   Table APPLY_RESULT created.
   ---------------------------
   ```

   Examine the script:

   `APPLY_RESULT`: is a table that stores the results of the prediction.

   `TABLE(PREDICTION_SET(MODEL_RF USING *))`: is a table that has results from the `PREDICTION_SET` query. The `PREDICTION_SET` query returns probabilities for each row.

2. Compute lift by using the `DBMS_DATA_MINING.APPLY` and the `DBMS_DATA_MINING.COMPUTE_LIFT` procedures:

   ```
   %script

   BEGIN EXECUTE IMMEDIATE 'DROP TABLE APPLY_RESULT PURGE';
   EXCEPTION WHEN OTHERS THEN NULL; END;
   ```

```
/

BEGIN
  DBMS_DATA_MINING.APPLY('MODEL_RF','TEST_DATA','CUST_ID','APPLY_RESULT');


       DBMS_DATA_MINING.COMPUTE_LIFT (
          apply_result_table_name           => 'APPLY_RESULT',
          target_table_name                  => 'TEST_DATA',
          case_id_column_name               => 'CUST_ID',
          target_column_name                 => 'HOME_THEATER_PACKAGE',
          lift_table_name                       => 'LIFT_TABLE',
          positive_target_value           =>  to_char(1),
          score_column_name                  => 'PREDICTION',
          score_criterion_column_name    => 'PROBABILITY',
          num_quantiles                       =>  10,
          cost_matrix_table_name             =>  null,
          apply_result_schema_name         =>  null,
          target_schema_name                 =>  null,
          cost_matrix_schema_name           =>  null,
          score_criterion_type             =>  'PROBABILITY');


END;
```

```
PL/SQL procedure successfully completed.
---------------------------
PL/SQL procedure successfully completed.
```

Examine the script:

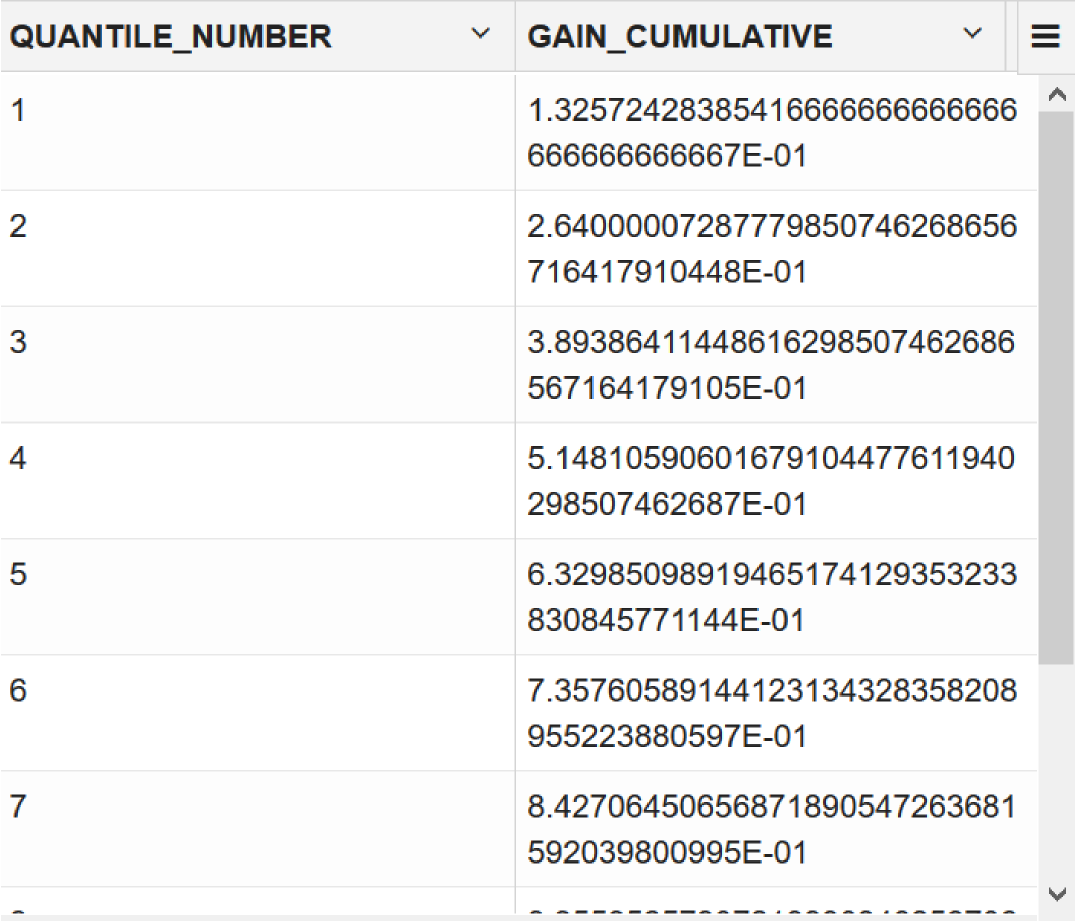
- `DBMS_DATA_MINING.APPLY`: This procedure creates a table in the user's schema to hold the results. The `APPLY` procedure generates predictions (scores) in a target column. The `APPLY` procedure has the following parameters:
  - `model_name`: Name of the model in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is `MODEL_RF`.
  - `data_table_name`: Name of table or view containing the data to be scored. Here, you are using `TEST_DATA`.
  - `case_id_column_name`: Name of the case identifier column. The case ID is `CUST_ID`.
  - `result_table_name`: Name of the table in which to store apply results. Here, the result table name is `APPLY_RESULT`.
- `DBMS_DATA_MINING.COMPUTE_LIFT`: This procedure computes lift and stores them in the user's schema. To compute lift, one of the target values must be designated as the positive class.
  The `COMPUTE_LIFT` procedure has the following parameters:
  - `apply_result_table_name`: Table containing the predictions. For this use case, it is `APPLY_RESULT`.
  - `target_table_name`: Table containing the known target values from the test data. In this use case, the target table name is `TEST_DATA`.

- case_id_column_name: Case ID column in the apply results table. Must match the case identifier in the targets table. The case ID column is CUST_ID.
- target_column_name: Target column in the targets table. Contains the known target values from the test data. In this use case, the target is HOME_THEATER_PACKAGE.
- lift_table_name: Table containing the lift statistics. The table will be created by the procedure in the user's schema. Type LIFT_TABLE.
- positive_target_value: The positive class. This should be the class of interest, for which you want to calculate lift. If the target column is a NUMBER, you can use the TO_CHAR() operator to provide the value as a string.
- score_column_name: Column containing the predictions in the apply results table. The default column name is 'PREDICTION', which is the default name created by the APPLY procedure.
- score_criterion_column_name: Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, the class with the lowest cost is predicted. The score_criterion_type parameter indicates whether probabilities or costs will be used for scoring. The default column name is 'PROBABILITY', which is the default name created by the APPLY procedure.
- num_quantiles: Number of quantiles to be used in calculating lift. The default is 10.
- cost_matrix_table_name: (Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the score_criterion_type parameter is set to 'COST', the costs will be used as the scoring criteria.
- apply_result_schema_name: Schema of the apply results table. If null, the user's schema is assumed.
- target_schema_name: Schema of the table containing the known targets. If null, the user's schema is assumed.
- cost_matrix_schema_name: Schema of the cost matrix table, if one is provided. If null, the user's schema is assumed.
- score_criterion_type: Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the score_criterion_column_name parameter. The default value of score_criterion_type is 'PROBABILITY'. To use costs as the scoring criterion, specify 'COST'. If score_criterion_type is set to 'COST' but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring.

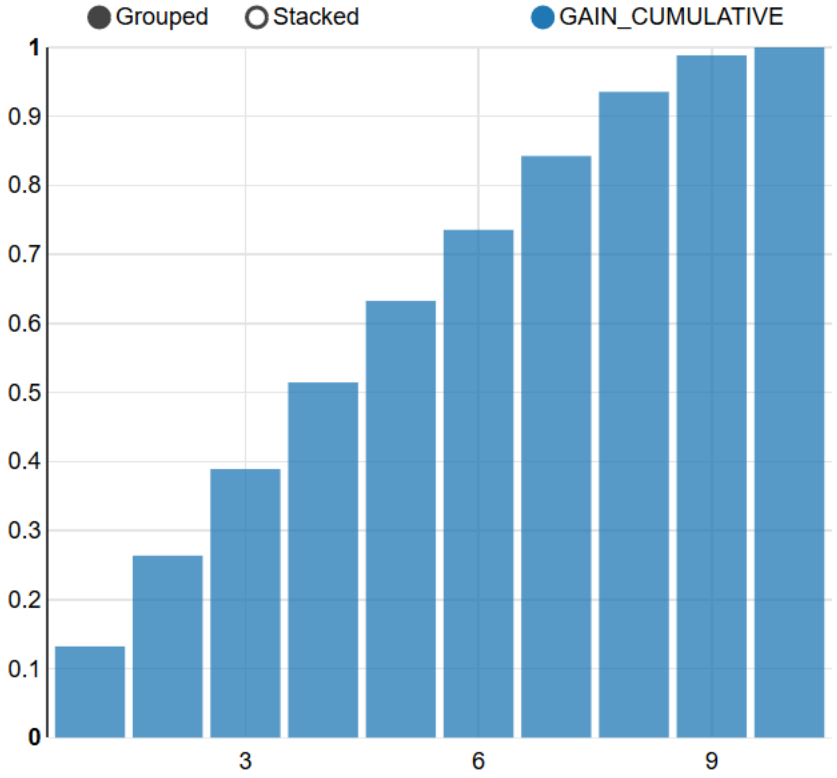
3. To view the cumulative gains, run the following statement:

   Cumulative gain is the ratio of the cumulative number of positive targets (HOME_THEATER_PACKAGE) to the total number of positive targets of a quantile. Cumulative gains act as a visual aid for measuring performance of a model. The chart consists of a curve and a baseline. The greater the area between the curve and the baseline, the better the model.

   ```
   %sql
   SELECT QUANTILE_NUMBER, GAIN_CUMULATIVE FROM LIFT_TABLE;
   ```



QUANTILE_NUMBER	GAIN_CUMULATIVE
1	1.32572428385416666666666666666666666667E-01
2	2.64000007287779850746268656716417910448E-01
3	3.89386411448616298507462686567164179105E-01
4	5.14810590601679104477611940298507462687E-01
5	6.32985098919465174129353233830845771144E-01
6	7.35760589144123134328358208955223880597E-01
7	8.42706450656871890547263681592039800995E-01



4. To compute confusion matrix, run the following statement:

   A confusion matrix evaluates the prediction results. It makes it easy to understand and estimate the effects of wrong predictions. You can observe the number and percentages in each cell of this matrix and notice how often the model predicted accurately.

```
%script

DECLARE
   v_accuracy NUMBER;
   BEGIN
        DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX (
                   accuracy => v_accuracy,
                   apply_result_table_name => 'apply_result',
                   target_table_name => 'test_data',
                   case_id_column_name => 'cust_id',
                   target_column_name => 'HOME_THEATER_PACKAGE',
                   confusion_matrix_table_name => 'confusion_matrix',
                   score_column_name => 'PREDICTION',
                   score_criterion_column_name => 'PROBABILITY',
                   cost_matrix_table_name => null,
                   apply_result_schema_name => null,
                   target_schema_name => null,
                   cost_matrix_schema_name => null,
                   score_criterion_type => 'PROBABILITY');
        DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' ||
ROUND(v_accuracy,4));
      END;
      /


**** MODEL ACCURACY ****: .696
---------------------------
PL/SQL procedure successfully completed.
---------------------------
```

   Examine the script:

   `v_accuracy` is a variable declared for this procedure to store and output the model accuracy percentage.

   The `COMPUTE_CONFUSION_MATRIX` procedure has the following parameters:

   - `accuracy`: Output parameter containing the overall percentage accuracy of the predictions. Here, it is `v_accuracy`.
   - `apply_result_table_name`: Table containing the predictions. In this use case, it is `APPLY_RESULT`.
   - `target_table_name`: Table containing the known target values from the test data. In this use case, you are using `TEST_DATA`.
   - `case_id_column_name`: Case ID column in the apply results table. Must match the case identifier in the targets table. Here, it is `CUST_ID`.
   - `target_column_name`: Target column in the targets table. Contains the known target values from the test data. In this use case, the target column is `HOME_THEATER_PACKAGE`.

- `confusion_matrix_table_name`: Table containing the confusion matrix. The table will be created by the procedure in the user's schema. Here set it as `confusion_matrix`.
- `score_column_name`: Column containing the predictions in the apply results table. The default column name is `PREDICTION`, which is the default name created by the `APPLY` procedure.
- `score_criterion_column_name`: Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, the class with the lowest cost is predicted. The `score_criterion_type` parameter indicates whether probabilities or costs will be used for scoring. The default column name is `'PROBABILITY'`, which is the default name created by the `APPLY` procedure.
- `cost_matrix_table_name`: (Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the `score_criterion_type` parameter is set to `'COSTS'`, the costs in this table will be used as the scoring criteria. Otherwise, set it as `null`.
- `apply_result_schema_name`: Schema of the apply results table. If null, the user's schema is assumed.
- `target_schema_name`: Schema of the table containing the known targets. If null, the user's schema is assumed.
- `cost_matrix_schema_name`: Schema of the cost matrix table, if one is provided. If null, the user's schema is assumed.
- `score_criterion_type`: Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the `score_criterion_column_name` parameter. The default value of `score_criterion_type` is `'PROBABILITY'`. To use costs as the scoring criterion, specify `'COST'`. If `score_criterion_type` is set to `'COST'` but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring.

`DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' || ROUND(v_accuracy,4))`: Outputs the model accuracy percentage rounded to 4 digits after the decimal.

5. To check the confusion matrix with predicted values and actual values, run the following statement:

```
select * from confusion_matrix;



ACTUAL_TARGET_VALUE   PREDICTED_TARGET_VALUE   VALUE
                  0                        1     501
                  0                        0     282
                  1                        0      38
                  1                        1     952

---------------------------
```

The value column here indicates classification. From this confusion matrix, the model has predicted actual positive class (also called as True Positive (TP)) for this use case 952 times and incorrectly predicted (also called as False Negative (FN)) for this use case 38 times. The model correctly predicted the negative class (also called true negative (TN)) for this use case 282 times and incorrectly predicted (also called false positive (FP)) for this use case 501 times.

The accuracy percentage of 69% shows that the model is fairly good for this use case.

Related Topics

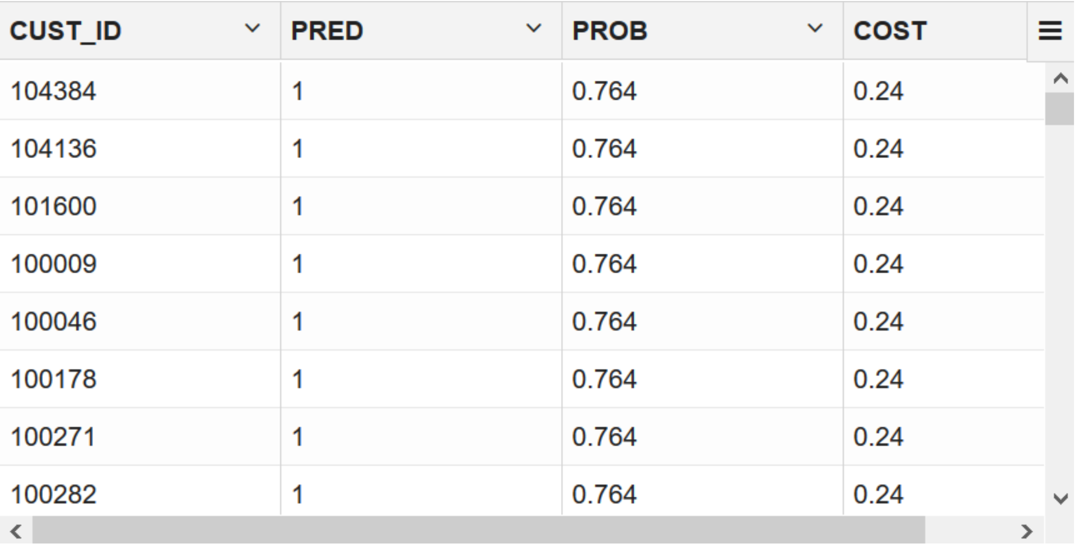
- PREDICTION_SET

Score

You are ready to predict the likely customers for the `HOME_THEATER_PACKAGE` responders. For classification problems, you can use `PREDICTION`, `PREDICTION_PROBABILITY`, or use analytic syntax to arrive at predictions.

1. To view customers who have more than 50% chance of buying a home theater package, run the following statement:

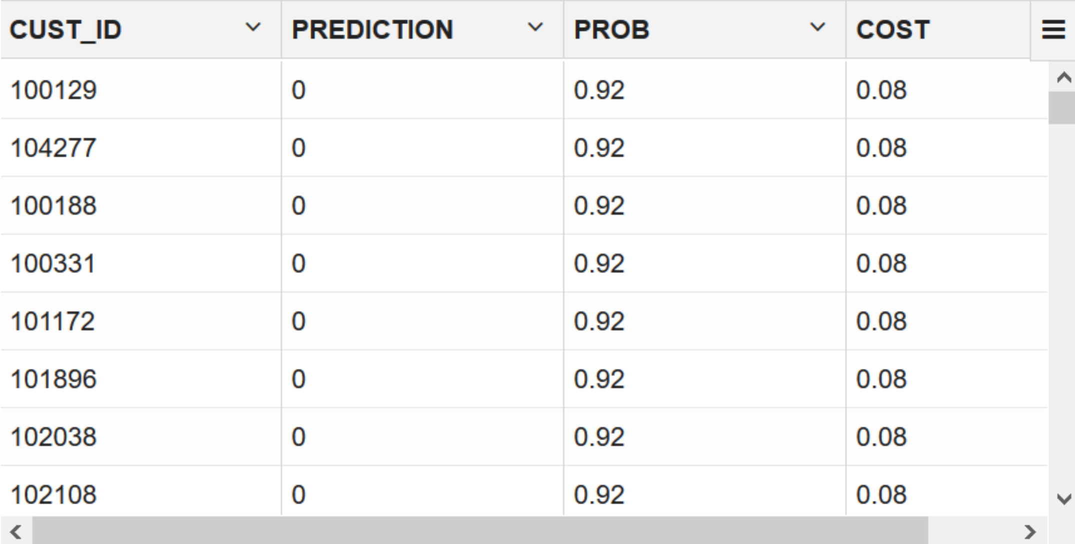
```
%sql
SELECT CUST_ID, PREDICTION PRED, ROUND(PROBABILITY,3) PROB, ROUND(COST,2) COST
  FROM APPLY_RESULT WHERE PREDICTION = 1 AND PROBABILITY > 0.5
  ORDER BY PROBABILITY DESC;
```



CUST_ID	PRED	PROB	COST
104384	1	0.764	0.24
104136	1	0.764	0.24
101600	1	0.764	0.24
100009	1	0.764	0.24
100046	1	0.764	0.24
100178	1	0.764	0.24
100271	1	0.764	0.24
100282	1	0.764	0.24

2. You can score on multiple rows of test data. This is called batch scoring. This step shows how you can view and select customers who are likely or unlikely to respond to `HOME_THEATER_PACKAGE` with a probability of more than 50% and a cost matrix.

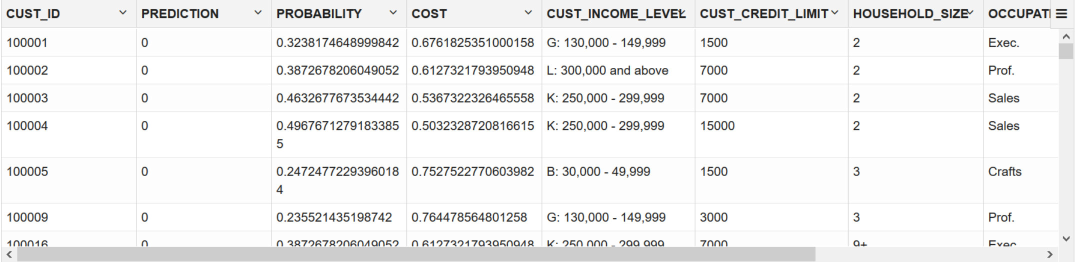
```
%sql

SELECT CUST_ID, PREDICTION, ROUND(PROBABILITY,2) PROB, ROUND(COST,2) COST
  FROM APPLY_RESULT WHERE PREDICTION = ${PREDICTION='1','1'|'0'}
  AND PROBABILITY > 0.5 ORDER BY PROBABILITY DESC;
```



CUST_ID	PREDICTION	PROB	COST
100129	0	0.92	0.08
104277	0	0.92	0.08
100188	0	0.92	0.08
100331	0	0.92	0.08
101172	0	0.92	0.08
101896	0	0.92	0.08
102038	0	0.92	0.08
102108	0	0.92	0.08

3. To interactively view probability of `HOME_THEATER_PACKAGE` respondents, run the following statement:

```
%sql
SELECT A.*, B.*
  FROM APPLY_RESULT A, TEST_DATA B
  WHERE PREDICTION = ${PREDICTION='1','1'|'0'} AND A.CUST_ID = B.CUST_ID;
```

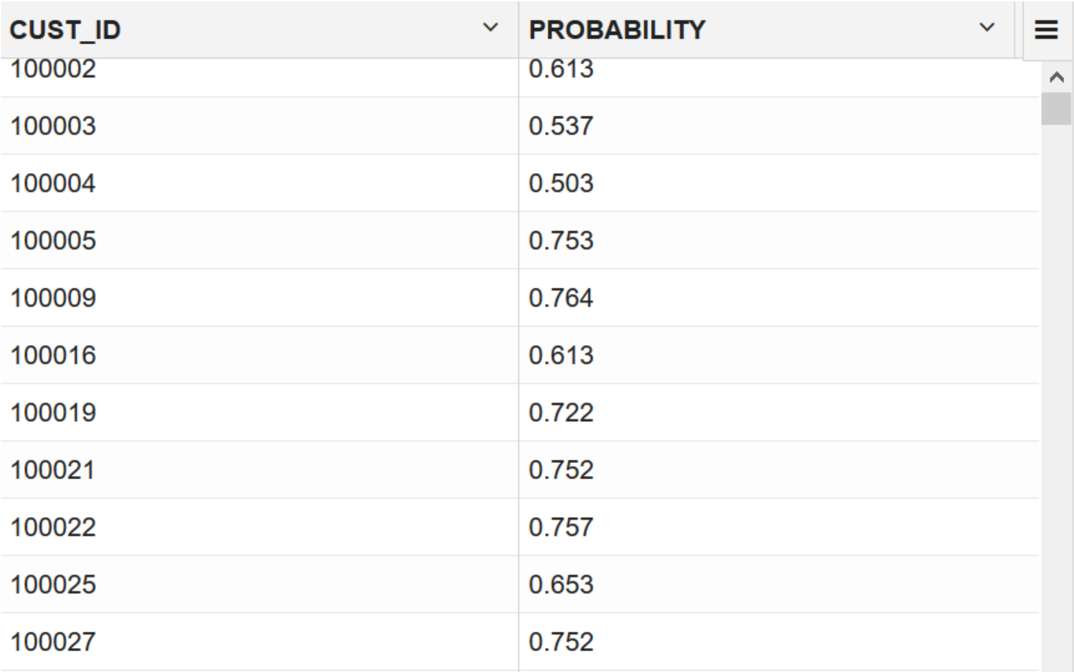


CUST_ID	PREDICTION	PROBABILITY	COST	CUST_INCOME_LEVEL	CUST_CREDIT_LIMIT	HOUSEHOLD_SIZE	OCCUPATI
100001	0	0.3238174648999842	0.6761825351000158	G: 130,000 - 149,999	1500	2	Exec.
100002	0	0.3872678206049052	0.6127321793950948	L: 300,000 and above	7000	2	Prof.
100003	0	0.4632677673534442	0.5367322326465558	K: 250,000 - 299,999	7000	2	Sales
100004	0	0.49676712791833855	0.5032328720816615	K: 250,000 - 299,999	15000	2	Sales
100005	0	0.24724772293960184	0.7527522770603982	B: 30,000 - 49,999	1500	3	Crafts
100009	0	0.235521435198742	0.764478564801258	G: 130,000 - 149,999	3000	3	Prof.
100016	0	0.3872678206049052	0.6127321793950948	K: 250,000 - 299,999	7000	9+	Exec.

4. To dynamically score and select customers with more than 50% chance of purchasing a home theater package, run the following statement:

```
%sql

SELECT *
FROM (  SELECT CUST_ID, ROUND(PREDICTION_PROBABILITY(MODEL_RF, '1'  USING A.*),3)
PROBABILITY
    FROM TEST_DATA A)
WHERE PROBABILITY > 0.5;
```

You can use `PREDICTION_PROBABILITY` to score in real-time.



CUST_ID	PROBABILITY
100002	0.613
100003	0.537
100004	0.503
100005	0.753
100009	0.764
100016	0.613
100019	0.722
100021	0.752
100022	0.757
100025	0.653
100027	0.752

5. To apply the model to a single record (singleton scoring), run the following statement:

```
%script
SELECT ROUND(PREDICTION_PROBABILITY(MODEL_RF, '1' USING
                                    '3' AS HOUSEHOLD_SIZE,
                                     5 AS YRS_RESIDENCE,
                                     1 AS CUST_INCOME_LEVEL),3)
PROBABILITY_HOME_THEATER_PACKAGE_RESPONDER
  FROM DUAL;
```

This may be useful if you want to test the model manually and see how the model works.

```
PROBABILITY_HOME_TEATER_PACKAGE_RESPONDER
                                     0.65

---------------------------
```

To conclude, you have successfully identified customers who are likely to purchase `HOME_THEATER_PACKAGE`. This prediction helps to promote and offer home theater package to the target customers.

Clustering Use Case Scenario

You're a game data scientist. Marketing team wants to promote a new game and want customers who bought a gaming product with a high credit limit. They want to segment customers based on game purchases and credit level. You help them identify target customers and segment the population using *k*-Means.

Related Content

Topic	Link
OML4SQL GitHub Example	Clustering - k-Means

Topic	Link
`CREATE_MODEL2` Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING - Model Settings
k-Means Settings	DBMS_DATA_MINING - Algorithm Settings: k-Means
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views
k-Means - Model Detail Views	Model Detail Views for k-Means
About Clustering	About Clustering
About *k*-Means	About k-Means

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data Set
  The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.
- Database
  Select or create database out of the following options:
  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements
  Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Related Topics

- Create a Notebook
- Edit your Notebook
- Installing Sample Schemas

Load Data

Access the data set from the SH Schema and explore the data to understand the attributes.

> **Remember:**
>
> The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.

To understand the data, you will perform the following:

- Access the data.
- Examine the various attributes or columns of the data set.
- Assess data quality (by exploring the data).

Access Data

You will use `CUSTOMERS` and `SUPPLEMENTARY_DEMOGRAPHICS` table data from the SH schema.

Examine Data

The following table displays information about the attributes from `SUPPLEMENTARY_DEMOGRAPHICS`:

Attribute Name	Information
`CUST_ID`	The ID of the customer
`EDUCATION`	Educational information of the customer
`OCCUPATION`	Occupation of the customer
`HOUSEHOLD_SIZE`	People per house
`YRS_RESIDENCE`	Number of years of residence
`AFFINITY_CARD`	Whether the customer holds an affinity card
`BULK_PACK_DISKETTES`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`FLAT_PANEL_MONITOR`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`HOME_THEATER_PACKAGE`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`BOOKKEEPING_APPLICATION`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`PRINTER_SUPPLIES`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`Y_BOX_GAMES`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`OS_DOC_SET_KANJI`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No
`COMMENTS`	Product. Indicates whether the customer already owns the product. 1 means Yes. 0 means No

Explore Data

Once the data is accessible, explore the data to understand and assess the quality of the data. At this stage assess the data to identify data types and noise in the data. Look for missing values and numeric outlier values.

Assess Data Quality

To assess the data, first, you must be able to view the data in your database. For this reason, you will use SQL statements to query the SH.CUSTOMERS and the SH.SUPPLEMENTARY_DEMOGRAPHICS table.

If you are working with Oracle Autonomous Database, you can use the Oracle Machine Learning (OML) Notebooks for your data science project, including assessing data quality. If you are using on-premise Oracle Database, you can use the Oracle SQL Developer to assess data quality. Query the SH schema as described.

> **Note:**
>
> Each record in the database is called a case and each case is identified by a case_id. In this use case, CUST_ID is the case_id.

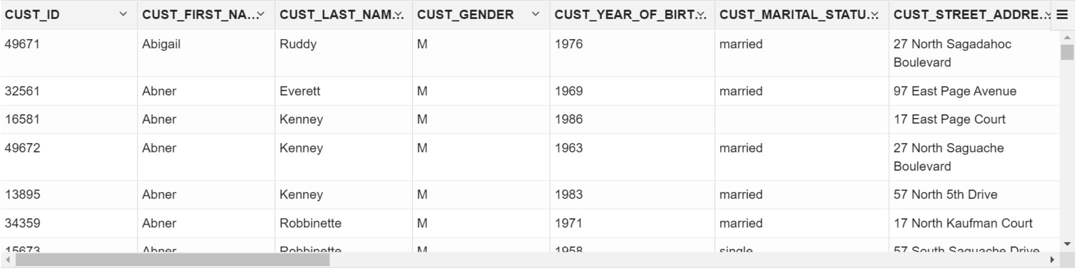
The following steps help you with the exploratory analysis of the data:

1. View the data in the SH.CUSTOMERS table by running the following query:

```
SELECT * FROM SH.CUSTOMERS;
```

2. To see distinct data from the table, run the following query:

```
SELECT DISTINCT * FROM SH.CUSTOMERS;
```



CUST_ID	CUST_FIRST_NA...	CUST_LAST_NAM...	CUST_GENDER	CUST_YEAR_OF_BIRT...	CUST_MARITAL_STATU...	CUST_STREET_ADDRE...
49671	Abigail	Ruddy	M	1976	married	27 North Sagadahoc Boulevard
32561	Abner	Everett	M	1969	married	97 East Page Avenue
16581	Abner	Kenney	M	1986		17 East Page Court
49672	Abner	Kenney	M	1963	married	27 North Saguache Boulevard
13895	Abner	Kenney	M	1983	married	57 North 5th Drive
34359	Abner	Robbinette	M	1971	married	17 North Kaufman Court
15673	Abner	Robbinette	M	1958	single	57 South Saguache Drive

3. Find the COUNT rows in the data set, run the following statement:

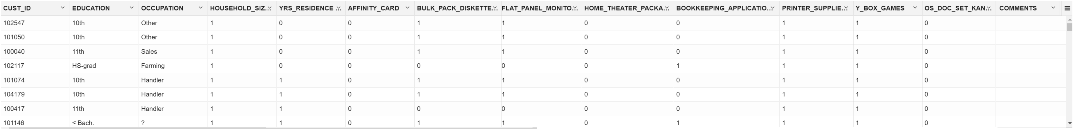
```
SELECT DISTINCT COUNT(*) from SH.CUSTOMERS;


COUNT(*)
     55500
---------------------------
```

4. To find distinct or unique customers in the table, run the following statement:

```
%script
SELECT COUNT (DISTINCT CUST_ID) FROM SH.CUSTOMERS;


COUNT(DISTINCTCUST_ID)
                 55500
---------------------------
```

5. Similarly, query the `SH.SUPPLEMENTARY_DEMOGRAPHICS` table.

```
SELECT * FROM SH.SUPPLEMENTARY_DEMOGRAPHICS;
```



CUST_ID	EDUCATION	OCCUPATION	HOUSEHOLD_SIZ..	YRS_RESIDENCE ..	AFFINITY_CARD	BULK_PACK_DISKETTE..	FLAT_PANEL_MONITO..	HOME_THEATER_PACKA..	BOOKKEEPING_APPLICATIO..	PRINTER_SUPPLIE..	Y_BOX_GAMES	OS_DOC_SET_KAN..	COMMENTS
102547	10th	Other	1	0	0	1	1	0	0	1	1	0	
101050	10th	Other	1	0	0	1	1	0	0	1	1	0	
100040	11th	Sales	1	0	0	1	1	0	0	1	1	0	
102117	HS-grad	Farming	1	0	0	0	0	0	1	1	1	0	
101074	10th	Handler	1	1	0	1	1	0	0	1	1	0	
104179	10th	Handler	1	1	0	1	1	0	0	1	1	0	
100417	11th	Handler	1	1	0	0	0	0	0	1	1	0	
101146	< Bach.	?	1	1	0	1	1	0	1	1	1	0	

6. To view the count of rows in the `SH.SUPPLEMENTARY_DEMOGRAPHICS` table, run the following statement:

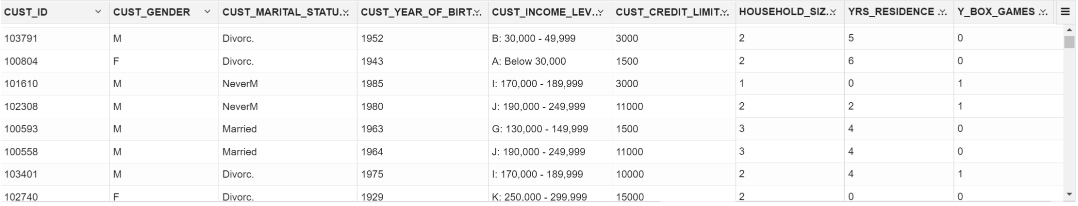
```
SELECT COUNT(*) from SH.SUPPLEMENTARY_DEMOGRAPHICS;


COUNT(*)
      4500
---------------------------
```

7. Create a table called `CUSTOMERDATA` by selecting the required columns from the `SH.CUSTOMERS` and the `SH.SUPPLIMENTARY_DEMOGRAPHICS` tables.

```
%script

CREATE OR REPLACE VIEW CUSTOMERDATA AS
   SELECT a.CUST_ID, a.CUST_GENDER, a.CUST_MARITAL_STATUS,
          a.CUST_YEAR_OF_BIRTH, a.CUST_INCOME_LEVEL, a.CUST_CREDIT_LIMIT,
          b.HOUSEHOLD_SIZE, b.YRS_RESIDENCE, b.Y_BOX_GAMES
   FROM SH.CUSTOMERS a, SH.SUPPLEMENTARY_DEMOGRAPHICS b
   WHERE a.CUST_ID = b.CUST_ID;



View CUSTOMERDATA created.
```

8. View the `CUSTOMERDATA` table.

```
SELECT * FROM CUSTOMERDATA;
```



CUST_ID	CUST_GENDER	CUST_MARITAL_STATU..	CUST_YEAR_OF_BIRT..	CUST_INCOME_LEV..	CUST_CREDIT_LIMIT..	HOUSEHOLD_SIZ..	YRS_RESIDENCE ..	Y_BOX_GAMES ..
103791	M	Divorc.	1952	B: 30,000 - 49,999	3000	2	5	0
100804	F	Divorc.	1943	A: Below 30,000	1500	2	6	0
101610	M	NeverM	1985	I: 170,000 - 189,999	3000	1	0	1
102308	M	NeverM	1980	J: 190,000 - 249,999	11000	2	2	1
100593	M	Married	1963	G: 130,000 - 149,999	1500	3	4	0
100558	M	Married	1964	J: 190,000 - 249,999	11000	3	4	0
103401	M	Divorc.	1975	I: 170,000 - 189,999	10000	2	4	1
102740	F	Divorc.	1929	K: 250,000 - 299,999	15000	2	0	0

9. Find the count of rows in the new `CUSTOMERDATA` table:

```
SELECT COUNT(*) FROM CUSTOMERDATA;


COUNT(*)
      4500
---------------------------
```

10. To view the data type of the columns, run the following statement:

```
%script
DESCRIBE CUSTOMERDATA;



Name                Null?    Type
-------------------- -------- ------------
CUST_ID              NOT NULL  NUMBER
CUST_GENDER        NOT NULL  CHAR(1)
CUST_MARITAL_STATUS          VARCHAR2(20)
CUST_YEAR_OF_BIRTH NOT NULL  NUMBER(4)
CUST_INCOME_LEVEL            VARCHAR2(30)
CUST_CREDIT_LIMIT            NUMBER
HOUSEHOLD_SIZE               VARCHAR2(21)
YRS_RESIDENCE                NUMBER
Y_BOX_GAMES               NUMBER(10)

---------------------------
```

11. To check if there are any missing values (NULL values), run the following statement:

```
SELECT COUNT(*) FROM CUSTOMERDATA WHERE CUST_ID=NULL OR CUST_GENDER=NULL
 OR CUST_MARITAL_STATUS=NULL OR CUST_YEAR_OF_BIRTH=NULL OR CUST_INCOME_LEVEL=NULL
 OR CUST_CREDIT_LIMIT=NULL OR HOUSEHOLD_SIZE=NULL OR YRS_RESIDENCE=NULL OR
Y_BOX_GAMES=NULL;


COUNT(*)
         0
---------------------------
```

NULLs, if found, are automatically handled by the OML algorithms. Alternately, you can manually replace NULLs with `NVL` SQL function.

This completes the data exploration stage. OML supports Automatic Data Preparation (ADP). ADP is enabled through the model settings. When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model. This step is done during the Build Model stage. The commonly used methods of data preparation are binning, normalization, and missing value treatment.

Related Topics

- How ADP Transforms the Data

Build Model

Build your model using your data set. Use the DBMS_DATA_MINING.CREATE_MODEL2 procedure to build your model and specify the model settings.

To evaluate the model and to accurately assess the performance of the model on the same data, you generally split or separate the data into training and test data. For an unsupervised learning, like Clustering, you do not have labels or predictors to calculate the accuracy or assess the performance. Thus, you can create a model using your data set without splitting. For an unsupervised learning, you don't have a real way of knowing how good your model is. So, a training or a test split is not useful.

Algorithm Selection

Before you build a model, choose the suitable algorithm. You can choose one of the following algorithms to solve a clustering problem:

- *k*-Means
- Expectation Maximization (EM)
- Orthogonal Cluster (O-Cluster)

K-Means does not assume a particular distribution of the data. The *k*-Means algorithm is a distance-based clustering algorithm that partitions the data into a specified number of clusters. The EM algorithm is a probability density estimation technique. EM method is based on assumption that the data has several clusters and each cluster is distributed according to a certain Gaussian distribution. O-Cluster is a neighbor based method. It identifies areas of high density in the data and separates the dense areas into clusters. It is able to cluster data points that forms a certain shape, which sometimes can be a complex pattern like a circle, spiral, or even a tie shape.

K-Means tends to cluster points only close to each other and does not necessarily cluster the data based on the shapes. Therefore, *K*-Means method is the one with the simplest assumption. Thus, it is the clustering method to start with.

The following steps guide you to build your model with the selected algorithm.

- Build your model using the CREATE_MODEL2 procedure. First, declare a variable to store model settings or hyperparameters. Run the following script:

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('KM_SH_CLUS1');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlist DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlist('ALGO_NAME')        := 'ALGO_KMEANS';
    V_setlist('PREP_AUTO')        := 'ON';
    V_setlist('KMNS_DISTANCE')    := 'KMNS_EUCLIDEAN';
    V_setlist('KMNS_DETAILS')     := 'KMNS_DETAILS_ALL';
    V_setlist('KMNS_ITERATIONS')  := '10';
    V_setlist('KMNS_NUM_BINS')    := '10';
    v_setlist('CLUS_NUM_CLUSTERS'):= '1';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'KM_SH_CLUS1',
        MINING_FUNCTION     => 'CLUSTERING',
        DATA_QUERY          => 'select * from CUSTOMERDATA',
        SET_LIST            => v_setlist,
        CASE_ID_COLUMN_NAME => 'CUST_ID');
END;




PL/SQL procedure successfully completed.
---------------------------
PL/SQL procedure successfully completed.
```

Examine the script:

- v_setlist is a variable to store SETTING_LIST.
- SETTING_LIST specifies model settings or hyperparameters for our model.
- DBMS_DATA_MINING is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- ALGO_NAME specifies the algorithm name. Since you are using the *k*-Means as your algorithm, set ALGO_KMEANS.
- PREP_AUTO is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is ON.
- KMNS_DISTANCE is the distance function that measures the similarity between the cases for *k*-Means. The value here is KMNS_EUCLIDEAN. This is the default value.
- KMNS_DETAILS determines the level of cluster details. KMNS_DETAILS_ALL computes cluster hierarchy, record counts, descriptive statistics (means, variances, modes, histograms, and rules).
- KMNS_ITERATIONS defines the maximum number of iterations for *k*-Means. The algorithm iterates until either the maximum number of iterations are reached or the minimum Convergence Tolerance, specified in KMNS_CONV_TOLERANCE, is satisfied. The default number of iterations is 20.
- KMNS_NUM_BINS provides a number of bins in the attribute histogram produced by *k*-Means.
- CLUS_NUM_CLUSTERS is the maximum number of leaf clusters generated by a clustering algorithm. The algorithm may return fewer clusters, depending on the data. Enhanced *k*-Means usually produces the exact number of clusters specified by CLUS_NUM_CLUSTERS, unless there are fewer distinct data points.

The CREATE_MODEL2 procedure takes the following parameters:

- MODEL_NAME: A unique model name that you will give to your model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is KM_SH_CLUS1.
- MINING_FUNCTION: Specifies the machine learning function. Since you are solving a clustering problem in this use case, select CLUSTERING.
- DATA_QUERY: A query that provides training data for building the model. Here, the query is SELECT * FROM CUSTOMERDATA.
- SET_LIST: Specifies SETTING_LIST.
- CASE_ID_COLUMN_NAME: A unique case identifier column in the build data. In this use case, case_id is CUST_ID. If there is a composite key, you must create a new attribute

before creating the model. This may involve concatenating values from the columns, or mapping a unique identifier to each distinct combination of values. The `CASE_ID` assists with reproducible results, joining scores for individual customers with other data in, example, scoring data table.

> **Note:**
>
> Any parameters or settings not specified are either system-determined or default values are used.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

Sometimes querying dictionary views and model detail views is sufficient to measure your model's performance. However, you can evaluate your model by computing test metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), confusion matrix, lift statistics, cost matrix, and so on. For Association Rules, you can inspect various rules to see if they reveal new insights for item dependencies (antecedent itemset implying consequent) or for unexpected relationships among items.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views.

The following steps help you to view different dictionary views and model detail views.

1. Run the following statement to view the settings in `USER_MINING_MODEL_SETTINGS`:

   ```
   %script
   SELECT SETTING_NAME, SETTING_VALUE
   FROM USER_MINING_MODEL_SETTINGS
   ```

```
WHERE MODEL_NAME = 'KM_SH_CLUS1'
ORDER BY SETTING_NAME;


SETTING_NAME                   SETTING_VALUE
ALGO_NAME                      ALGO_KMEANS
CLUS_NUM_CLUSTERS              1
KMNS_CONV_TOLERANCE            .001
KMNS_DETAILS                   KMNS_DETAILS_ALL
KMNS_DISTANCE                  KMNS_EUCLIDEAN
KMNS_ITERATIONS                3
KMNS_MIN_PCT_ATTR_SUPPORT      .1
KMNS_NUM_BINS                  10
KMNS_RANDOM_SEED               0
KMNS_SPLIT_CRITERION           KMNS_VARIANCE
ODMS_DETAILS                   ODMS_ENABLE
ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE
PREP_AUTO                      ON



14 rows selected.


---------------------------
```

2. Run the following statement to see attribute information in `USER_MINING_MODEL_ATTRIBUTES` view:

```
%script
SELECT ATTRIBUTE_NAME, ATTRIBUTE_TYPE
FROM USER_MINING_MODEL_ATTRIBUTES
WHERE MODEL_NAME = 'KM_SH_CLUS1'
ORDER BY ATTRIBUTE_NAME;



ATTRIBUTE_NAME        ATTRIBUTE_TYPE
CUST_CREDIT_LIMIT     NUMERICAL
CUST_GENDER           CATEGORICAL
CUST_INCOME_LEVEL     CATEGORICAL
CUST_MARITAL_STATUS   CATEGORICAL
CUST_YEAR_OF_BIRTH    NUMERICAL
HOUSEHOLD_SIZE        CATEGORICAL
YRS_RESIDENCE         NUMERICAL
Y_BOX_GAMES           NUMERICAL


8 rows selected.
---------------------------
```

3. Run the following statement to see information on various views in `USER_MINING_MODEL_VIEWS`:

```
%script
SELECT VIEW_NAME, VIEW_TYPE FROM USER_MINING_MODEL_VIEWS
```

```
WHERE MODEL_NAME='KM_SH_CLUS1'
ORDER BY VIEW_NAME;


VIEW_NAME           VIEW_TYPE
DM$VAKM_SH_CLUS1    Clustering Attribute Statistics
DM$VCKM_SH_CLUS1    k-Means Scoring Centroids
DM$VDKM_SH_CLUS1    Clustering Description
DM$VGKM_SH_CLUS1    Global Name-Value Pairs
DM$VHKM_SH_CLUS1    Clustering Histograms
DM$VNKM_SH_CLUS1    Normalization and Missing Value Handling
DM$VRKM_SH_CLUS1    Clustering Rules
DM$VSKM_SH_CLUS1    Computed Settings
DM$VWKM_SH_CLUS1    Model Build Alerts


9 rows selected.


---------------------------
```

4. Now, view the Clustering Description model detail view:

```
SELECT CLUSTER_ID CLU_ID, RECORD_COUNT REC_CNT, PARENT,
       TREE_LEVEL, ROUND(TO_NUMBER(DISPERSION),3) DISPERSION
FROM   DM$VDKM_SH_CLUS1
ORDER BY CLUSTER_ID;


CLU_ID   REC_CNT   PARENT   TREE_LEVEL   DISPERSION
       1      4500                     1        6.731

---------------------------
```

5. To see the leaf cluster IDs, run the following query:

Oracle supports hierarchical clustering. In hierarchical clustering, the data points having similar characteristics are grouped together. The cluster hierarchy is represented as a tree structure. The leaf clusters are the final clusters generated by the algorithm. Clusters higher up in the hierarchy are intermediate clusters.

```
SELECT CLUSTER_ID
FROM   DM$VDKM_SH_CLUS1
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL
ORDER BY CLUSTER_ID;


CLUSTER_ID
           1
---------------------------
```

Examine the query:

`LEFT_CHILD_ID IS NULL`: Outputs the leaf nodes on the left of the hierarchical tree

`RIGHT_CHILD_ID IS NULL`: Outputs the leaf nodes on the right of the hierarchical tree

6. View the dispersion details or the cluster description for the leaf cluster IDs:

Dispersion is a measure of cluster quality and computationally it is the sum of squared error. This also indicates the quality of the cluster model.

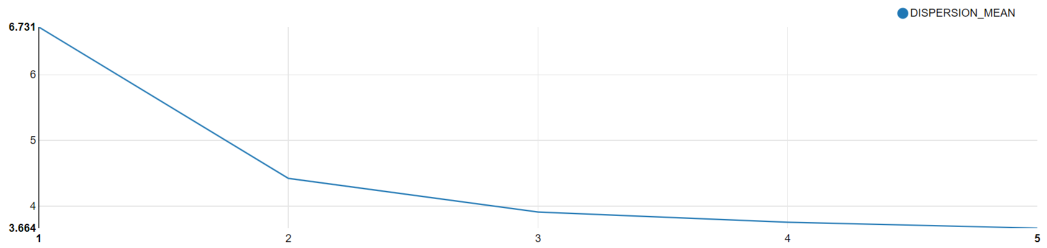
```
%script
SELECT CLUSTER_ID CLU_ID, RECORD_COUNT REC_CNT, PARENT,
       TREE_LEVEL, ROUND(TO_NUMBER(DISPERSION),3) DISPERSION
FROM   DM$VDKM_SH_CLUS1
WHERE CLUSTER_ID IN (SELECT CLUSTER_ID
                     FROM   DM$VDKM_SH_CLUS1
                     WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL)
ORDER BY CLUSTER_ID;
```

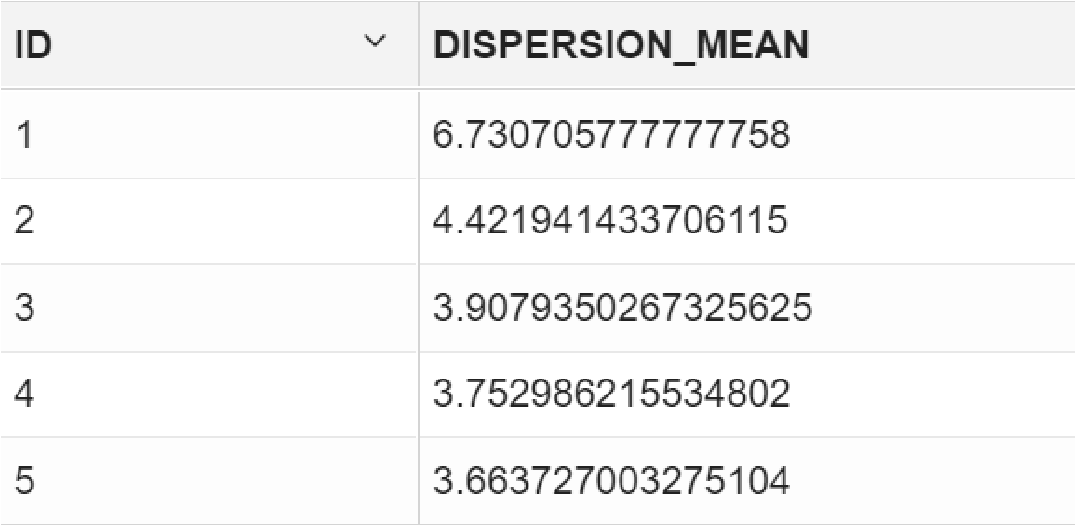
```


CLU_ID   REC_CNT   PARENT   TREE_LEVEL   DISPERSION
       1      4500                     1        6.731

---------------------------
```

7. To determine the optimal value of K (or the number of clusters) for the data, visualize the data with an Elbow method.

   The Elbow method is done with the leaf clusters. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the variance (or dispersion) as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

```
%sql
SELECT 1 ID, AVG(DISPERSION) DISPERSION_MEAN
FROM   DM$VDKM_SH_CLUS1
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL
UNION
SELECT 2 ID, AVG(DISPERSION) DISPERSION_MEAN
FROM   DM$VDKM_SH_CLUS2
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL
UNION
SELECT 3 ID, AVG(DISPERSION) DISPERSION_MEAN
FROM   DM$VDKM_SH_CLUS3
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL
UNION
SELECT 4 ID, AVG(DISPERSION) DISPERSION_MEAN
FROM   DM$VDKM_SH_CLUS4
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL
UNION
SELECT 5 ID, AVG(DISPERSION) DISPERSION_MEAN
FROM   DM$VDKM_SH_CLUS5
WHERE LEFT_CHILD_ID IS NULL AND RIGHT_CHILD_ID IS NULL;
```





ID	DISPERSION_MEAN
1	6.73070577777758
2	4.421941433706115
3	3.9079350267325625
4	3.752986215534802
5	3.663727003275104

From the resultant graph, the curve flattens after 3 or the dispersion value flattens after ID 3, which means that the optimal value of K (or the most suitable number of clusters that the data must be segmented into) is 3.

> **Note:**
>
> In Oracle SQL Developer, a visual aid to view the graph is not applicable. You can only compute the dispersion scores.

8. To view the Attribute details of the `KM_SH_CLUS3` model, run the following statement:

   The Attribute Details view displays statistics like mean, median, and mode of your model.

```
%script
SELECT CLUSTER_ID, ATTRIBUTE_NAME, ATTRIBUTE_SUBNAME, MEAN, VARIANCE, MODE_VALUE
FROM  DM$VAKM_SH_CLUS3;


CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME
MEAN                 VARIANCE                MODE_VALUE
           1 CUST_CREDIT_LIMIT
7924.22222222223     15914238.670321768
           1 CUST_YEAR_OF_BIRTH
1964.6244444444444   187.1267639722414
           1 YRS_RESIDENCE
4.021999999999995    3.617430984663253
           1 Y_BOX_GAMES
0.31244444444444447  0.2148706626163839
           1
CUST_GENDER
```

```
          M
           1
CUST_INCOME_LEVEL
          J: 190,000 - 249,999
           1
CUST_MARITAL_STATUS
          Married
           1
HOUSEHOLD_SIZE
          3
           2 CUST_CREDIT_LIMIT
7833.002645502645     15543554.858080933
           2 CUST_YEAR_OF_BIRTH
1957.631283068783     121.54941469457282
           2 YRS_RESIDENCE
4.8611111111111045    2.7838791487484835
           2 Y_BOX_GAMES
0.0                   0.0
           2
CUST_GENDER
          M
           2
CUST_INCOME_LEVEL
          J: 190,000 - 249,999

CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME
MEAN                 VARIANCE              MODE_VALUE
           2
CUST_MARITAL_STATUS
          Married
           2
HOUSEHOLD_SIZE
          3
           3 CUST_CREDIT_LIMIT
8111.111111111114    16632730.696798513
           3 CUST_YEAR_OF_BIRTH
1978.9518970189702   15.976667585319932
           3 YRS_RESIDENCE
2.3028455284552827   0.9272054568003305
           3 Y_BOX_GAMES
0.9525745257452575   0.04520692664553768
           3
CUST_GENDER
          M
           3
CUST_INCOME_LEVEL
          J: 190,000 - 249,999
           3
CUST_MARITAL_STATUS
          NeverM
           3
HOUSEHOLD_SIZE
          1
           4 CUST_CREDIT_LIMIT
3126.6094420600857   2978559.2320826976
           4 CUST_YEAR_OF_BIRTH
```

```
1978.4978540772531   22.143006137800537
           4 YRS_RESIDENCE
2.270386266094421    0.8944759795099003
           4 Y_BOX_GAMES
0.8819742489270386   0.10431953481932726

CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME
MEAN                 VARIANCE              MODE_VALUE
           4
CUST_GENDER
          F
           4
CUST_INCOME_LEVEL
          B: 30,000 - 49,999
           4
CUST_MARITAL_STATUS
          NeverM
           4
HOUSEHOLD_SIZE
          1
           5 CUST_CREDIT_LIMIT
10410.891089108914   6172923.883072166
           5 CUST_YEAR_OF_BIRTH
1979.1613861386138   13.011589751641117
           5 YRS_RESIDENCE
2.3178217821782146   0.9424967372852242
           5 Y_BOX_GAMES
0.9851485148514851   0.01464541895220246
           5
CUST_GENDER
          M
           5
CUST_INCOME_LEVEL
          J: 190,000 - 249,999
           5
CUST_MARITAL_STATUS
          NeverM
           5
HOUSEHOLD_SIZE
          1


40 rows selected.
```

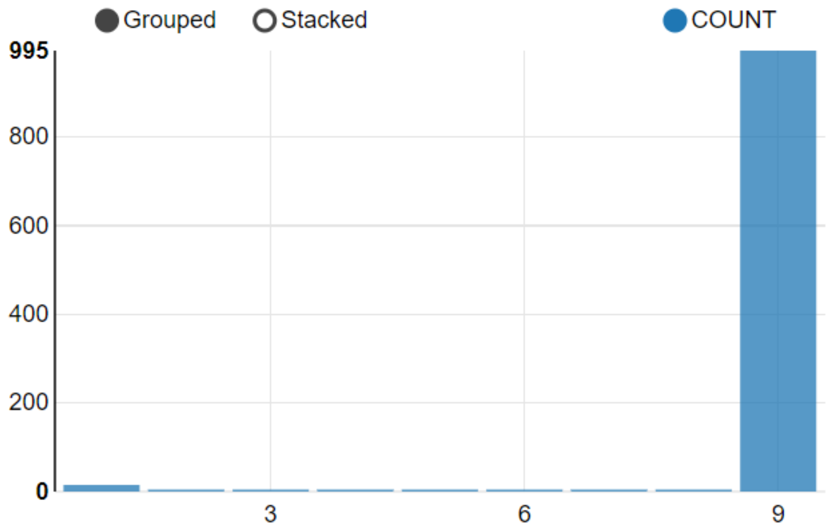
Notice that Cluster ID 5 has the highest mean for `Y_BOX_GAMES` users and has the highest `CUST_CREDIT_LIMIT`.

9. Now, for the model `KM_SH_CLUS3`, view the histogram details with specific attributes for each leaf cluster. For this use-case, view the histogram details for `Y_BOX_GAMES` and `CUST_INCOME_LEVEL` attributes. In this step, leaf cluster ID 5 and the attribute `Y_BOX_GAMES` are picked.

```
%sql

SELECT CLUSTER_ID, ATTRIBUTE_NAME, ATTRIBUTE_SUBNAME,
       BIN_ID, LOWER_BIN_BOUNDARY, UPPER_BIN_BOUNDARY, ATTRIBUTE_VALUE, COUNT
FROM DM$VHKM_SH_CLUS3
```

```
WHERE CLUSTER_ID = 5 AND ATTRIBUTE_NAME = 'Y_BOX_GAMES'
ORDER BY BIN_ID;
```

In OML Notebooks, click the bar plot icon and expand settings. Drag `BIN_ID` to **keys** and **COUNT** to values.

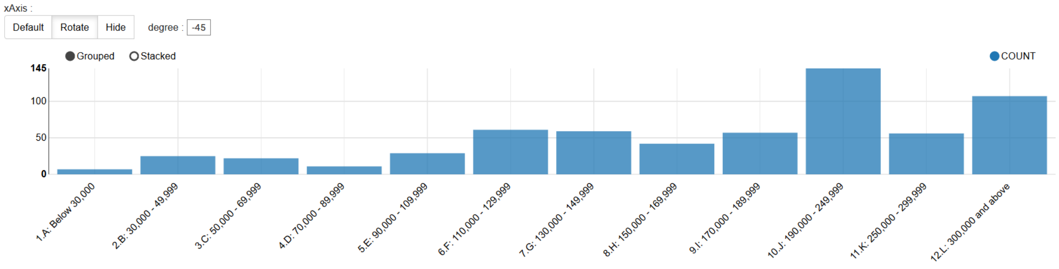


From this histogram, you can see that Cluster ID 5 is grouped into bins showing the count of `Y_BOX_GAMES` users. Bin 9 has the highest count of `Y_BOX_GAMES` users.

10. Similarly, for Cluster ID 5, view the histogram details for the `CUST_INCOME_LEVEL` attribute.

```
%sql

SELECT CLUSTER_ID, ATTRIBUTE_NAME, ATTRIBUTE_SUBNAME,
       BIN_ID, LOWER_BIN_BOUNDARY, UPPER_BIN_BOUNDARY, ATTRIBUTE_VALUE, COUNT
FROM DM$VHKM_SH_CLUS3
WHERE CLUSTER_ID = 5 AND ATTRIBUTE_NAME = 'CUST_INCOME_LEVEL'
ORDER BY BIN_ID;
```

In OML Notebooks, click the bar plot icon and expand settings. Drag `BIN_ID` and `ATTRIBUTE_VALUE` to **keys** and **COUNT** to values. In the xAxis options, click **Rotate**.


xAxis : Default Rotate Hide degree : -45

In this histogram, Cluster ID 5 is grouped into bins showing the count of customers with `CUST_INCOME_LEVEL` and indicates that the highest number of customers draw a salary package between 190,000 - 249,999 yearly.

11. Now, view the Rule details of leaf clusters (2, 4, and 5) to check the support and confidence level.

Support and confidence are metrics that describe the relationships between clustering rules and cases. Support is the percentage of cases for which the rule holds. Confidence is the probability that a case described by this rule is actually assigned to the cluster.

```
%script

SELECT CLUSTER_ID, ATTRIBUTE_NAME, ATTRIBUTE_SUBNAME, OPERATOR,
       NUMERIC_VALUE, ATTRIBUTE_VALUE, SUPPORT, ROUND(CONFIDENCE,3) CONFIDENCE
FROM DM$VRKM_SH_CLUS3
WHERE cluster_id IN (SELECT cluster_id
                     FROM DM$VDKM_SH_CLUS3
                     WHERE LEFT_CHILD_ID is NULL and RIGHT_CHILD_ID is NULL)
ORDER BY CLUSTER_ID, ATTRIBUTE_NAME, ATTRIBUTE_SUBNAME, OPERATOR, NUMERIC_VALUE,
ATTRIBUTE_VALUE;
```

```
CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE   ATTRIBUTE_VALUE        SUPPORT   CONFIDENCE
           2 CUST_CREDIT_LIMIT                         <=
15000.0                                     3024           0
           2 CUST_CREDIT_LIMIT                         >=
1500.0                                      3024           0
           2 CUST_GENDER
IN                         F                           3024        0.002
           2 CUST_GENDER
IN                         M                           3024        0.002
           2 CUST_INCOME_LEVEL
IN                         B: 30,000 - 49,999          2750            0
           2 CUST_INCOME_LEVEL
IN                         E: 90,000 - 109,999         2750            0
           2 CUST_INCOME_LEVEL
IN                         F: 110,000 - 129,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         G: 130,000 - 149,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         H: 150,000 - 169,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         I: 170,000 - 189,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         J: 190,000 - 249,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         K: 250,000 - 299,999        2750            0
           2 CUST_INCOME_LEVEL
IN                         L: 300,000 and above        2750            0
           2 CUST_MARITAL_STATUS
IN                         Divorc.                     2720        0.014

CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE        ATTRIBUTE_VALUE   SUPPORT   CONFIDENCE
           2 CUST_MARITAL_STATUS
IN                              Married                2720        0.014
           2 CUST_MARITAL_STATUS
IN                              NeverM                 2720        0.014
           2 CUST_YEAR_OF_BIRTH                        <=
1977.888888888889                           2854        0.041
           2 CUST_YEAR_OF_BIRTH                        >
1937.3333333333333                          2854        0.041
           2 HOUSEHOLD_SIZE
IN                              2                      2699        0.016
           2 HOUSEHOLD_SIZE
IN                              3                      2699        0.016
```

```
           2 HOUSEHOLD_SIZE
IN                              9+                     2699        0.016
           2 YRS_RESIDENCE                             <=
7.777777777777778                           2804        0.019
           2 YRS_RESIDENCE                             >
1.5555555555555556                          2804        0.019
           2 Y_BOX_GAMES                               <=
0.1111111111111111                          3024        0.056
           2 Y_BOX_GAMES                               >=
0.0                                         3024        0.056
           4 CUST_CREDIT_LIMIT                         <=
7500.0                                       466        0.128
           4 CUST_CREDIT_LIMIT                         >=
1500.0                                       466        0.128
           4 CUST_GENDER
IN                              F                       466        0.023

CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE        ATTRIBUTE_VALUE        SUPPORT   CONFIDENCE
           4 CUST_GENDER
IN                              M                           466
0.023
           4 CUST_INCOME_LEVEL
IN                              A: Below 30,000             466
0.079
           4 CUST_INCOME_LEVEL
IN                              B: 30,000 - 49,999          466
0.079
           4 CUST_INCOME_LEVEL
IN                              C: 50,000 - 69,999          466
0.079
           4 CUST_INCOME_LEVEL
IN                              D: 70,000 - 89,999          466
0.079
           4 CUST_INCOME_LEVEL
IN                              E: 90,000 - 109,999         466
0.079
           4 CUST_INCOME_LEVEL
IN                              F: 110,000 - 129,999        466
0.079
           4 CUST_INCOME_LEVEL
IN                              G: 130,000 - 149,999        466
0.079
           4 CUST_INCOME_LEVEL
IN                              H: 150,000 - 169,999        466
0.079
           4 CUST_INCOME_LEVEL
IN                              I: 170,000 - 189,999        466
0.079
           4 CUST_MARITAL_STATUS
IN                              Married                     413
0.043
           4 CUST_MARITAL_STATUS
IN                              NeverM                      413
0.043
           4 CUST_YEAR_OF_BIRTH                        <=
1986.0                                            451        0.103
           4 CUST_YEAR_OF_BIRTH                        >
1969.7777777777778                                451        0.103

CLUSTER_ID   ATTRIBUTE_NAME      ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE       ATTRIBUTE_VALUE        SUPPORT   CONFIDENCE
           4 HOUSEHOLD_SIZE
IN                             1                            418
0.043
           4 HOUSEHOLD_SIZE
IN                             2                            418
0.043
           4 HOUSEHOLD_SIZE
IN                             3                            418
0.043
           4 HOUSEHOLD_SIZE
IN                             9+                           418
0.043
           4 YRS_RESIDENCE                               <=
4.666666666666667                                464        0.086
           4 YRS_RESIDENCE                               >=
0.0                                              464        0.086
           4 Y_BOX_GAMES                                 <=
1.0                                              466        0.083
           4 Y_BOX_GAMES                                 >=
0.0                                              466        0.083
           5 CUST_CREDIT_LIMIT                           <=
15000.0                                         1010        0.056
           5 CUST_CREDIT_LIMIT                           >
6000.0                                          1010        0.056
           5 CUST_GENDER
IN                             F                           1010
0.002
           5 CUST_GENDER
IN                             M                           1010
0.002
           5 CUST_INCOME_LEVEL
IN                             F: 110,000 - 129,999         906
0.024
           5 CUST_INCOME_LEVEL
IN                             G: 130,000 - 149,999         906
0.024

CLUSTER_ID   ATTRIBUTE_NAME        ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE        ATTRIBUTE_VALUE        SUPPORT   CONFIDENCE
           5 CUST_INCOME_LEVEL
IN                              I: 170,000 - 189,999         906
0.024
           5 CUST_INCOME_LEVEL
IN                              J: 190,000 - 249,999         906
0.024
           5 CUST_INCOME_LEVEL
IN                              K: 250,000 - 299,999         906
0.024
           5 CUST_INCOME_LEVEL
IN                              L: 300,000 and above         906
```

```
0.024
           5 CUST_MARITAL_STATUS
IN                              Married                      944
0.046
           5 CUST_MARITAL_STATUS
IN                              NeverM                       944
0.046
           5 CUST_YEAR_OF_BIRTH                                    <=
1986.0                                           1003         0.12
           5 CUST_YEAR_OF_BIRTH                                    >
1969.7777777777778                               1003         0.12
           5 HOUSEHOLD_SIZE
IN                              1                            859
0.036
           5 HOUSEHOLD_SIZE
IN                              2                            859
0.036
           5 HOUSEHOLD_SIZE
IN                              3                            859
0.036
           5 YRS_RESIDENCE                                         <=
4.666666666666667                                 993        0.079
           5 YRS_RESIDENCE                                         >=
0.0                                               993        0.079
           5 Y_BOX_GAMES                                           <=
1.0                                               995        0.136

CLUSTER_ID   ATTRIBUTE_NAME   ATTRIBUTE_SUBNAME   OPERATOR
NUMERIC_VALUE        ATTRIBUTE_VALUE   SUPPORT   CONFIDENCE
           5 Y_BOX_GAMES                                  >
0.8888888888888888                           995        0.136


71 rows selected.

---------------------------
```

12. To view the size of each cluster, run the following statement:

    In OML Notebooks, you can also click the bar icon or the pie chart icon to view the bar graph or the pie chart.

```
%sql
SELECT CLUSTER_ID(KM_SH_CLUS3 USING *) AS CLUS, COUNT(*) AS CNT
FROM CUSTOMERDATA
GROUP BY CLUSTER_ID(KM_SH_CLUS3 USING *)
ORDER BY CNT DESC;
```

```
CLUS   CNT
     2   3024
     5   1010
     4    466
---------------------------
```

Score

Scoring involves applying the model to the target data. Use CLUSTER_PROBABILITY function to predict the clusters. For Clustering, "scoring" involves assigning each record to a cluster, with a certain probability. However, one can also obtain the probability of a record belonging to each cluster.

1. In the following step, you are scoring the probability of the top 10 customers that belong to cluster 5.

```
%script

SELECT CUST_ID,
       ROUND(CLUSTER_PROBABILITY(KM_SH_CLUS3, 5 USING *),3)
       PROB
FROM CUSTOMERDATA
WHERE rownum < 10
ORDER BY PROB DESC;



CUST_ID    PROB
   102308   0.539
   101232   0.502
   101610   0.374
   102828   0.303
   100134   0.302
   103948   0.297
   100696    0.25
   103791   0.141
   100804   0.104


9 rows selected.

---------------------------
```

2. To score the cluster ID of a given CUST_ID (customer), for this use case, you must target customers who have already purchased Y_BOX_GAMES and with high credit limit, to sell the new game product. In the previous stage, you have identified that cluster 5 has highest customers who have already purchased Y_BOX_GAMES with mean CUST_CREDIT_LIMIT of 10410. So, the target group is cluster ID 5. To score for a given CUST_ID (102308) and display the probability score, run the following query :

```
%sql
SELECT CLUSTER_ID(KM_SH_CLUS3 USING *) AS CLUSTER_ID, round (CLUSTER_PROBABILITY
(KM_SH_CLUS3 USING *),3) AS PROB
  FROM CUSTOMERDATA
where cust_id = 102308;



CLUSTER_ID    PROB
            5   0.539
---------------------------
```

Examine the query:

- CLUSTER_ID(KM_SH_CLUS3 USING *) AS CLUSTER_ID: Provides CLUSTER_ID from the KM_SH_CLUS3 model.
- round(CLUSTER_PROBABILITY(KM_SH_CLUS3 USING *),2) AS PROB: Provides cluster probability using KM_SH_CLUS3 model. ROUND (n,integer) returns results of CLUSTER_PROBABILITY rounded to n integer places to the right. Here, it is four places.

3. Additionally, you can obtain the probability of a record belonging to each cluster (such as 5, 3, 2) by running the following query:

```
%script
select CLUSTER_PROBABILITY(KM_SH_CLUS3,
      5 USING *) from CUSTOMERDATA;



CLUSTER_PROBABILITY(KM_SH_CLUS3,5USING*)
0.30701266050607
0.3064062868515786
0.2862730847381108
0.2868527181838429
0.3721982825972361
0.2816026555211009
0.30936576857241027
0.3051489029060863
0.1915573544647028
0.25158448263351973
0.37204422449011026
0.3064062868515786
0.35693390244389295
0.1902596096427133
...
```

To conclude, you have successfully segmented the population into different clusters and determined that cluster 5 has the target population for the use case. You can safely target customers in cluster 5 to sell a new game product. You can select the customer IDs from Step 1. You can also display a full list of target customers by removing the WHERE clause.

Time Series Use Case Scenario

You work in an electronic store, and sales of laptops and tablets have increased over the last two quarters. You want to forecast your product sales for the next four quarters using historical timestamped data. You forecast sales using the Exponential Smoothing algorithm, predicting changes over evenly spaced intervals of time using historical data.

Related Content

Topic	Link
OML4SQL GitHub Example	Time Series - Exponential Smoothing
CREATE_MODEL2 Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING - Model Settings
Exponential Smoothing Model (ESM) Settings	DBMS_DATA_MINING — Algorithm Settings:Exponential Smoothing
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views

Topic	Link
Exponential Smoothing Model - Model Detail Views	Model Detail Views for Exponential Smoothing
About Time Series	About Time Series

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data Set
  The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.

  You will use the `SALES` table from the `SH` schema. You can access the table by running the `SELECT` statements in OML Notebooks.
- Database
  Select or create database out of the following options:
  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements
  Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Related Topics

- Create a Notebook
- Edit your Notebook
- Uninstalling HR Schema

Load Data

Access the data set from the `SH` schema and explore the data to understand the attributes.

> **Remember:**
>
> The data set used for this use case is from the SH schema. The SH schema can be readily accessed in Oracle Autonomous Database. For on-premises databases, the schema is installed during the installation or can be manually installed by downloading the scripts. See Installing the Sample Schemas.

To understand the data, you will perform the following:

- Access the data.
- Examine the various attributes or columns of the data set.
- Assess data quality (by exploring the data).

Access Data

You will use SALES table data from the SH schema.

Examine Data

The following table displays information about the attributes from SALES:

Attribute Name	Information
PROD_ID	The ID of the product
CUST_ID	The ID of the customer
TIME_ID	The timestamp of the purchase of the product in yyy-mm-dd hh:mm:ss format
CHANNEL_ID	The channel ID of the channel sales data
PROMO_ID	The product promotion ID
QUANTITY_SOLD	The number of items sold
AMOUNT_SOLD	The amount or sales data

Identify Target Variable

In this use case, the task is to train a model that predicts the amount sold. Therefore, the target variable is the attribute AMOUNT_SOLD.

Explore Data

Explore the data to understand and assess the quality of the data. At this stage assess the data to identify data types and noise in the data. Look for missing values and numeric outlier values.

If you are working with Oracle Autonomous Database, you can use the Oracle Machine Learning (OML) Notebooks for your data science project, including assessing data quality. If you are using an on-premise Oracle Database, you can use the Oracle SQL Developer to assess data quality. Query the SH schema as described.

> **Note:**
>
> Each record in the database is called a case and each case is identified by a case_id. In this use case TIME_ID is the case_id as it is an independent variable and you are forecasting the sales for evenly spaced time.

The following steps help you with exploratory analysis of the data.

1. View the data in the SH.SALES table by running the following statement:

```
SELECT * FROM SH.SALES;
```

2. To find the number of rows in `SH.SALES` table, run the following statement:

   ```
   %script
   SELECT COUNT(*) from SH.SALES;
   ```

   ```
   COUNT(*)
       918843
   ---------------------------
   ```

3. Find the distinct users in the table, run the following query:

   ```
   %sql SELECT COUNT (DISTINCT CUST_ID) FROM SH.SALES;
   ```

   ```
   COUNT(DISTINCTCUST_ID)
                       7059
   ---------------------------
   ```

4. To view the datatype of the sales table, run the following query:

   ```
   %script
   DESCRIBE SH.SALES;
   ```

   ```
   Name              Null?        Type
   -------------     --------     ------------
   PROD_ID     NOT NULL     NUMBER
   CUST_ID     NOT NULL     NUMBER
   TIME_ID      NOT NULL     DATE
   CHANNEL_ID     NOT NULL     NUMBER
   PROMO_ID     NOT NULL     NUMBER
   QUANTITY_SOLD     NOT NULL     NUMBER(10,2)
   AMOUNT_SOLD     NOT NULL     NUMBER(10,2)

   ---------------------------
   ```

5. To view all the NULLs and missing values, run the following query:

   ```
   %sql SELECT COUNT(*) FROM SH.SALES WHERE PROD_ID=NULL OR CUST_ID=NULL OR
       TIME_ID=NULL OR CHANNEL_ID=NULL OR PROMO_ID=NULL OR QUANTITY_SOLD=NULL
   OR
       AMOUNT_SOLD=NULL;
   ```

   ```
   COUNT(*)
            0
   ---------------------------
   ```

   NULLs, if found, are automatically handled by the OML algorithms.

6. Now, prepare a view called `ESM_SH_DATA` by selecting the necessary columns from `SH.SALES` table. For this use case, select `TIME_ID` and `AMOUNT_SOLD`.

```
%script
CREATE OR REPLACE VIEW ESM_SH_DATA AS 
  SELECT TIME_ID, AMOUNT_SOLD FROM SH.SALES;



View ESM_SH_DATA created.
---------------------------
```

7. Count the number of rows to ensure that we have the same amount of data. Run the following query:

```
%script
SELECT count(*) from ESM_SH_DATA;



COUNT(*)
    918843
---------------------------
```

This completes the data understanding and data exploration stage. Time series data can contain missing values. The setting `EXSM_SETMISSING` can be used to specify how to handle missing values. The special value `EXSM_MISS_AUTO` indicates that, if the series contains missing values it is to be treated as an irregular time series. The Automatic Data Preparation (ADP) setting does not impact this data for time series. See How ADP Transforms the Data to understand how ADP prepares the data for some algorithms.

Build Model

To build a model using the time series data, you will use Exponential Smoothing algorithm on the `ESM_SH_DATA` view that is generated during the exploratory stage.

Oracle offers the Exponential Smoothing algorithm for time series.
Exponential smoothing is a forecasting method for time series data. It is a moving average method where exponentially decreasing weights are assigned to past observations. Components of Exponential Smoothing Model (ESM) such as trend and seasonality extensions, can have an additive or multiplicative form. For additive forms, the amplitude of the variation is independent of the level, whereas for multiplicative forms, the variation is connected to the level. The simpler additive models assume that error or noise, trend, and seasonality are linear effects within the recursive formulation.

To build a model using a supervised learning algorithm you may use a subset of the data into training and test data. Time series models usually use historical data to predict the future. This is different from model validation for classification and regression, which normally involves splitting data randomly into training and test sets. In this use case, there is no need to split the data set because the model is always predicting the current value based on information from the past. This means that although it seems that you train and test on the same data set, but when the model is applied, the forecast is always based on the previous date. In this use case, you will use the `ESM_SH_DATA` view.

1. To see the data in the `ESM_SH_DATA` view, run the following statement:

   ```
   %sql
   SELECT * from ESM_SH_DATA;


   TIME_ID     AMOUNT_SOLD
   20-JAN-98         1205.99
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-APR-98         1250.25
   05-JUL-98         1210.21
   05-JUL-98         1210.21
   05-JUL-98         1210.21
   05-JUL-98         1210.21
   05-JUL-98         1210.21 ...
   ```

2. Build a model with the `ESM_SH_DATA` table, run the following script:

   ```
    %script

       BEGIN DBMS_DATA_MINING.DROP_MODEL('ESM_SALES_FORECAST_1');
       EXCEPTION WHEN OTHERS THEN NULL; END;
       /
       DECLARE
           v_setlist DBMS_DATA_MINING.SETTING_LIST;
       BEGIN
           v_setlist('ALGO_NAME')            := 'ALGO_EXPONENTIAL_SMOOTHING';
           V_setlist('EXSM_INTERVAL')        := 'EXSM_INTERVAL_QTR';
           V_setlist('EXSM_PREDICTION_STEP') := '4';
           V_setlist('EXSM_MODEL')           := 'EXSM_WINTERS';
           V_setlist('EXSM_SEASONALITY')     := '4';
        V_setlist('EXSM_SETMISSING')    := 'EXSM_MISS_AUTO';

           DBMS_DATA_MINING.CREATE_MODEL2(
               MODEL_NAME          => 'ESM_SALES_FORECAST_1',
               MINING_FUNCTION     => 'TIME_SERIES',
               DATA_QUERY          => 'select * from ESM_SH_DATA',
               SET_LIST            => v_setlist,
               CASE_ID_COLUMN_NAME => 'TIME_ID',
               TARGET_COLUMN_NAME  =>'AMOUNT_SOLD');
       END;
       /


   PL/SQL procedure successfully completed.
       ---------------------------
       PL/SQL procedure successfully completed.
   ```

   Examine the script:

- v_setlist is a variable to store SETTING_LIST.
- SETTING_LIST specifies model settings or hyperparameters for the model.
- DBMS_DATA_MINING is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- ALGO_NAME specifies the algorithm name. Since you are using Exponential Smoothing as the algorithm, the value of the setting is ALGO_EXPONENTIAL_SMOOTHING.
- EXSM_INTERVAL indicates the interval of the data set or a unit of interval size. For example, day, week, month, and so on. You want to predict for quarterly sales. Hence, the setting is EXSM_INTERVAL_QTR. This setting applies only to the time column with datetime type.
- EXSM_PREDICTION_STEP specifies how many predictions to make. You want to display each value representing a quarter. Hence, a value of 4 gives four values ahead prediction.
- EXSM_MODEL specifies the type of exponential smoothing model to be used. Here the value is EXSM_HW. The Holt-Winters triple exponential smoothing model with additive trend and multiplicative seasonality is applied. This type of model considers various combinations of additive and multiplicative trend, seasonality and error, with and without trend damping. Other options are EXSM_SIMPLE, EXSM_SIMPLE_MULT, EXSM_HOLT, EXSM_HOLT_DMP, EXSM_MUL_TRND, EXSM_MULTRD_DMP, EXSM_SEAS_ADD, EXSM_SEAS_MUL, EXSM_HW, EXSM_HW_DMP, EXSM_HW_ADDSEA, EXSM_DHW_ADDSEA, EXSM_HWMT, EXSM_HWMT_DMP.
- EXSM_SEASONALITY indicates how long a season lasts. The parameter specifies a positive integer value as the length of seasonal cycle. The value it takes must be larger than 1. For example, 4 means that every group of four values forms a seasonal cycle.
- EXSM_SETMISSING specifies how to handle missing values. In time series, the special value EXSM_MISS_AUTO indicates that, if the series contains missing values it is to be treated as an irregular time series.

The CREATE_MODEL2 procedure has the following settings:

- MODEL_NAME: A unique name that you will give to the model. Name of the model in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is ESM_SALES_FORECAST_1.
- MINING_FUNCTION: Specifies the machine learning function. Since it is a time series problem, select TIME_SERIES.
- DATA_QUERY: A query that provides training data for building the model. Here, the query is SELECT * FROM ESM_SH_DATA.
- SET_LIST: Specifies SETTING_LIST.
- CASE_ID_COLUMN_NAME: A unique case identifier column in the training data. In this use case, case_id is TIME_ID. If there is a composite key, you must create a new attribute before creating the model.
- TARGET_COLUMN_NAME: Specifies the column that is to be predicted. Also referred to as the target variable of the model. In other words, the value the model predicts. In this use case, you are predicting the sale of products in terms of their dollar price. Therefore, in this use case, the TARGET_COLUMN_NAME is AMOUNT_SOLD.

> **Note:**
>
> Any parameters or settings not specified are either system-determined or default values are used.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

Sometimes querying dictionary views and model detail views is sufficient to measure your model's performance. However, you can evaluate your model by computing test metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), confusion matrix, lift statistics, cost matrix, and so on. For Association Rules, you can inspect various rules to see if they reveal new insights for item dependencies (antecedent itemset implying consequent) or for unexpected relationships among items.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

By examining various statistics in the model detail views, you can compare models to arrive at one model that satisfies your evaluation criteria.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views.

1. You can review the model settings by running the following query:

```
 %sql

SELECT SETTING_NAME, SETTING_VALUE
  FROM USER_MINING_MODEL_SETTINGS
  WHERE MODEL_NAME = UPPER('ESM_SALES_FORECAST_1')
  ORDER BY SETTING_NAME;
```

```
SETTING_NAME                   SETTING_VALUE
ALGO_NAME                      ALGO_EXPONENTIAL_SMOOTHING
EXSM_ACCUMULATE                EXSM_ACCU_TOTAL
EXSM_CONFIDENCE_LEVEL          .95
EXSM_INTERVAL                  EXSM_INTERVAL_QTR
EXSM_MODEL                     EXSM_WINTERS
EXSM_NMSE                      3
EXSM_OPTIMIZATION_CRIT         EXSM_OPT_CRIT_LIK
EXSM_PREDICTION_STEP           4
EXSM_SEASONALITY               4
EXSM_SETMISSING                EXSM_MISS_AUTO
ODMS_DETAILS                   ODMS_ENABLE
ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE
PREP_AUTO                      ON

14 rows selected.
---------------------------
```

2. To view the `DM$VP` model view, run the following statement:

   The `DM$VP` view for time series contains the result of an ESM model. The output has a set of records such as partition, `CASE_ID`, value, prediction, lower, upper, and so on and ordered by partition and `CASE_ID` (time).

```
%script
SELECT CASE_ID, VALUE, PREDICTION, LOWER, UPPER FROM
DM$VPESM_SALES_FORECAST_1
ORDER BY CASE_ID;
```

```
CASE_ID     VALUE                PREDICTION           LOWER   UPPER
01-JAN-98     6480684.0000011446   6452375.7547333492
01-APR-98     5593994.1400007578   5848724.7899219571
01-JUL-98     6071823.1000010688   6214546.3092128271
01-OCT-98     5937413.7100012964   5869219.4189072186
01-JAN-99      6093747.209999715    6132016.410793812
01-APR-99     4925471.6299999086   5385954.0785653945
01-JUL-99     5827050.1500000218   5350240.2540956484
01-OCT-99     5373678.6700002998   5304626.0456054937
01-JAN-00     5984889.4899995513   5541123.2442497462
01-APR-00     5371730.9200002486     5236126.09628068
01-JUL-00     6121239.2899996703   5955258.7436284116
01-OCT-00     6287646.9199997969   6089446.4024073323
01-JAN-01     6547097.4400001625   6837567.1739504253
01-APR-01     6922468.3900004178   6188944.0536819538                ...

---------------------------------------------------------------------------
-----------
```

Examine the statement:

- CASE_ID: Specifies the timestamp.
- VALUE: Specifies the AMOUNT_SOLD.
- PREDICTION: Indicates the predicted value for the model.
- LOWER and UPPER: Indicate the confidence bounds.

3. To view the model diagonistic view, DM$VG, and evaluate the model, run the following query:

   The DM$VG view for time series contains the global information of the model along with the estimated smoothing constants, the estimated initial state, and global diagnostic measures.

```
%sql
SELECT NAME, round(NUMERIC_VALUE,4), STRING_VALUE
  FROM DM$VGESM_SALES_FORECAST_1
  ORDER BY NAME;



NAME                ROUND(NUMERIC_VALUE,4)   STRING_VALUE
-2 LOG-LIKELIHOOD                   450.7508
AIC                                 466.7508
AICC                                487.3223
ALPHA                                 0.4525
AMSE                       157764777942.4555
BETA                                  0.4195
BIC                                 472.9315
CONVERGED                                    YES
GAMMA                                 0.0001
INITIAL LEVEL                   6110212.8741
INITIAL SEASON 1                      0.9939
INITIAL SEASON 2                      1.0231
INITIAL SEASON 3                      0.9366
INITIAL SEASON 4                      1.0465

NAME            ROUND(NUMERIC_VALUE,4)   STRING_VALUE
INITIAL TREND                 55478.0794
MAE                               0.0424
MSE                    104400146583.6485
NUM_ROWS                          918843
SIGMA                              0.054
STD                          323110.1153


20 rows selected.

---------------------------
```

- NAME: Indicates the diagnostic attribute name.
- NUMERIC_VALUE: Indicates the calculated statistical value for the model.
- STRING_VALUE: Indicates alphanumeric values for the diagnostic parameter.
  A few parameters to note for an exponential smoothing algorithm are:
  - ALPHA: Indicates the smoothing constant.

- BETA: Indicates the trend smoothing constant.
- GAMMA: Indicates the seasonal smoothing constant.
- MAE: Indicates Mean Absolute Error.
- MSE: Indicates Mean Square Error.

Exponential smoothing assumes that a series extends infinitely into the past, but that influence of past on future, decays smoothly and exponentially fast. The smooth rate of decay is expressed by one or more smoothing constants. The *smoothing constants* are parameters that the model estimates. These smoothing constants are represented as α, β, and γ. Values of a smoothing constant near one put almost all weight on the most recent observations. Values of a smoothing constant near zero allow the distant past observations to have a large influence.

Note that α is associated with the error or noise of the series, β is associated with the trend, and γ is associated with the seasonality factors. The γ value is closest to zero which means seasonality has an influence on the data set.

The MAE and MSE values are low which means that the model is good. The MSE magnitude depends on the actual scale of your original data. In this case, the STD is around 10^5. The square of it is roughly in the scale of 10^{10}. The error percentage is low and hence, the model is good.

Score

You are ready to forecast sales for the next four quarters.

For a time series model, you can use the DM$VP view to perform scoring or prediction.

1. Query the DM$VP model detail view to see the forecast (sales for four quarters). Run the following statement:

```
%sql
SELECT TO_CHAR(CASE_ID,'YYYY-MON') DATE_ID,
       round(VALUE,2) ACTUAL_SOLD,
       round(PREDICTION,2) FORECAST_SOLD,
       round(LOWER,2) LOWER_BOUND, round(UPPER,2) UPPER_BOUND
  FROM DM$VPESM_SALES_FORECAST_1
  ORDER BY CASE_ID;
```

   In this step, the prediction shows amount sold along with the case_id. The predictions display upper and lower confidence bounds showing that the estimates can vary between those values.

   Examine the statement:

   - TO_CHAR(CASE_ID,'YYYY-MON') DATE_ID: The DATE_ID column has timestamp or case_id extracted in year-month (yyyy-mon) format.
   - round(VALUE,2) ACTUAL_SOLD: Specifies the AMOUNT_SOLD value as ACTUAL_SOLD rounded to two numericals after the decimal.
   - round(PREDICTION,2) FORECAST_SOLD: Specifies the predicted value as FORECAST_SOLD rounded to two numericals after the decimal.
   - round(LOWER,2) LOWER_BOUND, round(UPPER,2) UPPER_BOUND: Specifies the lower and upper confidence levels rounded to two numericals after the decimal.

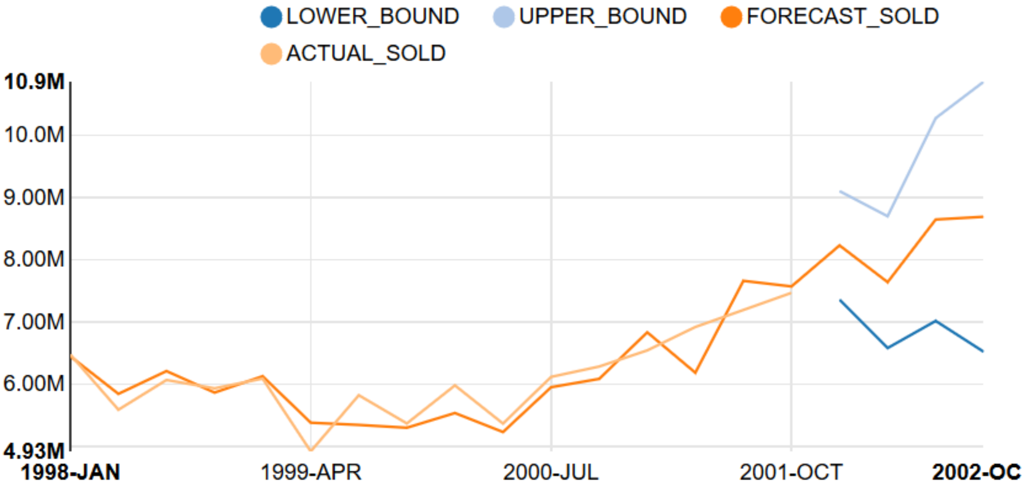
```
DATE_ID     ACTUAL_SOLD    FORECAST_SOLD    LOWER_BOUND    UPPER_BOUND
1998-JAN          6480684       6452375.75
```

```
1998-APR      5593994.14      5848724.79
1998-JUL       6071823.1      6214546.31
1998-OCT      5937413.71      5869219.42
1999-JAN      6093747.21      6132016.41
1999-APR      4925471.63      5385954.08
1999-JUL      5827050.15      5350240.25
1999-OCT      5373678.67      5304626.05
2000-JAN      5984889.49      5541123.24
2000-APR      5371730.92       5236126.1
2000-JUL      6121239.29      5955258.74
2000-OCT      6287646.92       6089446.4
2001-JAN      6547097.44      6837567.17
2001-APR      6922468.39      6188944.05

DATE_ID    ACTUAL_SOLD   FORECAST_SOLD   LOWER_BOUND   UPPER_BOUND
2001-JUL      7195998.63      7663836.77
2001-OCT      7470897.52      7573926.96
2002-JAN                      8232820.51    7360847.49    9104793.54
2002-APR                      7642694.94    6584565.24    8700824.63
2002-JUL                      8648402.54    7019914.28   10276890.81
2002-OCT                      8692842.46    6523676.33    10862008.6


20 rows selected.


---------------------------
```

2. To see a visual representation of the predictions in OML Notebooks, run the above same query with the following settings:

   Click **settings** and drag `DATE_ID` to **keys** and `FORECASTED_SOLD (avg)`, `ACTUAL_SOLD (avge)`, `LOWER_BOUND (avg)`, and `UPPER_BOUND(avg)` to **values**.

   ```
   %sql
   SELECT TO_CHAR(CASE_ID,'YYYY-MON') DATE_ID, VALUE ACTUAL_SOLD,
          round(PREDICTION,2) FORECAST_SOLD,
          round(LOWER,2) LOWER_BOUND, round(UPPER,2) UPPER_BOUND
     FROM DM$VPESM_SALES_FORECAST_1
     ORDER BY CASE_ID;
   ```



This completes the prediction step. The model has successfully forecast sales for the next four quarters. This helps in tracking the sales and also gives us an idea on stocking our products.

Association Rules Use Case Scenario

A popular movie rental website is being updated. The movie rental company wishes to provide movie recommendations to their customers based on their frequently rented movies and purchase transaction history. They approach you, a data scientist, for assistance with movie recommendations. Using the Apriori algorithm, you solve this problem by analysing popular movies that are frequently watched together.

Related Content

Topic	**Link**
OML4SQL GitHub Example	Apriori - Association Rules
`CREATE_MODEL2` Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING - Model Settings
Apriori Settings	DBMS_DATA_MINING - Machine Learning Function Settings
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views
Association Rules - Model Detail Views	• Model Detail Views for Association Rules • Model Detail View for Frequent Itemsets • Model Detail Views for Transactional Itemsets • Model Detail View for Transactional Rule
About Association	About Association
About Apriori	About Apriori

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data set
  The data set used for this use case is called MovieStream data set.

> **Note:**
>
> This data set is used for illustrative purpose only.

- Database
  Select or create database out of the following options:
  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements
  Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Related Topics

- Analyze Moviestream data in Oracle Autonomous Data Warehouse using SQL

Load Data

Examine the data set and its attributes. Load the data in your database.

In this use case, you will load the data set to your database. If you are using Oracle Autonomous Database, you will use an existing data file from the Oracle Cloud Infrastructure (OCI) Object Storage. You will create a sample table, load data into the sample table from files on the OCI Object Storage, and explore the data. If you are using the on-premises database, you will use Oracle SQL developer to import the data set and explore the data.

To understand the data, you will perform the following:

- Access the data.
- Examine the various attributes or columns of the data set.
- Assess data quality (by exploring the data).

Examine Data

The following table displays information about the attributes from `MOVIES_SALES_FACT`:

Attribute Name	Information
`ORDER_NUM`	Specifies the order number
`ACTUAL_PRICE`	Specifies the actual price of the movie
`AGE`	Specifies the age of the customer
`AGE_BAND`	Specifies the age band of the customer. The possible values are 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and so on.
`APP`	Specifies the application used for the movie
`CITY`	Specifies the name of the city
`CITY_ID`	Specifies the city ID
`COMMUTE_DISTANCE`	Specifies the commute distance
`COMMUTE_DISTANCE_BAND`	Specifies the commute distance band

Attribute Name	Information
CONTINENT	Specifies the continent name
COUNTRY	Specifies the country name
COUNTRY_CODE	Specifies the country code
COUNTRY_ID	Specifies the country ID
CREDIT_BALANCE	Specifies the credit balance of the customer
CUSTOMER_ID	Specifies the customer ID
CUSTOMER_NAME	Specifies the customer name
DAY	Specifies the day of the week in YYYY-mm-dd hh:mm:ss format
DAY_NAME	Specifies the day of the week
DAY_NUM_OF_WEEK	Specifies the day number of the week
DEVICE	Specifies the device information used by the customer
DISCOUNT_PERCENT	Specifies the discount percent
DISCOUNT_TYPE	Specifies the discount type availed by the customer. Possible values are referral, coupon, promotion, volume, none
EDUCATION	Specifies customer's education
EMAIL	Specifies email ID of the customer
FULL_TIME	Specifies customer's employment status such as full time, not employed, part time
GENDER	Specifies the gender of the customer
GENRE	Specifies the genre of the movie
HOUSEHOLD_SIZE	Specifies the household size of the customer
HOUSEHOLD_SIZE_BAND	Specifies the household size band
INCOME	Specifies the income of the customer
INCOME_BAND	Specifies the income band of the customer
INSUFF_FUNDS_INCIDENTS	Specifies the number of insufficient funds incidents that the customer had
JOB_TYPE	Specifies the cusotmer's job
LATE_MORT_RENT_PMTS	Specifies is the customer had any late mortgage or rent payment
LIST_PRICE	Specifies the list price of the movie
MARITAL_STATUS	Specifies the marital status of the customer
MONTH	Specifies the month in MON-YYYY format
MONTH_NAME	Specifies the month. For example, January.
MONTH_NUM_OF_YEAR	Specifies the month number of the year
MORTGAGE_AMT	Specifies the mortgage amount
MOVIE_ID	Specifies the movie ID
NUM_CARS	Specifies the number of the cars that the customer owns
NUM_MORTGAGES	Specifies the number of mortgages
OS	Specifies the OS information
PAYMENT_METHOD	Specifies the payment method
PET	Specifies if the customer owns a pet

Attribute Name	Information
POSTAL_CODE	Specifies the postal code of the address
PROMOTION_RESPONSE	Specifies the response of the customer to a promotional offer
QUANTITY_SOLD	Specifies the quantity sold
QUARTER_NAME	Specifies the quarter name in Qn-YYYY format. For example, Q1-2001.
QUARTER_NUM_OF_YEAR	Specifies the quarter number of the year
RENT_OWN	Specifies if the customer is living at a rented place or own place
SEARCH_GENRE	Specifies the genre of the movies searched
SEGMENT_DESCRIPTION	Describes the population segment
SEGMENT_NAME	Specifies the population segment name
SKU	Specifies the SKU ID
STATE_PROVINCE	Specifies the province
STATE_PROVINCE_ID	Specifies the province ID
STREET_ADDRESS	Specifies the customer's address
TITLE	Specifies the movie title
USERNAME	Specifies the username provided by the customer
WORK_EXPERIENCE	Specifies the work experience of the customer
WORK_EXPERIENCE_BAND	Specifies the work experience band of the customer
YEAR	Specifies the year
YEARS_CURRENT_EMPLOYER	Specifies the current employer of the customer
YEARS_CURRENT_EMPLOYER_BAND	Specifies the customer's employment band in years with the current employer
YEARS_RESIDENCE	Specifies the number of years the customer has been residing at a place
YEARS_RESIDENCE_BAND	Specifies the residence band

Create a Table

Create a table called MOVIE_SALES_FACT. This table is used in DBMS_CLOUD.COPY_DATA procedure to access the data set.

Enter the following code in the OML Notebooks and run the notebook.

```
%sql
CREATE TABLE MOVIE_SALES_FACT
( ORDER_NUM NUMBER(38,0),
 DAY DATE,
 DAY_NUM_OF_WEEK NUMBER(38,0),
 DAY_NAME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 MONTH VARCHAR2(12 BYTE) COLLATE USING_NLS_COMP,
 MONTH_NUM_OF_YEAR NUMBER(38,0),
 MONTH_NAME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 QUARTER_NAME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 QUARTER_NUM_OF_YEAR NUMBER(38,0),
 YEAR NUMBER(38,0),
 CUSTOMER_ID NUMBER(38,0),
 USERNAME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 CUSTOMER_NAME VARCHAR2(250 BYTE) COLLATE USING_NLS_COMP,
 STREET_ADDRESS VARCHAR2(250 BYTE) COLLATE USING_NLS_COMP,
 POSTAL_CODE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 CITY_ID NUMBER(38,0),
 CITY VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 STATE_PROVINCE_ID NUMBER(38,0),
 STATE_PROVINCE VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 COUNTRY_ID NUMBER(38,0),
 COUNTRY VARCHAR2(126 BYTE) COLLATE USING_NLS_COMP,
 COUNTRY_CODE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 CONTINENT VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 SEGMENT_NAME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 SEGMENT_DESCRIPTION VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 CREDIT_BALANCE NUMBER(38,0),
 EDUCATION VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 EMAIL VARCHAR2(128 BYTE) COLLATE USING_NLS_COMP,
 FULL_TIME VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 GENDER VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 HOUSEHOLD_SIZE NUMBER(38,0),
 HOUSEHOLD_SIZE_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 WORK_EXPERIENCE NUMBER(38,0),
 WORK_EXPERIENCE_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 INSUFF_FUNDS_INCIDENTS NUMBER(38,0),
 JOB_TYPE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 LATE_MORT_RENT_PMTS NUMBER(38,0),
 MARITAL_STATUS VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 MORTGAGE_AMT NUMBER(38,0),
 NUM_CARS NUMBER(38,0),
 NUM_MORTGAGES NUMBER(38,0),
 PET VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 PROMOTION_RESPONSE NUMBER(38,0),
 RENT_OWN VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 YEARS_CURRENT_EMPLOYER NUMBER(38,0),
 YEARS_CURRENT_EMPLOYER_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 YEARS_CUSTOMER NUMBER(38,0),
 YEARS_CUSTOMER_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 YEARS_RESIDENCE NUMBER(38,0),
 YEARS_RESIDENCE_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 AGE NUMBER(38,0),
 AGE_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 COMMUTE_DISTANCE NUMBER(38,0),
 COMMUTE_DISTANCE_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 INCOME NUMBER(38,0),
 INCOME_BAND VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 MOVIE_ID NUMBER(38,0),
 SEARCH_GENRE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 TITLE VARCHAR2(4000 BYTE) COLLATE USING_NLS_COMP,
 GENRE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 SKU NUMBER(38,0),
 LIST_PRICE NUMBER(38,2),
 APP VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 DEVICE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 OS VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 PAYMENT_METHOD VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 DISCOUNT_TYPE VARCHAR2(26 BYTE) COLLATE USING_NLS_COMP,
 DISCOUNT_PERCENT NUMBER(38,1),
 ACTUAL_PRICE NUMBER(38,2),
 QUANTITY_SOLD NUMBER(38,0)
 )
;
```

Load Data in the Table

Load the data set stored in object storage to the `MOVIE_SALES_FACT` table.

Before you load this data ensure that you set Compute Resources to Medium or High. If you select High, then, set **Memory** field to `16` for High Resource Service. You must have Administrator privilege to configure the memory settings. See Compute Resource.

Add a new paragraph in your OML notebook and run the following statement:

```
%script
BEGIN
 DBMS_CLOUD.COPY_DATA (table_name => 'MOVIE_SALES_FACT',file_uri_list =>
'https://objectstorage.uk-london-1.oraclecloud.com/n/adwc4pm/b/
moviestream_kl/o/d801_movie_sales_fact_m-*.csv', format => '{"delimiter":",",
"recorddelimiter":"newline", "skipheaders":"1", "quote":"\\\"",
"rejectlimit":"1000", "trimspaces":"rtrim", "ignoreblanklines":"false",
"ignoremissingcolumns":"true", "dateformat":"DD-MON-YYYY HH24:MI:SS"}');
END;
/



PL/SQL procedure successfully completed.


---------------------------
```

Examine the statement:

- `table_name`: is the target table's name.
- `credential_name`: is the name of the credential created earlier.
- `file_uri_list`: is a comma delimited list of the source files you want to load. The special character `*` in the file `d801_movie_sales_fact_m-*.csv` means you are bulk loading the MovieStream data set containing sales data for 2018-2020.
- `format`: defines the options you can specify to describe the format of the source file, including whether the file is of type text, ORC, Parquet, or Avro.
  - `delimiter`: Specifies the field delimiter (special character). Here, it is specified as "," (comma)
  - `recorddelimiter`: Specifies the record delimiter. The default value is `newline`. By default, `DBMS_CLOUD` tries to automatically find the correct newline character as the delimiter.
  - `skipheaders`: Specifies how many rows should be skipped from the start of the file. In this use case, it is `1`.

  - `quote`: Specifies the quote character for the fields.
  - `rejectlimit`: The operation will error out after specified number of rows are rejected. Here, the value is `1000`.
  - `trimspaces`: Specifies how the leading and trailing spaces of the fields are trimmed. Here it is `rtrim`. The `rtrim` value indicates that you want trailing spaces trimmed.
  - `ignoreblanklines`: Blank lines are ignored when set to true. The default value is `false`.
  - `ignoremissingcolumns`: If there are more columns in the `field_list` than there are in the source files, the extra columns are stored as null. The default value is `false`. In this use case, it is set to `true`.
  - `dateformat`: Specifies the date format in the source file.

In this example, *adwc4pm* is the Oracle Cloud Infrastructure object storage namespace and *moviestream_kl* is the bucket name.

Related Topics

- DBMS_CLOUD.COPY_DATA Procedure

Explore Data

Once the data is loaded into the table, explore the data to understand and assess the quality of the data. At this stage assess the data to identify data types and noise in the data. Look for missing values and numeric outlier values.

Assess Data Quality

To assess the data, first, you must be able to view the data in your database. For this reason, you will use SQL statements to query the table.

If you are working with Oracle Autonomous Database, you can use the Oracle Machine Learning (OML) Notebooks for your data science project, including assessing data quality. If you are using the on-premises Oracle Database, you can use the Oracle SQL Developer to assess data quality. Query the data as described.

> **Note:**
>
> Each record in the database is called a case and each case is identified by a `case_id`. In this use case, `CUSTOMER_ID` is the `case_id`.

The following steps help you with the exploratory analysis of the data:

1. View the data in the `MOVIE_SALES_FACT` table by running the following query:

   ```
   SELECT * FROM MOVIE_SALES_FACT;
   ```

2. Find the `COUNT` rows in the data set, run the following statement:

```
SELECT DISTINCT COUNT(*) from MOVIE_SALES_FACT;



COUNT(*)
 97890562
---------------------------
```

3. To find distinct or unique customers in the table, run the following statement:

```
%script SELECT COUNT (DISTINCT CUST_ID) FROM MOVIE_SALES_FACT;


COUNT(DISTINCTCUST_ID)
4845
---------------------------
```

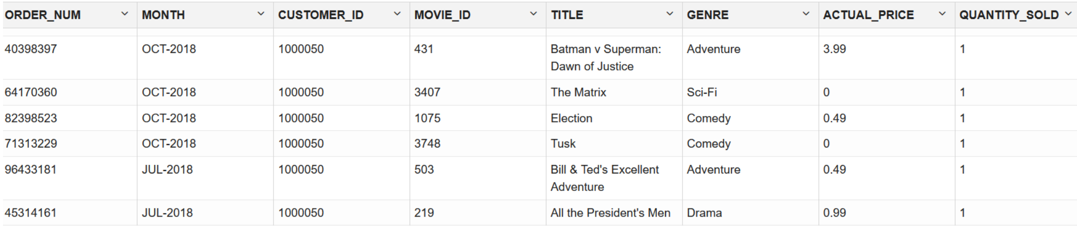
4. To view the data type of the columns, run the following statement:

```
%script
DESCRIBE MOVIE_SALES_FACT;



Name                 Null?      Type
--------------------------- ----- --------------
ORDER_NUM                 NUMBER(38)
DAY DATE
DAY_NUM_OF_WEEK                          NUMBER(38)
DAY_NAME                     VARCHAR2(26)
MONTH                         VARCHAR2(12)
MONTH_NUM_OF_YEAR                   NUMBER(38)
MONTH_NAME                 VARCHAR2(26)
QUARTER_NAME                      VARCHAR2(26)
QUARTER_NUM_OF_YEAR                  NUMBER(38)
YEAR                     NUMBER(38)
CUSTOMER_ID                 NUMBER(38)
USERNAME                 VARCHAR2(26)
CUSTOMER_NAME                     VARCHAR2(250)
STREET_ADDRESS                     VARCHAR2(250)
POSTAL_CODE                 VARCHAR2(26)
CITY_ID                 NUMBER(38)
CITY                     VARCHAR2(128)
STATE_PROVINCE_ID             NUMBER(38)
STATE_PROVINCE                 VARCHAR2(128)
COUNTRY_ID                 NUMBER(38)
COUNTRY                 VARCHAR2(126)
COUNTRY_CODE                 VARCHAR2(26)
CONTINENT                 VARCHAR2(128)
SEGMENT_NAME                 VARCHAR2(26)
SEGMENT_DESCRIPTION             VARCHAR2(128)
CREDIT_BALANCE                 NUMBER(38)
EDUCATION                 VARCHAR2(128)
EMAIL                     VARCHAR2(128)
FULL_TIME                 VARCHAR2(26)
```

```
GENDER                     VARCHAR2(26)
HOUSEHOLD_SIZE                 NUMBER(38)
HOUSEHOLD_SIZE_BAND             VARCHAR2(26)
WORK_EXPERIENCE             NUMBER(38)
WORK_EXPERIENCE_BAND             VARCHAR2(26)
INSUFF_FUNDS_INCIDENTS             NUMBER(38)
JOB_TYPE                 VARCHAR2(26)
LATE_MORT_RENT_PMTS             NUMBER(38)
MARITAL_STATUS                 VARCHAR2(26)
MORTGAGE_AMT                 NUMBER(38)
NUM_CARS                 NUMBER(38)
NUM_MORTGAGES                 NUMBER(38)
PET                     VARCHAR2(26)
PROMOTION_RESPONSE             NUMBER(38)
RENT_OWN                 VARCHAR2(26)
YEARS_CURRENT_EMPLOYER             NUMBER(38)
YEARS_CURRENT_EMPLOYER_BAND           VARCHAR2(26)
YEARS_CUSTOMER                 NUMBER(38)
YEARS_CUSTOMER_BAND             VARCHAR2(26)
YEARS_RESIDENCE             NUMBER(38)
YEARS_RESIDENCE_BAND             VARCHAR2(26)
AGE                     NUMBER(38)
AGE_BAND                 VARCHAR2(26)
COMMUTE_DISTANCE             NUMBER(38)
COMMUTE_DISTANCE_BAND             VARCHAR2(26)
INCOME                     NUMBER(38)
INCOME_BAND                 VARCHAR2(26)
MOVIE_ID                 NUMBER(38)
SEARCH_GENRE                 VARCHAR2(26)
TITLE                     VARCHAR2(4000)
GENRE                     VARCHAR2(26)
SKU                     NUMBER(38)
LIST_PRICE                 NUMBER(38,2)
APP                     VARCHAR2(26)
DEVICE                     VARCHAR2(26)
OS                     VARCHAR2(26)
PAYMENT_METHOD                 VARCHAR2(26)
DISCOUNT_TYPE                 VARCHAR2(26)
DISCOUNT_PERCENT             NUMBER(38,1)
ACTUAL_PRICE                 NUMBER(38,2)
QUANTITY_SOLD                 NUMBER(38)

---------------------------
```

5. Select the required columns from `MOVIE_SALES_FACT` table.

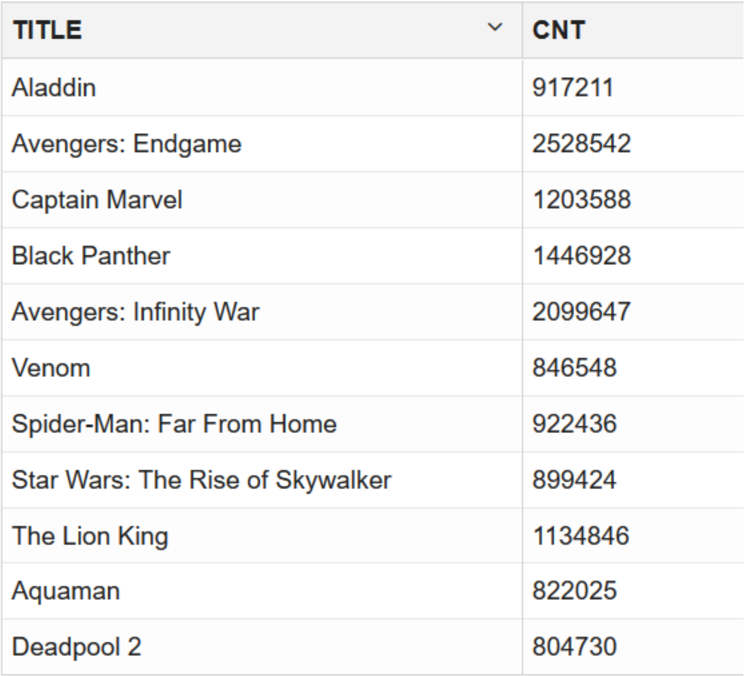
```
%sql
 SELECT ORDER_NUM, MONTH, CUSTOMER_ID, MOVIE_ID, TITLE, GENRE, ACTUAL_PRICE,
QUANTITY_SOLD FROM MOVIE_SALES_FACT
 ORDER BY CUSTOMER_ID;
```



ORDER_NUM	MONTH	CUSTOMER_ID	MOVIE_ID	TITLE	GENRE	ACTUAL_PRICE	QUANTITY_SOLD
40398397	OCT-2018	1000050	431	Batman v Superman: Dawn of Justice	Adventure	3.99	1
64170360	OCT-2018	1000050	3407	The Matrix	Sci-Fi	0	1
82398523	OCT-2018	1000050	1075	Election	Comedy	0.49	1
71313229	OCT-2018	1000050	3748	Tusk	Comedy	0	1
96433181	JUL-2018	1000050	503	Bill & Ted's Excellent Adventure	Adventure	0.49	1
45314161	JUL-2018	1000050	219	All the President's Men	Drama	0.99	1

6. Select customers who watched, for example, the movie "Titanic" and check other popular movies watched among those customers.

```
%sql
select title, count(1) cnt
from movie_sales_fact a
join (
select distinct customer_id
from movie_sales_fact
where title = 'Titanic' ) b
on a.customer_id = b.customer_id
group by title
having count(1) > 800000
```



TITLE	CNT
Aladdin	917211
Avengers: Endgame	2528542
Captain Marvel	1203588
Black Panther	1446928
Avengers: Infinity War	2099647
Venom	846548
Spider-Man: Far From Home	922436
Star Wars: The Rise of Skywalker	899424
The Lion King	1134846
Aquaman	822025
Deadpool 2	804730

7. The data set is huge with millions of records. Create a view called `MOVIES` to select a smaller data set by providing a customer ID range.

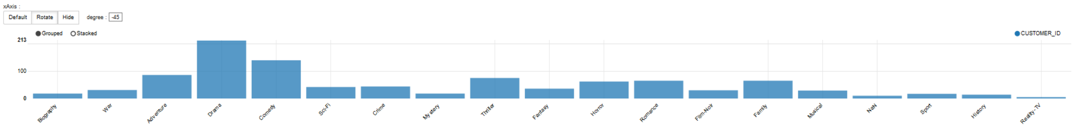
```
%script
CREATE OR REPLACE VIEW MOVIES AS
SELECT DISTINCT CUSTOMER_ID, MOVIE_ID, TITLE, GENRE
```

```
FROM MOVIE_SALES_FACT
WHERE CUSTOMER_ID BETWEEN 1000000 AND 1000120;


View MOVIES created.
 ---------------------------
```

8. You can check the distribution of genre from the new view `MOVIES`:

   ```
   %sql
   SELECT * FROM MOVIES;
   ```

   In OML Notebooks, click the bar icon and expand settings. Drag GENRE to **keys** and CUSTOMER_ID to **values** and select **COUNT**.

9. Now, check the count of rows by running the following statement:

   ```
   %script
   SELECT DISTINCT COUNT (*) FROM MOVIES;



   COUNT(*)
    10194
   ---------------------------
   ```

10. To check if there are any missing values (NULL values), run the following statement:

    ```
    SELECT COUNT(*) FROM MOVIES WHERE CUSTOMER_ID=NULL OR MOVIE_ID=NULL OR TITLE=NULL OR
    GENRE=NULL;


    COUNT(*)
    0
    ---------------------------
    ```

    NULLs, if found, are automatically handled by the OML algorithms. Alternately, you can manually replace NULLs with `NVL` SQL function.

This completes the data exploration stage. OML supports Automatic Data Preparation (ADP). ADP is enabled through the model settings. When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model. This step is done during the Build Model stage. The commonly used methods of data preparation are binning, normalization, and missing value treatment.

Related Topics

- How ADP Transforms the Data

Build Model

Build your model using your data set. Use the DBMS_DATA_MINING.CREATE_MODEL2 procedure to build your model and specify the model settings.

For unsupervised learning, like Association Rules, you do not have labels or predictors to calculate the accuracy or assess the performance. So you don't need to train your model on a separate training data set and then evaluate it on a test set. The entire data set can be used to build the model. For an unsupervised learning, you don't have an objective way to assess your model. So, a training or a test split is not useful.

Algorithm Selection

Oracle supports the Apriori algorithm to build an Association Rules model.

Apriori calculates the probability of an item being present in a frequent itemset, given that another item or group of items is present. An itemset is any combination of two or more items in a transaction. Frequent itemsets are those that occur with a minimum frequency that the user specifies. An association rule states that an item or group of items implies the presence of another item with some probability and support.

The following steps guide you to build your model with the Apriori algorithm.

- Build your model using the CREATE_MODEL2 procedure. First, declare a variable to store model settings or hyperparameters. Run the following script:

```
%script
 BEGIN DBMS_DATA_MINING.DROP_MODEL('AR_MOVIES');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
 v_setlist DBMS_DATA_MINING.SETTING_LIST;
BEGIN
 v_setlist('ALGO_NAME') := 'ALGO_APRIORI_ASSOCIATION_RULES';
 V_setlist('PREP_AUTO') := 'ON';
 V_setlist('ASSO_MIN_SUPPORT') := '0.02';
 V_setlist('ASSO_MIN_CONFIDENCE') := '0.1';
 V_setlist('ASSO_MAX_RULE_LENGTH'):= '2';
 V_setlist('ODMS_ITEM_ID_COLUMN_NAME'):= 'TITLE';

 DBMS_DATA_MINING.CREATE_MODEL2(MODEL_NAME         => 'AR_MOVIES',
                   MINING_FUNCTION     => 'ASSOCIATION',
                   DATA_QUERY         => 'select * from MOVIES',
                           SET_LIST         => v_setlist,
                   CASE_ID_COLUMN_NAME => 'CUSTOMER_ID');
END;



PL/SQL procedure successfully completed.
---------------------------
PL/SQL procedure successfully completed.
```

  Examine the script:

  - v_setlist is a variable to store SETTING_LIST.

- `DBMS_DATA_MINING` is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- `SETTING_LIST` specifies model settings or hyperparameters for our model.
- `ALGO_NAME` specifies the algorithm name. Since you are using Apriori as your algorithm, set `ALGO_APRIORI_ASSOCIATION_RULES`.
- `PREP_AUTO` is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is `ON`.
- `ASSO_MIN_SUPPORT` is minimum support for association rules (in percentage) that limits the number of itemsets used for association rules. An itemset must appear in at least this percentage of all the transactions if it is to be used as a basis for rules. Apriori discovers patterns with frequencies above the minimum support threshold. This is the minimum threshold that each rule must satisfy. Here, the algorithms finds patterns with frequenqies above 0.02. Increase the minimum support if you want to decrease the build time for the model and generate fewer rules.
- `ASSO_MIN_CONFIDENCE` determines minimum confidence for association rules. It is a conditional probability that the consequent occurs given the occurrence of an antecedent. In other words, the confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction. The default value is 0.1.
- `ASSO_MAX_RULE_LENGTH` specifies the maximum number of items in an itemset. If the maximum is two, all the item pairs are counted. In this use case, if you want to increase the value to 3, consider working with a smaller data set since each customer would watch lot of movies. If the maximum is greater than two, all the item pairs, all the item triples, and all the item combinations up to the specified maximum are counted. Increasing this value increases the run time and complexity significantly. Hence, for demonstration purposes on this data set, it is recommended to set the value to 2.

  > **Tip:**
  >
  > One way to limit the number of rules produced is to raise the support and confidence. Support is the joint probability of two items that are purchased together. For instance, item beer and diaper happens together with probability of 0.1, vodka and ice cream are purchased together with the probability of 0.05. If you raise the support threshold to 0.1. You will not see vodka and ice cream in the rules. Similarly, the confidence is the probability of people purchasing item A given they have purchased B. The probability of people who purchase beer given that they have already purchased a diaper is 0.2; The probability of people who purchase ice cream given that they have purchased vodka is 0.6. Using the threshold 0.6, you can remove the rule of people purchasing beer given that they already purchased diaper.

- `ODMS_ITEM_ID_COLUMN_NAME` name of a column that contains the items in a transaction. In this use case, it is `TITLE`. When this setting is specified, the algorithm expects the data to be presented in a native transactional format, consisting of two columns:
  - Case ID, either categorical or numeric
  - Item ID, either categorical or numeric

The `CREATE_MODEL2` procedure takes the following parameters:

- `MODEL_NAME`: Specify a unique name for your model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is `AR_MOVIES`.
- `MINING_FUNCTION`: Specifies the machine learning function. Since you are solving an association problem in this use case, select `ASSOCIATION`.
- `DATA_QUERY`: A query that provides training data for building the model. Here, the query is `SELECT * FROM MOVIES`.
- `SET_LIST`: Specifies `SETTING_LIST` variable. Here, it is `v_setlist`.
- `CASE_ID_COLUMN_NAME`: A unique case identifier column in the build data. In this use case, case_id is `CUSTOMER_ID`. If there is a composite key, you must create a new attribute before creating the model. This may involve concatenating values from the columns, or mapping a unique identifier to each distinct combination of values. The `CASE_ID` assists with reproducible results, joining scores for individual customers with other data in, example, scoring data table.

Note:

Any parameters or settings not specified are either system-determined or default values are used.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

Sometimes querying dictionary views and model detail views is sufficient to measure your model's performance. However, you can evaluate your model by computing test metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), confusion matrix, lift statistics, cost matrix, and so on. For Association Rules, you can inspect various rules to see if they reveal new insights for item dependencies (antecedent itemset implying consequent) or for unexpected relationships among items.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

By examining various statistics in the model detail views, you can compare models to arrive at one model that satisfies your evaluation criteria.
The results of an association model are the rules that identify patterns of association within the data. Oracle Machine Learning for SQL does not support a scoring operation for association modeling. Instead, support and confidence are the primary metrics for evaluating the quality of the rules that the model generates. These statistical measures can be used to rank the rules and hence the usefulness of the predictions.

Association rules can be applied as follows:

- Support: How often do these items occur together in the data when you apply Association Rules?
- Confidence: How frequently the consequent occurs in transactions that contain the antecedent.

- Value: How much business value is connected to item associations

Additionally, Oracle Machine Learning for SQL supports lift for association rules. Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support. Lift provides information about the improvement, the increase in probability of the consequent given the antecedent. Lift is defined as confidence of the combination of items divided by the support of the consequent. Any rule with an improvement of less than 1 does not indicate a real cross-selling opportunity, no matter how high its support and confidence, because it actually offers less ability to predict a purchase than does random chance.

The data dictionary views for Oracle Machine Learning are listed in the following table. A database administrator (DBA) and USER versions of the views are also available.

View Name	Description
ALL_MINING_MODELS	Provides information about all accessible machine learning models
ALL_MINING_MODEL_ATTRIBUTES	Provides information about the attributes of all accessible machine learning models
ALL_MINING_MODEL_SETTINGS	Provides information about the configuration settings for all accessible machine learning models
ALL_MINING_MODEL_VIEWS	Provides information about the model views for all accessible machine learning models
ALL_MINING_MODEL_XFORMS	Provides the user-specified transformations embedded in all accessible machine learning models.

Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views.

1. You can review the model settings in `USER_MINING_MODEL_SETTINGS` by running the following query:

```
SELECT SETTING_NAME, SETTING_VALUE
 FROM USER_MINING_MODEL_SETTINGS
 WHERE MODEL_NAME = 'AR_MOVIES'
 ORDER BY SETTING_NAME;



SETTING_NAME                    SETTING_VALUE
ALGO_NAME                       ALGO_APRIORI_ASSOCIATION_RULES
ASSO_MAX_RULE_LENGTH            2
ASSO_MIN_CONFIDENCE             0.1
ASSO_MIN_REV_CONFIDENCE         0
ASSO_MIN_SUPPORT                0.02
ASSO_MIN_SUPPORT_INT            1
ODMS_DETAILS                    ODMS_ENABLE
ODMS_ITEM_ID_COLUMN_NAME        TITLE
ODMS_MISSING_VALUE_TREATMENT    ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING                   ODMS_SAMPLING_DISABLE
PREP_AUTO                       ON


11 rows selected.
```

```
---------------------------
```

2. Run the following statement to see information on various views in `USER_MINING_MODEL_VIEWS`:

```
SELECT view_name, view_type FROM USER_MINING_MODEL_VIEWS
 WHERE MODEL_NAME = 'AR_MOVIES'
 ORDER BY VIEW_NAME;



VIEW_NAME        VIEW_TYPE
DM$VAAR_MOVIES   Association Rules For Transactional Data
DM$VGAR_MOVIES   Global Name-Value Pairs
DM$VIAR_MOVIES   Association Rule Itemsets
DM$VRAR_MOVIES   Association Rules
DM$VSAR_MOVIES   Computed Settings
DM$VTAR_MOVIES   Association Rule Itemsets For Transactional Data
DM$VWAR_MOVIES   Model Build Alerts


7 rows selected.


---------------------------
```

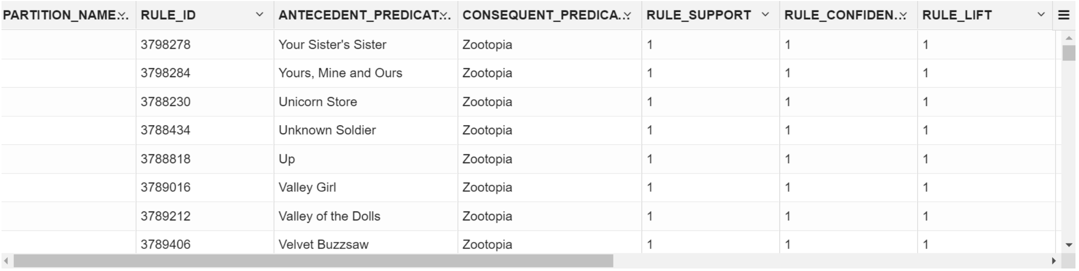
3. To view the Association Rules Itemsets For Transactional Data (DM$VTxx) model detail view, run the following script:

```
%script
SELECT ITEM_NAME, SUPPORT, NUMBER_OF_ITEMS
FROM DM$VTAR_MOVIES;




ITEM_NAME                SUPPORT               NUMBER_OF_ITEMS
Dallas Buyers Club                           1                 2
Dallas Buyers Club         0.66666666666666663                 2
Dallas Buyers Club         0.33333333333333331                 2
Elvira's Haunted Hills                       1                 2
Elvira's Haunted Hills     0.66666666666666663                 2
Elvira's Haunted Hills     0.33333333333333331                 2
Elvira's Haunted Hills                       1                 2
Elvira's Haunted Hills                       1                 2
Ghostbusters {{nbsp II                       1                 2
Ghostbusters {{nbsp II     0.66666666666666663                 2
Ghostbusters {{nbsp II     0.33333333333333331                 2
Ghostbusters {{nbsp II                       1                 2
Ghostbusters {{nbsp II                       1                 2
Hits                       0.33333333333333331                 2
...
```

This view provides the itemsets information in transactional format. In the first transaction, *Dallas Buyers Club* and another movie are purchased or rented together with 100% support (support 1).

4. Now, view the Association Rules for Transactional Data (DM$VAxx) model detail view:

```
%sql SELECT * FROM DM$VAAR_MOVIES;
```



PARTITION_NAME	RULE_ID	ANTECEDENT_PREDICAT..	CONSEQUENT_PREDICA..	RULE_SUPPORT	RULE_CONFIDEN..	RULE_LIFT
	3798278	Your Sister's Sister	Zootopia	1	1	1
	3798284	Yours, Mine and Ours	Zootopia	1	1	1
	3788230	Unicorn Store	Zootopia	1	1	1
	3788434	Unknown Soldier	Zootopia	1	1	1
	3788818	Up	Zootopia	1	1	1
	3789016	Valley Girl	Zootopia	1	1	1
	3789212	Valley of the Dolls	Zootopia	1	1	1
	3789406	Velvet Buzzsaw	Zootopia	1	1	1

From this view, you can see that both antecedent and consequent are purchased together frequently (Support =1). You can expect the consequent to be present whenever the listed antecedent is present (Confidence=1). You can say that the probability of purchasing the consequent increases with the presence of the listed antecedent (Lift=1).

5. To see top 10 association rules, run the following query:

The IF component of an association rule is known as the **antecedent**. The THEN component is known as the **consequent**. The antecedent and the consequent are disjoint; they have no items in common. Oracle Machine Learning for SQL supports association rules that have one or more items in the antecedent and a single item in the consequent.

```
%script
SELECT * FROM
 (SELECT RULE_ID, ANTECEDENT_PREDICATE ANTECEDENT,
CONSEQUENT_PREDICATE CONSEQUENT,
ROUND(RULE_SUPPORT,3) SUPP, ROUND(RULE_CONFIDENCE,3) CONF, NUMBER_OF_ITEMS
NUM_ITEMS
FROM DM$VAAR_MOVIES
ORDER BY RULE_CONFIDENCE DESC, RULE_SUPPORT DESC)
WHERE ROWNUM <= 10
ORDER BY RULE_ID;



RULE_ID   ANTECEDENT                   CONSEQUENT   SUPP   CONF
NUM_ITEMS
    10759 101 Dalmatians               10                1
1           2
    10761 12 Years a Slave             10                1
1           2
    10763 127 Hours                    10                1
1           2
    10771 1984                         10                1
1           2
    10773 2-Headed Shark Attack        10                1
1           2
    10777 20,000 Leagues Under the Sea 10                1
1           2
    10779 2001: A Space Odyssey        10                1
```

```
1          2
    10781 2012                          10                1
1          2
    10785 3 Ninjas                      10                1
1          2
    10787 3 from Hell                   10                1
1          2


10 rows selected.


---------------------------
```

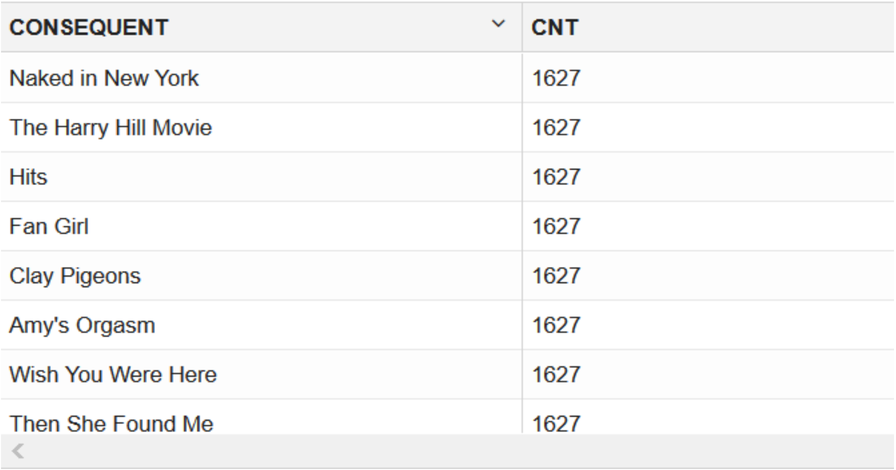
Examine the statement:

- RULE_ID is the rule identifier.
- ANTECEDENT_PREDICATE: provides the name of the antecedent.
- CONSEQUENT_PREDICATE: provides name of the consequent item.
- ROUND (RULE_SUPPORT, 3) SUPP: provides support of the rule rounded to 3 digits after the decimal.
- ROUND(RULE_CONFIDENCE, 3) CONF: the likelihood a transaction satisfying the rule when it contains the antecedent, rounded to 3 digits after the decimal.
- NUM_OF_ITEMS: specifies number of items in a rule.

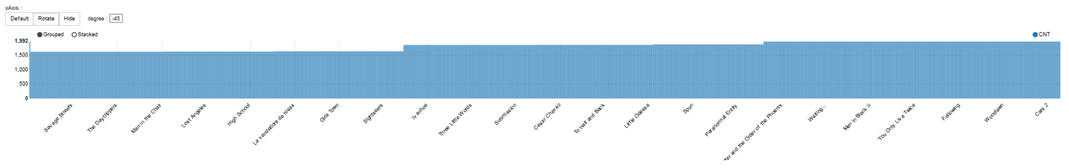
6. You can also view which consequent items occur most frequently or which consequent items are included in most rules. To do so, run the following query:

```
%sql
SELECT CONSEQUENT, COUNT(1) CNT FROM
(SELECT ANTECEDENT_PREDICATE ANTECEDENT,
CONSEQUENT_PREDICATE CONSEQUENT,
RULE_SUPPORT SUPP, RULE_CONFIDENCE CONF, NUMBER_OF_ITEMS NUM
FROM DM$VAAR_MOVIES
ORDER BY RULE_CONFIDENCE DESC)
GROUP BY  CONSEQUENT
ORDER BY CNT;
```

In OML Notebooks, click **settings** and click the **Bar Chart** icon to visualize the result. Click **Rotate** to rotate the bar graph to 45 degrees.



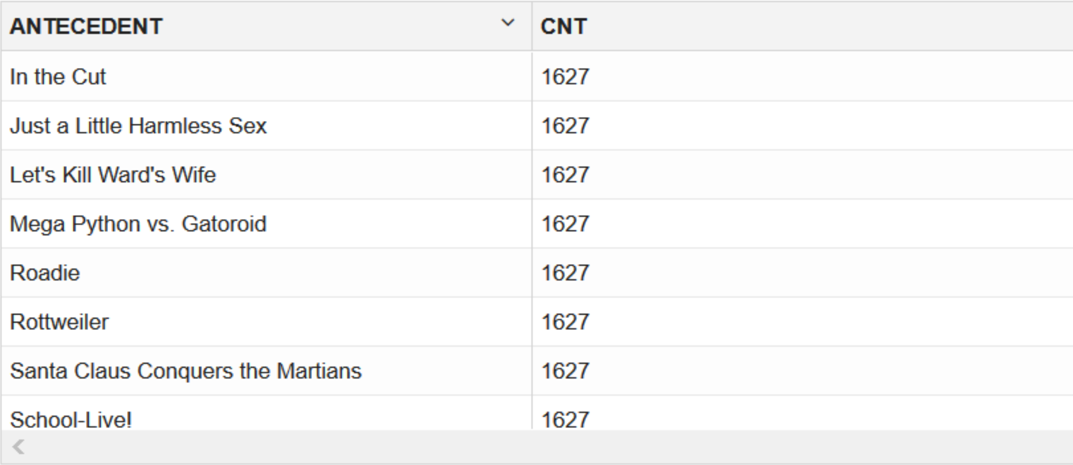
CONSEQUENT	CNT
Naked in New York	1627
The Harry Hill Movie	1627
Hits	1627
Fan Girl	1627
Clay Pigeons	1627
Amy's Orgasm	1627
Wish You Were Here	1627
Then She Found Me	1627



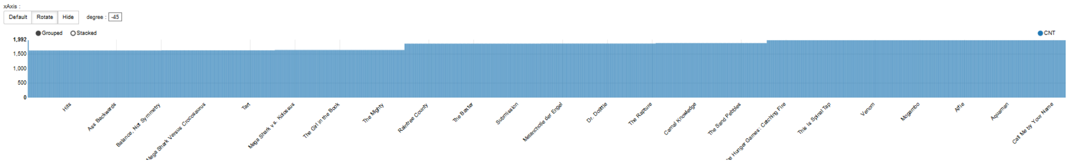
7. To view which antecedent items occur most frequently or which antecedent items are included in most rules, run the following script:

```
SELECT ANTECEDENT, COUNT(1) CNT
FROM
(SELECT ANTECEDENT_PREDICATE ANTECEDENT,
CONSEQUENT_PREDICATE CONSEQUENT,
RULE_SUPPORT SUPP, RULE_CONFIDENCE CONF, NUMBER_OF_ITEMS NUM
FROM DM$VAAR_MOVIES
ORDER BY RULE_CONFIDENCE DESC)
GROUP BY  ANTECEDENT
ORDER BY CNT
```

In OML Notebooks, click **settings** and click the **Bar Chart** icon to visualize the result. Click **Rotate** to rotate the bar graph to 45 degrees.



ANTECEDENT	CNT
In the Cut	1627
Just a Little Harmless Sex	1627
Let's Kill Ward's Wife	1627
Mega Python vs. Gatoroid	1627
Roadie	1627
Rottweiler	1627
Santa Claus Conquers the Martians	1627
School-Live!	1627



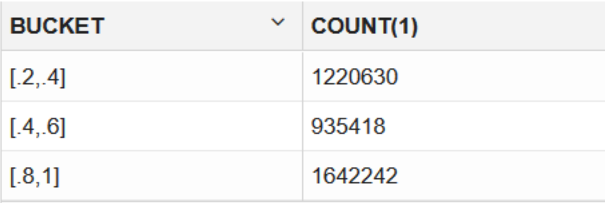
8. To check how many rules show up in each band of support, run the following query:

```
%sql
SELECT '['|| (SUPP_BIN -1)*0.2 ||','||SUPP_BIN*0.2||']' BUCKET, COUNT(1)
FROM (
SELECT ANTECEDENT_PREDICATE ANTECEDENT,
CONSEQUENT_PREDICATE CONSEQUENT,
RULE_SUPPORT SUPP, RULE_CONFIDENCE CONF, NUMBER_OF_ITEMS NUM,
WIDTH_BUCKET(RULE_SUPPORT, 0, 1, 4) SUPP_BIN
 FROM DM$VAAR_MOVIES ) a
GROUP BY SUPP_BIN
ORDER BY SUPP_BIN;
```

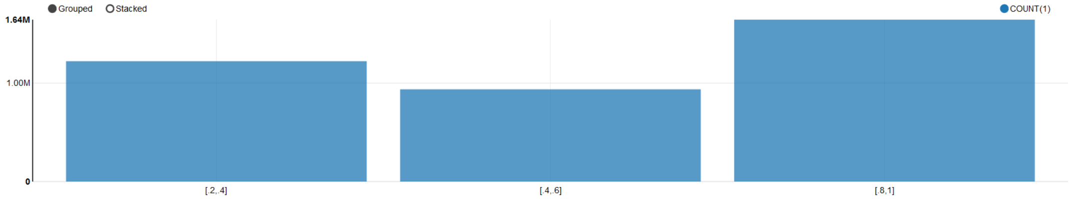
Examine the query:

- `SELECT '['|| (SUPP_BIN -1)*0.2 ||','||SUPP_BIN*0.2||']' BUCKET, COUNT(1)` creates the intervals for the buckets.
- The function `WIDTH_BUCKET` lets you construct equiwidth histograms, in which the histogram range is divided into intervals that have identical size. Here it produces buckets ranging from 0 to 1 and assigns number 1, …, 5, with identical size of 0.2. For instance the first bucket has the value = 1, for the range [0, 0.2].

In OML Notebooks, click **settings** and click the **Bar Chart** icon to visualize the result.

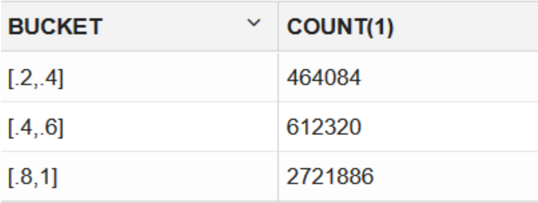


BUCKET	COUNT(1)
[.2,.4]	1220630
[.4,.6]	935418
[.8,1]	1642242



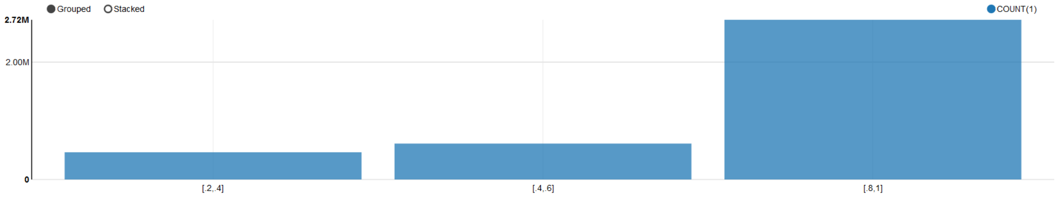
9. To check how many rules show up in each band of confidence, run the following query:

```
%sql
SELECT '['|| (CONF_BIN -1)*0.2 ||','||CONF_BIN*0.2||']' BUCKET, COUNT(1)
FROM (
SELECT ANTECEDENT_PREDICATE ANTECEDENT,
CONSEQUENT_PREDICATE CONSEQUENT,
RULE_SUPPORT SUPP, RULE_CONFIDENCE CONF, NUMBER_OF_ITEMS NUM,
WIDTH_BUCKET(RULE_CONFIDENCE, 0, 1, 4) CONF_BIN
 FROM DM$VAAR_MOVIES ) a
GROUP BY CONF_BIN
ORDER BY CONF_BIN;
```



BUCKET	COUNT(1)
[.2,.4]	464084
[.4,.6]	612320
[.8,1]	2721886

In OML Notebooks, click **settings** and click the **Bar Char**t icon to visualize the result.



10. To recommend top five movies based on customer's selection, use the `NUMBER_OF_ITEMS` and `EXTRACT` as predicate and query the Association Rules model detail view (`DM$VRxx`).

Association Rules support only a single consequent item.

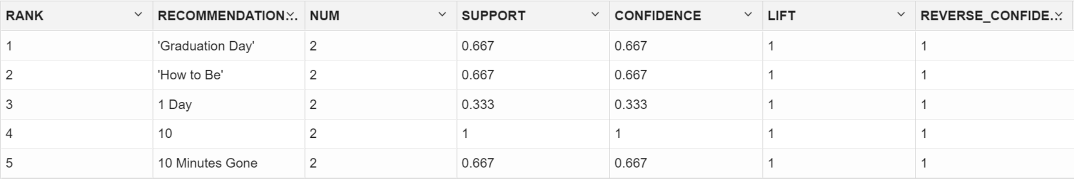
```
%sql

SELECT ROWNUM RANK,
 CONSEQUENT_NAME RECOMMENDATION,
 NUMBER_OF_ITEMS NUM,
 ROUND(RULE_SUPPORT, 3) SUPPORT,
 ROUND(RULE_CONFIDENCE, 3) CONFIDENCE,
 ROUND(RULE_LIFT, 3) LIFT,
 ROUND(RULE_REVCONFIDENCE, 3) REVERSE_CONFIDENCE
 FROM (SELECT * FROM DM$VRAR_MOVIES
 WHERE NUMBER_OF_ITEMS = 2
 AND EXTRACT(antecedent, '//item[item_name="101 Dalmatians"]') IS NOT NULL
 ORDER BY NUMBER_OF_ITEMS
 )
 WHERE ROWNUM <= 5;
```

Examine the query:

- `ROUND(RULE_LIFT, 3) LIFT`: The degree of improvement in the prediction over random chance when the rule is satisfied.
- `ROUND(RULE_REVCONFIDENCE, 3) REVERSE_CONFIDENCE`: The number of transactions in which the rule occurs divided by the number of transactions in which the consequent occurs rounded to 3 digits after the decimal.
- `NUMBER_OF_ITEMS`: Here, this parameter controls the size of the rule.

  > **Note:**
  >
  > In this use case, since you are looking for `ASSO_MAX_RULE_LENGTH` =2, you can skip this parameter.

- `EXTRACT`: Filters on the antecedent. If the antecedent must include “101 Dalmatians”, then use `extract(antecedent, '//item[item_name=“101 Dalmatians”]') IS NOT NULL`



RANK	RECOMMENDATION..	NUM	SUPPORT	CONFIDENCE	LIFT	REVERSE_CONFIDE..
1	'Graduation Day'	2	0.667	0.667	1	1
2	'How to Be'	2	0.667	0.667	1	1
3	1 Day	2	0.333	0.333	1	1
4	10	2	1	1	1	1
5	10 Minutes Gone	2	0.667	0.667	1	1

In this step, if the customer's cart has 101 Dalmatians movie, the customer is 66.7% likely to rent or buy *Graduation Day*, *How to Be*, and *10 Minutes Gone* and there are 100% chances that they will buy *10*.

To conclude, you have successfully examined association rules and provided top movie recommendations to customers based on their frequently purchased and/or rented movies.

Feature Extraction Use Case Scenario

You are developing a software application to recognize handwritten digits, which can be used to scan student answer sheets or forms. You are using the feature extraction technique to reduce the dimensionality of the dataset to produce a new feature space. This feature space concentrates the signal of the original data as linear combinations of the original data.

In other scenarios, feature extraction can be used to extract document themes, classify features, and so on.

The reduced features can be used with other machine learning algorithms, for example, classification and clustering algorithms.

In this use case, you'll use the neural network algorithm to recognize handwritten digits on the transformed space and contrast this with the accuracy using the original data.

You are using the default feature extraction algorithm Non-Negative Matrix Factorization (NMF) in two ways:

1. Creating the Projections of the Top 16 Features, and feeding a Neural Networks (NN) model with those features
2. Using the correlation between the various attributes and the Top 6 feature vectors to do an Attribute Selection manually and then feeding a Neural Network model with those features.

You are creating Neural Networks models to try to predict the correct handwritten digits based on an 8x8 image matrix (64 input attributes).

Related Content

Topic	Link
OML4SQL GitHub Example	Feature Extraction - Non-Negative Matrix Factorization
`CREATE_MODEL2` Procedure	CREATE_MODEL2 Procedure
Generic Model Settings	DBMS_DATA_MINING — Model Settings
Non-negative Matrix Factorization (NMF) Settings	DBMS_DATA_MINING - Algorithm Settings: Non-Negative Matrix Factorization
Neural Network Settings	DBMS_DATA_MINING — Algorithm Settings: Neural Network
Data Dictionary Settings	Oracle Machine Learning Data Dictionary Views
NMF - Model Detail Views	Model Detail Views for Non-Negative Matrix Factorization
Neural Network - Model Detail Views	Model Detail Views for Neural Network
About Feature Extraction	Feature Extraction
About NMF	Non-Negative Matrix Factorization
About Classification	Classification
About Neural Network	Neural Network

Before you start your OML4SQL use case journey, ensure that you have the following:

- Data Set
  You are using the DIGITS data set from the Scikit library. In this example, a DDL script is used to create and load the DIGITS table into the database. You can also download the raw data set here: https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/datasets/data/digits.csv.gz
- Database
  Select or create database out of the following options:
  - Get your FREE cloud account. Go to https://cloud.oracle.com/database and select Oracle Database Cloud Service (DBCS), or Oracle Autonomous Database. Create an account and create an instance. See Autonomous Database Quick Start Workshop.
  - Download the latest version of Oracle Database (on premises).
- Machine Learning Tools
  Depending on your database selection,
  - Use OML Notebooks for Oracle Autonomous Database.
  - Install and use Oracle SQL Developer connected to an on-premises database or DBCS. See Installing and Getting Started with SQL Developer.
- Other Requirements

Data Mining Privileges (this is automatically set for ADW). See System Privileges for Oracle Machine Learning for SQL.

Related Topics

- Create a Notebook
- Edit your Notebook
- Installing Sample Schemas

Load Data

Create a table called DIGITS. This table is used to access the data set.

Perform these steps to load the data into your database.

1. Download the DDL script, https://objectstorage.us-ashburn-1.oraclecloud.com/n/adwc4pm/b/OML_Data/o/digits.sql on your system.
2. Open the file with a text editor and remove OML_USER02. in all instances of OML_USER02.DIGITS.
3. Save the file.
4. Oracle recommends using Oracle SQL Developer with an on-premises Database or Cloud Database connection for loading the DIGITS data as the size of the digits.sql is too large for a notebook paragraph.

To understand the data, you will perform the following:

- Access the data.
- Examine the various attributes or columns of the data set.
- Assess data quality (by exploring the data).

Examine Data

Digits data set has 64 numerical features or columns (8x8 pixel images). Each image is of a hand-written digit. The digits 0-9 are used in this data set.

Explore Data

Once the data is accessible, explore the data to understand and assess the quality of the data. .

Assess Data Quality

Because this is a well-curated data set, it is free of noise, missing values (systemic or random), and outlier numeric values.

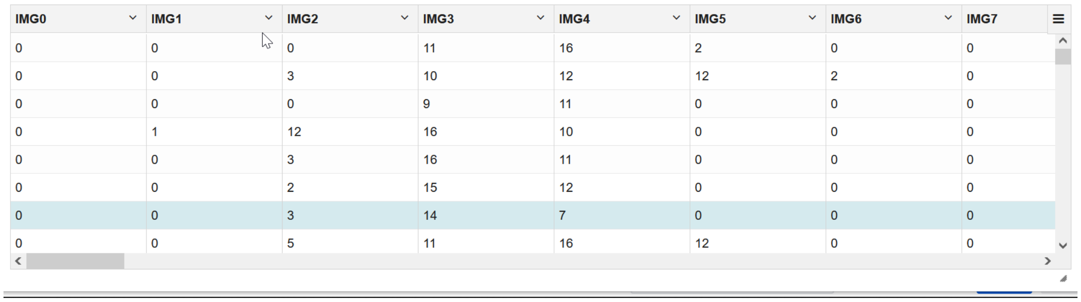
The following steps help you with the exploratory analysis of the data:

1. View the data in the DIGITS data by running the following statement:

   ```
   SELECT * FROM DIGITS;
   ```

2. To see distinct data from the table, run the following query:

   ```
   SELECT DISTINCT * FROM DIGITS;
   ```



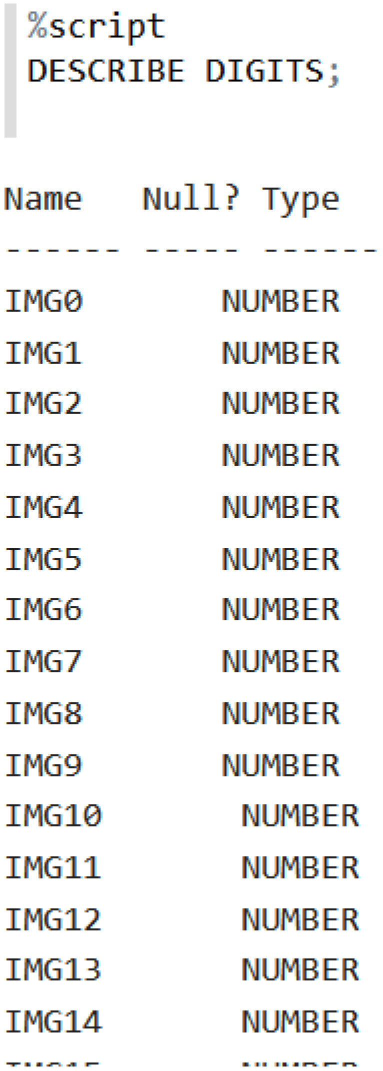
IMG0	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	IMG7
0	0	0	11	16	2	0	0
0	0	3	10	12	12	2	0
0	0	0	9	11	0	0	0
0	1	12	16	10	0	0	0
0	0	3	16	11	0	0	0
0	0	2	15	12	0	0	0
0	0	3	14	7	0	0	0
0	0	5	11	16	12	0	0

3. Find the `COUNT` of rows in the data set, run the following statement:

```
SELECT COUNT(*) from DIGITS;


COUNT(*)
     1797
---------------------------
```

4. To view the data type of the columns, run the following statement:

```
%script
DESCRIBE DIGITS;
```



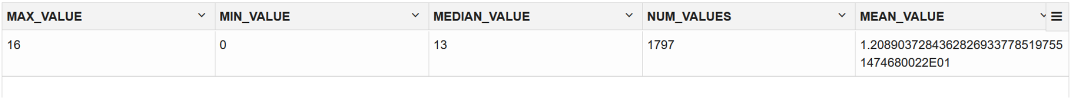
```
%script
DESCRIBE DIGITS;
```

```
Name   Null? Type
------ ----- ------
IMG0         NUMBER
IMG1         NUMBER
IMG2         NUMBER
IMG3         NUMBER
IMG4         NUMBER
IMG5         NUMBER
IMG6         NUMBER
IMG7         NUMBER
IMG8         NUMBER
IMG9         NUMBER
IMG10         NUMBER
IMG11         NUMBER
IMG12         NUMBER
IMG13         NUMBER
IMG14         NUMBER
```

5. This SQL query will select the maximum, minimum, median, count, and mean of the "IMG59" column in the `DIGITS` table.

   The median value is calculated using the `PERCENTILE_CONT` function, which takes the percentile (in this case, the 50th percentile or median) as an argument.

```
SELECT
    MAX(IMG59) as max_value,
    MIN(IMG59) as min_value,
    PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY IMG59) as median_value,
    COUNT(*) as num_values,
    AVG(IMG59) as mean_value
FROM DIGITS;
```

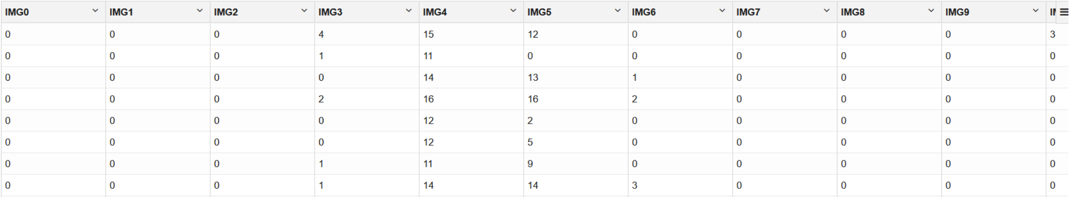


MAX_VALUE	MIN_VALUE	MEDIAN_VALUE	NUM_VALUES	MEAN_VALUE
16	0	13	1797	1.20890372843628269337785197551474680022E01

6. One way to find outliers is you can use the above query to calculate the mean and standard deviation of the data set, and then use those values to identify values that are

outside of a certain number of standard deviations from the mean. Here, you are checking for one column - IMG59 mean and standard deviation.

```
WITH outliers AS (
  SELECT
    AVG(IMG59) AS avg,
    STDDEV(IMG59) AS stddev
  FROM DIGITS
)
SELECT *
FROM DIGITS a
CROSS JOIN outliers b
WHERE ABS(a.IMG59 - b.avg) > 2 * b.stddev;
```



IMG0	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	IMG7	IMG8	IMG9	II
0	0	0	4	15	12	0	0	0	0	3
0	0	0	1	11	0	0	0	0	0	0
0	0	0	0	14	13	1	0	0	0	0
0	0	0	2	16	16	2	0	0	0	0
0	0	0	0	12	2	0	0	0	0	0
0	0	0	0	12	5	0	0	0	0	0
0	0	0	1	11	9	0	0	0	0	0
0	0	0	1	14	14	3	0	0	0	0

This completes the data exploration stage. OML supports Automatic Data Preparation (ADP). ADP is enabled through the model settings. When ADP is enabled, the transformations required by the algorithm are performed automatically and embedded in the model. This step is done during the Build Model stage. The commonly used methods of data preparation are binning, normalization, and missing value treatment.

Related Topics

- How ADP Transforms the Data

Build Model

Build your model using your data set. Use the `DBMS_DATA_MINING.CREATE_MODEL2` procedure to build your model and specify the model settings.

Algorithm Selection

You can choose one of the following algorithms to solve a Feature Extraction problem:

- Explicit Semantic Analysis (ESA) - this algorithm is not applicable for this use case data set.
- Non-Negative Matrix Factorization (NMF)
- Singular Value Decomposition (SVG)

Non-Negative matrix factorization (NMF) is now a popular tool for analyzing high-dimensional data because it automatically extracts sparse (missing values with mostly zero; many cells or pixels in this data set likely have zeros, however they are not truly missing) and meaningful features from a set of non-negative data vectors. NMF uses a low-rank matrix approximation to approximate a matrix **X** such that **X** is approximately equal to **WH**. The sub-matrix **W** contains the NMF basis column vectors; the sub-matrix **H** contains the associated coefficients (weights). The ability of NMF to automatically extract sparse and easily interpretable factors has led to its popularity. In the case of image recognition, such as digit images, the base images depict various handwritten digit prototypes and the columns of **H** indicate which feature is present in which image. Oracle Machine Learning uses NMF as the default algorithm for Feature Extraction.

For this use case, split the data into 60/40 as training and test data to further use it to compare the NMF model with that of another model using Neural Network (NN). You are splitting the data because you want to see how the model performs on data that you haven't seen before. If you put the whole data set into the original NMF model and then split it before giving it to NN, the NMF model has already seen the data when you try to test it. When we have completely new data, the extract features will not be based on it. You build the model using the training data and once the model is built, score the test data using the model.

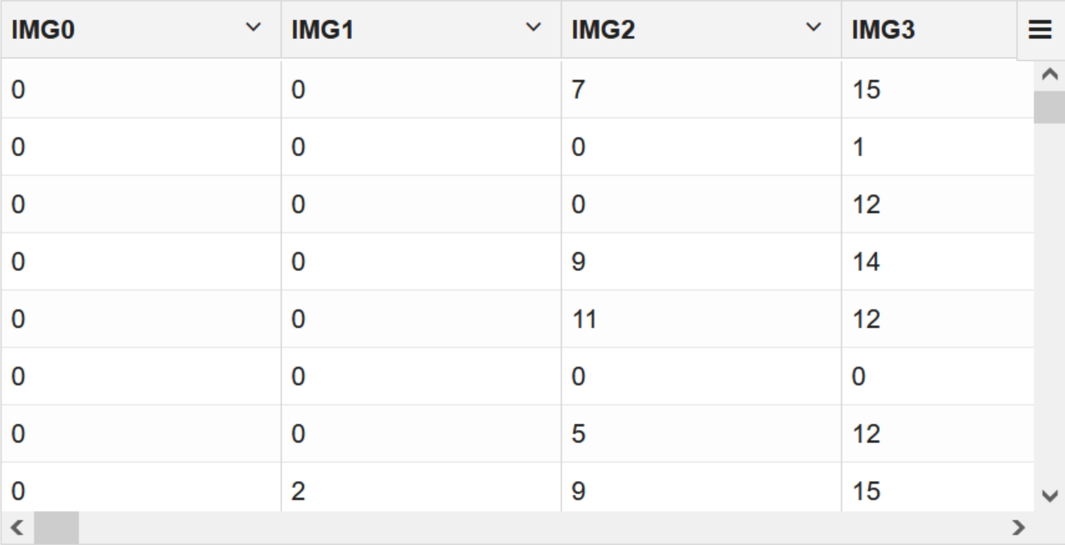
The following steps guide you to build your model with the selected algorithm.

1. To create the training and test data with 60/40 split, run the following statement:

   ```
   %script
   CREATE OR REPLACE VIEW TRAIN_DIGITS AS SELECT * FROM DIGITS SAMPLE (60)
   SEED (1);
   CREATE OR REPLACE VIEW TEST_DIGITS AS SELECT * FROM DIGITS MINUS SELECT *
   FROM TRAIN_DIGITS;



   View TRAIN_DIGITS created.

   ---------------------------

   View TEST_DIGITS created.
   ```

2. To view the data in the `TRAIN_DIGITS` view, run the following statement:

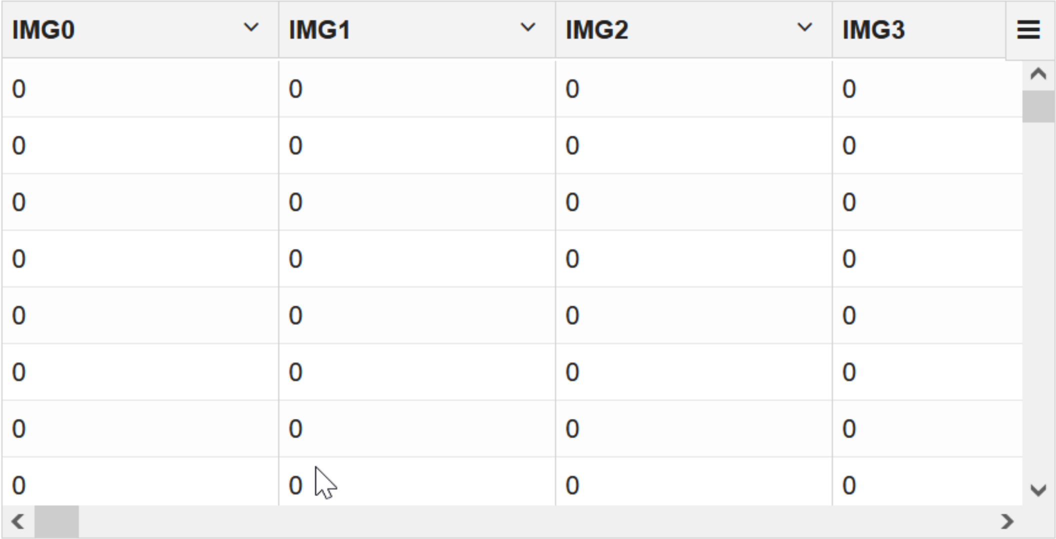
   ```
   %sql

   SELECT * FROM TRAIN_DIGITS;
   ```



IMG0	IMG1	IMG2	IMG3
0	0	7	15
0	0	0	1
0	0	0	12
0	0	9	14
0	0	11	12
0	0	0	0
0	0	5	12
0	2	9	15

3. To view the data in the `TEST_DIGITS` view, run the following statement:

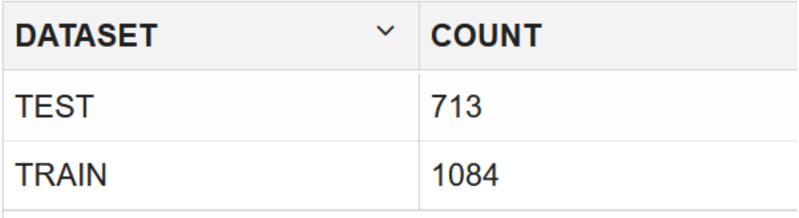
```
%sql

SELECT * FROM TEST_DIGITS;
```



IMG0	IMG1	IMG2	IMG3
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

4. To find the count of rows in `TRAIN_DIGITS` and `TEST_DIGITS`, run the following statement:

```
%sql
select 'TRAIN' dataset, count(*)  count from TRAIN_DIGITS
union
select 'TEST' dataset, count(*) count from TEST_DIGITS;
```



DATASET	COUNT
TEST	713
TRAIN	1084

5. Build your model using the `CREATE_MODEL2` procedure. First, declare a variable to store model settings or hyperparameters. Run the following script:

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('NMF_DIGITS');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlist DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlist('FEAT_NUM_FEATURES') := '16';
    v_setlist('PREP_AUTO')        := 'ON';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'NMF_DIGITS',
        MINING_FUNCTION     => 'FEATURE_EXTRACTION',
        DATA_QUERY          => 'SELECT * FROM DIGITS',
        SET_LIST            => v_setlist);
        CASE_ID_COLUMN_NAME => '"target"');
END;
/
```

```
PL/SQL procedure successfully completed.
---------------------------
PL/SQL procedure successfully completed.
```

Examine the script:

- `v_setlist` is a variable to store `SETTING_LIST`.
- `SETTING_LIST` specifies model settings or hyperparameters for your model.
- `DBMS_DATA_MINING` is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- `PREP_AUTO` is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is `ON`.
- `FEAT_NUM_FEATURES` is the number of features you want to extract by using the feature extraction model. In this use case, 16 features are used for illustrative purposes.

The `CREATE_MODEL2` procedure takes the following parameters:

- `MODEL_NAME`: A unique model name that you will give to your model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is `NMF_DIGITS`.
- `MINING_FUNCTION`: Specifies the machine learning function or mining technique. Since in this use case you are performing feature extraction or dimensionality reduction, select the mining function as `FEATURE_EXTRACTION`.
- `DATA_QUERY`: A query that provides training data for building the model. Here, the query is `SELECT * FROM DIGITS`.
- `SET_LIST`: Specifies `SETTING_LIST`.
- `CASE_ID_COLUMN_NAME`: A unique case identifier column in the build data. In this use case, case_id is `CUST_ID`. If there is a composite key, you must create a new attribute before creating the model. This may involve concatenating values from the columns, or mapping a unique identifier to each distinct combination of values. The `CASE_ID` assists with reproducible results, joining scores for individual rows with other data in, example, scoring data table. Here, since it an unsupervised technique, target is unknown usually, here it is not required. So, to exclude it from processing, you are declaring it as case_id. Note that `target` column appears in small case. That is why you must have it as `"target"`.

Note:

Any parameters or settings not specified are either system-determined or default values are used.

Evaluate

Evaluate your model by viewing diagnostic metrics and performing quality checks.

There is no specific set of testing parameters for feature extraction. In this use case, the evaluation mostly consists of comparing the NN models with and among the NMF models.

Dictionary and Model Views

To obtain information about the model and view model settings, you can query data dictionary views and model detail views. Specific views in model detail views display model statistics which can help you evaluate the model.

You'll be querying dictionary views. A database administrator (DBA) and USER versions of the views are also available. See Oracle Machine Learning Data Dictionary Views to learn more about the available dictionary views. Model detail views are specific to the algorithm. You can obtain more insights about the model you created by viewing the model detail views. The names of model detail views begin with DM$xx where xx corresponds to the view prefix. See Model Detail Views for more information.

The following steps help you to view different dictionary views and model detail views.

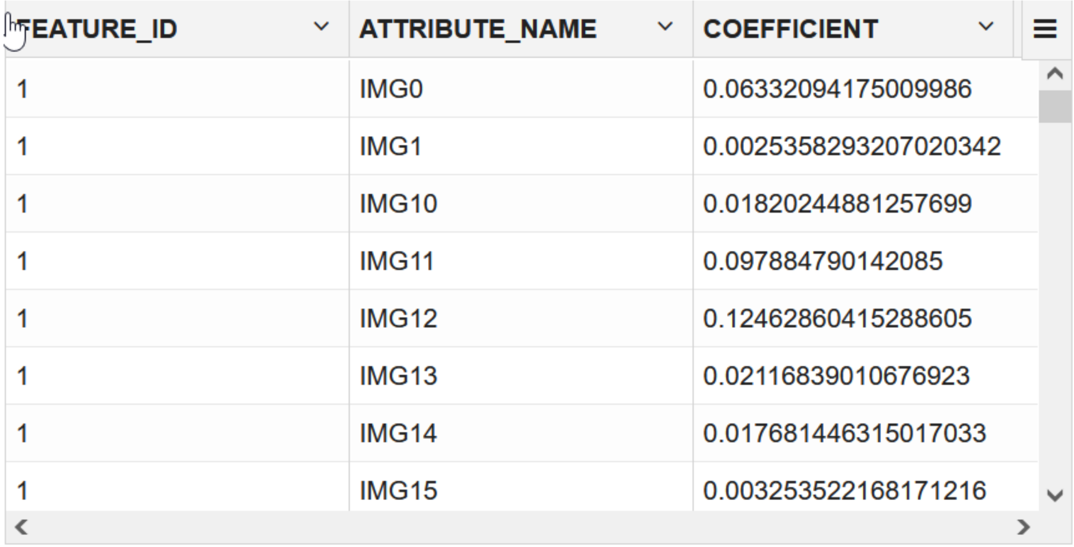
1. Run the following statement to view the settings in `USER_MINING_MODEL_SETTINGS`:

```
%script

SELECT SETTING_NAME, SETTING_VALUE
  FROM USER_MINING_MODEL_SETTINGS
  WHERE MODEL_NAME='NMF_DIGITS'
  ORDER BY SETTING_NAME;




SETTING_NAME                   SETTING_VALUE
ALGO_NAME                      ALGO_NONNEGATIVE_MATRIX_FACTOR
FEAT_NUM_FEATURES              16
NMFS_CONV_TOLERANCE            .05
NMFS_NONNEGATIVE_SCORING       NMFS_NONNEG_SCORING_ENABLE
NMFS_NUM_ITERATIONS            50
NMFS_RANDOM_SEED               -1
ODMS_DETAILS                   ODMS_ENABLE
ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE
PREP_AUTO                      ON


10 rows selected.


---------------------------
```

2. The matrix of attribute values, or pixel values in this use case, is factored into two sub-matrices. Say the factorization is represented by the product of sub-matrices W and H, as

WH. The sub-matrix H contains the coefficients (or weights) of the column vectors of sub-matrix W. To query the Non-Negative Matrix Factorization H Matrix, use the `DM$VENMF` view.

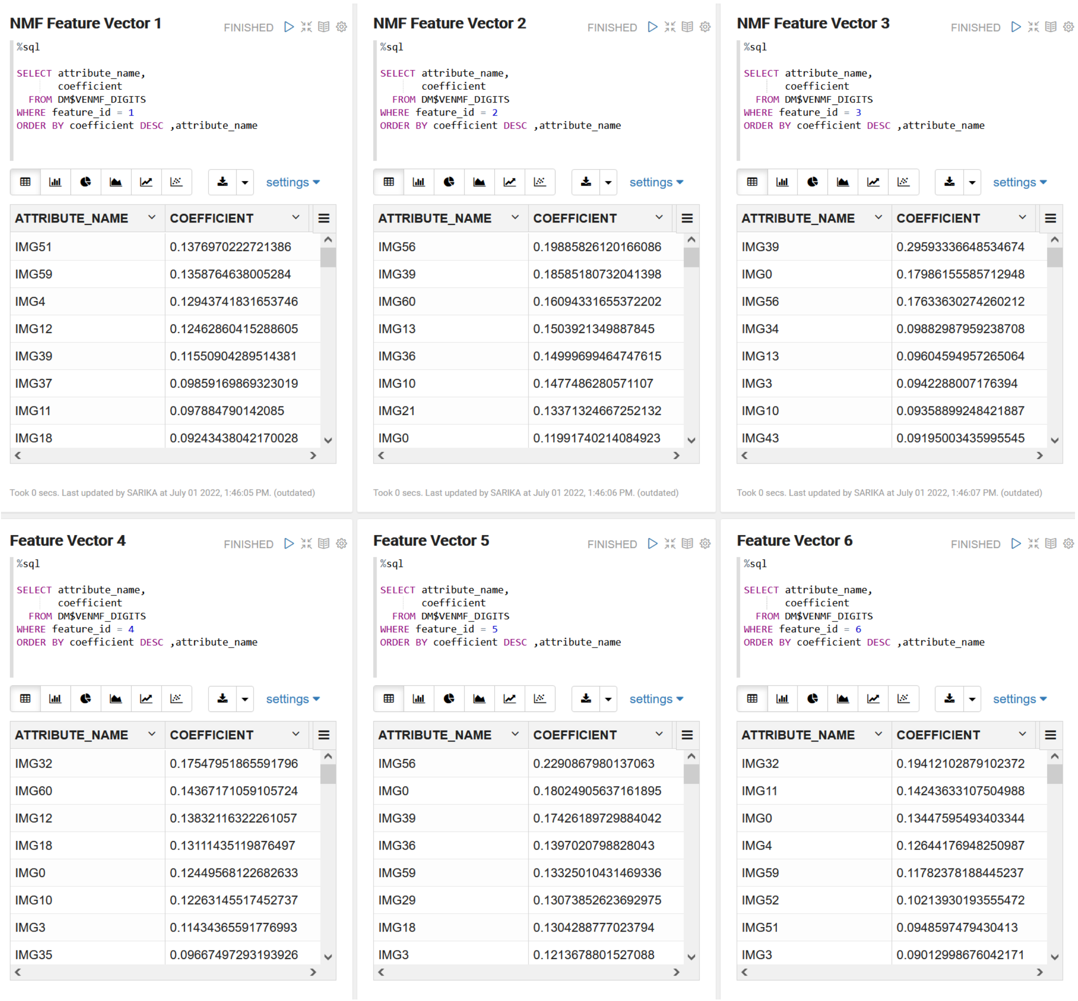
```
%sql
SELECT FEATURE_ID, ATTRIBUTE_NAME,
ATTRIBUTE_VALUE, COEFFICIENT
FROM DM$VENMF_DIGITS
ORDER BY FEATURE_ID, ATTRIBUTE_NAME;
```



FEATURE_ID	ATTRIBUTE_NAME	COEFFICIENT
1	IMG0	0.06332094175009986
1	IMG1	0.0025358293207020342
1	IMG10	0.01820244881257699
1	IMG11	0.097884790142085
1	IMG12	0.12462860415288605
1	IMG13	0.02116839010676923
1	IMG14	0.017681446315017033
1	IMG15	0.003253522168171216

3. Now, to understand the relationship between the original attribute set and the feature vectors, use the `DM$VENMF` view for each NMF feature vector. Each feature is a linear combination of the original attribute set. The coefficients of these linear combinations are non-negative. The model details return for each feature the coefficients associated with each one of the original attributes. This gives an idea of how the attributes are contributing to constructing each feature vector. For example, to view the attributes and coefficients for feature vector 1, use the `WHERE` clause and an `ORDER BY` clause in the query. Similarly, examine the attributes and their coefficients for feature vectors 2, 3, 4, 5, and 6 by changing the `WHERE` clause.

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 1
ORDER BY coefficient DESC ,attribute_name
```



NMF Feature Vector 1

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 1
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG51	0.1376970222721386
IMG59	0.1358764638005284
IMG4	0.12943741831653746
IMG12	0.12462860415288605
IMG39	0.11550904289514381
IMG37	0.09859169869323019
IMG11	0.097884790142085
IMG18	0.09243438042170028

Took 0 secs. Last updated by SARIKA at July 01 2022, 1:46:05 PM. (outdated)

NMF Feature Vector 2

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 2
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG56	0.19885826120166086
IMG39	0.18585180732041398
IMG60	0.16094331655372202
IMG13	0.1503921349887845
IMG36	0.14999699464747615
IMG10	0.147748628057107
IMG21	0.13371324667252132
IMG0	0.11991740214084923

Took 0 secs. Last updated by SARIKA at July 01 2022, 1:46:06 PM. (outdated)

NMF Feature Vector 3

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 3
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG39	0.29593336648534674
IMG0	0.17986155585712948
IMG56	0.17633630274260212
IMG34	0.09882987959238708
IMG13	0.09604594957265064
IMG3	0.0942288007176394
IMG10	0.09358899248421887
IMG43	0.09195003435995545

Took 0 secs. Last updated by SARIKA at July 01 2022, 1:46:07 PM. (outdated)

Feature Vector 4

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 4
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG32	0.17547951865591796
IMG60	0.14367171059105724
IMG12	0.13832116322261057
IMG18	0.13111435119876497
IMG0	0.12449568122682633
IMG10	0.12263145517452737
IMG3	0.11434365591776993
IMG35	0.09667497293193926

Feature Vector 5

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 5
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG56	0.2290867980137063
IMG0	0.18024905637161895
IMG39	0.17426189729884042
IMG36	0.1397020798828043
IMG59	0.13325010431469336
IMG29	0.13073852623692975
IMG18	0.1304288777023794
IMG3	0.1213678801527088

Feature Vector 6

```
%sql

SELECT attribute_name,
       coefficient
  FROM DM$VENMF_DIGITS
WHERE feature_id = 6
ORDER BY coefficient DESC ,attribute_name
```

ATTRIBUTE_NAME	COEFFICIENT
IMG32	0.19412102879102372
IMG11	0.14243633107504988
IMG0	0.13447595493403344
IMG4	0.12644176948250987
IMG59	0.11782378188445237
IMG52	0.10213930193555472
IMG51	0.0948597479430413
IMG3	0.09012998676042171

4. Now, create another model using the Neural Network algorithm. You are first building a model with the original feature set (`TRAIN_DIGITS`) for comparison purposes before the extracted features are used.

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('NN_ORIG_DIGITS');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;

BEGIN
    v_setlst('PREP_AUTO') := 'ON';
    v_setlst('ALGO_NAME') := 'ALGO_NEURAL_NETWORK';
    v_setlst('NNET_NODES_PER_LAYER') := '40';
    v_setlst('NNET_ACTIVATIONS') := '''NNET_ACTIVATIONS_TANH''';
    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'NN_ORIG_DIGITS',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'SELECT * FROM TRAIN_DIGITS',
        SET_LIST            => v_setlst,
        TARGET_COLUMN_NAME  => '"target"');
END;
/
```

Examine the script:

- v_setlist is a variable to store SETTING_LIST.
- SETTING_LIST specifies model settings or hyperparameters for your model.
- DBMS_DATA_MINING is the PL/SQL package used for machine learning. These settings are described in DBMS_DATA_MINING - Model Settings.
- PREP_AUTO is the setting used for Automatic Data Preparation. Here, enable Automatic Data Preparation. The value of the setting is ON.
- ALGO_NAME specifies the algorithm name. Since you are using the Neural Network as your algorithm, set ALGO_NEURAL_NETWORK.
- NET_NODES_PER_LAYRER defines the topology by the number of nodes per layer. Different layers can have different numbers of nodes. To specify the same number of nodes for each layer, you can provide a single value, which is then applied to each layer. The default number of nodes per layer is the number of attributes or 50 (if the number of attributes > 50). In this use case, the defined value is 40, used for illustrative purposes.
- NNET_ACTIVATIONS specifies the activation functions for the hidden layers. The activation function determines the output of neural networks. The activation functions map the outputs from a previous layer to values between 0 to 1 or -1 to 1, and so on, depending on the activation function applied. You can specify a single activation function, which is then applied to each hidden layer, or you can specify an activation function for each layer individually. See DBMS_DATA_MINING - Algorithm Settings: Neural Network to learn more about Neural Network settings. Different layers can have different activation functions. Here, you are using NNET_ACTIVATIONS_TANH. The range of the tanh function is from -1 to 1. The default value is the sigmoid function NNET_ACTIVATIONS_LOG_SIG, which has the range 0 to 1.

The CREATE_MODEL2 procedure takes the following parameters:

- MODEL_NAME: A unique model name that you will give to your model. The name of the model is in the form [schema_name.]model_name. If you do not specify a schema, then your own schema is used. Here, the model name is NN_ORIG_DIGITS.
- MINING_FUNCTION: Specifies the machine learning function or mining technique. Since in this Since it is a Classification problem, for this model you selected CLASSIFICATION.
- DATA_QUERY: A query that provides training data for building the model. Here, the query is SELECT * FROM TRAIN_DIGITS.
- SET_LIST: Specifies SETTING_LIST.
- TARGET_COLUMN_NAME: Specifies the column that needs to be predicted. Also referred to as the target variable of the model. In this use case, you are predicting the "target" value.

> **Note:**
>
> Any parameters or settings not specified are either system-determined or default values are used.

Evaluate and Compare Models

You used the NMF feature extraction algorithm to transform your data and feed it into a NN model with the intention of improving predictive accuracy. You will now evaluate these classification models and compare the model accuracy.

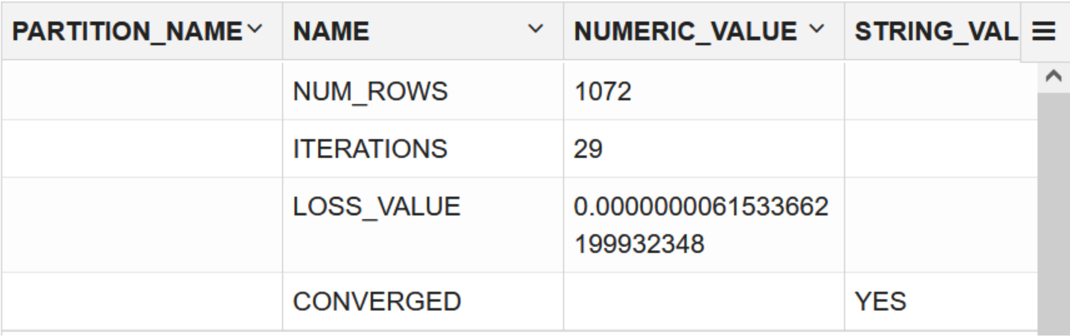
You built a neural network classification model using the original data set, and then you used the data set transformed by the NMF model to build another neural network model. You will evaluate these classification models. When comparing these metrics, consider the model's quality by looking at prediction accuracy. Metrics can also be compared between models created by different feature extraction algorithms, the same extraction algorithm with different settings, or different classification algorithms and settings.

1. To check if the model is converged, view the model views of the Neural Network model.

```
%script
  SELECT VIEW_NAME, VIEW_TYPE
  FROM USER_MINING_MODEL_VIEWS
  WHERE MODEL_NAME='NN_ORIG_DIGITS'
  ORDER BY VIEW_NAME;



VIEW_NAME              VIEW_TYPE
DM$VANN_ORIG_DIGITS    Neural Network Weights
DM$VCNN_ORIG_DIGITS    Scoring Cost Matrix
DM$VGNN_ORIG_DIGITS    Global Name-Value Pairs
DM$VNNN_ORIG_DIGITS    Normalization and Missing Value Handling
DM$VSNN_ORIG_DIGITS    Computed Settings
DM$VTNN_ORIG_DIGITS    Classification Targets
DM$VWNN_ORIG_DIGITS    Model Build Alerts


7 rows selected.


---------------------------
```

2. Display the view `DM$VGNN_ORIG_DIGITS` to check the global name-value pairs to see if the model is converged which means that the model iterates until it reaches a point where no improvement in the result could be seen.

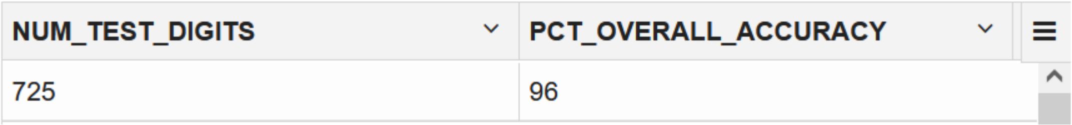
```
%sql

SELECT * from DM$VGNN_ORIG_DIGITS;
```



PARTITION_NAME	NAME	NUMERIC_VALUE	STRING_VAL
	NUM_ROWS	1072	
	ITERATIONS	29	
	LOSS_VALUE	0.0000000061533662 199932348	
	CONVERGED		YES

3. To check the quality of the model, run the following PCT accuracy code:

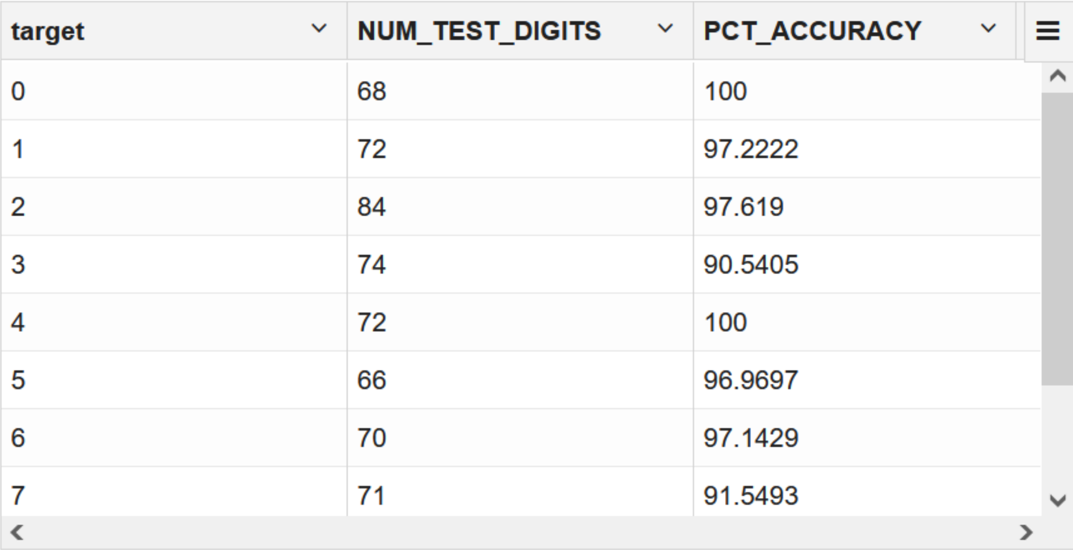
```
%sql

SELECT count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_OVERALL_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_ORIG_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_DIGITS)
```



NUM_TEST_DIGITS	PCT_OVERALL_ACCURACY
725	96

4. To check the quality of the model per target digit, run the following PCT accuracy code:

```
%sql

SELECT "target",
       count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_ORIG_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_DIGITS)
        GROUP BY "target"
        ORDER BY "target"
```



target	NUM_TEST_DIGITS	PCT_ACCURACY
0	68	100
1	72	97.2222
2	84	97.619
3	74	90.5405
4	72	100
5	66	96.9697
6	70	97.1429
7	71	91.5493

5. To evaluate your model, use the following SQL `PREDICTION` function to generate a Confusion Matrix:

```
%script
SELECT "target" AS actual_target_value,
       PREDICTION(NN_ORIG_DIGITS USING *) AS predicted_target_value,
       COUNT(*) AS value
  FROM TEST_DIGITS
 GROUP BY "target", PREDICTION(NN_ORIG_DIGITS USING *)
 ORDER BY 1, 2;
```

```
ACTUAL_TARGET_VALUE   PREDICTED_TARGET_VALUE   VALUE
                  0                        0      81
                  1                        1      83
                  1                        8       1
                  2                        1       1
                  2                        2      72
                  3                        2       1
                  3                        3      68
                  3                        5       1
                  4                        4      65
                  4                        7       1
                  5                        5      63
                  5                        6       1
                  5                        7       1
                  5                        9       1

ACTUAL_TARGET_VALUE   PREDICTED_TARGET_VALUE   VALUE
                  6                        5       2
                  6                        6      64
                  7                        7      63
                  7                        9       1
                  8                        1       1
                  8                        2       1
                  8                        7       1
```

```
                         8                            8        55
                         8                            9         1
                         9                            3         1
                         9                            5         1
                         9                            8         1
                         9                            9        81


27 rows selected.


---------------------------
```

Score

Scoring an NMF model produces data projections in the new feature space. The magnitude of a projection indicates how strongly a record maps to a feature.

In this use case, there are two options to produce projections to compare using the Neural Network model. One is to build a neural network model on the top 16 features of NMF to predict the digits. Another is to build a neural network model by manually selecting the original attributes with the highest coefficients for each feature vector.

For option one, create a new `TRAIN` and `TEST` view and then build a NN model.

1. Create a new `TRAIN` view with the top 16 NMF features.

   ```
   %sql

   CREATE OR REPLACE VIEW TRAIN_NMF_FEATURES AS
   SELECT "target",
          FEATURE_VALUE(NMF_DIGITS, 1 USING *) PROJ1,
          FEATURE_VALUE(NMF_DIGITS, 2 USING *) PROJ2,
          FEATURE_VALUE(NMF_DIGITS, 3 USING *) PROJ3,
          FEATURE_VALUE(NMF_DIGITS, 4 USING *) PROJ4,
          FEATURE_VALUE(NMF_DIGITS, 5 USING *) PROJ5,
          FEATURE_VALUE(NMF_DIGITS, 6 USING *) PROJ6,
          FEATURE_VALUE(NMF_DIGITS, 7 USING *) PROJ7,
          FEATURE_VALUE(NMF_DIGITS, 8 USING *) PROJ8,
          FEATURE_VALUE(NMF_DIGITS, 9 USING *) PROJ9,
          FEATURE_VALUE(NMF_DIGITS, 10 USING *) PROJ10,
          FEATURE_VALUE(NMF_DIGITS, 11 USING *) PROJ11,
          FEATURE_VALUE(NMF_DIGITS, 12 USING *) PROJ12,
          FEATURE_VALUE(NMF_DIGITS, 13 USING *) PROJ13,
          FEATURE_VALUE(NMF_DIGITS, 14 USING *) PROJ14,
          FEATURE_VALUE(NMF_DIGITS, 15 USING *) PROJ15,
          FEATURE_VALUE(NMF_DIGITS, 16 USING *) PROJ16
   FROM TRAIN_DIGITS;


   TRAIN_DIGITS view created
   ---------------------------
   ```

2. Create a new `TEST` view with the top 16 NMF features.

   ```
   %sql

   CREATE OR REPLACE VIEW TEST_NMF_FEATURES AS
   SELECT "target",
   ```

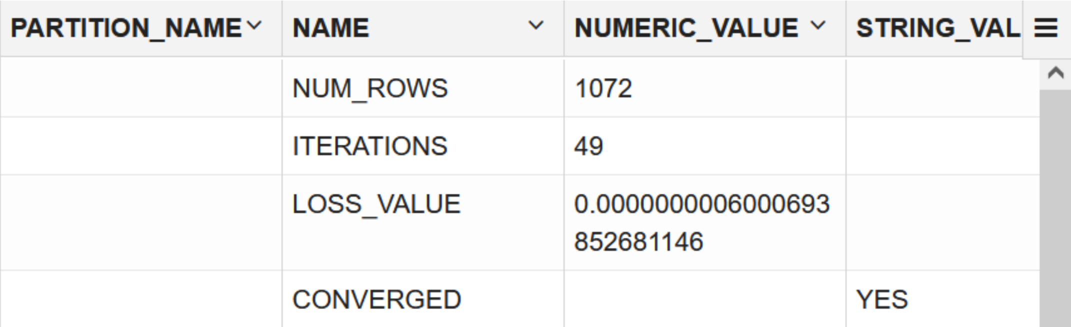
```
       FEATURE_VALUE(NMF_DIGITS, 1 USING *) PROJ1,
       FEATURE_VALUE(NMF_DIGITS, 2 USING *) PROJ2,
       FEATURE_VALUE(NMF_DIGITS, 3 USING *) PROJ3,
       FEATURE_VALUE(NMF_DIGITS, 4 USING *) PROJ4,
       FEATURE_VALUE(NMF_DIGITS, 5 USING *) PROJ5,
       FEATURE_VALUE(NMF_DIGITS, 6 USING *) PROJ6,
       FEATURE_VALUE(NMF_DIGITS, 7 USING *) PROJ7,
       FEATURE_VALUE(NMF_DIGITS, 8 USING *) PROJ8,
       FEATURE_VALUE(NMF_DIGITS, 9 USING *) PROJ9,
       FEATURE_VALUE(NMF_DIGITS, 10 USING *) PROJ10,
       FEATURE_VALUE(NMF_DIGITS, 11 USING *) PROJ11,
       FEATURE_VALUE(NMF_DIGITS, 12 USING *) PROJ12,
       FEATURE_VALUE(NMF_DIGITS, 13 USING *) PROJ13,
       FEATURE_VALUE(NMF_DIGITS, 14 USING *) PROJ14,
       FEATURE_VALUE(NMF_DIGITS, 15 USING *) PROJ15,
       FEATURE_VALUE(NMF_DIGITS, 16 USING *) PROJ16
FROM TEST_DIGITS;



TEST_DIGITS view created
---------------------------
```

3. Build a Neural Network model using the new `TRAIN` data set.

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('NN_NMF_DIGITS');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;

BEGIN
    v_setlst('PREP_AUTO') := 'ON';
    v_setlst('ALGO_NAME') := 'ALGO_NEURAL_NETWORK';
    v_setlst('NNET_NODES_PER_LAYER') := '40';
    v_setlst('NNET_ACTIVATIONS') := '''NNET_ACTIVATIONS_TANH''';
    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'NN_NMF_DIGITS',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'SELECT * FROM TRAIN_NMF_FEATURES',
        SET_LIST            => v_setlst,
        TARGET_COLUMN_NAME  => '"target"'
        );

END
```

4. Check the convergence of the model.

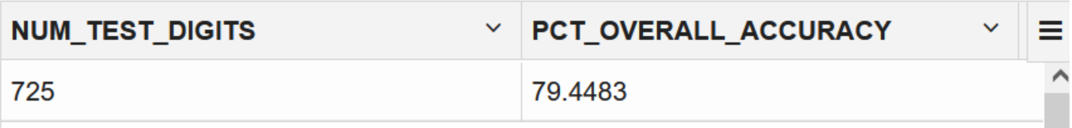
```
%sql

SELECT * from DM$VGNN_NMF_DIGITS;
```



PARTITION_NAME	NAME	NUMERIC_VALUE	STRING_VAL
	NUM_ROWS	1072	
	ITERATIONS	49	
	LOSS_VALUE	0.0000000006000693852681146	
	CONVERGED		YES

5. Check the quality of the model using PCT accuracy.

```
%sql

SELECT count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_OVERALL_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_NMF_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_NMF_FEATURES)
```



NUM_TEST_DIGITS	PCT_OVERALL_ACCURACY
725	79.4483

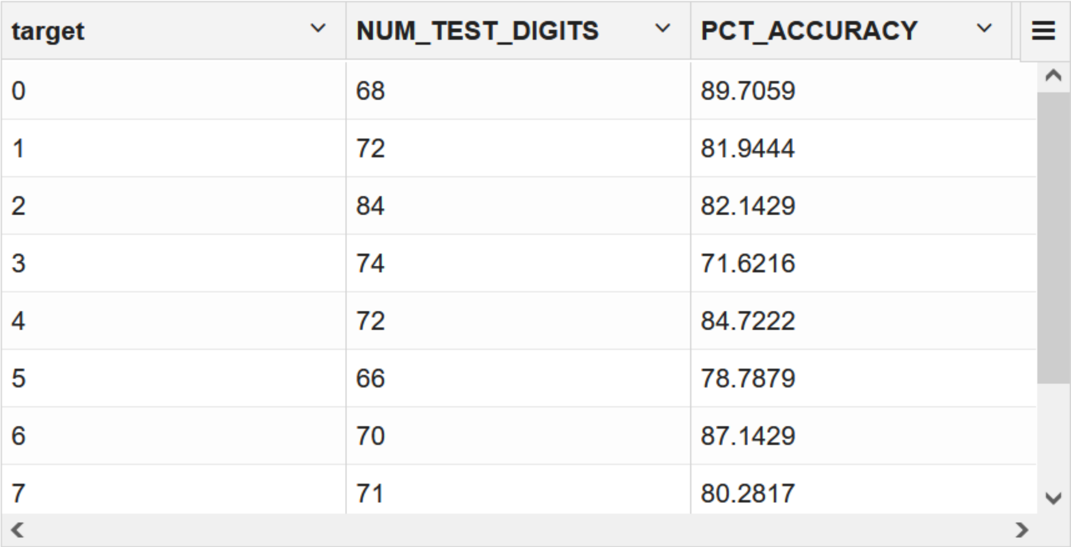
Compare the overall accuracy of the model with the original dataset and the overall accuracy of the model with the top 16 NMF features.

Tip:

The overall accuracy score looks poor here. This is for illustrative purposes. You may consider increasing the number to top 32 NMF features. Alternately, you can try a different algorithm.

6. Check the PCT accuracy of the target digit.

```
%sql

SELECT "target",
       count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_NMF_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_NMF_FEATURES)
        GROUP BY "target"
        ORDER BY "target"
```



target	NUM_TEST_DIGITS	PCT_ACCURACY
0	68	89.7059
1	72	81.9444
2	84	82.1429
3	74	71.6216
4	72	84.7222
5	66	78.7879
6	70	87.1429
7	71	80.2817

You'll notice that the model with the original dataset has better accuracy than the one with the top 16 NMF features. Further, let's see another model built by manually selecting the attributes for each feature vector having the highest coefficients and comparing the quality of all the models.

Score with Selected Attributes

You are creating another Neural Network model by manually selecting the attributes in each extracted feature vector based on their coefficients.

That means selecting all the unique attributes that have the highest coefficients with the feature vectors. Earlier you have seen the relationship of the original attributes and attribute coefficients with six feature vectors. See Step 3 of Dictionary and Model Views. For example, for Feature 1 select the following attributes: IMG51, IMG12, IMG59, IMG4, IMG27, IMG28, IMG44, IMG11, IMG37, IMG61, IMG50. The unique attributes for the six feature vectors are:

- Feature 1: IMG51, IMG12, IMG59, IMG4, IMG27, IMG28, IMG44, IMG11, IMG37, IMG61, IMG50
- Feature 2: IMG32, IMG39, IMG56, IMG0, IMG20, IMG19,
- Feature 3: IMG18, IMG36, IMG26, IMG21, IMG42
- Feature 4: IMG60, IMG3, IMG43, IMG34,
- Feature 5: IMG53, IMG13, IMG58, IMG10
- Feature 6: IMG29, IMG35, IMG52

Now create a new `TRAIN` and `TEST` view and then build a neural network model.

1. Create a new `TRAIN_NMF_ATTR` view with the top unique attributes.

```
%sql

CREATE OR REPLACE VIEW TRAIN_NMF_ATTR AS
SELECT "target",
       IMG51, IMG12, IMG59, IMG4, IMG27, IMG28, IMG44, IMG11, IMG37, IMG61, IMG50
       IMG32, IMG39, IMG56, IMG0, IMG20, IMG19,
       IMG18, IMG36, IMG26, IMG21, IMG42
       IMG60, IMG3, IMG43, IMG34,
       IMG53, IMG13, IMG58, IMG10,
       IMG29, IMG35, IMG52
FROM TRAIN_DIGITS;



TRAIN_NMF_ATTR view created
---------------------------
```

2. Create a `TEST_NMF_FEATURES` view with the top unique attributes.

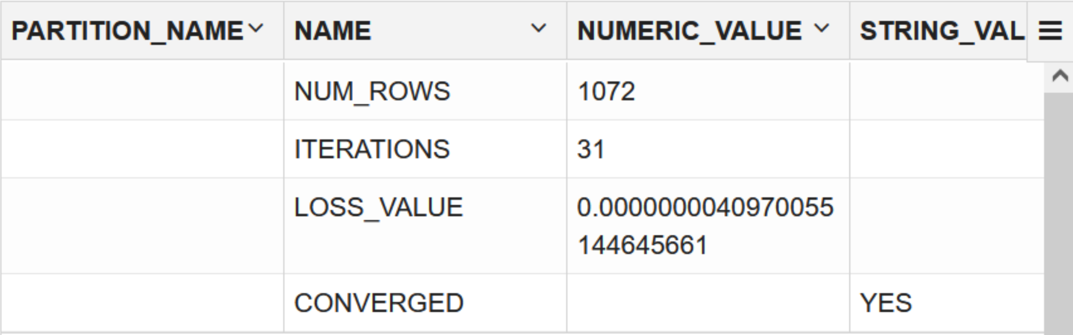
```
%sql

CREATE OR REPLACE VIEW TEST_NMF_ATTR AS
SELECT "target",
       IMG51, IMG12, IMG59, IMG4, IMG27, IMG28, IMG44, IMG11, IMG37, IMG61, IMG50
       IMG32, IMG39, IMG56, IMG0, IMG20, IMG19,
       IMG18, IMG36, IMG26, IMG21, IMG42
       IMG60, IMG3, IMG43, IMG34,
       IMG53, IMG13, IMG58, IMG10,
       IMG29, IMG35, IMG52
FROM TEST_DIGITS;



TEST_NMF_FEATURES view created
---------------------------
```

3. Build a Neural Network model using the new `TRAIN_NMF_ATTR` data set.

```
%script

BEGIN DBMS_DATA_MINING.DROP_MODEL('NN_NMF_ATT_DIGITS');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;

BEGIN
    v_setlst('PREP_AUTO') := 'ON';
    v_setlst('ALGO_NAME') := 'ALGO_NEURAL_NETWORK';
    v_setlst('NNET_NODES_PER_LAYER') := '40';
    v_setlst('NNET_ACTIVATIONS') := '''NNET_ACTIVATIONS_TANH''';
    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'NN_NMF_ATT_DIGITS',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'SELECT * FROM TRAIN_NMF_ATTR',
        SET_LIST            => v_setlst,
        TARGET_COLUMN_NAME  => '"target"'
        );

END;


PL/SQL procedure successfully completed.

---------------------------

PL/SQL procedure successfully completed.
```

4. Check the convergence of the model.

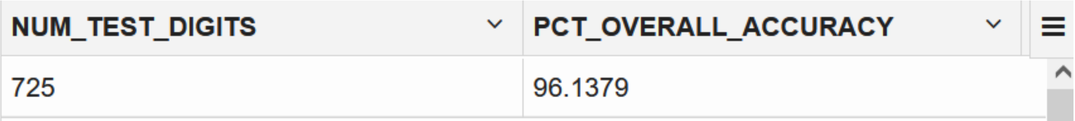
```
%sql

SELECT * from DM$VGNN_NMF_ATT_DIGITS;
```



PARTITION_NAME	NAME	NUMERIC_VALUE	STRING_VAL
	NUM_ROWS	1072	
	ITERATIONS	31	
	LOSS_VALUE	0.0000000040970055 144645661	
	CONVERGED		YES

5. Check the quality of the model using PCT accuracy.

```
%sql

SELECT count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_OVERALL_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_NMF_ATT_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_NMF_ATTR)
```

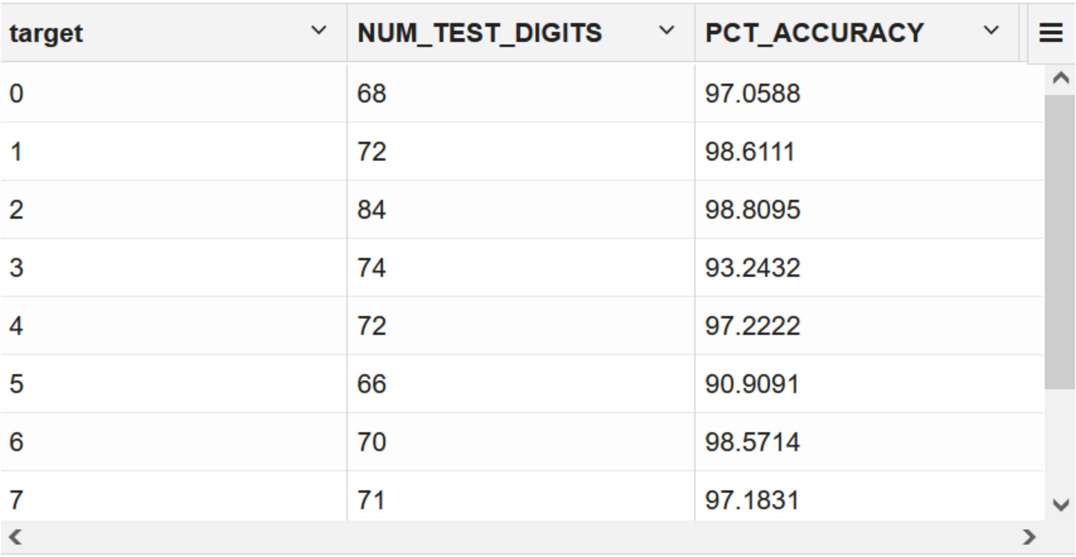


NUM_TEST_DIGITS	PCT_OVERALL_ACCURACY
725	96.1379

Compare the overall accuracy of the model with manually selected attributes with those of the earlier models.

6. Check the PCT accuracy of the target digit.

```
%sql

SELECT "target",
       count(*) NUM_TEST_DIGITS,
       ROUND((SUM(CASE WHEN ("target" - PRED_TARGET) = 0 THEN 1 ELSE 0 END)
              / COUNT(*))*100,4) PCT_ACCURACY FROM
       (SELECT "target",
               ROUND(PREDICTION(NN_NMF_ATT_DIGITS USING *), 1) PRED_TARGET
               FROM TEST_NMF_ATTR)
        GROUP BY "target"
        ORDER BY "target"
```



target	NUM_TEST_DIGITS	PCT_ACCURACY
0	68	97.0588
1	72	98.6111
2	84	98.8095
3	74	93.2432
4	72	97.2222
5	66	90.9091
6	70	98.5714
7	71	97.1831

Table 4-1 PCT Accuracy Comparison

Model	NUM_TEST_DIGITS	PCT_OVERALL_ACCURACY
NN model 1 with extracted features as input	725	96
NN model 2 with top 16 features of NMF data	725	79.4483
NN model 3 with original attributes with highest coeffient for each feature vector	725	96.1379

You found that the NN model using the 16 NMF Feature projections had a lower overall accuracy. However, the model using a reduced set of Attributes as input to the NN (using 33 out of the 64 total attributes, as suggested by NMF as being the most important) has shown a slightly better overall accuracy when compared to the original.

This way you can use one of these models for your app to read student sheets or forms to recognize handwritten numbers.

5
Examples

The OML4SQL examples are available on GitHub and some scenarios with those examples are illustrated.

- About the OML4SQL Examples
- PL/SQL API
- Example: Predicting Likely Candidates for a Sales Promotion
- Example: Analyzing Preferred Customers
- Example: Segmenting Customer Data
- Example : Comparison of Texts Using an ESA Model

About the OML4SQL Examples

The OML4SQL examples illustrate typical approaches to data preparation, algorithm selection, algorithm tuning, testing, and scoring.

You can learn a great deal about the OML4SQL application programming interface from the OML4SQL examples. The examples are simple. They include extensive inline comments to help you understand the code. They delete all temporary objects on exit so that you can run the examples repeatedly without setup or cleanup.

The OML4SQL examples are available on GitHub at https://github.com/oracle/oracle-db-examples/tree/master/machine-learning/sql/. Select the Database release (for example 23ai) to see the examples.

The OML4SQL examples create a set of machine learning models in the user's schema. The following table lists the file name of the example and the `mining_function` value and algorithm the example uses.

Table 5-1 Models Created by Examples

File Name	MINING_FUNCTION	Algorithm
`oml4sql-anomaly-detection-1class-svm.sql`	`CLASSIFICATION`	`ALGO_SUPPORT_VECTOR_MACHINE`
`oml4sql-anomaly-detection-em.sql`	`CLASSIFICATION`	`ALGO_EXPECTATION_MAXIMIZATION`
`oml4sql-association-rules.sql`	`ASSOCIATION`	`ALGO_APRIORI_ASSOCIATION_RULES`
`oml4sql-classification-decision-tree.sql`	`CLASSIFICATION`	`ALGO_DECISION_TREE`
`oml4sql-classification-glm.sql`	`CLASSIFICATION`	`ALGO_GENERALIZED_LINEAR_MODEL`
`oml4sql-classification-naive-bayes.sql`	`CLASSIFICATION`	`ALGO_NAIVE_BAYES`
`oml4sql-classification-neural-networks.sql`	`CLASSIFICATION`	`ALGO_NEURAL_NETWORK`

Table 5-1 (Cont.) Models Created by Examples

File Name	MINING_FUNCTION	Algorithm
oml4sql-classification-random-forest.sql	CLASSIFICATION	ALGO_RANDOM_FOREST
oml4sql-classification-regression-xgboost.sql	CLASSIFICATION	ALGO_XGBOOST
oml4sql-classification-svm.sql	CLASSIFICATION	ALGO_SUPPORT_VECTOR_MACHINES
oml4sql-classification-text-analysis-svm.sql	CLASSIFICATION	ALGO_SUPPORT_VECTOR_MACHINES
oml4sql-clustering-expectation-maximization.sql	CLUSTERING	ALGO_EXPECTATION_MAXIMIZATION
oml4sql-clustering-kmeanms-star-schema.sql	CLUSTERING	ALGO_KMEANS
oml4sql-clustering-kmeans.sql	CLUSTERING	ALGO_KMEANS
oml4sql-clustering-ocluster.sql	CLUSTERING	ALGO_O_CLUSTER
oml4sql-cross-validation-decision-tree.sql	CLASSIFICATION	ALGO_DECISION_TREE
oml4sql-feature-extraction-cur.sql	ATTRIBUTE_IMPORTANCE	ALGO_CUR_DECOMPOSITION
oml4sql-feature-extraction-nmf.sql	FEATURE_EXTRACTION	ALGO_NONNEGATIVE_MATRIX_FACTOR
oml4sql-feature-extraction-svd.sql	FEATURE_EXTRACTION	ALGO_SINGULAR_VALUE_DECOMP
oml4sql-feature-extraction-text-mining-esa.sql	FEATURE_EXTRACTION	ALGO_EXPLICIT_SEMANTIC_ANALYS
oml4sql-feature-extraction-text-mining-nmf.sql	FEATURE_EXTRACTION	ALGO_NONNEGATIVE_MATRIX_FACTOR
oml4sql-feature-extraction-text-term-extraction.sql	FEATURE_EXTRACTION	ALGO_EXPLICIT_SEMANTIC_ANALYSIS
oml4sql-partitioned-models-svm.sql	CLASSIFICATION	ALGO_SUPPORT_VECTOR_MACHINES
oml4sql-regression-glm.sql	REGRESSION	ALGO_GENERALIZED_LINEAR_MODEL
oml4sql-regression-neural-networks.sql	REGRESSION	ALGO_NEURAL_NETWORK
oml4sql-regression-random-forest.sql	REGRESSION	ALGO_RANDOM_FOREST
oml4sql-regression-svm.sql	REGRESSION	ALGO_SUPPORT_VECTOR_MACHINES
oml4sql-singular-value-decomposition.sql	REGRESSION	ALGO_SINGULAR_VALUE_DECOMPOSITION
oml4sql-survival-analysis-xgboost.sql	REGRESSION	ALGO_XGBOOST
oml4sql-time-series-esm-auto-model-search.sql	TIME_SERIES	ALGO_EXPONENTIAL_SMOOTHING
oml4sql-time-series-exponential-smoothing.sql	TIME_SERIES	ALGO_EXPONENTIAL_SMOOTHING
oml4sql-time-series-mset.sql	CLASSIFICATION	ALGO_MSET_SPRT

Table 5-1 (Cont.) Models Created by Examples

File Name	MINING_FUNCTION	Algorithm
oml4sql-time-series-regression-dataset.sql	-	This is a dataset to construct time series regression model.
oml4sql-time-series-regression.sql	TIME_SERIES and REGRESSION	Uses ALGO_EXPONENTIAL_SMOOTHING, ALGO_GENERALIZED_MODEL, and ALGO_XGBOOST

A few examples other than those listed in the table above are: oml4sql-attribute-importance.sql, which uses the DBMS_PREDICTIVE_ANALYTICS.EXPLAIN procedure to find the importance of attributes that independently impact the target attribute. oml4sql-feature-extraction-text-term-extraction.sql example, which uses the CTX.DDL package for text extraction.

Another set of examples demonstrates the use of the ALGO_EXTENSIBLE_LANG algorithm to register R language functions and create R models. The following table lists the R Extensibility examples. It shows the file name of the example and the MINING_FUNCTION value and R function used.

File Name	MINING_FUNCTION	R Function
oml4sql-r-extensible-algorithm-registration.sql	CLASSIFICATION	glm
oml4sql-r-extensible-association-rules.sql	ASSOCIATION	apriori
oml4sql-r-extensible-attribute-importance-via-rf.sql	REGRESSION	randomForest
oml4sql-r-extensible-glm.sql	REGRESSION	glm
oml4sql-r-extensible-kmeans.sql	CLUSTERING	kmeans
oml4sql-r-extensible-principal-components.sql	FEATURE_EXTRACTION	prcomp
oml4sql-r-extensible-regression-tree.sql	REGRESSION	rpart
oml4sql-r-extensible-regression-neural-networks.sql	REGRESSION	nnet

Install the OML4SQL Examples

Learn how to install OML4SQL examples.

The OML4SQL examples require:

- Oracle Database (on-premises, Oracle Database Cloud Service, or Oracle Autonomous Database)
- Oracle Database sample schemas
- A user account with the privileges described in Grant Privileges for Oracle Machine Learning for SQL.
- Running of dmshgrants.sql by a system administrator
- Running of dmsh.sql by the OML4SQL user

Follow these steps to install the OML4SQL examples:

1. Install or obtain access to an Oracle Database 23ai instance. To install the database, see the installation instructions for your platform at Oracle Database 23ai.
2. Ensure that the sample schemas are installed in the database. See *Oracle Database Sample Schemas* for details about the sample schemas.
3. Download the example code files from GitHub at https://github.com/oracle/oracle-db-examples/tree/master/machine-learning/sql. Select the Database edition. Place the files in a directory to which you have access on the Oracle Database server. For example, `$ORACLE_HOME/demo/schema`. `$ORACLE_HOME` is the home path where you have installed the database. Typically, `/scratch/u01/app/oracle/product/23.0.0/dbhome_1`.
4. Verify that your user account has the required privileges described in Grant Privileges for Oracle Machine Learning for SQL.
5. Ask your system administrator to run the `dmshgrants.sql` script, or run it yourself if you have administrative privileges. The script grants the privileges that are required for running the examples. These include `SELECT` access to tables in the `SH` schema as described in OML4SQL Sample Data and the system privileges.

   Connect as `SYSDBA`:

   ```
   CONNECT sys / as sysdba
   Enter password: sys_password
   Connected.
   ```

   Pass the name of the OML4SQL user to `dmshgrants`:

   ```
   @<location_of_examples>/dmshgrants oml_user
   ```

6. Connect to the database and run the `dmsh.sql` script. This script creates views of the sample data in the schema of the OML4SQL user.

   ```
   CONNECT oml_user
   Enter password: oml_user_password
   Connected.
   ```

   Issue the following to run the script:

   ```
   @<location_of_examples>/dmsh.sql
   ```

Related Topics

- *Oracle Database Sample Schemas*

OML4SQL Sample Data

The data used by the OML4SQL examples is based on these tables in the `SH` schema.

Those tables are:

```
SH.CUSTOMERS
SH.SALES
SH.PRODUCTS
```

```
SH.SUPPLEMENTARY_DEMOGRAPHICS
SH.COUNTRIES
```

The `dmshgrants` script grants `SELECT` access to the tables in the `SH` schema. The `dmsh.sql` script creates views of the `SH` tables in the schema of the OML4SQL user. The views are described in the following table.

Table 5-2   Views Created by dmsh.sql

View Name	Description
MINING_DATA	Joins and filters data
MINING_DATA_BUILD_V	Data for building models
MINING_DATA_TEST_V	Data for testing models
MINING_DATA_APPLY_V	Data to be scored
MINING_BUILD_TEXT	Data for building models that include text
MINING_TEST_TEXT	Data for testing models that include text
MINING_APPLY_TEXT	Data, including text columns, to be scored
MINING_DATA_ONE_CLASS_V	Data for anomaly detection

The association rules example creates its own transactional data.

PL/SQL API

The OML4SQL PL/SQL API is built into the `DBMS_DATA_MINING` PL/SQL package, which has routines for building, testing, and maintaining machine learning models. This package also has a batch apply operation.

The following example shows part of a simple PL/SQL script for creating an SVM classification model called SVMC_SH_Clas_sample. The model build uses weights, specified in a weights table, and settings, specified in a settings table. The weights influence the weighting of target classes. The settings override default behavior. The model uses Automatic Data Preparation (`prep_auto_on` setting). The model is trained on the data in mining_data_build_v.

Example 5-1   Creating a Classification Model

```
-----------------------  CREATE AND POPULATE A CLASS WEIGHTS TABLE  ------------
CREATE TABLE svmc_sh_sample_class_wt (
  target_value NUMBER,
  class_weight NUMBER);
INSERT INTO svmc_sh_sample_class_wt VALUES (0,0.35);
INSERT INTO svmc_sh_sample_class_wt VALUES (1,0.65);
COMMIT;
-----------------------  CREATE AND POPULATE A SETTINGS TABLE ------------------
CREATE TABLE svmc_sh_sample_settings (
  setting_name  VARCHAR2(30),
  setting_value VARCHAR2(4000));
BEGIN
INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
  (dbms_data_mining.algo_name, dbms_data_mining.algo_support_vector_machines);
INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
  (dbms_data_mining.svms_kernel_function, dbms_data_mining.svms_linear);
INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
  (dbms_data_mining.clas_weights_table_name, 'svmc_sh_sample_class_wt');
INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
```

```
  (dbms_data_mining.prep_auto, dbms_data_mining.prep_auto_on);
END;
/
------------------------  CREATE THE MODEL -------------------------------------
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'SVMC_SH_Clas_sample',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_data_build_v',
    case_id_column_name => 'cust_id',
    target_column_name  => 'affinity_card',
    settings_table_name => 'svmc_sh_sample_settings');
END;
/
```

Example: Predicting Likely Candidates for a Sales Promotion

This example shows `PREDICTION` query to target customers in Brazil for a special promotion that offers coupons and an affinity card.

The query uses data on marital status, education, and income to predict the customers who are most likely to take advantage of the incentives. The query applies a Decision Tree model called dt_sh_clas_sample to score the customer data. The model is created by the `oml4sql-classification-decision-tree.sql` example.

Example 5-2   Predict Best Candidates for an Affinity Card

```
SELECT cust_id
  FROM mining_data_apply_v
  WHERE
      PREDICTION(dt_sh_clas_sample
                   USING cust_marital_status, education, cust_income_level ) = 1
  AND country_name IN 'Brazil';
```

The output is as follows:

```
CUST_ID
----------
    100404
    100607
    101113
```

The same query, but with a bias to favor false positives over false negatives, is shown here.

```
SELECT cust_id
  FROM mining_data_apply_v
  WHERE
      PREDICTION(dt_sh_clas_sample COST MODEL
                   USING cust_marital_status, education, cust_income_level ) = 1
  AND country_name IN 'Brazil';
```

The output is as follows:

```
CUST_ID
----------
```

```
    100139
    100163
    100275
    100404
    100607
    101113
    101170
    101463
```

The `COST MODEL` keywords cause the cost matrix associated with the model to be used in making the prediction. The cost matrix, stored in a table called `dt_sh_sample_costs`, specifies that a false negative is eight times more costly than a false positive. Overlooking a likely candidate for the promotion is far more costly than including an unlikely candidate.

```
SELECT * FROM dt_sh_sample_cost;
```

The output is as follows:

```
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE       COST
------------------- ---------------------- ----------
                  0                      0          0
                  0                      1          1
                  1                      0          8
                  1                      1          0
```

Example: Analyzing Preferred Customers

The examples in this section reveal information about customers who use affinity cards or are likely to use affinity cards.

Example 5-3   Find Demographic Information About Preferred Customers

This query returns the gender, age, and length of residence of typical affinity card holders. The anomaly detection model, SVMO_SH_Clas_sample, returns `1` for typical cases and `0` for anomalies. The demographics are predicted for typical customers only; outliers are not included in the sample. The model is created by the `oml4sql-anomaly-detection-1class-svm.sql` example.

```
SELECT cust_gender, round(avg(age)) age,
       round(avg(yrs_residence)) yrs_residence,
       count(*) cnt
FROM mining_data_one_class_v
WHERE PREDICTION(SVMO_SH_Clas_sample using *) = 1
GROUP BY cust_gender
ORDER BY cust_gender;
```

The output is as follows:

```
CUST_GENDER         AGE YRS_RESIDENCE        CNT
------------ ---------- ------------- ----------
F                    40             4         36
M                    45             5        304
```

Example 5-4 Dynamically Identify Customers Who Resemble Preferred Customers

This query identifies customers who do not currently have an affinity card, but who share many of the characteristics of affinity card holders. The PREDICTION and PREDICTION_PROBABILITY functions use an OVER clause instead of a predefined model to classify the customers. The predictions and probabilities are computed dynamically.

```
SELECT cust_id, pred_prob
 FROM
  (SELECT cust_id, affinity_card,
    PREDICTION(FOR TO_CHAR(affinity_card) USING *) OVER () pred_card,
    PREDICTION_PROBABILITY(FOR TO_CHAR(affinity_card),1 USING *) OVER () pred_prob
   FROM mining_data_build_v)
 WHERE affinity_card = 0
  AND pred_card = 1
 ORDER BY pred_prob DESC;
```

The output is as follows:

```


  CUST_ID  PRED_PROB
---------- ---------
    102434       .96
    102365       .96
    102330       .96
    101733       .95
    102615       .94
    102686       .94
    102749       .93
.
.
.
.
    102580       .52
    102269       .52
    102533       .51
    101604       .51
    101656       .51

226 rows selected.
```

Example 5-5 Predict the Likelihood that a New Customer Becomes a Preferred Customer

This query computes the probability of a first-time customer becoming a preferred customer (an affinity card holder). This query can be run in real time at the point of sale.

The new customer is a 44-year-old American executive who has a bachelors degree and earns more than $300,000/year. He is married, lives in a household of 3, and has lived in the same residence for the past 6 years. The probability of this customer becoming a typical affinity card holder is only 5.8%.

```
SELECT PREDICTION_PROBABILITY(SVMO_SH_Clas_sample, 1 USING
                             44 AS age,
                             6 AS yrs_residence,
                             'Bach.' AS education,
                             'Married' AS cust_marital_status,
                             'Exec.' AS occupation,
                             'United States of America' AS country_name,
                             'M' AS cust_gender,
                             'L: 300,000 and above' AS cust_income_level,
                             '3' AS houshold_size
                             ) prob_typical
FROM DUAL;
```

The output is as follows:

```
PROB_TYPICAL
------------
  5.8
```

Example 5-6 Use Predictive Analytics to Find Top Predictors

The DBMS_PREDICTIVE_ANALYTICS PL/SQL package contains routines that perform simple machine learning operations without a predefined model. In this example, the EXPLAIN routine computes the top predictors for affinity card ownership. The procedure does not create a model that can be stored in the database for further exploration. Automatic Data Preparation is also performed behind the scenes. The results show that household size, marital status, and age are the top three predictors.

```
BEGIN
    DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
        data_table_name      => 'mining_data_test_v',
        explain_column_name  => 'affinity_card',
        result_table_name    => 'cust_explain_result');
END;
/

SELECT * FROM cust_explain_result
  WHERE rank < 4;
```

The output is as follows:

```
ATTRIBUTE_NAME            ATTRIBUTE_SUBNAME    EXPLANATORY_VALUE       RANK
------------------------- -------------------- ----------------- ----------
HOUSEHOLD_SIZE                                        .209628541          1
CUST_MARITAL_STATUS                                   .199794636          2
AGE                                                   .111683067          3
```

Another way to arrive at top predictors for affinity ownership is by using attribute importance mining function. Create a model with the Minimum Description Length algorithm. Define mining_function as ATTRIBUTE_IMPORTANCE. You can then query the DM$VA model detail view to get the top three predictors.

```
BEGIN DBMS_DATA_MINING.DROP_MODEL('AI_EXPLAIN_OUTPUT');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME') := 'ALGO_AI_MDL';
    V_setlst('PREP_AUTO') := 'ON';
    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME => 'AI_EXPLAIN_OUTPUT',
        MINING_FUNCTION => 'ATTRIBUTE_IMPORTANCE',
        DATA_QUERY => 'select * from mining_data_test_v',
        SET_LIST => v_setlst,
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        TARGET_COLUMN_NAME => 'AFFINITY_CARD');
END;


Find the top 3 predictors from the DM$VA model detail view:
SELECT ATTRIBUTE_NAME, ATTRIBUTE_IMPORTANCE_VALUE, ATTRIBUTE_RANK FROM
DM$VAAI_EXPLAIN_OUTPUT;
```

The output is as follows:

```
ATTRIBUTE_NAME             ATTRIBUTE_IMPORTANCE_VALUE   ATTRIBUTE_RANK
HOUSEHOLD_SIZE             0.16154338717879052                       1
CUST_MARITAL_STATUS        0.1561477632217005                        2
AGE                        0.08440594628406521                       3
```

Example: Segmenting Customer Data

The examples in this section use an Expectation Maximization clustering model to segment the customer data based on common characteristics.

Example 5-7 Compute Customer Segments

This query computes natural groupings of customers and returns the number of customers in each group. The em_sh_clus_sample model is created by the `oml4sql-clustering-expectation-maximization.sql` example.

```
SELECT CLUSTER_ID(em_sh_clus_sample USING *) AS clus, COUNT(*) AS cnt
  FROM mining_data_apply_v
GROUP BY CLUSTER_ID(em_sh_clus_sample USING *)
ORDER BY cnt DESC;
```

The output is as follows:

```
      CLUS        CNT
---------- ----------
         9        311
         3        294
         7        215
        12        201
        17        123
        16        114
        14         86
        19         64
        15         56
        18         36
```

Example 5-8 Find the Customers Who Are Most Likely To Be in the Largest Segment

The query in Example 5-7 shows that segment 9 has the most members. The following query lists the five customers who are most likely to be in segment 9.

```
SELECT cust_id
FROM (SELECT cust_id, RANK() over (ORDER BY prob DESC, cust_id) rnk_clus2
  FROM (SELECT cust_id,
          ROUND(CLUSTER_PROBABILITY(em_sh_clus_sample, 9 USING *),3) prob
          FROM mining_data_apply_v))
WHERE rnk_clus2 <= 5
ORDER BY rnk_clus2;
```

The output is as follows:

```
   CUST_ID
----------
    100002
    100012
    100016
    100019
    100021
```

Example 5-9 Find Key Characteristics of the Most Representative Customer in the Largest Cluster

The query in Example 5-8 lists customer 100002 first in the list of likely customers for segment 9. The following query returns the five characteristics that are most significant in determining the assignment of customer 100002 to segments with probability > 20% (only segment 9 for this customer).

```
SELECT S.cluster_id, probability prob,
       CLUSTER_DETAILS(em_sh_clus_sample, S.cluster_id, 5 using T.*) det
 FROM
  (SELECT v.*, CLUSTER_SET(em_sh_clus_sample, NULL, 0.2 USING *) pset
    FROM mining_data_apply_v v
    WHERE cust_id = 100002) T,
 TABLE(T.pset) S
 ORDER BY 2 desc;
```

The output is as follows:

```
CLUSTER_ID    PROB DET
---------- --------
------------------------------------------------------------------------------
--
         9  1.0000 <Details algorithm="Expectation Maximization" cluster="9">
                   <Attribute name="YRS_RESIDENCE" actualValue="4" weight="1"
rank="1"/>
                   <Attribute name="EDUCATION" actualValue="Bach." weight="0"
rank="2"/>
                   <Attribute name="AFFINITY_CARD" actualValue="0" weight="0"
rank="3"/>
                   <Attribute name="BOOKKEEPING_APPLICATION" actualValue="1"
weight="0" rank="4"/>
                   <Attribute name="Y_BOX_GAMES" actualValue="0" weight="0"
```

```
rank="5"/>
                   </Details>
```

Example : Comparison of Texts Using an ESA Model

The examples shows the FEATURE_COMPARE function comparing texts for semantic relatedness (similarity) using the Explicit Semantic Analysis (ESA) prebuilt Wikipedia-based model, which extracts topics and compares text.

The examples shows an ESA model built against a prebuilt Wiki data set rendering over 200,000 features. The documents are analyzed as text and the document titles are given as the feature IDs. In the first example, the pair of sentence scores higher because Nick Price is a golfer born in South Africa.

Similar Texts

```
SELECT 1-FEATURE_COMPARE(esa_wiki_mod USING 'There are several PGA tour
golfers from South Africa' text AND USING 'Nick Price won the 2002 Mastercard
Colonial Open' text) similarity FROM DUAL;
```

The output is as follows:

```
SIMILARITY
----------
      .110
```

The output metric shows distance calculation. Therefore, smaller number represent more similar texts. So, 1 minus the distance in the queries result in similarity.

Dissimilar Texts

```
SELECT 1-FEATURE_COMPARE(esa_wiki_mod USING 'There are several PGA tour
golfers from South Africa' text AND USING 'John Elway played quarterback for
the Denver Broncos' text) similarity FROM DUAL;
```

The output is as follows:

```
SIMILARITY
----------
      .004
```

Example: Using Vector Data for Dimensionality Reduction and Clustering

The example demonstrates how to use vector data for dimensionality reduction and clustering, using Principal Component Analysis (PCA) and *k*-Means.

1. Assume that there is a data set called `datavec` containing one `ID` column and a vector column with 100 dimensions.

```
Name                   Null?    Type
---------------------- -------- ----------------------------
ID                              NUMBER
PROD_DATA                       VECTOR(100, FLOAT32, DENSE)
```

2. Build a PCA feature extraction model. The following step creates a model that uses PCA scoring to reduce dimensionality.

```
DECLARE
  v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
  v_setlst('ALGO_NAME')         := 'ALGO_SINGULAR_VALUE_DECOMP';
  v_setlst('SVDS_SCORING_MODE') := 'SVDS_SCORING_PCA';

  DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'pca_model',
        MINING_FUNCTION     => 'FEATURE_EXTRACTION',
        DATA_QUERY          => 'SELECT * FROM DATAVEC',
        CASE_ID_COLUMN_NAME => 'id',
        SET_LIST            => v_setlst);
END;
/
```

3. Transform PCA results into a vector table `pca_data` with reduced dimensions by using the `VECTOR_EMBEDDING()` operator.

```
CREATE table pca_data as SELECT id, VECTOR_EMBEDDING(pca_model using *)
embedding FROM datavec;
```

4. The new `pca_data` contains one ID column and one vector with 10 dimensions based on the data characteristics.

```
DESC pca_data;
Name             Null?    Type
---------------- -------- ----------------------------
ID                          NUMBER
EMBEDDING                 VECTOR(10, FLOAT64, DENSE)
```

5. Build a *k*-Means clustering model on `pca_data`, leveraging its reduced dimensions.

```
DECLARE
  v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
  v_setlst('ALGO_NAME')         := 'ALGO_KMEANS';
  v_setlst('KMNS_DETAILS')      := 'KMNS_DETAILS_ALL';
  v_setlst('CLUS_NUM_CLUSTERS') := '2';

  DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'km_model',
        MINING_FUNCTION     => 'CLUSTERING',
        DATA_QUERY          => 'SELECT * FROM PCA_DATA',
        CASE_ID_COLUMN_NAME => 'id',
        SET_LIST            => v_setlst);
END;
/
```

6. Check the data dictionary settings.

```
SELECT model_name, attribute_name, data_type, target, vector_info
FROM   USER_MINING_MODEL_ATTRIBUTES
WHERE  model_name='KM_MODEL’ ORDER BY attribute_name;

MODEL_NAME ATTRIBUTE_NAME  DATA_TYPE            TAR VECTOR_INFO
---------- --------------- ------------------------ ---
--------------------
KM_MODEL   EMBEDDING       VECTOR                 NO  VECTOR(10,FLOAT64)
```

7. You can check the model detail views for `KM_MODEL` model.

```
SELECT model_name, view_name, view_type
FROM USER_MINING_MODEL_VIEWS
WHERE model_name='KM_MODEL’ ORDER BY view_name;

MODEL_NAME VIEW_NAME         VIEW_TYPE
---------- --------------- ----------------------------------------
KM_MODEL   DM$VAKM_MODEL   Clustering Attribute Statistics
KM_MODEL   DM$VCKM_MODEL   k-Means Scoring Centroids
KM_MODEL   DM$VDKM_MODEL   Clustering Description
KM_MODEL   DM$VGKM_MODEL   Global Name-Value Pairs
KM_MODEL   DM$VHKM_MODEL   Clustering Histograms
KM_MODEL   DM$VNKM_MODEL   Normalization and Missing Value Handling
KM_MODEL   DM$VRKM_MODEL   Clustering Rules
KM_MODEL   DM$VSKM_MODEL   Computed Settings
KM_MODEL   DM$VWKM_MODEL   Model Build Alerts
```

8. You can also view each vector dimension as a predictor from the model details.

```
SELECT * FROM(SELECT cluster_id, attribute_name, attribute_subname,
       mean, variance, mode_value
FROM DM$VAKM_MODEL ORDER BY cluster_id, attribute_name,attribute_subname)

CLUSTER_ID ATTRIBUTE_NAME  ATTRIBUTE_SUBNAME    MEAN
VARIANCE       MODE_VALUE
---------- --------------- -------------------- -------------
------------- --------------------
         1 EMBEDDING          DM$$VEC1             28.9538       3.4382
         2 EMBEDDING          DM$$VEC1             27.9580       5.5661
         3 EMBEDDING          DM$$VEC1             29.9495       2.1698
```

9. Use scoring operators `CLUSTER_ID` and `CLUSTER_PROBABILITY` to find cluster assignments and probabilities for each record in `pca_data`.

```
SELECT id, cluster_id(km_model using *) cluster_id,
cluster_probability(km_model using *)probability FROM pca_data ORDER BY id;

ID         CLUSTER_ID      PROBABILITY
----------      ----------      -----------
```

```
1                 1               .617
2                 2               .584
3                 1               .579
4                 1               .605
5                 1               .621
6                 1               .642
7                 2               .598
8                 2               .614
9                 2               .650
10                2               .618
```

6
Reference

Specify Model Settings

You can configure your model by specifying model settings.

Numerous configuration settings are available for configuring machine learning models at build time. Specify your model settings in `CREATE_MODEL` or `CREATE_MODEL2` procedures. To specify settings in `CREATE_MODEL` procedure, create a settings table with the columns shown in the following table and pass the table to in the procedure.

You can also use `CREATE_MODEL2` procedure where you can directly pass the model settings to a variable that can be used in the procedure. The variable can be declared with `DBMS_DATA_MINING.SETTING_LIST` procedure.

Table 6-1 Settings Table Required Columns

Column Name	Data Type
`setting_name`	`VARCHAR2(30)`
`setting_value`	`VARCHAR2(4000)`

Example 6-1 creates a settings table for a Support Vector Machine (SVM) classification model. Since SVM is not the default classifier, the `ALGO_NAME` setting is used to specify the algorithm. Setting the `SVMS_KERNEL_FUNCTION` to `SVMS_LINEAR` causes the model to be built with a linear kernel. If you do not specify the kernel function, the algorithm chooses the kernel based on the number of attributes in the data.

Example 6-2 creates a model with the model settings that are stored in a variable from `SETTING_LIST`.

Some settings apply generally to the model, others are specific to an algorithm. Model settings are referenced in Table 6-2 and Table 6-3.

Table 6-2   General Model Settings

Settings	Description
Machine learning function settings	Machine Learning Technique Settings
Algorithm names	Algorithm Names
Global model characteristics	Global Settings
Automatic Data Preparation	Automatic Data Preparation

Table 6-3   Algorithm-Specific Model Settings

Algorithm	Description
CUR Matrix Decomposition	DBMS_DATA_MINING —Algorithm Settings: CUR Matrix Decomposition
Decision Tree	DBMS_DATA_MINING —Algorithm Settings: Decision Tree
Expectation Maximization	DBMS_DATA_MINING —Algorithm Settings: Expectation Maximization
Explicit Semantic Analysis	DBMS_DATA_MINING —Algorithm Settings: Explicit Semantic Analysis
Exponential Smoothing	DBMS_DATA_MINING —Algorithm Settings: Exponential Smoothing Models
Generalized Linear Model	DBMS_DATA_MINING —Algorithm Settings: Generalized Linear Models
k-Means	DBMS_DATA_MINING —Algorithm Settings: *k*-Means
Multivariate State Estimation Technique - Sequential Probability Ratio Test	DBMS_DATA_MINING - Algorithm Settings: Multivariate State Estimation Technique - Sequential Probability Ratio Test
Naive Bayes	Algorithm Settings: Naive Bayes
Neural Network	DBMS_DATA_MINING —Algorithm Settings: Neural Network
Non-Negative Matrix Factorization	DBMS_DATA_MINING —Algorithm Settings: Non-Negative Matrix Factorization
O-Cluster	Algorithm Settings: O-Cluster
Random Forest	DBMS_DATA_MINING — Algorithm Settings: Random Forest
Singular Value Decomposition	DBMS_DATA_MINING —Algorithm Settings: Singular Value Decomposition
Support Vector Machine	DBMS_DATA_MINING —Algorithm Settings: Support Vector Machine
XGBoost	DBMS_DATA_MINING — Algorithm Settings: XGBoost

Note:

Some XGBoost objectives apply only to classification function models and other objectives apply only to regression function models. If you specify an incompatible `objective` value, an error is raised. In the `DBMS_DATA_MINING.CREATE_MODEL` procedure, if you specify `DBMS_DATA_MINING.CLASSIFICATION` as the function, then the only objective values that you can use are the `binary` and `multi` values. The one exception is `binary: logitraw`, which produces a continuous value and applies only to a regression model. If you specify `DBMS_DATA_MINING.REGRESSION` as the function, then you can specify `binary: logitraw` or any of the `count`, `rank`, `reg`, and `survival` values as the objective.

The values for the XGBoost objective setting are listed in the Settings for Learning Tasks table in DBMS_DATA_MINING — Algorithm Settings: XGBoost.

Example 6-1 Creating a Settings Table and Creating an SVM Classification Model Using CREATE.MODEL procedure

```
CREATE TABLE svmc_sh_sample_settings (
  setting_name VARCHAR2(30),
  setting_value VARCHAR2(4000));

BEGIN
  INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
    (dbms_data_mining.algo_name, dbms_data_mining.algo_support_vector_machines);
  INSERT INTO svmc_sh_sample_settings (setting_name, setting_value) VALUES
    (dbms_data_mining.svms_kernel_function, dbms_data_mining.svms_linear);
  COMMIT;
END;
/
-- Create the model using the specified settings
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'svm_model',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_data_build_v',
    case_id_column_name => 'cust_id',
    target_column_name  => 'affinity_card',
    settings_table_name => 'svmc_sh_sample_settings');
END;
```

Example 6-2 Specify Model Settings for a SVM Classification Model Using CREATE_MODEL2 procedure

```
DECLARE
    v_setlist DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlist('PREP_AUTO') := 'ON';
    v_setlist('ALGO_NAME') := 'ALGO_SUPPORT_VECTOR_MACHINES';
    v_setlist('SVMS_KERNEL_FUNCTION') := 'SVMS_LINEAR';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'SVM_MODEL',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'select * from mining_data_build_v',
        SET_LIST            => v_setlist,
        CASE_ID_COLUMN_NAME => 'CUST_ID,
    TARGET_COLUMN_NAME  => 'AFFINITY_CARD');
END;
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Specify Oracle Machine Learning Model Settings for an R Model

This topic applies only to Oracle on-premises.

The machine learning model settings for an R language model determine the characteristics of the model and are specified in the model settings table.

You can build a machine learning model in the R language by specifying R as the value of the ALGO_EXTENSIBLE_LANG setting in the model settings table. You can create a model by combining in the settings table generic settings that do not require an algorithm, such as ODMS_PARTITION_COLUMNS and ODMS_SAMPLING. You can also specify the following settings, which are exclusive to an R machine learning model.

- ALGO_EXTENSIBLE_LANG
- RALG_BUILD_FUNCTION
- RALG_BUILD_PARAMETER
- RALG_DETAILS_FORMAT
- RALG_DETAILS_FUNCTION
- RALG_SCORE_FUNCTION
- RALG_WEIGHT_FUNCTION

Related Topics

- Registered R Scripts
  The RALG_*_FUNCTION settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

ALGO_EXTENSIBLE_LANG

Use the ALGO_EXTENSIBLE_LANG setting to specify the language for the Oracle Machine Learning for SQL extensible algorithm framework.

Currently, R is the only valid value for the ALGO_EXTENSIBLE_LANG setting. When you set the value for ALGO_EXTENSIBLE_LANG to R, the machine learning models are built using the R language. You can use the following settings in the settings table to specify the characteristics of the R model.

- RALG_BUILD_FUNCTION
- RALG_BUILD_PARAMETER
- RALG_DETAILS_FUNCTION
- RALG_DETAILS_FORMAT
- RALG_SCORE_FUNCTION
- RALG_WEIGHT_FUNCTION

Related Topics

- Registered R Scripts
  The RALG_*_FUNCTION settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

RALG_BUILD_FUNCTION

Use the RALG_BUILD_FUNCTION setting to specify the name of an existing registered R script for building an Oracle Machine Learning for SQL model using the R language.

You must specify both the RALG_BUILD_FUNCTION and ALGO_EXTENSIBLE_LANG settings in the model settings table. The R script defines an R function that has as the first input argument an R data.frame object for training data. The function returns an Oracle Machine Learning model object. The first data argument is mandatory. The RALG_BUILD_FUNCTION can accept additional model build parameters.

> **Note:**
>
> The valid inputs for input parameters are numeric and string scalar data types.

Example 6-3 Example of RALG_BUILD_FUNCTION

This example shows how to specify the name of the R script *MY_LM_BUILD_SCRIPT* that is used to build the model.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_build_function,'MY_LM_BUILD_SCRIPT');
End;
/
```

The R script *MY_LM_BUILD_SCRIPT* defines an R function that builds the LM model. You must register the script *MY_LM_BUILD_SCRIPT* in the Oracle Machine Learning for R script repository which uses the existing Oracle Machine Learning for R security restrictions. You can use the Oracle Machine Learning for R sys.rqScriptCreate procedure to register the script. Oracle Machine Learning for R requires the RQADMIN role to register R scripts.

For example:

```
Begin
sys.rqScriptCreate('MY_LM_BUILD_SCRIPT', 'function(data, formula,
model.frame) {lm(formula = formula, data=data, model =
as.logical(model.frame)}');
End;
/
```

For Clustering and Feature Extraction machine learning function model builds, the R attributes dm$nclus and dm$nfeat must be set on the return R model to indicate the number of clusters and features respectively.

The R script MY_KM_BUILD_SCRIPT defines an R function that builds the *k*-Means model for clustering. The R attribute dm$nclus is set with the number of clusters for the returned clustering model.

```
'function(dat) {dat.scaled <- scale(dat)
     set.seed(6543); mod <- list()
     fit <- kmeans(dat.scaled, centers = 3L)
     mod[[1L]] <- fit
     mod[[2L]] <- attr(dat.scaled, "scaled:center")
     mod[[3L]] <- attr(dat.scaled, "scaled:scale")
     attr(mod, "dm$nclus") <- nrow(fit$centers)
     mod}'
```

The R script MY_PCA_BUILD_SCRIPT defines an R function that builds the PCA model. The R attribute dm$nfeat is set with the number of features for the returned feature extraction model.

```
'function(dat) {
     mod <- prcomp(dat, retx = FALSE)
     attr(mod, "dm$nfeat") <- ncol(mod$rotation)
     mod}'
```

Related Topics

- RALG_BUILD_PARAMETER
  The RALG_BUILD_FUNCTION input parameter specifies a list of numeric and string scalar values in SQL SELECT query statement format.
- Registered R Scripts
  The RALG_*_FUNCTION settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

RALG_BUILD_PARAMETER

The RALG_BUILD_FUNCTION input parameter specifies a list of numeric and string scalar values in SQL SELECT query statement format.

Example 6-4 Example of RALG_BUILD_PARAMETER

The RALG_BUILD_FUNCTION input parameters must be a list of numeric and string scalar values. The input parameters are optional.

The syntax of the parameter is:

```
'SELECT value parameter name ...FROM dual'
```

This example shows how to specify a formula for the input argument 'formula' and a numeric value of zero for input argument 'model.frame' using the RALG_BUILD_PARAMETER. These input arguments must match with the function signature of the R script used in the RALG_BUILD_FUNCTION parameter.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_build_parameter, 'select ''AGE ~ .'' as "formula", 0
as "model.frame" from dual');
```

```
End;
/
```

Related Topics

- RALG_BUILD_FUNCTION
  Use the RALG_BUILD_FUNCTION setting to specify the name of an existing registered R script for building an Oracle Machine Learning for SQL model using the R language.

RALG_DETAILS_FUNCTION

The RALG_DETAILS_FUNCTION specifies the R model metadata that is returned in the R data.frame.

Use the RALG_DETAILS_FUNCTION to specify an existing registered R script that generates model information. The script defines an R function that contains the first input argument for the R model object. The output of the R function must be a data.frame. The columns of the data.frame are defined by the RALG_DETAILS_FORMAT setting, and may contain only numeric or string scalar types.

Example 6-5 Example of RALG_DETAILS_FUNCTION

This example shows how to specify the name of the R script MY_LM_DETAILS_SCRIPT in the model settings table. This script defines the R function that is used to provide the model information.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_details_function, 'MY_LM_DETAILS_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, the script *MY_LM_DETAILS_SCRIPT* is registered as:

```
 'function(mod) data.frame(name=names(mod$coefficients),
    coef=mod$coefficients)'
```

Related Topics

- Registered R Scripts
  The RALG_*_FUNCTION settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.
- RALG_DETAILS_FORMAT
  Use the RALG_DETAILS_FORMAT setting to specify the names and column types in the model view.

RALG_DETAILS_FORMAT

Use the RALG_DETAILS_FORMAT setting to specify the names and column types in the model view.

The value of the setting is a string that contains a SELECT statement to specify a list of numeric and string scalar data types for the name and type of the model view columns.

When the RALG_DETAILS_FORMAT and RALG_DETAILS_FUNCTION settings are both specified, a model view by the name DM$VD *<model_name>* is created along with an R model in the current schema. The first column of the model view is PARTITION_NAME. It has the value NULL for non-partitioned models. The other columns of the model view are defined by RALG_DETAILS_FORMAT setting.

Example 6-6 Example of RALG_DETAILS_FORMAT

This example shows how to specify the name and type of the columns for the generated model view. The model view contains the varchar2 column attr_name and the number column coef_value after the first column partition_name.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_details_format, 'select cast(''a'' as varchar2(20)) as
attr_name, 0 as coef_value from dual');
End;
/
```

Related Topics

- RALG_DETAILS_FUNCTION
  The RALG_DETAILS_FUNCTION specifies the R model metadata that is returned in the R data.frame.

RALG_SCORE_FUNCTION

Use the RALG_SCORE_FUNCTION setting to specify an existing registered R script for R algorithm machine learning model to use for scoring data.

The specified R script defines an R function. The first input argument defines the model object. The second input argument defines the R data.frame that is used for scoring data.

Example 6-7 Example of RALG_SCORE_FUNCTION

This example shows how the R function takes the Linear Model model and scores the data in the data.frame. The function argument object is the LM model. The argument newdata is a data.frame containing the data to score.

```
function(object, newdata) {res <- predict.lm(object, newdata = newdata,
se.fit = TRUE); data.frame(fit=res$fit, se=res$se.fit,
df=summary(object)$df[1L])}
```

The output of the R function must be a data.frame. Each row represents the prediction for the corresponding scoring data from the input data.frame. The columns of the data.frame are specific to machine learning functions, such as:

Regression: A single numeric column for the predicted target value, with two optional columns containing the standard error of the model fit, and the degrees of freedom number. The optional columns are needed for the SQL function PREDICTION_BOUNDS to work.

Example 6-8 Example of RALG_SCORE_FUNCTION for Regression

This example shows how to specify the name of the R script *MY_LM_PREDICT_SCRIPT* that is used to score the model in the model settings table `model_setting_table`.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_score_function, 'MY_LM_PREDICT_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, the script *MY_LM_PREDICT_SCRIPT* is registered as:

```
function(object, newdata) {data.frame(pre = predict(object, newdata =
newdata))}
```

Classification: Each column represents the predicted probability of one target class. The column name is the target class name.

Example 6-9 Example of RALG_SCORE_FUNCTION for Classification

This example shows how to specify the name of the R script *MY_LOGITGLM_PREDICT_SCRIPT* that is used to score the logit Classification model in the model settings table `model_setting_table`.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_score_function, 'MY_LOGITGLM_PREDICT_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, *MY_LOGITGLM_PREDICT_SCRIPT* is registered as follows. It is a logit Classification with two target classes, "0" and "1".

```
'function(object, newdata) {
   pred <- predict(object, newdata = newdata, type="response");
   res <- data.frame(1-pred, pred);
   names(res) <- c("0", "1");
   res}'
```

Clustering: Each column represents the predicted probability of one cluster. The columns are arranged in order of cluster ID. Each cluster is assigned a cluster ID, and they are consecutive values starting from 1. To support `CLUSTER_DISTANCE` in the R model, the output of R score function returns an extra column containing the value of the distance to each cluster in order of cluster ID after the columns for the predicted probability.

Example 6-10 Example of RALG_SCORE_FUNCTION for Clustering

This example shows how to specify the name of the R script *MY_CLUSTER_PREDICT_SCRIPT* that is used to score the model in the model settings table `model_setting_table`.

```
Begin
insert into model_setting_table values
```

```
(dbms_data_mining.ralg_score_function, 'MY_CLUSTER_PREDICT_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, the script *MY_CLUSTER_PREDICT_SCRIPT* is registered as:

```
'function(object, dat){
     mod <- object[[1L]]; ce <- object[[2L]]; sc <- object[[3L]];
     newdata = scale(dat, center = ce, scale = sc);
     centers <- mod$centers;
     ss <- sapply(as.data.frame(t(centers)),
     function(v) rowSums(scale(newdata, center=v, scale=FALSE)^2));
     if (!is.matrix(ss)) ss <- matrix(ss, ncol=length(ss));
     disp <- -1 / (2* mod$tot.withinss/length(mod$cluster));
     distr <- exp(disp*ss);
     prob <- distr / rowSums(distr);
     as.data.frame(cbind(prob, sqrt(ss)))}'
```

Feature Extraction: Each column represents the coefficient value of one feature. The columns are arranged in order of feature ID. Each feature is assigned a feature ID, which are consecutive values starting from 1.

Example 6-11 Example of RALG_SCORE_FUNCTION for Feature Extraction

This example shows how to specify the name of the R script *MY_FEATURE_EXTRACTION_SCRIPT* that is used to score the model in the model settings table `model_setting_table`.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_score_function, 'MY_FEATURE_EXTRACTION_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, the script *MY_FEATURE_EXTRACTION_SCRIPT* is registered as:

```
 'function(object, dat) { as.data.frame(predict(object, dat)) }'
```

The function fetches the centers of the features from the R model, and computes the feature coefficient based on the distance of the score data to the corresponding feature center.

Related Topics

- Registered R Scripts
  The `RALG_*_FUNCTION` settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

RALG_WEIGHT_FUNCTION

Use the `RALG_WEIGHT_FUNCTION` setting to specify the name of an existing registered R script that computes the weight or contribution for each attribute in scoring. The specified R script is used in the SQL function `PREDICTION_DETAILS` to evaluate attribute contribution.

The specified R script defines an R function containing the first input argument for a model object, and the second input argument of an R `data.frame` for scoring data. When the machine learning function is Classification, Clustering, or Feature Extraction, the target class name, cluster ID, or feature ID is passed by the third input argument to compute the weight for that particular class, cluster, or feature. The script returns a `data.frame` containing the contributing weight for each attribute in a row. Each row corresponds to that input scoring `data.frame`.

Example 6-12 Example of RALG_WEIGHT_FUNCTION

This example specifies the name of the R script *`MY_PREDICT_WEIGHT_SCRIPT`* that computes the weight or contribution of R model attributes in the `model_setting_table`.

```
Begin
insert into model_setting_table values
(dbms_data_mining.ralg_weight_function, 'MY_PREDICT_WEIGHT_SCRIPT');
End;
/
```

In the Oracle Machine Learning for R script repository, the script *`MY_PREDICT_WEIGHT_SCRIPT`* for Regression is registered as:

```
'function(mod, data) { coef(mod)[-1L]*data }'
```

In the Oracle Machine Learning for R script repository, the script *`MY_PREDICT_WEIGHT_SCRIPT`* for logit Classification is registered as:

```
'function(mod, dat, clas) {
   v <- predict(mod, newdata=dat, type = "response");
   v0 <- data.frame(v, 1-v); names(v0) <- c("0", "1");
   res <- data.frame(lapply(seq_along(dat),
   function(x, dat) {
   if(is.numeric(dat[[x]])) dat[,x] <- as.numeric(0)
   else dat[,x] <- as.factor(NA);
   vv <- predict(mod, newdata = dat, type = "response");
   vv = data.frame(vv, 1-vv); names(vv) <- c("0", "1");
   v0[[clas]] / vv[[clas]]}, dat = dat));
   names(res) <- names(dat);
   res}'
```

Related Topics

- Registered R Scripts
  The `RALG_*_FUNCTION` settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

Registered R Scripts

The RALG_*_FUNCTION settings must specify R scripts that exist in the Oracle Machine Learning for R script repository.

You can register the R scripts using the Oracle Machine Learning for R SQL procedure sys.rqScriptCreate. To register a scripts, you must have the RQADMIN role.

The RALG_*_FUNCTION settings include the following functions:

- RALG_BUILD_FUNCTION
- RALG_DETAILS_FUNCTION
- RALG_SCORE_FUNCTION
- RALG_WEIGHT_FUNCTION

> **Note:**
>
> The R scripts must exist in the Oracle Machine Learning for R script repository for an R model to function.

After an R model is built, the name of the specified R script become a model setting. These R script must exist in the Oracle Machine Learning for R script repository for an R model to remain functional.

You can manage the R memory that is used to build, score, and view the R models through Oracle Machine Learning for R as well.

Algorithm Metadata Registration

Algorithm metadata registration allows for a uniform and consistent approach of registering new algorithm functions and their settings.

User have the ability to add new algorithms through the REGISTER_ALGORITHM procedure registration process. The new algorithms can appear as available within Oracle Machine Learning for SQL for their appropriate machine learning functions. Based on the registration metadata, the settings page is dynamically rendered. Algorithm metadata registration extends the machine learning model capability of Oracle Machine Learning for SQL.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*
- FETCH_JSON_SCHEMA Procedure
- REGISTER_ALGORITHM Procedure
- JSON Schema for R Extensible Algorithm

Model Settings in the Data Dictionary

Explains about ALL/USER/DBA_MINING_MODEL_SETTINGS in data dictionary view.

Information about Oracle Machine Learning model settings can be obtained from the data dictionary view ALL/USER/DBA_MINING_MODEL_SETTINGS. When used with the ALL prefix, this view returns information about the settings for the models accessible to the current user. When used with the `USER` prefix, it returns information about the settings for the models in the user's schema. The `DBA` prefix is only available for DBAs.

The columns of `ALL_MINING_MODEL_SETTINGS` are described as follows and explained in the following table.

```
describe all_mining_model_settings
```

The output is as follows:

```
Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                     NOT NULL VARCHAR2(30)
 MODEL_NAME                                NOT NULL VARCHAR2(30)
 SETTING_NAME                              NOT NULL VARCHAR2(30)
 SETTING_VALUE                                      VARCHAR2(4000)
 SETTING_TYPE                                       VARCHAR2(7)
```

Table 6-4   ALL_MINING_MODEL_SETTINGS

Column	Description
`owner`	Owner of the machine learning model.
`model_name`	Name of the machine learning model.
`setting_name`	Name of the setting.
`setting_value`	Value of the setting.
`setting_type`	`INPUT` if the value is specified by a user. `DEFAULT` if the value is system-generated.

The following query lists the settings for the Support Vector Machine (SVM) classification model SVMC_SH_CLAS_SAMPLE. The `ALGO_NAME`, `CLAS_WEIGHTS_TABLE_NAME`, and `SVMS_KERNEL_FUNCTION` settings are user-specified. These settings have been specified in a settings table for the model. The SVMC_SH_CLAS_SAMPLE model is created by the `oml4sql-classification-svm.sql` example.

Example 6-13   ALL_MINING_MODEL_SETTINGS

```
COLUMN setting_value FORMAT A35
 SELECT setting_name, setting_value, setting_type
            FROM all_mining_model_settings
            WHERE model_name in 'SVMC_SH_CLAS_SAMPLE';
```

The output is as follows:

```
SETTING_NAME                   SETTING_VALUE                       SETTING
------------------------------ ----------------------------------- -------
SVMS_ACTIVE_LEARNING           SVMS_AL_ENABLE                      DEFAULT
PREP_AUTO                      OFF                                 DEFAULT
SVMS_COMPLEXITY_FACTOR         0.244212                            DEFAULT
SVMS_KERNEL_FUNCTION           SVMS_LINEAR                         INPUT
```

```
CLAS_WEIGHTS_TABLE_NAME        svmc_sh_sample_class_wt           INPUT
SVMS_CONV_TOLERANCE            .001                              DEFAULT
ALGO_NAME                      ALGO_SUPPORT_VECTOR_MACHINES      INPUT
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

ALL_MINING_MODELS

Describes an example of ALL_MINING_MODELS and shows a sample query.

The following example describes ALL_MINING_MODELS and shows a sample query.

Example 6-14   ALL_MINING_MODELS

```
 describe ALL_MINING_MODELS
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                     NOT NULL VARCHAR2(128)
 MODEL_NAME                                NOT NULL VARCHAR2(128)
 MINING_FUNCTION                                    VARCHAR2(30)
 ALGORITHM                                          VARCHAR2(30)
 CREATION_DATE                             NOT NULL DATE
 BUILD_DURATION                                     NUMBER
 MODEL_SIZE                                         NUMBER
 BUILD_SOURCE                        CLOB
 PARTITIONED                                        VARCHAR2(3)
 COMMENTS                                           VARCHAR2(4000)
```

The following query returns the models accessible to you that use the Support Vector Machine algorithm.

```
SELECT mining_function, model_name
    FROM all_mining_models
    WHERE algorithm = 'SUPPORT_VECTOR_MACHINES'
    ORDER BY mining_function, model_name;



MINING_FUNCTION
MODEL_NAME
-------------------------
--------------------
CLASSIFICATION
PART2_CLAS_SAMPLE
CLASSIFICATION
PART_CLAS_SAMPLE
CLASSIFICATION
SVMC_SH_CLAS_SAMPLE
CLASSIFICATION
SVMO_SH_CLAS_SAMPLE
CLASSIFICATION
```

```
T_SVM_CLAS_SAMPLE
REGRESSION                SVMR_SH_REGR_SAMPLE
```

The models are created by the following examples:

- PART2_CLAS_SAMPLE by `oml4sql-partitioned-models-svm.sql`
- PART_CLAS_SAMPLE by `oml4sql-partitioned-models-svm.sql`
- SVMC_SH_CLAS_SAMPLE by `oml4sql-classification-svm.sql`
- SVMO_SH_CLAS_SAMPLE by `oml4sql-anomaly-detection-1class-svm.sql`
- T_SVM_CLAS_SAMPLE by `oml4sql-classification-text-mining-svm.sql`
- SVMR_SH_REGR_SAMPLE by `oml4sql-regression-svm.sql`

ALL_MINING_MODEL_ATTRIBUTES

Describes an example of `ALL_MINING_MODEL_ATTRIBUTES` and shows a sample query.

The following example describes `ALL_MINING_MODEL_ATTRIBUTES` and shows a sample query. Attributes are the predictors or conditions that are used to create models and score data.

Example 6-15 ALL_MINING_MODEL_ATTRIBUTES

```
describe ALL_MINING_MODEL_ATTRIBUTES
```

The output is as follows:

```

 Name                                       Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                      NOT NULL VARCHAR2(128)
 MODEL_NAME                                 NOT NULL VARCHAR2(128)
 ATTRIBUTE_NAME                             NOT NULL VARCHAR2(128)
 ATTRIBUTE_TYPE                                      VARCHAR2(11)
 DATA_TYPE                                           VARCHAR2(106)
 DATA_LENGTH                                         NUMBER
 DATA_PRECISION                                      NUMBER
 DATA_SCALE                                          NUMBER
 USAGE_TYPE                                          VARCHAR2(8)
 TARGET                                              VARCHAR2(3)
 ATTRIBUTE_SPEC                                      VARCHAR2(4000)
```

The following query returns the attributes of an SVM classification model named T_SVM_CLAS_SAMPLE. The model has both categorical and numerical attributes and includes one attribute that is unstructured text. The model is created by the `oml4sql-classification-text-mining-svm.sql` example

```
SELECT attribute_name, attribute_type, target
    FROM all_mining_model_attributes
    WHERE model_name = 'T_SVM_CLAS_SAMPLE'
    ORDER BY attribute_name;
```

The output is as follows:

```
ATTRIBUTE_NAME            ATTRIBUTE_TYPE
TAR
------------------------- --------------------
---
AFFINITY_CARD             CATEGORICAL
YES
AGE                       NUMERICAL
NO
BOOKKEEPING_APPLICATION   NUMERICAL
NO
BULK_PACK_DISKETTES       NUMERICAL
NO
COMMENTS                  TEXT
NO
COUNTRY_NAME              CATEGORICAL
NO
CUST_GENDER               CATEGORICAL
NO
CUST_INCOME_LEVEL         CATEGORICAL
NO
CUST_MARITAL_STATUS       CATEGORICAL
NO
EDUCATION                 CATEGORICAL
NO
FLAT_PANEL_MONITOR        NUMERICAL
NO
HOME_THEATER_PACKAGE      NUMERICAL
NO
HOUSEHOLD_SIZE            CATEGORICAL
NO
OCCUPATION                CATEGORICAL
NO
OS_DOC_SET_KANJI          NUMERICAL
NO
PRINTER_SUPPLIES          NUMERICAL
NO
YRS_RESIDENCE             NUMERICAL
NO
Y_BOX_GAMES               NUMERICAL            NO
```

ALL_MINING_MODEL_PARTITIONS

Describes an example of `ALL_MINING_MODEL_PARTITIONS` and shows a sample query.

The following example describes `ALL_MINING_MODEL_PARTITIONS` and shows a sample query.

Example 6-16   ALL_MINING_MODEL_PARTITIONS

```
describe ALL_MINING_MODEL_PARTITIONS
```

The output is as follows:

```
Name                                       Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                      NOT NULL VARCHAR2(128)
 MODEL_NAME                                 NOT NULL VARCHAR2(128)
 PARTITION_NAME                                      VARCHAR2(128)
 POSITION                                            NUMBER
 COLUMN_NAME                                NOT NULL VARCHAR2(128)
 COLUMN_VALUE                                        VARCHAR2(4000)
```

The following query returns the partition names and partition key values for two partitioned models. Model PART2_CLAS_SAMPLE has a two column partition key with system-generated partition names. The models are created by the `oml4sql-partitioned-models-svm.sql` example.

```
SELECT model_name, partition_name, position, column_name, column_value
    FROM all_mining_model_partitions
    ORDER BY model_name, partition_name, position;
```

The output is as follows:

```
MODEL_NAME           PARTITION_ POSITION COLUMN_NAME
COLUMN_VALUE
-------------------- ---------- -------- --------------------
----------------
PART2_CLAS_SAMPLE    DM$$_P0           1 CUST_GENDER
F
PART2_CLAS_SAMPLE    DM$$_P0           2 CUST_INCOME_LEVEL
HIGH
PART2_CLAS_SAMPLE    DM$$_P1           1 CUST_GENDER
F
PART2_CLAS_SAMPLE    DM$$_P1           2 CUST_INCOME_LEVEL
LOW
PART2_CLAS_SAMPLE    DM$$_P2           1 CUST_GENDER
F
PART2_CLAS_SAMPLE    DM$$_P2           2 CUST_INCOME_LEVEL
MEDIUM
PART2_CLAS_SAMPLE    DM$$_P3           1 CUST_GENDER
M
PART2_CLAS_SAMPLE    DM$$_P3           2 CUST_INCOME_LEVEL
HIGH
PART2_CLAS_SAMPLE    DM$$_P4           1 CUST_GENDER
M
PART2_CLAS_SAMPLE    DM$$_P4           2 CUST_INCOME_LEVEL
LOW
PART2_CLAS_SAMPLE    DM$$_P5           1 CUST_GENDER
M
PART2_CLAS_SAMPLE    DM$$_P5           2 CUST_INCOME_LEVEL
MEDIUM
```

```
PART_CLAS_SAMPLE     F                  1 CUST_GENDER
F
PART_CLAS_SAMPLE     M                  1 CUST_GENDER
M
PART_CLAS_SAMPLE     U                  1 CUST_GENDER          U
```

ALL_MINING_MODEL_SETTINGS

Describes an example of `ALL_MINING_MODEL_SETTINGS` and shows a sample query.

The following example describes `ALL_MINING_MODEL_SETTINGS` and shows a sample query. Settings influence model behavior. Settings may be specific to an algorithm or to a machine learning technique, or they may be general.

Example 6-17 ALL_MINING_MODEL_SETTINGS

```
 describe ALL_MINING_MODEL_SETTINGS
```

The output is as follows:

```
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                     NOT NULL VARCHAR2(128)
 MODEL_NAME                                NOT NULL VARCHAR2(128)
 SETTING_NAME                              NOT NULL VARCHAR2(30)
 SETTING_VALUE                                      VARCHAR2(4000)
 SETTING_TYPE                                       VARCHAR2(7)
```

The following query returns the settings for a model named SVD_SH_SAMPLE. The model uses the Singular Value Decomposition algorithm for feature extraction. The model is created by the `oml4sql-singular-value-decomposition.sql` example.

```
SELECT setting_name, setting_value, setting_type
    FROM all_mining_model_settings
    WHERE model_name = 'SVD_SH_SAMPLE'
    ORDER BY setting_name;
```

The output is as follows:

```
SETTING_NAME                   SETTING_VALUE
SETTING
------------------------------ ------------------------------
-------
ALGO_NAME                      ALGO_SINGULAR_VALUE_DECOMP
INPUT
ODMS_DETAILS                   ODMS_ENABLE                    DEFAULT
ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
DEFAULT
ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE
DEFAULT
PREP_AUTO                      OFF
INPUT
```

```
SVDS_SCORING_MODE               SVDS_SCORING_SVD
DEFAULT
SVDS_U_MATRIX_OUTPUT            SVDS_U_MATRIX_ENABLE           INPUT
```

ALL_MINING_MODEL_VIEWS

Describes an example of `ALL_MINING_MODEL_VIEWS` and shows a sample query.

The following example describes `ALL_MINING_MODEL_VIEWS` and shows a sample query. Model views provide details on the models.

Example 6-18   ALL_MINING_MODEL_VIEWS

```
describe ALL_MINING_MODEL_VIEWS
```

The output is as follows:

```
 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                     NOT NULL VARCHAR2(128)
 MODEL_NAME                                NOT NULL VARCHAR2(128)
 VIEW_NAME                                 NOT NULL VARCHAR2(128)
 VIEW_TYPE                                          VARCHAR2(128)
```

The following query returns the model views for the SVD_SH_SAMPLE model. The model uses the Singular Value Decomposition algorithm for feature extraction. The model is created by the `oml4sql-singular-value-decomposition.sql` example.

```
SELECT view_name, view_type
    FROM all_mining_model_views
    WHERE model_name = 'SVD_SH_SAMPLE'
    ORDER BY view_name;
```

The output is as follows:

```
VIEW_NAME
VIEW_TYPE
-------------------------
--------------------------------------------------
DM$VESVD_SH_SAMPLE        Singular Value Decomposition S
Matrix
DM$VGSVD_SH_SAMPLE        Global Name-Value
Pairs
DM$VNSVD_SH_SAMPLE        Normalization and Missing Value
Handling
DM$VSSVD_SH_SAMPLE        Computed
Settings
DM$VUSVD_SH_SAMPLE        Singular Value Decomposition U
Matrix
DM$VVSVD_SH_SAMPLE        Singular Value Decomposition V
```

```
Matrix
DM$VWSVD_SH_SAMPLE        Model Build Alerts
```

ALL_MINING_MODEL_XFORMS

Describes an example of `ALL_MINING_MODEL_XFORMS` and provides a sample query.

The following example describes `ALL_MINING_MODEL_XFORMS` and provides a sample query.

Example 6-19 ALL_MINING_MODEL_XFORMS

```
describe ALL_MINING_MODEL_XFORMS



 Name                                      Null?    Type
 ----------------------------------------- --------
----------------------------
 OWNER                                     NOT NULL VARCHAR2(128)
 MODEL_NAME                                NOT NULL VARCHAR2(128)
 ATTRIBUTE_NAME                                     VARCHAR2(128)
 ATTRIBUTE_SUBNAME                                  VARCHAR2(4000)
 ATTRIBUTE_SPEC                                     VARCHAR2(4000)
 EXPRESSION                                         CLOB
 REVERSE                                            VARCHAR2(3)
```

The following query returns the embedded transformations for a model `PART2_CLAS_SAMPLE`. The model is created by the `oml4sql-partitioned-models-svm.sql` example.

```
SELECT attribute_name, expression
    FROM all_mining_model_xforms
    WHERE model_name = 'PART2_CLAS_SAMPLE'
    ORDER BY attribute_name;
```

The output is as follows:

```

ATTRIBUTE_NAME

-------------------------

EXPRESSION

------------------------------------------------------------------------------
--
CUST_INCOME_LEVEL

CASE CUST_INCOME_LEVEL WHEN 'A: Below 30,000' THEN
'LOW'
    WHEN 'L: 300,000 and above' THEN
'HIGH'
    ELSE 'MEDIUM' END
```

DBMS_DATA_MINING

The `DBMS_DATA_MINING` package is the application programming interface for creating, evaluating, and querying Oracle Machine Learning for SQL models.

In Oracle Database Release 21c, Oracle Data Mining has been rebranded to Oracle Machine Learning for SQL (Oracle Machine Learning for SQL). The PL/SQL package name, however, has not changed and remains `DBMS_DATA_MINING`.

This chapter contains the following topics:

- Overview
- Security Model
- Mining Functions
- Model Settings
- Algorithm Specific Settings
- Solver Settings
- Datatypes
- Summary of DBMS_DATA_MINING Subprograms

> **See Also:**
>
> - *Oracle Machine Learning for SQL Concepts*
> - *Oracle Machine Learning for SQL User’s Guide*
> - DBMS_DATA_MINING_TRANSFORM
> - DBMS_PREDICTIVE_ANALYTICS

DBMS_DATA_MINING Overview

Oracle Machine Learning for SQL supports both supervised and unsupervised machine learning. Supervised machine learning predicts a target value based on historical data. Unsupervised machine learning discovers natural groupings and does not use a target. You can use Oracle Machine Learning for SQL procedures on structured data and unstructured text.

Supervised machine learning techniques include:

- Classification
- Regression
- Feature Selection (Attribute Importance)
- Time Series

Unsupervised machine learning techniques include:

- Clustering
- Association

- Feature Extraction
- Anomaly Detection

The steps you use to build and apply a machine learning model depend on the machine learning technique and the algorithm being used. The algorithms supported by Oracle Machine Learning for SQL are listed in the following table.

Table 6-5 Oracle Machine Learning for SQL Algorithms

Algorithm	Abbreviation	Function
Apriori	AR	Association
CUR Matrix Decomposition	CUR	Attribute importance
Decision Tree	DT	Classification
Expectation Maximization	EM	Clustering
Explicit Semantic Analysis	ESA	Feature extraction, classification
Exponential Smoothing	ESM	Time series
Generalized Linear Models	GLM	Classification, regression
k-Means	KM	Clustering
Minimum Descriptor Length	MDL	Attribute importance
Multivariate State Estimation Technique - Sequential Probability Ratio Test	MSET-SPRT	Anomaly detection, classification
Naive Bayes	NB	Classification
Neural Network	NN	Classification, regression
Non-Negative Matrix Factorization	NMF	Feature extraction
Orthogonal Partitioning Clustering	O-Cluster	Clustering
Random Forest	RF	Classification
Singular Value Decomposition and Principal Component Analysis	SVD and PCA	Feature extraction
Support Vector Machine	SVM	Classification, regression, anomaly detection
XGBoost	XGBoost	Classification, regression

Oracle Machine Learning for SQL supports more than one algorithm for the classification, regression, clustering, and feature extraction machine learning techniques. Each of these machine learning techniques has a default algorithm, as shown in the following table.

Table 6-6 Oracle Machine Learning for SQL Default Algorithms

Mining Function	Default Algorithm
Classification	Naive Bayes
Clustering	*k*-Means
Feature Extraction	Non-Negative Matrix Factorization
Feature Selection	Minimum Descriptor Length
Regression	Support Vector Machine
Time Series	Exponential Smoothing

DBMS_DATA_MINING Security Model

The DBMS_DATA_MINING package is owned by user SYS and is installed as part of database installation. Execution privilege on the package is granted to public. The routines in the package are run with invokers' rights (run with the privileges of the current user).

The DBMS_DATA_MINING package exposes APIs that are leveraged by the Oracle Machine Learning for SQL. Users who wish to create machine learning models in their own schema require the CREATE MINING MODEL system privilege. Users who wish to create machine learning models in other schemas require the CREATE ANY MINING MODEL system privilege.

Users have full control over managing models that exist within their own schema. Additional system privileges necessary for managing machine learning models in other schemas include ALTER ANY MINING MODEL, DROP ANY MINING MODEL, SELECT ANY MINING MODEL, COMMENT ANY MINING MODEL, and AUDIT ANY.

Individual object privileges on machine learning models, ALTER MINING MODEL and SELECT MINING MODEL, can be used to selectively grant privileges on a model to a different user.

See Also:

Oracle Data Mining User's Guide for more information about the security features of Oracle Machine Learning for SQL

DBMS_DATA_MINING — Machine Learning Functions

A machine learning **function** refers to the methods for solving a given class of machine learning problems.

The machine learning function must be specified when a model is created. You specify a machine learning function with the mining_function parameter of the CREATE_MODEL Procedure or the CREATE_MODEL2 Procedure.

Table 6-7 Machine Learning Functions

Value	Description
ASSOCIATION	Association is a descriptive machine learning function. An association model identifies relationships and the probability of their occurrence within a data set. Association models use the Apriori algorithm.
ATTRIBUTE_IMPORTANCE	Attribute importance is a predictive machine learning function, also known as feature selection. An attribute importance model identifies the relative importance of an attribute in predicting a given outcome. Attribute importance models can use Minimum Description Length (MDL) or CUR Matrix Decomposition. MDL is the default.

Table 6-7 (Cont.) Machine Learning Functions

Value	Description
`CLASSIFICATION`	Classification is a predictive machine learning function. A classification model uses historical data to predict a categorical target. Classification models can use: Decision Tree, logistic regression, Multivariate State Estimation Technique - Sequential Probability Ratio Test, Naive Bayes, Support Vector Machine (SVM), or XGBoost. The default is Naive Bayes. The classification function can also be used for **anomaly detection**. For anomaly detection, you can use the Multivariate State Estimation Technique - Sequential Probability Ratio Test algorithm or the SVM algorithm with a null target (One-Class SVM), or the EM algorithm with a null target (EM Anomaly).
`CLUSTERING`	Clustering is a descriptive machine learning function. A clustering model identifies natural groupings within a data set. Clustering models can use *k*-Means, O-Cluster, or Expectation Maximization. The default is *k*-Means.
`FEATURE_EXTRACTION`	Feature extraction is a descriptive machine learning function. A feature extraction model creates an optimized data set on which to base a model. Feature extraction models can use Explicit Semantic Analysis, Non-Negative Matrix Factorization, Singular Value Decomposition, or Principal Component Analysis. Non-Negative Matrix Factorization is the default.
`REGRESSION`	Regression is a predictive machine learning function. A regression model uses historical data to predict a numerical target. Regression models can use linear regression, Support Vector Machine, or XGBoost. The default is Support Vector Machine.
`TIME_SERIES`	Time series is a predictive machine learning function. A time series model forecasts the future values of a time-ordered series of historical numeric data over a user-specified time window. Time series models use the Exponential Smoothing algorithm.

See Also:

Oracle Machine Learning for SQL Concepts for more information about mining functions

DBMS_DATA_MINING — Model Settings

Oracle Machine Learning for SQL uses settings to specify the algorithm and other characteristics of a model. Some settings are general, some are specific to a machine learning function, and some are specific to an algorithm.

All settings have default values. If you want to override one or more of the settings for a model, then you must create a settings table. The settings table must have the column names and data types shown in the following table.

Table 6-8 Required Columns in the Model Settings Table

Column Name	Data Type
`SETTING_NAME`	`VARCHAR2(30)`
`SETTING_VALUE`	`VARCHAR2(4000)`

The information you provide in the settings table is used by the model at build time. The name of the settings table is an optional argument to the CREATE_MODEL Procedure. You can also provide these settings through the CREATE_MODEL2 Procedure.

The settings used by a model can be found by querying the data dictionary view `ALL_MINING_MODEL_SETTINGS`. This view displays the model settings used by the machine learning models to which you have access. All of the default and user-specified setting values are included in the view.

> **See Also:**
>
> - `ALL_MINING_MODEL_SETTINGS` in *Oracle Database Reference*
> - *Oracle Machine Learning for SQL User’s Guide* for information about specifying model settings

DBMS_DATA_MINING — Algorithm Names

The `ALGO_NAME` setting specifies the model algorithm.

The values for the `ALGO_NAME` setting are listed in the following table.

Table 6-9 Algorithm Names

ALGO_NAME Value	Description	Machine Learning Function
`ALGO_AI_MDL`	Minimum Description Length	Attribute importance
`ALGO_APRIORI_ASSOCIATION_RULES`	Apriori	Association rules
`ALGO_CUR_DECOMPOSITION`	CUR Matrix Decomposition	Attribute importance
`ALGO_DECISION_TREE`	Decision Tree	Classification
`ALGO_EXPECTATION_MAXIMIZATION`	Expectation Maximization	Clustering, Classification
`ALGO_EXPLICIT_SEMANTIC_ANALYS`	Explicit Semantic Analysis	Feature extraction Classification
`ALGO_EXPONENTIAL_SMOOTHING`	Exponential Smoothing	Time series
`ALGO_EXTENSIBLE_LANG`	Language used for extensible algorithm	All mining functions supported
`ALGO_GENERALIZED_LINEAR_MODEL`	Generalized Linear Model	Classification, regression; also feature selection and generation
`ALGO_KMEANS`	Enhanced *k*-Means	Clustering

Table 6-9 (Cont.) Algorithm Names

ALGO_NAME Value	Description	Machine Learning Function
ALGO_MSET_SPRT	Multivariate State Estimation Technique - Sequential Probability Ratio Test	Classification
ALGO_NAIVE_BAYES	Naive Bayes	Classification
ALGO_NEURAL_NETWORK	Neural Network	Classification
ALGO_NONNEGATIVE_MATRIX_FACTOR	Non-Negative Matrix Factorization	Feature extraction
ALGO_O_CLUSTER	O-Cluster	Clustering
ALGO_RANDOM_FOREST	Random Forest	Classification
ALGO_SINGULAR_VALUE_DECOMP	Singular Value Decomposition	Feature extraction
ALGO_SUPPORT_VECTOR_MACHINES	Support Vector Machine	Classification and regression
ALGO_XGBOOST	XGBoost	Classification and regression

See Also:

Oracle Machine Learning for SQL Concepts for information about algorithms

DBMS_DATA_MINING — Automatic Data Preparation

Oracle Machine Learning for SQL supports fully Automatic Data Preparation (ADP), user-directed general data preparation, and user-specified embedded data preparation. The PREP_* settings enable the user to request fully automated or user-directed general data preparation. By default, fully Automatic Data Preparation (PREP_AUTO_ON) is enabled.

When you enable ADP, the model uses heuristics to transform the build data according to the requirements of the algorithm. Instead of fully ADP, the user can request that the data be shifted and/or scaled with the PREP_SCALE* and PREP_SHIFT* settings. The transformation instructions are stored with the model and reused whenever the model is applied. The model settings can be viewed in USER_MINING_MODEL_SETTINGS.

You can choose to supplement Automatic Data Preparations by specifying additional transformations in the xform_list parameter when you build the model. See "CREATE_MODEL Procedure" and "CREATE_MODEL2 Procedure".

If you do not use ADP *and* do not specify transformations in the xform_list parameter to CREATE_MODEL, you must implement your own transformations separately in the build, test, and scoring data. You must take special care to implement the exact same transformations in each data set.

If you do not use ADP, but you *do* specify transformations in the xform_list parameter to CREATE_MODEL, OML4SQL embeds the transformation definitions in the model and prepares the test and scoring data to match the build data.

The values for the PREP_* setting are described in the following table.

The **Constant Value** column specifies constants using the prefix DBMS_DATA_MINING. For example, DBMS_DATA_MINING.PREP_AUTO_ON. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the DBMS_DATA_MINING prefix, in single quotes. For example, 'ON'.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-10 PREP_* Setting

Setting Name	Constant Value	String Value Equivalent	Description
PREP_AUTO	PREP_AUTO_ON	ON	This setting enables fully automated data preparation. The default is PREP_AUTO_ON.
	PREP_AUTO_OFF	OFF	Disables fully automated data preparation.
PREP_SCALE_2DNUM	PREP_SCALE_STDDEV	PREP_SCALE_STDDEV	This setting enables scaling data preparation for two-dimensional numeric columns. PREP_AUTO must be OFF for this setting to take effect. PREP_SCALE_STDDEV: A request to divide the column values by the standard deviation of the column and is often provided together with PREP_SHIFT_MEAN to yield z-score normalization.
	PREP_SCALE_RANGE	PREP_SCALE_RANGE	A request to divide the column values by the range of values and is often provided together with PREP_SHIFT_MIN to yield a range of [0,1].
PREP_SCALE_NNUM	PREP_SCALE_MAXABS	PREP_SCALE_MAXABS	This setting enables scaling data preparation for nested numeric columns. PREP_AUTO must be OFF for this setting to take effect. If specified, then the valid value for this setting is PREP_SCALE_MAXABS, which yields data in the range of [-1,1].
PREP_SHIFT_2DNUM	PREP_SHIFT_MEAN	PREP_SHIFT_MEAN	This setting enables centering data preparation for two-dimensional numeric columns. PREP_AUTO must be OFF for this setting to take effect. PREP_SHIFT_MEAN: Results in subtracting the average of the column from each value.
	PREP_SHIFT_MIN	PREP_SHIFT_MIN	Results in subtracting the minimum of the column from each value.

See Also:

Oracle® Machine Learning for SQL for information about data transformations

DBMS_DATA_MINING — Machine Learning Function Settings

The settings described in this table apply to a machine learning function.

Table 6-11 Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description
Association	`ASSO_MAX_RULE_LENGTH`	`TO_CHAR( 2<=` *`numeric_expr`* `<=20)`	Maximum rule length for association rules. Default is `4`.
Association	`ASSO_MIN_CONFIDENCE`	`TO_CHAR( 0<=` *`numeric_expr`* `<=1)`	Minimum confidence for association rules. Default is `0.1`.
Association	`ASSO_MIN_SUPPORT`	`TO_CHAR( 0<=` *`numeric_expr`* `<=1)`	Minimum support for association rules Default is `0.1`.
Association	`ASSO_MIN_SUPPORT_INT`	a positive integer	Minimum absolute support that each rule must satisfy. The value must be an integer. The default is `1`.
Association	`ASSO_MIN_REV_CONFIDENCE`	`TO_CHAR( 0<=` *`numeric_expr`* `<=1)`	Sets the Minimum Reverse Confidence that each rule should satisfy. The Reverse Confidence of a rule is defined as the number of transactions in which the rule occurs divided by the number of transactions in which the consequent occurs. The value is real number between 0 and 1. The default is `0`.

Association	`ASSO_IN_RULES`	`NULL`	Sets Including Rules applied for each association rule: it specifies the list of items that at least one of them must appear in each reported association rule, either as antecedent or as consequent. It is a comma separated string containing the list of including items.
			If not set, the default behavior is, the filtering is not applied.
			For example,
			`INSERT INTO sett_tab (setting_name, setting_value) VALUES`
			`      (dbms_data_mining.asso_in_rules, '''a'',''b''');`

Table 6-11 (Cont.) Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description

Association	`ASSO_EX_RULES`	`NULL`	Sets Excluding Rules applied for each association rule: it specifies the list of items that none of them can appear in each reported association rules. It is a comma separated string containing the list of excluded items. No rule can contain any item in the list.
			The default is `NULL`.
			For example,
			`INSERT INTO sett_tab (setting_name,`
			`setting_value) VALUES`
			`      (dbms_data_mining.asso_ex_rules,`
			`'''a'',''b''');`

Association	`ASSO_ANT_IN_RULES`	`NULL`	Sets Including Rules for the antecedent: it specifies the list of items that at least one of them must appear in the antecedent part of each reported association rule. It is a comma separated string containing the list of including items. The antecedent part of each rule must contain at least one item in the list.
			The default is `NULL`.
			For example,
			`INSERT INTO sett_tab (setting_name,`
			`setting_value) VALUES`
			
			`(dbms_data_mining.asso_ant_in_rules,`
			`'''a'',''b''');`

Association	`ASSO_ANT_EX_RULES`	`NULL`	Sets Excluding Rules for the antecedent: it specifies the list of items that none of them can appear in the antecedent part of each reported association rule. It is a comma separated string containing the list of excluded items. No rule can contain any item in the list in its antecedent part.
			The default is `NULL`.
			For example,
			`INSERT INTO sett_tab (setting_name,`
			`setting_value) VALUES`
			
			`(dbms_data_mining.asso_ant_ex_rules,`
			`'''a'',''b''');`

Table 6-11 (Cont.) Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description

Association	`ASSO_CONS_IN_RULES`	`NULL`	Sets Including Rules for the consequent: it specifies the list of items that at least one of them must appear in the consequent part of each reported association rule. It is a comma separated string containing the list of including items. The consequent of each rule must be an item in the list.
			The default is `NULL`.
			For example,
			`INSERT INTO sett_tab (setting_name, setting_value) VALUES`
			`(dbms_data_mining.asso_cons_in_rules, '''a'',''b''');`

Association	`ASSO_CONS_EX_RULES`	`NULL`	Sets Excluding Rules for the consequent: it specifies the list of items that none of them can appear in the consequent part of each reported association rule. It is a comma separated string containing the list of excluded items. No rule can have any item in the list as its consequent.
			The excluding rule can be used to reduce the data that must be stored, but the user may be required to build an extra model for executing different including or Excluding Rules.
			The default is `NULL`.
			For example,
			`INSERT INTO sett_tab (setting_name, setting_value) VALUES`
			`(dbms_data_mining.asso_cons_ex_rules, '''a'',''b''');`

Table 6-11 (Cont.) Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description
Association	ASSO_AGGREGATES	NULL	Specifies the columns to be aggregated. It is a comma separated string containing the names of the columns for aggregation. The number of columns in the list must be <= 10. You can set ASSO_AGGREGATES if ODMS_ITEM_ID_COLUMN_NAME is set indicating transactional input data. See DBMS_DATA_MINING - Global Settings. The data table must have valid column names such as ITEM_ID and CASE_ID which are derived from ODMS_ITEM_ID_COLUMN_NAME and case_id_column_name respectively. Numeric values are supported. ITEM_VALUE is not a mandatory value. The default is NULL. For each item, the user may supply several columns to aggregate. It requires more memory to buffer the extra data. Also, the performance impact can be seen because of the larger input data set and more operation.
Association	ASSO_ABS_ERROR	0<ASSO_ABS_ERRORMAX(ASSO_MIN_SUPPORT, ASSO_MIN_CONFIDENCE).	Specifies the absolute error for the association rules sampling. A smaller value of ASSO_ABS_ERROR obtains a larger sample size which gives accurate results but takes longer computational time. Set a reasonable value for ASSO_ABS_ERROR, such as its default value, to avoid large sample size. The default value is 0.5 * MAX(ASSO_MIN_SUPPORT, ASSO_MIN_CONFIDENCE).
Association	ASSO_CONF_LEVEL	0 ASSO_CONF_LEVEL 1	Specifies the confidence level for an association rules sample. A larger value of ASSO_CONF_LEVEL obtains a larger sample size. Any value between 0.9 and 1 is suitable. The default value is 0.95.
Classification	CLAS_COST_TABLE_NAME	*table_name*	(Decision tree only) Name of a table that stores a cost matrix to be used by the algorithm in building the model. The cost matrix specifies the costs associated with misclassifications. Only decision tree models can use a cost matrix at build time. All classification algorithms can use a cost matrix at apply time. The cost matrix table is user-created. See "ADD_COST_MATRIX Procedure" for the column requirements. See *Oracle Machine Learning for SQL Concepts* for information about costs.

Table 6-11 (Cont.) Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description
Classification	`CLAS_PRIORS_TABLE_NAME`	*table_name*	(Naive Bayes) Name of a table that stores prior probabilities to offset differences in distribution between the build data and the scoring data. The priors table is user-created. See *Oracle Machine Learning for SQL User's Guide* for the column requirements. See *Oracle Machine Learning for SQL Concepts* for additional information about priors.
Classification	`CLAS_WEIGHTS_TABLE_NAME`	*table_name*	(GLM and SVM only) Name of a table that stores weighting information for individual target values in SVM classification and GLM logistic regression models. The weights are used by the algorithm to bias the model in favor of higher weighted classes. The class weights table is user-created. See *Oracle Machine Learning for SQL User's Guide* for the column requirements. See *Oracle Machine Learning for SQL Concepts* for additional information about class weights.
Classification	`CLAS_WEIGHTS_BALANCED`	`ON` `OFF`	This setting indicates that the algorithm must create a model that balances the target distribution. This setting is most relevant in the presence of rare targets, as balancing the distribution may enable better average accuracy (average of per-class accuracy) instead of overall accuracy (which favors the dominant class). The default value is `OFF`.

Classification	`CLAS_MAX_SUP_BINS`	For Decision Tree:
		`2 <=` *`a number`* `<=2147483647`
		For Random Forest:
		`2 <=` *`a number`* `<=254`
			This parameter specifies the maximum number of bins for each attribute.
			The default value is `32`.
			See, DBMS_DATA_MINING — Automatic Data Preparation

Table 6-11 (Cont.) Machine Learning Function Settings

Machine Learning Function	Setting Name	Setting Value	Description
Clustering	`CLUS_NUM_CLUSTERS`	`TO_CHAR( numeric_expr >=1)`	The maximum number of leaf clusters generated by a clustering algorithm. The algorithm may return fewer clusters, depending on the data. Enhanced *k*-Means usually produces the exact number of clusters specified by `CLUS_NUM_CLUSTERS`, unless there are fewer distinct data points. When Expectation maximization (EM) is used for clustering, it may return fewer clusters than the number specified by `CLUS_NUM_CLUSTERS` depending on the data. The number of clusters returned by EM cannot be greater than the number of components, which is governed by algorithm-specific settings. (See *Expectation Maximization Settings for Learning* table) Depending on these settings, there may be fewer clusters than components. If component clustering is disabled, the number of clusters equals the number of components. The setting can be used only for EM Clustering algorithm. For EM Clustering algorithm, the default value of `CLUS_NUM_CLUSTERS` is system-determined. For *k*-Means and O-Cluster, the default is `10`.
Feature extraction	`FEAT_NUM_FEATURES`	`TO_CHAR( numeric_expr >=1)`	The number of features to be extracted by a feature extraction model. The default is estimated from the data by the algorithm. If the matrix rank is smaller than this number, fewer features will be returned. For CUR Matrix Decomposition, the `FEAT_NUM_FEATURES` value is the same as the `CURS_SVD_RANK` value.

See Also:

Oracle Machine Learning for SQL Concepts for information about machine learning functions

DBMS_DATA_MINING — Global Settings

The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

Table 6-12 Global Settings

Setting Name	Setting Value	Description
`ODMS_BOXCOX`	`ODMS_BOXCOX_ENABLE` `ODMS_BOXCOX_DISABLE`	This setting enables the Box-Cox variance-stabilization transformation. It is useful when the variance increases as the target value increases. It reduces variance and transforms a multiplicative relationship with the target, with a simpler additive relationship. This setting is applicable only to the Exponential Smoothing algorithm. When a value for `EXSM_MODEL` setting is not specified, the default value is `ODMS_BOXCOX_ENABLE` and when a value for the `EXSM_MODEL` setting is provided, the default value is `ODMS_BOXCOX_DISABLE`.
`ODMS_EXPLOSION_MIN_SUPP`	A positive integer	It is the minimum required support for categorical values that must be included in the explosion mapping. It removes categorical values with insufficient row instances to have a statistically significant effect on the model, however, they could potentially degrade performance. The default is system determined depending on the number of rows in the dataset. A value of `1` results into mapping all categorical values.
`ODMS_ITEM_ID_COLUMN_NAME`	*column_name*	(Association rules only) Name of a column that contains the items in a transaction. When this setting is specified, the algorithm expects the data to be presented in a native transactional format, consisting of two columns: • Case ID, either categorical or numeric • Item ID, either categorical or numeric **Note:** Oracle Machine Learning does not support `BOOLEAN` values for this setting. A typical example of transactional data is market basket data, wherein a case represents a basket that may contain many items. Each item is stored in a separate row, and many rows may be needed to represent a case. The case ID values do not uniquely identify each row. Transactional data is also called multi-record case data. Association rules function is normally used with transactional data, but it can also be applied to single-record case data (similar to other algorithms). For more information about single-record and multi-record case data, see *Oracle SQL Developer Data Modeler User's Guide*.

Table 6-12   (Cont.) Global Settings

Setting Name	Setting Value	Description
ODMS_ITEM_VALUE_COLUMN_NAME	*column_name*	(Association rules only) Name of a column that contains a value associated with each item in a transaction. This setting is only used when a value has been specified for ODMS_ITEM_ID_COLUMN_NAME indicating that the data is presented in native transactional format. If ASSO_AGGREGATES is used, then the build data must include the following three columns and the columns specified in the AGGREGATES setting. • Case ID, either categorical or numeric • Item ID, either categorical or numeric, specified by ODMS_ITEM_ID_COLUMN_NAME • Item value, either categorical or numeric, specified by ODMS_ITEM_VALUE_COLUMN_NAME **Note:** Oracle Machine Learning does not support BOOLEAN values for this setting. If ASSO_AGGREGATES, Case ID, and Item ID column are present, then the Item Value column may or may not appear. The Item Value column may specify information such as the number of items (for example, three apples) or the type of the item (for example, macintosh apples). For details on ASSO_AGGREGATES, see DBMS_DATA_MINING - Mining Function Settings.
ODMS_MISSING_VALUE_TREATMENT	ODMS_MISSING_VALUE_MEAN_MODE ODMS_MISSING_VALUE_DELETE_ROW ODMS_MISSING_VALUE_AUTO	Indicates how to treat missing values in the training data. This setting does not affect the scoring data. The default value is ODMS_MISSING_VALUE_AUTO. ODMS_MISSING_VALUE_MEAN_MODE replaces missing values with the mean (numeric attributes) or the mode (categorical attributes) both at build time and apply time where appropriate. ODMS_MISSING_VALUE_AUTO performs different strategies for different algorithms. When ODMS_MISSING_VALUE_TREATMENT is set to ODMS_MISSING_VALUE_DELETE_ROW, the rows in the training data that contain missing values are deleted. However, if you want to replicate this missing value treatment in the scoring data, then you must perform the transformation explicitly. The value ODMS_MISSING_VALUE_DELETE_ROW applies to all algorithms.

Table 6-12 (Cont.) Global Settings

Setting Name	Setting Value	Description
ODMS_ROW_WEIGHT_COLUMN_NAME	*column_name*	(GLM only) Name of a column in the training data that contains a weighting factor for the rows. The column data type must be numeric. Oracle Machine Learning does not support BOOLEAN values for this setting. Row weights can be used as a compact representation of repeated rows, as in the design of experiments where a specific configuration is repeated several times. Row weights can also be used to emphasize certain rows during model construction. For example, to bias the model towards rows that are more recent and away from potentially obsolete data.
ODMS_TEXT_POLICY_NAME	The name of an Oracle Text POLICY created using CTX_DDL.CREATE_POLICY.	Affects how individual tokens are extracted from unstructured text. For details about CTX_DDL.CREATE_POLICY, see *Oracle Text Reference*.
ODMS_TEXT_MAX_FEATURES	1 <= *value*	The maximum number of distinct features, across all text attributes, to use from a document set passed to CREATE_MODEL. The default is 3000. ESA has the default value of 300000.
ODMS_TEXT_MIN_DOCUMENTS	Non-negative value	This is a text processing setting the controls how in how many documents a token needs to appear to be used as a feature. The default is 1. ESA has a default of 3.
ODMS_PARTITION_COLUMNS	Comma separated list of machine learning attributes	This setting indicates a request to build a partitioned model. The setting value is a comma-separated list of the machine learning attributes used to determine the in-list partition key values. Oracle Machine Learning supports numeric and categorical values including BOOLEAN for this setting. These machine learning attributes are taken from the input columns unless an XFORM_LIST parameter is passed to CREATE_MODEL or CREATE_MODEL2. If the XFORM_LIST parameter is passed to during model building, then the machine learning attributes are taken from the attributes produced by these transformations.
ODMS_MAX_PARTITIONS	1< value <= 1000000	This setting indicates the maximum number of partitions allowed for the model. The default is 1000.
ODMS_SAMPLING	ODMS_SAMPLING_ENABLE ODMS_SAMPLING_DISABLE	This setting allows the user to request a sampling of the build data. The default is ODMS_SAMPLING_DISABLE.
ODMS_SAMPLE_SIZE	0 < Value	This setting determines how many rows will be sampled (approximately). It can be set only if ODMS_SAMPLING is enabled. The default value is the system determined.

Table 6-12 (Cont.) Global Settings

Setting Name	Setting Value	Description
ODMS_PARTITION_BUILD_TYPE	ODMS_PARTITION_BUILD_INTRA ODMS_PARTITION_BUILD_INTER ODMS_PARTITION_BUILD_HYBRID	This setting controls the parallel build of partitioned models. ODMS_PARTITION_BUILD_INTRA — Each partition is built in parallel using all replicas. ODMS_PARTITION_BUILD_INTER — Each partition is built entirely in a single slave, but multiple partitions may be built at the same time since multiple replicas are active. ODMS_PARTITION_BUILD_HYBRID — It is a combination of the other two types and is recommended for most situations to adapt to dynamic environments. The default mode is ODMS_PARTITION_BUILD_HYBRID
ODMS_TABLESPACE_NAME	*tablespace_name*	This setting controls the storage specifications. If you explicitly sets this to the name of a tablespace (for which you have sufficient quota), then the specified tablespace storage creates the resulting model content. If you do not provide this setting, then the default tablespace of the user creates the resulting model content.
ODMS_RANDOM_SEED	The value must be a non-negative integer	The hash function with a random number seed generates a random number with uniform distribution. Users can control the random number seed by this setting. The default is 0. This setting is used by Random Forest, Neural Network, and CUR Matrix Decomposition.
ODMS_DETAILS	• ODMS_ENABLE • ODMS_DISABLE	This setting reduces the space that is used while creating a model, especially a partitioned model. The default value is ODMS_ENABLE. When the setting is ODMS_ENABLE, it creates model tables and views when the model is created. You can query the model with SQL. When the setting is ODMS_DISABLE, model views are not created and tables relevant to model details are not created either. The reduction in space depends on the model. Reduction on the order of 10x can be achieved.

See Also:

Oracle Machine Learning for SQL Concepts for information about GLM

Oracle Machine Learning for SQL Concepts for information about association rules

Oracle Machine Learning for SQL User's Guide for information about machine learning unstructured text

DBMS_DATA_MINING — Algorithm Specific Model Settings

Oracle Machine Learning for SQL uses algorithm specific settings to define the characteristics of a model.

All settings have default values. If you want to override one or more of the settings for a model, then you must specify those settings.

The information you provide in the settings table is used by the model at build time. The name of the settings table is an optional argument to the CREATE_MODEL Procedure. You can also provide these settings through the CREATE_MODEL2 Procedure.

The settings used by a model can be found by querying the data dictionary view `ALL_MINING_MODEL_SETTINGS`. This view displays the model settings used by the machine learning models to which you have access. All of the default and user-specified setting values are included in the view.

> **See Also:**
>
> - `ALL_MINING_MODEL_SETTINGS` in *Oracle Database Reference*
> - *Oracle Machine Learning for SQL User's Guide* for information about specifying model settings

DBMS_DATA_MINING — Algorithm Settings: ALGO_EXTENSIBLE_LANG

The settings listed in the following table configure the behavior of the machine learning model with an extensible algorithm. The model is built in the R language.

The `RALG_*_FUNCTION` specifies the R script that is used to build, score, and view an R model and must be registered in the Oracle Machine Learning for R script repository. The R scripts are registered through Oracle Machine Learning for R with special privileges. When `ALGO_EXTENSIBLE_LANG` is set to R in the `MINING_MODEL_SETTING` table, the machine learning model is built in the R language. After the R model is built, the names of the R scripts are recorded in the `MINING_MODEL_SETTING` table in the `SYS` schema. The scripts must exist in the script repository for the R model to function. The amount of R memory used to build, score, and view the R model through these R scripts can be controlled by Oracle Machine Learning for R.

All algorithm-independent `DBMS_DATA_MINING` subprograms can operate on an R model for machine learning functions such as association, attribute importance, classification, clustering, feature extraction, and regression.

The supported `DBMS_DATA_MINING` subprograms include, but are not limited, to the following:

- ADD_COST_MATRIX Procedure
- COMPUTE_CONFUSION_MATRIX Procedure
- COMPUTE_LIFT Procedure
- COMPUTE_ROC Procedure
- CREATE_MODEL Procedure
- DROP_MODEL Procedure

- EXPORT_MODEL Procedure
- GET_MODEL_COST_MATRIX Function
- IMPORT_MODEL Procedure
- REMOVE_COST_MATRIX Procedure
- RENAME_MODEL Procedure

Table 6-13 ALGO_EXTENSIBLE_LANG Settings

Setting Name	Setting Value	Description
`RALG_BUILD_FUNCTION`	`R_BUILD_FUNCTION_SCRIPT_NAME`	Specifies the name of an existing registered R script for the R algorithm machine learning model build function. The R script defines an R function for the first input argument for training data and returns an R model object. For clustering and feature extraction machine learning function model build, the R attributes `dm$nclus` and `dm$nfeat` must be set on the R model to indicate the number of clusters and features respectively. The `RALG_BUILD_FUNCTION` must be set along with `ALGO_EXTENSIBLE_LANG` in the `model_setting_table`.
`RALG_BUILD_PARAMETER`	`SELECT` *value* `param_name`, ...`FROM DUAL`	Specifies a list of numeric and string scalar for optional input parameters of the model build function.
`RALG_SCORE_FUNCTION`	`R_SCORE_FUNCTION_SCRIPT_NAME`	Specifies the name of an existing registered R script to score data. The script returns a `data.frame` containing the corresponding prediction results. The setting is used to score data for machine learning functions such as regression, classification, clustering, and feature extraction. This setting does not apply to the association and the attribute importance functions.
`RALG_WEIGHT_FUNCTION`	`R_WEIGHT_FUNCTION_SCRIPT_NAME`	Specifies the name of an existing registered R script for the R algorithm that computes the weight (contribution) for each attribute in scoring. The script returns a `data.frame` containing the contributing weight for each attribute in a row. This function setting is needed for the `PREDICTION_DETAILS` SQL function.
`RALG_DETAILS_FUNCTION`	`R_DETAILS_FUNCTION_SCRIPT_NAME`	Specifies the name of an existing registered R script for the R algorithm that produces the model information. This setting is required to generate a model view.
`RALG_DETAILS_FORMAT`	`SELECT` *type_value* *column_name*, ... `FROM DUAL`	Specifies the `SELECT` query for the list of numeric and string scalars for the output column type and the column name of the generated model view. This setting is required to generate a model view.

See Also:

Oracle Machine Learning for SQL User's Guide

DBMS_DATA_MINING — Algorithm Settings: CUR Matrix Decomposition

The following settings affects the behavior of the CUR Matrix Decomposition algorithm.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.CURS_ROW_IMP_DISABLE`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'CURS_ROW_IMP_DISABLE'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-14 CUR Matrix Decomposition Settings

Setting Name	Constant Value	String Value Equivalent	Description
`CURS_APPROX_ATTR_NUM`	A positive integer	A positive integer	Defines the approximate number of attributes to be selected. The default value is the number of attributes.
`CURS_ROW_IMPORTANCE`	`CURS_ROW_IMP_ENABLE`	`CURS_ROW_IMP_ENABLE`	Defines the flag indicating whether or not to perform row selection. Enables row selection. The default value is `CURS_ROW_IMP_DISABLE`.
	`CURS_ROW_IMP_DISABLE`	`CURS_ROW_IMP_DISABLE`	Disables row selection.
`CURS_APPROX_ROW_NUM`	A positive integer	A positive integer	Defines the approximate number of rows to be selected. This parameter is only used when users decide to perform row selection (`CURS_ROW_IMP_ENABLE`). The default value is the total number of rows.
`CURS_SVD_RANK`	A positive integer	A positive integer	Defines the rank parameter used in the column/row leverage score calculation. If users do not provide an input value, the system determines the value.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

See Also:

Oracle Machine Learning for SQL Concepts

DBMS_DATA_MINING — Algorithm Settings: Decision Tree

These settings configure the behavior of the Decision Tree algorithm. Note that the Decision Tree settings are also used to configure the behavior of Random Forest as it constructs each individual decision tree.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.TREE_IMPURITY_ENTROPY`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'TREE_IMPURITY_ENTROPY'`.

Note:

The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-15 Decision Tree Settings

Setting Name	Constant Value	String Value Equivalent	Description
`TREE_IMPURITY_METRIC`	`TREE_IMPURITY_ENTROPY`	`TREE_IMPURITY_ENTROPY`	Tree impurity metric for Decision Tree. Tree algorithms seek the best test question for splitting data at each node. The best splitter and split values are those that result in the largest increase in target value homogeneity (purity) for the entities in the node. Purity is by a metric. By default, the algorithm uses `TREE_IMPURITY_GINI`.
	`TREE_IMPURITY_GINI`	`TREE_IMPURITY_GINI`	Decision trees can use either Gini (`TREE_IMPURITY_GINI`) or entropy (`TREE_IMPURITY_ENTROPY`) as the purity metric.

`TREE_TERM_MAX_DEPTH`	For Decision Tree:
	`2<=` *a number* `<=20`
	For Random Forest:
	`2<=` *a number* `<=100`
		For Decision Tree:
		`2<=` *a number* `<=20`
		For Random Forest:
		`2<=` *a number* `<=100`
			Criteria for splits: maximum tree depth (the maximum number of nodes between the root and any leaf node, including the leaf node).
			For Decision Tree, the default is `7`.
			For Random Forest, the default is `16`.

`TREE_TERM_MINPCT_NODE`	`0<=` *a number*`<=10`	`0<=` *a number*`<=10`	The minimum number of training rows in a node expressed as a percentage of the rows in the training data. Default is `0.05`, indicating 0.05%.

Table 6-15 (Cont.) Decision Tree Settings

Setting Name	Constant Value	String Value Equivalent	Description
TREE_TERM_MINPCT_SPLIT	0 < *a number* <=20	0 < *a number* <=20	The minimum number of rows required to consider splitting a node expressed as a percentage of the training rows. Default is 0.1, indicating 0.1%.
TREE_TERM_MINREC_NODE	*a number*>=0	*a number*>=0	The minimum number of rows in a node. Default is 10.
TREE_TERM_MINREC_SPLIT	a number > 1	a number > 1	Criteria for splits: minimum number of records in a parent node expressed as a value. No split is attempted if the number of records is below this value. Default is 20.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

See Also:

Oracle Machine Learning for SQL Concepts for information about Decision Tree

DBMS_DATA_MINING — Algorithm Settings: Expectation Maximization

These algorithm settings configure the behavior of the Expectation Maximization algorithm.

See Also:

Oracle Data Mining Concepts for information about Expectation Maximization

Table 6-16 Expectation Maximization Settings for Data Preparation and Analysis

Setting Name	Constant Value	String Value Equivalent	Description
EMCS_ATTRIBUTE_FILTER	EMCS_ATTR_FILTER_ENABLE	EMCS_ATTR_FILTER_ENABLE	Whether or not to include uncorrelated attributes in the model. When EMCS_ATTRIBUTE_FILTER is enabled, uncorrelated attributes are not included. **Note:** This setting applies only to attributes that are not nested. For Clustering, the default is system-determined. For anomaly detection, the default is EMCS_ATTR_FILTER_DISABLE.
	EMCS_ATTR_FILTER_DISABLE	EMCS_ATTR_FILTER_DISABLE	Includes uncorrelated attributes in the model.
EMCS_MAX_NUM_ATTR_2D	An integer greater than or equal to 1, represented as a character string.	An integer greater than or equal to 1, represented as a character string.	Maximum number of correlated attributes to include in the model. Note: This setting applies only to attributes that are not nested (2D). Default is 50. Expression: TO_CHAR(40)
EMCS_NUM_DISTRIBUTION	EMCS_NUM_DISTR_BERNOULLI	EMCS_NUM_DISTR_BERNOULLI	The distribution for modeling numeric attributes. Applies to the input table or view as a whole and does not allow per-attribute specifications. The options include Bernoulli, Gaussian, or system-determined distribution. When Bernoulli or Gaussian distribution is chosen, all numeric attributes are modeled using the same type of distribution. Default is EMCS_NUM_DISTR_SYSTEM.
	EMCS_NUM_DISTR_GAUSSIAN	EMCS_NUM_DISTR_GAUSSIAN	Models all numeric attributes using Gaussian distribution.
	EMCS_NUM_DISTR_SYSTEM	EMCS_NUM_DISTR_SYSTEM	When the distribution is system-determined, individual attributes may use different distributions (either Bernoulli or Gaussian), depending on the data.
EMCS_NUM_EQUIWIDTH_BINS	An integer between 1 to 255, inclusive, represented as a character string.	An integer between 1 to 255, inclusive, represented as a character string.	Number of equi-width bins that will be used for gathering cluster statistics for numeric columns. Default is 11. Expression: TO_CHAR(20)

Table 6-16 (Cont.) Expectation Maximization Settings for Data Preparation and Analysis

Setting Name	Constant Value	String Value Equivalent	Description
EMCS_NUM_PROJECTIONS	An integer greater than or equal to 1, represented as a character string.	An integer greater than or equal to 1, represented as a character string.	Specifies the number of projections that will be used for each nested column. If a column has fewer distinct attributes than the specified number of projections, the data will not be projected. The setting applies to all nested columns. Default is 50. Expression: TO_CHAR(40)
EMCS_NUM_QUANTILE_BINS	An integer between 1 to 255, inclusive, represented as a character string.	An integer between 1 to 255, inclusive, represented as a character string.	Specifies the number of quantile bins that will be used for modeling numeric columns with multivalued Bernoulli distributions. Default is system-determined. Expression: TO_CHAR(20)
EMCS_NUM_TOPN_BINS	An integer between 1 to 255, inclusive, represented as a character string.	An integer between 1 to 255, inclusive, represented as a character string.	Specifies the number of top-N bins that will be used for modeling categorical columns with multivalued Bernoulli distributions. Default is system-determined. Expression: TO_CHAR(10)

Table 6-17 Expectation Maximization Settings for Learning

Setting Name	Constant Value	String Value Equivalent	Description
EMCS_CONVERGENCE_CRITERION	EMCS_CONV_CRIT_HELDASIDE	EMCS_CONV_CRIT_HELDASIDE	The convergence criterion for EM. The convergence criterion may be based on a held-aside data set, or it may be Bayesian Information Criterion. EMCS_CONV_CRIT_HELDASIDE: Uses a held-aside data set for convergence criterion. Default is system determined.
	EMCS_CONV_CRIT_BIC	EMCS_CONV_CRIT_BIC	Uses the Bayesian Information Criterion (BIC) for convergence.
EMCS_LOGLIKE_IMPROVEMENT	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	When the convergence criterion is based on a held-aside data set (EMCS_CONVERGENCE_CRITERION = EMCS_CONV_CRIT_HELDASIDE), this setting specifies the percentage improvement in the value of the log likelihood function that is required for adding a new component to the model. Default value is 0.001. Expression: TO_CHAR(0.003)

Table 6-17 (Cont.) Expectation Maximization Settings for Learning

Setting Name	Constant Value	String Value Equivalent	Description
`EMCS_NUM_COMPONENTS`	An integer greater than or equal to 1, represented as a character string	An integer greater than or equal to 1, represented as a character string	Maximum number of components in the model. If model search is enabled, the algorithm automatically determines the number of components based on improvements in the likelihood function or based on regularization, up to the specified maximum. For EM Clustering, the number of components must be greater than or equal to the number of clusters. Default is 20 for both EM Clustering and EM Anomaly. Expression: `TO_CHAR(20)`
`EMCS_NUM_ITERATIONS`	An integer greater than or equal to 1, represented as a character string	An integer greater than or equal to 1, represented as a character string	Specifies the maximum number of iterations in the EM algorithm. Default is `100`. Expression: `TO_CHAR(50)`
`EMCS_MODEL_SEARCH`	`EMCS_MODEL_SEARCH_ENABLE`	`EMCS_MODEL_SEARCH_ENABLE`	This setting enables model search in EM where different model sizes are explored and a best size is selected. The default is `EMCS_MODEL_SEARCH_DISABLE`.
	`EMCS_MODEL_SEARCH_DISABLE` (default).	`EMCS_MODEL_SEARCH_DISABLE` (default).	The model search in EM is disabled.
`EMCS_REMOVE_COMPONENTS`	`EMCS_REMOVE_COMPS_ENABLE` (default)	`EMCS_REMOVE_COMPS_ENABLE` (default)	This setting allows the EM algorithm to remove a small component from the solution. The default is `EMCS_REMOVE_COMPS_ENABLE`.
	`EMCS_REMOVE_COMPS_DISABLE`	`EMCS_REMOVE_COMPS_DISABLE`	Prevents the EM algorithm from removing small components.
`EMCS_RANDOM_SEED`	Non-negative integer	Non-negative integer	This setting controls the seed of the random generator used in EM. The default is `0`.

Table 6-18 Expectation Maximization Settings for Component Clustering

Setting Name	Constant Value	String Value Equivalent	Description
`EMCS_CLUSTER_COMPONENTS`	`EMCS_CLUSTER_COMP_ENABLE`	`EMCS_CLUSTER_COMP_ENABLE`	Enables or disables the grouping of EM components into high-level clusters. When disabled, the components themselves are treated as clusters. The setting can be used only for EM Clustering. When component clustering is enabled, model scoring through the SQL `CLUSTER` function will produce assignments to the higher level clusters. Default is `EMCS_CLUSTER_COMP_ENABLE`.

Table 6-18 (Cont.) Expectation Maximization Settings for Component Clustering

Setting Name	Constant Value	String Value Equivalent	Description
	EMCS_CLUSTER_COMP_DISABLE	EMCS_CLUSTER_COMP_DISABLE	When clustering is disabled, the CLUSTER function will produce assignments to the original components.
EMCS_CLUSTER_THRESH	Specify an integer greater than or equal to 1, represented as a character string	Specify an integer greater than or equal to 1, represented as a character string	Dissimilarity threshold that controls the clustering of EM components. When the dissimilarity measure is less than the threshold, the components are combined into a single cluster. The setting can be used only for EM Clustering. A lower threshold may produce more clusters that are more compact. A higher threshold may produce fewer clusters that are more spread out. Default is 2. Expression: TO_CHAR(3)
EMCS_LINKAGE_FUNCTION	EMCS_LINKAGE_SINGLE	EMCS_LINKAGE_SINGLE	Allows the specification of a linkage function for the agglomerative clustering step. EMCS_LINKAGE_SINGLE uses the nearest distance within the branch. The clusters tend to be larger and have arbitrary shapes. Default is EMCS_LINKAGE_SINGLE.
	EMCS_LINKAGE_AVERAGE	EMCS_LINKAGE_AVERAGE	EMCS_LINKAGE_AVERAGE uses the average distance within the branch. There is less chaining effect and the clusters are more compact.
	EMCS_LINKAGE_COMPLETE	EMCS_LINKAGE_COMPLETE	EMCS_LINKAGE_COMPLETE uses the maximum distance within the branch. The clusters are smaller and require strong component overlap.

Table 6-19 Expectation Maximization Settings for Cluster Statistics

Setting Name	Constant Value	Description
EMCS_CLUSTER_STATISTICS	EMCS_CLUS_STATS_ENABLE EMCS_CLUS_STATS_DISABLE	Enables or disables the gathering of descriptive statistics for clusters (centroids, histograms, and rules). When statistics are disabled, model size is reduced, and GET_MODEL_DETAILS_EM only returns taxonomy (hierarchy) and cluster counts. The setting can be used only for EM Clustering. Default is EMCS_CLUS_STATS_ENABLE.
EMCS_MIN_PCT_ATTR_SUPPORT	A floating point number between 0 and 1 expressed as a character string	Minimum support required for including an attribute in the cluster rule. The support is the percentage of the data rows assigned to a cluster that must have non-null values for the attribute. The setting can be used only for EM Clustering. Default is 0.1. Expression: TO_CHAR(0.9)

Table 6-20 Expectation Maximization Settings for Anomaly Detection

Setting Name	Constant Value	String Value Equivalent	Description
`EMCS_OUTLIER_RATE`	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	The desired rate of outliers in the training data. The setting can be used only for EM Anomaly. Default is 0.05. Expression: `TO_CHAR(0.07)`

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

DBMS_DATA_MINING — Algorithm Settings: Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) is a useful technique for extracting meaningful and interpretable features.

The settings listed in the following table configure the ESA values.

Table 6-21 Explicit Semantic Analysis Settings

Setting Name	Setting Value	String Value Equivalent	Description
`ESAS_EMBEDDINGS`	`ESAS_EMBEDDINGS_ENABLE`	`ESAS_EMBEDDINGS_ENABLE`	This setting applies to feature extraction models. The default value is `ESAS_EMBEDDINGS_DISABLE`. When you set `ESAS_EMBEDDINGS_ENABLE`: • ESA generates embeddings during scoring • The FEATURE_ID of the generated embeddings is of the data type NUMBER • The `CASE_ID_COLUMN_NAME` argument of the `DBMS_DATA_MINING.CREATE_MODEL` and `DBMS_DATA_MINING.CREATE_MODEL2` function is optional.
	`ESAS_EMBEDDINGS_DISABLE`	`ESAS_EMBEDDINGS_DISABLE`	Disables the use of embeddings for ESA. This setting is useful when embeddings are not required or desired for the analysis

Table 6-21 (Cont.) Explicit Semantic Analysis Settings

Setting Name	Setting Value	String Value Equivalent	Description
ESAS_EMBEDDING_SIZE	A positive integer less than or equal to 4096	A positive integer less than or equal to 4096	This setting applies to feature extraction models. This setting specifies the size of the vectors representing embeddings. You can set this parameter only if you have enabled ESAS_EMBEDDINGS. The default size is 1024. If this value is less than the number of distinct features in the training set, then the actual number of explicit features is used as the size of embedding vectors instead.
ESAS_MIN_ITEMS	Text input 100 Non-text input is 0	Text input 100 Non-text input is 0	This setting determines the minimum number of non-zero entries that need to be present in an input row. The default is 100 for text input and 0 for non-text input.
ESAS_TOPN_FEATURES	A positive integer	A positive integer	This setting controls the maximum number of features per attribute. The default is 1000.
ESAS_VALUE_THRESHOLD	Non-negative number	Non-negative number	This setting thresholds a small value for attribute weights in the transformed build data. The default is 1e-8.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

See Also:

Oracle Machine Learning for SQL Concepts for information about ESA.

DBMS_DATA_MINING — Algorithm Settings: Exponential Smoothing

These settings configure the behavior of the Exponential Smoothing (ESM) algorithm.

The settings listed in the following table specify the setting names and possible values for Exponential Smoothing. You can specify the Setting Value using the prefix DBMS_DATA_MINING. For example, DBMS_DATA_MINING.EXSM_SIMPLE. Alternatively, you can specify the Setting Value without the DBMS_DATA_MINING prefix, in single quotes. For example, 'EXSM_SIMPLE'.

For Global settings, see DBMS_DATA_MINING — Global Settings.

Table 6-22 Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
EXSM_MODEL	EXSM_SIMPLE	EXSM_SIMPLE	This setting specifies the model. EXSM_SIMPLE: Forecasts data as a weighted moving average, with the influence of past observations declining exponentially with the length of time since the observation occurred. Errors in estimation are assumed to be normally distributed, with constant mean and variance. It is appropriate for data with no clear trend or seasonal pattern. The default value is EXSM_SIMPLE.
	EXSM_SIMPLE_MULT_ERR	EXSM_SIMPLE_MULT_ERR	Forecasts data as a weighted moving average, with the influence of past observations declining exponentially with the length of time since the observation occurred. Errors in estimation are assumed to be proportional to the level of the prior estimate.
	EXSM_HOLT	EXSM_HOLT	Applies Holt's linear exponential smoothing method, designed to forecast data with an underlying linear trend.
	EXSM_HOLT_DAMPED	EXSM_HOLT_DAMPED	Applies Holt's linear exponential smoothing with a damping factor to progressively reduce the strength of the trend over time.
	EXSM_MULT_TREND	EXSM_MULT_TREND	Applies an exponential smoothing framework with a multiplicative trend component, effectively capturing data where trends are not linear but grow or decay over time.
	EXSM_MULT_TREND_DAMPED	EXSM_MULT_TREND_DAMPED	Applies an exponential smoothing algorithm with a multiplicative trend that diminishes over time, providing a conservative approach to trend estimation.
	EXSM_SEASON_ADD	EXSM_SEASON_ADD	Applies an exponential smoothing with an additive seasonal component, isolating and accounting for seasonal variations without incorporating a trend.
	EXSM_SEASON_MUL	EXSM_SEASON_MUL	Executes exponential smoothing with a multiplicative seasonal component, capturing seasonal effects that increase or decrease in proportion to the level of the series.
	EXSM_WINTERS	EXSM_WINTERS	Applies the Holt-Winters method with additive trends and multiplicative seasonality, offering a robust model for data with both linear trend and proportional seasonal variation.
	EXSM_WINTERS_DAMPED	EXSM_WINTERS_DAMPED	Applies the Holt-Winters method with a damped trend and multiplicative seasonality, moderating the linear trend over time while still capturing proportional seasonal changes.
	EXSM_ADDWINTERS	EXSM_ADDWINTERS	Applies the Holt-Winters additive model to simultaneously smooth data with linear trends and additive seasonal effects.

Table 6-22 (Cont.) Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
	EXSM_ADDWINTERS_DAMPED	EXSM_ADDWINTERS_DAMPED	Applies the Holt-Winters additive approach with a damping mechanism, reducing the impact of the trend and seasonal components over time.
	EXSM_WINTERS_MUL_TREND	EXSM_WINTERS_MUL_TREND	Applies the Holt-Winters model with both trend and seasonality components being multiplicative, suited for series where the seasonal variations and trends are both increasing or decreasing proportional to level.
	EXSM_WINTERS_MUL_TREND_DMP	EXSM_WINTERS_MUL_TREND_DMP	Applies the Holt-Winters model with a damped multiplicative trend, effectively moderating the exponential increase or decrease of both trend and seasonal components over time.
EXSM_SEASONALITY	positive integer > 1	positive integer > 1	This setting specifies a positive integer value as the length of seasonal cycle. The value it takes must be larger than 1. For example, setting value 4 means that every group of four observations forms a seasonal cycle. This setting is only applicable and must be provided for models with seasonality, otherwise the model throws an error. When EXSM_INTERVAL is not set, this setting applies to the original input time series. When EXSM_INTERVAL is set, this setting applies to the accumulated time series.
EXSM_INTERVAL	EXSM_INTERVAL_YEAR	EXSM_INTERVAL_YEAR	This setting only applies and must be provided when the time column (case_id column) has datetime type. It specifies the spacing interval of the accumulated equally spaced time series. The model throws an error if the time column of input table is of datetime type and setting EXSM_INTERVAL is not provided. The model throws an error if the time column of input table is of oracle number type and setting EXSM_INTERVAL is provided. EXSM_INTERVAL_YEAR: This option sets the spacing interval of the accumulated time series to one year. When selected, the data is aggregated or summarized on a yearly basis.
	EXSM_INTERVAL_QTR	EXSM_INTERVAL_QTR	This option sets the spacing interval to a quarter, aggregating the data for every three months.
	EXSM_INTERVAL_MONTH	EXSM_INTERVAL_MONTH	This option adjusts the spacing interval to one month. The accumulated time series represent aggregated or summarized data for each month.
	EXSM_INTERVAL_WEEK	EXSM_INTERVAL_WEEK	With this option data is aggregated or summarized on a weekly basis, setting the spacing interval to one week.
	EXSM_INTERVAL_DAY	EXSM_INTERVAL_DAY	This option adjusts the spacing interval to one day. It's suitable for scenarios where daily aggregated insights are required.

Table 6-22 (Cont.) Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
	EXSM_INTERVAL_HOUR	EXSM_INTERVAL_HOUR	For more granular insights, this option sets the spacing interval to one hour. It's especially useful when analyzing data that changes significantly within a day.
	EXSM_INTERVAL_MINUTE	EXSM_INTERVAL_MINUTE	With this option the spacing is set to one minute. This provides a very detailed view of data, suitable for applications like high-frequency trading or real-time monitoring systems.
	EXSM_INTERVAL_SECOND	EXSM_INTERVAL_SECOND	For most granular details, this options sets the spacing interval to one second. It's tailored for scenarios requiring real-time or near-real-time analysis.
EXSM_INITVL_OPTIMIZE	EXSM_INITVL_OPTIMIZE_ENABLE	EXSM_INITVL_OPTIMIZE_ENABLE	The setting EXSM_INITVL_OPTIMIZE determines whether initial values are optimized during model build. The default value is EXSM_INITVL_OPTIMIZE_ENABLE.
	EXSM_INITVL_OPTIMIZE_DISABLE	EXSM_INITVL_OPTIMIZE_DISABLE	Note: EXSM_INITVL_OPTIMIZE can only be set to EXSM_INITVL_OPTIMIZE_DISABLE if the user has set EXSM_MODEL to EXSM_HW or EXSM_HW_ADDSEA. If EXSM_MODEL is set to another model type or is not specified, you get an error 40213 (conflicting settings) and the model is not built.
EXSM_ACCUMULATE	EXSM_ACCU_TOTAL	EXSM_ACCU_TOTAL	This setting only applies and must be provided when the time column has datetime type. It specifies how to generate the value of the accumulated time series from the input time series. EXSM_ACCU_TOTAL: This option calculates the total sum of the time series values within a specified interval. When selected, it will aggregate the data by summing up all the individual values in the datetime range. The default value is EXSM_ACCU_TOTAL.
	EXSM_ACCU_STD	EXSM_ACCU_STD	This option computes the standard deviation of the time series values within a specified interval. It helps you understand the amount of variation or dispersion in your data.
	EXSM_ACCU_MAX	EXSM_ACCU_MAX	By selecting this option, the maximum value of the time series within a specified interval will be determined. It helps in identifying the peak value in the given range.
	EXSM_ACCU_MIN	EXSM_ACCU_MIN	This option focuses on determining the minimum value of the time series within a specified interval. It is useful for identifying the lowest value in the time series for the given datetime range.

Table 6-22 (Cont.) Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
	EXSM_ACCU_AVG	EXSM_ACCU_AVG	This specifies the average value of your time series within a specified interval. It calculates the mean value of all data points in the specified range.
	EXSM_ACCU_MEDIAN	EXSM_ACCU_MEDIAN	This option provides the median of the time series values within the given interval. The median gives a central value, which can be especially useful if your data contains outliers.
	EXSM_ACCU_COUNT	EXSM_ACCU_COUNT	This option counts the number of time series values within the specified interval. It is helpful if you want to know how many data points are present in a certain datetime range.
EXSM_SETMISSING	Specify an option: EXSM_MISS_MIN	EXSM_MISS_MIN	This setting specifies how to handle missing values, which may come from input data and/or the accumulation process of time series. You can specify either a number or an option. If a number is specified, all the missing values are set to that number. EXSM_MISS_MIN: Replaces missing value with minimum of the accumulated time series. If EXSM_SETMISSING setting is not provided, EXSM_MISS_AUTO is the default value. In such a case, the model treats the input time series as irregular time series, viewing missing values as gaps.
	EXSM_MISS_MAX	EXSM_MISS_MAX	Replaces missing value with maximum of the accumulated time series.
	EXSM_MISS_AVG	EXSM_MISS_AVG	Replaces missing value with average of the accumulated time series.
	EXSM_MISS_MEDIAN	EXSM_MISS_MEDIAN	Replaces missing value with median of the accumulated time series.
	EXSM_MISS_LAST	EXSM_MISS_LAST	Replaces missing value with last non-missing value of the accumulated time series.
	EXSM_MISS_FIRST	EXSM_MISS_FIRST	Replaces missing value with first non-missing value of the accumulated time series.
	EXSM_MISS_PREV	EXSM_MISS_PREV	Replaces missing value with the previous non-missing value of the accumulated time series.
	EXSM_MISS_NEXT	EXSM_MISS_NEXT	Replaces missing value with the next non-missing value of the accumulated time series.
	EXSM_MISS_AUTO	EXSM_MISS_AUTO	EXSM model treats the input data as an irregular (non-uniformly spaced) time series.
EXSM_PREDICTION_STEP	A number between 1-30.	A number between 1-30.	This setting specifies how many steps ahead the predictions are to be made. If it is not set, the default value is 1: the model gives one-step-ahead prediction. A value greater than 30 results in an error.

Table 6-22 (Cont.) Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
EXSM_CONFIDENCE_LEVEL	A number between 0 and 1, exclusive.	A number between 0 and 1, exclusive.	This setting specifies the desired confidence level for prediction. The lower and upper bounds of the specified confidence interval is reported. If this setting is not specified, the default confidence level is 95%.
EXSM_OPT_CRITERION	EXSM_OPT_CRIT_LIK	EXSM_OPT_CRIT_LIK	This setting specifies the desired optimization criterion. The optimization criterion is useful as a diagnostic for comparing models' fit to the same data. EXSM_OPT_CRIT_LIK: This represents the negative double of the logarithm of the likelihood associated with a given model. The default value is EXSM_OPT_CRIT_LIK.
	EXSM_OPT_CRIT_MSE	EXSM_OPT_CRIT_MSE	This provides the mean squared error pertaining to the model.
	EXSM_OPT_CRIT_AMSE	EXSM_OPT_CRIT_AMSE	This denotes the average of the mean squared error over a time window as specified by the user.
	EXSM_OPT_CRIT_SIG	EXSM_OPT_CRIT_SIG	This metric captures the standard deviation of the residuals of the model.
	EXSM_OPT_CRIT_MAE	EXSM_OPT_CRIT_MAE	This metric conveys the average absolute error associated with the model. It measures the size of the error.
EXSM_NMSE	A positive integer	A positive integer	This setting specifies the length of the window used in computing the error metric average mean square error (AMSE).

Table 6-22 (Cont.) Exponential Smoothing Settings

Setting Name	Setting Value	String Value Equivalent	Description
EXSM_SERIES_LIST	Comma delimited list of time series columns	Comma delimited list of time series columns	This setting allows you to forecast up to twenty predictor series in addition to the target series. The column names in EXSM_SERIES_LIST are enclosed in single quotes. **Note:** The list is enclosed in single quotes, not the individual column names. For example: `INSERT INTO <settings_table_name VALUES(dbms_data_mining.exsm_series_list, ‘<column1>,<column2>,<column3>,<column4>’);` The prefix DM$ must be added to the build and scoring data sets. The column names must be less than 125 characters long. See Model Detail Views for Exponential Smoothing.
EXSM_BACKCAST_OUTPUT	EXSM_BACKCAST_OUTPUT_ENABLE	EXSM_BACKCAST_OUTPUT_ENABLE	This setting enables the user to optionally suppress the output of backcast values. Backcasts are the model estimates for historical data. See Backcasts in Time Series for information on backcasts. The default value is EXSM_BACKCAST_OUTPUT_ENABLE.
	EXSM_BACKCAST_OUTPUT_DISABLE	EXSM_BACKCAST_OUTPUT_DISABLE	This setting disables the output of backcast values. Suppressing the output of backcast values can provide a potentially large reduction in the memory and storage requirements for a partitioned ESM model with a huge number of partitions.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

See Also:

Oracle Machine Learning for SQL Concepts for information about ESM.

https://github.com/oracle-samples/oracle-db-examples/tree/main/machine-learning/sql browse to the release folder and click the `oml4sql-time-series-exponential-smoothing.sql` example.

DBMS_DATA_MINING — Algorithm Settings: Generalized Linear Model

The settings listed in the following table configure the behavior of the Generalized Linear Model algorithm.

The settings listed in the following table specify the setting names and possible values for Generalized Linear Model. The Constant Value column specifies constants using the prefix `DBMS_DATA_MINING`. Alternatively, you can specify the corresponding string value from the String Value Equivalent column.

For Global settings, see DBMS_DATA_MINING — Global Settings.

For generic machine learning function settings, see DBMS_DATA_MINING — Machine Learning Functions.

Table 6-23 DBMS_DATA_MINING GLM Settings

Setting Name	Constant Value	String Value Equivalent	Description
`GLMS_CONF_LEVEL`	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	The confidence level for coefficient confidence intervals. The default confidence level is `0.95`. Expression: `TO_CHAR(0.98)`
`GLMS_FTR_GEN_METHOD`	`GLMS_FTR_GEN_QUADRATIC`	`GLMS_FTR_GEN_QUADRATIC`	Whether feature generation is quadratic or cubic. When feature generation is enabled, the algorithm automatically chooses the most appropriate feature generation method based on the data. `GLMS_FTR_GEN_QUADRATIC`: Generates features using a quadratic method.
	`GLMS_FTR_GEN_CUBIC`	`GLMS_FTR_GEN_CUBIC`	Generates features using a cubic method.
`GLMS_FTR_GENERATION`	`GLMS_FTR_GENERATION_ENABLE`	`GLMS_FTR_GENERATION_ENABLE`	Whether or not feature generation is enabled for GLM. By default, feature generation is not enabled. **Note:** Feature generation can only be enabled when feature selection is also enabled.
	`GLMS_FTR_GENERATION_DISABLE`	`GLMS_FTR_GENERATION_DISABLE`	Disables feature generation for GLM (default).

Table 6-23 (Cont.) DBMS_DATA_MINING GLM Settings

Setting Name	Constant Value	String Value Equivalent	Description
GLMS_FTR_SEL_CRIT	GLMS_FTR_SEL_AIC	GLMS_FTR_SEL_AIC	Feature selection penalty criterion for adding a feature to the model. When feature selection is enabled, the algorithm automatically chooses the penalty criterion based on the data. GLMS_FTR_SEL_AIC: Uses Akaike Information Criterion for feature selection.
	GLMS_FTR_SEL_SBIC	GLMS_FTR_SEL_SBIC	Uses Schwarz Bayesian Information Criterion for feature selection.
	GLMS_FTR_SEL_RIC	GLMS_FTR_SEL_RIC	Uses Risk Inflation Criterion for feature selection.
	GLMS_FTR_SEL_ALPHA_INV	GLMS_FTR_SEL_ALPHA_INV	Uses Alpha Inverse Criterion for feature selection.
GLMS_FTR_SELECTION	GLMS_FTR_SELECTION_ENABLE	GLMS_FTR_SELECTION_ENABLE	Whether or not feature selection is enabled for GLM. By default, feature selection is not enabled.
	GLMS_FTR_SELECTION_DISABLE	GLMS_FTR_SELECTION_DISABLE	Disables feature selection.
GLMS_MAX_FEATURES	An integer greater than 0 and less than or equal to 2000, represented as a character string	An integer greater than 0 and less than or equal to 2000, represented as a character string	When feature selection is enabled, this setting specifies the maximum number of features that can be selected for the final model. By default, the algorithm limits the number of features to ensure sufficient memory. Expression: TO_CHAR(200)
GLMS_PRUNE_MODEL	GLMS_PRUNE_MODEL_ENABLE	GLMS_PRUNE_MODEL_ENABLE	Prune enable or disable for features in the final model. Pruning is based on T-Test statistics for linear regression, or Wald Test statistics for logistic regression. Features are pruned in a loop until all features are statistically significant with respect to the full data. When feature selection is enabled, the algorithm automatically prunes as per the description. When feature selection is disabled, you cannot specify pruning.
	GLMS_PRUNE_MODEL_DISABLE	GLMS_PRUNE_MODEL_DISABLE	Disables pruning of features.

Table 6-23 (Cont.) DBMS_DATA_MINING GLM Settings

Setting Name	Constant Value	String Value Equivalent	Description
GLMS_REFERENCE_CLASS_NAME	*target_value*	*target_value*	The target value used as the reference class in a binary logistic regression model. Probabilities are produced for the other class. By default, the algorithm chooses the value with the highest prevalence (the most cases) for the reference class.
GLMS_RIDGE_REGRESSION	GLMS_RIDGE_REG_ENABLE	GLMS_RIDGE_REG_ENABLE	Enable or disable ridge regression. Ridge applies to both regression and classification machine learning functions. When ridge is enabled, prediction bounds are not produced by the PREDICTION_BOUNDS SQL function. **Note**: Ridge may only be enabled when feature selection is not specified, or has been explicitly disabled. If ridge regression is enabled, you cannot enable feature selection and an exception is raised.
	GLMS_RIDGE_REG_DISABLE	GLMS_RIDGE_REG_DISABLE	Disables ridge regression.
GLMS_RIDGE_VALUE	An integer greater than 0 represented as a character string	An integer greater than 0 represented as a character string	The value of the ridge parameter. This setting is only used when the algorithm is configured to use ridge regression. If ridge regression is enabled internally by the algorithm, then the ridge parameter is determined by the algorithm. Expression: TO_CHAR(5)
GLMS_ROW_DIAGNOSTICS	GLMS_ROW_DIAG_ENABLE	GLMS_ROW_DIAG_ENABLE	GLMS_ROW_DIAG_ENABLE: Enables row diagnostics.
	GLMS_ROW_DIAG_DISABLE (default).	GLMS_ROW_DIAG_DISABLE (default).	Disables row diagnostics.
GLMS_CONV_TOLERANCE	The range is (0, 1) non-inclusive.	The range is (0, 1) non-inclusive.	Convergence Tolerance setting of the GLM algorithm The default value is system-determined.
GLMS_NUM_ITERATIONS	A positive integer	A positive integer	Maximum number of iterations for the GLM algorithm. The default value is system-determined.

Table 6-23 (Cont.) DBMS_DATA_MINING GLM Settings

Setting Name	Constant Value	String Value Equivalent	Description
GLMS_BATCH_ROWS	0 or a positive integer	0 or a positive integer	Number of rows in a batch used by the SGD solver. The value of this parameter sets the size of the batch for the SGD solver. An input of 0 triggers a data driven batch size estimate. The default is 2000
GLMS_SOLVER	GLMS_SOLVER_SGD (StochasticGradient Descent)	GLMS_SOLVER_SGD (StochasticGradient Descent)	This setting allows the user to choose the GLM solver. The solver cannot be selected if GLMS_FTR_SELECTION setting is enabled. GLMS_SOLVER_SGD: Optimizes generalized linear models by iteratively updating parameters using a subset of the data to minimize errors. The default value is system determined. **See Also:** GLM Solvers
	GLMS_SOLVER_CHOL (Cholesky)	GLMS_SOLVER_CHOL (Cholesky)	Solves generalized linear models using the Cholesky decomposition method, which provides a stable and efficient solution by transforming the right-hand of the equation into a lower triangular matrix and its conjugate transpose.
	GLMS_SOLVER_QR	GLMS_SOLVER_QR	Utilizes the QR decomposition technique to solve generalized linear models, ensuring numerical stability and accuracy by decomposing the problem into an orthonormal matrix Q and upper triangular matrix R.
	GLMS_SOLVER_LBFGS_ADMM	GLMS_SOLVER_LBFGS_ADMM	Combines L-BFGS, an approximation of the Broyden-Fletcher-Goldfarb-Shanno optimization algorithm, with ADMM for solving large-scale generalized linear model problems efficiently.

Table 6-23 (Cont.) DBMS_DATA_MINING GLM Settings

Setting Name	Constant Value	String Value Equivalent	Description
GLMS_SPARSE_SOLVER	GLMS_SPARSE_SOLVER_ENABLE	GLMS_SPARSE_SOLVER_ENABLE	This setting allows the user to use sparse solver if it is available. The default value is GLMS_SPARSE_SOLVER_DISABLE.
	GLMS_SPARSE_SOLVER_DISABLE (default).	GLMS_SPARSE_SOLVER_DISABLE (default).	Disables sparse solver.
GLMS_LINK_FUNCTION	GLMS_IDENTITY_LINK	GLMS_IDENTITY_LINK	This setting allows the user to specify the link function for building a GLM model. The link functions are specific to the mining function. For classification, the following are applicable: • GLMS_LOGIT_LINK (default) • GLMS_PROBIT_LINK • GLMS_CLOGLOG_LINK • GLMS_CAUCHIT_LINK For regression, the following is applicable: GLMS_IDENTITY_LINK (default) GLMS_IDENTITY_LINK: Employs the identity link function for GLM regression, directly relating the response variable to the linear predictor without transformation. This is the default setting for Regression.
	GLMS_LOGIT_LINK	GLMS_LOGIT_LINK	Implements the logit link function for GLM classification, mapping probabilities onto the log-odds scale, commonly used for logistic regression.
	GLMS_PROBIT_LINK	GLMS_PROBIT_LINK	Uses the probit link function for GLM classification, assuming a normal cumulative distribution to model binary outcomes.
	GLMS_CLOGLOG_LINK	GLMS_CLOGLOG_LINK	Applies the complementary log-log (cloglog) link function for GLM classification, designed for modeling asymmetric probability distributions.
	GLMS_CAUCHIT_LINK	GLMS_CAUCHIT_LINK	Utilizes the Cauchit link function for GLM classification, leveraging the Cauchy cumulative distribution for handling heavy-tailed data.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.

- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.
- DBMS_DATA_MINING — Algorithm Settings: Neural Network
  The settings listed in the following table configure the behavior of the Neural Network algorithm.
- DBMS_DATA_MINING — Solver Settings: LBFGS
  The settings listed in the following table configure the behavior of L-BFGS. Neural Network and Generalized Linear Model (GLM) use these settings.
- DBMS_DATA_MINING — Solver Settings: ADMM
  The settings listed in the following table configure the behavior of Alternating Direction Method of Multipliers (ADMM). The Generalized Linear Model (GLM) algorithm uses these settings.
- *Oracle Machine Learning for SQL Concepts*

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about GLM.

DBMS_DATA_MINING — Algorithm Settings: *k*-Means

The settings listed in the following table configure the behavior of the *k*-Means algorithm.

You can specify the Constant Value using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.KMNS_CONV_TOLERANCE`. Alternatively, you can specify the String Value Equivalent without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'KMNS_CONV_TOLERANCE'`

Table 6-24 k-Means Settings

Setting Name	Constant Value	String Value Equivalent	Description
`KMNS_CONV_TOLERANCE`	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	Minimum Convergence Tolerance for *k*-Means. The algorithm iterates until the minimum Convergence Tolerance is satisfied or until the maximum number of iterations, specified in `KMNS_ITERATIONS`, is reached. Decreasing the Convergence Tolerance produces a more accurate solution but may result in longer run times. The default Convergence Tolerance is `0.001`. Expression: `TO_CHAR(0.001)`

Table 6-24 (Cont.) k-Means Settings

Setting Name	Constant Value	String Value Equivalent	Description
KMNS_DISTANCE	KMNS_COSINE	KMNS_COSINE	Distance function for *k*-Means. Specifies that the K-Means clustering algorithm will use the cosine similarity metric to measure the distance between points. Cosine similarity evaluates how similar two vectors are, based on the cosine of the angle between them. This is particularly useful for high-dimensional data such as text and document clustering. The default distance function is KMNS_EUCLIDEAN.
	KMNS_EUCLIDEAN	KMNS_EUCLIDEAN	Specifies that the K-Means clustering algorithm will use the Euclidean distance metric to measure the distance between points. Euclidean distance is the straight-line distance between two points in space and is widely used for clustering numerical data.
KMNS_ITERATIONS	A positive integer represented as a character string	A positive integer represented as a character string	Maximum number of iterations for *k*-Means. The algorithm iterates until either the maximum number of iterations is reached or the minimum Convergence Tolerance, specified in KMNS_CONV_TOLERANCE, is satisfied. The default number of iterations is 20. Expression: TO_CHAR(10)
KMNS_MIN_PCT_ATTR_SUPPORT	A floating point number between 0 and 1, inclusive, expressed as a character string	A floating point number between 0 and 1, inclusive, expressed as a character string	Minimum percentage of attribute values that must be non-null in order for the attribute to be included in the rule description for the cluster. If the data is sparse or includes many missing values, a minimum support that is too high can cause very short rules or even empty rules. The default minimum support is 0.1. Expression: TO_CHAR(0.5)
KMNS_NUM_BINS	A positive integer greater than 0 expressed as a character string	A positive integer greater than 0 expressed as a character string	Number of bins in the attribute histogram produced by *k*-means. The bin boundaries for each attribute are computed globally on the entire training data set. The binning method is equi-width. All attributes have the same number of bins with the exception of attributes with a single value that have only one bin. The default number of histogram bins is 11. Expression: TO_CHAR(15)

Table 6-24 (Cont.) k-Means Settings

Setting Name	Constant Value	String Value Equivalent	Description
KMNS_SPLIT_CRITERION	KMNS_SIZE	KMNS_SIZE	Split criterion for *k*-means. The split criterion controls the initialization of new *k*-Means clusters. The algorithm builds a binary tree and adds one new cluster at a time. When the split criterion is based on size, the new cluster is placed in the area where the largest current cluster is located.
	KMNS_VARIANCE	KMNS_VARIANCE	When the split criterion is based on the variance, the new cluster is placed in the area of the most spread-out cluster. The default split criterion is the KMNS_VARIANCE.
KMNS_RANDOM_SEED	Non-negative integer	Non-negative integer	This setting controls the seed of the random generator used during the *k*-Means initialization. It must be a non-negative integer value. The default is 0.
KMNS_DETAILS	KMNS_DETAILS_NONE	KMNS_DETAILS_NONE	This setting determines the level of cluster detail that are computed during the build. KMNS_DETAILS_NONE: No cluster details are computed. Only the scoring information is persisted.
	KMNS_DETAILS_HIERARCHY	KMNS_DETAILS_HIERARCHY	Cluster hierarchy and cluster record counts are computed.
	KMNS_DETAILS_ALL	KMNS_DETAILS_ALL	Cluster hierarchy, record counts, descriptive statistics (means, variances, modes, histograms, and rules) are computed. This is the default value.
KMNS_WINSORIZE	KMNS_WINSORIZE_ENABLE	KMNS_WINSORIZE_ENABLE	To winorize data, enable or disable this parameter. Data is restricted in a window size of six standard deviations around the mean value when winsorize is enabled. This functionality can be used with AUTO_DATA_PREP turned ON and OFF. The values outside the range are replaced with the ends of the interval. Winsorize is not enabled by default.

Note:

Winsorize is only available when the KMNS_EUCLIDEAN distance function is used. An exception is raised if Winsorize is enabled and other distance functions are set.

Table 6-24 (Cont.) k-Means Settings

Setting Name	Constant Value	String Value Equivalent	Description
	`KMNS_WINSORIZE_DISABLE`	`KMNS_WINSORIZE_DISABLE`	Disables winsorization for K-Means clustering. When disabled, extreme values in the data are not adjusted, potentially leading to sensitivity to outliers.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> - For generic machine learning function settings related to Clustering, see DBMS_DATA_MINING — Machine Learning Functions.
> - *Oracle Machine Learning for SQL Concepts* for information about *k*-Means

DBMS_DATA_MINING - Algorithm Settings: Multivariate State Estimation Technique - Sequential Probability Ratio Test

Settings that configure the training calibration behavior of the Multivariate State Estimation Technique - Sequential Probability Ratio Test algorithm.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.MSET_ADB_HEIGHT`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'MSET_ADB_HEIGHT'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-25 MSET-SPRT Settings

Setting Name	Setting Value	String Value Equivalent	Description
MSET_ADB_HEIGHT	A positive double	A positive double	Estimates the band within which signal values normally oscillate. The default value is 0.05.
MSET_ALERT_COUNT	A positive integer	A positive integer	The number of the last *n* signals (the alert window) that should have passed the threshold to raise an alert. The alert count should be lower or equal to the alert window. The default value is 5.
MSET_ALERT_WINDOW	A positive integer greater than or equal to MSET_ALERT_COUNT	A positive integer greater than or equal to MSET_ALERT_COUNT	The number of signals to consider in the SPRT hypothesis consolidation logic. The default value is 5.
MSET_ALPHA_PROB	A positive double between 0 and 1	A positive double between 0 and 1	False Alarm Probability FAP (false positive). The default is 0.01.
MSET_BETA_PROB	A positive double between 0 and 1	A positive double between 0 and 1	Missed Alarm Probability MAP (false negative). The default is 0.10.
MSET_HELDASIDE	A positive integer	A positive integer	The approximate number of data rows used for MSET model calibration. You can use ODMS_RANDOM_SEED to change the held-aside sample. The default value is 10000.
MSET_MEMORY_VECTORS	A positive integer	A positive integer	The default value is data driven.
MSET_PROJECTION_THRESHOLD	A positive integer >0, <=10000	A positive integer >0, <=10000	Specifies whether to use random projections. When the number of sensors exceeds the setting value, random projections are used. To turn off random projections, set the threshold to a value that is equal to or greater than the number of sensors. The default value is 500.
MSET_STD_TOLERANCE	A positive integer	A positive integer	The tolerance in standard deviations used in the SPRT calculation. The default value is 3.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

DBMS_DATA_MINING — Algorithm Settings: Naive Bayes

The settings listed in the following table configure the behavior of the Naive Bayes algorithm.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.NABS_PAIRWISE_THRESHOLD`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'NABS_PAIRWISE_THRESHOLD'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-26 Naive Bayes Settings

Setting Name	Setting Value	String Value Equivalent	Description
`NABS_PAIRWISE_THRESHOLD`	A floating point number between 0 and 1, inclusive, expressed as a character string	A floating point number between 0 and 1, inclusive, expressed as a character string	Value of pairwise threshold for NB algorithm Default is `0`. Expression: `TO_CHAR(0.5)`
`NABS_SINGLETON_THRESHOLD`	A floating point number between 0 and 1, inclusive, expressed as a character string	A floating point number between 0 and 1, inclusive, expressed as a character string	Value of singleton threshold for NB algorithm Default value is `0`. Expression: `TO_CHAR(0.5)`

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about Naive Bayes

DBMS_DATA_MINING — Algorithm Settings: Neural Network

The settings listed in the following table configure the behavior of the Neural Network algorithm.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.NNET_SOLVER_ADAM`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the DBMS_DATA_MINING prefix, in single quotes. For example, 'NNET_SOLVER_ADAM'.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-27 DBMS_DATA_MINING Neural Network Settings

Setting Name	Constant Value	String Value Equivalents	Description
NNET_SOLVER	One of the following strings: NNET_SOLVER_ADAM	NNET_SOLVER_ADAM	Specifies the method of optimization. The default value is system determined. NNET_SOLVER_ADAM: Uses the Adam optimization method.
	NNET_SOLVER_LBFGS	NNET_SOLVER_LBFGS	Uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) optimization method.

Table 6-27 (Cont.) DBMS_DATA_MINING Neural Network Settings

Setting Name	Constant Value	String Value Equivalents	Description

`NNET_ACTIVATIONS`	One or more of the following strings:
	`NNET_ACTIVATIONS_ARCTAN`
		`NNET_ACTIVATIONS_ARCTAN`	Specifies the activation functions for the hidden layers. You can specify a single activation function, which is then applied to each hidden layer, or you can specify an activation function for each layer individually. Different layers can have different activation functions.
			To apply a different activation function to one or more of the layers, you must specify an activation function for each layer. The number of activation functions you specify must be consistent with the `NNET_HIDDEN_LAYERS` and `NNET_NODES_PER_LAYER` values.
			For example, if you have three hidden layers, you could specify the use of the same activation function for all three layers with the following settings value:
			`('NNET_ACTIVATIONS',`
			`'NNET_ACTIVATIONS_TANH')`
			The following settings value specifies a different activation function for each layer:
			`('NNET_ACTIVATIONS',`
			`'''NNET_ACTIVATIONS_TANH'',`
			`''NNET_ACTIVATIONS_LOG_SIG'',`
			`''NNET_ACTIVATIONS_ARCTAN''')`
			Note:
			You specify the different activation functions as strings within a single string. All quotes are single and two single quotes are used to escape a single quote in SQL statements and PL/SQL blocks.
			`NNET_ACTIVATIONS_ARCTAN`: Uses the arctangent activation function.
			The default value is `NNET_ACTIVATIONS_LOG_SIG`.

Table 6-27 (Cont.) DBMS_DATA_MINING Neural Network Settings

Setting Name	Constant Value	String Value Equivalents	Description
	NNET_ACTIVATIONS_BIPOLAR_SIG	NNET_ACTIVATIONS_BIPOLAR_SIG	Uses the bipolar sigmoid activation function.
	NNET_ACTIVATIONS_LINEAR	NNET_ACTIVATIONS_LINEAR	Uses the linear activation function.
	NNET_ACTIVATIONS_LOG_SIG	NNET_ACTIVATIONS_LOG_SIG	Uses the logistic sigmoid activation function.
	NNET_ACTIVATIONS_RELU	NNET_ACTIVATIONS_RELU	Uses the rectified linear unit activation function.
	NNET_ACTIVATIONS_TANH	NNET_ACTIVATIONS_TANH	Uses the hyperbolic tangent activation function.
NNET_HELDASIDE_MAX_FAIL	A positive integer	A positive integer	With NNET_REGULARIZER_HELDASIDE, the training process is stopped early if the network performance on the validation data fails to improve or remains the same for NNET_HELDASIDE_MAX_FAIL epochs in a row. The default value is 6.
NNET_HELDASIDE_RATIO	An integer greater than 0 and less than or equal to 1, represented as a character string	An integer greater than 0 and less than or equal to 1, represented as a character string	Define the held ratio for the held-aside method. The default value is 0.25. Expression: TO_CHAR(0.45)
NNET_HIDDEN_LAYERS	A positive integer	A positive integer	Defines the topology by the number of hidden layers. The default value is 1.
NNET_ITERATIONS	A positive integer	A positive integer	Specifies the maximum number of iterations in the Neural Network algorithm. For the DMSSET_NN_SOLVER_LBFGS solver, the default value is 200. For the DMSSET_NN_SOLVER_ADAM solver, the default value is 10000.
NNET_NODES_PER_LAYER	A positive integer or a list of positive integers	A positive integer or a list of positive integers	Defines the topology by the number of nodes per layer. Different layers can have different numbers of nodes. To specify the same number of nodes for each layer, you can provide a single value, which is then applied to each layer. To specify a different number of nodes for one or more layers, provide a list of comma-separated positive integers, one for each layer. For example, '10, 20, 5' for three layers. The setting values must be consistent with the NNET_HIDDEN_LAYERS value. The default number of nodes per layer is the number of attributes or 50 (if the number of attributes > 50).

Table 6-27 (Cont.) DBMS_DATA_MINING Neural Network Settings

Setting Name	Constant Value	String Value Equivalents	Description
NNET_REG_LAMBDA	An integer greater than or equal to 0 represented as a character string	An integer greater than or equal to 0 represented as a character string	Defines the L2 regularization parameter lambda. This can not be set together with NNET_REGULARIZER_HELDASIDE. The default value is 1. Expression: TO_CHAR(2)
NNET_REGULARIZER	One of the following strings: NNET_REGULARIZER_HELDASIDE	NNET_REGULARIZER_HELDASIDE	Regularization setting for Neural Network algorithm. NNET_REGULARIZER_HELDASIDE: Uses a held-aside method for regularization. If the total number of training rows is greater than 50000, the default is NNET_REGULARIZER_HELDASIDE.
	NNET_REGULARIZER_L2	NNET_REGULARIZER_L2	Applies L2 regularization, which penalizes the sum of squared weights.
	NNET_REGULARIZER_NONE	NNET_REGULARIZER_NONE	Disables regularization. If the total number of training rows is less than or equal to 50000, the default is NNET_REGULARIZER_NONE.
NNET_TOLERANCE	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	Defines the convergence tolerance setting of the Neural Network algorithm. The default value is 0.000001. Expression: TO_CHAR(0.00004)
NNET_WEIGHT_LOWER_BOUND	A real number	A real number	The setting specifies the lower bound of the region where weights are randomly initialized. NNET_WEIGHT_LOWER_BOUND and NNET_WEIGHT_UPPER_BOUND must be set together. Setting one and not setting the other raises an error. NNET_WEIGHT_LOWER_BOUND must not be greater than NNET_WEIGHT_UPPER_BOUND. The default value is -sqrt(6/(l_nodes+r_nodes)). The value of l_nodes for: • input layer dense attributes is (1+number of dense attributes) • input layer sparse attributes is number of sparse attributes • each hidden layer is (1+number of nodes in that hidden layer) The value of r_nodes is the number of nodes in the layer that the weight is connecting to.

Table 6-27 (Cont.) DBMS_DATA_MINING Neural Network Settings

Setting Name	Constant Value	String Value Equivalents	Description
NNET_WEIGHT_UPPER_BOUND	A real number	A real number	This setting specifies the upper bound of the region where weights are initialized. It should be set in pairs with NNET_WEIGHT_LOWER_BOUND and its value must not be smaller than the value of NNET_WEIGHT_LOWER_BOUND. If not specified, the values of NNET_WEIGHT_LOWER_BOUND and NNET_WEIGHT_UPPER_BOUND are system determined. The default value is sqrt(6/(l_nodes+r_nodes)). See NNET_WEIGHT_LOWER_BOUND.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.
- DBMS_DATA_MINING — Solver Settings: LBFGS
  The settings listed in the following table configure the behavior of L-BFGS. Neural Network and Generalized Linear Model (GLM) use these settings.

See Also:

Oracle Machine Learning for SQL Concepts for information about Neural Network.

DBMS_DATA_MINING — Algorithm Settings: Non-Negative Matrix Factorization

The settings listed in the following table configure the behavior of the Non-negative Matrix Factorization algorithm.

The **Constant Value** column specifies constants using the prefix DBMS_DATA_MINING. For example, DBMS_DATA_MINING.NMFS_NONNEG_SCORING_ENABLE. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the DBMS_DATA_MINING prefix, in single quotes. For example, 'NMFS_NONNEG_SCORING_ENABLE'.

Note:

The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

You can query the data dictionary view *_MINING_MODEL_SETTINGS (using the ALL, USER, or DBA prefix) to find the setting values for a model. See *Oracle Database Reference* for information about *_MINING_MODEL_SETTINGS.

Table 6-28 NMF Settings

Setting Name	Constant Value	String Value Equivalent	Description
NMFS_CONV_TOLERANCE	A floating point number between 0 and 0.5 expressed as a character string	A floating point number between 0 and 0.5 expressed as a character string	Convergence tolerance for NMF algorithm Default is 0.05 Expression: TO_CHAR(0.02)
NMFS_NONNEGATIVE_SCORING	NMFS_NONNEG_SCORING_ENABLE	NMFS_NONNEG_SCORING_ENABLE	Whether negative numbers should be allowed in scoring results. When set to NMFS_NONNEG_SCORING_ENABLE, negative feature values will be replaced with zeros. Default is NMFS_NONNEG_SCORING_ENABLE
	NMFS_NONNEG_SCORING_DISABLE	NMFS_NONNEG_SCORING_DISABLE	When set to NMFS_NONNEG_SCORING_DISABLE, negative feature values will be allowed.
NMFS_NUM_ITERATIONS	An integer between 1 to 500, inclusive, represented as a character string	An integer between 1 to 500, inclusive, represented as a character string	Number of iterations for NMF algorithm Default is 50 Expression: TO_CHAR(80)
NMFS_RANDOM_SEED	An integer represented as a character string	An integer represented as a character string	Random seed for NMF algorithm. Default is -1. Expression: TO_CHAR(2)

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about NMF

DBMS_DATA_MINING — Algorithm Settings: O-Cluster

The settings in the table configure the behavior of the O-Cluster algorithm.

The **Constant Value** column specifies constants using the prefix DBMS_DATA_MINING. For example, DBMS_DATA_MINING.OCLT_SENSITIVITY. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'OCLT_SENSITIVITY'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-29 O-CLuster Settings

Setting Name	Constant Value	String Value Equivalent	Description
`OCLT_SENSITIVITY`	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	A fraction that specifies the peak density required for separating a new cluster. The fraction is related to the global uniform density. Default is `0.5`. Example: `TO_CHAR(0.9)`

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about O-Cluster

DBMS_DATA_MINING — Algorithm Settings: Random Forest

These settings configure the behavior of the Random Forest algorithm. Random Forest makes use of the Decision Tree settings to configure the construction of individual trees.

The **Constant Value** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.RFOR_MTRY`. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'RFOR_MTRY'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-30 Random Forest Settings

Setting Name	Constant Value	String Value Equivalent	Description
`RFOR_MTRY`	`a number >= 0`	`a number >= 0`	Size of the random subset of columns to be considered when choosing a split at a node. For each node, the size of the pool remains the same, but the specific candidate columns change. The default is half of the columns in the model signature. The special value `0` indicates that the candidate pool includes all columns.
`RFOR_NUM_TREES`	`1<= a number <=65535`	`1<= a number <=65535`	Number of trees in the forest Default is `20`.
`RFOR_SAMPLING_RATIO`	`0< a fraction<=1`	`0< a fraction<=1`	Fraction of the training data to be randomly sampled for use in the construction of an individual tree. The default is half of the number of rows in the training data.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.
- DBMS_DATA_MINING — Algorithm Settings: Decision Tree
  These settings configure the behavior of the Decision Tree algorithm. Note that the Decision Tree settings are also used to configure the behavior of Random Forest as it constructs each individual decision tree.

See Also:

Oracle Machine Learning for SQL Concepts for information about Random Forest

DBMS_DATA_MINING — Algorithm Constants and Settings: Singular Value Decomposition

The following settings configure the behavior of the Singular Value Decomposition algorithm.

Table 6-31 Singular Value Decomposition Settings

Setting Name	Constant Value	String Value Equivalent	Description
`SVDS_U_MATRIX_OUTPUT`	`SVDS_U_MATRIX_ENABLE`	`SVDS_U_MATRIX_ENABLE`	Indicates whether or not to persist the **U** Matrix produced by SVD. The U matrix in SVD has as many rows as the number of rows in the build data. To avoid creating a large model, the U matrix is persisted only when `SVDS_U_MATRIX_OUTPUT` is enabled. When `SVDS_U_MATRIX_OUTPUT` is enabled, the build data must include a case ID. If no case ID is present and the U matrix is requested, then an exception is raised. Default is `SVDS_U_MATRIX_DISABLE`.
	`SVDS_U_MATRIX_DISABLE`	`SVDS_U_MATRIX_DISABLE`	Does not persist the U Matrix.
`SVDS_SCORING_MODE`	`SVDS_SCORING_SVD`	`SVDS_SCORING_SVD`	Whether to use SVD or PCA scoring for the model. When the build data is scored with SVD, the projections will be the same as the U matrix. Default is `SVDS_SCORING_SVD`.
	`SVDS_SCORING_PCA`	`SVDS_SCORING_PCA`	When the build data is scored with PCA, the projections will be the product of the U and S matrices.

Table 6-31 (Cont.) Singular Value Decomposition Settings

Setting Name	Constant Value	String Value Equivalent	Description
SVDS_SOLVER	SVDS_SOLVER_TSSVD	SVDS_SOLVER_TSSVD	This setting indicates the solver to be used for computing SVD of the data. In the case of PCA, the solver setting indicates the type of SVD solver used to compute the PCA for the data. When this setting is not specified the solver type selection is data driven. If the number of attributes is greater than 3240, then the default wide solver is used. Otherwise, the default narrow solver is selected. The following are the group of solvers: • Narrow data solvers: for matrices with up to 11500 attributes (TSEIGEN) or up to 8100 attributes (TSSVD). • Wide data solvers: for matrices up to 1 million attributes. For narrow data solvers: • Tall-Skinny SVD uses QR computation TSVD (SVDS_SOLVER_TSSVD) • Tall-Skinny SVD uses eigenvalue computation, TSEIGEN (SVDS_SOLVER_TSEIGEN), is the default solver for narrow data. For wide data solvers: • Stochastic SVD uses QR computation SSVD (SVDS_SOLVER_SSVD), is the default solver for wide data solvers. • Stochastic SVD uses eigenvalue computations, STEIGEN (SVDS_SOLVER_STEIGEN).
	SVDS_SOLVER_TSEIGEN	SVDS_SOLVER_TSEIGEN	Tall-Skinny SVD using eigenvalue computation for matrices with up to 11500 attributes. This is the default solver for narrow data.
	SVDS_SOLVER_SSVD	SVDS_SOLVER_SSVD	Stochastic SVD using QR computation for matrices with up to 1 million attributes. This is the default solver for wide data.
	SVDS_SOLVER_STEIGEN	SVDS_SOLVER_STEIGEN	Stochastic SVD using eigenvalue computations for matrices with up to 1 million attributes.
SVDS_TOLERANCE	Range [0, 1]	Range [0, 1]	This setting is used to prune features. Define the minimum value the eigenvalue of a feature as a share of the first eigenvalue to not to prune. Default value is data driven.
SVDS_RANDOM_SEED	Range [0 - 4,294,967,296]	Range [0 - 4,294,967,296]	The random seed value is used for initializing the sampling matrix used by the Stochastic SVD solver. The default is 0. The SVD Solver must be set to SSVD or STEIGEN.

Table 6-31 (Cont.) Singular Value Decomposition Settings

Setting Name	Constant Value	String Value Equivalent	Description
SVDS_OVER_SAMPLING	Range [1, 5000].	Range [1, 5000].	This setting is configures the number of columns in the sampling matrix used by the Stochastic SVD solver. The number of columns in this matrix is equal to the requested number of features plus the oversampling setting. The SVD Solver must be set to SSVD or STEIGEN.
SVDS_POWER_ITERATIONS	Range [0, 20].	Range [0, 20].	The power iteration setting improves the accuracy of the SSVD solver. The default is 2. The SVD Solver must be set to SSVD or STEIGEN.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

See Also:

Oracle Machine Learning for SQL Concepts

DBMS_DATA_MINING — Algorithm Settings: Support Vector Machine

The settings listed in the following table configure the behavior of the Support Vector Machine algorithm.

The **Constant Value** column specifies constants using the prefix DBMS_DATA_MINING. For example, DBMS_DATA_MINING.SVMS_GAUSSIAN. Alternatively, you can specify the corresponding string value from the **String Value Equivalent** column without the DBMS_DATA_MINING prefix, in single quotes. For example, 'SVMS_GAUSSIAN'.

Note:

The distinction between **Constant Value** and **String Value Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

Table 6-32 SVM Settings

Setting Name	Constant Value	String Value Equivalent	Description
SVMS_COMPLEXITY_FACTOR	An integer greater than 0 represented as a character string	An integer greater than 0 represented as a character string	Regularization setting that balances the complexity of the model against model robustness to achieve good generalization on new data. SVM uses a data-driven approach to finding the complexity factor. Value of complexity factor for SVM algorithm (both classification and regression). Default value estimated from the data by the algorithm. Expression: TO_CHAR(20)
SVMS_CONV_TOLERANCE	An integer greater than 0 represented as a character string	An integer greater than 0 represented as a character string	Convergence tolerance for SVM algorithm. Default is 0.0001. Expression: TO_CHAR(0.005)
SVMS_EPSILON	An integer greater than 0 represented as a character string	An integer greater than 0 represented as a character string	Regularization setting for regression, similar to complexity factor. Epsilon specifies the allowable residuals, or noise, in the data. Value of epsilon factor for SVM regression. Default is 0.1. Expression: TO_CHAR(0.5)
SVMS_KERNEL_FUNCTION	SVMS_GAUSSIAN	SVMS_GAUSSIAN	Kernel for Support Vector Machine. Linear or Gaussian. SVMS_GAUSSIAN: Uses the Gaussian kernel for SVM. The default value is SVMS_LINEAR.
	SVMS_LINEAR	SVMS_LINEAR	Uses the Linear kernel for SVM. This is the default option.
SVMS_OUTLIER_RATE	A floating point number between 0 and 1 expressed as a character string	A floating point number between 0 and 1 expressed as a character string	The desired rate of outliers in the training data. Valid for One-Class SVM models only (anomaly detection). Default is 0.01. Expression: TO_CHAR(0.04)
SVMS_STD_DEV	An integer greater than 0 represented as a character string	An integer greater than 0 represented as a character string	Controls the spread of the Gaussian kernel function. SVM uses a data-driven approach to find a standard deviation value that is on the same scale as distances between typical cases. Value of standard deviation for SVM algorithm. This is applicable only for Gaussian kernel. Default value estimated from the data by the algorithm. Expression: TO_CHAR(6)

Table 6-32 (Cont.) SVM Settings

Setting Name	Constant Value	String Value Equivalent	Description
SVMS_NUM_ITERATIONS	A positive integer	A positive integer	This setting sets an upper limit on the number of SVM iterations. The default is system determined because it depends on the SVM solver.
SVMS_NUM_PIVOTS	Range [1; 10000]	Range [1; 10000]	This setting sets an upper limit on the number of pivots used in the Incomplete Cholesky decomposition. It can be set only for non-linear kernels. The default value is 200.
SVMS_BATCH_ROWS	A positive integer	A positive integer	This setting applies to SVM models with linear kernel. This setting sets the size of the batch for the SGD solver. An input of 0 triggers a data driven batch size estimate. The default is 20000.
SVMS_REGULARIZER	SVMS_REGULARIZER_L1	SVMS_REGULARIZER_L1	This setting controls the type of regularization that the SGD SVM solver uses. The setting can be used only for linear SVM models. The default is system determined because it depends on the potential model size. SVMS_REGULARIZER_L1: Uses L1 regularization.
	SVMS_REGULARIZER_L2	SVMS_REGULARIZER_L2	Uses L2 regularization.
SVMS_SOLVER	SVMS_SOLVER_SGD (Sub-Gradient Descend)	SVMS_SOLVER_SGD (Sub-Gradient Descend)	Enables to choose the SVM solver. The SGD solver cannot be selected if the kernel is non-linear. The default value is system determined. SVMS_SOLVER_SGD: Uses Sub-Gradient Descent solver.
	SVMS_SOLVER_IPM (Interior Point Method)	SVMS_SOLVER_IPM (Interior Point Method)	Uses Interior Point Method solver.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about SVM

DBMS_DATA_MINING — Algorithm Settings: XGBoost

Settings that configure the behavior of the XGBoost gradient boosting algorithm.

The **Constant Name** column specifies constants using the prefix `DBMS_DATA_MINING`. For example, `DBMS_DATA_MINING.xgboost_booster`. Alternatively, you can specify the corresponding string value from the **String Name Equivalent** column without the `DBMS_DATA_MINING` prefix, in single quotes. For example, `'booster'`.

> **Note:**
>
> The distinction between **Constant Value** and **String Name Equivalent** for this algorithm is applicable to Oracle Database 19*c* and Oracle Database 21*c*.

The XGBoost settings are case sensitive. Enter the settings as they appear in the settings table. These settings match the XGBoost settings available in open source. OML4SQL XGBoost is based on the 1.7.4 version of XGBoost.
For Global settings, see DBMS_DATA_MINING — Global Settings.

For generic machine learning technique settings, see DBMS_DATA_MINING — Machine Learning Functions.

Table 6-33 General Settings

Constant Name	String Name Equivalent	Setting Value	Description

`xgboost_booster`	`booster`	A string that is one of the following:
		`dart`
		`gblinear`
		`gbtree`
			The booster to use:
			• `dart`
			• `gblinear`
			• `gbtree`
			The `dart` and `gbtree` boosters use tree-based models whereas `gblinear` uses linear functions.
			The default value is `gbtree`.

`xgboost_num_round`	`num_round`	A non-negative integer.	The number of rounds for boosting. The default value is `10`.

Table 6-34 Settings for Tree Boosting

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_alpha`	`alpha`	A non-negative number	L1 regularization term on weights. Increasing this value makes the model more conservative. The default value is `0`.
`xgboost_colsample_bylevel`	`colsample_bylevel`	A number in the range (0, 1]	Subsample ratio of columns for each split, in each level. Subsampling occurs each time a new split is made. This parameter has no effect when `tree_method` is set to `hist`. The default value is `1`.

Table 6-34 (Cont.) Settings for Tree Boosting

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_colsample_bynode`	`colsample_bynode`	A number in the range (0, 1]	The subsample ratio of columns for each node (split). Subsampling occurs once every time a new split is evaluated. Columns are subsampled from the set of columns chosen for the current level. The default value is `1`.
`xgboost_colsample_bytree`	`colsample_bytree`	A number in the range (0, 1]	Subsample ratio of columns when constructing each tree. Subsampling occurs once in every boosting iteration. The default value is `1`.
`xgboost_eta`	`eta`	A number in the range [0, 1]	Step-size shrinkage used in the update step to prevent overfitting. After each boosting step, `eta` shrinks the feature weights to make the boosting process more conservative. The default value is `0.3`.
`xgboost_gamma`	`gamma`	A number in the range [0, ∞]	Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma value is, the more conservative the algorithm is. The default value is `0`.
`xgboost_grow_policy`	`grow_policy`	A string; one of the following: • `depthwise` • `lossguide`	Controls the way new nodes are added to the tree: • `depthwise` splits at nodes closest to the root • `lossguide` splits at nodes with the highest loss change Valid only if `tree_method` is set to `hist`. The default value is `depthwise`.
`xgboost_interaction_constraints`	`interaction_constraints`	`[[x0,x1,x2], [x0,x4],[x5,x6]]` where `xn` are feature names or columns	This setting specifies permitted interactions in the model. Specify the constrains in the form of a nested list where each inner list is a group of features (column names) that are allowed to interact with each other. If a single column is passed in the interactions then, the input is ignored. Here, features x0, x1, and x2 are allowed to interact with each other but with no other feature. Similarly, x0 and x4 are allowed to interact with each other but with no other feature and so on. This setting is applicable to 2-Dimensional features. An error occurs if you pass columns of non-supported type and non-existing feature names.
`xgboost_lambda`	`lambda`	A non-negative number	L2 regularization term on weights. The default value is `1`.

Table 6-34 (Cont.) Settings for Tree Boosting

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_max_bin`	`max_bin`	A non-negative integer	Maximum number of discrete bins to bucket continuous features. Increasing this number improves the optimality of splits at the cost of higher computation time. This parameter is valid only when `tree_method` is set to `hist`. The default value is `256`.
`xgboost_max_delta_step`	`max_delta_step`	A number in the range [0, ∞]	Maximum delta step allowed for each leaf output. Setting this to a positive value can help make the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when the class is extremely imbalanced. Setting it to value from 1 to 10 might help control the update. The default value is `0`, which means there is no constraint.
`xgboost_max_depth`	`max_depth`	An integer in the range [0, ∞]	Maximum depth of a tree. Increasing this value makes the model more complex and more likely to overfit. Setting this to 0 indicates no limit. **Note:** You must set a `max_depth` limit when the `grow_policy` setting is `depthwise`. The default value is `6`.
`xgboost_max_leaves`	`max_leaves`	A non-negative number	Maximum number of nodes to add. Use this setting only when `grow_policy` is set to `lossguide`. The default value is `0`.
`xgboost_min_child_weight`	`min_child_weight`	A number in the range [0, ∞]	Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with a sum of instance weight less than `min_child_weight`, then the building process stops partitioning. In a linear regression task, this corresponds to the minimum number of instances needed in each node. The larger `min_child_weight` is, the more conservative the algorithm is. The default value is `1`.

Table 6-34 (Cont.) Settings for Tree Boosting

Constant Name	String Name Equivalent	Setting Value	Description
xgboost_monotone_decrease_constraints	monotone_decrease_constraints	'x4,x5'	This setting specifies the features (column names) that must obey decreasing constraint. The feature names are separated by a comma. For example, setting value 'x4,x5' sets decreasing constraint on features x4 and x5. This setting applies to numeric columns and 2-Dimensional features. An error occurs if you pass columns of non-supported type and non-existing feature names.
xgboost_monotone_increase_constraints	monotone_increase_constraints	'x0,x3'	This setting specifies the features (column names) that must obey increasing constraint. The feature names are separated by a comma. For example, setting value 'x0,x3' sets increasing constraint on features x0 and x3. This setting is applicable to 2-Dimensional features. An error occurs if you pass columns of non-supported type and non-existing feature names.
xgboost_num_parallel_tree	num_parallel_tree	A non-negative integer	Number of parallel trees constructed during each iteration. Use this option to support a boosted random forest. The default value is 1.
xgboost_scale_pos_weight	scale_pos_weight	A non-negative number	Controls the balance of positive and negative weights, which is useful for unbalanced classes. A typical value to consider: sum(negative cases) / sum(positive cases). The default value is 1.
xgboost_sketch_eps	sketch_eps	A number in the range (0, 1)	Increases enumeration accuracy. Valid only for the approximate greedy tree method. Compared to directly selecting the number of bins, this setting comes with a theoretical guarantee with sketch accuracy. You usually do not need to change this setting, but you might consider setting a lower number for more accurate enumeration. The default value is 0.03.
xgboost_subsample	subsample	A number in the range (0, 1]	Subsample ratio of the training instances. A setting of 0.5 means that XGBoost randomly samples half of the training data prior to growing trees, which prevents overfitting. Subsampling occurs once in every boosting iteration. The default value is 1.

Table 6-34 (Cont.) Settings for Tree Boosting

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_tree_method`	`tree_method`	A string that is one of the following: • `approx` • `auto` • `exact` • `hist`	Tree construction algorithm used in XGBoost: • `approx`: Approximate greedy algorithm using sketching and histogram. • `auto`: Use a heuristic to choose the faster algorithm: – For a small to medium sized data set, uses the exact greedy algorithm. – For a very large data set, uses the approximate greedy algorithm. • `exact`: Exact greedy algorithm. • `hist`: Fast histogram optimized approximate greedy algorithm; uses some performance improvements such as bins caching. The default value is `auto`.
`xgboost_updater`	`updater`	A comma-separated string; one or more of the following: • `grow_colmaker` • `grow_histmaker` • `grow_skmaker` • `grow_quantile_histmaker` • `prune` • `sync`	Defines the sequence of tree updaters to run, which provides a modular way to construct and to modify the trees. This is an advanced parameter that is usually set automatically, depending on some other parameters. However, you can also explicitly specify a settting. The setting values are: • `grow_colmaker`: Non-distributed column-based construction of trees. • `grow_histmaker`: Distributed tree construction with row-based data splitting based on a global proposal of histogram counting. • `grow_skmaker`: Uses the approximate sketching algorithm. • `grow_quantile_histmaker`: Grow tree using quantized histogram. • `prune`: Prunes the splits where loss < `min_split_loss` (or `gamma`). • `sync`: Synchronizes trees in all distributed nodes.

Table 6-35 Settings for the Dart Booster

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_one_drop`	`one_drop`	A number that is 0 or 1	When set to 1, at least one tree is always dropped during the dropout. When set to 0, at least one tree is not always dropped during the dropout. The default value is `0`.

Table 6-35 (Cont.) Settings for the Dart Booster

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_normalize_type`	`normalize_type`	A string; either: • `forest` • `tree`	Type of normalization algorithm: • `forest`: New trees have the same weight as the sum of the dropped trees (forest): – The weight of new trees is `1 / (1 + learning_rate)` – Dropped trees are scaled by a factor of `1 / (1 + learning_rate)` • `tree`: New trees have the same weight as dropped trees: – The weight of new trees is `1 / (k + learning_rate)` – Dropped trees are scaled by a factor of `k / (k + learning_rate)` The default value is `tree`.
`xgboost_rate_drop`	`rate_drop`	A number in the range [0.0, 1.0]	Dropout rate (a fraction of the previous trees to drop during the dropout). The default value is `0.0`.
`xgboost_sample_type`	`sample_type`	A string; either: • `uniform` • `weighted`	Type of sampling algorithm: • `uniform`: Dropped trees are selected uniformly • `weighted`: Dropped trees are selected in proportion to weight The default value is `uniform`.
`xgboost_skip_drop`	`skip_drop`	A number in the range [0.0, 1.0]	Probability of skipping the dropout procedure during a boosting iteration. If a dropout is skipped, new trees are added in the same manner as `gbtree`. A non-zero `skip_drop` has higher priority than `rate_drop` or `one_drop`. The default value is `0.0`.

Table 6-36 Settings for the Linear Booster

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_alpha`	`alpha`	A non-negative number	L1 regularization term on weights, normalized to the number of training examples. Increasing this value makes the model more conservative. The default value is `0`.

Table 6-36 (Cont.) Settings for the Linear Booster

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_feature_selector`	`feature_selector`	A string that is one of the following: • `cyclic` • `greedy` • `random` • `shuffle` • `thrifty`	Feature selection and ordering method: • `cyclic`: Deterministic selection by cycling through the features one at a time. • `greedy`: Selects the coordinate with the greatest gradient magnitude. This method: – Has `O(num_feature^2)` complexity – Is fully deterministic – Allows restricting the selection to the `top_k` features per group with the largest magnitude of univariate weight change, by setting the `top_k` parameter; doing so reduces the complexity to `O(num_feature*top_k)`. • `random`: A random (with replacement) coordinate selector. • `shuffle`: Similar to `cyclic` but with random feature shuffling prior to each update. • `thrifty`: Thrifty, approximately-greedy feature selector. Prior to cyclic updates, reorders features in descending magnitude of their univariate weight changes. This operation is multithreaded and is a linear complexity approximation of the quadratic greedy selection. Restricts the selection per group to the `top_k` features with the largest magnitude of univariate weight change. The default value is `cyclic`.
`xgboost_lambda`	`lambda`	A non-negative number	L2 regularization term on weights, normalized to the number of training examples. Increasing this value makes the model more conservative. The default value is `0`.
`xgboost_top_k`	`top_k`	A non-negative integer	Number of top features to select for the `greedy` or `thrifty` feature selector. The value of 0 uses all of the features. The default value is `0`.
`xgboost_updater`	`updater`	A string that is one of the following: • `coord_descent` • `shotgun`	Algorithm to fit the linear model: • `coord_descent`: Ordinary coordinate descent algorithm; multithreaded but still produces a deterministic solution. • `shotgun`: Parallel coordinate descent algorithm based on the `shotgun` algorithm; uses "hogwild" parallelism and therefore produces a nondeterministic solution on each run. The default value is `shotgun`.

Table 6-37 Settings for Tweedie Regression

Constant Name	String Name Equivalent	Setting Value	Description
`xgboost_tweedie_variance_power`	`tweedie_variance_power`	A number in the range (1, 2)	Controls the variance of the Tweedie distribution `var(y) ~ E(y)^tweedie_variance_power`. A setting closer to 1 shifts towards a Poisson distribution. A setting closer to 2 shifts towards a gamma distribution. The default value is `1.5`.

Some XGBoost objectives apply only to classification function models and other objectives apply only to regression function models. If you specify an incompatible `objective` value, an error is raised. In the `DBMS_DATA_MINING.CREATE_MODEL` procedure, if you specify `DBMS_DATA_MINING.CLASSIFICATION` as the function, then the only objective values that you can use are the `binary` and `multi` values. The one exception is `binary: logitraw`, which produces a continuous value and applies only to a regression model. If you specify `DBMS_DATA_MINING.REGRESSION` as the function, then you can specify `binary: logitraw` or any of the `count`, `rank`, `reg`, and `survival` values as the objective.

Table 6-38 Settings for Learning Tasks

Setting Name	String Name Equivalent	Setting Value	Description
`xgboost_objective`	`objective`	For a classification model, a string that is one of the following: • `binary:hinge` • `binary:logistic` • `multi:softmax` • `multi:softprob` For a regression model, a string that is one of the following: • `binary:logitraw` • `count:poisson` • `rank:map` • `rank:ndcg` • `rank:pairwise` • `reg:gamma` • `reg:logistic` • `reg:tweedie` • `survival:aft` • `survival:cox` • `reg:squarederror` • `reg:squaredlogerror`	**Settings for a Classification model:** • `binary:hinge`: Hinge loss for binary classification. This setting makes predictions of 0 or 1, rather than producing probabilities. • `binary:logistic`: Logistic regression for binary classification. The output is the probability. • `multi:softmax`: Performs multiclass classification using the `softmax` objective; you must also set `num_class(`*`number_of_classes`*`)`. • `multi:softprob`: : Same as `softmax`, except the output is a vector of `ndata * nclass`, which can be further reshaped to an `ndata * nclass` matrix. The result contains the predicted probability of each data point belonging to each class. The default `objective` value for classification is `multi:softprob`. **Settings for a Regression model:** • `binary:logitraw`: Logistic regression for binary classification; the output is the score before logistic transformation. • `count:poisson`: Poisson regression for count data; the output is the mean of the Poisson distribution. The `max_delta_step` value is set to 0.7 by default in Poisson regression to safeguard optimization. • `rank:map`: Using `LambdaMART`, performs list-wise ranking in which the Mean Average Precision (MAP) is maximized. • `rank:ndcg`: Using `LambdaMART`, performs list-wise ranking in which the Normalized Discounted Cumulative Gain (NDCG) is maximized. • `rank:pairwise`: Performs ranking by minimizing the pairwise loss. • `reg:gamma`: Gamma regression with log-link; the output is the mean of the gamma distribution. This setting might be useful for any outcome that might be gamma-distributed, such as modeling insurance claims severity. • `reg:logistic`: Logistic regression. • `reg:tweedie`: Tweedie regression with log-link. This setting might be useful for any outcome that might be Tweedie-distributed, such as modeling total loss in insurance. • `survival:aft`: Applies the Accelerated Failure Time (AFT) model for censored survival time data. When you select this

Table 6-38 (Cont.) Settings for Learning Tasks

Setting Name	String Name Equivalent	Setting Value	Description
			option, `eval_metric` uses `aft-nloglik` as the default value. • `survival:cox`: Cox regression for right-censored survival time data (negative values are considered right-censored). Predictions are returned on the hazard ratio scale (that is, as `HR = exp(marginal_prediction)` in the proportional hazard function `h(t) = h0(t) * HR)`. • `reg:squarederror`: Regression with squared loss. • `reg:squaredlogerror`: Regression with squared log loss. All input labels must be greater than -1. The default `objective` value for regression is `reg:squarederror`.
`xgboost_aft_loss_distribution`	`aft_loss_distribution`	[normal, logistic, extreme]	Specifies the distribution of the Z term in the AFT model. It specifies the Probabilty Density Function used by `survival:aft` objective and `aft-nloglik` evaluation metric. The default value is `normal`.
`xgboost_aft_loss_distribution_scale`	`aft_loss_distribution_scale`	A positive number	Specifies the scaling factor σ, which scales the size of Z term in the AFT model. The default value is `1`.
`xgboost_aft_right_bound_column_name`	`aft_right_bound_column_name`	*column_name*	Specifies the column containing the right bounds of the labels for an AFT model. You cannot select this parameter for a non-AFT model. **Note:** Oracle Machine Learning does not support `BOOLEAN` values for this setting.
`xgboost_base_score`	`base_score`	A number	Initial prediction score of all instances, global bias. For a sufficient number of iterations, changing this value does not have much effect. The default value is `0.5`.

Table 6-38 (Cont.) Settings for Learning Tasks

Setting Name	String Name Equivalent	Setting Value	Description
`xgboost_eval_metric`	`eval_metric`	A comma-separated string; one or more of the following: • `aft-nloglik` • `auc` • `aucpr` • `cox-nloglik` • `error` • `error@`*`t`* • `gamma-deviance` • `gamma-nloglik` • `logloss` • `mae` • `map` • `map@`*`n`* • `merror` • `mlogloss` • `ndcg` • `ndcg@`*`n`* • `poisson-nloglik` • `rmse` • `tweedie-nloglik@`*`rho`* • `ndcg-` • `map-` • `rmsle`	Evaluation metrics for validation data. You can specify one or more of these evaluation metrics: • `aft-nloglik`: Sets the `eval_metric` to negative log likelihood of AFT model. • `auc`: Area under the curve. • `aucpr`: Area under the PR curve. • `cox-nloglik`: Negative partial log-likelihood for Cox proportional hazards regression. • `error`: Binary classification error rate, calculated as the number of wrong cases divided by the number of all cases. For the predictions, the evaluation regards the instances with a prediction value larger than 0.5 as positive instances, and the others as negative instances. • `error@`*`t`*: You can specify a binary classification threshold value other than 0.5 by specifying a numerical value *t*; for example, `error@0.8`. • `gamma-deviance`: Residual deviance for `gamma` regression. • `gamma-nloglik`: Negative log-likelihood for `gamma` regression. • `logloss`: Negative log-likelihood. • `mae`: Mean absolute error. • `map`: Mean average precision. • `map@`*`n`*: Assigns the integer *n* as the cut-off value for the top positions in the lists for evaluation. • `merror`: Multiclass classification error rate calculated as the number of wrong cases divided by the number of all cases; the objective must be `multi:softprob` or `multi:softmax`. • `mlogloss`: Multiclass `logloss`; the objective must be `multi:softprob` or `multi:softmax`. • `ndcg`: Normalized Discounted Cumulative Gain. • `ndcg@`*`n`*: Assigns the integer *n* as the cut-off value for the top positions in the lists for evaluation. • `poisson-nloglik`: Negative log-likelihood for Poisson regression • `rmse`: Root Mean Square Error. • `tweedie-nloglik@`*`rho`*: Negative log-likelihood for Tweedie regression (at a specified value `rho` of the `tweedie_variance_power` parameter);

Table 6-38 (Cont.) Settings for Learning Tasks

Setting Name	String Name Equivalent	Setting Value	Description
			`rho` must be a number in the range (1, 2); for example, `tweedie-nloglik@1.8`. • `ndcg-` and `map-`: In XGBoost, NDCG and MAP will evaluate the score of a list without any positive samples as 1. By adding "-" in the evaluation metric XGBoost will evaluate these score as 0 to be consistent under some conditions. • `rmsle`: It is root mean square log error. This is the default metric of `reg:squaredlogerror` objective. This metric reduces errors generated by outliers in dataset. But because log function is employed, `rmsle` might output nan when prediction value is less than -1. A default metric is assigned according to the objective: • `error` for classification • `mean average precision` for ranking • `rmse` for regression
`xgboost_seed`	`seed`	A non-negative integer	Random number seed. The default value is `0`.

Related Topics

- DBMS_DATA_MINING — Machine Learning Functions
  A machine learning **function** refers to the methods for solving a given class of machine learning problems.
- DBMS_DATA_MINING — Global Settings
  The configuration settings in this table are applicable to any type of model, but are currently only implemented for specific algorithms.

> **See Also:**
>
> https://github.com/oracle/oracle-db-examples/tree/master/machine-learning/sql/, select the release, and browse for an example of XGBoost.

DBMS_DATA_MINING — Solver Settings

Oracle Machine Learning for SQL algorithms can use different solvers. Solver settings can be provided at build time in the settings table.

Related Topics

- DBMS_DATA_MINING - Solver Settings: Adam
  These settings configure the behavior of the Adaptive Moment Estimation (Adam) solver.

- DBMS_DATA_MINING — Solver Settings: ADMM
  The settings listed in the following table configure the behavior of Alternating Direction Method of Multipliers (ADMM). The Generalized Linear Model (GLM) algorithm uses these settings.
- DBMS_DATA_MINING — Solver Settings: LBFGS
  The settings listed in the following table configure the behavior of L-BFGS. Neural Network and Generalized Linear Model (GLM) use these settings.

DBMS_DATA_MINING - Solver Settings: Adam

These settings configure the behavior of the Adaptive Moment Estimation (Adam) solver.

Neural Network models use these settings.

Table 6-39 DBMS_DATA_MINING Adam Settings

Setting Name	Setting Value	Description
`ADAM_ALPHA`	A non-negative double precision floating point number in the interval (0; 1]	The learning rate for Adam. The default value is `0.001`.
`ADAM_BATCH_ROWS`	A positive integer	The number of rows per batch. The default value is `10000`.
`ADAM_BETA1`	A positive double precision floating point number in the interval [0; 1)	The exponential decay rate for the 1st moment estimates. The default value is `0.9`.
`ADAM_BETA2`	A positive double precision floating point number in the interval [0; 1)	The exponential decay rate for the 2nd moment estimates. The default value is `0.99`.
`ADAM_GRADIENT_TOLERANCE`	A positive double precision floating point number	The gradient infinity norm tolerance for Adam. The default value is `1E-9`.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Neural Network
  The settings listed in the following table configure the behavior of the Neural Network algorithm.

DBMS_DATA_MINING — Solver Settings: ADMM

The settings listed in the following table configure the behavior of Alternating Direction Method of Multipliers (ADMM). The Generalized Linear Model (GLM) algorithm uses these settings.

Table 6-40 DBMS_DATA_MINING ADMM Settings

Settings Name	Setting Value	Description
`ADMM_CONSENSUS`	A positive integer	It is a ADMM's consensus parameter. The value must be a positive number. The default value is `0.1`.
`ADMM_ITERATIONS`	A positive integer	The number of `ADMM` iterations. The value must be a positive integer. The default value is `50`.

Table 6-40 (Cont.) DBMS_DATA_MINING ADMM Settings

Settings Name	Setting Value	Description
ADMM_TOLERANCE	A positive integer	It is a tolerance parameter. The value must be a positive number. The default value is 0.0001

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Generalized Linear Model
  The settings listed in the following table configure the behavior of the Generalized Linear Model algorithm.
- *Oracle Machine Learning for SQL Concepts*

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about neural network

DBMS_DATA_MINING — Solver Settings: LBFGS

The settings listed in the following table configure the behavior of L-BFGS. Neural Network and Generalized Linear Model (GLM) use these settings.

Table 6-41 DBMS_DATA_MINING L-BFGS Settings

Setting Name	Setting Value	Description
LBFGS_GRADIENT_TOLERANCE	An integer greater than 0 represented as a character string	Defines gradient infinity norm tolerance for L-BFGS. Default value is 1E-9. Expression: TO_CHAR (0.000000002)
LBFGS_HISTORY_DEPTH	A positive integer.	Defines the number of historical copies kept in L-BFGS solver. The default value is 20.
LBFGS_SCALE_HESSIAN	LBFGS_SCALE_HESSIAN_ENABLE LBFGS_SCALE_HESSIAN_DISABLE	Defines whether to scale Hessian in L-BFGS or not. Default value is LBFGS_SCALE_HESSIAN_ENABLE.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Neural Network
  The settings listed in the following table configure the behavior of the Neural Network algorithm.
- DBMS_DATA_MINING — Algorithm Settings: Generalized Linear Model
  The settings listed in the following table configure the behavior of the Generalized Linear Model algorithm.

See Also:

Oracle Machine Learning for SQL Concepts for information about neural network

DBMS_DATA_MINING Datatypes

The DBMS_DATA_MINING package defines object data types for processing transactional data. The package also defines a type for user-specified transformations. These types are called DM_NESTED_*n*, where *n* identifies the Oracle data type of the nested attributes.

The Oracle Machine Learning for SQL object data types are described in the following table:

Table 6-42 DBMS_DATA_MINING Summary of Data Types

Datatype	Description
DM_NESTED_BINARY_DOUBLE	The name and value of a numerical attribute of type BINARY_DOUBLE.
DM_NESTED_BINARY_DOUBLES	A collection of DM_NESTED_BINARY_DOUBLE.
DM_NESTED_BINARY_FLOAT	The name and value of a numerical attribute of type BINARY_FLOAT.
DM_NESTED_BINARY_FLOATS	A collection of DM_NESTED_BINARY_FLOAT.
DM_NESTED_CATEGORICAL	The name and value of a categorical attribute of type CHAR, VARCHAR, or VARCHAR2.
DM_NESTED_CATEGORICALS	A collection of DM_NESTED_CATEGORICAL.
DM_NESTED_NUMERICAL	The name and value of a numerical attribute of type NUMBER or FLOAT.
DM_NESTED_NUMERICALS	A collection of DM_NESTED_NUMERICAL.
ORA_MINING_VARCHAR2_NT	A table of VARCHAR2(4000).
TRANSFORM_LIST	A list of user-specified transformations for a model. Accepted as a parameter by the CREATE_MODEL Procedure. This collection type is defined in the DBMS_DATA_MINING_TRANSFORM package.

For more information about processing nested data, see *Oracle Machine Learning for SQL User's Guide*.

Note:

Starting from Oracle Database 12*c* Release 2, *GET_MODEL_DETAILS are deprecated and are replaced with *Model Detail Views*. See *Oracle Machine Learning for SQL User's Guide*.

Deprecated Types

This topic contains tables listing deprecated types.

The DBMS_DATA_MINING package defines object datatypes for storing information about model attributes. Most of these types are returned by the table functions GET_*n*, where *n* identifies the type of information to return. These functions take a model name as input and return the requested information as a collection of rows.

For a list of the GET functions, see "Summary of DBMS_DATA_MINING Subprograms".

All the table functions use pipelining, which causes each row of output to be materialized as it is read from model storage, without waiting for the generation of the complete table object. For more information on pipelined, parallel table functions, consult the *Oracle Database PL/SQL Language Reference.*

Table 6-43 DBMS_DATA_MINING Summary of Deprecated Datatypes

Datatype	Description
DM_CENTROID	The centroid of a cluster.
DM_CENTROIDS	A collection of DM_CENTROID. A member of DM_CLUSTER.
DM_CHILD	A child node of a cluster.
DM_CHILDREN	A collection of DM_CHILD. A member of DM_CLUSTER.
DM_CLUSTER	A cluster. A cluster includes DM_PREDICATES, DM_CHILDREN, DM_CENTROIDS, and DM_HISTOGRAMS. It also includes a DM_RULE. See also, DM_CLUSTER Fields.
DM_CLUSTERS	A collection of DM_CLUSTER. Returned by GET_MODEL_DETAILS_KM Function, GET_MODEL_DETAILS_OC Function, and GET_MODEL_DETAILS_EM Function. See also, DM_CLUSTER Fields.
DM_CONDITIONAL	The conditional probability of an attribute in a Naive Bayes model.
DM_CONDITIONALS	A collection of DM_CONDITIONAL. Returned by GET_MODEL_DETAILS_NB Function.
DM_COST_ELEMENT	The actual and predicted values in a cost matrix.
DM_COST_MATRIX	A collection of DM_COST_ELEMENT. Returned by GET_MODEL_COST_MATRIX Function.
DM_EM_COMPONENT	A component of an Expectation Maximization model.
DM_EM_COMPONENT_SET	A collection of DM_EM_COMPONENT. Returned by GET_MODEL_DETAILS_EM_COMP Function.
DM_EM_PROJECTION	A projection of an Expectation Maximization model.
DM_EM_PROJECTION_SET	A collection of DM_EM_PROJECTION. Returned by GET_MODEL_DETAILS_EM_PROJ Function.
DM_GLM_COEFF	The coefficient and associated statistics of an attribute in a Generalized Linear Model.
DM_GLM_COEFF_SET	A collection of DM_GLM_COEFF. Returned by GET_MODEL_DETAILS_GLM Function.
DM_HISTOGRAM_BIN	A histogram associated with a cluster.

Table 6-43 (Cont.) DBMS_DATA_MINING Summary of Deprecated Datatypes

Datatype	Description
DM_HISTOGRAMS	A collection of DM_HISTOGRAM_BIN. A member of DM_CLUSTER. See also, DM_CLUSTER Fields.
DM_ITEM	An item in an association rule.
DM_ITEMS	A collection of DM_ITEM.
DM_ITEMSET	A collection of DM_ITEMS.
DM_ITEMSETS	A collection of DM_ITEMSET. Returned by GET_FREQUENT_ITEMSETS Function.
DM_MODEL_GLOBAL_DETAIL	High-level statistics about a model.
DM_MODEL_GLOBAL_DETAILS	A collection of DM_MODEL_GLOBAL_DETAIL. Returned by GET_MODEL_DETAILS_GLOBAL Function.
DM_NB_DETAIL	Information about an attribute in a Naive Bayes model.
DM_NB_DETAILS	A collection of DM_DB_DETAIL. Returned by GET_MODEL_DETAILS_NB Function.
DM_NMF_ATTRIBUTE	An attribute in a feature of a Non-Negative Matrix Factorization model.
DM_NMF_ATTRIBUTE_SET	A collection of DM_NMF_ATTRIBUTE. A member of DM_NMF_FEATURE.
DM_NMF_FEATURE	A feature in a Non-Negative Matrix Factorization model.
DM_NMF_FEATURE_SET	A collection of DM_NMF_FEATURE. Returned by GET_MODEL_DETAILS_NMF Function.
DM_PREDICATE	Antecedent and consequent in a rule.
DM_PREDICATES	A collection of DM_PREDICATE. A member of DM_RULE and DM_CLUSTER. Predicates are returned by GET_ASSOCIATION_RULES Function, GET_MODEL_DETAILS_EM Function, GET_MODEL_DETAILS_KM Function, and GET_MODEL_DETAILS_OC Function. See also, DM_CLUSTER Fields.
DM_RANKED_ATTRIBUTE	An attribute ranked by its importance in an Attribute Importance model.
DM_RANKED_ATTRIBUTES	A collection of DM_RANKED_ATTRIBUTE. Returned by GET_MODEL_DETAILS_AI Function.
DM_RULE	A rule that defines a conditional relationship. The rule can be one of the association rules returned by GET_ASSOCIATION_RULES Function, or it can be a rule associated with a cluster in the collection of clusters returned by GET_MODEL_DETAILS_KM Function and GET_MODEL_DETAILS_OC Function. See also, DM_CLUSTER Fields.
DM_RULES	A collection of DM_RULE. Returned by GET_ASSOCIATION_RULES Function. See also, DM_CLUSTER Fields.
DM_SVD_MATRIX	A factorized matrix S, V, or U returned by a Singular Value Decomposition model.

Table 6-43 (Cont.) DBMS_DATA_MINING Summary of Deprecated Datatypes

Datatype	Description
DM_SVD_MATRIX_SET	A collection of DM_SVD_MATRIX. Returned by GET_MODEL_DETAILS_SVD Function.
DM_SVM_ATTRIBUTE	The name, value, and coefficient of an attribute in a Support Vector Machine model.
DM_SVM_ATTRIBUTE_SET	A collection of DM_SVM_ATTRIBUTE. Returned by GET_MODEL_DETAILS_SVM Function. Also a member of DM_SVM_LINEAR_COEFF.
DM_SVM_LINEAR_COEFF	The linear coefficient of each attribute in a Support Vector Machine model.
DM_SVM_LINEAR_COEFF_SET	A collection of DM_SVM_LINEAR_COEFF. Returned by GET_MODEL_DETAILS_SVM Function for an SVM model built using the linear kernel.
DM_TRANSFORM	The transformation and reverse transformation expressions for an attribute.
DM_TRANSFORMS	A collection of DM_TRANSFORM. Returned by GET_MODEL_TRANSFORMATIONS Function.

Return Values for Clustering Algorithms

The table contains description of DM_CLUSTER return value columns, nested table columns, and rows.

Table 6-44 DM_CLUSTER Return Values for Clustering Algorithms

Return Value	Description
DM_CLUSTERS	A set of rows of type DM_CLUSTER. The rows have the following columns: (id NUMBER, cluster_id VARCHAR2(4000), record_count NUMBER, parent NUMBER, tree_level NUMBER, dispersion NUMBER, split_predicate DM_PREDICATES, child DM_CHILDREN, centroid DM_CENTROIDS, histogram DM_HISTOGRAMS, rule DM_RULE)
DM_PREDICATE	The antecedent and consequent columns each return nested tables of type DM_PREDICATES. The rows, of type DM_PREDICATE, have the following columns: (attribute_name VARCHAR2(4000), attribute_subname VARCHAR2(4000), conditional_operator CHAR(2)/*=,<>,<,>,<=,>=*/, attribute_num_value NUMBER, attribute_str_value VARCHAR2(4000), attribute_support NUMBER, attribute_confidence NUMBER)

DM_CLUSTER Fields

The following table describes DM_CLUSTER fields.

Table 6-45 DM_CLUSTER Fields

Column Name	Description
id	Cluster identifier
cluster_id	The ID of a cluster in the model
record_count	Specifies the number of records
parent	Parent ID
tree_level	Specifies the number of splits from the root
dispersion	A measure used to quantify whether a set of observed occurrences are dispersed compared to a standard statistical model.
split_predicate	The split_predicate column of DM_CLUSTER returns a nested table of type DM_PREDICATES. Each row, of type DM_PREDICATE, has the following columns: (attribute_name VARCHAR2(4000), attribute_subname VARCHAR2(4000), conditional_operator CHAR(2) / *=,<>,<,>,<=,>=*/, attribute_num_value NUMBER, attribute_str_value VARCHAR2(4000), attribute_support NUMBER, attribute_confidence NUMBER) Note: The Expectation Maximization algorithm uses all the fields except dispersion and split_predicate.
child	The child column of DM_CLUSTER returns a nested table of type DM_CHILDREN. The rows, of type DM_CHILD, have a single column of type NUMBER, which contains the identifiers of each child.
centroid	The centroid column of DM_CLUSTER returns a nested table of type DM_CENTROIDS. The rows, of type DM_CENTROID, have the following columns: (attribute_name VARCHAR2(4000), attribute_subname VARCHAR2(4000), mean NUMBER, mode_value VARCHAR2(4000), variance NUMBER)

Table 6-45 (Cont.) DM_CLUSTER Fields

Column Name	Description

`histogram`	The `histogram` column of `DM_CLUSTER` returns a nested table of type `DM_HISTOGRAMS`. The rows, of type `DM_HISTOGRAM_BIN`, have the following columns:
	`(attribute_name    VARCHAR2(4000),`
	` attribute_subname  VARCHAR2(4000),`
	` bin_id             NUMBER,`
	` lower_bound        NUMBER,`
	` upper_bound        NUMBER,`
	` label              VARCHAR2(4000),`
	` count              NUMBER)`

`rule`	The `rule` column of `DM_CLUSTER` returns a single row of type `DM_RULE`. The columns are:
	`(rule_id            INTEGER,`
	` antecedent         DM_PREDICATES,`
	` consequent         DM_PREDICATES,`
	` rule_support       NUMBER,`
	` rule_confidence    NUMBER,`
	` rule_lift          NUMBER,`
	` antecedent_support NUMBER,`
	` consequent_support NUMBER,`
	` number_of_items    INTEGER)`

Usage Notes

- The table function pipes out rows of type `DM_CLUSTER`. For information on Oracle Machine Learning for SQL data types and piped output from table functions, see "Data Types".
- For descriptions of predicates (`DM_PREDICATE`) and rules (`DM_RULE`), see GET_ASSOCIATION_RULES Function.

Summary of DBMS_DATA_MINING Subprograms

This table summarizes the subprograms included in the `DBMS_DATA_MINING` package.

The `GET_*` interfaces are replaced by model views. Oracle recommends that users leverage model detail views instead. For more information, refer to Model Detail Views in *Oracle Machine Learning for SQL User’s Guide* and Static Data Dictionary Views: ALL_ALL_TABLES to ALL_OUTLINES in *Oracle Database Reference*.

Table 6-46 DBMS_DATA_MINING Package Subprograms

Subprogram	Purpose
ADD_COST_MATRIX Procedure	Adds a cost matrix to a classification model
ADD_PARTITION Procedure	Adds single or multiple partitions in an existing partition model
ALTER_REVERSE_EXPRESSION Procedure	Changes the reverse transformation expression to an expression that you specify
APPLY Procedure	Applies a model to a data set (scores the data)

Table 6-46 (Cont.) DBMS_DATA_MINING Package Subprograms

Subprogram	Purpose
COMPUTE_CONFUSION_MATRIX Procedure	Computes the confusion matrix for a classification model
COMPUTE_CONFUSION_MATRIX_PART Procedure	Computes the evaluation matrix for partitioned models
COMPUTE_LIFT Procedure	Computes lift for a classification model
COMPUTE_LIFT_PART Procedure	Computers lift for partitioned models
COMPUTE_ROC Procedure	Computes Receiver Operating Characteristic (ROC) for a classification model
COMPUTE_ROC_PART Procedure	Computes Receiver Operating Characteristic (ROC) for a partitioned model
CREATE_MODEL Procedure	Creates a model
CREATE_MODEL2 Procedure	Creates a model without extra persistent stages
Create Model Using Registration Information	Fetches setting information from JSON object
DROP_ALGORITHM Procedure	Drops the registered algorithm information.
DROP_PARTITION Procedure	Drops a single partition
DROP_MODEL Procedure	Drops a model
EXPORT_MODEL Procedure	Exports a model to a dump file
EXPORT_SERMODEL Procedure	Exports a model in a serialized format
FETCH_JSON_SCHEMA Procedure	Fetches and reads JSON schema from `all_mining_algorithms` view
GET_MODEL_COST_MATRIX Function	Returns the cost matrix for a model
IMPORT_MODEL Procedure	Imports a model into a user schema
IMPORT_ONNX_MODEL Procedure	Imports an ONNX model into the Database
IMPORT_SERMODEL Procedure	Imports a serialized model back into the database
JSON Schema for R Extensible Algorithm	Displays flexibility in creating JSON schema for R Extensible
REGISTER_ALGORITHM Procedure	Registers a new algorithm
RANK_APPLY Procedure	Ranks the predictions from the `APPLY` results for a classification model
REMOVE_COST_MATRIX Procedure	Removes a cost matrix from a model
RENAME_MODEL Procedure	Renames a model

Deprecated GET_MODEL_DETAILS

Starting from Oracle Database 12*c* Release 2, the following `GET_MODEL_DETAILS` are deprecated:

Table 6-47 Deprecated `GET_MODEL_DETAILS` Functions

Subprogram	Purpose
GET_ASSOCIATION_RULES Function	Returns the rules from an association model

Table 6-47 (Cont.) Deprecated `GET_MODEL_DETAILS` Functions

Subprogram	Purpose
GET_FREQUENT_ITEMSETS Function	Returns the frequent itemsets for an association model
GET_MODEL_DETAILS_AI Function	Returns details about an attribute importance model
GET_MODEL_DETAILS_EM Function	Returns details about an Expectation Maximization model
GET_MODEL_DETAILS_EM_COMP Function	Returns details about the parameters of an Expectation Maximization model
GET_MODEL_DETAILS_EM_PROJ Function	Returns details about the projects of an Expectation Maximization model
GET_MODEL_DETAILS_GLM Function	Returns details about a Generalized Linear Model model
GET_MODEL_DETAILS_GLOBAL Function	Returns high-level statistics about a model
GET_MODEL_DETAILS_KM Function	Returns details about a *k*-Means model
GET_MODEL_DETAILS_NB Function	Returns details about a Naive Bayes model
GET_MODEL_DETAILS_NMF Function	Returns details about a Non-Negative Matrix Factorization model
GET_MODEL_DETAILS_OC Function	Returns details about an O-Cluster model
GET_MODEL_SETTINGS Function	Returns the settings used to build the given model This function is replaced with `USER/ALL/DBA_MINING_MODEL_SETTINGS`
GET_MODEL_SIGNATURE Function	Returns the list of columns from the build input table This function is replaced with `USER/ALL/DBA_MINING_MODEL_ATTRIBUTES`
GET_MODEL_DETAILS_SVD Function	Returns details about a Singular Value Decomposition model
GET_MODEL_DETAILS_SVM Function	Returns details about a Support Vector Machine model with a linear kernel
GET_MODEL_TRANSFORMATIONS Function	Returns the transformations embedded in a model This function is replaced with `USER/ALL/DBA_MINING_MODEL_XFORMS`
GET_MODEL_DETAILS_XML Function	Returns details about a Decision Tree model
GET_TRANSFORM_LIST Procedure	Converts between two different transformation specification formats

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*
- *Oracle Database Reference*

ADD_COST_MATRIX Procedure

The ADD_COST_MATRIX procedure associates a cost matrix table with a classification model. The cost matrix biases the model by assigning costs or benefits to specific model outcomes.

The cost matrix is stored with the model and taken into account when the model is scored.

You can also specify a cost matrix inline when you invoke an Oracle Machine Learning for SQL function for scoring. To view the scoring matrix for a model, query the DM$VC prefixed model view. Refer to Model Detail View for Classification Algorithm.

To obtain the default scoring matrix for a model, query the DM$VC prefixed model view. To remove the default scoring matrix from a model, use the REMOVE_COST_MATRIX procedure. See REMOVE_COST_MATRIX Procedure.

> **See Also:**
>
> - "Biasing a Classification Model" in *Oracle Machine Learning for SQL Concepts* for more information about costs
> - *Oracle Database SQL Language Reference* for syntax of inline cost matrix
> - Specifying Costs in *Oracle Machine Learning for SQL User’s Guide*

Syntax

```
DBMS_DATA_MINING.ADD_COST_MATRIX (
       model_name                 IN VARCHAR2,
       cost_matrix_table_name     IN VARCHAR2,
       cost_matrix_schema_name    IN VARCHAR2 DEFAULT NULL);
       partition_name             IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-48 ADD_COST_MATRIX Procedure Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is assumed.
cost_matrix_table_name	Name of the cost matrix table (described in Table 6-49).
cost_matrix_schema_name	Schema of the cost matrix table. If no schema is specified, then the current schema is used.
partition_name	Name of the partition in a partitioned model

Usage Notes

1. If the model is not in your schema, then ADD_COST_MATRIX requires the ALTER ANY MINING MODEL system privilege or the ALTER object privilege for the machine learning model.
2. The cost matrix table must have the columns shown in Table 6-49.

Table 6-49 Required Columns in a Cost Matrix Table

Column Name	Data Type
ACTUAL_TARGET_VALUE	Valid target data type
PREDICTED_TARGET_VALUE	Valid target data type
COST	NUMBER, FLOAT, BINARY_DOUBLE, or BINARY_FLOAT

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for valid target data types

3. The types of the actual and predicted target values must be the same as the type of the model target. For example, if the target of the model is BINARY_DOUBLE, then the actual and predicted values must be BINARY_DOUBLE. If the actual and predicted values are CHAR or VARCHAR, then ADD_COST_MATRIX treats them as VARCHAR2 internally.

   If the types do not match, or if the actual or predicted value is not a valid target value, then the ADD_COST_MATRIX procedure raises an error.

> **Note:**
>
> If a reverse transformation is associated with the target, then the actual and predicted values must be consistent with the target after the reverse transformation has been applied.
>
> See “Reverse Transformations and Model Transparency” under the “About Transformation Lists” section in DBMS_DATA_MINING_TRANSFORM Operational Notes for more information.

4. Since a benefit can be viewed as a negative cost, you can specify a benefit for a given outcome by providing a negative number in the costs column of the cost matrix table.

5. All classification algorithms can use a cost matrix for scoring. The Decision Tree algorithm can also use a cost matrix at build time. If you want to build a Decision Tree model with a cost matrix, specify the cost matrix table name in the CLAS_COST_TABLE_NAME setting in the settings table for the model. See Table 6-11.

   The cost matrix used to create a Decision Tree model becomes the default scoring matrix for the model. If you want to specify different costs for scoring, use the REMOVE_COST_MATRIX procedure to remove the cost matrix and the ADD_COST_MATRIX procedure to add a new one.

6. Scoring on a partitioned model is partition-specific. Scoring cost matrices can be added to or removed from an individual partition in a partitioned model. If PARTITION_NAME is NOT NULL, then the model must be a partitioned model. The COST_MATRIX is added to that partition of the partitioned model.

   If the PARTITION_NAME is NULL, but the model is a partitioned model, then the COST_MATRIX table is added to every partition in the model.

Example

This example creates a cost matrix table called COSTS_NB and adds it to a Naive Bayes model called NB_SH_CLAS_SAMPLE. The model has a binary target: 1 means that the customer responds to a promotion; 0 means that the customer does not respond. The cost matrix assigns a cost of .25 to misclassifications of customers who do not respond and a cost of .75 to misclassifications of customers who do respond. This means that it is three times more costly to misclassify responders than it is to misclassify non-responders.

```
CREATE TABLE costs_nb (
  actual_target_value           NUMBER,
  predicted_target_value        NUMBER,
  cost                          NUMBER);
INSERT INTO costs_nb values (0, 0, 0);
INSERT INTO costs_nb values (0, 1, .25);
INSERT INTO costs_nb values (1, 0, .75);
INSERT INTO costs_nb values (1, 1, 0);
COMMIT;

EXEC dbms_data_mining.add_cost_matrix('nb_sh_clas_sample', 'costs_nb');

SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
   FROM mining_data_apply_v
   WHERE PREDICTION(nb_sh_clas_sample COST MODEL
      USING cust_marital_status, education, household_size) = 1
   GROUP BY cust_gender
   ORDER BY cust_gender;

C        CNT    AVG_AGE
- ---------- ----------
F         72         39
M        555         44
```

ADD_PARTITION Procedure

ADD_PARTITION procedure supports a single or multiple partition addition to an existing partitioned model.

The ADD_PARTITION procedure derives build settings and user-defined expressions from the existing model. The target column must exist in the input data query when adding partitions to a supervised model.

Syntax

```
DBMS_DATA_MINING.ADD_PARTITION (
      model_name                 IN VARCHAR2,
      data_query                 IN CLOB,
      add_options                IN VARCHAR2 DEFAULT ERROR);
```

Parameters

Table 6-50 ADD_PARTITION Procedure Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.

Table 6-50 (Cont.) ADD_PARTITION Procedure Parameters

Parameter	Description
data_query	An arbitrary SQL statement that provides data to the model build. The user must have privilege to evaluate this query.
add_options	Allows users to control the conditional behavior of ADD for cases where rows in the input dataset conflict with existing partitions in the model. The following are the possible values: • REPLACE: Replaces the existing partition for which the conflicting keys are found. • ERROR: Terminates the ADD operation without adding any partitions. • IGNORE: Eliminates the rows having the conflicting keys. **Note:** For better performance, Oracle recommends using DROP_PARTITION followed by the ADD_PARTITION instead of using the REPLACE option.

ALTER_REVERSE_EXPRESSION Procedure

This procedure replaces a reverse transformation expression with an expression that you specify. If the attribute does not have a reverse expression, the procedure creates one from the specified expression.

You can also use this procedure to customize the output of clustering, feature extraction, and anomaly detection models.

Syntax

```
DBMS_DATA_MINING.ALTER_REVERSE_EXPRESSION (
         model_name                VARCHAR2,
         expression                CLOB,
         attribute_name            VARCHAR2  DEFAULT NULL,
         attribute_subname         VARCHAR2  DEFAULT NULL);
```

Parameters

Table 6-51 ALTER_REVERSE_EXPRESSION Procedure Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, your own schema is used.
expression	An expression to replace the reverse transformation associated with the attribute.
attribute_name	Name of the attribute. Specify NULL if you wish to apply *expression* to a cluster, feature, or One-Class SVM prediction.
attribute_subname	Name of the nested attribute if *attribute_name* is a nested column, otherwise NULL.

Usage Notes

1. For purposes of model transparency, Oracle Machine Learning for SQL provides reverse transformations for transformations that are embedded in a model. Reverse transformations are applied to the attributes returned in model detail views and to the scored target of predictive models.

   See Also:

   - "About Transformation Lists" under DBMS_DATA_MINING_TRANSFORM Operational Notes
   - Model Detail Views in *Oracle Machine Learning for SQL User's Guide*

2. If you alter the reverse transformation for the target of a model that has a cost matrix, you must specify a transformation expression that has the same type as the actual and predicted values in the cost matrix. Also, the reverse transformation that you specify must result in values that are present in the cost matrix.

   See Also:

   "ADD_COST_MATRIX Procedure" and *Oracle Machine Learning for SQL Concepts* for information about cost matrixes.

3. To prevent reverse transformation of an attribute, you can specify `NULL` for *`expression`*.

4. The reverse transformation expression can contain a reference to a PL/SQL function that returns a valid Oracle data type. For example, you could define a function like the following for a categorical attribute named `blood_pressure` that has values 'Low', 'Medium' and 'High'.

   ```
   CREATE OR REPLACE FUNCTION numx(c char) RETURN NUMBER IS
     BEGIN
       CASE c WHEN ''Low'' THEN RETURN 1;
              WHEN ''Medium'' THEN RETURN 2;
              WHEN ''High'' THEN RETURN 3;
              ELSE RETURN null;
       END CASE;
     END numx;
   ```

   Then you could invoke `ALTER_REVERSE_EXPRESION` for `blood_pressure` as follows.

   ```
   EXEC dbms_data_mining.alter_reverse_expression(
                '<model_name>', 'NUMX(blood_pressure)', 'blood_pressure');
   ```

5. You can use `ALTER_REVERSE_EXPRESSION` to label clusters produced by clustering models and features produced by feature extraction.

   You can use `ALTER_REVERSE_EXPRESSION` to replace the zeros and ones returned by anomaly-detection models. By default, anomaly-detection models label anomalous records with 0 and all other records with 1.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for information about anomaly detection

Examples

1. In this example, the target (`affinity_card`) of the model `CLASS_MODEL` is manipulated internally as `yes` or `no` instead of `1` or `0` but returned as `1`s and `0`s when scored. The `ALTER_REVERSE_EXPRESSION` procedure causes the target values to be returned as `TRUE` or `FALSE`.

```
DECLARE
        v_xlst dbms_data_mining_transform.TRANSFORM_LIST;
  BEGIN
    dbms_data_mining_transform.SET_TRANSFORM(v_xlst,
          'affinity_card', NULL,
          'decode(affinity_card, 1, ''yes'', ''no'')',
          'decode(affinity_card, ''yes'', 1, 0)');
    dbms_data_mining.CREATE_MODEL(
      model_name             => 'CLASS_MODEL',
      mining_function        => dbms_data_mining.classification,
      data_table_name        => 'mining_data_build',
      case_id_column_name    => 'cust_id',
      target_column_name     => 'affinity_card',
      settings_table_name    => NULL,
      data_schema_name       => 'oml_user',
      settings_schema_name   => NULL,
      xform_list             => v_xlst );
  END;
/
SELECT cust_income_level, occupation,
           PREDICTION(CLASS_MODEL USING *) predict_response
      FROM mining_data_test WHERE age = 60 AND cust_gender IN 'M'
      ORDER BY cust_income_level;

CUST_INCOME_LEVEL              OCCUPATION            PREDICT_RESPONSE
------------------------------ --------------------- --------------------
A: Below 30,000                Transp.                                  1
E: 90,000 - 109,999            Transp.                                  1
E: 90,000 - 109,999            Sales                                    1
G: 130,000 - 149,999           Handler                                  0
G: 130,000 - 149,999           Crafts                                   0
H: 150,000 - 169,999           Prof.                                    1
J: 190,000 - 249,999           Prof.                                    1
J: 190,000 - 249,999           Sales                                    1

BEGIN
  dbms_data_mining.ALTER_REVERSE_EXPRESSION (
     model_name      => 'CLASS_MODEL',
     expression      => 'decode(affinity_card, ''yes'', ''TRUE'', ''FALSE'')',
     attribute_name  => 'affinity_card');
END;
/
column predict_response on
column predict_response format a20
SELECT cust_income_level, occupation,
             PREDICTION(CLASS_MODEL USING *) predict_response
      FROM mining_data_test WHERE age = 60 AND cust_gender IN 'M'
      ORDER BY cust_income_level;

CUST_INCOME_LEVEL              OCCUPATION            PREDICT_RESPONSE
------------------------------ --------------------- --------------------
A: Below 30,000                Transp.               TRUE
E: 90,000 - 109,999            Transp.               TRUE
E: 90,000 - 109,999            Sales                 TRUE
G: 130,000 - 149,999           Handler               FALSE
G: 130,000 - 149,999           Crafts                FALSE
H: 150,000 - 169,999           Prof.                 TRUE
J: 190,000 - 249,999           Prof.                 TRUE
J: 190,000 - 249,999           Sales                 TRUE
```

2. This example specifies labels for the clusters that result from the `sh_clus` model. The labels consist of the word "Cluster" and the internal numeric identifier for the cluster.

```
BEGIN
  dbms_data_mining.ALTER_REVERSE_EXPRESSION( 'sh_clus', '''Cluster ''||value');
END;
/

SELECT cust_id, cluster_id(sh_clus using *) cluster_id
   FROM sh_aprep_num
       WHERE cust_id < 100011
       ORDER by cust_id;

CUST_ID CLUSTER_ID
------- ------------------------------------------------
 100001 Cluster 18
 100002 Cluster 14
 100003 Cluster 14
 100004 Cluster 18
 100005 Cluster 19
 100006 Cluster 7
 100007 Cluster 18
 100008 Cluster 14
 100009 Cluster 8
 100010 Cluster 8
```

APPLY Procedure

The `APPLY` procedure applies a machine learning model to the data of interest, and generates the results in a table. The `APPLY` procedure is also referred to as **scoring**.

For predictive machine learning functions, the `APPLY` procedure generates predictions in a target column. For descriptive machine learning functions such as Clustering, the `APPLY` process assigns each case to a cluster with a probability.

In Oracle Machine Learning for SQL, the `APPLY` procedure is not applicable to Association models and Attribute Importance models.

Note:

Scoring can also be performed directly in SQL using the OML4SQL functions. See

- Oracle Machine Learning for SQL Functions in *Oracle Database SQL Language Reference*
- Scoring and Deployment in *Oracle Machine Learning for SQL User's Guide*

Syntax

```
DBMS_DATA_MINING.APPLY (
      model_name           IN VARCHAR2,
      data_table_name      IN VARCHAR2,
      case_id_column_name  IN VARCHAR2,
      result_table_name    IN VARCHAR2,
      data_schema_name     IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-52 ***APPLY Procedure Parameters***

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`data_table_name`	Name of table or view containing the data to be scored
`case_id_column_name`	Name of the case identifier column
`result_table_name`	Name of the table in which to store apply results
`data_schema_name`	Name of the schema containing the data to be scored

Usage Notes

1. The data provided for `APPLY` must undergo the same preprocessing as the data used to create and test the model. When you use Automatic Data Preparation, the preprocessing required by the algorithm is handled for you by the model: both at build time and apply time. (See "Automatic Data Preparation".)

2. `APPLY` creates a table in the user's schema to hold the results. The columns are algorithm-specific.

   The columns in the results table are listed in Table 6-53 through Table 6-57. The case ID column name in the results table will match the case ID column name provided by you. The type of the incoming case ID column is also preserved in `APPLY` output.

   > **Note:**
   >
   > Make sure that the case ID column does not have the same name as one of the columns that will be created by `APPLY`. For example, when applying a Classification model, the case ID in the scoring data must not be `PREDICTION` or `PROBABILITY` (See Table 6-53).

3. The data type for the `PREDICTION`, `CLUSTER_ID`, and `FEATURE_ID` output columns is influenced by any reverse expression that is embedded in the model by the user. If the user does not provide a reverse expression that alters the scored value type, then the types will conform to the descriptions in the following tables. See "ALTER_REVERSE_EXPRESSION Procedure".

4. If the model is partitioned, the `result_table_name` can contain results from different partitions depending on the data from the input data table. An additional column called `PARTITION_NAME` is added to the result table indicating the partition name that is associated with each row.

For a non-partitioned model, the behavior does not change.

Classification

The results table for Classification has the columns described in Table 6-53. If the target of the model is categorical, the `PREDICTION` column will have a `VARCHAR2` data type. If the target has a binary type, the `PREDICTION` column will have the binary type of the target.

Table 6-53   APPLY Results Table for Classification

Column Name	Data type
`Case ID column name`	Type of the case ID
`PREDICTION`	Type of the target
`PROBABILITY`	`BINARY_DOUBLE`

Anomaly Detection

The results table for Anomaly Detection has the columns described in Table 6-54.

Table 6-54   APPLY Results Table for Anomaly Detection

Column Name	Data Type
`Case ID column name`	Type of the case ID
`PREDICTION`	`NUMBER`
`PROBABILITY`	`BINARY_DOUBLE`

Regression

The results table for Regression has the columns described in APPLY Procedure.

Table 6-55   APPLY Results Table for Regression

Column Name	Data Type
`Case ID column name`	Type of the case ID
`PREDICTION`	Type of the target

Clustering

Clustering is an unsupervised machine learning function, and hence there are no targets. The results of an `APPLY` procedure contain simply the cluster identifier corresponding to a case, and the associated probability. The results table has the columns described in Table 6-56.

Table 6-56   APPLY Results Table for Clustering

Column Name	Data Type
`Case ID column name`	Type of the case ID
`CLUSTER_ID`	`NUMBER`
`PROBABILITY`	`BINARY_DOUBLE`

Feature Extraction

Feature Extraction is also an unsupervised machine learning function, hence there are no targets. The results of an APPLY procedure will contain simply the feature identifier corresponding to a case, and the associated match quality. The results table has the columns described in Table 6-57.

Table 6-57   APPLY Results Table for Feature Extraction

Column Name	Data Type
Case ID column name	Type of the case ID
FEATURE_ID	NUMBER
MATCH_QUALITY	BINARY_DOUBLE

Examples

This example applies the GLM Regression model GLMR_SH_REGR_SAMPLE to the data in the MINING_DATA_APPLY_V view. The APPLY results are output of the table REGRESSION_APPLY_RESULT.

```
SQL> BEGIN
       DBMS_DATA_MINING.APPLY (
       model_name     => 'glmr_sh_regr_sample',
       data_table_name     => 'mining_data_apply_v',
       case_id_column_name => 'cust_id',
       result_table_name   => 'regression_apply_result');
    END;
    /

SQL> SELECT * FROM regression_apply_result WHERE cust_id >  101485;

   CUST_ID PREDICTION
---------- ----------
    101486 22.8048824
    101487 25.0261101
    101488 48.6146619
    101489   51.82595
    101490 22.6220714
    101491 61.3856816
    101492 24.1400748
    101493  58.034631
    101494 45.7253149
    101495 26.9763318
    101496 48.1433425
    101497 32.0573434
    101498 49.8965531
    101499  56.270656
    101500 21.1153047
```

COMPUTE_CONFUSION_MATRIX Procedure

This procedure computes a confusion matrix, stores it in a table in the user's schema, and returns the model accuracy.

A confusion matrix is a test metric for classification models. It compares the predictions generated by the model with the actual target values in a set of test data. The confusion matrix lists the number of times each class was correctly predicted and the number of times it was predicted to be one of the other classes.

COMPUTE_CONFUSION_MATRIX accepts three input streams:

- The predictions generated on the test data. The information is passed in three columns:
  - Case ID column
  - Prediction column
  - Scoring criterion column containing either probabilities or costs
- The known target values in the test data. The information is passed in two columns:
  - Case ID column
  - Target column containing the known target values
- (Optional) A cost matrix table with predefined columns. See the Usage Notes for the column requirements.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more details about confusion matrixes and other test metrics for classification
>
> "COMPUTE_LIFT Procedure"
>
> "COMPUTE_ROC Procedure"

Syntax

```
DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX (
      accuracy                     OUT NUMBER,
      apply_result_table_name      IN  VARCHAR2,
      target_table_name            IN  VARCHAR2,
      case_id_column_name          IN  VARCHAR2,
      target_column_name           IN  VARCHAR2,
      confusion_matrix_table_name  IN  VARCHAR2,
      score_column_name            IN  VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name  IN  VARCHAR2 DEFAULT 'PROBABILITY',
      cost_matrix_table_name       IN  VARCHAR2 DEFAULT NULL,
      apply_result_schema_name     IN  VARCHAR2 DEFAULT NULL,
      target_schema_name           IN  VARCHAR2 DEFAULT NULL,
      cost_matrix_schema_name      IN  VARCHAR2 DEFAULT NULL,
      score_criterion_type         IN  VARCHAR2 DEFAULT 'PROBABILITY');
```

Parameters

Table 6-58 COMPUTE_CONFUSION_MATRIX Procedure Parameters

Parameter	Description
`accuracy`	Output parameter containing the overall percentage accuracy of the predictions.
`apply_result_table_name`	Table containing the predictions.
`target_table_name`	Table containing the known target values from the test data.
`case_id_column_name`	Case ID column in the apply results table. Must match the case identifier in the targets table.
`target_column_name`	Target column in the targets table. Contains the known target values from the test data.

Table 6-58 (Cont.) COMPUTE_CONFUSION_MATRIX Procedure Parameters

Parameter	Description
`confusion_matrix_table_name`	Table containing the confusion matrix. The table will be created by the procedure in the user's schema. The columns in the confusion matrix table are described in the Usage Notes.
`score_column_name`	Column containing the predictions in the apply results table. The default column name is `PREDICTION`, which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`score_criterion_column_name`	Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, the class with the lowest cost is predicted. The `score_criterion_type` parameter indicates whether probabilities or costs will be used for scoring. The default column name is '`PROBABILITY`', which is the default name created by the `APPLY` procedure (See "APPLY Procedure"). See the Usage Notes for additional information.
`cost_matrix_table_name`	(Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the `score_criterion_type` parameter is set to `'COSTS'`, the costs in this table will be used as the scoring criteria. The columns in a cost matrix table are described in the Usage Notes.
`apply_result_schema_name`	Schema of the apply results table. If null, the user's schema is assumed.
`target_schema_name`	Schema of the table containing the known targets. If null, the user's schema is assumed.
`cost_matrix_schema_name`	Schema of the cost matrix table, if one is provided. If null, the user's schema is assumed.
`score_criterion_type`	Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the `score_criterion_column_name` parameter. The default value of `score_criterion_type` is `'PROBABILITY'`. To use costs as the scoring criterion, specify `'COST'`. If `score_criterion_type` is set to '`COST`' but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring. See the Usage Notes and the Examples.

Usage Notes

- The predictive information you pass to `COMPUTE_CONFUSION_MATRIX` may be generated using SQL `PREDICTION` functions, the `DBMS_DATA_MINING.APPLY` procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the confusion matrix.

- Instead of passing a cost matrix to COMPUTE_CONFUSION_MATRIX, you can use a scoring cost matrix associated with the model. A scoring cost matrix can be embedded in the model or it can be defined dynamically when the model is applied. To use a scoring cost matrix, invoke the SQL PREDICTION_COST function to populate the score criterion column.
- The predictions that you pass to COMPUTE_CONFUSION_MATRIX are in a table or view specified in apply_result_table_name.

```
CREATE TABLE apply_result_table_name AS (
            case_id_column_name            VARCHAR2,
            score_column_name             VARCHAR2,
            score_criterion_column_name    VARCHAR2);
```

- A cost matrix must have the columns described in Table 6-59.

Table 6-59   Columns in a Cost Matrix

Column Name	Data Type
actual_target_value	Type of the target column in the build data
predicted_target_value	Type of the predicted target in the test data. The type of the predicted target must be the same as the type of the actual target unless the predicted target has an associated reverse transformation.
cost	BINARY_DOUBLE

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for valid target data types
>
> *Oracle Machine Learning for SQL Concepts* for more information about cost matrixes

- The confusion matrix created by COMPUTE_CONFUSION_MATRIX has the columns described in Table 6-60.

Table 6-60   Columns in a Confusion Matrix

Column Name	Data Type
actual_target_value	Type of the target column in the build data
predicted_target_value	Type of the predicted target in the test data. The type of the predicted target is the same as the type of the actual target unless the predicted target has an associated reverse transformation.
value	BINARY_DOUBLE

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more information about confusion matrixes

Examples

These examples use the Naive Bayes model `nb_sh_clas_sample`.

Compute a Confusion Matrix Based on Probabilities

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
       SELECT cust_id,
              PREDICTION(nb_sh_clas_sample USING *) prediction,
              PREDICTION_PROBABILITY(nb_sh_clas_sample USING *) probability
       FROM mining_data_test_v;
```

Using probabilities as the scoring criterion, you can compute the confusion matrix as follows.

```
DECLARE
   v_accuracy    NUMBER;
      BEGIN
        DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX (
                   accuracy                     => v_accuracy,
                   apply_result_table_name      => 'nb_apply_results',
                   target_table_name            => 'mining_data_test_v',
                   case_id_column_name          => 'cust_id',
                   target_column_name           => 'affinity_card',
                   confusion_matrix_table_name  => 'nb_confusion_matrix',
                   score_column_name            => 'PREDICTION',
                   score_criterion_column_name  => 'PROBABILITY'
                   cost_matrix_table_name       =>  null,
                   apply_result_schema_name     =>  null,
                   target_schema_name           =>  null,
                   cost_matrix_schema_name      =>  null,
                   score_criterion_type         => 'PROBABILITY');
        DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' || ROUND(v_accuracy,4));
      END;
      /
```

The confusion matrix and model accuracy are shown as follows.

```
 **** MODEL ACCURACY ****: .7847

SQL>SELECT * from nb_confusion_matrix;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         60
                  0                      0        891
                  1                      1        286
                  0                      1        263
```

Compute a Confusion Matrix Based on a Cost Matrix Table

The confusion matrix in the previous example shows a high rate of false positives. For 263 cases, the model predicted 1 when the actual value was 0. You could use a cost matrix to minimize this type of error.

The cost matrix table `nb_cost_matrix` specifies that a false positive is 3 times more costly than a false negative.

```
SQL> SELECT * from nb_cost_matrix;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE       COST
------------------- ---------------------- ----------
```

```
                  0                       0          0
                  0                       1        .75
                  1                       0        .25
                  1                       1          0
```

This statement shows how to generate the predictions using `APPLY`.

```
BEGIN
    DBMS_DATA_MINING.APPLY(
          model_name          => 'nb_sh_clas_sample',
          data_table_name     => 'mining_data_test_v',
          case_id_column_name => 'cust_id',
          result_table_name   => 'nb_apply_results');
 END;
/
```

This statement computes the confusion matrix using the cost matrix table. The score criterion column is named '`PROBABILITY`', which is the name generated by `APPLY`.

```
DECLARE
  v_accuracy    NUMBER;
     BEGIN
       DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX (
                accuracy                     => v_accuracy,
                apply_result_table_name      => 'nb_apply_results',
                target_table_name            => 'mining_data_test_v',
                case_id_column_name          => 'cust_id',
                target_column_name           => 'affinity_card',
                confusion_matrix_table_name  => 'nb_confusion_matrix',
                score_column_name            => 'PREDICTION',
                score_criterion_column_name  => 'PROBABILITY',
                cost_matrix_table_name       => 'nb_cost_matrix',
                apply_result_schema_name     => null,
                target_schema_name           => null,
                cost_matrix_schema_name      => null,
                score_criterion_type         => 'COST');
       DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' || ROUND(v_accuracy,4));
    END;
    /
```

The resulting confusion matrix shows a decrease in false positives (212 instead of 263).

```
**** MODEL ACCURACY ****: .798

SQL> SELECT * FROM nb_confusion_matrix;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         91
                  0                      0        942
                  1                      1        255
                  0                      1        212
```

Compute a Confusion Matrix Based on Embedded Costs

You can use the `ADD_COST_MATRIX` procedure to embed a cost matrix in a model. The embedded costs can be used instead of probabilities for scoring. This statement adds the previously-defined cost matrix to the model.

```
BEGIN    DBMS_DATA_MINING.ADD_COST_MATRIX ('nb_sh_clas_sample', 'nb_cost_matrix');END;/
```

The following statement applies the model to the test data using the embedded costs and stores the results in a table.

```
CREATE TABLE nb_apply_results AS
         SELECT cust_id,
              PREDICTION(nb_sh_clas_sample COST MODEL USING *) prediction,
              PREDICTION_COST(nb_sh_clas_sample COST MODEL USING *) cost
          FROM mining_data_test_v;
```

You can compute the confusion matrix using the embedded costs.

```
DECLARE
   v_accuracy         NUMBER;
   BEGIN
       DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX (
            accuracy                     => v_accuracy,
            apply_result_table_name      => 'nb_apply_results',
            target_table_name            => 'mining_data_test_v',
            case_id_column_name          => 'cust_id',
            target_column_name           => 'affinity_card',
            confusion_matrix_table_name  => 'nb_confusion_matrix',
            score_column_name            => 'PREDICTION',
            score_criterion_column_name  => 'COST',
            cost_matrix_table_name       => null,
            apply_result_schema_name     => null,
            target_schema_name           => null,
            cost_matrix_schema_name      => null,
            score_criterion_type         => 'COST');
   END;
   /
```

The results are:

```
**** MODEL ACCURACY ****: .798

SQL> SELECT * FROM nb_confusion_matrix;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         91
                  0                      0        942
                  1                      1        255
                  0                      1        212
```

COMPUTE_CONFUSION_MATRIX_PART Procedure

The `COMPUTE_CONFUSION_MATRIX_PART` procedure computes a confusion matrix, stores it in a table in the user's schema, and returns the model accuracy.

`COMPUTE_CONFUSION_MATRIX_PART` provides support to computation of evaluation metrics per-partition for partitioned models. For non-partitioned models, refer to COMPUTE_CONFUSION_MATRIX Procedure.

A confusion matrix is a test metric for classification models. It compares the predictions generated by the model with the actual target values in a set of test data. The confusion matrix lists the number of times each class was correctly predicted and the number of times it was predicted to be one of the other classes.

`COMPUTE_CONFUSION_MATRIX_PART` accepts three input streams:

- The predictions generated on the test data. The information is passed in three columns:
  - Case ID column
  - Prediction column
  - Scoring criterion column containing either probabilities or costs

- The known target values in the test data. The information is passed in two columns:
  - Case ID column
  - Target column containing the known target values
- (Optional) A cost matrix table with predefined columns. See the Usage Notes for the column requirements.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more details about confusion matrixes and other test metrics for classification
>
> "COMPUTE_LIFT_PART Procedure"
>
> "COMPUTE_ROC_PART Procedure"

Syntax

```
DBMS_DATA_MINING.compute_confusion_matrix_part(
      accuracy                    OUT DM_NESTED_NUMERICALS,
      apply_result_table_name     IN  VARCHAR2,
      target_table_name           IN  VARCHAR2,
      case_id_column_name         IN  VARCHAR2,
      target_column_name          IN  VARCHAR2,
      confusion_matrix_table_name IN  VARCHAR2,
      score_column_name           IN  VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name IN  VARCHAR2 DEFAULT 'PROBABILITY',
      score_partition_column_name IN  VARCHAR2 DEFAULT 'PARTITION_NAME',
      cost_matrix_table_name      IN  VARCHAR2 DEFAULT NULL,
      apply_result_schema_name    IN  VARCHAR2 DEFAULT NULL,
      target_schema_name          IN  VARCHAR2 DEFAULT NULL,
      cost_matrix_schema_name     IN  VARCHAR2 DEFAULT NULL,
      score_criterion_type        IN  VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-61 COMPUTE_CONFUSION_MATRIX_PART Procedure Parameters

Parameter	Description
`accuracy`	Output parameter containing the overall percentage accuracy of the predictions The output argument is changed from `NUMBER` to `DM_NESTED_NUMERICALS`
`apply_result_table_name`	Table containing the predictions
`target_table_name`	Table containing the known target values from the test data
`case_id_column_name`	Case ID column in the apply results table. Must match the case identifier in the targets table.
`target_column_name`	Target column in the targets table. Contains the known target values from the test data.
`confusion_matrix_table_name`	Table containing the confusion matrix. The table will be created by the procedure in the user's schema. The columns in the confusion matrix table are described in the Usage Notes.

Table 6-61 (Cont.) COMPUTE_CONFUSION_MATRIX_PART Procedure Parameters

Parameter	Description
`score_column_name`	Column containing the predictions in the apply results table. The default column name is `PREDICTION`, which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`score_criterion_column_name`	Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, then the class with the lowest cost is predicted. The `score_criterion_type` parameter indicates whether probabilities or costs will be used for scoring. The default column name is `PROBABILITY`, which is the default name created by the `APPLY` procedure (See "APPLY Procedure"). See the Usage Notes for additional information.
`score_partition_column_name`	(Optional) Parameter indicating the column which contains the name of the partition. This column slices the input test results such that each partition has independent evaluation matrices computed.
`cost_matrix_table_name`	(Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the `score_criterion_type` parameter is set to `COSTS`, the costs in this table will be used as the scoring criteria. The columns in a cost matrix table are described in the Usage Notes.
`apply_result_schema_name`	Schema of the apply results table. If null, then the user's schema is assumed.
`target_schema_name`	Schema of the table containing the known targets. If null, then the user's schema is assumed.
`cost_matrix_schema_name`	Schema of the cost matrix table, if one is provided. If null, then the user's schema is assumed.
`score_criterion_type`	Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the `score_criterion_column_name` parameter. The default value of `score_criterion_type` is `PROBABILITY`. To use costs as the scoring criterion, specify `COST`. If `score_criterion_type` is set to `COST` but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring. See the Usage Notes and the Examples.

Usage Notes

- The predictive information you pass to `COMPUTE_CONFUSION_MATRIX_PART` may be generated using SQL `PREDICTION` functions, the `DBMS_DATA_MINING.APPLY` procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the confusion matrix.

- Instead of passing a cost matrix to `COMPUTE_CONFUSION_MATRIX_PART`, you can use a scoring cost matrix associated with the model. A scoring cost matrix can be embedded in the model or it can be defined dynamically when the model is applied. To use a scoring cost matrix, invoke the SQL `PREDICTION_COST` function to populate the score criterion column.
- The predictions that you pass to `COMPUTE_CONFUSION_MATRIX_PART` are in a table or view specified in `apply_result_table_name`.

```
CREATE TABLE apply_result_table_name AS (
           case_id_column_name           VARCHAR2,
           score_column_name            VARCHAR2,
           score_criterion_column_name    VARCHAR2);
```

- A cost matrix must have the columns described in Table 6-59.

Table 6-62   Columns in a Cost Matrix

Column Name	Data Type
`actual_target_value`	Type of the target column in the test data
`predicted_target_value`	Type of the predicted target in the test data. The type of the predicted target must be the same as the type of the actual target unless the predicted target has an associated reverse transformation.
`cost`	`BINARY_DOUBLE`

See Also:

Oracle Machine Learning for SQL User's Guide for valid target data types

Oracle Machine Learning for SQL Concepts for more information about cost matrixes

- The confusion matrix created by `COMPUTE_CONFUSION_MATRIX_PART` has the columns described in Table 6-60.

Table 6-63   Columns in a Confusion Matrix Part

Column Name	Data Type
`actual_target_value`	Type of the target column in the test data
`predicted_target_value`	Type of the predicted target in the test data. The type of the predicted target is the same as the type of the actual target unless the predicted target has an associated reverse transformation.
`value`	`BINARY_DOUBLE`

See Also:

Oracle Machine Learning for SQL Concepts for more information about confusion matrixes

Examples

These examples use the Naive Bayes model `nb_sh_clas_sample`.

Compute a Confusion Matrix Based on Probabilities

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
       SELECT cust_id,
              PREDICTION(nb_sh_clas_sample USING *) prediction,
              PREDICTION_PROBABILITY(nb_sh_clas_sample USING *) probability
       FROM mining_data_test_v;
```

Using probabilities as the scoring criterion, you can compute the confusion matrix as follows.

```
DECLARE
   v_accuracy    NUMBER;
      BEGIN
        DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX_PART (
                   accuracy                     => v_accuracy,
                   apply_result_table_name      => 'nb_apply_results',
                   target_table_name            => 'mining_data_test_v',
                   case_id_column_name          => 'cust_id',
                   target_column_name           => 'affinity_card',
                   confusion_matrix_table_name  => 'nb_confusion_matrix',
                   score_column_name            => 'PREDICTION',
                   score_criterion_column_name  => 'PROBABILITY'
                   score_partition_column_name  => 'PARTITION_NAME'
                   cost_matrix_table_name       =>  null,
                   apply_result_schema_name     =>  null,
                   target_schema_name           =>  null,
                   cost_matrix_schema_name      =>  null,
                   score_criterion_type         => 'PROBABILITY');
        DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' || ROUND(v_accuracy,4));
      END;
      /
```

The confusion matrix and model accuracy are shown as follows.

```
 **** MODEL ACCURACY ****: .7847

SELECT * FROM NB_CONFUSION_MATRIX;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         60
                  0                      0        891
                  1                      1        286
                  0                      1        263
```

Compute a Confusion Matrix Based on a Cost Matrix Table

The confusion matrix in the previous example shows a high rate of false positives. For 263 cases, the model predicted 1 when the actual value was 0. You could use a cost matrix to minimize this type of error.

The cost matrix table `nb_cost_matrix` specifies that a false positive is 3 times more costly than a false negative.

```
 SELECT * from NB_COST_MATRIX;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE       COST
------------------- ---------------------- ----------
                  0                      0          0
                  0                      1        .75
                  1                      0        .25
                  1                      1          0
```

This statement shows how to generate the predictions using `APPLY`.

```
BEGIN
    DBMS_DATA_MINING.APPLY(
          model_name          => 'nb_sh_clas_sample',
          data_table_name     => 'mining_data_test_v',
          case_id_column_name => 'cust_id',
          result_table_name   => 'nb_apply_results');
 END;
/
```

This statement computes the confusion matrix using the cost matrix table. The score criterion column is named '`PROBABILITY`', which is the name generated by `APPLY`.

```
DECLARE
  v_accuracy    NUMBER;
     BEGIN
       DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX_PART (
                accuracy                     => v_accuracy,
                apply_result_table_name      => 'nb_apply_results',
                target_table_name            => 'mining_data_test_v',
                case_id_column_name          => 'cust_id',
                target_column_name           => 'affinity_card',
                confusion_matrix_table_name  => 'nb_confusion_matrix',
                score_column_name            => 'PREDICTION',
                score_criterion_column_name  => 'PROBABILITY',
                score_partition_column_name  => 'PARTITION_NAME'
                cost_matrix_table_name       => 'nb_cost_matrix',
                apply_result_schema_name     => null,
                target_schema_name           => null,
                cost_matrix_schema_name      => null,
                score_criterion_type         => 'COST');
       DBMS_OUTPUT.PUT_LINE('**** MODEL ACCURACY ****: ' || ROUND(v_accuracy,4));
    END;
    /
```

The resulting confusion matrix shows a decrease in false positives (212 instead of 263).

```
**** MODEL ACCURACY ****: .798

 SELECT * FROM NB_CONFUSION_MATRIX;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         91
                  0                      0        942
                  1                      1        255
                  0                      1        212
```

Compute a Confusion Matrix Based on Embedded Costs

You can use the ADD_COST_MATRIX procedure to embed a cost matrix in a model. The embedded costs can be used instead of probabilities for scoring. This statement adds the previously-defined cost matrix to the model.

```
BEGIN
DBMS_DATA_MINING.ADD_COST_MATRIX ('nb_sh_clas_sample', 'nb_cost_matrix');
END;/
```

The following statement applies the model to the test data using the embedded costs and stores the results in a table.

```
CREATE TABLE nb_apply_results AS
         SELECT cust_id,
              PREDICTION(nb_sh_clas_sample COST MODEL USING *) prediction,
              PREDICTION_COST(nb_sh_clas_sample COST MODEL USING *) cost
          FROM mining_data_test_v;
```

You can compute the confusion matrix using the embedded costs.

```
DECLARE
   v_accuracy         NUMBER;
   BEGIN
       DBMS_DATA_MINING.COMPUTE_CONFUSION_MATRIX_PART (
            accuracy                     => v_accuracy,
            apply_result_table_name      => 'nb_apply_results',
            target_table_name            => 'mining_data_test_v',
            case_id_column_name          => 'cust_id',
            target_column_name           => 'affinity_card',
            confusion_matrix_table_name  => 'nb_confusion_matrix',
            score_column_name            => 'PREDICTION',
            score_criterion_column_name  => 'COST',
            score_partition_column_name  => 'PARTITION_NAME'
            cost_matrix_table_name       => null,
            apply_result_schema_name     => null,
            target_schema_name           => null,
            cost_matrix_schema_name      => null,
            score_criterion_type         => 'COST');
   END;
   /
```

The results are:

```
**** MODEL ACCURACY ****: .798

 SELECT * FROM NB_CONFUSION_MATRIX;
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE      VALUE
------------------- ---------------------- ----------
                  1                      0         91
                  0                      0        942
                  1                      1        255
                  0                      1        212
```

COMPUTE_LIFT Procedure

This procedure computes lift and stores the results in a table in the user's schema.

Lift is a test metric for binary classification models. To compute lift, one of the target values must be designated as the positive class. COMPUTE_LIFT compares the predictions generated by the model with the actual target values in a set of test data. Lift measures the degree to which the model's predictions of the positive class are an improvement over random chance.

Lift is computed on scoring results that have been ranked by probability (or cost) and divided into quantiles. Each quantile includes the scores for the same number of cases.

COMPUTE_LIFT calculates quantile-based and cumulative statistics. The number of quantiles and the positive class are user-specified. Additionally, COMPUTE_LIFT accepts three input streams:

- The predictions generated on the test data. The information is passed in three columns:
  - Case ID column
  - Prediction column
  - Scoring criterion column containing either probabilities or costs associated with the predictions
- The known target values in the test data. The information is passed in two columns:
  - Case ID column
  - Target column containing the known target values
- (Optional) A cost matrix table with predefined columns. See the Usage Notes for the column requirements.

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more details about lift and test metrics for classification
>
> "COMPUTE_CONFUSION_MATRIX Procedure"
>
> "COMPUTE_ROC Procedure"

Syntax

```
DBMS_DATA_MINING.COMPUTE_LIFT (
      apply_result_table_name      IN VARCHAR2,
      target_table_name            IN VARCHAR2,
      case_id_column_name          IN VARCHAR2,
      target_column_name           IN VARCHAR2,
      lift_table_name              IN VARCHAR2,
      positive_target_value        IN VARCHAR2,
      score_column_name            IN VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name  IN VARCHAR2 DEFAULT 'PROBABILITY',
      num_quantiles                IN NUMBER DEFAULT 10,
      cost_matrix_table_name       IN VARCHAR2 DEFAULT NULL,
      apply_result_schema_name     IN VARCHAR2 DEFAULT NULL,
      target_schema_name           IN VARCHAR2 DEFAULT NULL,
      cost_matrix_schema_name      IN VARCHAR2 DEFAULT NULL
      score_criterion_type         IN VARCHAR2 DEFAULT 'PROBABILITY');
```

Parameters

Table 6-64 COMPUTE_LIFT Procedure Parameters

Parameter	Description
apply_result_table_name	Table containing the predictions.

Table 6-64 (Cont.) COMPUTE_LIFT Procedure Parameters

Parameter	Description
target_table_name	Table containing the known target values from the test data.
case_id_column_name	Case ID column in the apply results table. Must match the case identifier in the targets table.
target_column_name	Target column in the targets table. Contains the known target values from the test data.
lift_table_name	Table containing the lift statistics. The table will be created by the procedure in the user's schema. The columns in the lift table are described in the Usage Notes.
positive_target_value	The positive class. This should be the class of interest, for which you want to calculate lift. If the target column is a NUMBER, you can use the TO_CHAR() operator to provide the value as a string.
score_column_name	Column containing the predictions in the apply results table. The default column name is 'PREDICTION', which is the default name created by the APPLY procedure (See "APPLY Procedure").
score_criterion_column_name	Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, the class with the lowest cost is predicted. The score_criterion_type parameter indicates whether probabilities or costs will be used for scoring. The default column name is 'PROBABILITY', which is the default name created by the APPLY procedure (See "APPLY Procedure"). See the Usage Notes for additional information.
num_quantiles	Number of quantiles to be used in calculating lift. The default is 10.
cost_matrix_table_name	(Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the score_criterion_type parameter is set to 'COST', the costs will be used as the scoring criteria. The columns in a cost matrix table are described in the Usage Notes.
apply_result_schema_name	Schema of the apply results table. If null, the user's schema is assumed.
target_schema_name	Schema of the table containing the known targets. If null, the user's schema is assumed.
cost_matrix_schema_name	Schema of the cost matrix table, if one is provided. If null, the user's schema is assumed.

Table 6-64 (Cont.) COMPUTE_LIFT Procedure Parameters

Parameter	Description
score_criterion_type	Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the score_criterion_column_name parameter. The default value of score_criterion_type is 'PROBABILITY'. To use costs as the scoring criterion, specify 'COST'. If score_criterion_type is set to 'COST' but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring. See the Usage Notes and the Examples.

Usage Notes

- The predictive information you pass to COMPUTE_LIFT may be generated using SQL PREDICTION functions, the DBMS_DATA_MINING.APPLY procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the lift.
- Instead of passing a cost matrix to COMPUTE_LIFT, you can use a scoring cost matrix associated with the model. A scoring cost matrix can be embedded in the model or it can be defined dynamically when the model is applied. To use a scoring cost matrix, invoke the SQL PREDICTION_COST function to populate the score criterion column.
- The predictions that you pass to COMPUTE_LIFT are in a table or view specified in apply_results_table_name.

```
CREATE TABLE apply_result_table_name AS (
            case_id_column_name              VARCHAR2,
            score_column_name              VARCHAR2,
            score_criterion_column_name       VARCHAR2);
```

- A cost matrix must have the columns described in Table 6-65.

Table 6-65 Columns in a Cost Matrix

Column Name	Data Type
actual_target_value	Type of the target column in the build data
predicted_target_value	Type of the predicted target in the test data. The type of the predicted target must be the same as the type of the actual target unless the predicted target has an associated reverse transformation.
cost	NUMBER

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more information about cost matrixes

- The table created by COMPUTE_LIFT has the columns described in Table 6-66

Table 6-66   Columns in a Lift Table

Column Name	Data Type
quantile_number	NUMBER
probability_threshold	NUMBER
gain_cumulative	NUMBER
quantile_total_count	NUMBER
quantile_target_count	NUMBER
percent_records_cumulative	NUMBER
lift_cumulative	NUMBER
target_density_cumulative	NUMBER
targets_cumulative	NUMBER
non_targets_cumulative	NUMBER
lift_quantile	NUMBER
target_density	NUMBER

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for details about the information in the lift table

- When a cost matrix is passed to COMPUTE_LIFT, the cost threshold is returned in the probability_threshold column of the lift table.

Examples

This example uses the Naive Bayes model nb_sh_clas_sample.

The example illustrates lift based on probabilities. For examples that show computation based on costs, see "COMPUTE_CONFUSION_MATRIX Procedure".

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
    SELECT cust_id, t.prediction, t.probability
    FROM mining_data_test_v, TABLE(PREDICTION_SET(nb_sh_clas_sample USING *)) t;
```

Using probabilities as the scoring criterion, you can compute lift as follows.

```
BEGIN
       DBMS_DATA_MINING.COMPUTE_LIFT (
            apply_result_table_name            => 'nb_apply_results',
            target_table_name                  => 'mining_data_test_v',
            case_id_column_name                => 'cust_id',
            target_column_name                 => 'affinity_card',
            lift_table_name                        => 'nb_lift',
            positive_target_value              =>  to_char(1),
            score_column_name                  => 'PREDICTION',
            score_criterion_column_name    => 'PROBABILITY',
            num_quantiles                          =>  10,
            cost_matrix_table_name             =>  null,
            apply_result_schema_name          =>  null,
            target_schema_name                =>  null,
            cost_matrix_schema_name           =>  null,
            score_criterion_type              =>  'PROBABILITY');
      END;
      /
```

This query displays some of the statistics from the resulting lift table.

```
SQL>SELECT quantile_number, probability_threshold, gain_cumulative,
           quantile_total_count
           FROM nb_lift;

QUANTILE_NUMBER PROBABILITY_THRESHOLD GAIN_CUMULATIVE QUANTILE_TOTAL_COUNT
--------------- --------------------- --------------- --------------------
              1            .989335775       .15034965                   55
              2            .980534911       .26048951                   55
              3            .968506098      .374125874                   55
              4            .958975196      .493006993                   55
              5            .946705997      .587412587                   55
              6            .927454174       .66958042                   55
              7            .904403627      .748251748                   55
              8            .836482525      .839160839                   55
             10            .500184953               1                   54
```

COMPUTE_LIFT_PART Procedure

The `COMPUTE_LIFT_PART` procedure computes lift and stores the results in a table in the user's schema. This procedure provides support to the computation of evaluation metrics per-partition for partitioned models.

Lift is a test metric for binary classification models. To compute lift, one of the target values must be designated as the positive class. `COMPUTE_LIFT_PART` compares the predictions generated by the model with the actual target values in a set of test data. Lift measures the degree to which the model's predictions of the positive class are an improvement over random chance.

Lift is computed on scoring results that have been ranked by probability (or cost) and divided into quantiles. Each quantile includes the scores for the same number of cases.

`COMPUTE_LIFT_PART` calculates quantile-based and cumulative statistics. The number of quantiles and the positive class are user-specified. Additionally, `COMPUTE_LIFT_PART` accepts three input streams:

- The predictions generated on the test data. The information is passed in three columns:
  – Case ID column
  – Prediction column
  – Scoring criterion column containing either probabilities or costs associated with the predictions
- The known target values in the test data. The information is passed in two columns:
  – Case ID column
  – Target column containing the known target values
- (Optional) A cost matrix table with predefined columns. See the Usage Notes for the column requirements.

See Also:

Oracle Machine Learning for SQL Concepts for more details about Lift and test metrics for classification

"COMPUTE_LIFT Procedure"

"COMPUTE_CONFUSION_MATRIX Procedure"

"COMPUTE_CONFUSION_MATRIX_PART Procedure"

"COMPUTE_ROC Procedure"

"COMPUTE_ROC_PART Procedure"

Syntax

```
DBMS_DATA_MINING.COMPUTE_LIFT_PART (
      apply_result_table_name     IN VARCHAR2,
      target_table_name           IN VARCHAR2,
      case_id_column_name         IN VARCHAR2,
      target_column_name          IN VARCHAR2,
      lift_table_name             IN VARCHAR2,
      positive_target_value       IN VARCHAR2,
      score_column_name           IN VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name IN VARCHAR2 DEFAULT 'PROBABILITY',
      score_partition_column_name IN VARCHAR2 DEFAULT 'PARTITION_NAME',
      num_quantiles               IN NUMBER   DEFAULT 10,
      cost_matrix_table_name      IN VARCHAR2 DEFAULT NULL,
      apply_result_schema_name    IN VARCHAR2 DEFAULT NULL,
      target_schema_name          IN VARCHAR2 DEFAULT NULL,
      cost_matrix_schema_name     IN VARCHAR2 DEFAULT NULL,
      score_criterion_type        IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-67 COMPUTE_LIFT_PART Procedure Parameters

Parameter	Description
`apply_result_table_name`	Table containing the predictions
`target_table_name`	Table containing the known target values from the test data
`case_id_column_name`	Case ID column in the apply results table. Must match the case identifier in the targets table.
`target_column_name`	Target column in the targets table. Contains the known target values from the test data.
`lift_table_name`	Table containing the Lift statistics. The table will be created by the procedure in the user's schema. The columns in the Lift table are described in the Usage Notes.
`positive_target_value`	The positive class. This should be the class of interest, for which you want to calculate Lift. If the target column is a `NUMBER`, then you can use the `TO_CHAR()` operator to provide the value as a string.

Table 6-67 (Cont.) COMPUTE_LIFT_PART Procedure Parameters

Parameter	Description
score_column_name	Column containing the predictions in the apply results table. The default column name is PREDICTION, which is the default name created by the APPLY procedure (See "APPLY Procedure").
score_criterion_column_name	Column containing the scoring criterion in the apply results table. Contains either the probabilities or the costs that determine the predictions. By default, scoring is based on probability; the class with the highest probability is predicted for each case. If scoring is based on cost, then the class with the lowest cost is predicted. The score_criterion_type parameter indicates whether probabilities or costs will be used for scoring. The default column name is PROBABILITY, which is the default name created by the APPLY procedure (See "APPLY Procedure"). See the Usage Notes for additional information.
score_partition_column_name	Optional parameter indicating the column containing the name of the partition. This column slices the input test results such that each partition has independent evaluation matrices computed.
num_quantiles	Number of quantiles to be used in calculating Lift. The default is 10.
cost_matrix_table_name	(Optional) Table that defines the costs associated with misclassifications. If a cost matrix table is provided and the score_criterion_type parameter is set to COST, then the costs will be used as the scoring criteria. The columns in a cost matrix table are described in the Usage Notes.
apply_result_schema_name	Schema of the apply results table If null, then the user's schema is assumed.
target_schema_name	Schema of the table containing the known targets If null, then the user's schema is assumed.
cost_matrix_schema_name	Schema of the cost matrix table, if one is provided If null, then the user's schema is assumed.
score_criterion_type	Whether to use probabilities or costs as the scoring criterion. Probabilities or costs are passed in the column identified in the score_criterion_column_name parameter. The default value of score_criterion_type is PROBABILITY. To use costs as the scoring criterion, specify COST. If score_criterion_type is set to COST but no cost matrix is provided and if there is a scoring cost matrix associated with the model, then the associated costs are used for scoring. See the Usage Notes and the Examples.

Usage Notes

- The predictive information you pass to COMPUTE_LIFT_PART may be generated using SQL PREDICTION functions, the DBMS_DATA_MINING.APPLY procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the Lift.
- Instead of passing a cost matrix to COMPUTE_LIFT_PART, you can use a scoring cost matrix associated with the model. A scoring cost matrix can be embedded in the model or it can be defined dynamically when the model is applied. To use a scoring cost matrix, invoke the SQL PREDICTION_COST function to populate the score criterion column.
- The predictions that you pass to COMPUTE_LIFT_PART are in a table or view specified in apply_results_table_name.

```
CREATE TABLE apply_result_table_name AS (
            case_id_column_name            VARCHAR2,
            score_column_name              VARCHAR2,
            score_criterion_column_name    VARCHAR2);
```

- A cost matrix must have the columns described in Table 6-65.

Table 6-68 Columns in a Cost Matrix

Column Name	Data Type
actual_target_value	Type of the target column in the test data
predicted_target_value	Type of the predicted target in the test data. The type of the predicted target must be the same as the type of the actual target unless the predicted target has an associated reverse transformation.
cost	NUMBER

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for more information about cost matrixes

- The table created by COMPUTE_LIFT_PART has the columns described in Table 6-66

Table 6-69 Columns in a COMPUTE_LIFT_PART Table

Column Name	Data Type
quantile_number	NUMBER
probability_threshold	NUMBER
gain_cumulative	NUMBER
quantile_total_count	NUMBER
quantile_target_count	NUMBER
percent_records_cumulative	NUMBER
lift_cumulative	NUMBER
target_density_cumulative	NUMBER
targets_cumulative	NUMBER

Table 6-69 (Cont.) Columns in a COMPUTE_LIFT_PART Table

Column Name	Data Type
non_targets_cumulative	NUMBER
lift_quantile	NUMBER
target_density	NUMBER

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for details about the information in the Lift table

- When a cost matrix is passed to COMPUTE_LIFT_PART, the cost threshold is returned in the probability_threshold column of the Lift table.

Examples

This example uses the Naive Bayes model nb_sh_clas_sample.

The example illustrates Lift based on probabilities. For examples that show computation based on costs, see "COMPUTE_CONFUSION_MATRIX Procedure".

For a partitioned model example, see "COMPUTE_CONFUSION_MATRIX_PART Procedure".

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
    SELECT cust_id, t.prediction, t.probability
    FROM mining_data_test_v, TABLE(PREDICTION_SET(nb_sh_clas_sample USING *)) t;
```

Using probabilities as the scoring criterion, you can compute Lift as follows.

```
BEGIN
      DBMS_DATA_MINING.COMPUTE_LIFT_PART (
             apply_result_table_name     => 'nb_apply_results',
             target_table_name           => 'mining_data_test_v',
             case_id_column_name         => 'cust_id',
             target_column_name          => 'affinity_card',
             lift_table_name             => 'nb_lift',
             positive_target_value       =>  to_char(1),
             score_column_name           => 'PREDICTION',
             score_criterion_column_name => 'PROBABILITY',
             score_partition_column_name => 'PARTITITON_NAME',
             num_quantiles               =>  10,
             cost_matrix_table_name      =>  null,
             apply_result_schema_name    =>  null,
             target_schema_name          =>  null,
             cost_matrix_schema_name     =>  null,
             score_criterion_type        =>  'PROBABILITY');
END;
/
```

This query displays some of the statistics from the resulting Lift table.

```
SELECT quantile_number, probability_threshold, gain_cumulative,
           quantile_total_count
           FROM nb_lift;

QUANTILE_NUMBER PROBABILITY_THRESHOLD GAIN_CUMULATIVE QUANTILE_TOTAL_COUNT
--------------- --------------------- --------------- --------------------
              1            .989335775       .15034965                   55
              2            .980534911       .26048951                   55
              3            .968506098      .374125874                   55
              4            .958975196      .493006993                   55
              5            .946705997      .587412587                   55
              6            .927454174       .66958042                   55
              7            .904403627      .748251748                   55
              8            .836482525      .839160839                   55
             10            .500184953               1                   54
```

COMPUTE_ROC Procedure

This procedure computes the receiver operating characteristic (ROC), stores the results in a table in the user's schema, and returns a measure of the model accuracy.

ROC is a test metric for binary classification models. To compute ROC, one of the target values must be designated as the positive class. COMPUTE_ROC compares the predictions generated by the model with the actual target values in a set of test data.

ROC measures the impact of changes in the probability threshold. The probability threshold is the decision point used by the model for predictions. In binary classification, the default probability threshold is 0.5. The value predicted for each case is the one with a probability greater than 50%.

ROC can be plotted as a curve on an X-Y axis. The false positive rate is placed on the X axis. The true positive rate is placed on the Y axis. A false positive is a positive prediction for a case that is negative in the test data. A true positive is a positive prediction for a case that is positive in the test data.

COMPUTE_ROC accepts two input streams:

- The predictions generated on the test data. The information is passed in three columns:
  - Case ID column
  - Prediction column
  - Scoring criterion column containing probabilities
- The known target values in the test data. The information is passed in two columns:
  - Case ID column
  - Target column containing the known target values

See Also:

Oracle Machine Learning for SQL Concepts for more details about ROC and test metrics for classification

"COMPUTE_CONFUSION_MATRIX Procedure"

"COMPUTE_LIFT Procedure"

Syntax

```
DBMS_DATA_MINING.COMPUTE_ROC (
      roc_area_under_curve          OUT NUMBER,
      apply_result_table_name       IN  VARCHAR2,
      target_table_name             IN  VARCHAR2,
      case_id_column_name           IN  VARCHAR2,
      target_column_name            IN  VARCHAR2,
      roc_table_name                IN  VARCHAR2,
      positive_target_value         IN  VARCHAR2,
      score_column_name             IN  VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name   IN  VARCHAR2 DEFAULT 'PROBABILITY',
      apply_result_schema_name      IN  VARCHAR2 DEFAULT NULL,
      target_schema_name            IN  VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-70   COMPUTE_ROC Procedure Parameters

Parameter	Description
`roc_area_under_the_curve`	Output parameter containing the area under the ROC curve (AUC). The AUC measures the likelihood that an actual positive will be predicted as positive. The greater the AUC, the greater the flexibility of the model in accommodating trade-offs between positive and negative class predictions. AUC can be especially important when one target class is rarer or more important to identify than another.
`apply_result_table_name`	Table containing the predictions.
`target_table_name`	Table containing the known target values from the test data.
`case_id_column_name`	Case ID column in the apply results table. Must match the case identifier in the targets table.
`target_column_name`	Target column in the targets table. Contains the known target values from the test data.
`roc_table_name`	Table containing the ROC output. The table will be created by the procedure in the user's schema. The columns in the ROC table are described in the Usage Notes.
`positive_target_value`	The positive class. This should be the class of interest, for which you want to calculate ROC. If the target column is a `NUMBER`, you can use the `TO_CHAR()` operator to provide the value as a string.

Table 6-70 (Cont.) COMPUTE_ROC Procedure Parameters

Parameter	Description
`score_column_name`	Column containing the predictions in the apply results table. The default column name is '`PREDICTION`', which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`score_criterion_column_name`	Column containing the scoring criterion in the apply results table. Contains the probabilities that determine the predictions. The default column name is '`PROBABILITY`', which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`apply_result_schema_name`	Schema of the apply results table. If null, the user's schema is assumed.
`target_schema_name`	Schema of the table containing the known targets. If null, the user's schema is assumed.

Usage Notes

- The predictive information you pass to `COMPUTE_ROC` may be generated using SQL `PREDICTION` functions, the `DBMS_DATA_MINING.APPLY` procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the receiver operating characteristic.
- The predictions that you pass to `COMPUTE_ROC` are in a table or view specified in `apply_results_table_name`.

```
CREATE TABLE apply_result_table_name AS (
            case_id_column_name            VARCHAR2,
            score_column_name              VARCHAR2,
            score_criterion_column_name    VARCHAR2);
```

- The table created by `COMPUTE_ROC` has the columns shown in Table 6-71.

Table 6-71 COMPUTE_ROC Output

Column	Datatype
`probability`	`BINARY_DOUBLE`
`true_positives`	`NUMBER`
`false_negatives`	`NUMBER`
`false_positives`	`NUMBER`
`true_negatives`	`NUMBER`
`true_positive_fraction`	`NUMBER`
`false_positive_fraction`	`NUMBER`

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for details about the output of COMPUTE_ROC

- ROC is typically used to determine the most desirable probability threshold. This can be done by examining the true positive fraction and the false positive fraction. The true positive fraction is the percentage of all positive cases in the test data that were correctly predicted as positive. The false positive fraction is the percentage of all negative cases in the test data that were incorrectly predicted as positive.

  Given a probability threshold, the following statement returns the positive predictions in an apply result table ordered by probability.

  ```
  SELECT case_id_column_name
         FROM apply_result_table_name
         WHERE probability > probability_threshold
         ORDER BY probability DESC;
  ```

- There are two approaches to identifying the most desirable probability threshold. Which approach you use depends on whether or not you know the relative cost of positive versus negative class prediction errors.

  If the costs are known, you can apply the relative costs to the ROC table to compute the minimum cost probability threshold. Suppose the relative cost ratio is: Positive Class Error Cost / Negative Class Error Cost = 20. Then execute a query like this.

  ```
  WITH cost AS (
    SELECT probability_threshold, 20 * false_negatives + false_positives cost
      FROM ROC_table
    GROUP BY probability_threshold),
      minCost AS (
        SELECT min(cost) minCost
          FROM cost)
        SELECT max(probability_threshold)probability_threshold
          FROM cost, minCost
      WHERE cost = minCost;
  ```

  If relative costs are not well known, you can simply scan the values in the ROC table (in sorted order) and make a determination about which of the displayed trade-offs (misclassified positives versus misclassified negatives) is most desirable.

  ```
  SELECT * FROM ROC_table
           ORDER BY probability_threshold;
  ```

Examples

This example uses the Naive Bayes model `nb_sh_clas_sample`.

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
    SELECT cust_id, t.prediction, t.probability
    FROM mining_data_test_v, TABLE(PREDICTION_SET(nb_sh_clas_sample USING *)) t;
```

Using the predictions and the target values from the test data, you can compute ROC as follows.

```
DECLARE
     v_area_under_curve NUMBER;
BEGIN
     DBMS_DATA_MINING.COMPUTE_ROC (
         roc_area_under_curve        => v_area_under_curve,
         apply_result_table_name     => 'nb_apply_results',
         target_table_name           => 'mining_data_test_v',
         case_id_column_name         => 'cust_id',
         target_column_name          => 'mining_data_test_v',
         roc_table_name              => 'nb_roc',
         positive_target_value       => '1',
         score_column_name           => 'PREDICTION',
         score_criterion_column_name => 'PROBABILITY');
     DBMS_OUTPUT.PUT_LINE('**** AREA UNDER ROC CURVE ****: ' ||
     ROUND(v_area_under_curve,4));
END;
/
```

The resulting AUC and a selection of columns from the ROC table are shown as follows.

```
**** AREA UNDER ROC CURVE ****: .8212

 SELECT PROBABILITY, TRUE_POSITIVE_FRACTION, FALSE_POSITIVE_FRACTION
           FROM NB_ROC;

PROBABILITY  TRUE_POSITIVE_FRACTION  FALSE_POSITIVE_FRACTION
-----------  ----------------------  -----------------------
     .00000                       1                        1
     .50018              .826589595               .227902946
     .53851              .823699422               .221837088
     .54991              .820809249               .217504333
     .55628              .815028902               .215771231
     .55628              .817919075               .215771231
     .57563              .800578035               .214904679
     .57563              .812138728               .214904679
      .                   .                        .
      .                   .                        .
      .                   .                        .
```

COMPUTE_ROC_PART Procedure

The COMPUTE_ROC_PART procedure computes Receiver Operating Characteristic (ROC), stores the results in a table in the user's schema, and returns a measure of the model accuracy. This procedure provides support to computation of evaluation metrics per-partition for partitioned models.

ROC is a test metric for binary classification models. To compute ROC, one of the target values must be designated as the positive class. COMPUTE_ROC_PART compares the predictions generated by the model with the actual target values in a set of test data.

ROC measures the impact of changes in the probability threshold. The probability threshold is the decision point used by the model for predictions. In binary classification, the default probability threshold is 0.5. The value predicted for each case is the one with a probability greater than 50%.

ROC can be plotted as a curve on an x-y axis. The false positive rate is placed on the x-axis. The true positive rate is placed on the y-axis. A false positive is a positive prediction for a case that is negative in the test data. A true positive is a positive prediction for a case that is positive in the test data.

`COMPUTE_ROC_PART` accepts two input streams:

- The predictions generated on the test data. The information is passed in three columns:
  - Case ID column
  - Prediction column
  - Scoring criterion column containing probabilities
- The known target values in the test data. The information is passed in two columns:
  - Case ID column
  - Target column containing the known target values

See Also:

Oracle Machine Learning for SQL Concepts for more details about ROC and test metrics for Classification

"COMPUTE_ROC Procedure"

"COMPUTE_CONFUSION_MATRIX Procedure"

"COMPUTE_LIFT_PART Procedure"

"COMPUTE_LIFT Procedure"

Syntax

```
DBMS_DATA_MINING.compute_roc_part(
      roc_area_under_curve         OUT DM_NESTED_NUMERICALS,
      apply_result_table_name      IN  VARCHAR2,
      target_table_name            IN  VARCHAR2,
      case_id_column_name          IN  VARCHAR2,
      target_column_name           IN  VARCHAR2,
      roc_table_name               IN  VARCHAR2,
      positive_target_value        IN  VARCHAR2,
      score_column_name            IN  VARCHAR2 DEFAULT 'PREDICTION',
      score_criterion_column_name  IN  VARCHAR2 DEFAULT 'PROBABILITY',
      score_partition_column_name  IN  VARCHAR2 DEFAULT 'PARTITION_NAME',
      apply_result_schema_name     IN  VARCHAR2 DEFAULT NULL,
      target_schema_name           IN  VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-72 COMPUTE_ROC_PART Procedure Parameters

Parameter	Description
`roc_area_under_the_curve`	Output parameter containing the area under the ROC curve (AUC). The AUC measures the likelihood that an actual positive will be predicted as positive. The greater the AUC, the greater the flexibility of the model in accommodating trade-offs between positive and negative class predictions. AUC can be especially important when one target class is rarer or more important to identify than another. The output argument is changed from `NUMBER` to `DM_NESTED_NUMERICALS`.
`apply_result_table_name`	Table containing the predictions.
`target_table_name`	Table containing the known target values from the test data.
`case_id_column_name`	Case ID column in the apply results table. Must match the case identifier in the targets table.
`target_column_name`	Target column in the targets table. Contains the known target values from the test data.
`roc_table_name`	Table containing the ROC output. The table will be created by the procedure in the user's schema. The columns in the ROC table are described in the Usage Notes.
`positive_target_value`	The positive class. This should be the class of interest, for which you want to calculate ROC. If the target column is a `NUMBER`, then you can use the `TO_CHAR()` operator to provide the value as a string.
`score_column_name`	Column containing the predictions in the apply results table. The default column name is `PREDICTION`, which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`score_criterion_column_name`	Column containing the scoring criterion in the apply results table. Contains the probabilities that determine the predictions. The default column name is `PROBABILITY`, which is the default name created by the `APPLY` procedure (See "APPLY Procedure").
`score_partition_column_name`	Optional parameter indicating the column which contains the name of the partition. This column slices the input test results such that each partition has independent evaluation matrices computed.
`apply_result_schema_name`	Schema of the apply results table. If null, then the user's schema is assumed.
`target_schema_name`	Schema of the table containing the known targets. If null, then the user's schema is assumed.

Usage Notes

- The predictive information you pass to COMPUTE_ROC_PART may be generated using SQL PREDICTION functions, the DBMS_DATA_MINING.APPLY procedure, or some other mechanism. As long as you pass the appropriate data, the procedure can compute the receiver operating characteristic.
- The predictions that you pass to COMPUTE_ROC_PART are in a table or view specified in apply_results_table_name.

```
CREATE TABLE apply_result_table_name AS (
            case_id_column_name              VARCHAR2,
            score_column_name              VARCHAR2,
            score_criterion_column_name      VARCHAR2);
```

- The COMPUTE_ROC_PART table has the following columns:

Table 6-73   COMPUTE_ROC_PART Output

Column	Data Type
probability	BINARY_DOUBLE
true_positives	NUMBER
false_negatives	NUMBER
false_positives	NUMBER
true_negatives	NUMBER
true_positive_fraction	NUMBER
false_positive_fraction	NUMBER

> **See Also:**
>
> *Oracle Machine Learning for SQL Concepts* for details about the output of COMPUTE_ROC_PART

- ROC is typically used to determine the most desirable probability threshold. This can be done by examining the true positive fraction and the false positive fraction. The true positive fraction is the percentage of all positive cases in the test data that were correctly predicted as positive. The false positive fraction is the percentage of all negative cases in the test data that were incorrectly predicted as positive.

  Given a probability threshold, the following statement returns the positive predictions in an apply result table ordered by probability.

```
SELECT case_id_column_name
       FROM apply_result_table_name
       WHERE probability > probability_threshold
       ORDER BY probability DESC;
```

- There are two approaches to identify the most desirable probability threshold. The approach you use depends on whether you know the relative cost of positive versus negative class prediction errors.

If the costs are known, then you can apply the relative costs to the ROC table to compute the minimum cost probability threshold. Suppose the relative cost ratio is: Positive Class Error Cost / Negative Class Error Cost = 20. Then execute a query as follows:

```
WITH cost AS (
  SELECT probability_threshold, 20 * false_negatives + false_positives cost
    FROM ROC_table
  GROUP BY probability_threshold),
    minCost AS (
      SELECT min(cost) minCost
        FROM cost)
      SELECT max(probability_threshold)probability_threshold
        FROM cost, minCost
    WHERE cost = minCost;
```

If relative costs are not well known, then you can simply scan the values in the ROC table (in sorted order) and make a determination about which of the displayed trade-offs (misclassified positives versus misclassified negatives) is most desirable.

```
SELECT * FROM ROC_table
         ORDER BY probability_threshold;
```

Examples

This example uses the Naive Bayes model `nb_sh_clas_sample`.

The following statement applies the model to the test data and stores the predictions and probabilities in a table.

```
CREATE TABLE nb_apply_results AS
    SELECT cust_id, t.prediction, t.probability
    FROM mining_data_test_v, TABLE(PREDICTION_SET(nb_sh_clas_sample USING *)) t;
```

Using the predictions and the target values from the test data, you can compute ROC as follows.

```
DECLARE
     v_area_under_curve NUMBER;
BEGIN
     DBMS_DATA_MINING.COMPUTE_ROC_PART (
         roc_area_under_curve        => v_area_under_curve,
         apply_result_table_name     => 'nb_apply_results',
         target_table_name           => 'mining_data_test_v',
         case_id_column_name         => 'cust_id',
         target_column_name          => 'affinity_card',
         roc_table_name              => 'nb_roc',
         positive_target_value       => '1',
         score_column_name           => 'PREDICTION',
         score_criterion_column_name => 'PROBABILITY');
         score_partition_column_name => 'PARTITION_NAME'
     DBMS_OUTPUT.PUT_LINE('**** AREA UNDER ROC CURVE ****: ' ||
     ROUND(v_area_under_curve,4));
END;
/
```

The resulting AUC and a selection of columns from the ROC table are shown as follows.

```
**** AREA UNDER ROC CURVE ****: .8212

 SELECT PROBABILITY, TRUE_POSITIVE_FRACTION, FALSE_POSITIVE_FRACTION
            FROM NB_ROC;

PROBABILITY  TRUE_POSITIVE_FRACTION  FALSE_POSITIVE_FRACTION
-----------  ----------------------  -----------------------
     .00000                       1                        1
     .50018              .826589595               .227902946
     .53851              .823699422               .221837088
     .54991              .820809249               .217504333
     .55628              .815028902               .215771231
     .55628              .817919075               .215771231
     .57563              .800578035               .214904679
     .57563              .812138728               .214904679
      .                    .                        .
      .                    .                        .
      .                    .                        .
```

CREATE_MODEL Procedure

This procedure creates an Oracle Machine Learning for SQL model with a given machine learning function.

Syntax

```
DBMS_DATA_MINING.CREATE_MODEL (
      model_name            IN VARCHAR2,
      mining_function       IN VARCHAR2,
      data_table_name       IN VARCHAR2,
      case_id_column_name   IN VARCHAR2,
      target_column_name    IN VARCHAR2 DEFAULT NULL,
      settings_table_name   IN VARCHAR2 DEFAULT NULL,
      data_schema_name      IN VARCHAR2 DEFAULT NULL,
      settings_schema_name  IN VARCHAR2 DEFAULT NULL,
      xform_list            IN TRANSFORM_LIST DEFAULT NULL);
```

Parameters

Table 6-74 CREATE_MODEL Procedure Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used. See the Usage Notes for model naming restrictions.
`mining_function`	The machine learning function. Values are listed in Table 6-7.
`data_table_name`	Table or view containing the build data
`case_id_column_name`	Case identifier column in the build data.
`target_column_name`	For supervised models, the target column in the build data. `NULL` for unsupervised models.
`settings_table_name`	Table containing build settings for the model. `NULL` if there is no settings table (only default settings are used).
`data_schema_name`	Schema hosting the build data. If `NULL`, then the user's schema is assumed.

Table 6-74 (Cont.) CREATE_MODEL Procedure Parameters

Parameter	Description
settings_schema_name	Schema hosting the settings table. If NULLthen the user's schema is assumed.

xform_list	A list of transformations to be used in addition to or instead of automatic transformations, depending on the value of the PREP_AUTO setting. (See "Automatic Data Preparation".)
	The datatype of xform_list is TRANSFORM_LIST, which consists of records of type TRANSFORM_REC. Each TRANSFORM_REC specifies the transformation information for a single attribute.
	`TYPE`
	`  TRANFORM_REC      IS RECORD (`
	`     attribute_name       VARCHAR2(4000),`
	`     attribute_subname    VARCHAR2(4000),`
	`     expression           EXPRESSION_REC,`
	`     reverse_expression   EXPRESSION_REC,`
	`     attribute_spec       VARCHAR2(4000));`
	The expression field stores a SQL expression for transforming the attribute. The reverse_expression field stores a SQL expression for reversing the transformation in model details and, if the attribute is a target, in the results of scoring. The SQL expressions are manipulated by routines in the DBMS_DATA_MINING_TRANSFORM package:
	• SET_EXPRESSION Procedure
	• GET_EXPRESSION Function
	• SET_TRANSFORM Procedure
	The attribute_spec field identifies individualized treatment for the attribute. See the Usage Notes for details.
	See Table 6-127for details about the TRANSFORM_REC type.

Usage Notes

1. You can use the attribute_spec field of the xform_list argument to identify an attribute as unstructured text or to disable Automatic Data Preparation for the attribute. The attribute_spec can have the following values:

   - TEXT: Indicates that the attribute contains unstructured text. The TEXT value may optionally be followed by POLICY_NAME, TOKEN_TYPE, MAX_FEATURES, and MIN_DOCUMENTS parameters.

     TOKEN_TYPE has the following possible values: NORMAL, STEM, THEME, SYNONYM, BIGRAM, STEM_BIGRAM. SYNONYM may be optionally followed by a thesaurus name in square brackets.

     MAX_FEATURES specifies the maximum number of tokens extracted from the text.

     MIN_DOCUMENTS specifies the minimal number of documents in which every selected token shall occur. (For information about creating a text policy, see CTX_DDL.CREATE_POLICY in *Oracle Text Reference*).

     Oracle Machine Learning for SQL can process columns of VARCHAR2/CHAR, CLOB, BLOB, and BFILE as text. If the column is VARCHAR2 or CHAR and you do not specify TEXT, then OML4SQL processes the column as categorical data. If the column is CLOB, then OML4SQL processes it as text by default (You do not need to specify it as TEXT. However, you do need to provide an Oracle Text Policy in the settings). If the column is

BLOB or BFILE, then you must specify it as TEXT, otherwise CREATE_MODEL returns an error.

If you specify TEXT for a nested column or for an attribute in a nested column, then CREATE_MODEL returns an error.

- NOPREP: Disables ADP for the attribute. When ADP is OFF, the NOPREP value is ignored.

  You can specify NOPREP for a nested column, but not for an attribute in a nested column. If you specify NOPREP for an attribute in a nested column when ADP is on, then CREATE_MODEL will return an error.

2. You can obtain information about a model by querying the Data Dictionary views.

```
ALL/USER/DBA_MINING_MODELS
ALL/USER/DBA_MINING_MODEL_ATTRIBUTES
ALL/USER/DBA_MINING_MODEL_SETTINGS
ALL/USER/DBA_MINING_MODEL_VIEWS
ALL/USER/DBA_MINING_MODEL_PARTITIONS
ALL/USER/DBA_MINING_MODEL_XFORMS
```

   You can obtain information about model attributes by querying the model details through model views. Refer to *Oracle Machine Learning for SQL User's Guide*.

3. The naming rules for models are more restrictive than the naming rules for most database schema objects. A model name must satisfy the following additional requirements:

   - It must be 123 or fewer characters long.
   - It must be a nonquoted identifier. Oracle requires that nonquoted identifiers contain only alphanumeric characters, the underscore (_), dollar sign ($), and pound sign (#); the initial character must be alphabetic. Oracle strongly discourages the use of the dollar sign and pound sign in nonquoted literals.

   Naming requirements for schema objects are fully documented in *Oracle Database SQL Language Reference*.

4. To build a partitioned model, you must provide additional settings.

   The setting for partitioning columns are as follows:

```
INSERT INTO settings_table VALUES ('ODMS_PARTITION_COLUMNS', 'GENDER,
AGE');
```

   To set user-defined partition number for a model, the setting is as follows:

```
INSERT INTO settings_table VALUES ('ODMS_MAX_PARTITIONS', '10');
```

   The default value for maximum number of partitions is 1000.

5. By passing an xform_list to CREATE_MODEL, you can specify a list of transformations to be performed on the input data. If the PREP_AUTO setting is ON, the transformations are used in addition to the automatic transformations. If the PREP_AUTO setting is OFF, the specified transformations are the only ones implemented by the model. In both cases, transformation definitions are embedded in the model and run automatically whenever the model is applied. See "Automatic Data Preparation". Other transforms that can be specified with xform_list include FORCE_IN. Refer to *Oracle Machine Learning for SQL User's Guide*.

Examples

The first example builds a classification model using the Support Vector Machine algorithm.

```
-- Create the settings table
CREATE TABLE svm_model_settings (
  setting_name  VARCHAR2(30),
  setting_value VARCHAR2(30));

-- Populate the settings table
-- Specify SVM. By default, Naive Bayes is used for classification.
-- Specify ADP. By default, ADP is not used.
BEGIN
  INSERT INTO svm_model_settings (setting_name, setting_value) VALUES
     (dbms_data_mining.algo_name, dbms_data_mining.algo_support_vector_machines);
  INSERT INTO svm_model_settings (setting_name, setting_value) VALUES
     (dbms_data_mining.prep_auto,dbms_data_mining.prep_auto_on);
  COMMIT;
END;
/
-- Create the model using the specified settings
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'svm_model',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_data_build_v',
    case_id_column_name => 'cust_id',
    target_column_name  => 'affinity_card',
    settings_table_name => 'svm_model_settings');
END;
/
```

You can display the model settings with the following query:

```
SELECT * FROM user_mining_model_settings
       WHERE model_name IN 'SVM_MODEL';

MODEL_NAME     SETTING_NAME            SETTING_VALUE                  SETTING
-------------  ----------------------  -----------------------------  -------
SVM_MODEL      ALGO_NAME               ALGO_SUPPORT_VECTOR_MACHINES  INPUT

SVM_MODEL      SVMS_STD_DEV            3.004524                      DEFAULT
SVM_MODEL      PREP_AUTO               ON                            INPUT
SVM_MODEL      SVMS_COMPLEXITY_FACTOR  1.887389                      DEFAULT
SVM_MODEL      SVMS_KERNEL_FUNCTION    SVMS_LINEAR                      DEFAULT
SVM_MODEL      SVMS_CONV_TOLERANCE     .001                          DEFAULT
```

The following is an example of querying a model view instead of the older `GEL_MODEL_DETAILS_SVM` routine.

```
SELECT target_value, attribute_name, attribute_value, coefficient   FROM
DM$VLSVM_MODEL;
```

The second example creates an anomaly detection model. Anomaly detection uses SVM classification without a target. This example uses the same settings table created for the SVM classification model in the first example.

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'anomaly_detect_model',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'mining_data_build_v',
    case_id_column_name => 'cust_id',
    target_column_name  => null,
    settings_table_name => 'svm_model_settings');
END;
/
```

This query shows that the models created in these examples are the only ones in your schema.

```
SELECT model_name, mining_function, algorithm FROM user_mining_models;

MODEL_NAME               MINING_FUNCTION      ALGORITHM
----------------------   -------------------- ------------------------------
SVM_MODEL                CLASSIFICATION       SUPPORT_VECTOR_MACHINES
ANOMALY_DETECT_MODEL     CLASSIFICATION       SUPPORT_VECTOR_MACHINES
```

This query shows that only the SVM classification model has a target.

```
SELECT model_name, attribute_name, attribute_type, target
       FROM user_mining_model_attributes
       WHERE target = 'YES';

MODEL_NAME           ATTRIBUTE_NAME   ATTRIBUTE_TYPE     TARGET
------------------   ---------------  -----------------  ------
SVM_MODEL            AFFINITY_CARD    CATEGORICAL         YES
```

CREATE_MODEL2 Procedure

The CREATE_MODEL2 procedure is an alternate procedure to the CREATE_MODEL procedure, which enables creating a model without extra persistence stages. In the CREATE_MODEL procedure, the input is a table or a view and if such an object is not already present, the user must create it. By using the CREATE_MODEL2 procedure, the user does not need to create such transient database objects.

Syntax

```
DBMS_DATA_MINING.CREATE_MODEL2 (
     model_name             IN VARCHAR2,
     mining_function        IN VARCHAR2,
     data_query             IN CLOB,
     set_list               IN SETTING_LIST,
     case_id_column_name    IN VARCHAR2 DEFAULT NULL,
     target_column_name     IN VARCHAR2 DEFAULT NULL,
     xform_list             IN TRANSFORM_LIST DEFAULT NULL);
```

Parameters

Table 6-75 CREATE_MODEL2 Procedure Parameters

Parameter	Description
model_name	Name of the model in the form [schema_name.]model_name. If you do not specify a schema, then the current schema is used. See the Usage Notes, CREATE_MODEL Procedure for model naming restrictions.
mining_function	The machine learning function. Values are listed in DBMS_DATA_MINING — Machine Learning Function Settings.

Table 6-75 (Cont.) CREATE_MODEL2 Procedure Parameters

Parameter	Description
data_query	A query which provides training data for building the model.
set_list	Specifies the SETTING_LIST SETTING_LIST is a table of CLOB index by VARCHAR2(30); Where the index is the setting name and the CLOB is the setting value for that name.
case_id_column_name	Case identifier column in the build data.
target_column_name	For supervised models, the target column in the build data. NULL for unsupervised models.
xform_list	Refer to CREATE_MODEL Procedure.

Usage Notes

Refer to CREATE_MODEL Procedure for Usage Notes.

Examples

The following example uses the Support Vector Machine algorithm.

```
declare
 v_setlst DBMS_DATA_MINING.SETTING_LIST;

BEGIN
  v_setlst(dbms_data_mining.algo_name) :=
dbms_data_mining.algo_support_vector_machines;
  v_setlst(dbms_data_mining.prep_auto) := dbms_data_mining.prep_auto_on;

DBMS_DATA_MINING.CREATE_MODEL2(
    model_name          => 'svm_model',
    mining_function     => dbms_data_mining.classification,
    data_query          => 'select * from mining_data_build_v',
    data_table_name     => 'mining_data_build_v',
    case_id_column_name=> 'cust_id',
    target_column_name => 'affinity_card',
    set_list            => v_setlst,
    case_id_column_name=> 'cust_id',
    target_column_name => 'affinity_card');
END;
/
```

Create Model Using Registration Information

Create model function fetches the setting information from JSON object.

Usage Notes

If an algorithm is registered, user can create model using the registered algorithm name. Since all R scripts and default setting values are already registered, providing the value through the setting table is not necessary. This makes the use of this algorithm easier.

Examples

The first example builds a Classification model using the GLM algorithm.

```
CREATE TABLE GLM_RDEMO_SETTINGS_CL (

   setting_name  VARCHAR2(30),
   setting_value VARCHAR2(4000));
   BEGIN
        INSERT INTO GLM_RDEMO_SETTINGS_CL VALUES
         ('ALGO_EXTENSIBLE_LANG', 'R');
        INSERT INTO GLM_RDEMO_SETTINGS_CL VALUES
         (dbms_data_mining.ralg_registration_algo_name, 't1');
        INSERT INTO GLM_RDEMO_SETTINGS_CL VALUES
        (dbms_data_mining.odms_formula,
        'AGE + EDUCATION + HOUSEHOLD_SIZE + OCCUPATION');
        INSERT INTO GLM_RDEMO_SETTINGS_CL VALUES
         ('RALG_PARAMETER_FAMILY',   'binomial(logit)' );
   END;
   /
     BEGIN
          DBMS_DATA_MINING.CREATE_MODEL(
          model_name                    =>    'GLM_RDEMO_CLASSIFICATION',
          mining_function               =>     dbms_data_mining.classification,
          data_table_name               =>    'mining_data_build_v',
          case_id_column_name           =>    'CUST_ID',
          target_column_name            =>    'AFFINITY_CARD',
          settings_table_name           =>    'GLM_RDEMO_SETTINGS_CL');
      END;
      /
```

DROP_ALGORITHM Procedure

This function is used to drop the registered algorithm information.

Syntax

```
DBMS_DATA_MINING.DROP_ALGORITHM (algorithm_name  IN  VARCHAR2(30),
                                 cascade         IN  BOOLEAN default FALSE)
```

Parameters

Table 6-76 DROP_ALGORITHM Procedure Parameters

Parameter	Description
algorithm_name	Name of the algorithm.
cascade	If the cascade option is TRUE, all the models with this algorithms are forced to drop. There after, the algorithm is dropped. The default value is FALSE.

Usage Note

- To drop a machine learning model, you must be the owner or you must have the RQADMIN privilege. See *Oracle Machine Learning for SQL User's Guide* for information about privileges for machine learning.
- Make sure a model is not built on the algorithm, then drop the algorithm from the system table.

- If you try to drop an algorithm with a model built on it, then an error is displayed.

DROP_PARTITION Procedure

Syntax

```
DBMS_DATA_MINING.DROP_PARTITION (
      model_name                 IN VARCHAR2,
      partition_name             IN VARCHAR2);
```

Parameters

Table 6-77 DROP_PARTITION Procedure Parameters

Parameters	Description
`model_name`	Name of the machine learning model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Name of the partition that must be dropped.

DROP_MODEL Procedure

This procedure deletes the specified machine learning model.

Syntax

```
DBMS_DATA_MINING.DROP_MODEL (model_name IN VARCHAR2,
                             force      IN BOOLEAN DEFAULT FALSE);
```

Parameters

Table 6-78 DROP_MODEL Procedure Parameters

Parameter	Description
`model_name`	Name of the machine learning model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`force`	Forces the machine learning model to be dropped even if it is invalid. A machine learning model may be invalid if a serious system error interrupted the model build process.

Usage Note

To drop a machine learning model, you must be the owner or you must have the `DROP ANY MINING MODEL` privilege. See *Oracle Data Mining User's Guide* for information about privileges for Oracle Machine Learning for SQL.

Example

You can use the following command to delete a valid machine learning model named `nb_sh_clas_sample` that exists in your schema.

```
BEGIN
  DBMS_DATA_MINING.DROP_MODEL(model_name => 'nb_sh_clas_sample');
```

```
END;
/
```

EXPORT_MODEL Procedure

This procedure exports the specified machine learning models to a dump file set.

To import the models from the dump file set, use the IMPORT_MODEL Procedure. `EXPORT_MODEL` and `IMPORT_MODEL` use Oracle Data Pump technology.

When Oracle Data Pump is used to export/import an entire schema or database, the machine learning models in the schema or database are included. However, `EXPORT_MODEL` and `IMPORT_MODEL` are the only utilities that support the export/import of individual models.

> **See Also:**
>
> *Oracle Database Utilities* for information about Oracle Data Pump
>
> *Oracle Machine Learning for SQL User's Guide* for more information about exporting and importing machine learning models

Syntax

```
DBMS_DATA_MINING.EXPORT_MODEL (
      filename          IN VARCHAR2,
      directory         IN VARCHAR2,
      model_filter      IN VARCHAR2 DEFAULT NULL,
      filesize          IN VARCHAR2 DEFAULT NULL,
      operation         IN VARCHAR2 DEFAULT NULL,
      remote_link       IN VARCHAR2 DEFAULT NULL,
      jobname           IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-79 EXPORT_MODEL Procedure Parameters

Parameter	Description
`filename`	Name of the dump file set to which the models should be exported. The name must be unique within the schema. The dump file set can contain one or more files. The number of files in a dump file set is determined by the size of the models being exported (both metadata and data) and a specified or estimated maximum file size. You can specify the file size in the `filesize` parameter, or you can use the `operation` parameter to cause Oracle Data Pump to estimate the file size. If the size of the models to export is greater than the maximum file size, one or more additional files are created. When the export operation completes successfully, the name of the dump file set is automatically expanded to *`filename01.dmp`*, even if there is only one file in the dump set. If there are additional files, they are named sequentially as *`filename02.dmp`*, *`filename03.dmp`*, and so forth.

Table 6-79 (Cont.) EXPORT_MODEL Procedure Parameters

Parameter	Description
directory	Name of a pre-defined directory object that specifies where the dump file set should be created. The exporting user must have read/write privileges on the directory object and on the file system directory that it identifies. See *Oracle Database SQL Language Reference* for information about directory objects.
model_filter	Optional parameter that specifies which model or models to export. If you do not specify a value for model_filter, all models in the schema are exported. You can also specify NULL (the default) or 'ALL' to export all models. You can export individual models by name and groups of models based on machine learning function or algorithm. For instance, you could export all regression models or all Naive Bayes models. Examples are provided in Table 6-80.
filesize	Optional parameter that specifies the maximum size of a file in the dump file set. The size may be specified in bytes, kilobytes (K), megabytes (M), or gigabytes (G). The default size is 50 MB. If the size of the models to export is larger than filesize, one or more additional files are created within the dump set. See the description of the filename parameter for more information.
operation	Optional parameter that specifies whether or not to estimate the size of the files in the dump set. By default the size is not estimated and the value of the filesize parameter determines the size of the files. You can specify either of the following values for operation: • 'EXPORT' — Export all or the specified models. (Default) • 'ESTIMATE' — Estimate the size of the exporting models.
remote_link	Optional parameter that specifies the name of a database link to a remote system. The default value is NULL. A database link is a schema object in a local database that enables access to objects in a remote database. When you specify a value for remote_link, you can export the models in the remote database. The EXP_FULL_DATABASE role is required for exporting the remote models. The EXP_FULL_DATABASE privilege, the CREATE DATABASE LINK privilege, and other privileges may also be required.
jobname	Optional parameter that specifies the name of the export job. By default, the name has the form *username*_exp_*nnnn*, where *nnnn* is a number. For example, a job name in the SCOTT schema might be SCOTT_exp_134. If you specify a job name, it must be unique within the schema. The maximum length of the job name is 30 characters. A log file for the export job, named *jobname.log*, is created in the same directory as the dump file set.

Usage Notes

The model_filter parameter specifies which models to export. You can list the models by name, or you can specify all models that have the same machine learning function or algorithm. You can query the USER_MINING_MODELS view to list the models in your schema.

```
SQL> describe user_mining_models
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 MODEL_NAME                                NOT NULL VARCHAR2(30)
 MINING_FUNCTION                                    VARCHAR2(30)
```

```
ALGORITHM                                          VARCHAR2(30)
CREATION_DATE                             NOT NULL DATE
BUILD_DURATION                                     NUMBER
MODEL_SIZE                                         NUMBER
COMMENTS                                           VARCHAR2(4000)
```

Examples of model filters are provided in Table 6-80.

Table 6-80   Sample Values for the Model Filter Parameter

Sample Value	Meaning
`'mymodel'`	Export the model named `mymodel`
`'name= ''mymodel'''`	Export the model named `mymodel`
`'name IN (''mymodel2'',''mymodel3'')'`	Export the models named `mymodel2` and `mymodel3`
`'ALGORITHM_NAME = ''NAIVE_BAYES'''`	Export all Naive Bayes models. See Table 6-9 for a list of algorithm names.
`'FUNCTION_NAME =''CLASSIFICATION'''`	Export all classification models. See Table 6-7 for a list of machine learning functions.

Examples

1. The following statement exports all the models in the `oml_user3` schema to a dump file set called `models_out` in the directory `$ORACLE_HOME/rdbms/log`. This directory is mapped to a directory object called `DATA_PUMP_DIR`. The `oml_user3` user has read/write access to the directory and to the directory object.

   ```
   SQL>execute dbms_data_mining.export_model ('models_out', 'DATA_PUMP_DIR');
   ```

   You can exit SQL*Plus and list the resulting dump file and log file.

   ```
   SQL>EXIT
   >cd $ORACLE_HOME/rdbms/log
   >ls
   >oml_user3_exp_1027.log  models_out01.dmp
   ```

2. The following example uses the same directory object and is run by the same user. This example exports the models called `NMF_SH_SAMPLE` and `SVMR_SH_REGR_SAMPLE` to a different dump file set in the same directory.

   ```
   SQL>EXECUTE DBMS_DATA_MINING.EXPORT_MODEL ( 'models2_out', 'DATA_PUMP_DIR',
               'name in (''NMF_SH_SAMPLE'', ''SVMR_SH_REGR_SAMPLE'')');
   SQL>EXIT
   >cd $ORACLE_HOME/rdbms/log
   >ls
   >oml_user3_exp_1027.log  models_out01.dmp
    oml_user3_exp_924.log  models2_out01.dmp
   ```

3. The following examples show how to export models with specific algorithm and machine learning function names.

   ```
   SQL>EXECUTE DBMS_DATA_MINING.EXPORT_MODEL('algo.dmp','DM_DUMP',
           'ALGORITHM_NAME IN (''O_CLUSTER'',''GENERALIZED_LINEAR_MODEL'',
           ''SUPPORT_VECTOR_MACHINES'',''NAIVE_BAYES'')');

   SQL>EXECUTE DBMS_DATA_MINING.EXPORT_MODEL('func.dmp', 'DM_DUMP',
           'FUNCTION_NAME IN (CLASSIFICATION,CLUSTERING,FEATURE_EXTRACTION)');
   ```

EXPORT_SERMODEL Procedure

This procedure exports the model in a serialized format so that they can be moved to another platform for scoring.

When exporting a model in serialized format, the user must pass in an empty BLOB locator and specify the model name to be exported. If the model is partitioned, the user can optionally select an individual partition to export, otherwise all partitions are exported. The returned BLOB contains the content that can be deployed.

Syntax

```
DBMS_DATA_MINING.EXPORT_SERMODEL (
      model_data     IN OUT NOCOPY BLOB,
      model_name     IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-81 EXPORT_SERMODEL Procedure Parameters

Parameter	Description
model_data	Provides serialized model data.
model_name	Name of the machine learning model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
partition_name	Name of the partition that must be exported.

Examples

The following statement exports all of the models in a serialized format.

```
DECLARE
 v_blob blob;
BEGIN
 dbms_lob.createtemporary(v_blob, FALSE);
 dbms_data_mining.export_sermodel(v_blob, 'MY_MODEL');
-- save v_blob somewhere (e.g., bfile, etc.)
 dbms_lob.freetemporary(v_blob);
END;
/
```

See Also:

Oracle Machine Learning for SQL User's Guide for more information about exporting and importing machine learning models

FETCH_JSON_SCHEMA Procedure

User can fetch and read JSON schema from the `ALL_MINING_ALGORITHMS` view. This function returns the pre-registered JSON schema for R extensible algorithms.

Syntax

```
DBMS_DATA_MINING.FETCH_JSON_SCHEMA RETURN CLOB;
```

Parameters

Table 6-82 FETCH_JSON_SCHEMA Procedure Parameters

Parameter	Description
`RETURN`	This function returns the pre-registered JSON schema for R extensibility. The default value is `CLOB`.

Usage Note

If a user wants to register a new algorithm using the algorithm registration function, they must fetch and follow the pre-registered JSON schema using this function, when they create the required JSON object metadata, and then pass it to the registration function.

GET_ASSOCIATION_RULES Function

The `GET_ASSOCIATION_RULES` function returns the rules produced by an association model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

You can specify filtering criteria to `GET_ASSOCIATION_RULES` to return a subset of the rules. Filtering criteria can improve the performance of the table function. If the number of rules is large, then the greatest performance improvement will result from specifying the `topn` parameter.

Syntax

```
DBMS_DATA_MINING.get_association_rules(
      model_name       IN VARCHAR2,
      topn             IN NUMBER DEFAULT NULL,
      rule_id          IN INTEGER DEFAULT NULL,
      min_confidence   IN NUMBER DEFAULT NULL,
      min_support      IN NUMBER DEFAULT NULL,
      max_rule_length  IN INTEGER DEFAULT NULL,
      min_rule_length  IN INTEGER DEFAULT NULL,
      sort_order       IN ORA_MINING_VARCHAR2_NT DEFAULT NULL,
      antecedent_items IN DM_ITEMS DEFAULT NULL,
      consequent_items IN DM_ITEMS DEFAULT NULL,
      min_lift         IN NUMBER DEFAULT NULL,
      partition_name   IN VARCHAR2 DEFAULT NULL)
  RETURN DM_Rules PIPELINED;
```

Parameters

Table 6-83 GET_ASSOCIATION_RULES Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used. This is the only required parameter of `GET_ASSOCIATION_RULES`. All other parameters specify optional filters on the rules to return.
`topn`	Returns the *n* top rules ordered by confidence and then support, both descending. If you specify a sort order, then the top *n* rules are derived after the sort is performed. If `topn` is specified and no maximum or minimum rule length is specified, then the only columns allowed in the sort order are `RULE_CONFIDENCE` and `RULE_SUPPORT`. If `topn` is specified and a maximum or minimum rule length *is* specified, then `RULE_CONFIDENCE`, `RULE_SUPPORT`, and `NUMBER_OF_ITEMS` are allowed in the sort order.
`rule_id`	Identifier of the rule to return. If you specify a value for `rule_id`, do not specify values for the other filtering parameters.
`min_confidence`	Returns the rules with confidence greater than or equal to this number.
`min_support`	Returns the rules with support greater than or equal to this number.
`max_rule_length`	Returns the rules with a length less than or equal to this number. Rule length refers to the number of items in the rule (See `NUMBER_OF_ITEMS` in Table 6-84). For example, in the rule A=>B (if A, then B), the number of items is 2. If `max_rule_length` is specified, then the `NUMBER_OF_ITEMS` column is permitted in the sort order.
`min_rule_length`	Returns the rules with a length greater than or equal to this number. See `max_rule_length` for a description of rule length. If `min_rule_length` is specified, then the `NUMBER_OF_ITEMS` column is permitted in the sort order.
`sort_order`	Sorts the rules by the values in one or more of the returned columns. Specify one or more column names, each followed by `ASC` for ascending order or `DESC` for descending order. (See Table 6-84 for the column names.) For example, to sort the result set in descending order first by the `NUMBER_OF_ITEMS` column, then by the `RULE_CONFIDENCE` column, you must specify: `ORA_MINING_VARCHAR2_NT('NUMBER_OF_ITEMS DESC', 'RULE_CONFIDENCE DESC')` If you specify `topn`, the results will vary depending on the sort order. By default, the results are sorted by Confidence in descending order, then by Support in descending order.
`antecedent_items`	Returns the rules with these items in the antecedent.
`consequent_items`	Returns the rules with this item in the consequent.
`min_lift`	Returns the rules with lift greater than or equal to this number.
`partition_name`	Specifies a partition in a partitioned model.

Return Values

The object type returned by GET_ASSOCIATION_RULES is described in Table 6-84. For descriptions of each field, see the Usage Notes.

Table 6-84   GET_ASSOCIATION RULES Function Return Values

Return Value	Description

DM_RULES	A set of rows of type DM_RULE. The rows have the following columns:
	`(rule_id INTEGER,`
	` antecedent DM_PREDICATES,`
	` consequent DM_PREDICATES,`
	` rule_support NUMBER,`
	` rule_confidence NUMBER,`
	` rule_lift NUMBER,`
	` antecedent_support NUMBER,`
	` consequent_support NUMBER,`
	` number_of_items INTEGER )`

DM_PREDICATES	The antecedent and consequent columns each return nested tables of type DM_PREDICATES. The rows, of type DM_PREDICATE, have the following columns:
	`(attribute_name VARCHAR2(4000),`
	` attribute_subname VARCHAR2(4000),`
	` conditional_operator CHAR(2)/*=,<>,<,>,<=,>=*/,`
	` attribute_num_value NUMBER,`
	` attribute_str_value VARCHAR2(4000),`
	` attribute_support NUMBER,`
	` attribute_confidence NUMBER)`

Usage Notes

1. This table function pipes out rows of type DM_RULES. For information on machine learning data types and piped output from table functions, see "Datatypes".
2. The columns returned by GET_ASSOCIATION_RULES are described as follows:

Column in DM_RULES	Description
rule_id	Unique identifier of the rule

Column in DM_RULES	Description

antecedent	The independent condition in the rule. When this condition exists, the dependent condition in the consequent also exists.
	The condition is a combination of attribute values called a predicate (DM_PREDICATE). The predicate specifies a condition for each attribute. The condition may specify equality (=), inequality (<>), greater than (>), less than (<), greater than or equal to (>=), or less than or equal to (<=) a given value.
	Support and Confidence for each attribute condition in the antecedent is returned in the predicate. Support is the number of transactions that satisfy the antecedent. Confidence is the likelihood that a transaction will satisfy the antecedent.
	Note: The occurrence of the attribute as a DM_PREDICATE indicates the presence of the item in the transaction. The actual value for attribute_num_value or attribute_str_value is meaningless. For example, the following predicate indicates that 'Mouse Pad' is present in the transaction *even though* the attribute value is NULL.
	`DM_PREDICATE('PROD_NAME',`
	`                'Mouse Pad', '= ', NULL, NULL, NULL, NULL))`

consequent	The dependent condition in the rule. This condition exists when the antecedent exists. The consequent, like the antecedent, is a predicate (DM_PREDICATE). Support and confidence for each attribute condition in the consequent is returned in the predicate. Support is the number of transactions that satisfy the consequent. Confidence is the likelihood that a transaction will satisfy the consequent.
rule_support	The number of transactions that satisfy the rule.
rule_confidence	The likelihood of a transaction satisfying the rule.
rule_lift	The degree of improvement in the prediction over random chance when the rule is satisfied.
antecedent_support	The ratio of the number of transactions that satisfy the antecedent to the total number of transactions.
consequent_support	The ratio of the number of transactions that satisfy the consequent to the total number of transactions.
number_of_items	The total number of attributes referenced in the antecedent and consequent of the rule.

Examples

The following example demonstrates an association model build followed by several invocations of the GET_ASSOCIATION_RULES table function:

```
-- prepare a settings table to override default settings
CREATE TABLE market_settings AS
SELECT *
  FROM TABLE(DBMS_DATA_MINING.GET_DEFAULT_SETTINGS)
 WHERE setting_name LIKE 'ASSO_%';
BEGIN
-- update the value of the minimum confidence
UPDATE market_settings
   SET setting_value = TO_CHAR(0.081)
 WHERE setting_name = DBMS_DATA_MINING.asso_min_confidence;

-- build an AR model
```

```
DBMS_DATA_MINING.CREATE_MODEL(
  model_name => 'market_model',
  function => DBMS_DATA_MINING.ASSOCIATION,
  data_table_name => 'market_build',
  case_id_column_name => 'item_id',
  target_column_name => NULL,
  settings_table_name => 'market_settings');
END;
/
-- View the (unformatted) rules
SELECT rule_id, antecedent, consequent, rule_support,
       rule_confidence
  FROM TABLE(DBMS_DATA_MINING.GET_ASSOCIATION_RULES('market_model'));
```

In the previous example, you view all rules. To view just the top 20 rules, use the following statement.

```
-- View the top 20 (unformatted) rules
SELECT rule_id, antecedent, consequent, rule_support,
       rule_confidence
  FROM TABLE(DBMS_DATA_MINING.GET_ASSOCIATION_RULES('market_model', 20));
```

The following query uses the association model `AR_SH_SAMPLE`.

```
SELECT * FROM TABLE (
   DBMS_DATA_MINING.GET_ASSOCIATION_RULES (
      'AR_SH_SAMPLE', 10, NULL, 0.5, 0.01, 2, 1,
         ORA_MINING_VARCHAR2_NT (
         'NUMBER_OF_ITEMS DESC', 'RULE_CONFIDENCE DESC', 'RULE_SUPPORT DESC'),
         DM_ITEMS(DM_ITEM('CUSTPRODS', 'Mouse Pad', 1, NULL),
                  DM_ITEM('CUSTPRODS', 'Standard Mouse', 1, NULL)),
         DM_ITEMS(DM_ITEM('CUSTPRODS', 'Extension Cable', 1, NULL))));
```

The query returns three rules, shown as follows:

```
13  DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Mouse Pad', '= ', 1, NULL, NULL, NULL),
       DM_PREDICATE('CUSTPRODS', 'Standard Mouse', '= ', 1, NULL, NULL, NULL))
    DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Extension Cable', '= ', 1, NULL, NULL, NULL))
    .15532      .84393   2.7075     .18404     .3117   2

11  DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Standard Mouse', '= ', 1, NULL, NULL, NULL))
    DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Extension Cable', '= ', 1, NULL, NULL, NULL))
    .18085      .56291   1.8059     .32128     .3117   1

9   DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Mouse Pad', '= ', 1, NULL, NULL, NULL))
    DM_PREDICATES(
       DM_PREDICATE('CUSTPRODS', 'Extension Cable', '= ', 1, NULL, NULL, NULL))
      .17766    .55116   1.7682     .32234     .3117   1
```

See Also:

Table 6-84 for the DM_RULE column data types.

GET_FREQUENT_ITEMSETS Function

The GET_FREQUENT_ITEMSETS function returns a set of rows that represent the frequent itemsets from an association model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead..

See Model Detail Views in *Oracle Machine Learning for SQL User's Guide*.

For a detailed description of frequent itemsets, consult *Oracle Machine Learning for SQL Concepts*.

Syntax

```
DBMS_DATA_MINING.get_frequent_itemsets(
      model_name IN VARCHAR2,
      topn IN NUMBER DEFAULT NULL,
      max_itemset_length IN NUMBER DEFAULT NULL,
      partition_name     IN VARCHAR2 DEFAULT NULL)
  RETURN DM_ItemSets PIPELINED;
```

Parameters

Table 6-85 GET_FREQUENT_ITEMSETS Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
topn	When not NULL, return the top *n* rows ordered by support in descending order
max_itemset_length	Maximum length of an item set.
partition_name	Specifies a partition in a partitioned model.

Note:

The partition_name columns applies only when the model is partitioned.

Return Values

Table 6-86 GET_FREQUENT_ITEMSETS Function Return Values

Return Value	Description

`DM_ITEMSETS`	A set of rows of type `DM_ITEMSET`. The rows have the following columns:
	`(partition_name VARCHAR2(128)`
	`itemsets_id NUMBER,`
	`items DM_ITEMS,`
	`support NUMBER,`
	`number_of_items NUMBER)`
	Note: The `partition_name` columns applies only when the model is partitioned.
	The `items` column returns a nested table of type `DM_ITEMS`. The rows have type `DM_ITEM`:
	`(attribute_name VARCHAR2(4000),`
	`attribute_subname VARCHAR2(4000),`
	`attribute_num_value NUMBER,`
	`attribute_str_value VARCHAR2(4000))`

Usage Notes

This table function pipes out rows of type `DM_ITEMSETS`. For information on machine learning data types and piped output from table functions, see "Data Types".

Examples

The following example demonstrates an association model build followed by an invocation of `GET_FREQUENT_ITEMSETS` table function from Oracle SQL.

```
-- prepare a settings table to override default settings
CREATE TABLE market_settings AS

    SELECT *

  FROM TABLE(DBMS_DATA_MINING.GET_DEFAULT_SETTINGS)
 WHERE setting_name LIKE 'ASSO_%';
BEGIN
-- update the value of the minimum confidence
UPDATE market_settings
   SET setting_value = TO_CHAR(0.081)
 WHERE setting_name = DBMS_DATA_MINING.asso_min_confidence;

/* build a AR model */
DBMS_DATA_MINING.CREATE_MODEL(
  model_name           => 'market_model',
  function             => DBMS_DATA_MINING.ASSOCIATION,
  data_table_name      => 'market_build',
  case_id_column_name  => 'item_id',
  target_column_name   => NULL,
  settings_table_name  => 'market_settings');
END;
/

-- View the (unformatted) Itemsets from SQL*Plus
SELECT itemset_id, items, support, number_of_items
  FROM TABLE(DBMS_DATA_MINING.GET_FREQUENT_ITEMSETS('market_model'));
```

In the example above, you view all itemsets. To view just the top 20 itemsets, use the following statement:

```
-- View the top 20 (unformatted) Itemsets from SQL*Plus
SELECT itemset_id, items, support, number_of_items
  FROM TABLE(DBMS_DATA_MINING.GET_FREQUENT_ITEMSETS('market_model', 20));
```

GET_MODEL_COST_MATRIX Function

The GET_* interfaces are replaced by model views, and Oracle recommends that users leverage the views instead.

The GET_MODEL_COST_MATRIX function is replaced by the DM$VC prefixed view, Scoring Cost Matrix. The cost matrix used when building a Decision Tree is made available by the DM$VM prefixed view, Decision Tree build cost matrix.

Refer to Model Detail View for Classification Algorithm.

The GET_MODEL_COST_MATRIX function returns the rows of a cost matrix associated with the specified model.

By default, this function returns the scoring cost matrix that was added to the model with the ADD_COST_MATRIX procedure. If you wish to obtain the cost matrix used to create a model, specify cost_matrix_type_create as the matrix_type. See Table 6-87.

See also ADD_COST_MATRIX Procedure.

Syntax

```
DBMS_DATA_MINING.GET_MODEL_COST_MATRIX (
      model_name                IN VARCHAR2,
      matrix_type               IN VARCHAR2 DEFAULT cost_matrix_type_score)
      partition_name            IN VARCHAR2 DEFAULT NULL);
RETURN DM_COST_MATRIX PIPELINED;
```

Parameters

Table 6-87   GET_MODEL_COST_MATRIX Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
matrix_type	The type of cost matrix. COST_MATRIX_TYPE_SCORE — cost matrix used for scoring. (Default.) COST_MATRIX_TYPE_CREATE — cost matrix used to create the model (Decision Tree only).
partition_name	Name of the partition in a partitioned model

Return Values

Table 6-88 GET_MODEL_COST_MATRIX Function Return Values

Return Value	Description
DM_COST_MATRIX	A set of rows of type DM_COST_ELEMENT. The rows have the following columns: actual VARCHAR2(4000), NUMBER, predicted VARCHAR2(4000), cost NUMBER)

Usage Notes

Only Decision Tree models can be built with a cost matrix. If you want to build a Decision Tree model with a cost matrix, specify the cost matrix table name in the CLAS_COST_TABLE_NAME setting in the settings table for the model. See Table 6-11.

The cost matrix used to create a Decision Tree model becomes the default scoring matrix for the model. If you want to specify different costs for scoring, you can use the REMOVE_COST_MATRIX procedure to remove the cost matrix and the ADD_COST_MATRIX procedure to add a new one.

The GET_MODEL_COST_MATRIX may return either the build or scoring cost matrix defined for a model or model partition.

If you do not specify a partitioned model name, then an error is displayed.

Example

This example returns the scoring cost matrix associated with the Naive Bayes model NB_SH_CLAS_SAMPLE.

```
column actual format a10
column predicted format a10
SELECT *
    FROM TABLE(dbms_data_mining.get_model_cost_matrix('nb_sh_clas_sample'))
    ORDER BY predicted, actual;

ACTUAL     PREDICTED   COST
---------- ---------- -----
0          0            .00
1          0            .75
0          1            .25
1          1            .00
```

GET_MODEL_DETAILS_AI Function

The GET_MODEL_DETAILS_AI function returns a set of rows that provide the details of an attribute importance model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User's Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_ai(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN dm_ranked_attributes pipelined;
```

Parameters

Table 6-89 GET_MODEL_DETAILS_AI Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model.

Return Values

Table 6-90 GET_MODEL_DETAILS_AI Function Return Values

Return Value	Description

`DM_RANKED_ATTRIBUTES`	A set of rows of type `DM_RANKED_ATTRIBUTE`. The rows have the following columns:
	`(attribute_name        VARCHAR2(4000,`
	` attribute_subname     VARCHAR2(4000),`
	` importance_value      NUMBER,`
	` rank                  NUMBER(38))`

Examples

The following example returns model details for the attribute importance model `AI_SH_sample`, which was created by the sample program `dmaidemo.sql`.

```
SELECT attribute_name, importance_value, rank
    FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_AI('AI_SH_sample'))
    ORDER BY RANK;

ATTRIBUTE_NAME                           IMPORTANCE_VALUE       RANK
---------------------------------------- ---------------- ----------
HOUSEHOLD_SIZE                                 .151685183          1
CUST_MARITAL_STATUS                            .145294546          2
YRS_RESIDENCE                                   .07838928          3
AGE                                            .075027496          4
Y_BOX_GAMES                                    .063039952          5
EDUCATION                                      .059605314          6
HOME_THEATER_PACKAGE                           .056458722          7
OCCUPATION                                     .054652937          8
CUST_GENDER                                    .035264741          9
BOOKKEEPING_APPLICATION                        .019204751         10
PRINTER_SUPPLIES                                        0         11
OS_DOC_SET_KANJI                               -.00050013         12
FLAT_PANEL_MONITOR                             -.00509564         13
BULK_PACK_DISKETTES                            -.00540822         14
COUNTRY_NAME                                   -.01201116         15
CUST_INCOME_LEVEL                              -.03951311         16
```

GET_MODEL_DETAILS_EM Function

The GET_MODEL_DETAILS_EM function returns a set of rows that provide statistics about the clusters produced by an expectation maximization model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User's Guide*.

By default, the EM algorithm groups components into high-level clusters, and GET_MODEL_DETAILS_EM returns only the high-level clusters with their hierarchies. Alternatively, you can configure EM model to disable the grouping of components into high-level clusters. In this case, GET_MODEL_DETAILS_EM returns the components themselves as clusters with their hierarchies. See Table 6-16.

Syntax

```
DBMS_DATA_MINING.get_model_details_em(
      model_name VARCHAR2,
      cluster_id NUMBER   DEFAULT NULL,
      attribute  VARCHAR2 DEFAULT NULL,
      centroid   NUMBER   DEFAULT 1,
      histogram  NUMBER   DEFAULT 1,
      rules      NUMBER   DEFAULT 2,
      attribute_subname  VARCHAR2 DEFAULT NULL,
      topn_attributes NUMBER DEFAULT NULL,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN dm_clusters PIPELINED;
```

Parameters

Table 6-91 GET_MODEL_DETAILS_EM Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
cluster_id	The ID of a cluster in the model. When a valid cluster ID is specified, only the details of this cluster are returned. Otherwise, the details for all clusters are returned.
attribute	The name of an attribute. When a valid attribute name is specified, only the details of this attribute are returned. Otherwise, the details for all attributes are returned
centroid	This parameter accepts the following values: • 1: Details about centroids are returned (default) • 0: Details about centroids are not returned
histogram	This parameter accepts the following values: • 1: Details about histograms are returned (default) • 0: Details about histograms are not returned
rules	This parameter accepts the following values: • 2: Details about rules are returned (default) • 1: Rule summaries are returned • 0: No information about rules is returned

Table 6-91 (Cont.) GET_MODEL_DETAILS_EM Function Parameters

Parameter	Description
`attribute_subname`	The name of a nested attribute. The full name of a nested attribute has the form: *`attribute_name.attribute_subname`* where *`attribute_name`* is the name of the column and *`attribute_subname`* is the name of the nested attribute in that column. If the attribute is not nested, then `attribute_subname` is null.
`topn_attributes`	Restricts the number of attributes returned in the centroid, histogram, and rules objects. Only the *`n`* attributes with the highest confidence values in the rules are returned. If the number of attributes included in the rules is less than *`topn`*, then, up to *`n`* additional attributes in alphabetical order are returned. If both the `attribute` and `topn_attributes` parameters are specified, then `topn_attributes` is ignored.
`partition_name`	Specifies a partition in a partitioned model.

Usage Notes

1. For information on Oracle Machine Learning for SQL data types and return values for Clustering algorithms piped output from table functions, see "Data Types".
2. `GET_MODEL_DETAILS` functions preserve model transparency by automatically reversing the transformations applied during the build process. Thus the attributes returned in the model details are the original attributes (or a close approximation of the original attributes) used to build the model.
3. When cluster statistics are disabled (`EMCS_CLUSTER_STATISTICS` is set to `EMCS_CLUS_STATS_DISABLE`), `GET_MODEL_DETAILS_EM` does not return centroids, histograms, or rules. Only taxonomy (hierarchy) and cluster counts are returned.
4. When the `partition_name` is `NULL` for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

GET_MODEL_DETAILS_EM_COMP Function

he `GET_MODEL_DETAILS_EM_COMP` table function returns a set of rows that provide details about the parameters of an expectation maximization model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_em_comp(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_EM_COMPONENT_SET PIPELINED;
```

Parameters

Table 6-92 GET_MODEL_DETAILS_EM_COMP Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model to retrieve details for.

Return Values

Table 6-93 GET_MODEL_DETAILS_EM_COMP Function Return Values

Return Value	Description

`DM_EM_COMPONENT_SET`	A set of rows of type `DM_EM_COMPONENT`. The rows have the following columns:
	`(info_type           VARCHAR2(30),`
	` component_id        NUMBER,`
	` cluster_id          NUMBER,`
	` attribute_name      VARCHAR2(4000),`
	` covariate_name      VARCHAR2(4000),`
	` attribute_value     VARCHAR2(4000),`
	` value               NUMBER )`

Usage Notes

1. This table function pipes out rows of type `DM_EM_COMPONENT`. For information on Oracle Machine Learning for SQL data types and piped output from table functions, see "Data Types".

   The columns in each row returned by `GET_MODEL_DETAILS_EM_COMP` are described as follows:

Column in DM_EM_COMPONENT	Description
`info_type`	The type of information in the row. The following information types are supported: • `cluster` • `prior` • `mean` • `covariance` • `frequency`
`component_id`	Unique identifier of a component
`cluster_id`	Unique identifier of the high-level leaf cluster for each component
`attribute_name`	Name of an original attribute or a derived feature ID. The derived feature ID is used in models built on data with nested columns. The derived feature definitions can be obtained from the GET_MODEL_DETAILS_EM_PROJ Function.

Column in DM_EM_COMPONENT	Description
covariate_name	Name of an original attribute or a derived feature ID used in variance/covariance definition
attribute_value	Categorical value or bin interval for binned numerical attributes
value	Encodes different information depending on the value of info_type, as follows: • cluster — The value field is NULL • prior — The value field returns the component prior • mean — The value field returns the mean of the attribute specified in attribute_name • covariance — The value field returns the covariance of the attributes specified in attribute_name and covariate_name. Using the same attribute in attribute_name and covariate_name, returns the variance. • frequency— The value field returns the multivalued Bernoulli frequency parameter for the attribute/value combination specified by attribute_name and attribute_value See Usage Note 2 for details.

2. The following table shows which fields are used for each info_type. The blank cells represent NULLs.

info_type	component_id	cluster_id	attribute_name	covariate_name	attribute_value	value
cluster	X	X				
prior	X	X				X
mean	X	X	X			X
covariance	X	X	X	X		X
frequency	X	X	X		X	X

3. GET_MODEL_DETAILS functions preserve model transparency by automatically reversing the transformations applied during the build process. Thus the attributes returned in the model details are the original attributes (or a close approximation of the original attributes) used to build the model.

4. When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

GET_MODEL_DETAILS_EM_PROJ Function

The GET_MODEL_DETAILS_EM_PROJ function returns a set of rows that provide statistics about the projections produced by an expectation maximization model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_em_proj(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_EM_PROJECTION_SET PIPELINED;
```

Parameters

Table 6-94 GET_MODEL_DETAILS_EM_PROJ Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model

Return Values

Table 6-95 GET_MODEL_DETAILS_EM_PROJ Function Return Values

Return Value	Description

`DM_EM_PROJECTION_SET`	A set of rows of type `DM_EM_PROJECTION`. The rows have the following columns:
	`(feature_name VARCHAR2(4000),`
	` attribute_name VARCHAR2(4000),`
	` attribute_subname VARCHAR2(4000),`
	` attribute_value VARCHAR2(4000),`
	` coefficient NUMBER )`
	See Usage Notes for details.

Usage Notes

1. This table function pipes out rows of type `DM_EM_PROJECTION`. For information on machine learning data types and piped output from table functions, see "Datatypes".

   The columns in each row returned by `GET_MODEL_DETAILS_EM_PROJ` are described as follows:

Column in DM_EM_PROJECTION	Description
`feature_name`	Name of a derived feature. The feature maps to the attribute_name returned by the GET_MODEL_DETAILS_EM Function.
`attribute_name`	Name of a column in the build data
`attribute_subname`	Subname in a nested column
`attribute_value`	Categorical value
`coefficient`	Projection coefficient. The representation is sparse; only the non-zero coefficients are returned.

2. `GET_MODEL_DETAILS` functions preserve model transparency by automatically reversing the transformations applied during the build process. Thus the attributes returned in the model details are the original attributes (or a close approximation of the original attributes) used to build the model.

   The coefficients are related to the transformed, not the original, attributes. When returned directly with the model details, the coefficients may not provide meaningful information.

3. When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_GLM Function

The GET_MODEL_DETAILS_GLM function returns the coefficient statistics for a generalized linear model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

The same set of statistics is returned for both linear and logistic regression, but statistics that do not apply to the machine learning function are returned as NULL. For more details, see the Usage Notes.

Syntax

```
DBMS_DATA_MINING.get_model_details_glm(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_GLM_Coeff_Set PIPELINED;
```

Parameters

Table 6-96 GET_MODEL_DETAILS_GLM Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
partition_name	Specifies a partition in a partitioned model

Return Values

Table 6-97 GET_MODEL_DETAILS_GLM Return Values

Return Value	Description

`DM_GLM_COEFF_SET`	A set of rows of type `DM_GLM_COEFF`. The rows have the following columns:
	`(class                VARCHAR2(4000),`
	` attribute_name        VARCHAR2(4000),`
	` attribute_subname     VARCHAR2(4000),`
	` attribute_value       VARCHAR2(4000),`
	` feature_expression    VARCHAR2(4000),`
	` coefficient           NUMBER,`
	` std_error             NUMBER,`
	` test_statistic        NUMBER,`
	` p_value               NUMBER,`
	` VIF                   NUMBER,`
	` std_coefficient       NUMBER,`
	` lower_coeff_limit     NUMBER,`
	` upper_coeff_limit     NUMBER,`
	` exp_coefficient       BINARY_DOUBLE,`
	` exp_lower_coeff_limit BINARY_DOUBLE,`
	` exp_upper_coeff_limit BINARY_DOUBLE)`

`GET_MODEL_DETAILS_GLM` returns a row of statistics for each attribute and one extra row for the intercept, which is identified by a null value in the attribute name. Each row has the `DM_GLM_COEFF` data type. The statistics are described in Table 6-98.

Table 6-98 DM_GLM_COEFF Data Type Description

Column	Description
`class`	The non-reference target class for logistic regression. The model is built to predict the probability of this class. The other class (the reference class) is specified in the model setting `GLMS_REFERENCE_CLASS_NAME`. See Table 6-23. For Linear Regression, `class` is null.
`attribute_name`	The attribute name when there is no subname, or first part of the attribute name when there is a subname. The value of `attribute_name` is also the name of the column in the case table that is the source for this attribute. For the intercept, `attribute_name` is null. Intercepts are equivalent to the bias term in SVM models.
`attribute_subname`	The name of an attribute in a nested table. The full name of a nested attribute has the form: *`attribute_name.attribute_subname`* where *`attribute_name`* is the name of the nested column in the case table that is the source for this attribute. If the attribute is not nested, then `attribute_subname` is null. If the attribute is an intercept, then both the `attribute_name` and the `attribute_subname` are null.

Table 6-98 (Cont.) DM_GLM_COEFF Data Type Description

Column	Description
`attribute_value`	The value of the attribute (categorical attribute only). For numeric attributes, `attribute_value` is null.
`feature_expression`	The feature name constructed by the algorithm when feature generation is enabled and higher-order features are found. If feature selection is not enabled, then the feature name is simply the fully-qualified attribute name (*`attribute_name.attribute_subname`* if the attribute is in a nested column). For categorical attributes, the algorithm constructs a feature name that has the following form: *`fully-qualified_attribute_name.attribute_value`* For numeric attributes, the algorithm constructs a name for the higher-order feature by taking the product of the resulting values: (*`attrib1`*)*(*`attrib2`*))*...... where *`attrib1`* and *`attrib2`* are fully-qualified attribute names.
`coefficient`	The linear coefficient estimate.
`std_error`	Standard error of the coefficient estimate.
`test_statistic`	For linear regression, the t-value of the coefficient estimate. For logistic regression, the Wald chi-square value of the coefficient estimate.
`p-value`	Probability of the `test_statistic`. Used to analyze the significance of specific attributes in the model.
`VIF`	Variance Inflation Factor. The value is zero for the intercept. For logistic regression, `VIF` is null. VIF is not computed if the solver is Cholesky.
`std_coefficient`	Standardized estimate of the coefficient.
`lower_coeff_limit`	Lower confidence bound of the coefficient.
`upper_coeff_limit`	Upper confidence bound of the coefficient.
`exp_coefficient`	Exponentiated coefficient for logistic regression. For linear regression, `exp_coefficient` is null.
`exp_lower_coeff_limit`	Exponentiated coefficient for lower confidence bound of the coefficient for logistic regression. For linear regression, `exp_lower_coeff_limit` is null.
`exp_upper_coeff_limit`	Exponentiated coefficient for upper confidence bound of the coefficient for logistic regression. For linear regression, `exp_lower_coeff_limit` is null.

Usage Notes

Not all statistics are necessarily returned for each coefficient. Statistics will be null if:

- They do not apply to the machine learning function. For example, `exp_coefficient` does not apply to linear regression.
- They cannot be computed from a theoretical standpoint. For information on ridge regression, see Table 6-23.
- They cannot be computed because of limitations in system resources.
- Their values would be infinity.

- When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following example returns some of the model details for the GLM regression model `GLMR_SH_Regr_sample`.

```
SET line 120
SET pages 99
column attribute_name format a30
column attribute_subname format a20
column attribute_value format a20
col coefficient format 990.9999
col std_error format 990.9999
SQL> SELECT * FROM
(SELECT attribute_name, attribute_value, coefficient, std_error
  FROM DM$VDGLMR_SH_REGR_SAMPLE order by 1,2)
WHERE rownum < 11;

ATTRIBUTE_NAME                 ATTRIBUTE_VALUE     COEFFICIENT     STD_ERROR
------------------------------ -------------------- ----------- ---------
AFFINITY_CARD                                         -0.5797          0.5283
BOOKKEEPING_APPLICATION                             -0.4689          3.8872
BULK_PACK_DISKETTES                                 -0.9819          2.5430
COUNTRY_NAME                     Argentina              -1.2020          1.1876
COUNTRY_NAME                     Australia              -0.0071          5.1146
COUNTRY_NAME                     Brazil                  5.2931          1.9233
COUNTRY_NAME                     Canada                   4.0191          2.4108
COUNTRY_NAME                     China                   0.8706          3.5889
COUNTRY_NAME                     Denmark              -2.9822           3.1803
COUNTRY_NAME                     France                  -1.1044          7.1811
```

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_GLOBAL Function

The `GET_MODEL_DETAILS_GLOBAL` function returns statistics about the model as a whole. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

Global details are available for Generalized Linear Models, Association Rules, Singular Value Decomposition, and Expectation Maximization. There are new Global model views which show global information for all algorithms. Oracle recommends that users leverage the views instead. Refer to Model Details View Global.

Syntax

```
DBMS_DATA_MINING.get_model_details_global(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_model_global_details PIPELINED;
```

Parameters

Table 6-99 GET_MODEL_DETAILS_GLOBAL Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model.

Return Values

Table 6-100 GET_MODEL_DETAILS_GLOBAL Function Return Values

Return Value	Description

`DM_MODEL_GLOBAL_DETAILS`	A collection of rows of type `DM_MODEL_GLOBAL_DETAIL`. The rows have the following columns:
	`(global_detail_name   VARCHAR2(30),`
	` global_detail_value   NUMBER)`

Examples

The following example returns the global model details for the GLM regression model `GLMR_SH_Regr_sample`.

```
SELECT *
  FROM TABLE(dbms_data_mining.get_model_details_global(
              'GLMR_SH_Regr_sample'))
ORDER BY global_detail_name;
GLOBAL_DETAIL_NAME             GLOBAL_DETAIL_VALUE
------------------------------ -------------------
ADJUSTED_R_SQUARE                       .731412557
AIC                                       5931.814
COEFF_VAR                               18.1711243
CORRECTED_TOTAL_DF                            1499
CORRECTED_TOT_SS                        278740.504
DEPENDENT_MEAN                              38.892
ERROR_DF                                      1433
ERROR_MEAN_SQUARE                       49.9440956
ERROR_SUM_SQUARES                       71569.8891
F_VALUE                                 62.8492452
GMSEP                                    52.280819
HOCKING_SP                              .034877162
J_P                                     52.1749319
MODEL_CONVERGED                                  1
MODEL_DF                                        66
MODEL_F_P_VALUE                                  0
MODEL_MEAN_SQUARE                       3138.94871
MODEL_SUM_SQUARES                       207170.615
NUM_PARAMS                                      67
NUM_ROWS                                      1500
ROOT_MEAN_SQ                            7.06711367
R_SQ                                    .743238288
SBIC                                    6287.79977
VALID_COVARIANCE_MATRIX                          1
```

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_KM Function

The `GET_MODEL_DETAILS_KM` function returns a set of rows that provide the details of a *k*-means clustering model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

You can provide input to `GET_MODEL_DETAILS_KM` to request specific information about the model, thus improving the performance of the query. If you do not specify filtering parameters, then `GET_MODEL_DETAILS_KM` returns all the information about the model.

Syntax

```
DBMS_DATA_MINING.get_model_details_km(
      model_name VARCHAR2,
      cluster_id NUMBER   DEFAULT NULL,
      attribute  VARCHAR2 DEFAULT NULL,
      centroid   NUMBER   DEFAULT 1,
      histogram  NUMBER   DEFAULT 1,
      rules      NUMBER   DEFAULT 2,
      attribute_subname  VARCHAR2 DEFAULT NULL,
      topn_attributes NUMBER DEFAULT NULL,
      partition_name VARCHAR2 DEFAULT NULL)
  RETURN dm_clusters PIPELINED;
```

Parameters

Table 6-101 GET_MODEL_DETAILS_KM Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`cluster_id`	The ID of a cluster in the model. When a valid cluster ID is specified, only the details of this cluster are returned. Otherwise the details for all clusters are returned.
`attribute`	The name of an attribute. When a valid attribute name is specified, only the details of this attribute are returned. Otherwise, the details for all attributes are returned
`centroid`	This parameter accepts the following values: • 1: Details about centroids are returned (default) • 0: Details about centroids are not returned
`histogram`	This parameter accepts the following values: • 1: Details about histograms are returned (default) • 0: Details about histograms are not returned
`rules`	This parameter accepts the following values: • 2: Details about rules are returned (default) • 1: Rule summaries are returned • 0: No information about rules is returned

Table 6-101 (Cont.) GET_MODEL_DETAILS_KM Function Parameters

Parameter	Description
`attribute_subname`	The name of a nested attribute. The full name of a nested attribute has the form: *`attribute_name.attribute_subname`* where *`attribute_name`* is the name of the column and *`attribute_subname`* is the name of the nested attribute in that column. If the attribute is not nested, `attribute_subname` is null.
`topn_attributes`	Restricts the number of attributes returned in the centroid, histogram, and rules objects. Only the *`n`* attributes with the highest confidence values in the rules are returned. If the number of attributes included in the rules is less than *`topn`*, then up to *`n`* additional attributes in alphabetical order are returned. If both the `attribute` and `topn_attributes` parameters are specified, then `topn_attributes` is ignored.
`partition_name`	Specifies a partition in a partitioned model.

Usage Notes

1. The table function pipes out rows of type `DM_CLUSTERS`. For information on machine learning data types and Return Value for Clustering Algorithms piped output from table functions, see "Data Types".
2. When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following example returns model details for the *k*-means clustering model `KM_SH_Clus_sample`.

```
SELECT T.id           clu_id,
       T.record_count rec_cnt,
       T.parent       parent,
       T.tree_level   tree_level,
       T.dispersion   dispersion
  FROM (SELECT *
          FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_KM(
                     'KM_SH_Clus_sample'))
        ORDER BY id) T
 WHERE ROWNUM < 6;

    CLU_ID    REC_CNT     PARENT TREE_LEVEL DISPERSION
---------- ---------- ---------- ---------- ----------
         1       1500                     1  5.9152211
         2        638          1          2 3.98458982
         3        862          1          2 5.83732097
         4        376          3          3 5.05192137
         5        486          3          3 5.42901522
```

Related Topics

- *Oracle Machine Learning for SQL User's Guide*

GET_MODEL_DETAILS_NB Function

The GET_MODEL_DETAILS_NB function returns a set of rows that provide the details of a naive Bayes model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User's Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_nb(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_NB_Details PIPELINED;
```

Parameters

Table 6-102 GET_MODEL_DETAILS_NB Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
partition_name	Specifies a partition in a partitioned model

Return Values

Table 6-103 GET_MODEL_DETAILS_NB Function Return Values

Return Value	Description
DM_NB_DETAILS	A set of rows of type DM_NB_DETAIL. The rows have the following columns: (target_attribute_name VARCHAR2(30), target_attribute_str_value VARCHAR2(4000), target_attribute_num_value NUMBER, prior_probability NUMBER, conditionals DM_CONDITIONALS) The conditionals column of DM_NB_DETAIL returns a nested table of type DM_CONDITIONALS. The rows, of type DM_CONDITIONAL, have the following columns: (attribute_name VARCHAR2(4000), attribute_subname VARCHAR2(4000), attribute_str_value VARCHAR2(4000), attribute_num_value NUMBER, conditional_probability NUMBER)

Usage Notes

- The table function pipes out rows of type DM_NB_DETAILS. For information on machine learning data types and piped output from table functions, see "Data Types".
- When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following query is from the sample program `dmnbdemo.sql`. It returns model details about the model `NB_SH_Clas_sample`. For information about the sample programs, see *Oracle Machine Learning for SQL User's Guide*.

The query creates labels from the bin boundary tables that were used to bin the training data. It replaces the attribute values with the labels. For numeric bins, the labels are `(`*`lower_boundary`*`,`*`upper_boundary`*`]`; for categorical bins, the label matches the value it represents. (This method of categorical label representation will only work for cases where one value corresponds to one bin.) The target was not binned.

```
WITH
    bin_label_view AS (
    SELECT col, bin, (DECODE(bin,'1','[','(') || lv || ',' || val || ']') label
      FROM (SELECT col,
                   bin,
                   LAST_VALUE(val) OVER (
                   PARTITION BY col ORDER BY val
                   ROWS BETWEEN UNBOUNDED PRECEDING AND 1 PRECEDING) lv,
                   val
             FROM nb_sh_sample_num)
   UNION ALL
   SELECT col, bin, val label
     FROM nb_sh_sample_cat
   ),
   model_details AS (
   SELECT T.target_attribute_name                                        tname,
          NVL(TO_CHAR(T.target_attribute_num_value,T.target_attribute_str_value)) tval,
          C.attribute_name                                               pname,
          NVL(L.label, NVL(C.attribute_str_value, C.attribute_num_value)) pval,
          T.prior_probability                                           priorp,
          C.conditional_probability                                      condp
     FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_NB('NB_SH_Clas_sample')) T,
          TABLE(T.conditionals) C,
          bin_label_view L
    WHERE C.attribute_name = L.col (+) AND
          (NVL(C.attribute_str_value,C.attribute_num_value) = L.bin(+))
   ORDER BY 1,2,3,4,5,6
   )
   SELECT tname, tval, pname, pval, priorp, condp
     FROM model_details
    WHERE ROWNUM < 11;
```

```
TNAME          TVAL PNAME                     PVAL           PRIORP   CONDP
-------------- ---- ------------------------- ------------- ------- -------
AFFINITY_CARD  0    AGE                       (24,30]         .6500   .1714
AFFINITY_CARD  0    AGE                       (30,35]         .6500   .1509
AFFINITY_CARD  0    AGE                       (35,40]         .6500   .1125
AFFINITY_CARD  0    AGE                       (40,46]         .6500   .1134
AFFINITY_CARD  0    AGE                       (46,53]         .6500   .1071
AFFINITY_CARD  0    AGE                       (53,90]         .6500   .1312
AFFINITY_CARD  0    AGE                       [17,24]         .6500   .2134
AFFINITY_CARD  0    BOOKKEEPING_APPLICATION   0               .6500   .1500
AFFINITY_CARD  0    BOOKKEEPING_APPLICATION   1               .6500   .8500
AFFINITY_CARD  0    BULK_PACK_DISKETTES       0               .6500   .3670
```

Related Topics

- *Oracle Machine Learning for SQL User's Guide*

GET_MODEL_DETAILS_NMF Function

The GET_MODEL_DETAILS_NMF function returns a set of rows that provide the details of a non-negative matrix factorization model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_nmf(
      model_name IN VARCHAR2,
      partition_name VARCHAR2 DEFAULT NULL)
   RETURN DM_NMF_Feature_Set PIPELINED;
```

Parameters

Table 6-104   GET_MODEL_DETAILS_NMF Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
partition_name	Specifies a partition in a partitioned model

Return Values

Table 6-105   GET_MODEL_DETAILS_NMF Function Return Values

Return Value	Description
DM_NMF_FEATURE_SET	A set of rows of DM_NMF_FEATURE. The rows have the following columns: (feature_id NUMBER, mapped_feature_id VARCHAR2(4000), attribute_set DM_NMF_ATTRIBUTE_SET) The attribute_set column of DM_NMF_FEATURE returns a nested table of type DM_NMF_ATTRIBUTE_SET. The rows, of type DM_NMF_ATTRIBUTE, have the following columns: (attribute_name VARCHAR2(4000), attribute_subname VARCHAR2(4000), attribute_value VARCHAR2(4000), coefficient NUMBER)

Usage Notes

- The table function pipes out rows of type DM_NMF_FEATURE_SET. For information on machine learning data types and piped output from table functions, see "Data Types".
- When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following example returns model details for the feature extraction model NMF_SH_Sample.

```
SELECT * FROM (
SELECT F.feature_id,
       A.attribute_name,
       A.attribute_value,
       A.coefficient
  FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_NMF('NMF_SH_Sample')) F,
       TABLE(F.attribute_set) A
ORDER BY feature_id,attribute_name,attribute_value
) WHERE ROWNUM < 11;

FEATURE_ID ATTRIBUTE_NAME          ATTRIBUTE_VALUE          COEFFICIENT
--------- -----------------------  ---------------- -------------------
        1 AFFINITY_CARD                                 .051208078859308
        1 AGE                                          .0390513260041573
        1 BOOKKEEPING_APPLICATION                      .0512734004239326
        1 BULK_PACK_DISKETTES                           .232471260895683
        1 COUNTRY_NAME             Argentina          .00766817464479959
        1 COUNTRY_NAME             Australia         .000157637881096675
        1 COUNTRY_NAME             Brazil              .0031409632415604
        1 COUNTRY_NAME             Canada             .00144213099311427
        1 COUNTRY_NAME             China             .000102279310968754
        1 COUNTRY_NAME             Denmark           .000242424084307513
```

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_OC Function

The GET_MODEL_DETAILS_OC function returns a set of rows that provide the details of an O-cluster clustering model. The rows are an enumeration of the clustering patterns generated during the creation of the model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

You can provide input to GET_MODEL_DETAILS_OC to request specific information about the model, thus improving the performance of the query. If you do not specify filtering parameters, then GET_MODEL_DETAILS_OC returns all the information about the model.

Syntax

```
DBMS_DATA_MINING.get_model_details_oc(
      model_name VARCHAR2,
      cluster_id NUMBER   DEFAULT NULL,
      attribute  VARCHAR2 DEFAULT NULL,
      centroid   NUMBER   DEFAULT 1,
      histogram  NUMBER   DEFAULT 1,
      rules      NUMBER   DEFAULT 2,
      topn_attributes NUMBER DEFAULT NULL,
      partition_name VARCHAR2 DEFAULT NULL)
  RETURN dm_clusters PIPELINED;
```

Parameters

Table 6-106 GET_MODEL_DETAILS_OC Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`cluster_id`	The ID of a cluster in the model. When a valid cluster ID is specified, only the details of this cluster are returned. Otherwise the details for all clusters are returned.
`attribute`	The name of an attribute. When a valid attribute name is specified, only the details of this attribute are returned. Otherwise, the details for all attributes are returned
`centroid`	This parameter accepts the following values: • 1: Details about centroids are returned (default) • 0: Details about centroids are not returned
`histogram`	This parameter accepts the following values: • 1: Details about histograms are returned (default) • 0: Details about histograms are not returned
`rules`	This parameter accepts the following values: • 2: Details about rules are returned (default) • 1: Rule summaries are returned • 0: No information about rules is returned
`topn_attributes`	Restricts the number of attributes returned in the centroid, histogram, and rules objects. Only the *n* attributes with the highest confidence values in the rules are returned. If the number of attributes included in the rules is less than *topn*, then up to *n* additional attributes in alphabetical order are returned. If both the `attribute` and `topn_attributes` parameters are specified, then `topn_attributes` is ignored.
`partition_name`	Specifies a partition in a partitioned model.

Usage Notes

1. For information about machine learning data types and return values for clustering algorithms piped output from table functions, see "Data Types".
2. When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following example returns model details for the clustering model `OC_SH_Clus_sample`.

For each cluster in this example, the split predicate indicates the attribute and the condition used to assign records to the cluster's children during model build. It provides an important piece of information on how the population within a cluster can be divided up into two smaller clusters.

```
SELECT clu_id, attribute_name, op, s_value
     FROM (SELECT a.id clu_id, sp.attribute_name, sp.conditional_operator op,
                  sp.attribute_str_value s_value
             FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_OC(
                     'OC_SH_Clus_sample')) a,
                  TABLE(a.split_predicate) sp
           ORDER BY a.id, op, s_value)
     WHERE ROWNUM < 11;

     CLU_ID ATTRIBUTE_NAME       OP S_VALUE
----------- -------------------- ---------------------------------
          1 OCCUPATION           IN ?
          1 OCCUPATION           IN Armed-F
          1 OCCUPATION           IN Cleric.
          1 OCCUPATION           IN Crafts
          2 OCCUPATION           IN ?
          2 OCCUPATION           IN Armed-F
          2 OCCUPATION           IN Cleric.
          3 OCCUPATION           IN Exec.
          3 OCCUPATION           IN Farming
          3 OCCUPATION           IN Handler
```

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_SETTINGS Function

The `GET_MODEL_SETTINGS` function returns the settings used to build the given model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. See "Static Data Dictionary Views: `ALL_ALL_TABLES` to `ALL_OUTLINES`" in *Oracle Database Reference*.

Syntax

```
FUNCTION get_model_settings(model_name IN VARCHAR2)
  RETURN DM_Model_Settings PIPELINED;
```

Parameters

Table 6-107 GET_MODEL_SETTINGS Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.

Return Values

Table 6-108 GET_MODEL_SETTINGS Function Return Values

Return Value	Description

`DM_MODEL_SETTINGS`	A set of rows of type `DM_MODEL_SETTINGS`. The rows have the following columns:
	`DM_MODEL_SETTINGS TABLE OF SYS.DM_MODEL_SETTING`
	`Name                    Type`
	`---------------------- --------------------`
	`SETTING_NAME            VARCHAR2(30)`
	`SETTING_VALUE           VARCHAR2(4000)`

Usage Notes

1. This table function pipes out rows of type DM_MODEL_SETTINGS. For information on machine learning data types and piped output from table functions, see "DBMS_DATA_MINING Datatypes".
2. The setting names/values include both those specified by the user and any defaults assigned by the build process.

Examples

The following example returns model model settings for an example naive Bayes model.

```
SETTING_NAME                   SETTING_VALUE
------------------------------ ------------------------------
ALGO_NAME                       ALGO_NAIVE_BAYES
PREP_AUTO                       ON
ODMS_MAX_PARTITIONS             1000
NABS_SINGLETON_THRESHOLD       0
CLAS_WEIGHTS_BALANCED          OFF
NABS_PAIRWISE_THRESHOLD        0
ODMS_PARTITION_COLUMNS         GENDER,Y_BOX_GAMES
ODMS_MISSING_VALUE_TREATMENT   ODMS_MISSING_VALUE_AUTO
ODMS_SAMPLING                  ODMS_SAMPLING_DISABLE

9 rows selected.
```

Related Topics

- *Oracle Database Reference*

GET_MODEL_SIGNATURE Function

The GET_MODEL_SIGNATURE function returns the list of columns from the build input table that were used by the build process to train the model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. See "Static Data Dictionary Views: ALL_ALL_TABLES to ALL_OUTLINES" in *Oracle Database Reference*.

Syntax

```
FUNCTION get_model_signature (model_name IN VARCHAR2)
RETURN DM_Model_Signature PIPELINED;
```

Parameters

Table 6-109 GET_MODEL_SIGNATURE Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.

Return Values

Table 6-110 GET_MODEL_SIGNATURE Function Return Values

Return Value	Description

`DM_MODEL_SIGNATURE`	A set of rows of type `DM_MODEL_SIGNATURE`. The rows have the following columns:
	`DM_MODEL_SIGNATURE TABLE OF SYS.DM_MODEL_SIGNATURE_ATTRIBUTE`
	`Name                   Type`
	`------------------     --------------------`
	`ATTRIBUTE_NAME         VARCHAR2(130)`
	`ATTRIBUTE_TYPE         VARCHAR2(106)`

Usage Notes

1. This table function pipes out rows of type `DM_MODEL_SIGNATURE`. For information on machine learning data types and piped output from table functions, see "DBMS_DATA_MINING Datatypes".
2. The signature names or types include only those attributes used by the build process.

Examples

The following example returns model settings for an example naive Bayes model.

```
ATTRIBUTE_NAME                 ATTRIBUTE_TYPE
------------------------------ ------------------
AGE                            NUMBER
ANNUAL_INCOME                  NUMBER
AVERAGE___ITEMS_PURCHASED      NUMBER
BOOKKEEPING_APPLICATION        NUMBER
BULK_PACK_DISKETTES            NUMBER
BULK_PURCH_AVE_AMT             NUMBER
DISABLE_COOKIES                NUMBER
EDUCATION                      VARCHAR2
FLAT_PANEL_MONITOR             NUMBER
GENDER                         VARCHAR2
HOME_THEATER_PACKAGE           NUMBER
HOUSEHOLD_SIZE                 VARCHAR2
MAILING_LIST                   NUMBER
MARITAL_STATUS                 VARCHAR2
NO_DIFFERENT_KIND_ITEMS        NUMBER
OCCUPATION                     VARCHAR2
OS_DOC_SET_KANJI               NUMBER
PETS                           NUMBER
PRINTER_SUPPLIES               NUMBER
PROMO_RESPOND                  NUMBER
SHIPPING_ADDRESS_COUNTRY       VARCHAR2
SR_CITIZEN                     NUMBER
TOP_REASON_FOR_SHOPPING        VARCHAR2
WKS_SINCE_LAST_PURCH           NUMBER
WORKCLASS                      VARCHAR2
YRS_RESIDENCE                  NUMBER
Y_BOX_GAMES                    NUMBER

27 rows selected.
```

Related Topics

- *Oracle Database Reference*

GET_MODEL_DETAILS_SVD Function

The GET_MODEL_DETAILS_SVD function returns a set of rows that provide the details of a singular value decomposition model. Oracle recommends to use model details view settings. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

Refer to Model Details View for Singular Value Decomposition.

Syntax

```
DBMS_DATA_MINING.get_model_details_svd(
      model_name IN VARCHAR2,
      matrix_type IN VARCHAR2 DEFAULT NULL,
      partition_name VARCHAR2 DEFAULT NULL)
   RETURN DM_SVD_MATRIX_Set PIPELINED;
```

Parameters

Table 6-111 GET_MODEL_DETAILS_SVD Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
matrix_type	Specifies which of the three SVD matrix types to return. Values are: U, S, V, and NULL. When matrix_type is null (default), all matrices are returned. The U matrix is only computed when the SVDS_U_MATRIX_OUTPUT setting is enabled. It is not computed by default. If the model does not contain U matrices and you set matrix_type to U, an empty set of rows is returned. See Table 6-31.
partition_name	A partition in a partitioned model.

Return Values

Table 6-112 GET_MODEL_DETAILS_SVD Function Return Values

Return Value	Description

DM_SVD_MATRIX_SET	A set of rows of type DM_SVD_MATRIX. The rows have the following columns:
	`(matrix_type          CHAR(1),`
	` feature_id          NUMBER,`
	` mapped_feature_id   VARCHAR2(4000),`
	` attribute_name      VARCHAR2(4000),`
	` attribute_subname   VARCHAR2(4000),`
	` case_id             VARCHAR2(4000),`
	` value               NUMBER,`
	` variance            NUMBER,`
	` pct_cum_variance    NUMBER)`
	See Usage Notes for details.

Usage Notes

1. This table function pipes out rows of type DM_SVD_MATRIX. For information on machine learning data types and piped output from table functions, see "Data Types".

   The columns in each row returned by GET_MODEL_DETAILS_SVD are described as follows:

Column in DM_SVD_MATRIX_SET	Description
matrix_type	The type of matrix. Possible values are **S**, **V**, and **U**. This field is never null.
feature_id	The feature that the matrix entry refers to.
mapped_feature_id	A descriptive name for the feature.
attribute_name	Column name in the **V** matrix component bases. This field is null for the **S** and **U** matrices.
attribute_subname	Subname in the **V** matrix component bases. This is relevant only in the case of a nested column. This field is null for the **S** and **U** matrices.
case_id	Unique identifier of the row in the build data described by the **U** matrix projection. This field is null for the **S** and **V** matrices.
value	The matrix entry value.
variance	The variance explained by a component. It is non-null only for **S** matrix entries. This column is non-null only for **S** matrix entries and for SVD models with setting dbms_data_mining.svds_scoring_mode set to dbms_data_mining.svds_scoring_pca and the build data is centered, either manually or because the setting dbms_data_mining.prep_auto is set to dbms_data_mining.prep_auto_on.
pct_cum_variance	The percent cumulative variance explained by the components thus far. The components are ranked by the explained variance in descending order. This column is non-null only for **S** matrix entries and for SVD models with setting dbms_data_mining.svds_scoring_mode set to dbms_data_mining.svds_scoring_pca and the build data is centered, either manually or because the setting dbms_data_mining.prep_auto is set to dbms_data_mining.prep_auto_on.

2. The output of GET_MODEL_DETAILS is in sparse format. Zero values are not returned. Only the diagonal elements of the **S** matrix, the non-zero coefficients in the **V** matrix bases, and the non-zero **U** matrix projections are returned.

   There is one exception: If the data row does not produce non-zero **U** Matrix projections, the case ID for that row is returned with NULL for the feature_id and value. This is done to avoid losing any records from the original data.

3. GET_MODEL_DETAILS functions preserve model transparency by automatically reversing the transformations applied during the build process. Thus the attributes returned in the model details are the original attributes (or a close approximation of the original attributes) used to build the model.

4. When the value is NULL for a partitioned model, an exception is thrown. When the value is not null, it must contain the preferred partition name.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_SVM Function

The GET_MODEL_DETAILS_SVM function returns a set of rows that provide the details of a linear support vector machines (SVM) model. If invoked for nonlinear SVM, it returns ORA-40215. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views in *Oracle Machine Learning for SQL User’s Guide*.

In linear SVM models, only nonzero coefficients are stored. This reduces storage and speeds up model loading. As a result, if an attribute is missing in the coefficient list returned by GET_MODEL_DETAILS_SVM, then the coefficient of this attribute should be interpreted as zero.

Syntax

```
DBMS_DATA_MINING.get_model_details_svm(
      model_name   VARCHAR2,
      reverse_coef NUMBER DEFAULT 0,
      partition_name VARCHAR2 DEFAULT NULL)
  RETURN DM_SVM_Linear_Coeff_Set PIPELINED;
```

Parameters

Table 6-113 GET_MODEL_DETAILS_SVM Function Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
reverse_coef	Whether or not GET_MODEL_DETAILS_SVM should transform the attribute coefficients using the original attribute transformations. When reverse_coef is set to 0 (default), GET_MODEL_DETAILS_SVM returns the coefficients directly from the model without applying transformations. When reverse_coef is set to 1, GET_MODEL_DETAILS_SVM transforms the coefficients and bias by applying the normalization shifts and scales that were generated using automatic data preparation. See Usage Note 4.
partition_name	Specifies a partition in a partitioned model.

Return Values

Table 6-114 GET_MODEL_DETAILS_SVM Function Return Values

Return Value	Description

`DM_SVM_LINEAR_COEFF_SET`	A set of rows of type `DM_SVM_LINEAR_COEFF`. The rows have the following columns:
	`(class           VARCHAR2(4000),`
	` attribute_set    DM_SVM_ATTRIBUTE_SET)`
	The `attribute_set` column returns a nested table of type `DM_SVM_ATTRIBUTE_SET`. The rows, of type `DM_SVM_ATTRIBUTE`, have the following columns:
	`(attribute_name       VARCHAR2(4000),`
	` attribute_subname  VARCHAR2(4000),`
	` attribute_value      VARCHAR2(4000),`
	` coefficient            NUMBER)`
	See Usage Notes.

Usage Notes

1. This table function pipes out rows of type `DM_SVM_LINEAR_COEFF`. For information on machine learning data types and piped output from table functions, see "Data Types".
2. The `class` column of `DM_SVM_LINEAR_COEFF` contains classification target values. For SVM Regression models, `class` is null. For each classification target value, a set of coefficients is returned. For binary classification, one-class classification, and regression models, only a single set of coefficients is returned.
3. The `attribute_value` column in `DM_SVM_ATTRIBUTE_SET` is used for categorical attributes.
4. `GET_MODEL_DETAILS` functions preserve model transparency by automatically reversing the transformations applied during the build process. Thus the attributes returned in the model details are the original attributes (or a close approximation of the original attributes) used to build the model.

   The coefficients are related to the transformed, not the original, attributes. When returned directly with the model details, the coefficients may not provide meaningful information. If you want `GET_MODEL_DETAILS_SVM` to transform the coefficients such that they relate to the original attributes, set the `reverse_coef` parameter to 1.
5. When the value is `NULL` for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Examples

The following example returns model details for the SVM classification model `SVMC_SH_Clas_sample`, which was created by the sample program `dmsvcdem.sql`. For information about the sample programs, see *Oracle Machine Learning for SQL User's Guide*.

```
WITH
  mod_dtls AS (
  SELECT *
    FROM TABLE(DBMS_DATA_MINING.GET_MODEL_DETAILS_SVM('SVMC_SH_Clas_sample'))
  ),
  model_details AS (
  SELECT D.class, A.attribute_name, A.attribute_value, A.coefficient
    FROM mod_dtls D,
          TABLE(D.attribute_set) A
    ORDER BY D.class, ABS(A.coefficient) DESC
  )
  SELECT class, attribute_name aname, attribute_value aval, coefficient coeff
    FROM model_details
    WHERE ROWNUM < 11;

CLASS      ANAME                     AVAL                      COEFF
---------- ------------------------- ------------------------- -----
1                                                              -2.85
1          BOOKKEEPING_APPLICATION                              1.11
1          OCCUPATION                Other                      -.94
1          HOUSEHOLD_SIZE            4-5                         .88
1          CUST_MARITAL_STATUS       Married                     .82
1          YRS_RESIDENCE                                         .76
1          HOUSEHOLD_SIZE            6-8                        -.74
1          OCCUPATION                Exec.                       .71
1          EDUCATION                 11th                       -.71
1          EDUCATION                 Masters                     .63
```

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

GET_MODEL_DETAILS_XML Function

This function returns an XML object that provides the details of a decision tree model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. Use model detail views instead.

See Model Detail Views for Decision Tree in *Oracle Machine Learning for SQL User’s Guide*.

Syntax

```
DBMS_DATA_MINING.get_model_details_xml(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN XMLType;
```

Parameters

Table 6-115   GET_MODEL_DETAILS_XML Function Parameters

Parameter	Description
`model_name`	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model.

Return Values

Table 6-116 GET_MODEL_DETAILS_XML Function Return Value

Return Value	Description
`XMLTYPE`	The XML definition for the decision tree model. See "XMLTYPE" for details. The XML definition conforms to the Data Mining Group Predictive Model Markup Language (PMML) version 2.1 specification. The specification is available at `https://dmg.org`. If a nested attribute is used as a splitter, the attribute will appear in the XML document as field="'<column_name>'.<subname>", as opposed to the non-nested attributes which appear in the document as field="<column_name>". **Note:** The column names are surrounded by single quotes and a period separates the column_name from the subname. The rest of the document style remains unchanged.

Usage Notes

Special characters that cannot be displayed by Oracle XML are converted to '#'.

Examples

The following statements in SQL*Plus return the details of the decision tree model `dt_sh_clas_sample`.

Note: The "&quot" characters you will see in the XML output are a result of SQL*Plus behavior. To display the XML in proper format, cut and past it into a file and open the file in a browser.

```
column dt_details format a320
SELECT
 dbms_data_mining.get_model_details_xml('dt_sh_clas_sample')
 AS DT_DETAILS
FROM dual;


DT_DETAILS
--------------------------------------------------------------------------------
<PMML version="2.1">
  <Header copyright="Copyright (c) 2004, Oracle Corporation. All rights
      reserved."/>
  <DataDictionary numberOfFields="9">
    <DataField name="AFFINITY_CARD" optype="categorical"/>
    <DataField name="AGE" optype="continuous"/>
    <DataField name="BOOKKEEPING_APPLICATION" optype="continuous"/>
    <DataField name="CUST_MARITAL_STATUS" optype="categorical"/>
    <DataField name="EDUCATION" optype="categorical"/>
    <DataField name="HOUSEHOLD_SIZE" optype="categorical"/>
    <DataField name="OCCUPATION" optype="categorical"/>
    <DataField name="YRS_RESIDENCE" optype="continuous"/>
    <DataField name="Y_BOX_GAMES" optype="continuous"/>
  </DataDictionary>
  <TreeModel modelName="DT_SH_CLAS_SAMPLE" functionName="classification"
      splitCharacteristic="binarySplit">
    <Extension name="buildSettings">
      <Setting name="TREE_IMPURITY_METRIC" value="TREE_IMPURITY_GINI"/>
      <Setting name="TREE_TERM_MAX_DEPTH" value="7"/>
      <Setting name="TREE_TERM_MINPCT_NODE" value=".05"/>
      <Setting name="TREE_TERM_MINPCT_SPLIT" value=".1"/>
      <Setting name="TREE_TERM_MINREC_NODE" value="10"/>
      <Setting name="TREE_TERM_MINREC_SPLIT" value="20"/>
      <costMatrix>
        <costElement>
          <actualValue>0</actualValue>
          <predictedValue>0</predictedValue>
          <cost>0</cost>
        </costElement>
        <costElement>
          <actualValue>0</actualValue>
          <predictedValue>1</predictedValue>
          <cost>1</cost>
        </costElement>
        <costElement>
          <actualValue>1</actualValue>
          <predictedValue>0</predictedValue>
          <cost>8</cost>
        </costElement>
        <costElement>
          <actualValue>1</actualValue>
          <predictedValue>1</predictedValue>
          <cost>0</cost>
        </costElement>
      </costMatrix>
    </Extension>
    <MiningSchema>
      .
      .
      .
      .
      .
      .
      </Node>
    </Node>
  </TreeModel>
</PMML>
```

GET_MODEL_TRANSFORMATIONS Function

This function returns the transformation expressions embedded in the specified model. Starting from Oracle Database 12*c* Release 2, this function is deprecated. See "Static Data Dictionary Views: `ALL_ALL_TABLES` to `ALL_OUTLINES`" in *Oracle Database Reference*.

All `GET_*` interfaces are replaced by model views, and Oracle recommends that users reference the model views to retrieve the relevant information. The `GET_MODEL_TRANSFORMATIONS` function is replaced by the following:

- USER(/DBA/ALL)_MINING_MODEL_XFORMS: provides the user-embedded transformations
- DM$VX prefixed model view: provides text feature extraction information
- D$VN prefixed mode view: provides normalization and missing value information
- DM$VB: provides binning information

See Also:

"About Transformation Lists" in DBMS_DATA_MINING_TRANSFORM Operational Notes

"GET_TRANSFORM_LIST Procedure"

"CREATE_MODEL Procedure"

"ALL_MINING_MODEL_XFORMS" in *Oracle Database Reference*

"DBA_MINING_MODEL_XFORMS" in *Oracle Database Reference*

"USER_MINING_MODEL_XFORMS" in *Oracle Database Reference*

Model Details View for Binning

Normalization and Missing Value Handling

Data Preparation for Text Features

Syntax

```
DBMS_DATA_MINING.get_model_transformations(
      model_name IN VARCHAR2,
      partition_name IN VARCHAR2 DEFAULT NULL)
  RETURN DM_Transforms PIPELINED;
```

Parameters

Table 6-117 GET_MODEL_TRANSFORMATIONS Function Parameters

Parameter	Description
`model_name`	Indicates the name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.
`partition_name`	Specifies a partition in a partitioned model

Return Values

Table 6-118 GET_MODEL_TRANSFORMATIONS Function Return Value

Return Value	Description

`DM_TRANSFORMS`	The transformation expressions embedded in *`model_name`*.
	The `DM_TRANSFORMS` type is a table of `DM_TRANSFORM` objects. Each `DM_TRANSFORM` has these fields:
	`attribute_name        VARCHAR2(4000)`
	`attribute_subname     VARCHAR2(4000)`
	`expression            CLOB`
	`reverse_expression    CLOB`

Usage Notes

When Automatic Data Preparation (ADP) is enabled, both automatic and user-defined transformations may be associated with an attribute. In this case, the user-defined transformations are evaluated before the automatic transformations.

When invoked for a partitioned model, the `partition_name` parameter must be specified.

Examples

In this example, several columns in the `SH.CUSTOMERS` table are used to create a naive Bayes model. A transformation expression is specified for one of the columns. The model does not use ADP.

```
CREATE OR REPLACE VIEW mining_data AS
   SELECT cust_id, cust_year_of_birth, cust_income_level,cust_credit_limit
   FROM sh.customers;

describe mining_data
 Name                                   Null?    Type
 -------------------------------------- -------- --------------------------
 CUST_ID                                NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                     NOT NULL NUMBER(4)
 CUST_INCOME_LEVEL                               VARCHAR2(30)
 CUST_CREDIT_LIMIT                               NUMBER

CREATE TABLE settings_nb(
      setting_name  VARCHAR2(30),
      setting_value VARCHAR2(30));
BEGIN
     INSERT INTO settings_nb (setting_name, setting_value) VALUES
           (dbms_data_mining.algo_name, dbms_data_mining.algo_naive_bayes);
     INSERT INTO settings_nb (setting_name, setting_value) VALUES
           (dbms_data_mining.prep_auto, dbms_data_mining.prep_auto_off);
     COMMIT;
END;
/
DECLARE
    mining_data_xforms   dbms_data_mining_transform.TRANSFORM_LIST;
  BEGIN
    dbms_data_mining_transform.SET_TRANSFORM (
         xform_list           =>  mining_data_xforms,
         attribute_name       => 'cust_year_of_birth',
         attribute_subname    =>  null,
         expression           => 'cust_year_of_birth + 10',
         reverse_expression   => 'cust_year_of_birth - 10');
    dbms_data_mining.CREATE_MODEL (
        model_name           =>  'new_model',
        mining_function      =>   dbms_data_mining.classification,
        data_table_name      =>  'mining_data',
        case_id_column_name  =>  'cust_id',
        target_column_name   =>  'cust_income_level',
        settings_table_name  =>  'settings_nb',
        data_schema_name     =>   nulL,
        settings_schema_name =>   null,
        xform_list           =>   mining_data_xforms );
  END;
 /
SELECT attribute_name, TO_CHAR(expression), TO_CHAR(reverse_expression)
      FROM TABLE (dbms_data_mining.GET_MODEL_TRANSFORMATIONS('new_model'));
```

```
ATTRIBUTE_NAME      TO_CHAR(EXPRESSION)       TO_CHAR(REVERSE_EXPRESSION)
------------------  ------------------------  -----------------------------
CUST_YEAR_OF_BIRTH  cust_year_of_birth + 10   cust_year_of_birth - 10
```

Related Topics

- *Oracle Database Reference*

GET_TRANSFORM_LIST Procedure

This procedure converts transformation expressions specified as DM_TRANSFORMS to a transformation list (TRANSFORM_LIST) that can be used in creating a model. DM_TRANSFORMS is returned by the GET_MODEL_TRANSFORMATIONS function.

You can also use routines in the DBMS_DATA_MINING_TRANSFORM package to construct a transformation list.

> **See Also:**
>
> “About Transformation Lists” in DBMS_DATA_MINING_TRANSFORM
>
> "GET_MODEL_TRANSFORMATIONS Function"
>
> "CREATE_MODEL Procedure"

Syntax

```
DBMS_DATA_MINING.GET_TRANSFORM_LIST (
      xform_list            OUT NOCOPY TRANSFORM_LIST,
      model_xforms          IN  DM_TRANSFORMS);
```

Parameters

Table 6-119   GET_TRANSFORM_LIST Procedure Parameters

Parameter	Description

xform_list	A list of transformation specifications that can be embedded in a model. Accepted as a parameter to the CREATE_MODEL Procedure.
	The TRANSFORM_LIST type is a table of TRANSFORM_REC objects. Each TRANSFORM_REC has these fields:
	`attribute_name      VARCHAR2(30)`
	`attribute_subname   VARCHAR2(4000)`
	`expression          EXPRESSION_REC`
	`reverse_expression  EXPRESSION_REC`
	`attribute_spec      VARCHAR2(4000)`
	For details about the TRANSFORM_LIST collection type, see Table 6-127.

Table 6-119 (Cont.) GET_TRANSFORM_LIST Procedure Parameters

Parameter	Description

`model_xforms`	A list of embedded transformation expressions returned by the GET_MODEL_TRANSFORMATIONS Function for a specific model.
	The `DM_TRANSFORMS` type is a table of `DM_TRANSFORM` objects. Each `DM_TRANSFORM` has these fields:
	`attribute_name       VARCHAR2(4000)`
	`attribute_subname    VARCHAR2(4000)`
	`expression           CLOB`
	`reverse_expression   CLOB`

Examples

In this example, a model `mod1` is trained using several columns in the `SH.CUSTOMERS` table. The model uses ADP, which automatically bins one of the columns.

A second model `mod2` is trained on the same data without ADP, but it uses a transformation list that was obtained from `mod1`. As a result, both `mod1` and `mod2` have the same embedded transformation expression.

```
CREATE OR REPLACE VIEW mining_data AS
     SELECT cust_id, cust_year_of_birth, cust_income_level, cust_credit_limit
     FROM sh.customers;

describe mining_data
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_INCOME_LEVEL                                  VARCHAR2(30)
 CUST_CREDIT_LIMIT                                  NUMBER

CREATE TABLE setmod1(setting_name  VARCHAR2(30),setting_value VARCHAR2(30));
BEGIN
   INSERT INTO setmod1 VALUES (dbms_data_mining.algo_name, dbms_data_mining.algo_naive_bayes);
   INSERT INTO setmod1 VALUES (dbms_data_mining.prep_auto,dbms_data_mining.prep_auto_on);
   dbms_data_mining.CREATE_MODEL (
               model_name            => 'mod1',
               mining_function       => dbms_data_mining.classification,
               data_table_name       => 'mining_data',
               case_id_column_name   => 'cust_id',
               target_column_name    => 'cust_income_level',
               settings_table_name   => 'setmod1');
    COMMIT;
END;
/
CREATE TABLE setmod2(setting_name  VARCHAR2(30),setting_value VARCHAR2(30));
BEGIN
  INSERT INTO setmod2
      VALUES (dbms_data_mining.algo_name, dbms_data_mining.algo_naive_bayes);
  COMMIT;
END;
/
DECLARE
  v_xform_list       dbms_data_mining_transform.TRANSFORM_LIST;
  dmxf               DM_TRANSFORMS;
```

```
BEGIN
   EXECUTE IMMEDIATE
    'SELECT dm_transform(attribute_name, attribute_subname,expression, reverse_expression)
     FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS (''mod1''))'
     BULK COLLECT INTO dmxf;
   dbms_data_mining.GET_TRANSFORM_LIST (
        xform_list              =>  v_xform_list,
        model_xforms            =>  dmxf);
   dbms_data_mining.CREATE_MODEL(
         model_name             => 'mod2',
         mining_function        =>  dbms_data_mining.classification,
         data_table_name        => 'mining_data',
         case_id_column_name    => 'cust_id',
         target_column_name     => 'cust_income_level',
         settings_table_name    => 'setmod2',
         xform_list             =>  v_xform_list);
END;
/

-- Transformation expression embedded in mod1
SELECT TO_CHAR(expression) FROM TABLE (dbms_data_mining.GET_MODEL_TRANSFORMATIONS('mod1'));

TO_CHAR(EXPRESSION)
--------------------------------------------------------------------------------
CASE WHEN "CUST_YEAR_OF_BIRTH"<1915 THEN 0 WHEN "CUST_YEAR_OF_BIRTH"<=1915 THEN 0
WHEN "CUST_YEAR_OF_BIRTH"<=1920.5 THEN 1 WHEN "CUST_YEAR_OF_BIRTH"<=1924.5 THEN 2
.
.
.
.5 THEN 29 WHEN "CUST_YEAR_OF_BIRTH" IS NOT NULL THEN 30 END

-- Transformation expression embedded in mod2
SELECT TO_CHAR(expression) FROM TABLE (dbms_data_mining.GET_MODEL_TRANSFORMATIONS('mod2'));

TO_CHAR(EXPRESSION)
--------------------------------------------------------------------------------
CASE WHEN "CUST_YEAR_OF_BIRTH"<1915 THEN 0 WHEN "CUST_YEAR_OF_BIRTH"<=1915 THEN 0
WHEN "CUST_YEAR_OF_BIRTH"<=1920.5 THEN 1 WHEN "CUST_YEAR_OF_BIRTH"<=1924.5 THEN 2
.
.
.
.5 THEN 29 WHEN "CUST_YEAR_OF_BIRTH" IS NOT NULL THEN 30 END

-- Reverse transformation expression embedded in mod1
SELECT TO_CHAR(reverse_expression)FROM TABLE (dbms_data_mining.GET_MODEL_TRANSFORMATIONS('mod1'));

TO_CHAR(REVERSE_EXPRESSION)
--------------------------------------------------------------------------------
DECODE("CUST_YEAR_OF_BIRTH",0,'( ; 1915), [1915; 1915]',1,'(1915; 1920.5]',2,'(1
920.5; 1924.5]',3,'(1924.5; 1928.5]',4,'(1928.5; 1932.5]',5,'(1932.5; 1936.5]',6
.
.
.
8,'(1987.5; 1988.5]',29,'(1988.5; 1989.5]',30,'(1989.5;  )',NULL,'NULL')

-- Reverse transformation expression embedded in mod2
SELECT TO_CHAR(reverse_expression) FROM TABLE (dbms_data_mining.GET_MODEL_TRANSFORMATIONS('mod2'));

TO_CHAR(REVERSE_EXPRESSION)
--------------------------------------------------------------------------------
DECODE("CUST_YEAR_OF_BIRTH",0,'( ; 1915), [1915; 1915]',1,'(1915; 1920.5]',2,'(1
920.5; 1924.5]',3,'(1924.5; 1928.5]',4,'(1928.5; 1932.5]',5,'(1932.5; 1936.5]',6
.
.
.
8,'(1987.5; 1988.5]',29,'(1988.5; 1989.5]',30,'(1989.5;  )',NULL,'NULL')
```

IMPORT_MODEL Procedure

This procedure imports one or more machine learning models. The procedure is overloaded. You can call it to import machine learning models from a dump file set, or you can call it to import a single machine learning model from a PMML document.

Import from a dump file set

You can import machine learning models from a dump file set that was created by the EXPORT_MODEL Procedure. IMPORT_MODEL and EXPORT_MODEL use Oracle Data Pump technology to export to and import from a dump file set.

When Oracle Data Pump is used directly to export/import an entire schema or database, the machine learning models in the schema or database are included. EXPORT_MODEL and IMPORT_MODEL export/import machine learning models only.

Import from PMML

You can import a machine learning model represented in Predictive Model Markup Language (PMML). The model must be of type RegressionModel, either linear regression or binary logistic regression.

PMML is an XML-based standard specified by the Data Mining Group (https://dmg.org). Applications that are PMML-compliant can deploy PMML-compliant models that were created by any vendor. Oracle Machine Learning for SQL supports the core features of PMML 3.1 for regression models.

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for more information about exporting and importing machine learning models
>
> *Oracle Database Utilities* for information about Oracle Data Pump
>
> https://dmg.org/dmg-faq.html for more information about PMML

Syntax

Imports a machine learning model from a dump file set:

```
DBMS_DATA_MINING.IMPORT_MODEL (
      filename          IN  VARCHAR2,
      directory         IN  VARCHAR2,
      model_filter      IN  VARCHAR2 DEFAULT NULL,
      operation         IN  VARCHAR2 DEFAULT NULL,
      remote_link       IN  VARCHAR2 DEFAULT NULL,
      jobname           IN  VARCHAR2 DEFAULT NULL,
      schema_remap      IN  VARCHAR2 DEFAULT NULL,
      tablespace_remap  IN  VARCHAR2 DEFAULT NULL);
```

Imports a machine learning model from a PMML document:

```
DBMS_DATA_MINING.IMPORT_MODEL (
      model_name        IN  VARCHAR2,
      pmmldoc           IN  XMLTYPE
      strict_check      IN  BOOLEAN DEFAULT FALSE);
```

Parameters

Table 6-120   IMPORT_MODEL Procedure Parameters

Parameter	Description
`filename`	Name of the dump file set from which the models should be imported. The dump file set must have been created by the `EXPORT_MODEL` procedure or the `expdp` export utility of Oracle Data Pump. The dump file set can contain one or more files. (Refer to "EXPORT_MODEL Procedure" for details.) If the dump file set contains multiple files, you can specify '*filename*%U' instead of listing them. For example, if your dump file set contains 3 files, `archive01.dmp`, `archive02.dmp`, and `archive03.dmp`, you can import them by specifying `'archive%U'`.
`directory`	Name of a pre-defined directory object that specifies where the dump file set is located. Both the exporting and the importing user must have read/write access to the directory object and to the file system directory that it identifies. Note: The target database must have also have read/write access to the file system directory.

`model_filter`	Optional parameter that specifies one or more models to import. If you do not specify a value for `model_filter`, all models in the dump file set are imported. You can also specify `NULL` (the default) or 'ALL' to import all models.
	The value of `model_filter` can be one or more model names. The following are valid filters.
	`'mymodel1'`
	`'name IN (''mymodel2'',''mymodel3'')'`
	The first causes `IMPORT_MODEL` to import a single model named `mymodel1`. The second causes `IMPORT_MODEL` to import two models, `mymodel2` and `mymodel3`.

`operation`	Optional parameter that specifies whether to import the models or the SQL statements that create the models. By default, the models are imported. You can specify either of the following values for `operation`: • `'IMPORT'` — Import the models (Default) • `'SQL_FILE'`— Write the SQL DDL for creating the models to a text file. The text file is named *job_name*`.sql` and is located in the dump set directory.
`remote_link`	Optional parameter that specifies the name of a database link to a remote system. The default value is `NULL`. A database link is a schema object in a local database that enables access to objects in a remote database. When you specify a value for `remote_link`, you can import models into the local database from the remote database. The import is fileless; no dump file is involved. The `IMP_FULL_DATABASE` role is required for importing the remote models. The `EXP_FULL_DATABASE` privilege, the `CREATE DATABASE LINK` privilege, and other privileges may also be required. (See Example 2.)
`jobname`	Optional parameter that specifies the name of the import job. By default, the name has the form *username*`_imp_`*nnnn*, where *nnnn* is a number. For example, a job name in the `SCOTT` schema might be `SCOTT_imp_134`. If you specify a job name, it must be unique within the schema. The maximum length of the job name is 30 characters. A log file for the import job, named *jobname*`.log,` is created in the same directory as the dump file set.

Table 6-120 (Cont.) IMPORT_MODEL Procedure Parameters

Parameter	Description
`schema_remap`	Optional parameter for importing into a different schema. By default, models are exported and imported within the same schema. If the dump file set belongs to a different schema, you must specify a schema mapping in the form *`export_user`*:*`import_user`*. For example, you would specify `'SCOTT:MARY'` to import a model exported by `SCOTT` into the `MARY` schema. Note: In some cases, you may need to have the `IMP_FULL_DATABASE` privilege or the `SYS` role to import a model from a different schema.
`tablespace_remap`	Optional parameter for importing into a different tablespace. By default, models are exported and imported within the same tablespace. If the dump file set belongs to a different tablespace, you must specify a tablespace mapping in the form *`export_tablespace`*:*`import_tablespace`*. For example, you would specify `'TBLSPC01:TBLSPC02'` to import a model that was exported from tablespace `TBLSPC01` into tablespace `TBLSPC02`. Note: In some cases, you may need to have the `IMP_FULL_DATABASE` privilege or the `SYS` role to import a model from a different tablespace.
`model_name`	Name for the new model that will be created in the database as a result of an import from PMML The name must be unique within the user's schema.
`pmmldoc`	The PMML document representing the model to be imported. The PMML document has an `XMLTYPE` object type. See "XMLTYPE" for details.
`strict_check`	Whether or not an error occurs when the PMML document contains sections that are not part of core PMML (for example, Output or Targets). OML4SQL supports only core PMML; any non-core features may affect the scoring representation. If the PMML does not strictly conform to core PMML and `strict_check` is set to `TRUE`, then `IMPORT_MODEL` returns an error. If strict_check is `FALSE` (the default), then the error is suppressed. The model may be imported and scored.

Examples

1. This example shows a model being exported and imported within the schema `oml_user2`. Then the same model is imported into the `oml_user3` schema. The `oml_user3` user has the `IMP_FULL_DATABASE` privilege. The `oml_user2` user has been assigned the `USER2` tablespace; `oml_user3` has been assigned the `USER3` tablespace.

```
SQL> connect oml_user2
Enter password: oml_user2_password
Connected.
SQL> select model_name from user_mining_models;

MODEL_NAME
------------------------------
NMF_SH_SAMPLE
SVMO_SH_CLAS_SAMPLE
SVMR_SH_REGR_SAMPLE

-- export the model called NMF_SH_SAMPLE to a dump file in same schema
SQL>EXECUTE DBMS_DATA_MINING.EXPORT_MODEL (
            filename =>'NMF_SH_SAMPLE_out',
            directory =>'DATA_PUMP_DIR',
            model_filter => 'name = ''NMF_SH_SAMPLE''');

-- import the model back into the same schema
```

```
SQL>EXECUTE DBMS_DATA_MINING.IMPORT_MODEL (
            filename => 'NMF_SH_SAMPLE_out01.dmp',
            directory => 'DATA_PUMP_DIR',
            model_filter => 'name = ''NMF_SH_SAMPLE''');

-- connect as different user
-- import same model into that schema
SQL> connect oml_user3
Enter password: oml_user3_password
Connected.
SQL>EXECUTE DBMS_DATA_MINING.IMPORT_MODEL (
            filename => 'NMF_SH_SAMPLE_out01.dmp',
            directory => 'DATA_PUMP_DIR',
            model_filter => 'name = ''NMF_SH_SAMPLE''',
            operation =>'IMPORT',
            remote_link => NULL,
            jobname => 'nmf_imp_job',
            schema_remap => 'oml_user2:oml_user3',
            tablespace_remap => 'USER2:USER3');
```

The following example shows user MARY importing all models from a dump file, model_exp_001.dmp, which was created by user SCOTT. User MARY has been assigned a tablespace named USER2; user SCOTT was assigned the tablespace USERS when the models were exported into the dump file model_exp_001.dmp.The dump file is located in the file system directory mapped to a directory object called DM_DUMP. If user MARY does not have IMP_FULL_DATABASE privileges, IMPORT_MODEL will raise an error.

```
-- import all models
DECLARE
  file_name  VARCHAR2(40);
BEGIN
  file_name := 'model_exp_001.dmp';
  DBMS_DATA_MINING.IMPORT_MODEL(
            filename=> 'file_name',
            directory=>'DM_DUMP',
            schema_remap=>'SCOTT:MARY',
            tablespace_remap=>'USERS:USER2');
  DBMS_OUTPUT.PUT_LINE(
           'DBMS_DATA_MINING.IMPORT_MODEL of all models from SCOTT done!');
END;
/
```

2. This example shows how the user xuser could import the model oml_user.r1mod from a remote database. The SQL*Net connection alias for the remote database is R1DB. The user xuser is assigned the SYSAUX tablespace; the user oml_user is assigned the TBS_1 tablespace.

```
CONNECT / AS SYSDBA;
GRANT CREATE DATABASE LINK TO xuser;
GRANT imp_full_database TO xuser;
CONNECT xuser/xuserpassword
CREATE DATABASE LINK oml_user_link
         CONNECT TO oml_user IDENTIFIED BY oml_userpassword USING 'R1DB';
EXEC dbms_data_mining.import_model (
    NULL,
   'oml_user_DIR',
   'R1MOD',
    remote_link => 'oml_user_LINK', schema_remap => 'oml_user:XUSER',
                   tablespace_remap => 'TBS_1:SYSAUX' );
SELECT name FROM dm_user_models;

NAME
```

```
-----------------------------------------------------------------------------
R1MOD
```

3. This example shows how a PMML document called `SamplePMML1.xml` could be imported from a location referenced by directory object `PMMLDIR` into the schema of the current user. The imported model will be called `PMMLMODEL1`.

```
BEGIN
    dbms_data_mining.import_model ('PMMLMODEL1',
        XMLType (bfilename ('PMMLDIR', 'SamplePMML1.xml'),
          nls_charset_id ('AL32UTF8')
        ));
END;
```

IMPORT_ONNX_MODEL Procedure

This procedure enables you to import an ONNX model into the Database.

Syntax

```
DBMS_DATA_MINING.IMPORT_ONNX_MODEL(
model_name   IN  VARCHAR2,
model_data   IN  BLOB,
metadata     IN  JSON);
```

Parameters

Table 6-121 IMPORT_ONNX_MODEL Procedure Parameters

Parameter	Description
`model_name`	Name of the model in the form `[schema_name.]model_name`. If you do not specify a schema, then your own schema is used.
`model_data`	It is a `BLOB` holding the ONNX representation of the model. The `BLOB` contains the identical byte sequence as the one stored in an ONNX file.
`metadata`	A JSON description of the metadata describing the model. The metadata at minimum must describe the machine learning function supported by the model. The model's metadata parameters are described in JSON Metadata Parameters for ONNX Models .

Example

The following example illustrates a code snippet of using the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure. The complete step-by-step example is illustrated in Import ONNX Models and Generate Embeddings and Alternate Method to Import ONNX Models.

```
DBMS_DATA_MINING.IMPORT_ONNX_MODEL('my_embedding_model.onnx',
                                              :blob_bind_variable,
                                               JSON('{"function" :
"embedding",
                                                      "embeddingOutput" :
"embedding" ,
                                                       "input":{"input":
["DATA"]}}'));
```

For a complete example to illustrate how you can define a BLOB variable and use it in the IMPORT_ONNX_MODEL procedure, you can have the following:

```
CREATE OR REPLACE MY_LOAD_EMBEDDING_MODEL(embedding_model_name VARCHAR2,
onnx_blob BLOB) IS
BEGIN
DBMS_DATA_MINING.IMPORT_ONNX_MODEL(embedding_model_name,
                            onnx_blob,
                            JSON('{"function" : "embedding",
                                   "embeddingOutput" : "embedding" ,
                                   "input":{"input": ["DATA"]}}'));
END;
/
```

Usage Notes

The name of the model follows the same restrictions as those used for other machine learning models, namely:

- The schema name, if provided, is limited to 128 characters.
- The model name is limited to 123 characters and must follow the rules of unquoted identifiers: they contain only alphanumeric characters, the underscore (_), dollar sign ($), and pound sign (#). The initial character must be alphabetic.
- The model size is limited to 1 gigabyte.
- The model must not depend on external initializers. To know more about initializers and other ONNX concepts, see https://onnx.ai/onnx/intro/concepts.html.

IMPORT_SERMODEL Procedure

This procedure imports the serialized format of the model back into a database.

The import routine takes the serialized content in the BLOB and the name of the model to be created with the content. This import does not create model views or tables that are needed for querying model details. The import procedure only provides the ability to score the model.

Syntax

```
DBMS_DATA_MINING.IMPORT_SERMODEL (
      model_data     IN BLOB,
      model_name     IN VARCHAR2,);
```

Parameters

Table 6-122 IMPORT_SERMODEL Procedure Parameters

Parameter	Description
model_data	Provides model data in BLOB format.
model_name	Name of the machine learning model in the form [*schema_name*.]*model_name*. If you do not specify a schema, then your own schema is used.

Examples

The following statement imports the serialized format of the models.

```
declare
 v_blob blob;
BEGIN
 dbms_lob.createtemporary(v_blob, FALSE);
-- fill in v_blob from somewhere (e.g., bfile, etc.)
 dbms_data_mining.import_sermodel(v_blob, 'MY_MODEL');
 dbms_lob.freetemporary(v_blob);
END;
/
```

Related Topics

- EXPORT_SERMODEL Procedure
  This procedure exports the model in a serialized format so that they can be moved to another platform for scoring.

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for more information about exporting and importing machine learning models

JSON Schema for R Extensible Algorithm

Provides some flexibility when creating a new JSON object following the JSON schema.

Usage Note

Some flexibility when creating a new JSON object is as follows:

- Partial registration is allowed. For example, the detail function can be missing.
- Different orders are allowed. For example, the detail function can be written before the build function or after it.

Example 6-20   JSON Schema

JSON schema 1.1 for R extensible algorithm:

```
{
    "type": "object",
    "properties": {
        "algo_name_display": { "type" : "object",
                                               "properties" : {
                                               "language" : { "type" :
"string",

"enum" : ["English", "Spanish", "French"],

"default" : "English"},
                                               "name" : { "type" : "string"}}
```

```
                                             },

        "function_language": {"type": "string" },
        "mining_function": {
                 "type" : "array",
                 "items" : [
                     { "type" : "object",
                        "properties" : {
                           "mining_function_name"  : { "type" : "string"},
                           "build_function": {
                                   "type": "object",
                                   "properties": {
                                        "function_body": { "type": "CLOB" }
                                                        }
                                    },

        "detail_function": {
                 "type" : "array",
                  "items" : [
                      {"type": "object",
                        "properties": {
                             "function_body": { "type": "CLOB" },
                             "view_columns": { "type" : "array",
                                                                   "items" : {

"type" : "object",

"properties" : {

 "name" : { "type" : "string"},

 "type" : { "type" : "string",

                 "enum" : ["VARCHAR2",

                                  "NUMBER",

                                  "DATE",

                                  "BOOLEAN"]

               }
                                                                             }
                                                          }
                                            }
                                 }
                     ]
        },

       "score_function": {
                 "type": "object",
                 "properties": {
                       "function_body": { "type": "CLOB" }
                        }
                 },
        "weight_function": {
```

```
                        "type": "object",
                        "properties": {
                            "function_body": { "type": "CLOB" },
                        }
                 }
                               }
           }]
        },

       "algo_setting": {
                "type" : "array",
                "items" : [
                    { "type" : "object",
                       "properties" : {
                          "name"              : { "type" : "string"},
                          "name_display": { "type" : "object",
                                                         "properties" : {
                                                         "language" :
{ "type" : "string",

   "enum" : ["English", "Spanish", "French"],

   "default" : "English"},
                                                         "name" : { "type" :
"string"}}
                                                      },
                          "type" : { "type" : "string",
                                          "enum" : ["string", "integer",
"number", "boolean"]},

                          "optional": {"type" : "BOOLEAN",
                                               "default" : "FALSE"},

                          "value" : { "type" :  "string"},

                          "min_value" : { "type": "object",
                                                      "properties": {
                                                            "min_value":
{"type": "number"},
                                                             "inclusive":
{ "type": "boolean",

     "default" : TRUE},
                                                       }
                                                  },
                           "max_value" : {"type": "object",
                                                     "properties": {
                                                          "max_value":
{"type": "number"},
                                                          "inclusive":
{ "type": "boolean",

   "default" : TRUE},
                                                              }
                                                    },
                          "categorical choices" : { "type": "array",
                                                                  "items": {
                                                                      "type":
"string"
                                                                   }
                                                               },

                          "description_display": { "type" : "object",

"properties" : {

"language" : { "type" : "string",

            "enum" : ["English", "Spanish", "French"],

            "default" : "English"},
                                                                  "name" :
{ "type" : "string"}}
                                                               }
                        }
                    }
                 ]
          }
    }
}
```

Example 6-21 JSON object example

The following is an JSON object example that must be passed to the registration procedure:

```
{  "algo_name_display"    :     {"English", "t1"},
                         "function_language"    :      "R",
                         "mining_function" : {
  "mining_function_name" : "CLASSIFICATION",
                         "build_function" : {"function_body": "function(dat,
formula, family)
{
                                                         set.seed(1234);
                                          mod <- glm(formula = formula,
data=dat,
                                                      family=
eval(parse(text=family))); mod}"},
           "score_function" :  { "function_body": "function(mod, dat) {
                                             res <- predict(mod, newdata =
dat,
type=''response
                                         '');
                                             res2=data.frame(1-res, res);
res2}"}}
                          },
                          "algo_setting" :   [{"name"                :
"dbms_data_mining.odms_m


                                                  issing_value_treatment",
                            "name_display"    : {"English",
"dbms_data_mining.odms_missing_value
_treatment"},
                            "type"                  : "string",
                            "optional"            :  "TRUE",
                            "value"                :
"dbms_data_mining.odms_missing_value_mean_mode",
                            "categorical choices"   :
[    "dbms_data_mining.odms_missing_value_mean_mode",

"dbms_data_mining.odms_missing_value_auto",

"dbms_data_mining.odms_missing_value_delete_row"],
                            "description"                : {"English",
                                                                     "how to
treat missing values"}
                         },

{"name"                : "RALG_PARAMETER_FAMILY",
                            "name_display"   : {"English",
"RALG_PARAMETER_FAMILY"},
                            "type"                  : "string",
                            "optional"            :  "TRUE",
                            "value"                :  "",
                            "description"       : {"English", "R family
parameter in build function"}
                         }
],
                        }
```

REGISTER_ALGORITHM Procedure

Use this function to register a new algorithm by providing the algorithm name, machine learning function, and all other algorithm metadata.

Syntax

```
 DBMS_DATA_MINING.REGISTER_ALGORITHM (
                     algorithm_name          IN VARCHAR2,
                     algorithm_metadata      IN CLOB,
                     algorithm_description   IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-123 ***REGISTER_ALGORITHM Procedure Parameters***

Parameter	Description
algorithm_name	Name of the algorithm.
algorithm_metadata	Metadata of the algorithm.
algorithm_description	Description of the algorithm.

Usage Notes

The registration procedure performs the following:

- Checks whether algorithm_metadata has correct JSON syntax.

- Checks whether the input JSON object follows the predefined JSON schema.
- Checks whether current user has RQADMIN privilege.
- Checks duplicate algorithms so that the same algorithm is not registered twice.
- Checks for missing entries. For example, algorithm name, algorithm type, metadata, and build function.

Register Algorithms After the JSON Object Is Created

SQL users can register new algorithms by creating a JSON object following the JSON schema and passing it to the REGISTER_ALGORITHM procedure.

```
BEGIN
  DBMS_DATA_MINING.register_algorithm(
    algorithm_name                   =>   't1',
    algorithm_metadata              =>
    '{"function_language" : "R",
      "mining_function" :
        { "mining_function_name" : "CLASSIFICATION",
           "build_function" : {"function_body": "function(dat, formula,
family) { set.seed(1234);
                                          mod <- glm(formula = formula,
data=dat,

family=eval(parse(text=family)));
mod}"},
           "score_function" :  {"function_body": "function(mod, dat) {
                                             res <- predict(mod, newdata =
dat, type=''response'');
                                             res2=data.frame(1-res, res);
res2}"}}
    }',
    algorithm_description  => 't1');
END;
/
```

RANK_APPLY Procedure

This procedure ranks the results of an APPLY operation based on a top-N specification for predictive and descriptive model results.

For classification models, you can provide a cost matrix as input, and obtain the ranked results with costs applied to the predictions.

Syntax

```
DBMS_DATA_MINING.RANK_APPLY (
      apply_result_table_name        IN VARCHAR2,
      case_id_column_name            IN VARCHAR2,
      score_column_name              IN VARCHAR2,
      score_criterion_column_name    IN VARCHAR2,
      ranked_apply_table_name        IN VARCHAR2,
      top_N                          IN NUMBER (38) DEFAULT 1,
      cost_matrix_table_name         IN VARCHAR2    DEFAULT NULL,
      apply_result_schema_name       IN VARCHAR2    DEFAULT NULL,
      cost_matrix_schema_name        IN VARCHAR2    DEFAULT NULL);
```

Parameters

Table 6-124 RANK_APPLY Procedure Parameters

Parameter	Description
apply_result_table_name	Name of the table or view containing the results of an APPLY operation on the test data set (see Usage Notes)
case_id_column_name	Name of the case identifier column. This must be the same as the one used for generating APPLY results.
score_column_name	Name of the prediction column in the apply results table
score_criterion_column_name	Name of the probability column in the apply results table
ranked_apply_result_tab_name	Name of the table containing the ranked apply results
top_N	Top N predictions to be considered from the APPLY results for precision recall computation
cost_matrix_table_name	Name of the cost matrix table
apply_result_schema_name	Name of the schema hosting the APPLY results table
cost_matrix_schema_name	Name of the schema hosting the cost matrix table

Usage Notes

You can use RANK_APPLY to generate ranked apply results, based on a top-N filter and also with application of cost for predictions, if the model was built with costs.

The behavior of RANK_APPLY is similar to that of APPLY with respect to other DDL-like operations such as CREATE_MODEL, DROP_MODEL, and RENAME_MODEL. The procedure does not depend on the model; the only input of relevance is the apply results generated in a fixed schema table from APPLY.

The main intended use of RANK_APPLY is for the generation of the final APPLY results against the scoring data in a production setting. You can apply the model against test data using APPLY, compute various test metrics against various cost matrix tables, and use the candidate cost matrix for RANK_APPLY.

The schema for the apply results from each of the supported algorithms is listed in subsequent sections. The case_id column will be the same case identifier column as that of the apply results.

Classification Models — NB and SVM

For numerical targets, the ranked results table will have the definition as shown:

```
(case_id        VARCHAR2/NUMBER,
prediction      NUMBER,
probability     NUMBER,
cost            NUMBER,
rank            INTEGER)
```

For categorical targets, the ranked results table will have the following definition:

```
(case_id        VARCHAR2/NUMBER,
prediction      VARCHAR2,
probability     NUMBER,
cost            NUMBER,
rank            INTEGER)
```

Clustering Using *k*-Means or O-Cluster

Clustering is an unsupervised machine learning function, and hence there are no targets. The results of an `APPLY` operation contains simply the cluster identifier corresponding to a case, and the associated probability. Cost matrix is not considered here. The ranked results table will have the definition as shown, and contains the cluster ids ranked by `top-N`.

```
(case_id        VARCHAR2/NUMBER,
cluster_id      NUMBER,
probability     NUMBER,
rank            INTEGER)
```

Feature Extraction using NMF

Feature extraction is also an unsupervised machine learning function, and hence there are no targets. The results of an `APPLY` operation contains simply the feature identifier corresponding to a case, and the associated match quality. Cost matrix is not considered here. The ranked results table will have the definition as shown, and contains the feature ids ranked by `top-N`.

```
(case_id         VARCHAR2/NUMBER,
feature_id       NUMBER,
match_quality    NUMBER,
rank             INTEGER)
```

Examples

```
BEGIN
/* build a model with name census_model.
 * (See example under CREATE_MODEL)
 */

/* if training data was pre-processed in any manner,
 * perform the same pre-processing steps on apply
 * data also.
 * (See examples in the section on DBMS_DATA_MINING_TRANSFORM)
 */

/* apply the model to data to be scored */
DBMS_DATA_MINING.RANK_APPLY(
  apply_result_table_name       => 'census_apply_result',
  case_id_column_name           => 'person_id',
  score_column_name             => 'prediction',
  score_criterion_column_name   => 'probability
  ranked_apply_result_tab_name  => 'census_ranked_apply_result',
  top_N                         => 3,
  cost_matrix_table_name        => 'census_cost_matrix');
END;
/

-- View Ranked Apply Results
SELECT *
  FROM census_ranked_apply_result;
```

REMOVE_COST_MATRIX Procedure

The REMOVE_COST_MATRIX procedure removes the default scoring matrix from a classification model.

> **See Also:**
>
> - "ADD_COST_MATRIX Procedure"
> - "REMOVE_COST_MATRIX Procedure"

Syntax

```
DBMS_DATA_MINING.REMOVE_COST_MATRIX (
      model_name   IN  VARCHAR2);
```

Parameters

Table 6-125 Remove_Cost_Matrix Procedure Parameters

Parameter	Description
model_name	Name of the model in the form [*schema_name*.]*model_name*. If you do not specify a schema, your own schema is used.

Usage Notes

If the model is not in your schema, then REMOVE_COST_MATRIX requires the ALTER ANY MINING MODEL system privilege or the ALTER object privilege for the machine learning model.

Example

The naive Bayes model NB_SH_CLAS_SAMPLE has an associated cost matrix that can be used for scoring the model.

```
SQL>SELECT *
      FROM TABLE(dbms_data_mining.get_model_cost_matrix('nb_sh_clas_sample'))
      ORDER BY predicted, actual;

ACTUAL     PREDICTED        COST
---------- ---------- ----------
0          0                   0
1          0                 .75
0          1                 .25
1          1                   0
```

You can remove the cost matrix with REMOVE_COST_MATRIX.

```
SQL>EXECUTE dbms_data_mining.remove_cost_matrix('nb_sh_clas_sample');

SQL>SELECT *
      FROM TABLE(dbms_data_mining.get_model_cost_matrix('nb_sh_clas_sample'))
      ORDER BY predicted, actual;

no rows selected
```

RENAME_MODEL Procedure

This procedure changes the name of the machine learning model indicated by *model_name* to the name that you specify as *new_model_name*.

If a model with *new_model_name* already exists, then the procedure optionally renames *new_model_name* to *versioned_model_name* before renaming *model_name* to *new_model_name*.

The model name is in the form [*schema_name*.]*model_name*. If you do not specify a schema, your own schema is used. For machine learning model naming restrictions, see the Usage Notes for "CREATE_MODEL Procedure".

Syntax

```
DBMS_DATA_MINING.RENAME_MODEL (
     model_name            IN VARCHAR2,
     new_model_name        IN VARCHAR2,
     versioned_model_name  IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-126   RENAME_MODEL Procedure Parameters

Parameter	Description
`model_name`	Model to be renamed.
`new_model_name`	New name for the model *`model_name`*.
`versioned_model_name`	New name for the model *`new_model_name`* if it already exists.

Usage Notes

If you attempt to rename a model while it is being applied, then the model will be renamed but the apply operation will return indeterminate results.

Examples

1. This example changes the name of model `census_model` to `census_model_2012`.

```
BEGIN
  DBMS_DATA_MINING.RENAME_MODEL(
    model_name      => 'census_model',
    new_model_name  => 'census_model_2012');
END;
/
```

2. In this example, there are two classification models in the user's schema: `clas_mod`, the working model, and `clas_mod_tst`, a test model. The `RENAME_MODEL` procedure preserves `clas_mod` as `clas_mod_old` and makes the test model the new working model.

```
SELECT model_name FROM user_mining_models;
MODEL_NAME
-------------------------------------------------------------------
CLAS_MOD
CLAS_MOD_TST

BEGIN
  DBMS_DATA_MINING.RENAME_MODEL(
    model_name            => 'clas_mod_tst',
    new_model_name        => 'clas_mod',
    versioned_model_name  => 'clas_mod_old');
END;
/

SELECT model_name FROM user_mining_models;
MODEL_NAME
-------------------------------------------------------------------
CLAS_MOD
CLAS_MOD_OLD
```

DBMS_DATA_MINING_TRANSFORM

DBMS_DATA_MINING_TRANSFORM implements a set of transformations that are commonly used in machine learning.

This chapter contains the following topics:

- Overview
- Operational Notes
- Security Model
- Datatypes
- Constants
- Summary of DBMS_DATA_MINING_TRANSFORM Subprograms

See Also:

- DBMS_DATA_MINING
- *Oracle Machine Learning for SQL User's Guide*

DBMS_DATA_MINING_TRANSFORM Overview

A transformation is a SQL expression that modifies the data in one or more columns.

Data must typically undergo certain transformations before it can be used to build a machine learning model. Many machine learning algorithms have specific transformation requirements.

Data that will be scored must be transformed in the same way as the data that was used to create (train) the model.

External or Embedded Transformations

DBMS_DATA_MINING_TRANSFORM offers two approaches to implementing transformations. For a given model, you can either:

- Create a list of transformation expressions and pass it to the CREATE_MODEL Procedure

  or

- Create a view that implements the transformations and pass the name of the view to the CREATE_MODEL Procedure

If you create a transformation list and pass it to `CREATE_MODEL`, the transformation expressions are embedded in the model and automatically implemented whenever the model is applied.

If you create a view, the transformation expressions are external to the model. You will need to re-create the transformations whenever you apply the model.

> **Note:**
>
> Embedded transformations significantly enhance the model's usability while simplifying the process of model management.

Automatic Transformations

Oracle Machine Learning for SQL supports an Automatic Data Preparation (ADP) mode. When ADP is enabled, most algorithm-specific transformations are *automatically* embedded. Any additional transformations must be explicitly provided in an embedded transformation list or in a view.

If ADP is enabled and you create a model with a transformation list, both sets of transformations are embedded. The model will execute the user-specified transformations from the transformation list before executing the automatic transformations specified by ADP.

Within a transformation list, you can selectively disable ADP for individual attributes.

> **See Also:**
>
> "Automatic Data Preparation"
>
> *Oracle Machine Learning for SQL User’s Guide* for a more information about ADP
>
> "DBMS_DATA_MINING_TRANSFORM-About Transformation Lists"

Transformations in DBMS_DATA_MINING_TRANSFORM

The transformations supported by `DBMS_DATA_MINING_TRANSFORM` are summarized in this section.

Binning

Binning refers to the mapping of continuous or discrete values to discrete values of reduced cardinality.

- Supervised Binning (Categorical and Numerical)

  Binning is based on intrinsic relationships in the data as determined by a decision tree model.

  See "INSERT_BIN_SUPER Procedure".
- Top-N Frequency Categorical Binning

  Binning is based on the number of cases in each category.

  See "INSERT_BIN_CAT_FREQ Procedure"
- Equi-Width Numerical Binning

Binning is based on equal-range partitions.

See "INSERT_BIN_NUM_EQWIDTH Procedure".

- Quantile Numerical Binning

  Binning is based on quantiles computed using the SQL `NTILE` function.

  See "INSERT_BIN_NUM_QTILE Procedure".

Linear Normalization

Normalization is the process of scaling continuous values down to a specific range, often between zero and one. Normalization transforms each numerical value by subtracting a number (the **shift**) and dividing the result by another number (the **scale**).

```
x_new = (x_old-shift)/scale
```

- Min-Max Normalization

  Normalization is based on the minimum and maximum with the following shift and scale:

  ```
  shift = min
  scale = max-min
  ```

  See "INSERT_NORM_LIN_MINMAX Procedure".

- Scale Normalization

  Normalization is based on the minimum and maximum with the following shift and scale:

  ```
  shift = 0
  scale = max{abs(max), abs(min)}
  ```

  See "INSERT_NORM_LIN_SCALE Procedure".

- Z-Score Normalization

  Normalization is based on the mean and standard deviation with the following shift and scale:

  ```
  shift = mean
  scale = standard_deviation
  ```

  See "INSERT_NORM_LIN_ZSCORE Procedure".

Outlier Treatment

An outlier is a numerical value that is located far from the rest of the data. Outliers can artificially skew the results of machine learning.

- Winsorizing

  Outliers are replaced with the nearest value that is not an outlier.

  See "INSERT_CLIP_WINSOR_TAIL Procedure"

- Trimming

  Outliers are set to `NULL`.

  See "INSERT_CLIP_TRIM_TAIL Procedure".

Missing Value Treatment

Missing data may indicate sparsity or it may indicate that some values are missing at random. `DBMS_DATA_MINING_TRANSFORM` supports the following transformations for minimizing the effects of missing values:

- Missing numerical values are replaced with the mean.

  See "INSERT_MISS_NUM_MEAN Procedure".
- Missing categorical values are replaced with the mode.

  See "INSERT_MISS_CAT_MODE Procedure".

> **Note:**
>
> Oracle Machine Learning for SQL also has default mechanisms for handling missing data. See *Oracle Machine Learning for SQL User's Guide* for details.

DBMS_DATA_MINING_TRANSFORM Operational Notes

The `DBMS_DATA_MINING_TRANSFORM` package offers a flexible framework for specifying data transformations. If you choose to embed transformations in the model (the preferred method), you create a **transformation list** object and pass it to the CREATE_MODEL Procedure. If you choose to transform the data without embedding, you create a view.

When specified in a transformation list, the transformation expressions are run by the model. When specified in a view, the transformation expressions are run by the view.

Transformation Definitions

Transformation definitions are used to generate the SQL expressions that transform the data. For example, the transformation definitions for normalizing a numeric column are the shift and scale values for that data.

With the `DBMS_DATA_MINING_TRANSFORM` package, you can call procedures to compute the transformation definitions, or you can compute them yourself, or you can do both.

Transformation Definition Tables

`DBMS_DATA_MINING_TRANSFORM` provides **`INSERT`** procedures that compute transformation definitions and insert them in transformation definition tables. You can modify the values in the transformation definition tables or populate them yourself.

`XFORM` routines use populated definition tables to transform data in external views. **`STACK`** routines use populated definition tables to build transformation lists.

To specify transformations based on definition tables, follow these steps:

1. Use **`CREATE`** routines to create transformation definition tables.

   The tables have columns to hold the transformation definitions for a given type of transformation. For example, the CREATE_BIN_NUM Procedure creates a definition table that has a column for storing data values and another column for storing the associated bin identifiers.
2. Use **`INSERT`** routines to compute and insert transformation definitions in the tables.

   Each `INSERT` routine uses a specific technique for computing the transformation definitions. For example, the INSERT_BIN_NUM_EQWIDTH Procedure computes bin boundaries by identifying the minimum and maximum values then setting the bin boundaries at equal intervals.
3. Use **`STACK`** or **`XFORM`** routines to generate transformation expressions based on the information in the definition tables:

- Use **STACK** routines to add the transformation expressions to a transformation list. Pass the transformation list to the CREATE_MODEL Procedure. The transformation expressions will be assembled into one long SQL query and embedded in the model.
- Use **XFORM** routines to execute the transformation expressions within a view. The transformations will be external to the model and will need to be re-created whenever the model is applied to new data.

Transformations Without Definition Tables

STACK routines are not the only method for adding transformation expressions to a transformation list. You can also build a transformation list without using definition tables.

To specify transformations without using definition tables, follow these steps:

1. Write a SQL expression for transforming an attribute.
2. Write a SQL expression for reversing the transformation. (See "Reverse Transformations and Model Transparency" in "DBMS_DATA_MINING_TRANSFORM-About Transformation Lists".)
3. Determine whether or not to disable ADP for the attribute. By default ADP is enabled for the attribute if it is specified for the model. (See "Disabling Automatic Data Preparation" in "DBMS_DATA_MINING_TRANSFORM - About Transformation Lists".)
4. Specify the SQL expressions and ADP instructions in a call to the SET_TRANSFORM Procedure, which adds the information to a transformation list.
5. Repeat steps 1 through 4 for each attribute that you wish to transform.
6. Pass the transformation list to the CREATE_MODEL Procedure. The transformation expressions will be assembled into one long SQL query and embedded in the model.

> **Note:**
>
> SQL expressions that you specify with SET_TRANSFORM must fit within a VARCHAR2. To specify a longer expression, you can use the SET_EXPRESSION Procedure. With SET_EXPRESSION, you can build an expression by appending rows to a VARCHAR2 array.

About Stacking

Transformation lists are built by stacking transformation records. Transformation lists are evaluated from bottom to top. Each transformation expression depends on the result of the transformation expression below it in the stack.

Related Topics

- CREATE_MODEL Procedure
  This procedure creates an Oracle Machine Learning for SQL model with a given machine learning function.
- DBMS_DATA_MINING_TRANSFORM — About Transformation Lists
  The elements of a transformation list are **transformation records**. Each transformation record provides all the information needed by the model for managing the transformation of a single attribute.

- DBMS_DATA_MINING_TRANSFORM — About Stacking and Stack Procedures
  Transformation lists are built by stacking transformation records. Transformation lists are evaluated from bottom to top. Each transformation expression depends on the result of the transformation expression below it in the stack.
- DBMS_DATA_MINING_TRANSFORM — Nested Data Transformations
  The `CREATE` routines create transformation definition tables that include two columns, `col` and `att`, for identifying attributes.

DBMS_DATA_MINING_TRANSFORM — About Transformation Lists

The elements of a transformation list are **transformation records**. Each transformation record provides all the information needed by the model for managing the transformation of a single attribute.

Each transformation record includes the following fields:

- *`attribute_name`* — Name of the column of data to be transformed
- *`attribute_subname`* — Name of the nested attribute if *`attribute_name`* is a nested column, otherwise `NULL`
- *`expression`* — SQL expression for transforming the attribute
- *`reverse_expression`* — SQL expression for reversing the transformation
- *`attribute_spec`* — Identifies special treatment for the attribute during the model build. See Table 6-159 for details.

> **See Also:**
>
> - Table 6-127 for details about the `TRANSFORM_LIST` and `TRANSFORM_REC` object types
> - SET_TRANSFORM Procedure
> - CREATE_MODEL Procedure

Reverse Transformations and Model Transparency

An algorithm manipulates transformed attributes to train and score a model. The transformed attributes, however, may not be meaningful to an end user. For example, if attribute *x* has been transformed into bins 1 — 4, the bin names 1, 2 , 3, and 4 are manipulated by the algorithm, but a user is probably not interested in the model details about bins 1 — 4 or in predicting the numbers 1 — 4.

To return original attribute values in model details and predictions, you can provide a reverse expression in the transformation record for the attribute. For example, if you specify the transformation expression **`'log(10, y)'`** for attribute *y*, you could specify the reverse transformation expression **`'power(10, y)'`**.

Reverse transformations enable **model transparency**. They make internal processing transparent to the user.

Note:

STACK procedures automatically reverse normalization transformations, but they do not provide a mechanism for reversing binning, clipping, or missing value transformations.

You can use the DBMS_DATA_MINING.ALTER_REVERSE_EXPRESSION procedure to specify or update reverse transformations expressions for an existing model.

See Also:

Table 6-127

"ALTER_REVERSE_EXPRESSION Procedure"

"Summary of DBMS_DATA_MINING Subprograms" for links to the model details functions

Disabling Automatic Data Preparation

ADP is controlled by a model-specific setting (PREP_AUTO). The PREP_AUTO setting affects all model attributes unless you disable it for individual attributes.

If ADP is enabled and you set *attribute_spec* to NOPREP, only the transformations that you specify for that attribute will be evaluated. If ADP is enabled and you do *not* set *attribute_spec* to NOPREP, the automatic transformations will be evaluated *after* the transformations that you specify for the attribute.

If ADP is not enabled for the model, the *attribute_spec* field of the transformation record is ignored.

See Also:

"Automatic Data Preparation" for information about the PREP_AUTO setting

Adding Transformation Records to a Transformation List

A transformation list is a stack of transformation records. When a new transformation record is added, it is appended to the top of the stack. (See "About Stacking" for details.)

When you use SET_TRANSFORM to add a transformation record to a transformation list, you can specify values for all the fields in the transformation record.

When you use STACK procedures to add transformation records to a transformation list, only the transformation expression field is populated. For normalization transformations, the reverse transformation expression field is also populated.

You can use both STACK procedures and SET_TRANSFORM to build one transformation list. Each STACK procedure call adds transformation records for all the attributes in a specified transformation definition table. Each SET_TRANSFORM call adds a transformation record for a single attribute.

DBMS_DATA_MINING_TRANSFORM — About Stacking and Stack Procedures

Transformation lists are built by stacking transformation records. Transformation lists are evaluated from bottom to top. Each transformation expression depends on the result of the transformation expression below it in the stack.

Stack Procedures

STACK procedures create transformation records from the information in transformation definition tables. For example STACK_BIN_NUM builds a transformation record for each attribute specified in a definition table for numeric binning. STACK procedures stack the transformation records as follows:

- If an attribute is specified in the definition table but not in the transformation list, the STACK procedure creates a transformation record, computes the reverse transformation (if possible), inserts the transformation and reverse transformation in the transformation record, and appends the transformation record to the top of the transformation list.
- If an attribute is specified in the transformation list but not in the definition table, the STACK procedure takes no action.
- If an attribute is specified in the definition table *and* in the transformation list, the STACK procedure stacks the transformation expression from the definition table on top of the transformation expression in the transformation record and updates the reverse transformation. See Table 6-127and Example 6-25.

Example 6-22 Stacking a Clipping Transformation

This example shows how STACK_CLIP Procedure would add transformation records to a transformation list. Note that the clipping transformations are not reversed in COL1 and COL2 after stacking (as described in "Reverse Transformations and Model Transparency" in "DBMS_DATA_MINING_TRANSFORM-About Transformation Lists").

Refer to:

- CREATE_CLIP Procedure — Creates the definition table
- INSERT_CLIP_TRIM_TAIL Procedure — Inserts definitions in the table
- INSERT_CLIP_WINSOR_TAIL Procedure — Inserts definitions in the table
- Table 6-127 — Describes the structure of the transformation list (TRANSFORM_LIST object)

Assume a clipping definition table populated as follows.

col	att	lcut	lval	rcut	rval
COL1	null	-1.5	-1.5	4.5	4.5
COL2	null	0	0	1	1

Assume the following transformation list before stacking.

```
-------------------------
transformation record #1:
-------------------------
     attribute_name       = COL1
     attribute_subname    = null
     expression           = log(10, COL1)
     reverse_expression   = power(10, COL1)
-------------------------
transformation record #2:
-------------------------
     attribute_name       = COL3
     attribute_subname    = null
     expression           = ln(COL3)
     reverse_expression   = exp(COL3)
```

After stacking, the transformation list is as follows.

```
-------------------------
transformation record #1:
-------------------------
     attribute_name       = COL1
     attribute_subname    = null
     expression           = CASE WHEN log(10, COL1) < -1.5 THEN -1.5
                                 WHEN log(10, COL1) > 4.5  THEN 4.5
                                 ELSE log(10, COL1)
                            END;
     reverse_expression   = power(10, COL1)
-------------------------
transformation record #2:
-------------------------
     attribute_name       = COL3
     attribute_subname    = null
     expression           = ln(COL3)
     reverse_expression   = exp(COL3)
-------------------------
transformation record #3:
-------------------------
     attribute_name       = COL2
     attribute_subname    = null
     expression           = CASE WHEN COL2 < 0 THEN 0
                                 WHEN COL2 > 1 THEN 1
                                 ELSE COL2
                            END;
     reverse_expression   = null
```

DBMS_DATA_MINING_TRANSFORM — Nested Data Transformations

The `CREATE` routines create transformation definition tables that include two columns, `col` and `att`, for identifying attributes.

The column `col` holds the name of a column in the data table. If the data column is not nested, then `att` is null, and the name of the attribute is *`col`*. If the data column is nested, then `att` holds the name of the nested attribute, and the name of the attribute is *`col.att`*. The `INSERT` and `XFORM` routines ignore the `att` column in the definition tables. Neither the `INSERT` nor the `XFORM` routines support nested data.

Only the `STACK` procedures and `SET_TRANSFORM` support nested data. Nested data transformations are always embedded in the model.

Nested columns in Oracle Machine Learning for SQL can have the following types:

```
DM_NESTED_NUMERICALS
DM_NESTED_CATEGORICALS
DM_NESTED_BINARY_DOUBLES
DM_NESTED_BINARY_FLOATS
```

See Also:

"Constants"

Oracle Machine Learning for SQL User's Guide for details about nested attributes in Oracle Machine Learning for SQL

Specifying Nested Attributes in a Transformation Record

A transformation record (TRANSFORM_REC) includes two fields, attribute_name and attribute_subname, for identifying the attribute. The field attribute_name holds the name of a column in the data table. If the data column is not nested, then attribute_subname is null, and the name of the attribute is *attribute_name*. If the data column is nested, then attribute_subname holds the name of the nested attribute, and the name of the attribute is *attribute_name.attribute_subname*.

Transforming Individual Nested Attributes

You can specify different transformations for different attributes in a nested column, and you can specify a default transformation for all the remaining attributes in the column. To specify a default nested transformation, specify null in the attribute_name field and the name of the nested column in the attribute_subname field as shown in Example 6-23. Note that the keyword VALUE is used to represent the value of a nested attribute in a transformation expression.

Example 6-23   Transforming a Nested Column

The following statement transforms two of the nested attributes in COL_N1. Attribute ATTR1 is transformed with normalization; Attribute ATTR2 is set to null, which causes attribute removal transformation (ATTR2 is not used in training the model). All the remaining attributes in COL_N1 are divided by 10.

```
DECLARE
  stk dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
  dbms_data_mining_transform.SET_TRANSFORM(
      stk,'COL_N1', 'ATTR1', '(VALUE - (-1.5))/20', 'VALUE *20 + (-1.5)');
  dbms_data_mining_transform.SET_TRANSFORM(
      stk,'COL_N1', 'ATTR2', NULL, NULL);
  dbms_data_mining_transform.SET_TRANSFORM(
      stk, NULL, 'COL_N1', 'VALUE/10', 'VALUE*10');
END;
/
```

The following SQL is generated from this statement.

```
CAST(MULTISET(SELECT DM_NESTED_NUMERICAL(
                               "ATTRIBUTE_NAME",
                               DECODE("ATTRIBUTE_NAME",
                                 'ATTR1', ("VALUE" - (-1.5))/20,
                                 "VALUE"/10))
                        FROM TABLE("COL_N1")
                       WHERE "ATTRIBUTE_NAME" IS NOT IN ('ATTR2'))
             AS DM_NESTED_NUMERICALS)
```

If transformations are not specified for COL_N1.ATTR1 and COL_N1.ATTR2, then the default transformation is used for all the attributes in COL_N1, and the resulting SQL does not include a DECODE.

```
    CAST(MULTISET(SELECT DM_NESTED_NUMERICAL(
                               "ATTRIBUTE_NAME",
                               "VALUE"/10)
                        FROM TABLE("COL_N1"))
             AS DM_NESTED_NUMERICALS)
```

Since DECODE is limited to 256 arguments, multiple DECODE functions are nested to support an arbitrary number of individual nested attribute specifications.

Adding a Nested Column

You can specify a transformation that adds a nested column to the data, as shown in Example 6-24.

Example 6-24   Adding a Nested Column to a Transformation List

```
DECLARE
    v_xlst dbms_data_mining_transform.TRANSFORM_LIST;
  BEGIN
    dbms_data_mining_transform.SET_TRANSFORM(v_xlst,
      'YOB_CREDLIM', NULL,
      'dm_nested_numericals(
           dm_nested_numerical(
                 ''CUST_YEAR_OF_BIRTH'', cust_year_of_birth),
           dm_nested_numerical(
                 ''CUST_CREDIT_LIMIT'', cust_credit_limit))',
        NULL);
    dbms_data_mining_transform.SET_TRANSFORM(
              v_xlst, 'CUST_YEAR_OF_BIRTH', NULL, NULL, NULL);
    dbms_data_mining_transform.SET_TRANSFORM(
              v_xlst, 'CUST_CREDIT_LIMIT', NULL, NULL, NULL);
    dbms_data_mining_transform.XFORM_STACK(
              v_xlst, 'mining_data', 'mining_data_v');
END;
/

set long 2000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_V';

TEXT
---------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_POSTAL_CODE",dm_nested_numericals(
        dm_nested_numerical(
           'CUST_YEAR_OF_BIRTH', cust_year_of_birth),
        dm_nested_numerical(
           'CUST_CREDIT_LIMIT', cust_credit_limit)) "YOB_CREDLIM" FROM mining_data

SELECT * FROM mining_data_v WHERE cust_id = 104500;

CUST_ID CUST_POSTAL_CODE YOB_CREDLIM(ATTRIBUTE_NAME, VALUE)
------- ---------------- ---------------------------------------------------------
 104500 68524             DM_NESTED_NUMERICALS(DM_NESTED_NUMERICAL(
                         'CUST_YEAR_OF_BIRTH', 1962),
                          DM_NESTED_NUMERICAL('CUST_CREDIT_LIMIT', 15000))
```

Stacking Nested Transformations

Example 6-25 shows how the STACK_NORM_LIN Procedure would add transformation records for nested column COL_N to a transformation list.

Refer to:

- CREATE_NORM_LIN Procedure — Creates the definition table
- INSERT_NORM_LIN_MINMAX Procedure — Inserts definitions in the table
- INSERT_NORM_LIN_SCALE Procedure — Inserts definitions in the table
- INSERT_NORM_LIN_ZSCORE Procedure — Inserts definitions in the table
- Table 6-127 — Describes the structure of the transformation list

Example 6-25 Stacking a Nested Normalization Transformation

Assume a linear normalization definition table populated as follows.

col	att	shift	scale
COL_N	ATT2	0	20
null	COL_N	0	10

Assume the following transformation list before stacking.

```
-------------------------
transformation record #1:
-------------------------
     attribute_name       = COL_N
     attribute_subname    = ATT1
     expression           = log(10, VALUE)
     reverse_expression   = power(10, VALUE)
-------------------------
transformation record #2:
-------------------------
     attribute_name       = null
     attribute_subname    = COL_N
     expression           = ln(VALUE)
     reverse_expression   = exp(VALUE)
```

After stacking, the transformation list is as follows.

```
-------------------------
transformation record #1:
-------------------------
     attribute_name       = COL_N
     attribute_subname    = ATT1
     expression           = (log(10, VALUE) - 0)/10
     reverse_expression   = power(10, VALUE*10 + 0)
-------------------------
transformation record #2:
-------------------------
     attribute_name       = NULL
     attribute_subname    = COL_N
     expression           = (ln(VALUE)- 0)/10
     reverse_expression   = exp(VALUE *10 + 0)
-------------------------
transformation record #3:
-------------------------
```

```
attribute_name        = COL_N
attribute_subname     = ATT2
expression            = (ln(VALUE) - 0)/20
reverse_expression    = exp(VALUE * 20 + 0)
```

DBMS_DATA_MINING_TRANSFORM Security Model

The `DBMS_DATA_MINING_TRANSFORM` package is owned by user `SYS` and is installed as part of database installation. Execution privilege on the package is granted to public. The routines in the package are run with invokers' rights (run with the privileges of the current user).

The `DBMS_DATA_MINING_TRANSFORM.INSERT_*` procedures have a *`data_table_name`* parameter that enables the user to provide the input data for transformation purposes. The value of *`data_table_name`* can be the name of a physical table or a view. The *`data_table_name`* parameter can also accept an inline query.

> **Note:**
>
> Because an inline query can be used to specify the data for transformation, Oracle strongly recommends that the calling routine perform any necessary SQL injection checks on the input string.

> **See Also:**
>
> "Operational Notes" for a description of the `DBMS_DATA_MINING_TRANSFORM.INSERT_*` procedures

DBMS_DATA_MINING_TRANSFORM Datatypes

`DBMS_DATA_MINING_TRANSFORM` defines the datatypes described in the following table.

Table 6-127 Datatypes in DBMS_DATA_MINING_TRANSFORM

List Type	List Elements	Description
`COLUMN_LIST`	`VARRAY(1000) OF varchar2(32)`	`COLUMN_LIST` stores quoted and non-quoted identifiers for column names. `COLUMN_LIST` is the datatype of the *`exclude_list`* parameter in the `INSERT` procedures. See "INSERT_AUTOBIN_NUM_EQWIDTH Procedure" for an example. See *Oracle Database PL/SQL Language Reference* for information about populating `VARRAY` structures.

Table 6-127 (Cont.) Datatypes in DBMS_DATA_MINING_TRANSFORM

List Type	List Elements	Description
DESCRIBE_ LIST	DBMS_SQL.DESC_TAB2 TYPE desc_tab2 IS TABLE OF desc_rec2 INDEX BY BINARY_INTEGER TYPE desc_rec2 IS RECORD (col_type BINARY_INTEGER := 0, col_max_len BINARY_INTEGER := 0, col_name VARCHAR2(32767):= '', col_name_len BINARY_INTEGER := 0, col_schema_name VARCHAR2(32) := '', col_schema_name_len BINARY_INTEGER := 0, col_precision BINARY_INTEGER := 0, col_scale BINARY_INTEGER := 0, col_charsetid BINARY_INTEGER := 0, col_charsetform BINARY_INTEGER := 0, col_null_ok BOOLEAN := TRUE);	DESCRIBE_LIST describes the columns of the data table after the transformation list has been applied. A DESCRIBE_LIST is returned by the DESCRIBE_STACK Procedure. The DESC_TAB2 and DESC_REC2 types are defined in the DBMS_SQL package. See "DESC_REC2 Record Type". The col_type field of DESC_REC2 identifies the datatype of the column. The datatype is expressed as a numeric constant that represents a built-in datatype. For example, a 1 indicates a variable length character string. The codes for Oracle built-in datatypes are listed in *Oracle Database SQL Language Reference*. The codes for the Oracle Machine Learning for SQL nested types are described in "Constants". The col_name field of DESC_REC2 identifies the column name. It may be populated with a column name, an alias, or an expression. If the column name is a SELECT expression, it may be very long. If the expression is longer than 30 bytes, it cannot be used in a view unless it is given an alias.
TRANSFORM_ LIST	**TABLE OF transform_rec** **TYPE transform_rec IS RECORD** (attribute_name VARCHAR2(30), attribute_subname VARCHAR2(4000), expression **EXPRESSION_REC**, reverse_expression **EXPRESSION_REC**, attribute_spec VARCHAR2(4000)); **TYPE expression_rec IS RECORD** (lstmt DBMS_SQL.VARCHAR2A, lb BINARY_INTEGER DEFAULT 1, ub BINARY_INTEGER DEFAULT 0); TYPE varchar2a IS TABLE OF VARCHAR2(32767) INDEX BY BINARY_INTEGER;	TRANSFORM_LIST is a list of transformations that can be embedded in a model. A TRANSFORM_LIST is accepted as an argument by the CREATE_MODEL Procedure. Each element in a TRANSFORM_LIST is a TRANSFORM_REC that specifies how to transform a single attribute. The attribute_name is a column name. The attribute_subname is the nested attribute name if the column is nested, otherwise attribute_subname is null. The expression field holds a SQL expression for transforming the attribute. See "About Transformation Lists" for an explanation of reverse expressions. The attribute_spec field can be used to cause the attribute to be handled in a specific way during the model build. See Table 6-159 for details. The expressions in a TRANSFORM_REC have type EXPRESSION_REC. The lstmt field stores a VARCHAR2A, which is a table of VARCHAR2(32767). The VARCHAR2A datatype allows transformation expressions to be very long, as they can be broken up across multiple rows of VARCHAR2. The VARCHAR2A type is defined in the DBMS_SQL package. See "VARCHAR2A Table Type". The ub (upper bound) and lb (lower bound) fields indicate how many rows there are in the VARCHAR2A table. If ub < lb (default) the EXPRESSION_REC is empty; if lb=ub=1 there is one row; if lb=1 and ub=2 there are 2 rows, and so on.

DBMS_DATA_MINING_TRANSFORM Constants

DBMS_DATA_MINING_TRANSFORM defines the constants described in the following table.

Table 6-128 Constants in DBMS_DATA_MINING_TRANSFORM

Constant	Value	Description

`NEST_NUM_COL_TYPE`	100001	Indicates that an attribute in the transformation list comes from a row in a column of `DM_NESTED_NUMERICALS`.
		Nested numerical attributes are defined as follows:
		`attribute_name        VARCHAR2(4000)`
		`value                 NUMBER`

`NEST_CAT_COL_TYPE`	100002	Indicates that an attribute in the transformation list comes from a row in a column of `DM_NESTED_CATEGORICALS`.
		Nested categorical attributes are defined as follows:
		`attribute_name        VARCHAR2(4000)`
		`value                 VARCHAR2(4000)`

`NEST_BD_COL_TYPE`	100003	Indicates that an attribute in the transformation list comes from a row in a column of `DM_NESTED_BINARY_DOUBLES`.
		Nested binary double attributes are defined as follows:
		`attribute_name        VARCHAR2(4000)`
		`value                 BINARY_DOUBLE`

`NEST_BF_COL_TYPE`	100004	Indicates that an attribute in the transformation list comes from a row in a column of `DM_NESTED_BINARY_FLOATS`.
		`attribute_name        VARCHAR2(4000)`
		`value                 BINARY_FLOAT`

See Also:

Oracle Machine Learning for SQL User’s Guide for information about nested data in Oracle Machine Learning for SQL

Summary of DBMS_DATA_MINING_TRANSFORM Subprograms

This table lists the `DBMS_DATA_MINING_TRANSFORM` subprograms in alphabetical order and briefly describes them.

Table 6-129 DBMS_DATA_MINING_TRANSFORM Package Subprograms

Subprogram	Purpose
CREATE_BIN_CAT Procedure	Creates a transformation definition table for categorical binning
CREATE_BIN_NUM Procedure	Creates a transformation definition table for numerical binning
CREATE_CLIP Procedure	Creates a transformation definition table for clipping
CREATE_COL_REM Procedure	Creates a transformation definition table for column removal
CREATE_MISS_CAT Procedure	Creates a transformation definition table for categorical missing value treatment

Table 6-129 (Cont.) DBMS_DATA_MINING_TRANSFORM Package Subprograms

Subprogram	Purpose
CREATE_MISS_NUM Procedure	Creates a transformation definition table for numerical missing values treatment
CREATE_NORM_LIN Procedure	Creates a transformation definition table for linear normalization
DESCRIBE_STACK Procedure	Describes the transformation list
GET_EXPRESSION Function	Returns a VARCHAR2 chunk from a transformation expression
INSERT_AUTOBIN_NUM_EQWIDTH Procedure	Inserts numeric automatic equi-width binning definitions in a transformation definition table
INSERT_BIN_CAT_FREQ Procedure	Inserts categorical frequency-based binning definitions in a transformation definition table
INSERT_BIN_NUM_EQWIDTH Procedure	Inserts numeric equi-width binning definitions in a transformation definition table
INSERT_BIN_NUM_QTILE Procedure	Inserts numeric quantile binning expressions in a transformation definition table
INSERT_BIN_SUPER Procedure	Inserts supervised binning definitions in numerical and categorical transformation definition tables
INSERT_CLIP_TRIM_TAIL Procedure	Inserts numerical trimming definitions in a transformation definition table
INSERT_CLIP_WINSOR_TAIL Procedure	Inserts numerical winsorizing definitions in a transformation definition table
INSERT_MISS_CAT_MODE Procedure	Inserts categorical missing value treatment definitions in a transformation definition table
INSERT_MISS_NUM_MEAN Procedure	Inserts numerical missing value treatment definitions in a transformation definition table
INSERT_NORM_LIN_MINMAX Procedure	Inserts linear min-max normalization definitions in a transformation definition table
INSERT_NORM_LIN_SCALE Procedure	Inserts linear scale normalization definitions in a transformation definition table
INSERT_NORM_LIN_ZSCORE Procedure	Inserts linear zscore normalization definitions in a transformation definition table
SET_EXPRESSION Procedure	Adds a VARCHAR2 chunk to an expression
SET_TRANSFORM Procedure	Adds a transformation record to a transformation list
STACK_BIN_CAT Procedure	Adds a categorical binning expression to a transformation list
STACK_BIN_NUM Procedure	Adds a numerical binning expression to a transformation list
STACK_CLIP Procedure	Adds a clipping expression to a transformation list
STACK_COL_REM Procedure	Adds a column removal expression to a transformation list
STACK_MISS_CAT Procedure	Adds a categorical missing value treatment expression to a transformation list
STACK_MISS_NUM Procedure	Adds a numerical missing value treatment expression to a transformation list
STACK_NORM_LIN Procedure	Adds a linear normalization expression to a transformation list
XFORM_BIN_CAT Procedure	Creates a view of the data table with categorical binning transformations
XFORM_BIN_NUM Procedure	Creates a view of the data table with numerical binning transformations

Table 6-129 (Cont.) DBMS_DATA_MINING_TRANSFORM Package Subprograms

Subprogram	Purpose
XFORM_CLIP Procedure	Creates a view of the data table with clipping transformations
XFORM_COL_REM Procedure	Creates a view of the data table with column removal transformations
XFORM_EXPR_NUM Procedure	Creates a view of the data table with the specified numeric transformations
XFORM_EXPR_STR Procedure	Creates a view of the data table with the specified categorical transformations
XFORM_MISS_CAT Procedure	Creates a view of the data table with categorical missing value treatment
XFORM_MISS_NUM Procedure	Creates a view of the data table with numerical missing value treatment
XFORM_NORM_LIN Procedure	Creates a view of the data table with linear normalization transformations
XFORM_STACK Procedure	Creates a view of the transformation list

CREATE_BIN_CAT Procedure

This procedure creates a transformation definition table for categorical binning.

The columns are described in the following table.

Table 6-130 Columns in a Transformation Definition Table for Categorical Binning

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of categorical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User's Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`val`	`VARCHAR2(4000)`	Values of the attribute
`bin`	`VARCHAR2(4000)`	Bin assignments for the values

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_CAT (
     bin_table_name      IN VARCHAR2,
     bin_schema_name     IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-131 CREATE_BIN_CAT Procedure Parameters

Parameter	Description
`bin_table_name`	Name of the transformation definition table to be created

Table 6-131 (Cont.) CREATE_BIN_CAT Procedure Parameters

Parameter	Description
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about categorical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the following procedures to populate the transformation definition table:
   - INSERT_BIN_CAT_FREQ Procedure — frequency-based binning
   - INSERT_BIN_SUPER Procedure — supervised binning

> **See Also:**
>
> "Binning" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Examples

The following statement creates a table called `bin_cat_xtbl` in the current schema. The table has columns that can be populated with bin assignments for categorical attributes.

```
BEGIN
   DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_CAT('bin_cat_xtbl');
END;
/
DESCRIBE bin_cat_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 VAL                                                VARCHAR2(4000)
 BIN                                                VARCHAR2(4000)
```

CREATE_BIN_NUM Procedure

This procedure creates a transformation definition table for numerical binning.

The columns are described in the following table.

Table 6-132 Columns in a Transformation Definition Table for Numerical Binning

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of numerical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User's Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`val`	`NUMBER`	Values of the attribute
`bin`	`VARCHAR2(4000)`	Bin assignments for the values

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_NUM (
     bin_table_name     IN VARCHAR2,
     bin_schema_name    IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-133 CREATE_BIN_NUM Procedure Parameters

Parameter	Description
`bin_table_name`	Name of the transformation definition table to be created
`bin_schema_name`	Schema of *`bin_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the following procedures to populate the transformation definition table:
   - INSERT_AUTOBIN_NUM_EQWIDTH Procedure — automatic equi-width binning
   - INSERT_BIN_NUM_EQWIDTH Procedure — user-specified equi-width binning
   - INSERT_BIN_NUM_QTILE Procedure — quantile binning
   - INSERT_BIN_SUPER Procedure — supervised binning

> **See Also:**
>
> "Binning" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Examples

The following statement creates a table called `bin_num_xtbl` in the current schema. The table has columns that can be populated with bin assignments for numerical attributes.

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_NUM('bin_num_xtbl');
END;
/

DESCRIBE bin_num_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 VAL                                                NUMBER
 BIN                                                VARCHAR2(4000)
```

CREATE_CLIP Procedure

This procedure creates a transformation definition table for clipping or winsorizing to minimize the effect of outliers.

The columns are described in the following table.

Table 6-134 Columns in a Transformation Definition Table for Clipping or Winsorizing

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of numerical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User’s Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column of `DM_NESTED_NUMERICALS`. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`lcut`	`NUMBER`	The lowest typical value for the attribute. If the attribute values were plotted on an *xy* axis, *`lcut`* would be the left-most boundary of the range of values considered typical for this attribute. Any values to the left of *`lcut`* are outliers.
`lval`	`NUMBER`	Value assigned to an outlier to the left of *`lcut`*
`rcut`	`NUMBER`	The highest typical value for the attribute If the attribute values were plotted on an *xy* axis, *`rcut`* would be the right-most boundary of the range of values considered typical for this attribute. Any values to the right of *`rcut`* are outliers.
`rval`	`NUMBER`	Value assigned to an outlier to the right of *`rcut`*

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_CLIP (
     clip_table_name     IN VARCHAR2,
     clip_schema_name    IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-135 CREATE_CLIP Procedure Parameters

Parameter	Description
`clip_table_name`	Name of the transformation definition table to be created
`clip_schema_name`	Schema of *`clip_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the following procedures to populate the transformation definition table:
   - INSERT_CLIP_TRIM_TAIL Procedure — replaces outliers with nulls
   - INSERT_CLIP_WINSOR_TAIL Procedure — replaces outliers with an average value

> **See Also:**
>
> "Outlier Treatment" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Examples

The following statement creates a table called `clip_xtbl` in the current schema. The table has columns that can be populated with clipping instructions for numerical attributes.

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.CREATE_CLIP('clip_xtbl');
END;
/

DESCRIBE clip_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 LCUT                                               NUMBER
 LVAL                                               NUMBER
 RCUT                                               NUMBER
 RVAL                                               NUMBER
```

CREATE_COL_REM Procedure

This procedure creates a transformation definition table for removing columns from the data table.

The columns are described in the following table.

Table 6-136 Columns in a Transformation Definition Table for Column Removal

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User's Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is nested (`DM_NESTED_NUMERICALS` or `DM_NESTED_CATEGORICALS`). If col is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_COL_REM (
      rem_table_name           VARCHAR2,
      rem_schema_name          VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-137 CREATE_COL_REM Procedure Parameters

Parameter	Description
`rem_table_name`	Name of the transformation definition table to be created
`rem_schema_name`	Schema of *`rem_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See "Nested Data Transformations" for information about transformation definition tables and nested data.
2. See "Operational Notes".

Examples

The following statement creates a table called `rem_att_xtbl` in the current schema. The table has columns that can be populated with the names of attributes to exclude from the data to be mined.

```
BEGIN
    DBMS_DATA_MINING_TRANSFORM.CREATE_COL_REM ('rem_att_xtbl');
END;
 /
DESCRIBE rem_att_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
```

CREATE_MISS_CAT Procedure

This procedure creates a transformation definition table for replacing categorical missing values.

The columns are described in the following table.

Table 6-138 Columns in a Transformation Definition Table for Categorical Missing Value Treatment

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of categorical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User's Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column of `DM_NESTED_CATEGORICALS`. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`val`	`VARCHAR2(4000)`	Replacement for missing values in the attribute

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_CAT (
     miss_table_name       IN VARCHAR2,
     miss_schema_name      IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-139 CREATE_MISS_CAT Procedure Parameters

Parameter	Description
`miss_table_name`	Name of the transformation definition table to be created
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about categorical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the INSERT_MISS_CAT_MODE Procedure to populate the transformation definition table.

> **See Also:**
>
> "Missing Value Treatment" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Examples

The following statement creates a table called `miss_cat_xtbl` in the current schema. The table has columns that can be populated with values for missing data in categorical attributes.

```
BEGIN

  DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_CAT('miss_cat_xtbl');
END;
/

DESCRIBE miss_cat_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 VAL                                                VARCHAR2(4000)
```

CREATE_MISS_NUM Procedure

This procedure creates a transformation definition table for replacing numerical missing values.

The columns are described in Table 6-140.

Table 6-140 Columns in a Transformation Definition Table for Numerical Missing Value Treatment

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of numerical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User's Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column of `DM_NESTED_NUMERICALS`. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`val`	`NUMBER`	Replacement for missing values in the attribute

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_NUM (
     miss_table_name       IN VARCHAR2,
     miss_schema_name      IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-141 CREATE_MISS_NUM Procedure Parameters

Parameter	Description
`miss_table_name`	Name of the transformation definition table to be created
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User’s Guide* for details about numerical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the INSERT_MISS_NUM_MEAN Procedure to populate the transformation definition table.

> **See Also:**
>
> "Missing Value Treatment" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Example

The following statement creates a table called `miss_num_xtbl` in the current schema. The table has columns that can be populated with values for missing data in numerical attributes.

```
BEGIN
    DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_NUM('miss_num_xtbl');
END;
/

DESCRIBE miss_num_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 VAL                                                NUMBER
```

CREATE_NORM_LIN Procedure

This procedure creates a transformation definition table for linear normalization.

The columns are described in Table 6-142.

Table 6-142 Columns in a Transformation Definition Table for Linear Normalization

Name	Datatype	Description
`col`	`VARCHAR2(30)`	Name of a column of numerical data. If the column is not nested, the column name is also the attribute name. For information about attribute names, see *Oracle Machine Learning for SQL User’s Guide*.
`att`	`VARCHAR2(4000)`	The attribute subname if *`col`* is a nested column of `DM_NESTED_NUMERICALS`. If *`col`* is nested, the attribute name is *`col.att`*. If *`col`* is not nested, *`att`* is null.
`shift`	`NUMBER`	A constant to subtract from the attribute values
`scale`	`NUMBER`	A constant by which to divide the shifted values

Syntax

```
DBMS_DATA_MINING_TRANSFORM.CREATE_NORM_LIN (
     norm_table_name       IN VARCHAR2,
     norm_schema_name      IN VARCHAR2 DEFAULT NULL );
```

Parameters

Table 6-143 CREATE_NORM_LIN Procedure Parameters

Parameter	Description
`norm_table_name`	Name of the transformation definition table to be created
`norm_schema_name`	Schema of *`norm_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. See "Nested Data Transformations" for information about transformation definition tables and nested data.
3. You can use the following procedures to populate the transformation definition table:
   - INSERT_NORM_LIN_MINMAX Procedure — Uses linear min-max normalization
   - INSERT_NORM_LIN_SCALE Procedure — Uses linear scale normalization
   - INSERT_NORM_LIN_ZSCORE Procedure — Uses linear zscore normalization

> **See Also:**
>
> "Linear Normalization" in DBMS_DATA_MINING_TRANSFORM Overview
>
> "Operational Notes"

Examples

The following statement creates a table called `norm_xtbl` in the current schema. The table has columns that can be populated with shift and scale values for normalizing numerical attributes.

```
BEGIN
     DBMS_DATA_MINING_TRANSFORM.CREATE_NORM_LIN('norm_xtbl');
END;
/

DESCRIBE norm_xtbl
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 COL                                                VARCHAR2(30)
 ATT                                                VARCHAR2(4000)
 SHIFT                                              NUMBER
 SCALE                                              NUMBER
```

DESCRIBE_STACK Procedure

This procedure describes the columns of the data table after a list of transformations has been applied.

Only the columns that are specified in the transformation list are transformed. The remaining columns in the data table are included in the output without changes.

To create a view of the data table after the transformations have been applied, use the XFORM_STACK Procedure.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.DESCRIBE_STACK (
     xform_list           IN  TRANSFORM_LIST,
     data_table_name      IN  VARCHAR2,
     describe_list        OUT DESCRIBE_LIST,
     data_schema_name     IN  VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-144 DESCRIBE_STACK Procedure Parameters

Parameter	Description
xform_list	A list of transformations. See Table 6-127 for a description of the TRANSFORM_LIST object type.
data_table_name	Name of the table containing the data to be transformed
describe_list	Descriptions of the columns in the data table after the transformations specified in *xform_list* have been applied. See Table 6-127 for a description of the DESCRIBE_LIST object type.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes" for information about transformation lists and embedded transformations.

Examples

This example shows the column name and datatype, the column name length, and the column maximum length for the view oml_user.cust_info after the transformation list has been applied. All the transformations are user-specified. The results of DESCRIBE_STACK do not include one of the columns in the original table, because the SET_TRANSFORM procedure sets that column to NULL.

```
CREATE OR REPLACE VIEW cust_info AS
         SELECT a.cust_id, c.country_id, c.cust_year_of_birth,
         CAST(COLLECT(DM_Nested_Numerical(
                  b.prod_name, 1))
                AS DM_Nested_Numericals) custprods
                 FROM sh.sales a, sh.products b, sh.customers c
                  WHERE a.prod_id = b.prod_id AND
                        a.cust_id=c.cust_id and
                        a.cust_id between 100001 AND 105000
         GROUP BY a.cust_id, country_id, cust_year_of_birth;
```

```
describe cust_info
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 COUNTRY_ID                                NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUSTPRODS                                          SYS.DM_NESTED_NUMERICALS

DECLARE
  cust_stack   dbms_data_mining_transform.TRANSFORM_LIST;
  cust_cols    dbms_data_mining_transform.DESCRIBE_LIST;
BEGIN
  dbms_data_mining_transform.SET_TRANSFORM (cust_stack,
     'country_id', NULL, 'country_id/10', 'country_id*10');
  dbms_data_mining_transform.SET_TRANSFORM (cust_stack,
      'cust_year_of_birth', NULL, NULL, NULL);
  dbms_data_mining_transform.SET_TRANSFORM (cust_stack,
      'custprods', 'Mouse Pad', 'value*100', 'value/100');
  dbms_data_mining_transform.DESCRIBE_STACK(
       xform_list => cust_stack,
       data_table_name => 'cust_info',
       describe_list => cust_cols);
  dbms_output.put_line('====');
  for i in 1..cust_cols.COUNT loop
    dbms_output.put_line('COLUMN_NAME:     '||cust_cols(i).col_name);
    dbms_output.put_line('COLUMN_TYPE:     '||cust_cols(i).col_type);
    dbms_output.put_line('COLUMN_NAME_LEN: '||cust_cols(i).col_name_len);
    dbms_output.put_line('COLUMN_MAX_LEN:  '||cust_cols(i).col_max_len);
    dbms_output.put_line('====');
  END loop;
END;
/
====
COLUMN_NAME:     CUST_ID
COLUMN_TYPE:     2
COLUMN_NAME_LEN: 7
COLUMN_MAX_LEN:  22
====
COLUMN_NAME:     COUNTRY_ID
COLUMN_TYPE:     2
COLUMN_NAME_LEN: 10
COLUMN_MAX_LEN:  22
====
COLUMN_NAME:     CUSTPRODS
COLUMN_TYPE:     100001
COLUMN_NAME_LEN: 9
COLUMN_MAX_LEN:  40
====
```

GET_EXPRESSION Function

This function returns a row from a VARCHAR2 array that stores a transformation expression. The array is built by calls to the SET_EXPRESSION Procedure.

The array can be used for specifying SQL expressions that are too long to be used with the SET_TRANSFORM Procedure.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.GET_EXPRESSION (
     expression           IN EXPRESSION_REC,
     chunk_num            IN PLS_INTEGER DEFAULT NULL);
RETURN VARCHAR2;
```

Parameters

Table 6-145 GET_EXPRESSION Function Parameters

Parameter	Description
`expression`	An expression record (`EXPRESSION_REC`) that specifies a transformation expression or a reverse transformation expression for an attribute. Each expression record includes a `VARCHAR2` array and index fields for specifying upper and lower boundaries within the array. There are two `EXPRESSION_REC` fields within a transformation record (`TRANSFORM_REC`): one for the transformation expression; the other for the reverse transformation expression. See Table 6-127 for a description of the `EXPRESSION_REC` type.
`chunk`	A `VARCHAR2` chunk (row) to be appended to *`expression`*.

Usage Notes

1. Chunk numbering starts with one. For chunks outside of the range, the return value is null. When a chunk number is null the whole expression is returned as a string. If the expression is too big, a `VALUE_ERROR` is raised.
2. See "About Transformation Lists".
3. See "Operational Notes".

Examples

See the example for the SET_EXPRESSION Procedure.

Related Topics

- SET_EXPRESSION Procedure
  This procedure appends a row to a `VARCHAR2` array that stores a SQL expression.
- SET_TRANSFORM Procedure
  This procedure appends the transformation instructions for an attribute to a transformation list.

INSERT_AUTOBIN_NUM_EQWIDTH Procedure

This procedure performs numerical binning and inserts the transformation definitions in a transformation definition table. The procedure identifies the minimum and maximum values and computes the bin boundaries at equal intervals.

`INSERT_AUTOBIN_NUM_EQWIDTH` computes the number of bins separately for each column. If you want to use equi-width binning with the same number of bins for each column, use the INSERT_BIN_NUM_EQWIDTH Procedure.

`INSERT_AUTOBIN_NUM_EQWIDTH` bins all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_AUTOBIN_NUM_EQWIDTH (
     bin_table_name         IN VARCHAR2,
     data_table_name        IN VARCHAR2,
     bin_num                IN PLS_INTEGER DEFAULT 3,
     max_bin_num            IN PLS_INTEGER DEFAULT 100,
     exclude_list           IN COLUMN_LIST DEFAULT NULL,
     round_num              IN PLS_INTEGER DEFAULT 6,
     sample_size            IN PLS_INTEGER DEFAULT 50000,
     bin_schema_name        IN VARCHAR2 DEFAULT NULL,
     data_schema_name       IN VARCHAR2 DEFAULT NULL,
     rem_table_name         IN VARCHAR2 DEFAULT NULL,
     rem_schema_name        IN VARCHAR2 DEFAULT NULL));
```

Parameters

Table 6-146 INSERT_AUTOBIN_NUM_EQWIDTH Procedure Parameters

Parameter	Description

bin_table_name	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The following columns are required:
	`COL       VARCHAR2(30)`
	`VAL       NUMBER`
	`BIN       VARCHAR2(4000)`
	CREATE_BIN_NUM creates an additional column, ATT, which may be used for specifying nested attributes. This column is not used by INSERT_AUTOBIN_NUM_EQWIDTH.

data_table_name	Name of the table containing the data to be transformed
bin_num	Minimum number of bins. If *bin_num* is 0 or NULL, it is ignored. The default value of *bin_num* is 3.
max_bin_num	Maximum number of bins. If *max_bin_num* is 0 or NULL, it is ignored. The default value of *max_bin_num* is 100.
exclude_list	List of numerical columns to be excluded from the binning process. If you do not specify *exclude_list*, all numerical columns in the data source are binned. The format of *exclude_list* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`
round_num	Specifies how to round the number in the VAL column of the transformation definition table. When *round_num* is positive, it specifies the most significant digits to retain. When *round_num* is negative, it specifies the least significant digits to remove. In both cases, the result is rounded to the specified number of digits. See the Usage Notes for an example. The default value of *round_num* is 6.
sample_size	Size of the data sample. If *sample_size* is less than the total number of non-NULL values in the column, then *sample_size* is used instead of the SQL COUNT function in computing the number of bins. If *sample_size* is 0 or NULL, it is ignored. See the Usage Notes. The default value of *sample_size* is 50,000.
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.

Table 6-146 (Cont.) INSERT_AUTOBIN_NUM_EQWIDTH Procedure Parameters

Parameter	Description
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.
rem_table_name	Name of a transformation definition table for column removal. The table must have the columns described in "CREATE_COL_REM Procedure". INSERT_AUTOBIN_NUM_EQWIDTH ignores columns with all nulls or only one unique value. If you specify a value for *rem_table_name*, these columns are removed from the mining data. If you do not specify a value for *rem_table_name*, these unbinned columns remain in the data.
rem_schema_name	Schema of *rem_table_name*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. INSERT_AUTOBIN_NUM_EQWIDTH computes the number of bins for a column based on the number of non-null values (COUNT), the maximum (MAX), the minimum (MIN), the standard deviation (STDDEV), and the constant C=3.49/0.9:

   ```
   N=floor(power(COUNT,1/3)*(max-min)/(c*dev))
   ```

   If the *sample_size* parameter is specified, it is used instead of COUNT.

   See *Oracle Machine Learning for SQL User's Guide* for information about the COUNT, MAX, MIN, STDDEV, FLOOR, and POWER functions.
3. INSERT_AUTOBIN_NUM_EQWIDTH uses absolute values to compute the number of bins. The sign of the parameters *bin_num*, *max_bin_num*, and *sample_size* has no effect on the result.
4. In computing the number of bins, INSERT_AUTOBIN_NUM_EQWIDTH evaluates the following criteria in the following order:
   a. The minimum number of bins (*bin_num*)
   b. The maximum number of bins (*max_bin_num*)
   c. The maximum number of bins for integer columns, calculated as the number of distinct values in the range *max-min+1*.
5. The *round_num* parameter controls the rounding of column values in the transformation definition table, as follows:

   ```
   For a value of 308.162:
   when round_num =  1       result is 300
   when round_num =  2       result is 310
   when round_num =  3       result is 308
   when round_num =  0       result is 308.162
   when round_num = -1       result is 308.16
   when round_num = -2       result is 308.2
   ```

Examples

In this example, INSERT_AUTOBIN_NUM_EQWIDTH computes the bin boundaries for the cust_year_of_birth column in sh.customers and inserts the transformations in a transformation definition table. The STACK_BIN_NUM Procedure creates a transformation list from the contents of the definition table. The CREATE_MODEL Procedure embeds the transformation list in a new model called nb_model.

The transformation and reverse transformation expressions embedded in nb_model are returned by the GET_MODEL_TRANSFORMATIONS Function.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_year_of_birth, cust_postal_code
       FROM sh.customers;

DESCRIBE mining_data
 Name                          Null?    Type
 ----------------------------- -------- ----------------------------
 CUST_ID                       NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH            NOT NULL NUMBER(4)
 CUST_POSTAL_CODE              NOT NULL VARCHAR2(10)

BEGIN
  dbms_data_mining_transform.CREATE_BIN_NUM(
     bin_table_name   => 'bin_tbl');
  dbms_data_mining_transform.INSERT_AUTOBIN_NUM_EQWIDTH (
     bin_table_name   => 'bin_tbl',
     data_table_name  => 'mining_data',
     bin_num          => 3,
     max_bin_num      => 5,
     exclude_list     => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
END;
/

set numwidth 4
column val off
SELECT col, val, bin FROM bin_tbl
      ORDER BY val ASC;

COL                        VAL BIN
------------------------- ---- -----
CUST_YEAR_OF_BIRTH        1913
CUST_YEAR_OF_BIRTH        1928 1
CUST_YEAR_OF_BIRTH        1944 2
CUST_YEAR_OF_BIRTH        1959 3
CUST_YEAR_OF_BIRTH        1975 4
CUST_YEAR_OF_BIRTH        1990 5

DECLARE
     year_birth_xform   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
     dbms_data_mining_transform.STACK_BIN_NUM (
         bin_table_name           => 'bin_tbl',
         xform_list               =>  year_birth_xform);
     dbms_data_mining.CREATE_MODEL(
         model_name              => 'nb_model',
         mining_function         => dbms_data_mining.classification,
         data_table_name         => 'mining_data',
         case_id_column_name     => 'cust_id',
         target_column_name      => 'cust_postal_code',
         settings_table_name     => null,
         data_schema_name        => null,
         settings_schema_name    => null,
         xform_list              => year_birth_xform);
END;
/
```

```
SELECT attribute_name
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

ATTRIBUTE_NAME
------------------------
CUST_YEAR_OF_BIRTH

SELECT expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

EXPRESSION
--------------------------------------------------------------------------------
CASE WHEN "CUST_YEAR_OF_BIRTH"<1913 THEN NULL WHEN "CUST_YEAR_OF_BIRTH"<=1928.4
 THEN '1' WHEN "CUST_YEAR_OF_BIRTH"<=1943.8 THEN '2' WHEN "CUST_YEAR_OF_BIRTH"
<=1959.2 THEN '3' WHEN "CUST_YEAR_OF_BIRTH"<=1974.6 THEN '4' WHEN
"CUST_YEAR_OF_BIRTH" <=1990 THEN '5' END

SELECT reverse_expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

REVERSE_EXPRESSION
--------------------------------------------------------------------------------
DECODE("CUST_YEAR_OF_BIRTH",'5','(1974.6; 1990]','1','[1913; 1928.4]','2','(1928
.4; 1943.8]','3','(1943.8; 1959.2]','4','(1959.2; 1974.6]',NULL,'( ; 1913), (199
0;  ), NULL')
```

INSERT_BIN_CAT_FREQ Procedure

This procedure performs categorical binning and inserts the transformation definitions in a transformation definition table. The procedure computes the bin boundaries based on frequency.

`INSERT_BIN_CAT_FREQ` bins all the `CHAR` and `VARCHAR2` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_CAT_FREQ (
     bin_table_name        IN VARCHAR2,
     data_table_name       IN VARCHAR2,
     bin_num               IN PLS_INTEGER DEFAULT 9,
     exclude_list          IN COLUMN_LIST DEFAULT NULL,
     default_num           IN PLS_INTEGER DEFAULT 2,
     bin_support           IN NUMBER DEFAULT NULL,
     bin_schema_name       IN VARCHAR2 DEFAULT NULL,
     data_schema_name      IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-147 INSERT_BIN_CAT_FREQ Procedure Parameters

Parameter	Description

`bin_table_name`	Name of the transformation definition table for categorical binning. You can use the CREATE_BIN_CAT Procedure to create the definition table.The following columns are required:
	`COL        VARCHAR2(30)`
	`VAL        VARCHAR2(4000)`
	`BIN        VARCHAR2(4000)`
	`CREATE_BIN_CAT` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_BIN_CAT_FREQ`.

`data_table_name`	Name of the table containing the data to be transformed
`bin_num`	The number of bins to fill using frequency-based binning The total number of bins will be *bin_num*+1. The additional bin is the default bin. Classes that are not assigned to a frequency-based bin will be assigned to the default bin. The default binning order is from highest to lowest: the most frequently occurring class is assigned to the first bin, the second most frequently occurring class is assigned to the second bin, and so on.You can reverse the binning order by specifying a negative number for *bin_num*. The negative sign causes the binning order to be from lowest to highest. If the total number of distinct values (classes) in the column is less than *bin_num*, then a separate bin will be created for each value and the default bin will be empty. If you specify `NULL` or `0` for *bin_num*, no binning is performed. The default value of *bin_num* is 9.

`exclude_list`	List of categorical columns to be excluded from the binning process. If you do not specify *exclude_list*, all categorical columns in the data source are binned.
	The format of *exclude_list* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`                                        ...'coln')`

`default_num`	The number of class occurrences (rows of the same class) required for assignment to the default bin By default, *default_num* is the minimum number of occurrences required for assignment to the default bin. For example, if *default_num* is 3 and a given class occurs only once, it will not be assigned to the default bin. You can change the occurrence requirement from minimum to maximum by specifying a negative number for *default_num*. For example, if *default_num* is -3 and a given class occurs only once, it *will* be assigned to the default bin, but a class that occurs four or more times will not be included. If you specify `NULL` or `0` for *default_bin*, there are no requirements for assignment to the default bin. The default value of *default_num* is 2.

Table 6-147 (Cont.) INSERT_BIN_CAT_FREQ Procedure Parameters

Parameter	Description
bin_support	The number of class occurrences (rows of the same class) required for assignment to a frequency-based bin. *bin_support* is expressed as a fraction of the total number of rows. By default, *bin_support* is the minimum percentage required for assignment to a frequency-based bin. For example, if there are twenty rows of data and you specify.2 for *bin_support*, then there must be four or more occurrences of a class (.2*20) in order for it to be assigned to a frequency-based bin. You can change *bin_support* from a minimum percentage to a maximum percentage by specifying a negative number for *bin_support*. For example, if there are twenty rows of data and you specify -.2 for *bin_support*, then there must be four or less occurrences of a class in order for it to be assigned to a frequency-based bin. Classes that occur less than a positive *bin_support* or more than a negative *bin_support* will be assigned to the default bin. If you specify NULL or 0 for *bin_support*, then there is no support requirement for frequency-based binning. The default value of *bin_support* is NULL.
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about categorical data.
2. If values occur with the same frequency, INSERT_BIN_CAT_FREQ assigns them in descending order when binning is from most to least frequent, or in ascending order when binning is from least to most frequent.

Examples

1. In this example, INSERT_BIN_CAT_FREQ computes the bin boundaries for the cust_postal_code and cust_city columns in sh.customers and inserts the transformations in a transformation definition table. The STACK_BIN_CAT Procedure creates a transformation list from the contents of the definition table, and the CREATE_MODEL Procedure embeds the transformation list in a new model called nb_model.

   The transformation and reverse transformation expressions embedded in nb_model are returned by the GET_MODEL_TRANSFORMATIONS Function.

```
CREATE OR REPLACE VIEW mining_data AS
          SELECT cust_id, cust_year_of_birth, cust_postal_code, cust_city
          FROM sh.customers;

DESCRIBE mining_data
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_POSTAL_CODE                          NOT NULL VARCHAR2(10)
 CUST_CITY                                 NOT NULL VARCHAR2(30)
```

```
BEGIN
    dbms_data_mining_transform.CREATE_BIN_CAT(
       bin_table_name   => 'bin_tbl_1');
    dbms_data_mining_transform.INSERT_BIN_CAT_FREQ (
       bin_table_name   => 'bin_tbl_1',
       data_table_name  => 'mining_data',
       bin_num          => 4);
END;
/

column col format a18
column val format a15
column bin format a10
SELECT col, val, bin
       FROM bin_tbl_1
       ORDER BY col ASC, bin ASC;

COL                VAL             BIN
------------------ --------------- ----------
CUST_CITY          Los Angeles     1
CUST_CITY          Greenwich       2
CUST_CITY          Killarney       3
CUST_CITY          Montara         4
CUST_CITY                          5
CUST_POSTAL_CODE   38082           1
CUST_POSTAL_CODE   63736           2
CUST_POSTAL_CODE   55787           3
CUST_POSTAL_CODE   78558           4
CUST_POSTAL_CODE                   5

DECLARE
      city_xform   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
      dbms_data_mining_transform.STACK_BIN_CAT (
          bin_table_name           => 'bin_tbl_1',
          xform_list               =>  city_xform);
      dbms_data_mining.CREATE_MODEL(
           model_name              => 'nb_model',
           mining_function         => dbms_data_mining.classification,
           data_table_name         => 'mining_data',
           case_id_column_name     => 'cust_id',
           target_column_name      => 'cust_city',
           settings_table_name     => null,
           data_schema_name        => null,
           settings_schema_name    => null,
           xform_list              => city_xform);
END;
/

SELECT attribute_name
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

ATTRIBUTE_NAME
-----------------------------------------------------------------------------
CUST_CITY
CUST_POSTAL_CODE

SELECT expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

EXPRESSION
```

```
-----------------------------------------------------------------------------
DECODE("CUST_CITY",'Greenwich','2','Killarney','3','Los Angeles','1',
'Montara','4',NULL,NULL,'5')
DECODE("CUST_POSTAL_CODE",'38082','1','55787','3','63736','2','78558','4',NULL,NULL,'5')

SELECT reverse_expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('nb_model'));

REVERSE_EXPRESSION
-----------------------------------------------------------------------------
DECODE("CUST_CITY",'2','''Greenwich''','3','''Killarney''','1',
'''Los Angeles''','4','''Montara''',NULL,'NULL','5','DEFAULT')
DECODE("CUST_POSTAL_CODE",'1','''38082''','3','''55787''','2','''63736''',
'4','''78558''',NULL,'NULL','5','DEFAULT')
```

2. The binning order in example 1 is from most frequent to least frequent. The following example shows reverse order binning (least frequent to most frequent). The binning order is reversed by setting `bin_num` to -4 instead of 4.

```
BEGIN
    dbms_data_mining_transform.CREATE_BIN_CAT(
        bin_table_name   => 'bin_tbl_reverse');
    dbms_data_mining_transform.INSERT_BIN_CAT_FREQ (
        bin_table_name   => 'bin_tbl_reverse',
        data_table_name  => 'mining_data',
        bin_num          => -4);
 END;
 /

column col format a20
SELECT col, val, bin
       FROM bin_tbl_reverse
       ORDER BY col ASC, bin ASC;

COL                  VAL             BIN
-------------------- --------------- ----------
CUST_CITY            Tokyo           1
CUST_CITY            Sliedrecht      2
CUST_CITY            Haarlem         3
CUST_CITY            Diemen          4
CUST_CITY                            5
CUST_POSTAL_CODE     49358           1
CUST_POSTAL_CODE     80563           2
CUST_POSTAL_CODE     74903           3
CUST_POSTAL_CODE     71349           4
CUST_POSTAL_CODE                     5
```

INSERT_BIN_NUM_EQWIDTH Procedure

This procedure performs numerical binning and inserts the transformation definitions in a transformation definition table. The procedure identifies the minimum and maximum values and computes the bin boundaries at equal intervals.

`INSERT_BIN_NUM_EQWIDTH` computes a specified number of bins (*n*) and assigns *(max-min)/n* values to each bin. The number of bins is the same for each column. If you want to use equi-width binning, but you want the number of bins to be calculated on a per-column basis, use the INSERT_AUTOBIN_NUM_EQWIDTH Procedure.

`INSERT_BIN_NUM_EQWIDTH` bins all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_NUM_EQWIDTH (
     bin_table_name         IN VARCHAR2,
     data_table_name        IN VARCHAR2,
     bin_num                IN PLS_INTEGER DEFAULT 10,
     exclude_list           IN COLUMN_LIST DEFAULT NULL,
     round_num              IN PLS_INTEGER DEFAULT 6,
     bin_schema_name        IN VARCHAR2 DEFAULT NULL,
     data_schema_name       IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-148 INSERT_BIN_NUM_EQWIDTH Procedure Parameters

Parameter	Description

`bin_table_name`	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The following columns are required:
	`COL     VARCHAR2(30)`
	`VAL     NUMBER`
	`BIN     VARCHAR2(4000)`
	`CREATE_BIN_NUM` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_BIN_NUM_EQWIDTH`.

`data_table_name`	Name of the table containing the data to be transformed
`bin_num`	Number of bins. No binning occurs if *`bin_num`* is `0` or `NULL`. The default number of bins is 10.
`exclude_list`	List of numerical columns to be excluded from the binning process. If you do not specify *`exclude_list`*, all numerical columns in the data source are binned. The format of *`exclude_list`* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`
`round_num`	Specifies how to round the number in the `VAL` column of the transformation definition table. When *`round_num`* is positive, it specifies the most significant digits to retain. When *`round_num`* is negative, it specifies the least significant digits to remove. In both cases, the result is rounded to the specified number of digits. See the Usage Notes for an example. The default value of *`round_num`* is 6.
`bin_schema_name`	Schema of *`bin_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.

2. The *round_num* parameter controls the rounding of column values in the transformation definition table, as follows:

```
For a value of 308.162:
when round_num =  1       result is 300
when round_num =  2       result is 310
when round_num =  3       result is 308
when round_num =  0       result is 308.162
when round_num = -1       result is 308.16
when round_num = -2       result is 308.2
```

3. INSERT_BIN_NUM_EQWIDTH ignores columns with all NULL values or only one unique value.

Examples

In this example, INSERT_BIN_NUM_EQWIDTH computes the bin boundaries for the affinity_card column in mining_data_build and inserts the transformations in a transformation definition table. The STACK_BIN_NUM Procedure creates a transformation list from the contents of the definition table. The CREATE_MODEL Procedure embeds the transformation list in a new model called glm_model.

The transformation and reverse transformation expressions embedded in glm_model are returned by the GET_MODEL_TRANSFORMATIONS Function.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_income_level, cust_gender, affinity_card
       FROM mining_data_build;

DESCRIBE mining_data
 Name                          Null?    Type
 ------------------------- -------- -----------------
 CUST_ID                       NOT NULL NUMBER
 CUST_INCOME_LEVEL                      VARCHAR2(30)
 CUST_GENDER                            VARCHAR2(1)
 AFFINITY_CARD                          NUMBER(10)

BEGIN
    dbms_data_mining_transform.CREATE_BIN_NUM(
        bin_table_name   => 'bin_tbl');
    dbms_data_mining_transform.INSERT_BIN_NUM_EQWIDTH (
        bin_table_name   => 'bin_tbl',
        data_table_name  => 'mining_data',
        bin_num          => 4,
        exclude_list     => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
END;
/

set numwidth 10
column val off
column col format a20
column bin format a10
SELECT col, val, bin FROM bin_tbl
    ORDER BY val ASC;

COL                            VAL  BIN
-------------------- ----------  ----------
AFFINITY_CARD                    0
AFFINITY_CARD                  .25  1
AFFINITY_CARD                   .5  2
AFFINITY_CARD                  .75  3
AFFINITY_CARD                    1  4
```

```
CREATE TABLE glmsettings(
        setting_name  VARCHAR2(30),
        setting_value VARCHAR2(30));

BEGIN
   INSERT INTO glmsettings (setting_name, setting_value) VALUES
         (dbms_data_mining.algo_name, dbms_data_mining.algo_generalized_linear_model);
   COMMIT;
END;
/

DECLARE
     xforms   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
     dbms_data_mining_transform.STACK_BIN_NUM (
         bin_table_name           => 'bin_tbl',
         xform_list               =>  xforms,
         literal_flag             =>  TRUE);
     dbms_data_mining.CREATE_MODEL(
         model_name              => 'glm_model',
         mining_function         => dbms_data_mining.regression,
         data_table_name         => 'mining_data',
         case_id_column_name     => 'cust_id',
         target_column_name      => 'affinity_card',
         settings_table_name     => 'glmsettings',
         data_schema_name        => null,
         settings_schema_name    => null,
         xform_list              => xforms);
END;
/

SELECT attribute_name
      FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('glm_model'));

ATTRIBUTE_NAME
------------------------
AFFINITY_CARD

SELECT expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('glm_model'));

EXPRESSION
--------------------------------------------------------------------------------
CASE WHEN "AFFINITY_CARD"<0 THEN NULL WHEN "AFFINITY_CARD"<=.25 THEN 1 WHEN
"AFFINITY_CARD"<=.5 THEN 2 WHEN "AFFINITY_CARD"<=.75 THEN 3 WHEN
"AFFINITY_CARD"<=1 THEN 4 END

SELECT reverse_expression
       FROM TABLE(dbms_data_mining.GET_MODEL_TRANSFORMATIONS('glm_model'));

REVERSE_EXPRESSION
--------------------------------------------------------------------------------
DECODE("AFFINITY_CARD",4,'(.75; 1]',1,'[0; .25]',2,'(.25; .5]',3,'(.5; .75]',
NULL,'( ; 0), (1;  ), NULL')
```

INSERT_BIN_NUM_QTILE Procedure

This procedure performs numerical binning and inserts the transformation definitions in a transformation definition table. The procedure calls the SQL `NTILE` function to order the data and divide it equally into the specified number of bins (quantiles).

`INSERT_BIN_NUM_QTILE` bins all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_NUM_QTILE (
    bin_table_name       IN VARCHAR2,
    data_table_name      IN VARCHAR2,
    bin_num              IN PLS_INTEGER DEFAULT 10,
    exclude_list         IN COLUMN_LIST DEFAULT NULL,
    bin_schema_name      IN VARCHAR2 DEFAULT NULL,
    data_schema_name     IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-149 INSERT_BIN_NUM_QTILE Procedure Parameters

Parameter	Description

`bin_table_name`	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The following columns are required:
	`COL        VARCHAR2(30)`
	`VAL        NUMBER`
	`BIN        VARCHAR2(4000)`
	`CREATE_BIN_NUM` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_BIN_NUM_QTILE`.

`data_table_name`	Name of the table containing the data to be transformed
`bin_num`	Number of bins. No binning occurs if *`bin_num`* is `0` or `NULL`. The default number of bins is 10.

`exclude_list`	List of numerical columns to be excluded from the binning process. If you do not specify *`exclude_list`*, all numerical columns in the data source are binned.
	The format of *`exclude_list`* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`...'coln')`

`bin_schema_name`	Schema of *`bin_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.

2. After dividing the data into quantiles, the NTILE function distributes any remainder values one for each quantile, starting with the first. See *Oracle Database SQL Language Reference* for details.
3. Columns with all NULL values are ignored by INSERT_BIN_NUM_QTILE.

Examples

In this example, INSERT_BIN_NUM_QTILE computes the bin boundaries for the cust_year_of_birth and cust_credit_limit columns in sh.customers and inserts the transformations in a transformation definition table. The STACK_BIN_NUM Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in STACK_VIEW. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
        SELECT cust_id, cust_year_of_birth, cust_credit_limit, cust_city
        FROM sh.customers;

DESCRIBE mining_data
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_CREDIT_LIMIT                         NUMBER
 CUST_CITY                                 NOT NULL VARCHAR2(30)

BEGIN
   dbms_data_mining_transform.CREATE_BIN_NUM(
        bin_table_name   => 'bin_tbl');
   dbms_data_mining_transform.INSERT_BIN_NUM_QTILE (
        bin_table_name   => 'bin_tbl',
        data_table_name  => 'mining_data',
        bin_num          => 3,
        exclude_list     => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
END;
/

set numwidth 8
column val off
column col format a20
column bin format a10
SELECT col, val, bin
      FROM bin_tbl
      ORDER BY col ASC, val ASC;

COL                       VAL BIN
-------------------- -------- ----------
CUST_CREDIT_LIMIT        1500
CUST_CREDIT_LIMIT        3000 1
CUST_CREDIT_LIMIT        9000 2
CUST_CREDIT_LIMIT       15000 3
CUST_YEAR_OF_BIRTH       1913
CUST_YEAR_OF_BIRTH       1949 1
CUST_YEAR_OF_BIRTH       1965 2
CUST_YEAR_OF_BIRTH       1990 3

DECLARE
   xforms   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
   dbms_data_mining_transform.STACK_BIN_NUM (
        bin_table_name        => 'bin_tbl',
        xform_list            =>  xforms);
   dbms_data_mining_transform.XFORM_STACK (
        xform_list            =>  xforms,
        data_table_name       => 'mining_data',
        xform_view_name       => 'stack_view');
END;
/

set long 3000
SELECT text FROM user_views WHERE view_name in 'STACK_VIEW';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID",CASE WHEN "CUST_YEAR_OF_BIRTH"<1913 THEN NULL WHEN "CUST_YEAR_O
F_BIRTH"<=1949 THEN '1' WHEN "CUST_YEAR_OF_BIRTH"<=1965 THEN '2' WHEN "CUST_YEAR
_OF_BIRTH"<=1990 THEN '3' END "CUST_YEAR_OF_BIRTH",CASE WHEN "CUST_CREDIT_LIMIT"
<1500 THEN NULL WHEN "CUST_CREDIT_LIMIT"<=3000 THEN '1' WHEN "CUST_CREDIT_LIMIT"
<=9000 THEN '2' WHEN "CUST_CREDIT_LIMIT"<=15000 THEN '3' END "CUST_CREDIT_LIMIT"
,"CUST_CITY" FROM mining_data
```

INSERT_BIN_SUPER Procedure

This procedure performs numerical and categorical binning and inserts the transformation definitions in transformation definition tables. The procedure computes bin boundaries based on intrinsic relationships between predictors and a target.

INSERT_BIN_SUPER uses an intelligent binning technique known as **supervised binning**. It builds a single-predictor decision tree and derives the bin boundaries from splits within the tree.

INSERT_BIN_SUPER bins all the VARCHAR2, CHAR, NUMBER, and FLOAT columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_SUPER (
    num_table_name        IN VARCHAR2,
    cat_table_name        IN VARCHAR2,
    data_table_name       IN VARCHAR2,
    target_column_name    IN VARCHAR2,
    max_bin_num           IN PLS_INTEGER   DEFAULT 1000,
    exclude_list          IN COLUMN_LIST   DEFAULT NULL,
    num_schema_name       IN VARCHAR2      DEFAULT NULL,
    cat_schema_name       IN VARCHAR2      DEFAULT NULL,
    data_schema_name      IN VARCHAR2      DEFAULT NULL,
    rem_table_name        IN VARCHAR2      DEFAULT NULL,
    rem_schema_name       IN VARCHAR2      DEFAULT NULL);
```

Parameters

Table 6-150 INSERT_BIN_SUPER Procedure Parameters

Parameter	Description

num_table_name	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The following columns are required:
	`COL         VARCHAR2(30)`
	`VAL         VNUMBER`
	`BIN         VARCHAR2(4000)`
	CREATE_BIN_NUM creates an additional column, ATT, which may be used for specifying nested attributes. This column is not used by INSERT_BIN_SUPER.

cat_table_name	Name of the transformation definition table for categorical binning. You can use the CREATE_BIN_CAT Procedure to create the definition table. The following columns are required:
	`COL         VARCHAR2(30)`
	`VAL         VARCHAR2(4000)`
	`BIN         VARCHAR2(4000)`
	CREATE_BIN_CAT creates an additional column, ATT, which is used for specifying nested attributes. This column is not used by INSERT_BIN_SUPER.

data_table_name	Name of the table containing the data to be transformed
target_column_name	Name of a column to be used as the target for the decision tree models
max_bin_num	The maximum number of bins. The default is 1000.

exclude_list	List of columns to be excluded from the binning process. If you do not specify *exclude_list*, all numerical and categorical columns in the data source are binned.
	The format of *exclude_list* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`                                        ...'coln')`

num_schema_name	Schema of *num_table_name*. If no schema is specified, the current schema is used.
cat_schema_name	Schema of *cat_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.
rem_table_name	Name of a column removal definition table. The table must have the columns described in "CREATE_COL_REM Procedure". You can use CREATE_COL_REM to create the table. See Usage Notes.
rem_schema_name	Schema of *rem_table_name*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical and categorical data.

2. Columns that have no significant splits are not binned. You can remove the unbinned columns from the mining data by specifying a column removal definition table. If you do not specify a column removal definition table, the unbinned columns remain in the mining data.
3. See *Oracle Machine Learning for SQL Concepts* to learn more about decision trees in Oracle Machine Learning for SQL

Examples

In this example, `INSERT_BIN_SUPER` computes the bin boundaries for predictors of `cust_credit_limit` and inserts the transformations in transformation definition tables. One predictor is numerical, the other is categorical. (`INSERT_BIN_SUPER` determines that the `cust_postal_code` column is not a significant predictor.) `STACK` procedures create transformation lists from the contents of the definition tables.

The SQL expressions that compute the transformations are shown in the views `MINING_DATA_STACK_NUM` and `MINING_DATA_STACK_CAT`. The views are for display purposes only; they cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
     SELECT cust_id, cust_year_of_birth, cust_marital_status,
            cust_postal_code, cust_credit_limit
     FROM sh.customers;

DESCRIBE mining_data
 Name                             Null?    Type
 -------------------------------- -------- ------------------------------------
 CUST_ID                          NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH               NOT NULL NUMBER(4)
 CUST_MARITAL_STATUS                       VARCHAR2(20)
 CUST_POSTAL_CODE                 NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                         NUMBER

BEGIN
    dbms_data_mining_transform.CREATE_BIN_NUM(
         bin_table_name      => 'bin_num_tbl');
    dbms_data_mining_transform.CREATE_BIN_CAT(
         bin_table_name      => 'bin_cat_tbl');
    dbms_data_mining_transform.CREATE_COL_REM(
         rem_table_name      => 'rem_tbl');
END;
/

BEGIN
   COMMIT;
   dbms_data_mining_transform.INSERT_BIN_SUPER (
      num_table_name      => 'bin_num_tbl',
      cat_table_name      => 'bin_cat_tbl',
      data_table_name     => 'mining_data',
      target_column_name  => 'cust_credit_limit',
      max_bin_num         =>  4,
      exclude_list        =>  dbms_data_mining_transform.COLUMN_LIST('cust_id'),
      num_schema_name     => 'oml_user',
      cat_schema_name     => 'oml_user',
      data_schema_name    => 'oml_user',
      rem_table_name      => 'rem_tbl',
      rem_schema_name     => 'oml_user');
   COMMIT;
END;
/

set numwidth 8
```

```
column val off
SELECT col, val, bin FROM bin_num_tbl
      ORDER BY bin ASC;

COL                       VAL BIN
-------------------- -------- ----------
CUST_YEAR_OF_BIRTH     1923.5 1
CUST_YEAR_OF_BIRTH     1923.5 1
CUST_YEAR_OF_BIRTH     1945.5 2
CUST_YEAR_OF_BIRTH     1980.5 3
CUST_YEAR_OF_BIRTH            4

column val on
column val format a20
SELECT col, val, bin FROM bin_cat_tbl
      ORDER BY bin ASC;

COL                  VAL                  BIN
-------------------- -------------------- ----------
CUST_MARITAL_STATUS  married              1
CUST_MARITAL_STATUS  single               2
CUST_MARITAL_STATUS  Mar-AF               3
CUST_MARITAL_STATUS  Mabsent              3
CUST_MARITAL_STATUS  Divorc.              3
CUST_MARITAL_STATUS  Married              3
CUST_MARITAL_STATUS  Widowed              3
CUST_MARITAL_STATUS  NeverM               3
CUST_MARITAL_STATUS  Separ.               3
CUST_MARITAL_STATUS  divorced             4
CUST_MARITAL_STATUS  widow                4

SELECT col from rem_tbl;

COL
--------------------
CUST_POSTAL_CODE

DECLARE
    xforms_num      dbms_data_mining_transform.TRANSFORM_LIST;
    xforms_cat      dbms_data_mining_transform.TRANSFORM_LIST;
    BEGIN
       dbms_data_mining_transform.STACK_BIN_NUM (
            bin_table_name    => 'bin_num_tbl',
            xform_list        => xforms_num);
       dbms_data_mining_transform.XFORM_STACK (
            xform_list         => xforms_num,
            data_table_name    => 'mining_data',
            xform_view_name    => 'mining_data_stack_num');
       dbms_data_mining_transform.STACK_BIN_CAT (
             bin_table_name    => 'bin_cat_tbl',
             xform_list        => xforms_cat);
       dbms_data_mining_transform.XFORM_STACK (
             xform_list         => xforms_cat,
             data_table_name    => 'mining_data',
             xform_view_name    => 'mining_data_stack_cat');
   END;
 /

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK_NUM';

TEXT
```

```
--------------------------------------------------------------------------------
SELECT "CUST_ID",CASE WHEN "CUST_YEAR_OF_BIRTH"<1923.5 THEN '1' WHEN "CUST_YEAR_
OF_BIRTH"<=1923.5 THEN '1' WHEN "CUST_YEAR_OF_BIRTH"<=1945.5 THEN '2' WHEN "CUST
_YEAR_OF_BIRTH"<=1980.5 THEN '3' WHEN "CUST_YEAR_OF_BIRTH" IS NOT NULL THEN '4'
END "CUST_YEAR_OF_BIRTH","CUST_MARITAL_STATUS","CUST_POSTAL_CODE","CUST_CREDIT_L
IMIT" FROM mining_data


SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK_CAT';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_YEAR_OF_BIRTH",DECODE("CUST_MARITAL_STATUS",'Divorc.','3'
,'Mabsent','3','Mar-AF','3','Married','3','NeverM','3','Separ.','3','Widowed','3
','divorced','4','married','1','single','2','widow','4') "CUST_MARITAL_STATUS","
CUST_POSTAL_CODE","CUST_CREDIT_LIMIT" FROM mining_data
```

INSERT_CLIP_TRIM_TAIL Procedure

This procedure replaces numeric outliers with nulls and inserts the transformation definitions in a transformation definition table.

`INSERT_CLIP_TRIM_TAIL` computes the boundaries of the data based on a specified percentage. It removes the values that fall outside the boundaries (tail values) from the data. If you wish to replace the tail values instead of removing them, use the INSERT_CLIP_WINSOR_TAIL Procedure.

`INSERT_CLIP_TRIM_TAIL` clips all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_CLIP_TRIM_TAIL (
    clip_table_name     IN VARCHAR2,
    data_table_name     IN VARCHAR2,
    tail_frac           IN NUMBER DEFAULT 0.025,
    exclude_list        IN COLUMN_LIST DEFAULT NULL,
    clip_schema_name    IN VARCHAR2 DEFAULT NULL,
    data_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-151 INSERT_CLIP_TRIM_TAIL Procedure Parameters

Parameter	Description

`clip_table_name`	Name of the transformation definition table for numerical clipping. You can use the CREATE_CLIP Procedure to create the definition table. The following columns are required:
	`COL          VARCHAR2(30)`
	`LCUT         NUMBER`
	`LVAL         NUMBER`
	`RCUT         NUMBER`
	`RVAL         NUMBER`
	`CREATE_CLIP` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_CLIP_TRIM_TAIL`.

`data_table_name`	Name of the table containing the data to be transformed

Table 6-151 (Cont.) INSERT_CLIP_TRIM_TAIL Procedure Parameters

Parameter	Description
`tail_frac`	The percentage of non-null values to be designated as outliers at each end of the data. For example, if *`tail_frac`* is .01, then 1% of the data at the low end and 1% of the data at the high end will be treated as outliers. If *`tail_frac`* is greater than or equal to .5, no clipping occurs. The default value of *`tail_frac`* is 0.025.

`exclude_list`	List of numerical columns to be excluded from the clipping process. If you do not specify *`exclude_list`*, all numerical columns in the data are clipped.
	The format of *`exclude_list`* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`...'coln')`

`clip_schema_name`	Schema of *`clip_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. The `DBMS_DATA_MINING_TRANSFORM` package provides two clipping procedures: `INSERT_CLIP_TRIM_TAIL` and `INSERT_CLIP_WINSOR_TAIL`. Both procedures compute the boundaries as follows:
   - Count the number of non-null values, ***n***, and sort them in ascending order
   - Calculate the number of outliers, ***t***, as ***n*tail_frac***
   - Define the lower boundary ***lcut*** as the value at position ***1+floor(t)***
   - Define the upper boundary ***rcut*** as the value at position ***n-floor(t)***

     (The SQL `FLOOR` function returns the largest integer less than or equal to ***t***.)
   - All values that are <= ***lcut*** or => ***rcut*** are designated as outliers.

   `INSERT_CLIP_TRIM_TAIL` replaces the outliers with nulls, effectively removing them from the data.

   `INSERT_CLIP_WINSOR_TAIL` assigns ***lcut*** to the low outliers and ***rcut*** to the high outliers.

Examples

In this example, `INSERT_CLIP_TRIM_TAIL` trims 10% of the data in two columns (5% from the high end and 5% from the low end) and inserts the transformations in a transformation definition table. The STACK_CLIP Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the trimming is shown in the view `MINING_DATA_STACK`. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_year_of_birth, cust_credit_limit, cust_city
       FROM sh.customers;
```

```
DESCRIBE mining_data
 Name                             Null?    Type
 -------------------------------- -------- --------------------
 CUST_ID                          NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH               NOT NULL NUMBER(4)
 CUST_CREDIT_LIMIT                         NUMBER
 CUST_CITY                        NOT NULL VARCHAR2(30)

BEGIN
   dbms_data_mining_transform.CREATE_CLIP(
      clip_table_name    => 'clip_tbl');
   dbms_data_mining_transform.INSERT_CLIP_TRIM_TAIL(
      clip_table_name    => 'clip_tbl',
      data_table_name    => 'mining_data',
      tail_frac          => 0.05,
      exclude_list       => DBMS_DATA_MINING_TRANSFORM.COLUMN_LIST('cust_id'));
END;
/

SELECT col, lcut, lval, rcut, rval
      FROM clip_tbl
      ORDER BY col ASC;

COL                        LCUT     LVAL     RCUT     RVAL
-------------------- --------- -------- -------- --------
CUST_CREDIT_LIMIT          1500             11000
CUST_YEAR_OF_BIRTH         1934              1982

DECLARE
     xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
     dbms_data_mining_transform.STACK_CLIP (
          clip_table_name    => 'clip_tbl',
          xform_list         => xforms);
     dbms_data_mining_transform.XFORM_STACK (
          xform_list         => xforms,
          data_table_name    => 'mining_data',
          xform_view_name    => 'mining_data_stack');
 END;
 /

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID",CASE WHEN "CUST_YEAR_OF_BIRTH" < 1934 THEN NULL WHEN "CUST_YEAR
_OF_BIRTH" > 1982 THEN NULL ELSE "CUST_YEAR_OF_BIRTH" END "CUST_YEAR_OF_BIRTH",C
ASE WHEN "CUST_CREDIT_LIMIT" < 1500 THEN NULL WHEN "CUST_CREDIT_LIMIT" > 11000 T
HEN NULL ELSE "CUST_CREDIT_LIMIT" END "CUST_CREDIT_LIMIT","CUST_CITY" FROM minin
g_data
```

INSERT_CLIP_WINSOR_TAIL Procedure

This procedure replaces numeric outliers with the upper or lower boundary values. It inserts the transformation definitions in a transformation definition table.

`INSERT_CLIP_WINSOR_TAIL` computes the boundaries of the data based on a specified percentage. It replaces the values that fall outside the boundaries (tail values) with the related boundary value. If you wish to set tail values to null, use the INSERT_CLIP_TRIM_TAIL Procedure.

`INSERT_CLIP_WINSOR_TAIL` clips all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_CLIP_WINSOR_TAIL (
    clip_table_name    IN VARCHAR2,
    data_table_name    IN VARCHAR2,
    tail_frac          IN NUMBER DEFAULT 0.025,
    exclude_list       IN COLUMN_LIST DEFAULT NULL,
    clip_schema_name   IN VARCHAR2 DEFAULT NULL,
    data_schema_name   IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-152 INSERT_CLIP_WINSOR_TAIL Procedure Parameters

Parameter	Description

`clip_table_name`	Name of the transformation definition table for numerical clipping. You can use the CREATE_CLIP Procedure to create the definition table. The following columns are required:
	`COL          VARCHAR2(30)`
	`LCUT         NUMBER`
	`LVAL         NUMBER`
	`RCUT         NUMBER`
	`RVAL         NUMBER`
	`CREATE_CLIP` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_CLIP_WINSOR_TAIL`.

`data_table_name`	Name of the table containing the data to be transformed
`tail_frac`	The percentage of non-null values to be designated as outliers at each end of the data. For example, if *`tail_frac`* is .01, then 1% of the data at the low end and 1% of the data at the high end will be treated as outliers. If *`tail_frac`* is greater than or equal to .5, no clipping occurs. The default value of *`tail_frac`* is 0.025.

`exclude_list`	List of numerical columns to be excluded from the clipping process. If you do not specify *`exclude_list`*, all numerical columns in the data are clipped.
	The format of *`exclude_list`* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`                                        ...'coln')`

`clip_schema_name`	Schema of *`clip_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User’s Guide* for details about numerical data.
2. The `DBMS_DATA_MINING_TRANSFORM` package provides two clipping procedures: `INSERT_CLIP_WINSOR_TAIL` and `INSERT_CLIP_TRIM_TAIL`. Both procedures compute the boundaries as follows:

- Count the number of non-null values, ***`n`***, and sort them in ascending order
- Calculate the number of outliers, ***`t`***, as ***`n*tail_frac`***
- Define the lower boundary ***`lcut`*** as the value at position ***`1+floor(t)`***
- Define the upper boundary ***rcut*** as the value at position ***`n-floor(t)`***

  (The SQL `FLOOR` function returns the largest integer less than or equal to ***`t`***.)
- All values that are <= ***`lcut`*** or => ***`rcut`*** are designated as outliers.

`INSERT_CLIP_WINSOR_TAIL` assigns ***`lcut`*** to the low outliers and ***`rcut`*** to the high outliers.

`INSERT_CLIP_TRIM_TAIL` replaces the outliers with nulls, effectively removing them from the data.

Examples

In this example, `INSERT_CLIP_WINSOR_TAIL` winsorizes 10% of the data in two columns (5% from the high end, and 5% from the low end) and inserts the transformations in a transformation definition table. The STACK_CLIP Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view `MINING_DATA_STACK`. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
        SELECT cust_id, cust_year_of_birth, cust_credit_limit, cust_city
        FROM sh.customers;

describe mining_data
 Name                                      Null?    Type
 ----------------------------------------- -------- -------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_CREDIT_LIMIT                                  NUMBER
 CUST_CITY                                 NOT NULL VARCHAR2(30)

BEGIN
  dbms_data_mining_transform.CREATE_CLIP(
      clip_table_name   => 'clip_tbl');
  dbms_data_mining_transform.INSERT_CLIP_WINSOR_TAIL(
      clip_table_name   => 'clip_tbl',
      data_table_name   => 'mining_data',
      tail_frac         => 0.05,
      exclude_list      => DBMS_DATA_MINING_TRANSFORM.COLUMN_LIST('cust_id'));
END;
/

SELECT col, lcut, lval, rcut, rval FROM clip_tbl
   ORDER BY col ASC;
COL                                LCUT     LVAL     RCUT     RVAL
------------------------------ -------- -------- -------- --------
CUST_CREDIT_LIMIT                  1500     1500    11000    11000
CUST_YEAR_OF_BIRTH                 1934     1934     1982     1982

DECLARE
   xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
   dbms_data_mining_transform.STACK_CLIP (
   clip_table_name    => 'clip_tbl',
   xform_list        => xforms);
```

```
dbms_data_mining_transform.XFORM_STACK (
   xform_list          => xforms,
   data_table_name     => 'mining_data',
   xform_view_name     => 'mining_data_stack');
END;
/

set long 3000
SQL> SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID",CASE WHEN "CUST_YEAR_OF_BIRTH" < 1934 THEN 1934 WHEN "CUST_YEAR
_OF_BIRTH" > 1982 THEN 1982 ELSE "CUST_YEAR_OF_BIRTH" END "CUST_YEAR_OF_BIRTH",C
ASE WHEN "CUST_CREDIT_LIMIT" < 1500 THEN 1500 WHEN "CUST_CREDIT_LIMIT" > 11000 T
HEN 11000 ELSE "CUST_CREDIT_LIMIT" END "CUST_CREDIT_LIMIT","CUST_CITY" FROM mini
ng_data
```

INSERT_MISS_CAT_MODE Procedure

This procedure replaces missing categorical values with the value that occurs most frequently in the column (the mode). It inserts the transformation definitions in a transformation definition table.

`INSERT_MISS_CAT_MODE` replaces missing values in all `VARCHAR2` and `CHAR` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_MISS_CAT_MODE (
    miss_table_name     IN VARCHAR2,
    data_table_name     IN VARCHAR2,
    exclude_list        IN COLUMN_LIST DEFAULT NULL,
    miss_schema_name    IN VARCHAR2 DEFAULT NULL,
    data_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-153 INSERT_MISS_CAT_MODE Procedure Parameters

Parameter	Description

`miss_table_name`	Name of the transformation definition table for categorical missing value treatment. You can use the CREATE_MISS_CAT Procedure to create the definition table. The following columns are required:
	`COL          VARCHAR2(30)`
	`VAL          VARCHAR2(4000)`
	`CREATE_MISS_CAT` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_MISS_CAT_MODE`.

`data_table_name`	Name of the table containing the data to be transformed
`exclude_list`	List of categorical columns to be excluded from missing value treatment. If you do not specify *`exclude_list`*, all categorical columns are transformed. The format of *`exclude_list`* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`

Table 6-153 (Cont.) INSERT_MISS_CAT_MODE Procedure Parameters

Parameter	Description
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about categorical data.
2. If you wish to replace categorical missing values with a value other than the mode, you can edit the transformation definition table.

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for information about default missing value treatment in Oracle Machine Learning for SQL

Example

In this example, `INSERT_MISS_CAT_MODE` computes missing value treatment for `cust_city` and inserts the transformation in a transformation definition table. The STACK_MISS_CAT Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view `MINING_DATA_STACK`. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
         SELECT cust_id, cust_year_of_birth, cust_city
         FROM sh.customers;

describe mining_data
 Name                              Null?    Type
 -------------------------------- -------- ----------------
 CUST_ID                           NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                NOT NULL NUMBER(4)
 CUST_CITY                         NOT NULL VARCHAR2(30)

BEGIN
  dbms_data_mining_transform.create_miss_cat(
      miss_table_name   => 'missc_tbl');
  dbms_data_mining_transform.insert_miss_cat_mode(
      miss_table_name   => 'missc_tbl',
      data_table_name   => 'mining_data');
END;
/

SELECT stats_mode(cust_city) FROM mining_data;

STATS_MODE(CUST_CITY)
------------------------------
Los Angeles
```

```
SELECT col, val
    from missc_tbl;

COL                            VAL
------------------------------ ------------------------------
CUST_CITY                      Los Angeles

DECLARE
    xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
    dbms_data_mining_transform.STACK_MISS_CAT (
        miss_table_name    => 'missc_tbl',
        xform_list        => xforms);
    dbms_data_mining_transform.XFORM_STACK (
         xform_list         => xforms,
         data_table_name    => 'mining_data',
         xform_view_name    => 'mining_data_stack');
END;
/

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_YEAR_OF_BIRTH",NVL("CUST_CITY",'Los Angeles') "CUST_CITY"
 FROM mining_data
```

INSERT_MISS_NUM_MEAN Procedure

This procedure replaces missing numerical values with the average (the mean) and inserts the transformation definitions in a transformation definition table.

INSERT_MISS_NUM_MEAN replaces missing values in all NUMBER and FLOAT columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_MISS_NUM_MEAN (
    miss_table_name    IN VARCHAR2,
    data_table_name    IN VARCHAR2,
    exclude_list       IN COLUMN_LIST DEFAULT NULL,
    round_num          IN PLS_INTEGER DEFAULT 6,
    miss_schema_name   IN VARCHAR2 DEFAULT NULL,
    data_schema_name   IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-154 INSERT_MISS_NUM_MEAN Procedure Parameters

Parameter	Description

`miss_table_name`	Name of the transformation definition table for numerical missing value treatment. You can use the CREATE_MISS_NUM Procedure to create the definition table.
	The following columns are required by `INSERT_MISS_NUM_MEAN`:
	`COL          VARCHAR2(30)`
	`VAL          NUMBER`
	`CREATE_MISS_NUM` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_MISS_NUM_MEAN`.

`data_table_name`	Name of the table containing the data to be transformed

`exclude_list`	List of numerical columns to be excluded from missing value treatment. If you do not specify *`exclude_list`*, all numerical columns are transformed.
	The format of *`exclude_list`* is:
	`dbms_data_mining_transform.COLUMN_LIST('col1','col2',`
	`                                           ...'coln')`

`round_num`	The number of significant digits to use for the mean. The default number is 6.
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.
2. If you wish to replace numerical missing values with a value other than the mean, you can edit the transformation definition table.

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* for information about default missing value treatment in Oracle Machine Learning for SQL

Example

In this example, `INSERT_MISS_NUM_MEAN` computes missing value treatment for `cust_year_of_birth` and inserts the transformation in a transformation definition table. The STACK_MISS_NUM Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view `MINING_DATA_STACK`. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
    SELECT cust_id, cust_year_of_birth, cust_city
    FROM sh.customers;

DESCRIBE mining_data
 Name                                       Null?    Type
 ------------------------------------------ -------- -------------------
 CUST_ID                                    NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                         NOT NULL NUMBER(4)
 CUST_CITY                                  NOT NULL VARCHAR2(30)

BEGIN
   dbms_data_mining_transform.create_miss_num(
       miss_table_name   => 'missn_tbl');
   dbms_data_mining_transform.insert_miss_num_mean(
       miss_table_name   => 'missn_tbl',
       data_table_name   => 'mining_data',
       exclude_list      => DBMS_DATA_MINING_TRANSFORM.COLUMN_LIST('cust_id'));
END;
/

set numwidth 4
column val off
SELECT col, val
  FROM missn_tbl;

COL                   VAL
-------------------- ----
CUST_YEAR_OF_BIRTH   1957

SELECT avg(cust_year_of_birth) FROM mining_data;

AVG(CUST_YEAR_OF_BIRTH)
-----------------------
                   1957

DECLARE
    xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
    dbms_data_mining_transform.STACK_MISS_NUM (
         miss_table_name    => 'missn_tbl',
         xform_list        => xforms);
    dbms_data_mining_transform.XFORM_STACK (
         xform_list         => xforms,
         data_table_name    => 'mining_data',
         xform_view_name    => 'mining_data_stack');
END;
/

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID",NVL("CUST_YEAR_OF_BIRTH",1957.4) "CUST_YEAR_OF_BIRTH","CUST_CIT
Y" FROM mining_data
```

INSERT_NORM_LIN_MINMAX Procedure

This procedure performs linear normalization and inserts the transformation definitions in a transformation definition table.

INSERT_NORM_LIN_MINMAX computes the minimum and maximum values from the data and sets the value of *shift* and *scale* as follows:

```
shift = min
scale = max - min
```

Normalization is computed as:

```
x_new = (x_old - shift)/scale
```

INSERT_NORM_LIN_MINMAX rounds the value of *scale* to a specified number of significant digits before storing it in the transformation definition table.

INSERT_NORM_LIN_MINMAX normalizes all the NUMBER and FLOAT columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_MINMAX (
     norm_table_name      IN VARCHAR2,
     data_table_name      IN VARCHAR2,
     exclude_list         IN COLUMN_LIST DEFAULT NULL,
     round_num            IN PLS_INTEGER DEFAULT 6,
     norm_schema_name     IN VARCHAR2 DEFAULT NULL,
     data_schema_name     IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-155 *INSERT_NORM_LIN_MINMAX Procedure Parameters*

Parameter	Description
norm_table_name	Name of the transformation definition table for linear normalization. You can use the CREATE_NORM_LIN Procedure to create the definition table. The following columns are required: COL VARCHAR2(30) SHIFT NUMBER SCALE NUMBER CREATE_NORM_LIN creates an additional column, ATT, which may be used for specifying nested attributes. This column is not used by INSERT_NORM_LIN_MINMAX.
data_table_name	Name of the table containing the data to be transformed
exclude_list	List of numerical columns to be excluded from normalization. If you do not specify *exclude_list*, all numerical columns are transformed. The format of *exclude_list* is: dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')
round_num	The number of significant digits to use for the minimum and maximum. The default number is 6.

Table 6-155 (Cont.) *INSERT_NORM_LIN_MINMAX Procedure Parameters*

Parameter	Description
norm_schema_name	Schema of *norm_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.

Examples

In this example, INSERT_NORM_LIN_MINMAX normalizes the cust_year_of_birth column and inserts the transformation in a transformation definition table. The STACK_NORM_LIN Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view MINING_DATA_STACK. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
         SELECT cust_id, cust_gender, cust_year_of_birth
         FROM sh.customers;

describe mining_data
 Name                                 Null?    Type
 ------------------------------------ -------- ----------------
 CUST_ID                              NOT NULL NUMBER
 CUST_GENDER                          NOT NULL CHAR(1)
 CUST_YEAR_OF_BIRTH                   NOT NULL NUMBER(4)

BEGIN
       dbms_data_mining_transform.CREATE_NORM_LIN(
         norm_table_name  => 'norm_tbl');
       dbms_data_mining_transform.INSERT_NORM_LIN_MINMAX(
         norm_table_name  => 'norm_tbl',
         data_table_name  => 'mining_data',
         exclude_list     => dbms_data_mining_transform.COLUMN_LIST( 'cust_id'),
         round_num        => 3);
END;
/

SELECT col, shift, scale FROM norm_tbl;

COL                                 SHIFT      SCALE
------------------------------ ---------- ----------
CUST_YEAR_OF_BIRTH                   1910         77

DECLARE
     xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
     dbms_data_mining_transform.STACK_NORM_LIN (
          norm_table_name    => 'norm_tbl',
          xform_list        => xforms);
     dbms_data_mining_transform.XFORM_STACK (
          xform_list         => xforms,
          data_table_name    => 'mining_data',
          xform_view_name    => 'mining_data_stack');
END;
/

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_GENDER",("CUST_YEAR_OF_BIRTH"-1910)/77 "CUST_YEAR_OF_BIRT
H" FROM mining_data
```

INSERT_NORM_LIN_SCALE Procedure

This procedure performs linear normalization and inserts the transformation definitions in a transformation definition table.

`INSERT_NORM_LIN_SCALE` computes the minimum and maximum values from the data and sets the value of *`shift`* and *`scale`* as follows:

```
shift = 0
scale = max(abs(max), abs(min))
```

Normalization is computed as:

```
x_new = (x_old)/scale
```

`INSERT_NORM_LIN_SCALE` rounds the value of *`scale`* to a specified number of significant digits before storing it in the transformation definition table.

`INSERT_NORM_LIN_SCALE` normalizes all the `NUMBER` and `FLOAT` columns in the data source unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_SCALE (
     norm_table_name     IN VARCHAR2,
     data_table_name     IN VARCHAR2,
     exclude_list        IN COLUMN_LIST DEFAULT NULL,
     round_num           IN PLS_INTEGER DEFAULT 6,
     norm_schema_name    IN VARCHAR2 DEFAULT NULL,
     data_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-156   INSERT_NORM_LIN_SCALE Procedure Parameters

Parameter	Description

`norm_table_name`	Name of the transformation definition table for linear normalization. You can use the CREATE_NORM_LIN Procedure to create the definition table. The following columns are required:
	`COL       VARCHAR2(30)`
	`SHIFT     NUMBER`
	`SCALE     NUMBER`
	`CREATE_NORM_LIN` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_NORM_LIN_SCALE`.

Table 6-156   (Cont.) INSERT_NORM_LIN_SCALE Procedure Parameters

Parameter	Description
data_table_name	Name of the table containing the data to be transformed
exclude_list	List of numerical columns to be excluded from normalization. If you do not specify *exclude_list*, all numerical columns are transformed. The format of *exclude_list* is: dbms_data_mining_transform.COLUMN_LIST('col*1*','col*2*', ...'col*n*')
round_num	The number of significant digits to use for *scale*. The default number is 6.
norm_schema_name	Schema of *norm_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See *Oracle Machine Learning for SQL User’s Guide* for details about numerical data.

Examples

In this example, INSERT_NORM_LIN_SCALE normalizes the cust_year_of_birth column and inserts the transformation in a transformation definition table. The STACK_NORM_LIN Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view MINING_DATA_STACK. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
      SELECT cust_id, cust_gender, cust_year_of_birth
      FROM sh.customers;

DESCRIBE mining_data
 Name                               Null?    Type
 ---------------------------------- -------- ------------------
 CUST_ID                            NOT NULL NUMBER
 CUST_GENDER                        NOT NULL CHAR(1)
 CUST_YEAR_OF_BIRTH                 NOT NULL NUMBER(4)

BEGIN
   dbms_data_mining_transform.CREATE_NORM_LIN(
       norm_table_name  => 'norm_tbl');
       dbms_data_mining_transform.INSERT_NORM_LIN_SCALE(
       norm_table_name  => 'norm_tbl',
       data_table_name  => 'mining_data',
       exclude_list     => dbms_data_mining_transform.COLUMN_LIST( 'cust_id'),
       round_num        => 3);
 END;
 /

SELECT col, shift, scale FROM norm_tbl;

COL                  SHIFT SCALE
-------------------- ----- -----
```

```
CUST_YEAR_OF_BIRTH       0  1990

DECLARE
    xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
    dbms_data_mining_transform.STACK_NORM_LIN (
        norm_table_name    => 'norm_tbl',
        xform_list        => xforms);
    dbms_data_mining_transform.XFORM_STACK (
        xform_list         => xforms,
        data_table_name    => 'mining_data',
        xform_view_name    => 'mining_data_stack');
END;
/

set long 3000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_GENDER",("CUST_YEAR_OF_BIRTH"-0)/1990 "CUST_YEAR_OF_BIRTH
" FROM mining_data
```

INSERT_NORM_LIN_ZSCORE Procedure

This procedure performs linear normalization and inserts the transformation definitions in a transformation definition table.

`INSERT_NORM_LIN_ZSCORE` computes the mean and the standard deviation from the data and sets the value of *`shift`* and *`scale`* as follows:

```
shift = mean
scale = stddev
```

Normalization is computed as:

```
x_new = (x_old - shift)/scale
```

`INSERT_NORM_LIN_ZSCORE` rounds the value of *`scale`* to a specified number of significant digits before storing it in the transformation definition table.

`INSERT_NORM_LIN_ZSCORE` normalizes all the `NUMBER` and `FLOAT` columns in the data unless you specify a list of columns to ignore.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_ZSCORE (
     norm_table_name     IN VARCHAR2,
     data_table_name     IN VARCHAR2,
     exclude_list        IN COLUMN_LIST DEFAULT NULL,
     round_num           IN PLS_INTEGER DEFAULT 6,
     norm_schema_name    IN VARCHAR2 DEFAULT NULL,
     data_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-157 INSERT_NORM_LIN_ZSCORE Procedure Parameters

Parameter	Description

`norm_table_name`	Name of the transformation definition table for linear normalization. You can use the CREATE_NORM_LIN Procedure to create the definition table. The following columns are required:
	`COL      VARCHAR2(30)`
	`SHIFT    NUMBER`
	`SCALE    NUMBER`
	`CREATE_NORM_LIN` creates an additional column, `ATT`, which may be used for specifying nested attributes. This column is not used by `INSERT_NORM_LIN_ZSCORE`.

`data_table_name`	Name of the table containing the data to be transformed
`exclude_list`	List of numerical columns to be excluded from normalization. If you do not specify *`exclude_list`*, all numerical columns are transformed. The format of *`exclude_list`* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`
`round_num`	The number of significant digits to use for *`scale`*. The default number is 6.
`norm_schema_name`	Schema of *`norm_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

See *Oracle Machine Learning for SQL User's Guide* for details about numerical data.

Examples

In this example, `INSERT_NORM_LIN_ZSCORE` normalizes the `cust_year_of_birth` column and inserts the transformation in a transformation definition table. The STACK_NORM_LIN Procedure creates a transformation list from the contents of the definition table.

The SQL expression that computes the transformation is shown in the view `MINING_DATA_STACK`. The view is for display purposes only; it cannot be used to embed the transformations in a model.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_gender, cust_year_of_birth
       FROM sh.customers;

DESCRIBE mining_data
 Name                                   Null?    Type
 -------------------------------------- -------- ----------------------
 CUST_ID                                NOT NULL NUMBER
 CUST_GENDER                            NOT NULL CHAR(1)
 CUST_YEAR_OF_BIRTH                     NOT NULL NUMBER(4)

BEGIN
    dbms_data_mining_transform.CREATE_NORM_LIN(
       norm_table_name  => 'norm_tbl');
       dbms_data_mining_transform.INSERT_NORM_LIN_ZSCORE(
       norm_table_name  => 'norm_tbl',
       data_table_name  => 'mining_data',
       exclude_list     => dbms_data_mining_transform.COLUMN_LIST( 'cust_id'),
       round_num        => 3);
END;
/

SELECT col, shift, scale FROM norm_tbl;

COL                  SHIFT SCALE
-------------------- ----- -----
CUST_YEAR_OF_BIRTH    1960    15

DECLARE
    xforms      dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
    dbms_data_mining_transform.STACK_NORM_LIN (
        norm_table_name    => 'norm_tbl',
        xform_list        => xforms);
    dbms_data_mining_transform.XFORM_STACK (
        xform_list         => xforms,
        data_table_name    => 'mining_data',
        xform_view_name    => 'mining_data_stack');
END;
/

set long 3000
SQL> SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_STACK';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_GENDER",("CUST_YEAR_OF_BIRTH"-1960)/15 "CUST_YEAR_OF_BIRT
H" FROM mining_data
```

SET_EXPRESSION Procedure

This procedure appends a row to a VARCHAR2 array that stores a SQL expression.

The array can be used for specifying a transformation expression that is too long to be used with the SET_TRANSFORM Procedure.

The GET_EXPRESSION Function returns a row in the array.

When you use SET_EXPRESSION to build a transformation expression, you must build a corresponding reverse transformation expression, create a transformation record, and add the transformation record to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.SET_EXPRESSION (
           expression    IN OUT NOCOPY EXPRESSION_REC,
           chunk                       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-158 SET_EXPRESSION Procedure Parameters

Parameter	Description
`expression`	An expression record (`EXPRESSION_REC`) that specifies a transformation expression or a reverse transformation expression for an attribute. Each expression record includes a `VARCHAR2` array and index fields for specifying upper and lower boundaries within the array. There are two `EXPRESSION_REC` fields within a transformation record (`TRANSFORM_REC`): one for the transformation expression; the other for the reverse transformation expression. See Table 6-127 for a description of the `EXPRESSION_REC` type.
`chunk`	A `VARCHAR2` chunk (row) to be appended to *`expression`*.

Notes

1. You can pass `NULL` in the *`chunk`* argument to `SET_EXPRESSION` to clear the previous chunk. The default value of *`chunk`* is `NULL`.
2. See "About Transformation Lists".
3. See "Operational Notes".

Examples

In this example, two calls to `SET_EXPRESSION` construct a transformation expression and two calls construct the reverse transformation.

> **Note:**
>
> This example is for illustration purposes only. It shows how `SET_EXPRESSION` appends the text provided in *`chunk`* to the text that already exists in *`expression`*. The `SET_EXPRESSION` procedure is meant for constructing very long transformation expressions that cannot be specified in a `VARCHAR2` argument to `SET_TRANSFORM`.
>
> Similarly while transformation lists are intended for embedding in a model, the transformation list `v_xlst` is shown in an external view for illustration purposes.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_year_of_birth, cust_postal_code, cust_credit_limit
       FROM sh.customers;

DECLARE
        v_expr dbms_data_mining_transform.EXPRESSION_REC;
        v_rexp dbms_data_mining_transform.EXPRESSION_REC;
        v_xrec dbms_data_mining_transform.TRANSFORM_REC;
        v_xlst dbms_data_mining_transform.TRANSFORM_LIST :=
                                dbms_data_mining_transform.TRANSFORM_LIST(NULL);
BEGIN
    dbms_data_mining_transform.SET_EXPRESSION(
         EXPRESSION  => v_expr,
         CHUNK       => '("CUST_YEAR_OF_BIRTH"-1910)');
    dbms_data_mining_transform.SET_EXPRESSION(
          EXPRESSION  => v_expr,
          CHUNK       => '/77');
    dbms_data_mining_transform.SET_EXPRESSION(
          EXPRESSION  => v_rexp,
          CHUNK       => '"CUST_YEAR_OF_BIRTH"*77');
    dbms_data_mining_transform.SET_EXPRESSION(
          EXPRESSION  => v_rexp,
          CHUNK       => '+1910');

    v_xrec := null;
    v_xrec.attribute_name := 'CUST_YEAR_OF_BIRTH';
    v_xrec.expression := v_expr;
    v_xrec.reverse_expression := v_rexp;
    v_xlst.TRIM;
    v_xlst.extend(1);
    v_xlst(1) := v_xrec;

    dbms_data_mining_transform.XFORM_STACK (
        xform_list           =>  v_xlst,
        data_table_name      => 'mining_data',
        xform_view_name      => 'v_xlst_view');

    dbms_output.put_line('====');
    FOR i IN 1..v_xlst.count LOOP
      dbms_output.put_line('ATTR: '||v_xlst(i).attribute_name);
      dbms_output.put_line('SUBN: '||v_xlst(i).attribute_subname);
      FOR j IN v_xlst(i).expression.lb..v_xlst(i).expression.ub LOOP
        dbms_output.put_line('EXPR: '||v_xlst(i).expression.lstmt(j));
      END LOOP;
      FOR j IN v_xlst(i).reverse_expression.lb..
                v_xlst(i).reverse_expression.ub LOOP
        dbms_output.put_line('REXP: '||v_xlst(i).reverse_expression.lstmt(j));
      END LOOP;
      dbms_output.put_line('====');
    END LOOP;
  END;
/
====
ATTR: CUST_YEAR_OF_BIRTH
SUBN:
EXPR: ("CUST_YEAR_OF_BIRTH"-1910)
EXPR: /77
REXP: "CUST_YEAR_OF_BIRTH"*77
REXP: +1910
====
```

SET_TRANSFORM Procedure

This procedure appends the transformation instructions for an attribute to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.SET_TRANSFORM (
          xform_list                 IN OUT NOCOPY TRANSFORM_LIST,
          attribute_name             VARCHAR2,
          attribute_subname          VARCHAR2,
          expression                 VARCHAR2,
          reverse_expression         VARCHAR2,
          attribute_spec             VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-159 SET_TRANSFORM Procedure Parameters

Parameter	Description
`xform_list`	A transformation list. See Table 6-127for a description of the `TRANSFORM_LIST` object type.
`attribute_name`	Name of the attribute to be transformed
`attribute_subname`	Name of the nested attribute if *`attribute_name`* is a nested column, otherwise `NULL`.
`expression`	A SQL expression that specifies the transformation of the attribute.
`reverse_expression`	A SQL expression that reverses the transformation for readability in model details and in the target of a supervised model (if the attribute is a target)
`attribute_spec`	One or more keywords that identify special treatment for the attribute during model build. Values are: • `NOPREP` — When ADP is on, prevents automatic transformation of the attribute. If ADP is not on, this value has no effect. • `TEXT` — Causes the attribute to be treated as unstructured text data • `FORCE_IN` — Forces the inclusion of the attribute in the model build. Applies only to GLM models with feature selection enabled (`ftr_selection_enable` = yes). Feature selection is disabled by default. If the model is not using GLM with feature selection, this value has no effect. See "Specifying Transformation Instructions for an Attribute" in *Oracle Machine Learning for SQL User’s Guide*for more information about `attribute_spec`.

Usage Notes

1. See the following relevant sections in "Operational Notes":
   - About Transformation Lists
   - Nested Data Transformations
2. As shown in the following example, you can eliminate an attribute by specifying a null transformation expression and reverse expression. You can also use the STACK interface to remove a column (CREATE_COL_REM Procedure and STACK_COL_REM Procedure).

STACK_BIN_CAT Procedure

This procedure adds categorical binning transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_BIN_CAT (
     bin_table_name     IN            VARCHAR2,
     xform_list         IN OUT NOCOPY TRANSFORM_LIST,
     literal_flag       IN            BOOLEAN  DEFAULT FALSE,
     bin_schema_name    IN            VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-160 STACK_BIN_CAT Procedure Parameters

Parameter	Description
bin_table_name	Name of the transformation definition table for categorical binning. You can use the CREATE_BIN_CAT Procedure to create the definition table. The table must be populated with transformation definitions before you call STACK_BIN_CAT. To populate the table, you can use one of the INSERT procedures for categorical binning or you can write your own SQL. See Table 6-130
xform_list	A transformation list. See Table 6-127 for a description of the TRANSFORM_LIST object type.
literal_flag	Indicates whether the values in the bin column in the transformation definition table are valid SQL literals. When *literal_flag* is FALSE (the default), the bin identifiers will be transformed to SQL literals by surrounding them with single quotes. Set *literal_flag* to TRUE if the bin identifiers are numbers that should have a numeric datatype, as is the case for an O-Cluster model. See "INSERT_BIN_NUM_EQWIDTH Procedure" for an example.
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how a binning transformation for the categorical column cust_postal_code could be added to a stack called mining_data_stack.

> **Note:**
>
> This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. XFORM_STACK simply generates an external view of the transformed data. The actual purpose of the STACK procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to CREATE_MODEL in the xform_list parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE or REPLACE VIEW mining_data AS
    SELECT cust_id, cust_postal_code, cust_credit_limit
       FROM sh.customers
       WHERE cust_id BETWEEN 100050 AND 100100;
BEGIN
   dbms_data_mining_transform.CREATE_BIN_CAT ('bin_cat_tbl');
   dbms_data_mining_transform.INSERT_BIN_CAT_FREQ (
        bin_table_name   => 'bin_cat_tbl',
        data_table_name  => 'mining_data',
        bin_num          =>  3);
  END;
/
DECLARE
  MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
  dbms_data_mining_transform.STACK_BIN_CAT (
    bin_table_name       => 'bin_cat_tbl',
    xform_list           =>  mining_data_stack);
  dbms_data_mining_transform.XFORM_STACK (
    xform_list           =>  mining_data_stack,
    data_table_name      => 'mining_data',
    xform_view_name      => 'mining_data_stack_view');
  END;
/
-- Before transformation
column cust_postal_code format a16
SELECT * from mining_data
              WHERE cust_id BETWEEN 100050 AND 100053
              ORDER BY cust_id;

  CUST_ID CUST_POSTAL_CODE CUST_CREDIT_LIMIT
---------- ---------------- -----------------
    100050 76486                         1500
    100051 73216                         9000
    100052 69499                         5000
    100053 45704                         7000

-- After transformation
SELECT * FROM mining_data_stack_view
              WHERE cust_id BETWEEN 100050 AND 100053
              ORDER BY cust_id;

  CUST_ID CUST_POSTAL_CODE CUST_CREDIT_LIMIT
---------- ---------------- -----------------
    100050 4                             1500
    100051 1                             9000
    100052 4                             5000
    100053 4                             7000
```

STACK_BIN_NUM Procedure

This procedure adds numerical binning transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_BIN_NUM (
     bin_table_name     IN             VARCHAR2,
     xform_list         IN OUT  NOCOPY TRANSFORM_LIST,
     literal_flag       IN             BOOLEAN  DEFAULT FALSE,
     bin_schema_name    IN             VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-161 STACK_BIN_NUM Procedure Parameters

Parameter	Description
bin_table_name	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The table must be populated with transformation definitions before you call STACK_BIN_NUM. To populate the table, you can use one of the INSERT procedures for numerical binning or you can write your own SQL. See Table 6-132.
xform_list	A transformation list. See Table 6-127 for a description of the TRANSFORM_LIST object type.
literal_flag	Indicates whether the values in the bin column in the transformation definition table are valid SQL literals. When *literal_flag* is FALSE (the default), the bin identifiers will be transformed to SQL literals by surrounding them with single quotes. Set *literal_flag* to TRUE if the bin identifiers are numbers that should have a numeric datatype, as is the case for an O-Cluster model. See "INSERT_BIN_NUM_EQWIDTH Procedure" for an example.
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how a binning transformation for the numerical column cust_credit_limit could be added to a stack called mining_data_stack.

> **Note:**
>
> This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. XFORM_STACK simply generates an external view of the transformed data. The actual purpose of the STACK procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to CREATE_MODEL in the xform_list parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE OR REPLACE VIEW mining_data AS
    SELECT cust_id, cust_postal_code, cust_credit_limit
      FROM sh.customers
      WHERE cust_id BETWEEN 100050 and 100100;
BEGIN
  dbms_data_mining_transform.create_bin_num ('bin_num_tbl');
  dbms_data_mining_transform.insert_bin_num_qtile (
  bin_table_name    => 'bin_num_tbl',
  data_table_name   => 'mining_data',
  bin_num               => 5,
  exclude_list      =>  dbms_data_mining_transform.COLUMN_LIST('cust_id'));
END;
/
DECLARE
   MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
   dbms_data_mining_transform.STACK_BIN_CAT (
      bin_table_name        => 'bin_num_tbl',
      xform_list            =>  mining_data_stack);
   dbms_data_mining_transform.XFORM_STACK (
      xform_list            =>  mining_data_stack,
      data_table_name   => 'mining_data',
      xform_view_name   => 'mining_data_stack_view');
END;
/
-- Before transformation
SELECT cust_id, cust_postal_code, ROUND(cust_credit_limit) FROM mining_data
   WHERE cust_id BETWEEN 100050 AND 100055
   ORDER BY cust_id;
CUST_ID   CUST_POSTAL_CODE   ROUND(CUST_CREDIT_LIMIT)
-------   -----------------  -------------------------
100050    76486                                   1500
100051    73216                                   9000
100052    69499                                   5000
100053    45704                                   7000
100055    74673                                  11000
100055    74673                                  11000

-- After transformation
SELECT cust_id, cust_postal_code, ROUND(cust_credit_limit)
   FROM mining_data_stack_view
   WHERE cust_id BETWEEN 100050 AND 100055
   ORDER BY cust_id;
CUST_ID   CUST_POSTAL_CODE   ROUND(CUST_CREDIT_LIMITT)
-------   ----------------   -------------------------
100050    76486
100051    73216                                     2
100052    69499                                     1
100053    45704
100054    88021                                     3
100055    74673                                     3
```

STACK_CLIP Procedure

This procedure adds clipping transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_CLIP (
          clip_table_name     IN            VARCHAR2,
          xform_list          IN OUT NOCOPY TRANSFORM_LIST,
          clip_schema_name    IN            VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-162 STACK_CLIP Procedure Parameters

Parameter	Description
`clip_table_name`	Name of the transformation definition table for clipping.You can use the CREATE_CLIP Procedure to create the definition table. The table must be populated with transformation definitions before you call `STACK_CLIP`. To populate the table, you can use one of the `INSERT` procedures for clipping or you can write your own SQL. See Table 6-134
`xform_list`	A transformation list. See Table 6-127 for a description of the `TRANSFORM_LIST` object type.
`clip_schema_name`	Schema of *`clip_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

See DBMS_DATA_MINING_TRANSFORM Operational Notes. The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how a clipping transformation for the numerical column `cust_credit_limit` could be added to a stack called `mining_data_stack`.

> **Note:**
>
> This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. `XFORM_STACK` simply generates an external view of the transformed data. The actual purpose of the `STACK` procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to `CREATE_MODEL` in the `xform_list` parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, cust_postal_code, cust_credit_limit
       FROM sh.customers
       WHERE cust_id BETWEEN 100050 AND 100100;
BEGIN
   dbms_data_mining_transform.create_clip ('clip_tbl');
   dbms_data_mining_transform.insert_clip_winsor_tail (
       clip_table_name   => 'clip_tbl',
       data_table_name   => 'mining_data',
       tail_frac         => 0.25,
       exclude_list      => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
END;
/
```

```
DECLARE
      MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
      dbms_data_mining_transform.STACK_CLIP (
         clip_table_name     => 'clip_tbl',
         xform_list          =>  mining_data_stack);
      dbms_data_mining_transform.XFORM_STACK (
          xform_list        =>  mining_data_stack,
          data_table_name   => 'mining_data',
          xform_view_name   => 'mining_data_stack_view');
END;
/
-- Before transformation
SELECT cust_id, cust_postal_code, round(cust_credit_limit)
  FROM mining_data
    WHERE cust_id BETWEEN 100050 AND 100054
    ORDER BY cust_id;

CUST_ID   CUST_POSTAL_CODE   ROUND(CUST_CREDIT_LIMIT)
-------   ----------------   ------------------------
100050    76486                                  1500
100051    73216                                  9000
100052    69499                                  5000
100053    45704                                  7000
100054    88021                                 11000

-- After transformation
SELECT cust_id, cust_postal_code, round(cust_credit_limit)
  FROM mining_data_stack_view
    WHERE cust_id BETWEEN 100050 AND 100054
    ORDER BY cust_id;

CUST_ID   CUST_POSTAL_CODE   ROUND(CUST_CREDIT_LIMIT)
-------   ----------------   ------------------------
100050    76486                                  5000
100051    73216                                  9000
100052    69499                                  5000
100053    45704                                  7000
100054    88021                                 11000
```

STACK_COL_REM Procedure

This procedure adds column removal transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_COL_REM (
     rem_table_name     IN            VARCHAR2,
     xform_list         IN OUT NOCOPY TRANSFORM_LIST,
     rem_schema_name    IN            VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-163 STACK_COL_REM Procedure Parameters

Parameter	Description
`rem_table_name`	Name of the transformation definition table for column removal. You can use the CREATE_COL_REM Procedure to create the definition table. See Table 6-136. The table must be populated with column names before you call `STACK_COL_REM`. The INSERT_BIN_SUPER Procedure and the INSERT_AUTOBIN_NUM_EQWIDTH Procedure can optionally be used to populate the table. You can also use SQL `INSERT` statements.
`xform_list`	A transformation list. See Table 6-127 for a description of the `TRANSFORM_LIST` object type.
`rem_schema_name`	Schema of *`rem_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how the column `cust_credit_limit` could be removed in a transformation list called `mining_data_stack`.

> **Note:**
>
> This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. `XFORM_STACK` simply generates an external view of the transformed data. The actual purpose of the `STACK` procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to `CREATE_MODEL` in the `xform_list` parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE OR REPLACE VIEW mining_data AS
     SELECT cust_id, country_id, cust_postal_code, cust_credit_limit
        FROM sh.customers;

BEGIN
    dbms_data_mining_transform.create_col_rem ('rem_tbl');
END;
/

INSERT into rem_tbl VALUES (upper('cust_postal_code'), null);

DECLARE
  MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
```

```
     dbms_data_mining_transform.stack_col_rem (
          rem_table_name     => 'rem_tbl',
          xform_list         =>  mining_data_stack);
      dbms_data_mining_transform.XFORM_STACK (
          xform_list         =>  mining_data_stack,
          data_table_name    => 'mining_data',
          xform_view_name    => 'mining_data_stack_view');
END;
/

SELECT *  FROM mining_data
  WHERE cust_id BETWEEN 100050 AND 100051
  ORDER BY cust_id;

CUST_ID   COUNTRY_ID   CUST_POSTAL_CODE   CUST_CREDIT_LIMIT
-------   ----------   ----------------   -----------------
100050         52773   76486                           1500
100051         52790   73216                           9000

SELECT *  FROM mining_data_stack_view
  WHERE cust_id BETWEEN 100050 AND 100051
  ORDER BY cust_id;

CUST_ID   COUNTRY_ID   CUST_CREDIT_LIMIT
-------   ----------   -----------------
100050         52773              1500
100051         52790              9000
```

STACK_MISS_CAT Procedure

This procedure adds categorical missing value transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_MISS_CAT (
     miss_table_name     IN       VARCHAR2,
     xform_list          IN OUT   NOCOPY TRANSFORM_LIST,
     miss_schema_name    IN       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-164 STACK_MISS_CAT Procedure Parameters

Parameter	Description
`miss_table_name`	Name of the transformation definition table for categorical missing value treatment. You can use the CREATE_MISS_CAT Procedure to create the definition table. The table must be populated with transformation definitions before you call `STACK_MISS_CAT`. To populate the table, you can use the INSERT_MISS_CAT_MODE Procedure or you can write your own SQL. See Table 6-138.
`xform_list`	A transformation list. See Table 6-127 for a description of the `TRANSFORM_LIST` object type.
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how the missing values in the column `cust_marital_status` could be replaced with the mode in a transformation list called `mining_data_stack`.

> **Note:**
>
> This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. `XFORM_STACK` simply generates an external view of the transformed data. The actual purpose of the `STACK` procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to `CREATE_MODEL` in the `xform_list` parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE OR REPLACE VIEW mining_data AS
      SELECT cust_id, country_id, cust_marital_status
         FROM sh.customers
         where cust_id BETWEEN 1 AND 10;

BEGIN
  dbms_data_mining_transform.create_miss_cat ('miss_cat_tbl');
  dbms_data_mining_transform.insert_miss_cat_mode ('miss_cat_tbl', 'mining_data');
END;
/

DECLARE
  MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
     dbms_data_mining_transform.stack_miss_cat (
          miss_table_name   => 'miss_cat_tbl',
          xform_list        => mining_data_stack);
      dbms_data_mining_transform.XFORM_STACK (
          xform_list        =>  mining_data_stack,
          data_table_name   => 'mining_data',
          xform_view_name   => 'mining_data_stack_view');
END;
/
SELECT * FROM mining_data
  ORDER BY cust_id;

CUST_ID   COUNTRY_ID   CUST_MARITAL_STATUS
-------   ----------   --------------------
      1      52789
      2      52778
      3      52770
      4      52770
      5      52789
      6      52769     single
      7      52790     single
      8      52790     married
      9      52770     divorced
     10      52790     widow
```

```
SELECT * FROM mining_data_stack_view
   ORDER By cust_id;

CUST_ID   COUNTRY_ID   CUST_MARITAL_STATUS
-------   -----------  --------------------
      1       52789    single
      2       52778    single
      3       52770    single
      4       52770    single
      5       52789    single
      6       52769    single
      7       52790    single
      8       52790    married
      9       52770    divorced
     10       52790    widow
```

STACK_MISS_NUM Procedure

This procedure adds numeric missing value transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_MISS_NUM (
     miss_table_name     IN       VARCHAR2,
     xform_list          IN OUT   NOCOPY TRANSFORM_LIST,
     miss_schema_name    IN       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-165 STACK_MISS_NUM Procedure Parameters

Parameter	Description
miss_table_name	Name of the transformation definition table for numerical missing value treatment. You can use the CREATE_MISS_NUM Procedure to create the definition table. The table must be populated with transformation definitions before you call STACK_MISS_NUM. To populate the table, you can use the INSERT_MISS_NUM_MEAN Procedure or you can write your own SQL. See Table 6-140.
xform_list	A transformation list. See Table 6-127 for a description of the TRANSFORM_LIST object type.
miss_schema_name	Schema of *miss_table_name*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how the missing values in the column cust_credit_limit could be replaced with the mean in a transformation list called mining_data_stack.

Note:

This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. XFORM_STACK simply generates an external view of the transformed data. The actual purpose of the STACK procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to CREATE_MODEL in the xform_list parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
describe mining_data
 Name                                                  Null?    Type
 ----------------------------------------------------- -------- -----
 CUST_ID                                               NOT NULL NUMBER
 CUST_CREDIT_LIMIT                                              NUMBER

BEGIN
   dbms_data_mining_transform.create_miss_num ('miss_num_tbl');
   dbms_data_mining_transform.insert_miss_num_mean ('miss_num_tbl','mining_data');
END;
/
SELECT * FROM miss_num_tbl;

COL                  ATT      VAL
-------------------- ----- ------
CUST_ID                       5.5
CUST_CREDIT_LIMIT          185.71

DECLARE
    MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
  BEGIN
    dbms_data_mining_transform.STACK_MISS_NUM (
          miss_table_name   => 'miss_num_tbl',
          xform_list        => mining_data_stack);
    dbms_data_mining_transform.XFORM_STACK (
          xform_list        =>  mining_data_stack,
          data_table_name   => 'mining_data',
          xform_view_name   => 'mining_data_stack_view');
END;
/
-- Before transformation
SELECT * FROM mining_data
  ORDER BY cust_id;
CUST_ID CUST_CREDIT_LIMIT
------- -----------------
      1               100
      2
      3               200
      4
      5               150
      6               400
      7               150
      8
      9               100
     10               200

-- After transformation
SELECT * FROM mining_data_stack_view
  ORDER BY cust_id;
```

```
CUST_ID CUST_CREDIT_LIMIT
------- -----------------
      1               100
      2            185.71
      3               200
      4            185.71
      5               150
      6               400
      7               150
      8            185.71
      9               100
     10               200
```

STACK_NORM_LIN Procedure

This procedure adds linear normalization transformations to a transformation list.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.STACK_NORM_LIN (
     norm_table_name      IN       VARCHAR2,
     xform_list           IN OUT   NOCOPY TRANSFORM_LIST,
     norm_schema_name     IN       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-166 STACK_NORM_LIN Procedure Parameters

Parameter	Description
`norm_table_name`	Name of the transformation definition table for linear normalization. You can use the CREATE_NORM_LIN Procedure to create the definition table. The table must be populated with transformation definitions before you call `STACK_NORM_LIN`.To populate the table, you can use one of the `INSERT` procedures for normalization or you can write your own SQL. See Table 6-142.
`xform_list`	A transformation list. See Table 6-127 for a description of the `TRANSFORM_LIST` object type.
`norm_schema_name`	Schema of *`norm_table_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example shows how the column `cust_credit_limit` could be normalized in a transformation list called `mining_data_stack`.

Note:

This example invokes the XFORM_STACK Procedure to show how the data is transformed by the stack. `XFORM_STACK` simply generates an external view of the transformed data. The actual purpose of the `STACK` procedures is to assemble a list of transformations for embedding in a model. The transformations are passed to `CREATE_MODEL` in the `xform_list` parameter. See INSERT_BIN_NUM_EQWIDTH Procedure for an example.

```
CREATE OR REPLACE VIEW mining_data AS
       SELECT cust_id, country_id, cust_postal_code, cust_credit_limit
          FROM sh.customers;
BEGIN
   dbms_data_mining_transform.create_norm_lin ('norm_lin_tbl');
   dbms_data_mining_transform.insert_norm_lin_minmax (
       norm_table_name   => 'norm_lin_tbl',
       data_table_name   => 'mining_data',
       exclude_list      =>  dbms_data_mining_transform.COLUMN_LIST('cust_id',
                                                              'country_id'));
END;
/
SELECT * FROM norm_lin_tbl;
COL                  ATT    SHIFT  SCALE
-------------------- ----- ------ ------
CUST_CREDIT_LIMIT            1500  13500

DECLARE
   MINING_DATA_STACK   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
   dbms_data_mining_transform.stack_norm_lin (
        norm_table_name   => 'norm_lin_tbl',
        xform_list        => mining_data_stack);
   dbms_data_mining_transform.XFORM_STACK (
        xform_list        =>  mining_data_stack,
        data_table_name   => 'mining_data',
        xform_view_name   => 'mining_data_stack_view');
END;
/
SELECT * FROM mining_data
  WHERE cust_id between 1 and 10
  ORDER BY cust_id;
CUST_ID COUNTRY_ID CUST_POSTAL_CODE     CUST_CREDIT_LIMIT
------- ---------- -------------------- -----------------
      1      52789 30828                             9000
      2      52778 86319                            10000
      3      52770 88666                             1500
      4      52770 87551                             1500
      5      52789 59200                             1500
      6      52769 77287                             1500
      7      52790 38763                             1500
      8      52790 58488                             3000
      9      52770 63033                             3000
     10      52790 52602                             3000

SELECT * FROM mining_data_stack_view
  WHERE cust_id between 1 and 10
  ORDER BY cust_id;
CUST_ID COUNTRY_ID CUST_POSTAL_CODE     CUST_CREDIT_LIMIT
```

```
------- ---------- -------------------- -----------------
      1      52789 30828                           .55556
      2      52778 86319                           .62963
      3      52770 88666                                0
      4      52770 87551                                0
      5      52789 59200                                0
      6      52769 77287                                0
      7      52790 38763                                0
      8      52790 58488                           .11111
      9      52770 63033                           .11111
     10      52790 52602                           .11111
```

XFORM_BIN_CAT Procedure

This procedure creates a view that implements the categorical binning transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_BIN_CAT (
     bin_table_name       IN VARCHAR2,
     data_table_name      IN VARCHAR2,
     xform_view_name      IN VARCHAR2,
     literal_flag         IN BOOLEAN DEFAULT FALSE,
     bin_schema_name      IN VARCHAR2 DEFAULT NULL,
     data_schema_name     IN VARCHAR2 DEFAULT NULL,
     xform_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-167 XFORM_BIN_CAT Procedure Parameters

Parameter	Description
bin_table_name	Name of the transformation definition table for categorical binning. You can use the CREATE_BIN_CAT Procedure to create the definition table. The table must be populated with transformation definitions before you call XFORM_BIN_CAT. To populate the table, you can use one of the INSERT procedures for categorical binning or you can write your own SQL. See Table 6-130.
data_table_name	Name of the table containing the data to be transformed.
xform_view_name	Name of the view to be created. The view presents columns in *data_table_name* with the transformations specified in *bin_table_name*.
literal_flag	Indicates whether the values in the bin column in the transformation definition table are valid SQL literals. When *literal_flag* is FALSE (the default), the bin identifiers will be transformed to SQL literals by surrounding them with single quotes. Set *literal_flag* to TRUE if the bin identifiers are numbers that should have a numeric datatype, as is the case for an O-Cluster model. See "INSERT_BIN_NUM_EQWIDTH Procedure" for an example.
bin_schema_name	Schema of *bin_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Table 6-167 (Cont.) XFORM_BIN_CAT Procedure Parameters

Parameter	Description
xform_schema_name	Schema of *xform_view_name*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that bins the cust_postal_code column. The data source consists of three columns from sh.customer.

```
describe mining_data
 Name                                    Null?    Type
 -------------------------------------- -------- ------------------------
 CUST_ID                                 NOT NULL NUMBER
 CUST_POSTAL_CODE                        NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                                NUMBER

SELECT * FROM mining_data WHERE cust_id between 104066 and 104069;

   CUST_ID CUST_POSTAL_CODE
CUST_CREDIT_LIMIT
--------- --------------------
-----------------
    104066 69776
7000
    104067 52602
9000
    104068 55787
11000
    104069 55977
5000

BEGIN
  dbms_data_mining_transform.create_bin_cat(
     bin_table_name     => 'bin_cat_tbl');
  dbms_data_mining_transform.insert_bin_cat_freq(
     bin_table_name     => 'bin_cat_tbl',
     data_table_name    => 'mining_data',
     bin_num          => 10);
   dbms_data_mining_transform.xform_bin_cat(
     bin_table_name     => 'bin_cat_tbl',
     data_table_name    => 'mining_data',
     xform_view_name    => 'bin_cat_view');
END;
/

SELECT * FROM bin_cat_view WHERE cust_id between 104066 and 104069;

   CUST_ID CUST_POSTAL_CODE
CUST_CREDIT_LIMIT
---------- --------------------
-----------------
    104066 6
7000
```

```
    104067 11
9000
    104068 3
11000
    104069 11
5000

SELECT text FROM user_views WHERE view_name IN 'BIN_CAT_VIEW';

TEXT

--------------------------------------------------------------------------------

SELECT
"CUST_ID",DECODE("CUST_POSTAL_CODE",'38082','1','45704','9','48346','5','

55787','3','63736','2','67843','7','69776','6','72860','10','78558','4','80841',

'8',NULL,NULL,'11') "CUST_POSTAL_CODE","CUST_CREDIT_LIMIT" FROM
mining_data
```

XFORM_BIN_NUM Procedure

This procedure creates a view that implements the numerical binning transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_BIN_NUM (
     bin_table_name      IN VARCHAR2,
     data_table_name     IN VARCHAR2,
     xform_view_name     IN VARCHAR2,
     literal_flag        IN BOOLEAN DEFAULT FALSE,
     bin_schema_name     IN VARCHAR2 DEFAULT NULL,
     data_schema_name    IN VARCHAR2 DEFAULT NULL,
     xform_schema_name   IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-168 XFORM_BIN_NUM Procedure Parameters

Parameter	Description
`bin_table_name`	Name of the transformation definition table for numerical binning. You can use the CREATE_BIN_NUM Procedure to create the definition table. The table must be populated with transformation definitions before you call `XFORM_BIN_NUM`. To populate the table, you can use one of the `INSERT` procedures for numerical binning or you can write your own SQL. See "Table 6-132".
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view presents columns in *`data_table_name`* with the transformations specified in *`bin_table_name`*.

Table 6-168 (Cont.) XFORM_BIN_NUM Procedure Parameters

Parameter	Description
`literal_flag`	Indicates whether the values in the `bin` column in the transformation definition table are valid SQL literals. When *`literal_flag`* is `FALSE` (the default), the bin identifiers will be transformed to SQL literals by surrounding them with single quotes. Set *`literal_flag`* to `TRUE` if the bin identifiers are numbers that should have a numeric datatype, as is the case for an O-Cluster model. See "INSERT_BIN_NUM_EQWIDTH Procedure" for an example.
`bin_schema_name`	Schema of *`bin_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that bins the `cust_credit_limit` column. The data source consists of three columns from `sh.customer`.

```
describe mining_data
 Name                                      Null?    Type
 --------------------------------------- -------- ------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_POSTAL_CODE                          NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                                  NUMBER

column cust_credit_limit off
SELECT * FROM mining_data WHERE cust_id between 104066 and 104069;

   CUST_ID CUST_POSTAL_CODE   CUST_CREDIT_LIMIT
---------  ------------------
--------------------
    104066 69776                           7000
    104067 52602                           9000
    104068 55787                          11000
    104069 55977                           5000

BEGIN
   dbms_data_mining_transform.create_bin_num(
           bin_table_name     => 'bin_num_tbl');
   dbms_data_mining_transform.insert_autobin_num_eqwidth(
           bin_table_name     => 'bin_num_tbl',
           data_table_name    => 'mining_data',
           bin_num                => 5,
           max_bin_num            => 10,
           exclude_list       => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
  dbms_data_mining_transform.xform_bin_num(
          bin_table_name      => 'bin_num_tbl',
          data_table_name     => 'mining_data',
          xform_view_name     => 'mining_data_view');
```

```
END;
/
describe mining_data_view
 Name                                    Null?    Type
 ------------------------------------ -------- ------------------------
 CUST_ID                                 NOT NULL NUMBER
 CUST_POSTAL_CODE                        NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                                VARCHAR2(2)

col cust_credit_limit on
col cust_credit_limit format a25
SELECT * FROM mining_data_view WHERE cust_id between 104066 and 104069;

   CUST_ID CUST_POSTAL_CODE
CUST_CREDIT_LIMIT
---------- --------------------
-------------------------
    104066 69776
5
    104067 52602
6
    104068 55787
8
    104069 55977
3

set long 2000
SELECT text FROM user_views WHERE view_name IN 'MINING_DATA_VIEW';

TEXT

--------------------------------------------------------------------------------

SELECT "CUST_ID","CUST_POSTAL_CODE",CASE WHEN "CUST_CREDIT_LIMIT"<1500 THEN
NULL
 WHEN "CUST_CREDIT_LIMIT"<=2850 THEN '1' WHEN "CUST_CREDIT_LIMIT"<=4200 THEN
'2'
 WHEN "CUST_CREDIT_LIMIT"<=5550 THEN '3' WHEN "CUST_CREDIT_LIMIT"<=6900 THEN
'4'
 WHEN "CUST_CREDIT_LIMIT"<=8250 THEN '5' WHEN "CUST_CREDIT_LIMIT"<=9600 THEN
'6'
 WHEN "CUST_CREDIT_LIMIT"<=10950 THEN '7' WHEN "CUST_CREDIT_LIMIT"<=12300 THEN
'
8' WHEN "CUST_CREDIT_LIMIT"<=13650 THEN '9' WHEN "CUST_CREDIT_LIMIT"<=15000
THEN
 '10' END "CUST_CREDIT_LIMIT" FROM mining_data
```

XFORM_CLIP Procedure

This procedure creates a view that implements the clipping transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_CLIP (
    clip_table_name       IN VARCHAR2,
    data_table_name       IN VARCHAR2,
    xform_view_name       IN VARCHAR2,
    clip_schema_name      IN VARCHAR2 DEFAULT NULL,
    data_schema_name      IN VARCHAR2,DEFAULT NULL,
    xform_schema_name     IN VARCHAR2,DEFAULT NULL);
```

Parameters

Table 6-169 XFORM_CLIP Procedure Parameters

Parameter	Description
clip_table_name	Name of the transformation definition table for clipping. You can use the CREATE_CLIP Procedure to create the definition table. The table must be populated with transformation definitions before you call XFORM_CLIP. To populate the table, you can use one of the INSERT procedures for clipping you can write your own SQL. See Table 6-134.
data_table_name	Name of the table containing the data to be transformed
xform_view_name	Name of the view to be created. The view presents columns in *data_table_name* with the transformations specified in *clip_table_name*.
clip_schema_name	Schema of *clip_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.
xform_schema_name	Schema of *xform_view_name*. If no schema is specified, the current schema is used.

Examples

This example creates a view that clips the cust_credit_limit column. The data source consists of three columns from sh.customer.

```
describe mining_data
 Name                            Null?    Type
 ------------------------------ -------- -------------------------
 CUST_ID                         NOT NULL NUMBER
 CUST_POSTAL_CODE                NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                         NUMBER

BEGIN
   dbms_data_mining_transform.create_clip(
      clip_table_name    => 'clip_tbl');
   dbms_data_mining_transform.insert_clip_trim_tail(
      clip_table_name   => 'clip_tbl',
      data_table_name   => 'mining_data',
      tail_frac         => 0.05,
      exclude_list      => dbms_data_mining_transform.COLUMN_LIST('cust_id'));
   dbms_data_mining_transform.xform_clip(
      clip_table_name     => 'clip_tbl',
      data_table_name    => 'mining_data',
      xform_view_name    => 'clip_view');
END;
/
describe clip_view
 Name                           Null?    Type
 ----------------------------- -------- --------------------------
 CUST_ID                        NOT NULL NUMBER
 CUST_POSTAL_CODE               NOT NULL VARCHAR2(10)
 CUST_CREDIT_LIMIT                       NUMBER
```

```
SELECT MIN(cust_credit_limit), MAX(cust_credit_limit) FROM mining_data;

MIN(CUST_CREDIT_LIMIT) MAX(CUST_CREDIT_LIMIT)
---------------------- ----------------------
                  1500                  15000

SELECT MIN(cust_credit_limit), MAX(cust_credit_limit) FROM clip_view;

MIN(CUST_CREDIT_LIMIT) MAX(CUST_CREDIT_LIMIT)
---------------------- ----------------------
                  1500                  11000

set long 2000
SELECT text FROM user_views WHERE view_name IN 'CLIP_VIEW';

TEXT
--------------------------------------------------------------------------------
SELECT "CUST_ID","CUST_POSTAL_CODE",CASE WHEN "CUST_CREDIT_LIMIT" < 1500 THEN NU
LL WHEN "CUST_CREDIT_LIMIT" > 11000 THEN NULL ELSE "CUST_CREDIT_LIMIT" END "CUST
_CREDIT_LIMIT" FROM mining_data
```

XFORM_COL_REM Procedure

This procedure creates a view that implements the column removal transformations specified in a definition table. Only the columns that are specified in the definition table are removed; the remaining columns from the data table are present in the view.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_COL_REM (
     rem_table_name     IN       VARCHAR2,
     data_table_name    IN       VARCHAR2,
     xform_view_name    IN       VARCHAR2,
     rem_schema_name    IN       VARCHAR2 DEFAULT NULL,
     data_schema_name   IN       VARCHAR2 DEFAULT NULL,
     xform_schema_name  IN       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-170 XFORM_COL_REM Procedure Parameters

Parameter	Description
rem_table_name	Name of the transformation definition table for column removal. You can use the CREATE_COL_REM Procedure to create the definition table. See Table 6-136. The table must be populated with column names before you call XFORM_COL_REM. The INSERT_BIN_SUPER Procedure and the INSERT_AUTOBIN_NUM_EQWIDTH Procedure can optionally be used to populate the table. You can also use SQL INSERT statements.
data_table_name	Name of the table containing the data to be transformed
xform_view_name	Name of the view to be created. The view presents the columns in *data_table_name* that are not specified in *rem_table_name*.
rem_schema_name	Schema of *rem_table_name*. If no schema is specified, the current schema is used.
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.

Table 6-170 (Cont.) XFORM_COL_REM Procedure Parameters

Parameter	Description
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that includes all but one column from the table `customers` in the current schema.

```
describe customers
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_MARITAL_STATUS                                VARCHAR2(20)
 OCCUPATION                                         VARCHAR2(21)
 AGE                                                NUMBER
 YRS_RESIDENCE                                      NUMBER

BEGIN
    DBMS_DATA_MINING_TRANSFORM.CREATE_COL_REM ('colrem_xtbl');
END;
 /
INSERT INTO colrem_xtbl VALUES('CUST_MARITAL_STATUS', null);

NOTE: This currently doesn't work. See bug 9310319

BEGIN
   DBMS_DATA_MINING_TRANSFORM.XFORM_COL_REM (
     rem_table_name        => 'colrem_xtbl',
     data_table_name       => 'customers',
     xform_view_name       => 'colrem_view');
END;
/
describe colrem_view

 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 CUST_ID                                   NOT NULL NUMBER
 OCCUPATION                                         VARCHAR2(21)
 AGE                                                NUMBER
 YRS_RESIDENCE                                      NUMBER
```

XFORM_EXPR_NUM Procedure

This procedure creates a view that implements the specified numeric transformations. Only the columns that you specify are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_EXPR_NUM (
     expr_pattern       IN       VARCHAR2,
     data_table_name     IN       VARCHAR2,
     xform_view_name     IN       VARCHAR2,
     exclude_list        IN       COLUMN_LIST DEFAULT NULL,
     include_list        IN       COLUMN_LIST DEFAULT NULL,
     col_pattern         IN       VARCHAR2 DEFAULT ':col',
     data_schema_name    IN       VARCHAR2 DEFAULT NULL,
     xform_schema_name   IN       VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-171   XFORM_EXPR_NUM Procedure Parameters

Parameter	Description
`expr_pattern`	A numeric transformation expression
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view presents columns in *`data_table_name`* with the transformations specified in *`expr_pattern`* and *`col_pattern`*.
`exclude_list`	List of numerical columns to exclude. If `NULL`, no numerical columns are excluded. The format of *`exclude_list`* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`
`include_list`	List of numeric columns to include. If `NULL`, all numeric columns are included. The format of *`include_list`* is: `dbms_data_mining_transform.COLUMN_LIST('col1','col2', ...'coln')`
`col_pattern`	The value within *`expr_pattern`* that will be replaced with a column name. The value of *`col_pattern`* is case-sensitive. The default value of *`col_pattern`* is `':col'`
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

1. The `XFORM_EXPR_NUM` procedure constructs numeric transformation expressions from the specified expression pattern (*`expr_pattern`*) by replacing every occurrence of the specified column pattern (*`col_pattern`*) with an actual column name.

   `XFORM_EXPR_NUM` uses the SQL `REPLACE` function to construct the transformation expressions.

   ```
   REPLACE (expr_pattern,col_pattern,'"column_name"') || '"column_name"'
   ```

   If there is a column match, then the replacement is made in the transformation expression; if there is not a match, then the column is used without transformation.

> **See:**
>
> *Oracle Database SQL Language Reference* for information about the REPLACE function

2. Because of the include and exclude list parameters, the XFORM_EXPR_NUM and XFORM_EXPR_STR procedures allow you to easily specify individual columns for transformation within large data sets. The other XFORM_* procedures support an exclude list only. In these procedures, you must enumerate every column that you do not want to transform.
3. See "Operational Notes"

Examples

This example creates a view that transforms the datatype of numeric columns.

```
describe customers
 Name                                   Null?    Type
 -------------------------------------- -------- ------------------------
 CUST_ID                                NOT NULL NUMBER
 CUST_MARITAL_STATUS                             VARCHAR2(20)
 OCCUPATION                                      VARCHAR2(21)
 AGE                                             NUMBER
 YRS_RESIDENCE                                   NUMBER

BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_EXPR_NUM(
    expr_pattern         => 'to_char(:col)',
    data_table_name      => 'customers',
    xform_view_name      => 'cust_nonum_view',
    exclude_list         => dbms_data_mining_transform.COLUMN_LIST( 'cust_id'),
    include_list         => null,
    col_pattern          => ':col');
END;
/
describe cust_nonum_view
 Name                                   Null?    Type
 -------------------------------------- -------- ------------------------
 CUST_ID                                NOT NULL NUMBER
 CUST_MARITAL_STATUS                             VARCHAR2(20)
 OCCUPATION                                      VARCHAR2(21)
 AGE                                             VARCHAR2(40)
 YRS_RESIDENCE                                   VARCHAR2(40)
```

XFORM_EXPR_STR Procedure

This procedure creates a view that implements the specified categorical transformations. Only the columns that you specify are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_EXPR_STR (
     expr_pattern       IN       VARCHAR2,
     data_table_name    IN       VARCHAR2,
     xform_view_name    IN       VARCHAR2,
     exclude_list       IN       COLUMN_LIST DEFAULT NULL,
     include_list       IN       COLUMN_LIST DEFAULT NULL,
      col_pattern          IN        VARCHAR2 DEFAULT ':col',
      data_schema_name     IN        VARCHAR2 DEFAULT NULL,
      xform_schema_name    IN        VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-172 XFORM_EXPR_STR Procedure Parameters

Parameter	Description
expr_pattern	A character transformation expression
data_table_name	Name of the table containing the data to be transformed
xform_view_name	Name of the view to be created. The view presents columns in *data_table_name* with the transformations specified in *expr_pattern* and *col_pattern*.
exclude_list	List of categorical columns to exclude. If NULL, no categorical columns are excluded. The format of *exclude_list* is: dbms_data_mining_transform.COLUMN_LIST('col*1*','col*2*', ...'col*n*')
include_list	List of character columns to include. If NULL, all character columns are included. The format of *include_list* is: dbms_data_mining_transform.COLUMN_LIST('col*1*','col*2*', ...'col*n*')
col_pattern	The value within *expr_pattern* that will be replaced with a column name. The value of *col_pattern* is case-sensitive. The default value of *col_pattern* is ':col'
data_schema_name	Schema of *data_table_name*. If no schema is specified, the current schema is used.
xform_schema_name	Schema of *xform_view_name*. If no schema is specified, the current schema is used.

Usage Notes

1. The XFORM_EXPR_STR procedure constructs character transformation expressions from the specified expression pattern (*expr_pattern*) by replacing every occurrence of the specified column pattern (*col_pattern*) with an actual column name.

   XFORM_EXPR_STR uses the SQL REPLACE function to construct the transformation expressions.

   ```
   REPLACE (expr_pattern,col_pattern,'"column_name"') || '"column_name"'
   ```

   If there is a column match, then the replacement is made in the transformation expression; if there is not a match, then the column is used without transformation.

> **See:**
>
> *Oracle Database SQL Language Reference* for information about the REPLACE function

2. Because of the include and exclude list parameters, the XFORM_EXPR_STR and XFORM_EXPR_NUM procedures allow you to easily specify individual columns for transformation within large data sets. The other XFORM_* procedures support an exclude list only. In these procedures, you must enumerate every column that you do not want to transform.
3. See "Operational Notes"

Examples

This example creates a view that transforms character columns to upper case.

```
describe customers
 Name                                   Null?    Type
 ----------------------------------- -------- ------------------------
 CUST_ID                             NOT NULL NUMBER
 CUST_MARITAL_STATUS                          VARCHAR2(20)
 OCCUPATION                                   VARCHAR2(21)
 AGE                                          NUMBER
 YRS_RESIDENCE                                NUMBER

SELECT cust_id,  cust_marital_status, occupation FROM customers
    WHERE   cust_id > 102995
    ORDER BY cust_id desc;

CUST_ID CUST_MARITAL_STATUS  OCCUPATION
------- -------------------- ---------------------
 103000 Divorc.              Cleric.
 102999 Married              Cleric.
 102998 Married              Exec.
 102997 Married              Exec.
 102996 NeverM               Other

BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_EXPR_STR(
       expr_pattern            => 'upper(:col)',
       data_table_name         => 'customers',
       xform_view_name         => 'cust_upcase_view');
END;
/
describe cust_upcase_view
 Name                            Null?    Type
 ----------------------------- -------- --------------------
 CUST_ID                       NOT NULL NUMBER
 CUST_MARITAL_STATUS                    VARCHAR2(20)
 OCCUPATION                             VARCHAR2(21)
 AGE                                    NUMBER
 YRS_RESIDENCE                          NUMBER

SELECT cust_id,  cust_marital_status, occupation FROM cust_upcase_view
   WHERE   cust_id > 102995
   ORDER BY cust_id desc;

CUST_ID CUST_MARITAL_STATUS  OCCUPATION
```

```
------- -------------------- --------------------
 103000 DIVORC.              CLERIC.
 102999 MARRIED              CLERIC.
 102998 MARRIED              EXEC.
 102997 MARRIED              EXEC.
 102996 NEVERM               OTHER
```

XFORM_MISS_CAT Procedure

This procedure creates a view that implements the categorical missing value treatment transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_MISS_CAT (
    miss_table_name       IN VARCHAR2,
    data_table_name       IN VARCHAR2,
    xform_view_name       IN VARCHAR2,
    miss_schema_name      IN VARCHAR2 DEFAULT NULL,
    data_schema_name      IN VARCHAR2 DEFAULT NULL,
    xform_schema_name     IN VARCHAR2 DEFAULT NULL;
```

Parameters

Table 6-173 XFORM_MISS_CAT Procedure Parameters

Parameter	Description
`miss_table_name`	Name of the transformation definition table for categorical missing value treatment. You can use the CREATE_MISS_CAT Procedure to create the definition table. The table must be populated with transformation definitions before you call `XFORM_MISS_CAT`. To populate the table, you can use the INSERT_MISS_CAT_MODE Procedure or you can write your own SQL. See Table 6-138.
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view presents columns in *`data_table_name`* with the transformations specified in *`miss_table_name`*.
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that replaces missing categorical values with the mode.

```
SELECT * FROM geog;

REG_ID REGION
```

```
------ ------------------------------
     1 NE
     2 SW
     3 SE
     4 SW
     5
     6 NE
     7 NW
     8 NW
     9
    10
    11 SE
    12 SE
    13 NW
    14 SE
    15 SE

SELECT STATS_MODE(region) FROM geog;

STATS_MODE(REGION)
------------------------------
SE

BEGIN
  DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_CAT('misscat_xtbl');
  DBMS_DATA_MINING_TRANSFORM.INSERT_MISS_CAT_MODE (
    miss_table_name         => 'misscat_xtbl',
    data_table_name         => 'geog' );
END;
/

SELECT col, val FROM misscat_xtbl;

COL        VAL
---------- ----------
REGION     SE

BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_MISS_CAT (
     miss_table_name        => 'misscat_xtbl',
     data_table_name        => 'geog',
     xform_view_name        => 'geogxf_view');
END;
/

SELECT * FROM geogxf_view;

REG_ID REGION
------ ------------------------------
     1 NE
     2 SW
     3 SE
     4 SW
     5 SE
     6 NE
     7 NW
     8 NW
     9 SE
    10 SE
    11 SE
    12 SE
    13 NW
    14 SE
    15 SE
```

XFORM_MISS_NUM Procedure

This procedure creates a view that implements the numerical missing value treatment transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_MISS_NUM (
    miss_table_name       IN VARCHAR2,
    data_table_name       IN VARCHAR2,
    xform_view_name       IN VARCHAR2,
    miss_schema_name      IN VARCHAR2 DEFAULT NULL,
    data_schema_name      IN VARCHAR2 DEFAULT NULL,
    xform_schema_name     IN VARCHAR2 DEFAULT NULL;
```

Parameters

Table 6-174 XFORM_MISS_NUM Procedure Parameters

Parameter	Description
`miss_table_name`	Name of the transformation definition table for numerical missing value treatment. You can use the CREATE_MISS_NUM Procedure to create the definition table. The table must be populated with transformation definitions before you call `XFORM_MISS_NUM`. To populate the table, you can use the INSERT_MISS_NUM_MEAN Procedure or you can write your own SQL. See Table 6-140.
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view presents columns in *`data_table_name`* with the transformations specified in *`miss_table_name`*.
`miss_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that replaces missing numerical values with the mean.

```
SELECT * FROM items;

ITEM_ID       QTY
---------- ------
aa            200
```

```
bb            200
cc            250
dd
ee
ff            100
gg            250
hh            200
ii
jj            200

SELECT AVG(qty) FROM items;

AVG(QTY)
--------
     200

BEGIN
  DBMS_DATA_MINING_TRANSFORM.CREATE_MISS_NUM('missnum_xtbl');
  DBMS_DATA_MINING_TRANSFORM.INSERT_MISS_NUM_MEAN (
     miss_table_name          => 'missnum_xtbl',
     data_table_name         => 'items' );
END;
/

SELECT col, val FROM missnum_xtbl;

COL           VAL
---------- ------
QTY           200

BEGIN
    DBMS_DATA_MINING_TRANSFORM.XFORM_MISS_NUM (
        miss_table_name        => 'missnum_xtbl',
        data_table_name        => 'items',
        xform_view_name        => 'items_view');
END;
/

SELECT * FROM items_view;

ITEM_ID       QTY
---------- ------
aa            200
bb            200
cc            250
dd            200
ee            200
ff            100
gg            250
hh            200
ii            200
jj            200
```

XFORM_NORM_LIN Procedure

This procedure creates a view that implements the linear normalization transformations specified in a definition table. Only the columns that are specified in the definition table are transformed; the remaining columns from the data table are present in the view, but they are not changed.

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_NORM_LIN (
     norm_table_name       IN VARCHAR2,
     data_table_name       IN VARCHAR2,
     xform_view_name       IN VARCHAR2,
     norm_schema_name      IN VARCHAR2 DEFAULT NULL,
     data_schema_name      IN VARCHAR2 DEFAULT NULL,
     xform_schema_name     IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-175 XFORM_NORM_LIN Procedure Parameters

Parameter	Description
`norm_table_name`	Name of the transformation definition table for linear normalization. You can use the CREATE_NORM_LIN Procedure to create the definition table. The table must be populated with transformation definitions before you call `XFORM_NORM_LIN`. To populate the table, you can use one of the `INSERT` procedures for normalization or you can write your own SQL. See Table 6-138.
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view presents columns in *`data_table_name`* with the transformations specified in *`miss_table_name`*.
`norm_schema_name`	Schema of *`miss_table_name`*. If no schema is specified, the current schema is used.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes".

Examples

This example creates a view that normalizes the `cust_year_of_birth` and `cust_credit_limit` columns. The data source consists of three columns from `sh.customer`.

```
CREATE OR REPLACE VIEW mining_data AS
     SELECT cust_id, cust_year_of_birth, cust_credit_limit
     FROM sh.customers;

describe mining_data
 Name                                      Null?    Type
 ----------------------------------------- -------- --------------------------
 CUST_ID                                   NOT NULL NUMBER
 CUST_YEAR_OF_BIRTH                        NOT NULL NUMBER(4)
 CUST_CREDIT_LIMIT                                  NUMBER

SELECT * FROM mining_data WHERE cust_id > 104495
      ORDER BY cust_year_of_birth;
 CUST_ID CUST_YEAR_OF_BIRTH CUST_CREDIT_LIMIT
-------- ------------------ -----------------
  104496               1947              3000
  104498               1954             10000
  104500               1962             15000
  104499               1970              3000
  104497               1976              3000

BEGIN
  dbms_data_mining_transform.CREATE_NORM_LIN(
       norm_table_name        => 'normx_tbl');
 dbms_data_mining_transform.INSERT_NORM_LIN_MINMAX(
      norm_table_name     => 'normx_tbl',
      data_table_name     => 'mining_data',
      exclude_list     => dbms_data_mining_transform.COLUMN_LIST( 'cust_id'),
      round_num      => 3);
END;
/

SELECT col, shift, scale FROM normx_tbl;

COL                                SHIFT     SCALE
------------------------------ --------- ---------
CUST_YEAR_OF_BIRTH                  1910        77
CUST_CREDIT_LIMIT                   1500     13500

BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_NORM_LIN (
     norm_table_name     => 'normx_tbl',
     data_table_name     => 'mining_data',
     xform_view_name     => 'norm_view');
END;
/

SELECT * FROM norm_view WHERE cust_id > 104495
      ORDER BY cust_year_of_birth;

 CUST_ID CUST_YEAR_OF_BIRTH CUST_CREDIT_LIMIT
-------- ------------------ -----------------
  104496           .4805195          .1111111
  104498           .5714286          .6296296
  104500           .6753247                 1
  104499           .7792208          .1111111
  104497           .8571429          .1111111


set long 2000
SQL> SELECT text FROM user_views WHERE view_name IN 'NORM_VIEW';

TEXT
---------------------------------------------------------------------------
SELECT "CUST_ID",("CUST_YEAR_OF_BIRTH"-1910)/77 "CUST_YEAR_OF_BIRTH",("CUST
_CREDIT_LIMIT"-1500)/13500 "CUST_CREDIT_LIMIT" FROM mining_data
```

XFORM_STACK Procedure

This procedure creates a view that implements the transformations specified by the stack. Only the columns and nested attributes that are specified in the stack are transformed. Any remaining columns and nested attributes from the data table appear in the view without changes.

To create a list of objects that describe the transformed columns, use the DESCRIBE_STACK Procedure.

See Also:

"Overview"

Oracle Machine Learning for SQL User's Guide for more information about machine learning attributes

Syntax

```
DBMS_DATA_MINING_TRANSFORM.XFORM_STACK (
     xform_list          IN     TRANSFORM_list,
     data_table_name     IN     VARCHAR2,
     xform_view_name     IN     VARCHAR2,
     data_schema_name    IN     VARCHAR2 DEFAULT NULL,
     xform_schema_name   IN     VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-176 XFORM_STACK Procedure Parameters

Parameter	Description
`xform_list`	The transformation list. See Table 6-127 for a description of the `TRANSFORM_LIST` object type.
`data_table_name`	Name of the table containing the data to be transformed
`xform_view_name`	Name of the view to be created. The view applies the transformations in *`xform_list`* to *`data_table_name`*.
`data_schema_name`	Schema of *`data_table_name`*. If no schema is specified, the current schema is used.
`xform_schema_name`	Schema of *`xform_view_name`*. If no schema is specified, the current schema is used.

Usage Notes

See "Operational Notes". The following sections are especially relevant:

- "About Transformation Lists"
- "About Stacking"
- "Nested Data Transformations"

Examples

This example applies a transformation list to the view `oml_user.cust_info` and shows how the data is transformed. The `CREATE` statement for `cust_info` is shown in "DESCRIBE_STACK Procedure".

```
BEGIN
  dbms_data_mining_transform.CREATE_BIN_NUM ('birth_yr_bins');
  dbms_data_mining_transform.INSERT_BIN_NUM_QTILE (
        bin_table_name   => 'birth_yr_bins',
        data_table_name  =>  'cust_info',
        bin_num          =>   6,
        exclude_list     =>  dbms_data_mining_transform.column_list(
                                 'cust_id','country_id'));
END;
/
SELECT * FROM birth_yr_bins;

COL                  ATT      VAL BIN
-------------------- ----- ------ ----------
CUST_YEAR_OF_BIRTH           1922
CUST_YEAR_OF_BIRTH           1951 1
CUST_YEAR_OF_BIRTH           1959 2
CUST_YEAR_OF_BIRTH           1966 3
CUST_YEAR_OF_BIRTH           1973 4
CUST_YEAR_OF_BIRTH           1979 5
CUST_YEAR_OF_BIRTH           1986 6

DECLARE
      cust_stack   dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
      dbms_data_mining_transform.SET_TRANSFORM (cust_stack,
          'country_id', NULL, 'country_id/10', 'country_id*10');
      dbms_data_mining_transform.STACK_BIN_NUM ('birth_yr_bins',
          cust_stack);
      dbms_data_mining_transform.SET_TRANSFORM (cust_stack,
          'custprods', 'Mouse Pad', 'value*100', 'value/100');
      dbms_data_mining_transform.XFORM_STACK(
          xform_list        => cust_stack,
          data_table_name   => 'cust_info',
          xform_view_name   =>  'cust_xform_view');
  END;
/

-- Two rows of data without transformations
SELECT * from cust_info WHERE cust_id BETWEEN 100010 AND 100011;

CUST_ID COUNTRY_ID CUST_YEAR_OF_BIRTH CUSTPRODS(ATTRIBUTE_NAME, VALUE)
------- ---------- ------------------ -----------------------------------------
 100010      52790               1975 DM_NESTED_NUMERICALS(
                                        DM_NESTED_NUMERICAL(
                                         '18" Flat Panel Graphics Monitor', 1),
                                        DM_NESTED_NUMERICAL(
                                         'SIMM- 16MB PCMCIAII card', 1))
 100011      52775               1972 DM_NESTED_NUMERICALS(
                                       DM_NESTED_NUMERICAL(
                                         'External 8X CD-ROM', 1),
                                       DM_NESTED_NUMERICAL(
                                         'Mouse Pad', 1),
                                       DM_NESTED_NUMERICAL(
                                        'SIMM- 16MB PCMCIAII card', 1),
                                       DM_NESTED_NUMERICAL(
                                         'Keyboard Wrist Rest', 1),
                                       DM_NESTED_NUMERICAL(
                                         '18" Flat Panel Graphics Monitor', 1),
                                       DM_NESTED_NUMERICAL(
                                         'O/S Documentation Set - English', 1))
```

```
-- Same two rows of data with transformations
SELECT * FROM cust_xform_view WHERE cust_id BETWEEN 100010 AND 100011;

CUST_ID  COUNTRY_ID  C  CUSTPRODS(ATTRIBUTE_NAME, VALUE)
-------  ----------  -  --------------------------------------------------------
 100010       5279   5  DM_NESTED_NUMERICALS(
                          DM_NESTED_NUMERICAL(
                           '18" Flat Panel Graphics Monitor', 1),
                          DM_NESTED_NUMERICAL(
                            'SIMM- 16MB PCMCIAII card', 1))
 100011     5277.5   4  DM_NESTED_NUMERICALS(
                          DM_NESTED_NUMERICAL(
                            'External 8X CD-ROM', 1),
                          DM_NESTED_NUMERICAL(
                            'Mouse Pad', 100),
                          DM_NESTED_NUMERICAL(
                            'SIMM- 16MB PCMCIAII card', 1),
                          DM_NESTED_NUMERICAL(
                            'Keyboard Wrist Rest', 1),
                          DM_NESTED_NUMERICAL(
                            '18" Flat Panel Graphics Monitor', 1),
                          DM_NESTED_NUMERICAL(
                            'O/S Documentation Set - English', 1))
```

DBMS_PREDICTIVE_ANALYTICS

Machine learning can discover useful information buried in vast amounts of data. However, both the programming interfaces and the machine learning expertise required to obtain these results are too complex for use by the wide audiences that can obtain benefits from using Oracle Machine Learning for SQL.

The `DBMS_PREDICTIVE_ANALYTICS` package addresses both of these complexities by automating the entire machine learning process from data preprocessing through model building to scoring new data. This package provides an important tool that makes machine learning possible for a broad audience of users, in particular, business analysts.

This chapter contains the following topics:

- Overview
- Security Model
- Summary of DBMS_PREDICTIVE_ANALYTICS Subprograms

Using DBMS_PREDICTIVE_ANALYTICS

This section contains topics that relate to using the `DBMS_PREDICTIVE_ANALYTICS` package.

- Overview
- Security Model

DBMS_PREDICTIVE_ANALYTICS Overview

DBMS_PREDICTIVE_ANALYTICS automates parts of the machine learning process.

Machine learning, according to a commonly used process model, requires the following steps:

1. Understand the business problem.
2. Understand the data.

3. Prepare the data for mining.
4. Create models using the prepared data.
5. Evaluate the models.
6. Deploy and use the model to score new data.

DBMS_PREDICTIVE_ANALYTICS automates parts of step 3 — 5 of this process.

Predictive analytics procedures analyze and prepare the input data, create and test machine learning models using the input data, and then use the input data for scoring. The results of scoring are returned to the user. The models and supporting objects are not preserved after the operation completes.

DBMS_PREDICTIVE_ANALYTICS Security Model

The DBMS_PREDICTIVE_ANALYTICS package is owned by user SYS and is installed as part of database installation. Execution privilege on the package is granted to public. The routines in the package are run with invokers' rights (run with the privileges of the current user).

The DBMS_PREDICTIVE_ANALYTICS package exposes APIs which are leveraged by the Oracle Machine Learning for SQL option. Users who wish to invoke procedures in this package require the CREATE MINING MODEL system privilege (as well as the CREATE TABLE and CREATE VIEW system privilege).

Summary of DBMS_PREDICTIVE_ANALYTICS Subprograms

This table lists and briefly describes the DBMS_PREDICTIVE_ANALYTICS package subprograms.

Table 6-177 DBMS_PREDICTIVE_ANALYTICS Package Subprograms

Subprogram	Purpose
EXPLAIN Procedure	Ranks attributes in order of influence in explaining a target column.
PREDICT Procedure	Predicts the value of a target column based on values in the input data.
PROFILE Procedure	Generates rules that identify the records that have the same target value.

EXPLAIN Procedure

The EXPLAIN procedure identifies the attributes that are important in explaining the variation in values of a target column.

The input data must contain some records where the target value is known (not NULL). These records are used by the procedure to train a model that calculates the attribute importance.

Note:

EXPLAIN supports DATE and TIMESTAMP datatypes in addition to the numeric, character, and nested datatypes supported by Oracle Machine Learning for SQL models.

Data requirements for Oracle Machine Learning for SQL are described in *Oracle Machine Learning for SQL User's Guide*

The EXPLAIN procedure creates a result table that lists the attributes in order of their explanatory power. The result table is described in the Usage Notes.

Syntax

```
DBMS_PREDICTIVE_ANALYTICS.EXPLAIN (
     data_table_name     IN VARCHAR2,
     explain_column_name IN VARCHAR2,
     result_table_name   IN VARCHAR2,
     data_schema_name    IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-178   EXPLAIN Procedure Parameters

Parameter	Description
data_table_name	Name of input table or view
explain_column_name	Name of the column to be explained
result_table_name	Name of the table where results are saved
data_schema_name	Name of the schema where the input table or view resides and where the result table is created. Default: the current schema.

Usage Notes

The EXPLAIN procedure creates a result table with the columns described in Table 6-179.

Table 6-179   EXPLAIN Procedure Result Table

Column Name	Datatype	Description
ATTRIBUTE_NAME	VARCHAR2(30)	Name of a column in the input data; all columns except the explained column are listed in the result table.
EXPLANATORY_VALUE	NUMBER	Value indicating how useful the column is for determining the value of the explained column. Higher values indicate greater explanatory power. Value can range from 0 to 1. An individual column's explanatory value is independent of other columns in the input table. The values are based on how strong each individual column correlates with the explained column. The value is affected by the number of records in the input table, and the relations of the values of the column to the values of the explain column. An explanatory power value of 0 implies there is no useful correlation between the column's values and the explain column's values. An explanatory power of 1 implies perfect correlation; such columns should be eliminated from consideration for PREDICT. In practice, an explanatory power equal to 1 is rarely returned.
RANK	NUMBER	Ranking of explanatory power. Rows with equal values for explanatory_value have the same rank. Rank values are not skipped in the event of ties.

Example

The following example performs an EXPLAIN operation on the SUPPLEMENTARY_DEMOGRAPHICS table of Sales History.

```
--Perform EXPLAIN operation
BEGIN
```

```
    DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
        data_table_name      => 'supplementary_demographics',
        explain_column_name  => 'home_theater_package',
        result_table_name    => 'demographics_explain_result');
END;
/
--Display results
SELECT * FROM demographics_explain_result;

ATTRIBUTE_NAME                           EXPLANATORY_VALUE       RANK
---------------------------------------- ----------------- ----------
Y_BOX_GAMES                                     .524311073          1
YRS_RESIDENCE                                   .495987246          2
HOUSEHOLD_SIZE                                  .146208506          3
AFFINITY_CARD                                     .0598227          4
EDUCATION                                       .018462703          5
OCCUPATION                                      .009721543          6
FLAT_PANEL_MONITOR                               .00013733          7
PRINTER_SUPPLIES                                         0          8
OS_DOC_SET_KANJI                                         0          8
BULK_PACK_DISKETTES                                      0          8
BOOKKEEPING_APPLICATION                                  0          8
COMMENTS                                                 0          8
CUST_ID                                                  0          8
```

The results show that `Y_BOX_GAMES`, `YRS_RESIDENCE`, and `HOUSEHOLD_SIZE` are the best predictors of `HOME_THEATER_PACKAGE`.

PREDICT Procedure

The `PREDICT` procedure predicts the values of a target column.

The input data must contain some records where the target value is known (not `NULL`). These records are used by the procedure to train and test a model that makes the predictions.

> **Note:**
>
> `PREDICT` supports `DATE` and `TIMESTAMP` datatypes in addition to the numeric, character, and nested datatypes supported by Oracle Machine Learning for SQL models.
>
> Data requirements for Oracle Machine Learning for SQL are described in *Oracle Machine Learning for SQL User's Guide*

The `PREDICT` procedure creates a result table that contains a predicted target value for every record. The result table is described in the Usage Notes.

Syntax

```
DBMS_PREDICTIVE_ANALYTICS.PREDICT (
    accuracy                  OUT NUMBER,
    data_table_name           IN VARCHAR2,
    case_id_column_name       IN VARCHAR2,
    target_column_name        IN VARCHAR2,
    result_table_name         IN VARCHAR2,
    data_schema_name          IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-180 PREDICT Procedure Parameters

Parameter	Description
`accuracy`	Output parameter that returns the predictive confidence, a measure of the accuracy of the predicted values. The predictive confidence for a categorical target is the most common target value; the predictive confidence for a numerical target is the mean.
`data_table_name`	Name of the input table or view.
`case_id_column_name`	Name of the column that uniquely identifies each case (record) in the input data.
`target_column_name`	Name of the column to predict.
`result_table_name`	Name of the table where results will be saved.
`data_schema_name`	Name of the schema where the input table or view resides and where the result table is created. Default: the current schema.

Usage Notes

The `PREDICT` procedure creates a result table with the columns described in Table 6-181.

Table 6-181 PREDICT Procedure Result Table

Column Name	Datatype	Description
Case ID column name	`VARCHAR2` or `NUMBER`	The name of the case ID column in the input data.
`PREDICTION`	`VARCHAR2` or `NUMBER`	The predicted value of the target column for the given case.
`PROBABILITY`	`NUMBER`	For classification (categorical target), the probability of the prediction. For regression problems (numerical target), this column contains `NULL`.

> **Note:**
>
> Make sure that the name of the case ID column is not '`PREDICTION`' or '`PROBABILITY`'.

Predictions are returned for all cases whether or not they contained target values in the input.

Predicted values for known cases may be interesting in some situations. For example, you could perform deviation analysis to compare predicted values and actual values.

Example

The following example performs a `PREDICT` operation and displays the first 10 predictions. The results show an accuracy of 79% in predicting whether each customer has an affinity card.

```
--Perform PREDICT operation
DECLARE
    v_accuracy NUMBER(10,9);
BEGIN
```

```
    DBMS_PREDICTIVE_ANALYTICS.PREDICT(
        accuracy             => v_accuracy,
        data_table_name      => 'supplementary_demographics',
        case_id_column_name  => 'cust_id',
        target_column_name   => 'affinity_card',
        result_table_name    => 'pa_demographics_predict_result');
    DBMS_OUTPUT.PUT_LINE('Accuracy = ' || v_accuracy);
END;
/

Accuracy = .788696903

--Display results
SELECT * FROM pa_demographics_predict_result WHERE rownum < 10;

   CUST_ID PREDICTION PROBABILITY
---------- ---------- -----------
    101501          1  .834069848
    101502          0  .991269965
    101503          0   .99978311
    101504          1  .971643388
    101505          1  .541754127
    101506          0  .803719133
    101507          0  .999999303
    101508          0  .999999987
    101509          0  .999953074
```

PROFILE Procedure

The `PROFILE` procedure generates rules that describe the cases (records) from the input data.

For example, if a target column `CHURN` has values 'Yes' and 'No', `PROFILE` generates a set of rules describing the expected outcomes. Each profile includes a rule, record count, and a score distribution.

The input data must contain some cases where the target value is known (not `NULL`). These cases are used by the procedure to build a model that calculates the rules.

Note:

`PROFILE` does not support nested types or dates.

Data requirements for Oracle Machine Learning for SQL are described in *Oracle Machine Learning for SQL User’s Guide*

The `PROFILE` procedure creates a result table that specifies rules (profiles) and their corresponding target values. The result table is described in the Usage Notes.

Syntax

```
DBMS_PREDICTIVE_ANALYTICS.PROFILE (
     data_table_name          IN VARCHAR2,
     target_column_name       IN VARCHAR2,
     result_table_name        IN VARCHAR2,
     data_schema_name         IN VARCHAR2 DEFAULT NULL);
```

Parameters

Table 6-182 PROFILE Procedure Parameters

Parameter	Description
`data_table_name`	Name of the table containing the data to be analyzed.
`target_column_name`	Name of the target column.
`result_table_name`	Name of the table where the results will be saved.
`data_schema_name`	Name of the schema where the input table or view resides and where the result table is created. Default: the current schema.

Usage Notes

The `PROFILE` procedure creates a result table with the columns described in Table 6-183.

Table 6-183 PROFILE Procedure Result Table

Column Name	Datatype	Description
`PROFILE_ID`	`NUMBER`	A unique identifier for this profile (rule).
`RECORD_COUNT`	`NUMBER`	The number of records described by the profile.
`DESCRIPTION`	`SYS.XMLTYPE`	The profile rule. See "XML Schema for Profile Rules".

XML Schema for Profile Rules

The `DESCRIPTION` column of the result table contains XML that conforms to the following XSD:

```
<xs:element name="SimpleRule">
  <xs:complexType>
    <xs:sequence>
      <xs:group ref="PREDICATE"/>
      <xs:element ref="ScoreDistribution" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="id" type="xs:string" use="optional"/>
    <xs:attribute name="score" type="xs:string" use="required"/>
    <xs:attribute name="recordCount" type="NUMBER" use="optional"/>
  </xs:complexType>
</xs:element>
```

Example

This example generates a rule describing customers who are likely to use an affinity card (target value is 1) and a set of rules describing customers who are not likely to use an affinity card (target value is 0). The rules are based on only two predictors: education and occupation.

```
SET serveroutput ON
SET trimspool ON
SET pages 10000
SET long 10000
SET pagesize 10000
SET linesize 150
CREATE VIEW cust_edu_occ_view AS
              SELECT cust_id, education, occupation, affinity_card
              FROM sh.supplementary_demographics;
BEGIN
    DBMS_PREDICTIVE_ANALYTICS.PROFILE(
         DATA_TABLE_NAME    => 'cust_edu_occ_view',
         TARGET_COLUMN_NAME => 'affinity_card',
         RESULT_TABLE_NAME  => 'profile_result');
END;
/
```

This example generates eight rules in the result table `profile_result`. Seven of the rules suggest a target value of 0; one rule suggests a target value of 1. The `score` attribute on a rule identifies the target value.

This `SELECT` statement returns all the rules in the result table.

```
SELECT a.profile_id, a.record_count, a.description.getstringval()
 FROM profile_result a;
```

This `SELECT` statement returns the rules for a target value of 0.

```
SELECT *
  FROM profile_result t
  WHERE extractvalue(t.description, '/SimpleRule/@score') = 0;
```

The eight rules generated by this example are displayed as follows.

```
<SimpleRule id="1" score="0" recordCount="443">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Armed-F" "Exec." "Prof." "Protec."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"< Bach." "Assoc-V" "HS-grad"
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="297" />
  <ScoreDistribution value="1" recordCount="146" />
</SimpleRule>

<SimpleRule id="2" score="0" recordCount="18">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Armed-F" "Exec." "Prof." "Protec."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"10th" "11th" "12th" "1st-4th" "5th-6th" "7th-8th" "9th" "Presch."
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="18" />
</SimpleRule>

<SimpleRule id="3" score="0" recordCount="458">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Armed-F" "Exec." "Prof." "Protec."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"Assoc-A" "Bach."
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="248" />
  <ScoreDistribution value="1" recordCount="210" />
</SimpleRule>

<SimpleRule id="4" score="1" recordCount="276">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Armed-F" "Exec." "Prof." "Protec."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"Masters" "PhD" "Profsc"
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="1" recordCount="183" />
  <ScoreDistribution value="0" recordCount="93" />
</SimpleRule>

<SimpleRule id="5" score="0" recordCount="307">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"Assoc-A" "Bach." "Masters" "PhD" "Profsc"
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Crafts" "Sales" "TechSup" "Transp."
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="184" />
  <ScoreDistribution value="1" recordCount="123" />
</SimpleRule>

<SimpleRule id="6" score="0" recordCount="243">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">"Assoc-A" "Bach." "Masters" "PhD" "Profsc"
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"?" "Cleric." "Farming" "Handler" "House-s" "Machine" "Other"
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="197" />
  <ScoreDistribution value="1" recordCount="46" />
</SimpleRule>

<SimpleRule id="7" score="0" recordCount="2158">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">
        "10th" "11th" "12th" "1st-4th" "5th-6th" "7th-8th" "9th" "< Bach." "Assoc-V" "HS-grad"
        "Presch."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"?" "Cleric." "Crafts" "Farming" "Machine" "Sales" "TechSup" " Transp."
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
  <ScoreDistribution value="0" recordCount="1819"/>
  <ScoreDistribution value="1" recordCount="339"/>
</SimpleRule>

<SimpleRule id="8" score="0" recordCount="597">
  <CompoundPredicate booleanOperator="and">
    <SimpleSetPredicate field="EDUCATION" booleanOperator="isIn">
      <Array type="string">
        "10th" "11th" "12th" "1st-4th"  "5th-6th" "7th-8th" "9th" "< Bach." "Assoc-V" "HS-grad"
        "Presch."
      </Array>
    </SimpleSetPredicate>
    <SimpleSetPredicate field="OCCUPATION" booleanOperator="isIn">
      <Array type="string">"Handler" "House-s" "Other"
      </Array>
    </SimpleSetPredicate>
  </CompoundPredicate>
<ScoreDistribution value="0" recordCount="572"/>
<ScoreDistribution value="1" recordCount="25"/>
</SimpleRule>
```

Model Detail Views

Model detail views are algorithm-specific. Viewing the model detail views will provide you with additional information about the model you created. The names of model detail views begin with DM$. Some model views, such as Global Name-Value Pairs view (`DM$VG`*model_name*), Computed Settings view (`DM$VS`*model_name*), Model Build Alerts view (`DM$VW`*model_name*), and Normalization and Missing Value Handling view (`DM$VN`*model_name*), are shared by all algorithms and are documented separately. Aside from that, classification, clustering, and regression algorithms share some common views. The columns returned by these views may differ between algorithms.

The following are the model views, grouped by model function:

Association:

- Model Detail Views for Association Rules
- Model Detail View for Frequent Itemsets
- Model Detail Views for Transactional Itemsets
- Model Detail View for Transactional Rule

Classification, Regression, and Anomaly Detection:

- Model Detail Views for Classification Algorithms
- Model Detail Views for CUR Matrix Decomposition
- Model Detail Views for Decision Tree
- Model Detail Views for Generalized Linear Model
- Model Detail View for Multivariate State Estimation Technique - Sequential Probability Ratio Test
- Model Detail Views for Naive Bayes
- Model Detail Views for Neural Network
- Model Detail Views for Random Forest
- Model Detail View for Support Vector Machine

- Model Detail Views for XGBoost

Clustering:

- Model Detail Views for Clustering Algorithms
- Model Detail Views for Expectation Maximization
- Model Detail Views for *k*-Means
- Model Detail Views for O-Cluster

Feature Extraction:

- Model Detail Views for Explicit Semantic Analysis
- Model Detail Views for Non-Negative Matrix Factorization
- Model Detail Views for Singular Value Decomposition

Feature Selection:

- Model Detail Views for Minimum Description Length

Data Preparation and Other:

- Model Detail Views for Binning
- Model Detail Views for Global Information
- Model Detail Views for Normalization and Missing Value Handling

Time Series:

Model Detail Views for Exponential Smoothing

ONNX Models:

Model Detail Views for ONNX Models

Model Detail Views for Association Rules

The model detail view `DM$VR`*model_name* contains the generated rules for association models.

These are the available model views for Association Rules:

Model Views	Description
`DM$VA`*`model_name`*	Association Rules For Transactional Data
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VI`*model_name*:	Association Rule Itemsets
`DM$VR`*model_name*	Association Rules
`DM$VS`*model_name*	Computed Settings
`DM$VT`*model_name*	Association Rule Itemsets For Transactional Data
`DM$VW`*model_name*	Model Build Alerts

Depending on the settings of the model, this rule view (`DM$VR`*model_name*) different sets of columns. Settings `ODMS_ITEM_ID_COLUMN_NAME` and `ODMS_ITEM_VALUE_COLUMN_NAME` determine how each item is defined. If `ODMS_ITEM_ID_COLUMN_NAME` is set, the input format is called transactional input, otherwise, the input format is called 2-Dimensional input. With transactional input, if setting `ODMS_ITEM_VALUE_COLUMN_NAME` is not set, each item is defined by `ITEM_NAME`, otherwise, each item is defined by `ITEM_NAME` and `ITEM_VALUE`. With 2-Dimensional input, each item is defined by ITEM_NAME, ITEM_SUBNAME and ITEM_VALUE. Setting ASSO_AGGREGATES specifies the columns to aggregate, which is displayed in the view.

> **Note:**
>
> Setting ASSO_AGGREGATES is not allowed for 2-dimensional input.

The following shows the views with different settings.

Transactional Input Without ASSO_AGGREGATES Setting

When you sett ITEM_NAME (ODMS_ITEM_ID_COLUMN_NAME) and do not set ITEM_VALUE (ODMS_ITEM_VALUE_COLUMN_NAME), the view contains the following. The consequent item is defined with only the name field. If you also set ITEM_VALUE, the view has the additional column CONSEQUENT_VALUE that specifies the value field.

```
Name                                      Type
 ----------------------------------------- ----------------------------
 PARTITION_NAME                            VARCHAR2(128)
 RULE_ID                                   NUMBER
 RULE_SUPPORT                              NUMBER
 RULE_CONFIDENCE                           NUMBER
 RULE_LIFT                                 NUMBER
 RULE_REVCONFIDENCE                        NUMBER
 ANTECEDENT_SUPPORT                        NUMBER
 NUMBER_OF_ITEMS                           NUMBER
 CONSEQUENT_SUPPORT                        NUMBER
 CONSEQUENT_NAME                           VARCHAR2(4000)
 ANTECEDENT                                SYS.XMLTYPE
```

Table 6-184 Rule View Columns for Transactional Inputs

Column Name	Description
PARTITION_NAME	A partition in a partitioned model to retrieve details.
RULE_ID	The identifier of the rule.
RULE_SUPPORT	The number of transactions that satisfy the rule.
RULE_CONFIDENCE	The likelihood of a transaction satisfying the rule.
RULE_LIFT	The degree of improvement in the prediction over random chance when the rule is satisfied.
RULE_REVCONFIDENCE	The number of transactions in which the rule occurs divided by the number of transactions in which the consequent occurs.
ANTECEDENT_SUPPORT	The ratio of the number of transactions that satisfy the antecedent to the total number of transactions.
NUMBER_OF_ITEMS	The total number of attributes referenced in the antecedent and consequent of the rule.
CONSEQUENT_SUPPORT	The ratio of the number of transactions that satisfy the consequent to the total number of transactions.
CONSEQUENT_NAME	The name of the consequent.

Table 6-184 (Cont.) Rule View Columns for Transactional Inputs

Column Name	Description
CONSEQUENT_VALUE	The value of the consequent. This column is present when Item_value (ODMS_ITEM_VALUE_COLUMN_NAME) is set with TYPE as numerical or categorical.

ANTECEDENT	The antecedent is described as an itemset. At the itemset level, it specifies the number of aggregates, and if not zero, the names of the columns to be aggregated (as well as the mapping to ASSO_AGG*). The itemset contains >= 1 items.
	• When ODMS_ITEM_VALUE_COLUMN_NAME is not set, each item is defined by item_name. As an example, if the antecedent contains one item B, then it is represented as follows:
	`<itemset NUMAGGR="0"><item><item_name>B</item_name></item></itemset>`
	As another example, if the antecedent contains two items, A and C, then it is represented as follows:
	`<itemset NUMAGGR="0"><item><item_name>A</item_name></item><item><item_name>C</item_name></item></itemset>`
	• When setting ODMS_ITEM_VALUE_COLUMN_NAME is set, each item is defined by item_name and item_value. As an example, if the antecedent contains two items, (name A, value 1) and (name C, value 1), then it is represented as follows:
	`<itemset NUMAGGR="0"><item><item_name>A</item_name><item_value>1</item_value></item><item><item_name>C</item_name><item_value>1</item_value></item></itemset>`

Transactional Input With ASSO_AGGREGATES Setting

Similar to the view without an aggregates setting, there are three cases:

- Rule view when ODMS_ITEM_ID_COLUMN_NAME is set and Item_value (ODMS_ITEM_VALUE_COLUMN_NAME) is not set.
- Rule view when ODMS_ITEM_ID_COLUMN_NAME is set and Item_value (ODMS_ITEM_VALUE_COLUMN_NAME) is set with TYPE as numerical, the view has a CONSEQUENT_VALUE column.
- Rule view when ODMS_ITEM_ID_COLUMN_NAME is set and Item_value (ODMS_ITEM_VALUE_COLUMN_NAME) is set with TYPE as categorical, the view has a CONSEQUENT_VALUE column.

For the example that produces the following rules, see “Example: Calculating Aggregates” in *Oracle Machine Learning for SQL Concepts*.

The view reports two sets of aggregates results:

1. ANT_RULE_PROFIT refers to the total profit for the antecedent itemset with respect to the rule, the profit for each individual item of the antecedent itemset is shown in the ANTECEDENT(XMLtype) column, CON_RULE_PROFIT refers to the total profit for the consequent item with respect to the rule.

In the example, for rule (A, B) => C, the rule itemset (A, B, C) occurs in the transactions of customer 1 and customer 3. The ANT_RULE_PROFIT is $21.20, The ANTECEDENT is shown as follow, which tells that item A has profit 5.00 + 3.00 = $8.00 and item B has profit 3.20 + 10.00 = $13.20, which sum up to ANT_RULE_PROFIT.

```
<itemset NUMAGGR="1" ASSO_AGG0="profit"><item><item_name>A</
item_name><ASSO_AGG0>8.0E+000</ASSO_AGG0></item><item><item_name>B</
item_name><ASSO_AGG0>1.32E+001</ASSO_AGG0></item></itemset>
The CON_RULE_PROFIT is 12.00 + 14.00 = $26.00
```

2. ANT_PROFIT refers to the total profit for the antecedent itemset, while CON_PROFIT refers to the total profit for the consequent item. The difference between CON_PROFIT and CON_RULE_PROFIT (the same applies to ANT_PROFIT and ANT_RULE_PROFIT) is that CON_PROFIT counts all profit for the consequent item across all transactions where the consequent occurs, while CON_RULE_PROFIT only counts across transactions where the rule itemset occurs.

   For example, item C occurs in transactions for customer 1, 2 and 3, CON_PROFIT is 12.00 + 4.20 + 14.00 = $30.20, while CON_RULE_PROFIT only counts transactions for customer 1 and 3 where the rule itemset (A, B, C) occurs.

   Similarly, ANT_PROFIT counts all transactions where itemset (A, B) occurs, while ANT_RULE_PROFIT counts only transactions where the rule itemset (A, B, C) occurs. In this example, by coincidence, both count transactions for customer 1 and 3, and have the same value.

Example 6-26 Examples

The following example shows the view when setting ASSO_AGGREGATES specifies column profit and column sales to be aggregated. In this example, ITEM_VALUE column is not specified.

```
Name                                     Type
 ----------------------------------------- ----------------------------
 PARTITION_NAME                            VARCHAR2(128)
 RULE_ID                                   NUMBER
 RULE_SUPPORT                              NUMBER
 RULE_CONFIDENCE                           NUMBER
 RULE_LIFT                                 NUMBER
 RULE_REVCONFIDENCE                        NUMBER
 ANTECEDENT_SUPPORT                        NUMBER
 NUMBER_OF_ITEMS                           NUMBER
 CONSEQUENT_SUPPORT                        NUMBER
 CONSEQUENT_NAME                           VARCHAR2(4000)
 ANTECEDENT                                SYS.XMLTYPE
 ANT_RULE_PROFIT                           BINARY_DOUBLE
 CON_RULE_PROFIT                           BINARY_DOUBLE
 ANT_PROFIT                                BINARY_DOUBLE
 CON_PROFIT                                BINARY_DOUBLE
 ANT_RULE_SALES                            BINARY_DOUBLE
 CON_RULE_SALES                            BINARY_DOUBLE
 ANT_SALES                                 BINARY_DOUBLE
 CON_SALES                                 BINARY_DOUBLE
```

The rule view has a CONSEQUENT_VALUE column when ODMS_ITEM_ID_COLUMN_NAME is set and Item_value (ODMS_ITEM_VALUE_COLUMN_NAME) is set with TYPE as numerical or categorical.

2-Dimensional Inputs

In Oracle Machine Learning for SQL, association models can be built using either transactional or two-dimensional data formats. For two-dimensional input, each item is defined by three fields: NAME, VALUE and SUBNAME. The NAME field is the name of the column. The VALUE field is the content of the column. The SUBNAME field is used when the input data table contains a nested table. In that case, SUBNAME is the name of the nested table's column. See, Example: Creating a Nested Column for Market Basket Analysis. In this example, there is a nested column. The CONSEQUENT_SUBNAME is the ATTRIBUTE_NAME part of the nested column. That is, 'O/S Documentation Set - English' and CONSEQUENT_VALUE is the value part of the nested column, which is, 1.

The view uses three columns for the consequent. The rule view has the following columns:

```
Name                                      Type
 ----------------------- ---------------------
 PARTITION_NAME                            VARCHAR2(128)
 RULE_ID                                   NUMBER
 RULE_SUPPORT                              NUMBER
 RULE_CONFIDENCE                           NUMBER
 RULE_LIFT                                 NUMBER
 RULE_REVCONFIDENCE                        NUMBER
 ANTECEDENT_SUPPORT                        NUMBER
 NUMBER_OF_ITEMS                           NUMBER
 CONSEQUENT_SUPPORT                        NUMBER
 CONSEQUENT_NAME                           VARCHAR2(4000)
 CONSEQUENT_SUBNAME                        VARCHAR2(4000)
 CONSEQUENT_VALUE                          VARCHAR2(4000)
 ANTECEDENT                                SYS.XMLTYPE
```

> **Note:**
>
> All of the types for three columns for the consequent are VARCHAR2. ASSO_AGGREGATES is not applicable for 2-Dimensional input format.

The following table displays rule view columns for 2-Dimensional input with the descriptions of only the fields that are specific to 2-D inputs.

Table 6-185 Rule View for 2-Dimensional Input

Column Name	Description
CONSEQUENT_SUBNAME	For two-dimensional inputs, CONSEQUENT_SUBNAME is used for nested column in the input data table.
CONSEQUENT_VALUE	The value of the consequent when setting Item_value is set with TYPE as numerical or categorical.

Table 6-185 (Cont.) Rule View for 2-Dimensional Input

Column Name	Description
ANTECEDENT	The antecedent is described as an itemset. The itemset contains >= 1 items. Each item is defined using ITEM_NAME, ITEM_SUBNAME, and ITEM_VALUE: As an example, assuming that this is not a nested table input, and the antecedent contains one item: (name ADDR, value MA). The antecedent (XMLtype) is as follows: `<itemset NUMAGGR="0"><item><item_name>ADDR</item_name><item_subname></item_subname><item_value>MA</item_value></item></itemset>` For 2-Dimensional input with nested table, the subname field is filled.

Global Name-Value Pairs View for Association Rules

Global Name-Value Pairs View produces a single column for an association model. The following table describes the columns returned for association model.

Table 6-186 Global Name-Value Pairs View for an Association Model

Name	Description
ITEMSET_COUNT	The number of itemsets generated.
MAX_SUPPORT	The maximum support.
NUM_ROWS	The total number of rows used in the build.
RULE_COUNT	The number of association rules in the model generated.
TRANSACTION_COUNT	The number of the transactions in the input data.

Model Detail View for Frequent Itemsets

The model detail view DM$VI*model_name* contains information about frequent itemsets.

The Association Rule Itemsets view (DM$VI*model_name*) has the following columns:

```
Name                Type
-------------       ------------------
PARTITION_NAME      VARCHAR2 (128)
ITEMSET_ID          NUMBER
SUPPORT             NUMBER
NUMBER_OF_ITEMS     NUMBER
 ITEMSET             SYS.XMLTYPE
```

Table 6-187 Association Rule Itemsets View

Column Name	Description
PARTITION_NAME	A partition in a partitioned model

Table 6-187 (Cont.) Association Rule Itemsets View

Column Name	Description
`ITEMSET_ID`	Itemset identifier
`SUPPORT`	Support of the itemset
`NUMBER_OF_ITEMS`	Number of items in the itemset
`ITEMSET`	Frequent itemset The structure of the `SYS.XMLTYPE` column itemset is the same as the corresponding Antecedent column of the rule view.

Model Detail Views for Transactional Itemsets

The model detail view `DM$VT`*model_name* contains information about the transactional itemsets.

For the very common case of transactional data without aggregates, the Association Rule Itemsets For Transactional Data view (`DM$VT`*model_name*) provides the itemsets information in transactional format. This view can help improve performance for some queries as compared to the view with the XML column. The transactional itemsets view has the following columns:

```
Name                 Type
-----------------    -----------------
PARTITION_NAME       VARCHAR2(128)
ITEMSET_ID           NUMBER
ITEM_ID              NUMBER
SUPPORT              NUMBER
NUMBER_OF_ITEMS      NUMBER
ITEM_NAME            VARCHAR2(4000)
```

Table 6-188 Association Rule Itemsets For Transactional Data View

Column Name	Description
`PARTITION_NAME`	A partition in a partitioned model
`ITEMSET_ID`	Itemset identifier
`ITEM_ID`	Item identifier
`SUPPORT`	Support of the itemset
`NUMBER_OF_ITEMS`	Number of items in the itemset
`ITEM_NAME`	The name of the item

Model Detail View for Transactional Rule

The model detail view `DM$VA`*model_name* contains information about transactional rules and transactional itemsets.

Transactional data without aggregates also has an Association Rules For Transactional Data view (`DM$VA`*model_name*). This view can improve performance for some queries as compared to the view with the XML column. The transactional rule view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
RULE_ID                                   NUMBER
ANTECEDENT_PREDICATE                      VARCHAR2(4000)
CONSEQUENT_PREDICATE                      VARCHAR2(4000)
RULE_SUPPORT                              NUMBER
RULE_CONFIDENCE                           NUMBER
RULE_LIFT                                 NUMBER
RULE_REVCONFIDENCE                        NUMBER
RULE_ITEMSET_ID                           NUMBER
ANTECEDENT_SUPPORT                        NUMBER
CONSEQUENT_SUPPORT                        NUMBER
NUMBER_OF_ITEMS                           NUMBER
```

Table 6-189 Association Rules For Transactional Data View

Column Name	Description
`PARTITION_NAME`	A partition in a partitioned model
`RULE_ID`	Rule identifier
`ANTECEDENT_PREDICATE`	Name of the Antecedent item.
`CONSEQUENT_PREDICATE`	Name of the Consequent item
`RULE_SUPPORT`	Support of the rule
`RULE_CONFIDENCE`	The likelihood a transaction satisfies the rule when it contains the Antecedent.
`RULE_LIFT`	The degree of improvement in the prediction over random chance when the rule is satisfied
`RULE_REVCONFIDENCE`	The number of transactions in which the rule occurs divided by the number of transactions in which the consequent occurs
`RULE_ITEMSET_ID`	Itemset identifier
`ANTECEDENT_SUPPORT`	The ratio of the number of transactions that satisfy the antecedent to the total number of transactions
`CONSEQUENT_SUPPORT`	The ratio of the number of transactions that satisfy the consequent to the total number of transactions
`NUMBER_OF_ITEMS`	Number of items in the rule

Model Detail Views for Classification Algorithms

Model detail views for classification algorithms are the target map view and scoring cost view, which are applicable to all classification algorithms.

These are the available model views for Classification algorithm:

Model Views	Description
`DM$VA`*model_name*	Variable Importance
`DM$VC`*model_name*	Scoring Cost Matrix
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VS`*model_name*	Computed Settings
`DM$VT`*model_name*	Classification Targets
`DM$VW`*model_name*:	Model Build Alerts

The Classification Targets view (`DM$VT`*model_name*) describes the target distribution for classification models. The view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
TARGET_VALUE                              NUMBER/VARCHAR2
TARGET_COUNT                              NUMBER
TARGET_WEIGHT                             NUMBER
```

Table 6-190   Classification Targets View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`TARGET_VALUE`	Target value, numerical or categorical
`TARGET_COUNT`	Number of rows for a given `TARGET_VALUE`
`TARGET_WEIGHT`	Weight for a given `TARGET_VALUE`

The Scoring Cost Matrix view (`DM$VC`*model_name*) describes the scoring cost matrix for classification models. The view has the following columns:

```
Name                                      Type
----------------------------------------- --------------------------------
PARTITION_NAME                            VARCHAR2(128)
ACTUAL_TARGET_VALUE                       NUMBER/VARCHAR2
PREDICTED_TARGET_VALUE                    NUMBER/VARCHAR2
COST                                      NUMBER
```

Table 6-191   Scoring Cost Matrix View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model

Table 6-191 (Cont.) Scoring Cost Matrix View

Column Name	Description
ACTUAL_TARGET_VALUE	A valid target value
PREDICTED_TARGET_VALUE	Predicted target value
COST	Associated cost for the actual and predicted target value pair

Model Detail Views for CUR Matrix Decomposition

Model detail views for CUR Matrix Decomposition contain information about the scores and ranks of attributes and rows.

CUR Matrix Decomposition models have the following views:

Attribute importance and rank: DM$VC*model_name*

Row importance and rank: DM$VR*model_name*

Global statistics: DM$VG

The attribute importance and rank view DM$VC*model_name* has the following columns:

```
Name                           Type
-----------------              -----------------
PARTITION_NAME                 VARCHAR2(128)
ATTRIBUTE_NAME                 VARCHAR2(128)
ATTRIBUTE_SUBNAME              VARCHAR2(4000)
ATTRIBUTE_VALUE                VARCHAR2(4000)
ATTRIBUTE_IMPORTANCE           NUMBER
ATTRIBUTE_RANK                 NUMBER
```

Table 6-192 Attribute Importance and Rank View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
ATTRIBUTE_NAME	Attribute name
ATTRIBUTE_SUBNAME	Attribute subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Value of the attribute
ATTRIBUTE_IMPORTANCE	Attribute leverage score
ATTRIBUTE_RANK	Attribute rank based on leverage score

The view DM$VR*model_name* exposes the leverage scores and ranks of all selected rows through a view. This view is created when users decide to perform row importance and the CASE_ID column is present. The view has the following columns:

```
Name                       Type
---------------------         ------------------------
PARTITION_NAME                VARCHAR2(128)
CASE_ID                       Original cid data types,
                              including NUMBER, VARCHAR2,
                          DATE, TIMESTAMP,
                          TIMESTAMP WITH TIME ZONE,
                          TIMESTAMP WITH LOCAL TIME ZONE
ROW_IMPORTANCE                          NUMBER
ROW_RANK                                NUMBER
```

Table 6-193 Row Importance and Rank View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
CASE_ID	Case ID. The supported case ID types are the same as that supported for GLM, SVD, and ESA algorithms.
ROW_IMPORTANCE	Row leverage score
ROW_RANK	Row rank based on leverage score

The following table describes global statistics for CUR Matrix Decomposition.

Table 6-194 CUR Matrix Decomposition Statistics Information In Model Global View.

Name	Description
NUM_COMPONENTS	Number of SVD components (SVD rank)
NUM_ROWS	Number of rows used in the model build

Model Detail Views for Decision Tree

The model detail views specific to Decision Tree are the hierarchy view, node statistics view, node description view, and the cost matrix view.

These are the model views available for Decision Tree:

Model Views	Description
DM$VC*model_name*	Scoring Cost Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VI*model_name*	Decision Tree Statistics
DM$VM*model_name*	Decision Tree Build Cost Matrix
DM$VO*model_name*	Decision Tree Nodes
DM$VP*model_name*	Decision Tree Hierarchy
DM$VS*model_name*	Computed Settings
DM$VT*model_name*	Classification Targets
DM$VW*model_name*	Model Build Alerts

The Decision Tree Hierarchy view (DM$VP*model_name*) describes the decision tree hierarchy and the split information for each level in the decision tree. The view has the following columns:

```
Name                                   Type
----------------------------------     ---------------------------
PARTITION_NAME                         VARCHAR2(128)
```

```
PARENT                             NUMBER
SPLIT_TYPE                         VARCHAR2
NODE                               NUMBER
ATTRIBUTE_NAME                     VARCHAR2(128)
ATTRIBUTE_SUBNAME                  VARCHAR2(4000)
OPERATOR                           VARCHAR2
VALUE                              SYS.XMLTYPE
```

Table 6-195 Decision Tree Hierarchy View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`PARENT`	Node ID of the parent
`SPLIT_TYPE`	The main or surrogate split
`NODE`	The node ID
`ATTRIBUTE_NAME`	The attribute used as the splitting criterion at the parent node to produce this node.
`ATTRIBUTE_SUBNAME`	Split attribute subname. The value is null for non-nested columns.
`OPERATOR`	Split operator
`VALUE`	Value used as the splitting criterion. This is an XML element described using the `<Element>` tag. For example, `<Element>Windy</Element><Element>Hot</Element>`.

The Decision Tree Statistics view (`DM$VI`*model_name*) describes the statistics associated with individual tree nodes. The statistics include a target histogram for the data in the node. The view has the following columns:

```
Name                               Type
---------------------------------- ----------------------------
PARTITION_NAME                     VARCHAR2(128)
NODE                               NUMBER
NODE_SUPPORT                       NUMBER
PREDICTED_TARGET_VALUE             NUMBER/VARCHAR2
TARGET_VALUE                       NUMBER/VARCHAR2
TARGET_SUPPORT                     NUMBER
```

Table 6-196 Decision Tree Statistics View

Parameter	Description
`PARTITION_NAME`	Partition name in a partitioned model
`NODE`	The node ID
`NODE_SUPPORT`	Number of records in the training set that belong to the node
`PREDICTED_TARGET_VALUE`	Predicted Target value
`TARGET_VALUE`	A target value seen in the training data
`TARGET_SUPPORT`	The number of records that belong to the node and have the value specified in the `TARGET_VALUE` column

The Decision Tree Nodes (DM$VO*model_name*) view describes higher level node. The DM$VO*model_name* has the following columns:

```
Name                               Type
---------------------------------- ----------------------------
PARTITION_NAME                     VARCHAR2(128)
NODE                               NUMBER
NODE_SUPPORT                       NUMBER
PREDICTED_TARGET_VALUE             NUMBER/VARCHAR2
PARENT                             NUMBER
ATTRIBUTE_NAME                     VARCHAR2(128)
ATTRIBUTE_SUBNAME                  VARCHAR2(4000)
OPERATOR                           VARCHAR2
VALUE                              SYS.XMLTYPE
```

Table 6-197 Decision Tree Nodes View

Parameter	Description
PARTITION_NAME	Partition name in a partitioned model
NODE	The node ID
NODE_SUPPORT	Number of records in the training set that belong to the node
PREDICTED_TARGET_VALUE	Predicted Target value
PARENT	The ID of the parent
ATTRIBUTE_NAME	Specifies the attribute name
ATTRIBUTE_SUBNAME	Specifies the attribute subname
OPERATOR	Attribute predicate operator - a conditional operator taking the following values: *IN*, = , <>, < , >, <=, and >=
VALUE	Value used as the description criterion. This is an XML element described using the <Element> tag. For example, <Element>Windy</Element><Element>Hot</Element>.

The Decision Tree Build Cost Matrix view (DM$VM*model_name*) describes the cost matrix used by the Decision Tree build. The DM$VM*model_name* view has the following columns:

```
Name                            Type
----------------------------------------- --------------------------------
PARTITION_NAME                            VARCHAR2(128)
ACTUAL_TARGET_VALUE                       NUMBER/VARCHAR2
PREDICTED_TARGET_VALUE                    NUMBER/VARCHAR2
COST                                      NUMBER
```

Table 6-198 Decision Tree Build Cost Matrix View

Parameter	Description
PARTITION_NAME	Partition name in a partitioned model
ACTUAL_TARGET_VALUE	Valid target value

Table 6-198 (Cont.) Decision Tree Build Cost Matrix View

Parameter	Description
PREDICTED_TARGET_VALUE	Predicted Target value
COST	Associated cost for the actual and predicted target value pair

The following table describes the Global Name-Value Pairs view (DM$VG*model_name*) columns specific to a Decision Tree model.

Table 6-199 Global Name-Value Pairs View

Name	Description
NUM_ROWS	The total number of rows used in the build

Model Detail Views for Generalized Linear Model

Model detail views specific to Generalized Linear Model (GLM) such as details and row diagnostics for linear and logistic regression models are discussed.

The following model views are available for GLM:

Model Views	Description
DM$VA*model_name*	GLM Regression Row Diagnostics
DM$VD*model_name*	GLM Regression Attribute Diagnostics
DM$VG*model_name*	Global Name-Value Pairs
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

The GLM Regression Attribute Diagnostics view (DM$VD*model_name*) describes the final model information for both linear regression models and logistic regression models.

For linear regression, the view DM$VD*model_name* has the following columns:

```
Name                                  Type
------------------------------------ ----------------------------
PARTITION_NAME                        VARCHAR2(128)
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
ATTRIBUTE_VALUE                       VARCHAR2(4000)
FEATURE_EXPRESSION                    VARCHAR2(4000)
COEFFICIENT                           BINARY_DOUBLE
STD_ERROR                             BINARY_DOUBLE
TEST_STATISTIC                        BINARY_DOUBLE
P_VALUE                               BINARY_DOUBLE
VIF                                   BINARY_DOUBLE
STD_COEFFICIENT                       BINARY_DOUBLE
LOWER_COEFF_LIMIT                     BINARY_DOUBLE
UPPER_COEFF_LIMIT                     BINARY_DOUBLE
```

For logistic regression, the view DM$VD*model_name* has the following columns:

```
Name                                  Type
------------------------------------- ----------------------------
PARTITION_NAME                        VARCHAR2(128)
TARGET_VALUE                          NUMBER/VARCHAR2
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
ATTRIBUTE_VALUE                       VARCHAR2(4000)
FEATURE_EXPRESSION                    VARCHAR2(4000)
COEFFICIENT                           BINARY_DOUBLE
STD_ERROR                             BINARY_DOUBLE
TEST_STATISTIC                        BINARY_DOUBLE
P_VALUE                               BINARY_DOUBLE
STD_COEFFICIENT                       BINARY_DOUBLE
LOWER_COEFF_LIMIT                     BINARY_DOUBLE
UPPER_COEFF_LIMIT                     BINARY_DOUBLE
EXP_COEFFICIENT                       BINARY_DOUBLE
EXP_LOWER_COEFF_LIMIT                 BINARY_DOUBLE
EXP_UPPER_COEFF_LIMIT                 BINARY_DOUBLE
```

Table 6-200 Model View for Linear and Logistic Regression Models

Column Name	Description
PARTITION_NAME	The name of a feature in the model
TARGET_VALUE	Valid target value
ATTRIBUTE_NAME	The attribute name when there is no subname, or first part of the attribute name when there is a subname. ATTRIBUTE_NAME is the name of a column in the source table or view. If the column is a non-nested, numeric column, then ATTRIBUTE_NAME is the name of the machine learning attribute. For the intercept, ATTRIBUTE_NAME is null. Intercepts are equivalent to the bias term in SVM models.
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns. When the nested column is numeric, the machine learning attribute is identified by the combination ATTRIBUTE_NAME - ATTRIBUTE_SUBNAME. If the column is not nested, ATTRIBUTE_SUBNAME is null. If the attribute is an intercept, both the ATTRIBUTE_NAME and the ATTRIBUTE_SUBNAME are null.
ATTRIBUTE_VALUE	A unique value that can be assumed by a categorical column or nested categorical column. For categorical columns, a machine learning attribute is identified by a unique ATTRIBUTE_NAME.ATTRIBUTE_VALUE pair. For nested categorical columns, a machine learning attribute is identified by the combination: ATTRIBUTE_NAME.ATTRIBUTE_SUBNAME.ATTRIBUTE_VALUE. For numerical attributes, ATTRIBUTE_VALUE is null.

Table 6-200 (Cont.) Model View for Linear and Logistic Regression Models

Column Name	Description
FEATURE_EXPRESSION	The feature name constructed by the algorithm when feature selection is enabled. If feature selection is not enabled, the feature name is the fully-qualified attribute name (*attribute_name.attribute_subname* if the attribute is in a nested column). For categorical attributes, the algorithm constructs a feature name that has the following form: *fully-qualified_attribute_name.attribute_value* When feature generation is enabled, a term in the model can be a single machine learning attribute or the product of up to 3 machine learning attributes. Component machine learning attributes can be repeated within a single term. If feature generation is not enabled or, if feature generation is enabled, but no multiple component terms are discovered by the CREATE model process, then FEATURE_EXPRESSION is null. **Note:** In 12*c* Release 2, the algorithm does not subtract the mean from numerical components.
COEFFICIENT	The estimated coefficient.
STD_ERROR	Standard error of the coefficient estimate.
TEST_STATISTIC	For linear regression, the t-value of the coefficient estimate. For logistic regression, the Wald chi-square value of the coefficient estimate.
P_VALUE	Probability of the TEST_STATISTIC under the (NULL) hypothesis that the term in the model is not statistically significant. A low probability indicates that the term is significant, while a high probability indicates that the term can be better discarded. Used to analyze the significance of specific attributes in the model.
VIF	Variance Inflation Factor. The value is zero for the intercept. For logistic regression, VIF is null.
STD_COEFFICIENT	Standardized estimate of the coefficient.
LOWER_COEFF_LIMIT	Lower confidence bound of the coefficient.
UPPER_COEFF_LIMIT	Upper confidence bound of the coefficient.
EXP_COEFFICIENT	Exponentiated coefficient for logistic regression. For linear regression, EXP_COEFFICIENT is null.
EXP_LOWER_COEFF_LIMIT	Exponentiated coefficient for lower confidence bound of the coefficient for logistic regression. For linear regression, EXP_LOWER_COEFF_LIMIT is null.
EXP_UPPER_COEFF_LIMIT	Exponentiated coefficient for upper confidence bound of the coefficient for logistic regression. For linear regression, EXP_UPPER_COEFF_LIMIT is null.

The GLM Regression Row Diagnostics view `DM$VA`*model_name* describes row level information for both linear regression models and logistic regression models. For linear regression, the view `DM$VA`*model_name* has the following columns:

```
Name                                Type
----------------------------------- ----------------------------
PARTITION_NAME                      VARCHAR2(128)
CASE_ID                             NUMBER/VARHCAR2, DATE, TIMESTAMP,
                                    TIMESTAMP WITH TIME ZONE,
                                    TIMESTAMP WITH LOCAL TIME ZONE
TARGET_VALUE                        BINARY_DOUBLE
PREDICTED_TARGET_VALUE              BINARY_DOUBLE
Hat                                 BINARY_DOUBLE
RESIDUAL                            BINARY_DOUBLE
STD_ERR_RESIDUAL                    BINARY_DOUBLE
STUDENTIZED_RESIDUAL                BINARY_DOUBLE
PRED_RES                            BINARY_DOUBLE
COOKS_D                             BINARY_DOUBLE
```

Table 6-201 GLM Regression Row Diagnostics View for Linear Regression

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`CASE_ID`	Name of the case identifier
`TARGET_VALUE`	The actual target value as taken from the input row
`PREDICTED_TARGET_VALUE`	The model predicted target value for the row
`HAT`	The diagonal element of the n*n (n=number of rows) that the Hat matrix identifies with a specific input row. The model predictions for the input data are the product of the Hat matrix and vector of input target values. The diagonal elements (Hat values) represent the influence of the i^{th} row on the i^{th} fitted value. Large Hat values are indicators that the i^{th} row is a point of high leverage, a potential outlier.
`RESIDUAL`	The difference between the predicted and actual target value for a specific input row.
`STD_ERR_RESIDUAL`	The standard error residual, sometimes called the Studentized residual, re-scales the residual to have constant variance across all input rows in an effort to make the input row residuals comparable. The process multiplies the residual by square root of the row weight divided by the product of the model mean square error and 1 minus the Hat value.
`STUDENTIZED_RESIDUAL`	Studentized deletion residual adjusts the standard error residual for the influence of the current row.
`PRED_RES`	The predictive residual is the weighted square of the deletion residuals, computed as the row weight multiplied by the square of the residual divided by 1 minus the Hat value.
`COOKS_D`	Cook's distance is a measure of the combined impact of the i^{th} case on all of the estimated regression coefficients.

For logistic regression, the view `DM$VA`*model_name* has the following columns:

```
Name                                Type
----------------------------------- ----------------------------
PARTITION_NAME                      VARCHAR2(128)
```

```
CASE_ID                            NUMBER/VARHCAR2, DATE, TIMESTAMP,
                                   TIMESTAMP WITH TIME ZONE,
                                   TIMESTAMP WITH LOCAL TIME ZONE
TARGET_VALUE                       NUMBER/VARCHAR2
TARGET_VALUE_PROB                  BINARY_DOUBLE
Hat                                BINARY_DOUBLE
WORKING_RESIDUAL                   BINARY_DOUBLE
PEARSON_RESIDUAL                   BINARY_DOUBLE
DEVIANCE_RESIDUAL                  BINARY_DOUBLE
C                                  BINARY_DOUBLE
CBAR                                BINARY_DOUBLE
DIFDEV                             BINARY_DOUBLE
DIFCHISQ                           BINARY_DOUBLE
```

Table 6-202 GLM Regression Row Diagnostics View for Logistic Regression

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`CASE_ID`	Name of the case identifier
`TARGET_VALUE`	The actual target value as taken from the input row
`TARGET_VALUE_PROB`	Model estimate of the probability of the predicted target value.
`Hat`	The Hat value concept from linear regression is extended to logistic regression by multiplying the linear regression Hat value by the variance function for logistic regression, the predicted probability multiplied by 1 minus the predicted probability.
`WORKING_RESIDUAL`	The working residual is the residual of the working response. The working response is the response on the linearized scale. For logistic regression it has the form: the i^{th} row residual divided by the variance of the i^{th} row prediction. The variance of the prediction is the predicted probability multiplied by 1 minus the predicted probability. `WORKING_RESIDUAL` is the difference between the working response and the linear predictor at convergence.
`PEARSON_RESIDUAL`	The Pearson residual is a re-scaled version of the working residual, accounting for the weight. For logistic regression, the Pearson residual multiplies the residual by a factor that is computed as square root of the weight divided by the variance of the predicted probability for the i^{th} row. `RESIDUAL` is 1 minus the predicted probability of the actual target value for the row.
`DEVIANCE_RESIDUAL`	The `DEVIANCE_RESIDUAL` is the contribution to the model deviance of the i^{th} observation. For logistic regression it has the form the square root of 2 times the `log(1 + e^eta) - eta` for the non-reference class and - square root of 2 time the `log (1 + eta)` for the reference class, where `eta` is the linear prediction (the prediction as if the model were a linear regression).
`C`	Measures the overall change in the fitted logits due to the deletion of the i^{th} observation for all points including the one deleted (the i^{th} point). It is computed as the square of the Pearson residual multiplied by the Hat value divided by the square of 1 minus the Hat value. Confidence interval displacement diagnostics that provides scalar measure of the influence of individual observations.

Table 6-202 (Cont.) GLM Regression Row Diagnostics View for Logistic Regression

Column Name	Description
CBAR	C and CBAR are extensions of Cooks' distance for logistic regression. CBAR measures the overall change in the fitted logits due to the deletion of the ith observation for all points excluding the one deleted (the ith point). It is computed as the square of the Pearson residual multiplied by the Hat value divided by (1 minus the Hat value) Confidence interval displacement diagnostic which measures the influence of deleting an individual observation.
DIFDEV	A statistic that measures the change in deviance that occurs when an observation is deleted from the input. It is computed as the square of the deviance residual plus CBAR.
DIFCHISQ	A statistic that measures the change in the Pearson chi-square statistic that occurs when an observation is deleted from the input. It is computed as CBAR divided by the Hat value.

Global Details for GLM: Linear Regression

The following table describes Global Name-Value Pairs (DM$VG) for a linear regression model.

Table 6-203 Global Details for Linear Regression

Name	Description
ADJUSTED_R_SQUARE	Adjusted R-Square
AIC	Akaike's information criterion
COEFF_VAR	Coefficient of variation
CONVERGED	Indicates whether the model build process has converged to specified tolerance. The following are the possible values: • YES • NO
CORRECTED_TOTAL_DF	Corrected total degrees of freedom
CORRECTED_TOT_SS	Corrected total sum of squares
DEPENDENT_MEAN	Dependent mean
ERROR_DF	Error degrees of freedom
ERROR_MEAN_SQUARE	Error mean square
ERROR_SUM_SQUARES	Error sum of squares
F_VALUE	Model *F* value statistic
GMSEP	Estimated mean square error of the prediction, assuming multivariate normality
HOCKING_SP	Hocking Sp statistic
ITERATIONS	Tracks the number of SGD iterations. Applicable only when the solver is SGD.
J_P	JP statistic (the final prediction error)
MODEL_DF	Model degrees of freedom
MODEL_F_P_VALUE	Model *F* value probability

Table 6-203 (Cont.) Global Details for Linear Regression

Name	Description
MODEL_MEAN_SQUARE	Model mean square error
MODEL_SUM_SQUARES	Model sum of square errors
NUM_PARAMS	Number of parameters (the number of coefficients, including the intercept)
NUM_ROWS	Number of rows
R_SQ	R-Square
RANK_DEFICIENCY	The number of predictors excluded from the model due to multi-collinearity
ROOT_MEAN_SQ	Root mean square error
SBIC	Schwarz's Bayesian information criterion

Global Details for GLM: Logistic Regression

The following table returns Global Name-Value Pairs (DM$VG) for a logistic regression model.

Table 6-204 Global Details for Logistic Regression

Name	Description
AIC_INTERCEPT	Akaike's criterion for the fit of the baseline, intercept-only, model
AIC_MODEL	Akaike's criterion for the fit of the intercept and the covariates (predictors) mode
CONVERGED	Indicates whether the model build process has converged to specified tolerance. The following are the possible values: • YES • NO
DEPENDENT_MEAN	Dependent mean
ITERATIONS	Tracks the number of SGD iterations (number of IRLS iterations). Applicable only when the solver is SGD.
LR_DF	Likelihood ratio degrees of freedom
LR_CHI_SQ	Likelihood ratio chi-square value
LR_CHI_SQ_P_VALUE	Likelihood ratio chi-square probability value
NEG2_LL_INTERCEPT	-2 log likelihood of the baseline, intercept-only, model
NEG2_LL_MODEL	-2 log likelihood of the model
NUM_PARAMS	Number of parameters (the number of coefficients, including the intercept)
NUM_ROWS	Number of rows
PCT_CORRECT	Percent of correct predictions
PCT_INCORRECT	Percent of incorrectly predicted rows
PCT_TIED	Percent of cases where the estimated probabilities are equal for both target classes
PSEUDO_R_SQ_CS	Pseudo R-square Cox and Snell
PSEUDO_R_SQ_N	Pseudo R-square Nagelkerke

Table 6-204 (Cont.) Global Details for Logistic Regression

Name	Description
`RANK_DEFICIENCY`	The number of predictors excluded from the model due to multi-collinearity
`SC_INTERCEPT`	Schwarz's Criterion for the fit of the baseline, intercept-only, model
`SC_MODEL`	Schwarz's Criterion for the fit of the intercept and the covariates (predictors) model

> **Note:**
>
> - When ridge regression is enabled, fewer global details are returned. For information about ridge, see *Oracle Machine Learning for SQL Concepts*.
> - When the value is `NULL` for a partitioned model, an exception is thrown. When the value is not null, it must contain the desired partition name.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*
- Model Detail Views for Global Information
  Model detail views for global information contain information about global statistics, alerts, and computed settings.

Model Detail View for Multivariate State Estimation Technique - Sequential Probability Ratio Test

The model detail view specific to Multivariate State Estimation Technique - Sequential Probability Ratio Test contains information about Global Name-Value Paris.

The following are the available model views for MSET-SPRT:

Views	Description
`DM$VC`*model_name*	Scoring Cost Matrix
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VN`*model_name*	Normalization and Missing Value Handling
`DM$VS`*model_name*	Computed Settings
`DM$VT`*model_name*	Classification Targets
`DM$VW`*model_name*	Model Build Alerts

The following table lists the Global Name-Value Pairs (`DM$VG`*model_name*) for an MSET-SPRT. This statistic is included when due to memory constraints MSET-SPRT cannot use the `MSET_MEMORY_VECTORS` value set by the user.

Table 6-205 MSET-SPRT Information in the Model Global View

Name	Description
NUM_MVEC	The number of memory vectors used by the model.

Model Detail Views for Naive Bayes

The model detail views specific to Naive Bayes are the prior view and result view.

These the model views available for Naive Bayes:

Model Views	Description
DM$VB*model_name*	Automatic Data Preparation Binning
DM$VC*model_name*	Scoring Cost Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VP*model_name*	Naive Bayes Target Priors
DM$VS*model_name*	Computed Settings
DM$VT*model_name*	Classification Targets
DM$VV*model_name*	Naive Bayes Conditional Probabilities
DM$VW*model_name*	Model Build Alerts

The Naive Bayes Target Priors view (DM$VP*model_name*) describes the priors of the targets for a Naive Bayes model. The view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
TARGET_NAME                               VARCHAR2(128)
TARGET_VALUE                              NUMBER/VARCHAR2
PRIOR_PROBABILITY                         BINARY_DOUBLE
COUNT                                     NUMBER
```

Table 6-206 Naive Bayes Target Priors View for Naive Bayes

Column Name	Description
PARTITION_NAME	The name of a feature in the model
TARGET_NAME	Name of the target column
TARGET_VALUE	Target value, numerical or categorical
PRIOR_PROBABILITY	Prior probability for a given TARGET_VALUE
COUNT	Number of rows for a given TARGET_VALUE

The Naive Bayes Conditional Probabilities view (DM$VV*model_view*) describes the conditional probabilities of the Naive Bayes model. The view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
```

```
TARGET_NAME                              VARCHAR2(128)
TARGET_VALUE                             NUMBER/VARCHAR2
ATTRIBUTE_NAME                           VARCHAR2(128)
ATTRIBUTE_SUBNAME                        VARCHAR2(4000)
ATTRIBUTE_VALUE                          VARCHAR2(4000)
CONDITIONAL_PROBABILITY                  BINARY_DOUBLE
COUNT                                    NUMBER
```

Table 6-207 Naive Bayes Conditional Probabilities View for Naive Bayes

Column Name	Description
PARTITION_NAME	The name of a feature in the model
TARGET_NAME	Name of the target column
TARGET_VALUE	Target value, numerical or categorical
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Machine learning attribute value for the column ATTRIBUTE_NAME or the nested column ATTRIBUTE_SUBNAME (if any).
CONDITIONAL_PROBABILITY	Conditional probability of a machine learning attribute for a given target
COUNT	Number of rows for a given machine learning attribute and a given target

The following table describes the Global Name-Value Pairs view (DM$VG*model_name*) specific to a Naive Bayes model.

Table 6-208 Global Name-Value Pairs View for Naive Bayes

Name	Description
NUM_ROWS	The total number of rows used in the build

Model Detail Views for Neural Network

Model detail views specific to Neural Network contain information about the weights of the neurons: input layer and hidden layers.

These are the model views available for Neural Network:

Model Views	Description
DM$VA*model_name*	Neural Network Weights
DM$VC*model_name*	Scoring Cost Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VS*model_name*	Computed Settings
DM$VT*model_name*	Classification Targets
DM$VW*model_name*	Model Build Alerts

The Neural Network Weights view (`DM$VA`*model_name*) has the following columns:

```
Name
Type
----------------------      -----------------------
PARTITION_NAME              VARCHAR2(128)
LAYER                       NUMBER
IDX_FROM                    NUMBER
ATTRIBUTE_NAME              VARCHAR2(128)
ATTRIBUTE_SUBNAME           VARCHAR2(4000)
ATTRIBUTE_VALUE             VARCHAR2(4000)
IDX_TO                      NUMBER
TARGET_VALUE                NUMBER/VARCHAR2
WEIGHT                      BINARY_DOUBLE
```

Table 6-209   Neural Network Weights View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`LAYER`	Layer ID, 0 as an input layer
`IDX_FROM`	Node index that the weight connects from (attribute id for input layer)
`ATTRIBUTE_NAME`	Attribute name (only for the input layer)
`ATTRIBUTE_SUBNAME`	Attribute subname. The value is null for non-nested columns.
`ATTRIBUTE_VALUE`	Categorical attribute value
`IDX_TO`	Node index that the weights connects to
`TARGET_VALUE`	Target value. The value is null for regression.
`WEIGHT`	Value of the weight

The view Global Name-Value Pairs (`DM$VG`*model_name*) is a pre-existing view. The following name-value pairs are specific to a Neural Network view.

Table 6-210   Global Name-Value Pairs Viewfor Neural Network

Name	Description
`CONVERGED`	Indicates whether the model build process has converged to specified tolerance. The following are the possible values: • `YES` • `NO`
`ITERATIONS`	Number of iterations
`LOSS_VALUE`	Loss function value (if it is with `NNET_REGULARIZER_HELDASIDE` regularization, it is the loss function value on test data)
`NUM_ROWS`	Number of rows in the model (or partitioned model)

Model Detail Views for Random Forest

Model detail views specific to Random Forest contain variable importance measures and statistics.

The following model detail views are available for Random Forest:

Model View	Description
`DM$VA`*model_name*	Variable Importance
`DM$VC`*model_name*	Scoring Cost Matrix
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VS`*model_name*	Computed Settings
`DM$VT`*model_name*	Classification Targets
`DM$VW`*model_name*	Model Build Alerts

Model detail views and statistics specific to Random Forest are:

- Variable Importance statistics `DM$VA`*model_name*
- Random Forest statistics in the Global Name-Value Pairs `DM$VG`*model_name* view

One of the important outputs from a Random Forest model build is a ranking of attributes based on their relative importance. This is measured using Mean Decrease Gini. The `DM$VA`*model_name* view has the following columns:

```
Name                              Type
------------------------          ---------------
PARTITION_NAME                    VARCHAR2(128)
ATTRIBUTE_NAME                    VARCHAR2(128)
ATTRIBUTE_SUBNAME                 VARCHAR2(128)
ATTRIBUTE_IMPORTANCE              BINARY_DOUBLE
```

Table 6-211   Variable Importance Model View

Column Name	Description
`PARTITION_NAME`	Partition name. The value is null for models which are not partitioned.
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_SUBNAME`	Nested column subname. The value is null for non-nested columns.
`ATTRIBUTE_IMPORTANCE`	Measure of importance for an attribute in the forest (mean Decrease Gini value)

The Global Name-Value Pairs (`DM$VG`*model_name*) view is a pre-existing view. The following name-value pairs are added to the view.

Table 6-212 Random Forest Statistics Information In Model Global View

Name	Description
AVG_DEPTH	Average depth of the trees in the forest
AVG_NODECOUNT	Average number of nodes per tree
MAX_DEPTH	Maximum depth of the trees in the forest
MAX_NODECOUNT	Maximum number of nodes per tree
MIN_DEPTH	Minimum depth of the trees in the forest
MIN_NODECOUNT	Minimum number of nodes per tree
NUM_ROWS	The total number of rows used in the build

Model Detail View for Support Vector Machine

Model detail views specific to Support Vector Machine (SVM) contain linear coefficients and support vector statistics.

These model views are available for SVM:

Model Views	Description
DM$VCS*model_name*	Scoring Cost Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VS*model_name*	Computed Settings
DM$VT*model_name*	Classification Targets
DM$VW*model_name*	Model Build Alerts

The linear coefficient view DM$VL*model_name* describes the coefficients of a linear SVM algorithm. The *target_value* field in the view is present only for classification and has the type of the target. Regression models do not have a *target_value* field.

The *reversed_coefficient* field shows the value of the coefficient after reversing the automatic data preparation transformations. If data preparation is disabled, then *coefficient* and *reversed_coefficient* have the same value. The view has the following columns:

```
Name                                       Type
----------------------------------------- --------------------------------
PARTITION_NAME                            VARCHAR2(128)
TARGET_VALUE                              NUMBER/VARCHAR2
ATTRIBUTE_NAME                            VARCHAR2(128)
ATTRIBUTE_SUBNAME                         VARCHAR2(4000)
ATTRIBUTE_VALUE                           VARCHAR2(4000)
COEFFICIENT                               BINARY_DOUBLE
REVERSED_COEFFICIENT                      BINARY_DOUBLE
```

Table 6-213 Linear Coefficient View for Support Vector Machine

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
TARGET_VALUE	Target value, numerical or categorical
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Value of a categorical attribute
COEFFICIENT	Projection coefficient value
REVERSED_COEFFICIENT	Coefficient transformed on the original scale

The following table describes the SVM statistics global view.

Table 6-214 Support Vector Statistics Information In Model Global View

Name	Description
CONVERGED	Indicates whether the model build process has converged to specified tolerance: • YES • NO
ITERATIONS	Number of iterations performed during build
NUM_ROWS	Number of rows used for the build
REMOVED_ROWS_ZERO_NORM	Number of rows removed due to 0 norm. This applies to one-class linear models only.

Model Detail Views for XGBoost

The model detail views specific to XGBoost contain information about Feature Importance view and Global Name-Value Pairs view.

The following are the available model views for XGBoost Classification:

Model Views	Description
DM$VC*model_name*	Scoring Cost Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VI*model_name*	XGBoost Attribute Importance
DM$VS*model_name*	Computed Settings
DM$VT*model_name*	Classification Targets
DM$VW*model_name*	Model Build Alerts

The following are the available model views for XGBoost Regression:

Views	Description
DM$VG*model_name*	Global Name-Value Pairs
DM$VI*model_name*	XGBoost Attribute Importance

Views	Description
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

The DM$VI*model_name* view reports the feature importance values for each attribute of each partition of the model.

The view has the following columns for tree models (gbtree and dart boosters).

```
Name                 Type
-----------------    --------------
PNAME                VARCHAR2(128)
ATTRIBUTE_NAME       VARCHAR2(128)
ATTRIBUTE_SUBNAME    VARCHAR2(4000)
ATTRIBUTE_VALUE      VARCHAR2(4000)
GAIN                 BINARY_DOUBLE
COVER                BINARY_DOUBLE
FREQUENCY            BINARY_DOUBLE
```

Table 6-215 Feature Importance View for a Tree Model

Column Name	Description
PNAME	The name of a partition in a partitioned model.
ATTRIBUTE_NAME	The column name.
ATTRIBUTE_SUBNAME	The nested column subname; the value is null for non-nested columns.
ATTRIBUTE_VALUE	The value of a categorical attribute.
GAIN	The fractional contribution of each feature to the model based on the total gain of a feature’s splits; a higher percentage means a more important predictive feature.
COVER	The number of observation either seen by a split or collected by a leaf during training.
FREQUENCY	A percentage representing the relative number of times a feature has been used in trees.

For a linear model (gblinear) booster, the feature importance is the absolute magnitude of linear coefficients.

The view has the following columns for linear models.

```
Name            Type
-----------------    --------------
PNAME                VARCHAR2(128)
ATTRIBUTE_NAME       VARCHAR2(128)
ATTRIBUTE_SUBNAME    VARCHAR2(4000)
ATTRIBUTE_VALUE      VARCHAR2(4000)
WEIGHT               BINARY_DOUBLE
CLASS                BINARY_DOUBLE
```

Table 6-216 Feature Importance View for a Linear Model

Column Name	Description
PNAME	The name of a partition in a partitioned model.
ATTRIBUTE_NAME	The column name.
ATTRIBUTE_SUBNAME	The nested column subname; the value is null for non-nested columns.
ATTRIBUTE_VALUE	The value of a categorical attribute.
WEIGHT	The linear coefficient of the feature.
CLASS	The class label for a multiclass model.

The DM$VG*model_name* view reports global statistics for an XGBoost model. The statistics include an evaluation of the training data set using the evaluation metric you specified with the learning task eval_metric setting, or the default eval_metric if you didn't specify one. The view displays only the result of the last training iteration. When you specify more than one eval_metric, the view contains multiple rows, one for each eval_metric.

Model Detail Views for Clustering Algorithms

Oracle Machine Learning for SQL supports these clustering algorithms: Expectation Maximization (EM), *k*-Means (KM), and orthogonal partitioning clustering (O-Cluster, OC).

All clustering algorithms share the following views:

Model Views	Description
DM$VD*model_name*:	Clustering Description
DM$VA*model_name*	Clustering Attribute Statistics
DM$VH*model_name*	Clustering Histograms
DM$VR*model_name*	Clustering Rules

The Cluster Description view DM$VD*model_name* describes cluster level information about a clustering model. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
CLUSTER_ID                            NUMBER
CLUSTER_NAME                          NUMBER/VARCHAR2
RECORD_COUNT                          NUMBER
PARENT                                NUMBER
TREE_LEVEL                            NUMBER
LEFT_CHILD_ID                         NUMBER
RIGHT_CHILD_ID                        NUMBER
```

Table 6-217 Clustering Description View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model

Table 6-217 (Cont.) Clustering Description View

Column Name	Description
CLUSTER_ID	The ID of a cluster in the model
CLUSTER_NAME	Specifies the label of the cluster
RECORD_COUNT	Specifies the number of records
PARENT	The ID of the parent
TREE_LEVEL	Specifies the number of splits from the root
LEFT_CHILD_ID	The ID of the child cluster on the left side of the split
RIGHT_CHILD_ID	The ID of the child cluster on the right side of the split

The attribute view `DM$VA`*model_name* describes attribute level information about a clustering model. The values of the mean, variance, and mode for a particular cluster can be obtained from this view. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
CLUSTER_ID                            NUMBER
CLUSTER_NAME                          NUMBER/VARCHAR2
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
MEAN                                  BINARY_DOUBLE
VARIANCE                              BINARY_DOUBLE
MODE_VALUE                            VARCHAR2(4000)
```

Table 6-218 Clustering Attribute Statistics

Column Name	Description
PARTITION_NAME	A partition in a partitioned model
CLUSTER_ID	The ID of a cluster in the model
CLUSTER_NAME	Specifies the label of the cluster
ATTRIBUTE_NAME	Specifies the attribute name
ATTRIBUTE_SUBNAME	Specifies the attribute subname
MEAN	The field returns the average value of a numeric attribute
VARIANCE	The variance of a numeric attribute
MODE_VALUE	The mode is the most frequent value of a categorical attribute

The histogram view `DM$VH`*model_name* describes histogram level information about a clustering model. The bin information as well as bin counts can be obtained from this view. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
CLUSTER_ID                            NUMBER
CLUSTER_NAME                          NUMBER/VARCHAR2
```

```
ATTRIBUTE_NAME                     VARCHAR2(128)
ATTRIBUTE_SUBNAME                  VARCHAR2(4000)
BIN_ID                             NUMBER
LOWER_BIN_BOUNDARY                 BINARY_DOUBLE
UPPER_BIN_BOUNDARY                 BINARY_DOUBLE
ATTRIBUTE_VALUE                    VARCHAR2(4000)
COUNT                              NUMBER
```

Table 6-219 Clustering Histograms View

Column Name	Description
`PARTITION_NAME`	A partition in a partitioned model
`CLUSTER_ID`	The ID of a cluster in the model
`CLUSTER_NAME`	Specifies the label of the cluster
`ATTRIBUTE_NAME`	Specifies the attribute name
`ATTRIBUTE_SUBNAME`	Specifies the attribute subname
`BIN_ID`	Bin ID
`LOWER_BIN_BOUNDARY`	Numeric lower bin boundary
`UPPER_BIN_BOUNDARY`	Numeric upper bin boundary
`ATTRIBUTE_VALUE`	Categorical attribute value
`COUNT`	Histogram count

The rule view `DM$VR`*model_name* describes the rule level information about a clustering model. The information is provided at attribute predicate level. The view has the following columns:

```
Name                               Type
---------------------------------- ----------------------------
PARTITION_NAME                     VARCHAR2(128)
CLUSTER_ID                         NUMBER
CLUSTER_NAME                       NUMBER/VARCHAR2
ATTRIBUTE_NAME                     VARCHAR2(128)
ATTRIBUTE_SUBNAME                  VARCHAR2(4000)
OPERATOR                           VARCHAR2(2)
NUMERIC_VALUE                      NUMBER
ATTRIBUTE_VALUE                    VARCHAR2(4000)
SUPPORT                            NUMBER
CONFIDENCE                         BINARY_DOUBLE
RULE_SUPPORT                       NUMBER
RULE_CONFIDENCE                    BINARY_DOUBLE
```

Table 6-220 Clustering Rules View

Column Name	Description
`PARTITION_NAME`	A partition in a partitioned model
`CLUSTER_ID`	The ID of a cluster in the model
`CLUSTER_NAME`	Specifies the label of the cluster
`ATTRIBUTE_NAME`	Specifies the attribute name

Table 6-220 (Cont.) Clustering Rules View

Column Name	Description
ATTRIBUTE_SUBNAME	Specifies the attribute subname
OPERATOR	Attribute predicate operator - a conditional operator taking the following values: *IN*, = , <>, < , >, <=, and >=
NUMERIC_VALUE	Numeric lower bin boundary
ATTRIBUTE_VALUE	Categorical attribute value
SUPPORT	Attribute predicate support
CONFIDENCE	Attribute predicate confidence
RULE_SUPPORT	Rule level support
RULE_CONFIDENCE	Rule level confidence

Model Detail Views for Expectation Maximization

Model detail views specific to Expectation Maximization (EM) contain additional information about an EM model. Additional views are available for EM Clustering, but are absent for EM Anomaly.

These are the model views available for Expectation Maximization:

Model Views	Description
DM$VA*model_name*	Clustering Attribute Statistics
DM$VB*model_name*	Attribute Pair Kullback-Leibler Divergence
DM$VD*model_name*	Clustering Description
DM$VF*model_name*	Expectation Maximization Bernoulli parameters
DM$VG*model_name*	Global Name-Value Pairs
DM$VH*model_name*	Clustering Histograms
DM$VI*model_name*	Unsupervised Attribute Importance
DM$VM*model_name*	Expectation Maximization Gaussian parameters
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VO*model_name*	Expectation Maximization Components
DM$VP*model_name*	Expectation Maximization Projections
DM$VR*model_name*	Clustering Rules
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

For EM Clustering model, the following views contain information that is not in the clustering views. For the clustering views, refer to "Model Detail Views for Clustering Algorithms".

The Expectation Maximization Components view (DM$VO*model_name*) describes the EM Cluster components. The component view contains information about their prior probabilities and what cluster they map to. The view has the following columns:

```
Name                                   Type
-------------------------------------- ------------------------------
PARTITION_NAME                         VARCHAR2(128)
```

```
COMPONENT_ID                         NUMBER
CLUSTER_ID                           NUMBER
PRIOR_PROBABILITY                    BINARY_DOUBLE
```

Table 6-221   Expectation Maximization Components View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`COMPONENT_ID`	Unique identifier of a component
`CLUSTER_ID`	The ID of a cluster in the model
`PRIOR_PROBABILITY`	Component prior probability

The Expectation Maximization Gaussian view (`DM$VM`*model_name*) provides information about the mean and variance parameters for the attributes by Gaussian distribution models. The view has the following columns:

```
Name                                 Type
----------------------------------   ----------------------------
PARTITION_NAME                       VARCHAR2(128)
COMPONENT_ID                         NUMBER
ATTRIBUTE_NAME                       VARCHAR2(4000)
MEAN                                 BINARY_DOUBLE
VARIANCE                             BINARY_DOUBLE
```

The Expectation Maximization Bernoulli parameters view (`DM$VF`*model_name*) provides information about the parameters of the multi-valued Bernoulli distributions used by the EM model. The view has the following columns:

```
Name                                 Type
 ----------------------------------   ----------------------------
 PARTITION_NAME                       VARCHAR2(128)
 COMPONENT_ID                         NUMBER
 ATTRIBUTE_NAME                       VARCHAR2(4000)
 ATTRIBUTE_VALUE                      VARCHAR2(4000)
 FREQUENCY                            BINARY_DOUBLE
```

Table 6-222   Expectation Maximization Bernoulli parameters View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`COMPONENT_ID`	Unique identifier of a component
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_VALUE`	Categorical attribute value
`FREQUENCY`	The frequency of the multivalued Bernoulli distribution for the attribute/value combination specified by `ATTRIBUTE_NAME` and `ATTRIBUTE_VALUE`.

For 2-Dimensional columns, EM provides an attribute ranking similar to that of attribute importance. This ranking is based on a rank-weighted average over Kullback–Leibler divergence computed for pairs of columns. This unsupervised attribute importance is shown in the Unsupervised Attribute Importance view (DM$VI*model_name*) and has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
ATTRIBUTE_NAME                            VARCHAR2(128)
ATTRIBUTE_IMPORTANCE_VALUE                BINARY_DOUBLE
ATTRIBUTE_RANK                            NUMBER
```

Table 6-223 Unsupervised Attribute Importance View for Expectation Maximization

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
ATTRIBUTE_NAME	Column name
ATTRIBUTE_IMPORTANCE_VALUE	Importance value
ATTRIBUTE_RANK	An attribute rank based on the importance value

The pairwise Kullback–Leibler divergence is reported in the Attribute Pair Kullback-Leibler Divergence view (DM$VB*model_name*). This metric evaluates how much the observed joint distribution of two attributes diverges from the expected distribution under the assumption of independence. That is, the higher the value, the more dependent the two attributes are. The dependency value is scaled based on the size of the grid used for each pairwise computation. That ensures that all values fall within the [0; 1] range and are comparable. The view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
ATTRIBUTE_NAME_1                          VARCHAR2(128)
ATTRIBUTE_NAME_2                          VARCHAR2(128)
DEPENDENCY                                BINARY_DOUBLE
```

Table 6-224 Attribute Pair Kullback-Leibler Divergence View for Expectation Maximization

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
ATTRIBUTE_NAME_1	Name of the first attribute
ATTRIBUTE_NAME_2	Name of the second attribute
DEPENDENCY	Scaled pairwise Kullback-Leibler divergence

The projection table DM$VP*model_name* shows the coefficients used by random projections to map nested columns to a lower dimensional space. The view has rows only when nested or text data is present in the build data. The view has the following columns:

```
Name                               Type
---------------------------------- ----------------------------
PARTITION_NAME                     VARCHAR2(128)
```

```
FEATURE_NAME                       VARCHAR2(4000)
ATTRIBUTE_NAME                     VARCHAR2(128)
ATTRIBUTE_SUBNAME                  VARCHAR2(4000)
ATTRIBUTE_VALUE                    VARCHAR2(4000)
COEFFICIENT                        NUMBER
```

Table 6-225 Projection table for Expectation Maximization

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`FEATURE_NAME`	Name of feature
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_SUBNAME`	Nested column subname. The value is null for non-nested columns.
`ATTRIBUTE_VALUE`	Categorical attribute value
`COEFFICIENT`	Projection coefficient. The representation is sparse; only the non-zero coefficients are returned.

For EM Anomaly, currently there are no additional views other than the classification views. For the classification view, refer to “Model Detail Views for Classification Algorithms”.

Global Details for Expectation Maximization

The following table describes global details for EM.

Table 6-226 Global Details for Expectation Maximization

Name	Description
`CONVERGED`	Indicates whether the model build process has converged to specified tolerance. The possible values are: • `YES` • `NO`
`LOGLIKELIHOOD`	Loglikelihood on the build data
`NUM_COMPONENTS`	Number of components produced by the model
`NUM_CLUSTERS`	Number of clusters produced by the model (only available for EM Clustering)
`NUM_ROWS`	Number of rows used in the build
`RANDOM_SEED`	The random seed value used for the model build
`REMOVED_COMPONENTS`	The number of empty components excluded from the model

Related Topics

- Model Detail Views for Clustering Algorithms
  Oracle Machine Learning for SQL supports these clustering algorithms: Expectation Maximization (EM), *k*-Means (KM), and orthogonal partitioning clustering (O-Cluster, OC).

Model Detail Views for *k*-Means

Model detail views specific to *k*-Means (KM) contain clustering description view (DM$VG), and scoring information.

The following model views are available for *k*-Means algorithm.

Model Views	Description
DM$VA*model_name*	Clustering Attribute Statistics
DM$VC*model_name*	k-Means Scoring Centroids
DM$VD*model_name*	Clustering Description
DM$VG*model_name*	Global Name-Value Pairs
DM$VH*model_name*	Clustering Histograms
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VR*model_name*	Clustering Rules
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

"Model Detail Views for Clustering Algorithms" discusses common model views across clustering algorithms. Global Name-Value Pairs view (DM$VG), which contains information about Computed Settings view (DM$VS) and Model Build Alerts view (DM$VW), and Normalization and Missing Value Handling view (DM$VN) are addressed individually.

The following views contain information that is specific to *k*-Means model.

The *k*-Means Clustering Description view DM$VD*model_name* has an additional column:

```
Name                                Type
----------------------------------  ----------------------------
DISPERSION                          BINARY_DOUBLE
```

Table 6-227   Clustering Description for k-Means

Column Name	Description
DISPERSION	A measure used to quantify whether a set of observed occurrences are dispersed compared to a standard statistical model.

The *k*-Means Scoring Centroids view DM$VC*model_name* describes the centroid of each leaf clusters:

```
Name                                Type
 ----------------------------------  ----------------------------
 PARTITION_NAME                      VARCHAR2(128)
 CLUSTER_ID                          NUMBER
 CLUSTER_NAME                        NUMBER/VARCHAR2
 ATTRIBUTE_NAME                      VARCHAR2(128)
 ATTRIBUTE_SUBNAME                   VARCHAR2(4000)
 ATTRIBUTE_VALUE                     VARCHAR2(4000)
 VALUE                               BINARY_DOUBLE
```

Table 6-228 k-Means Scoring Centroids View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
CLUSTER_ID	The ID of a cluster in the model
CLUSTER_NAME	Specifies the label of the cluster
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Categorical attribute value
VALUE	Specifies the centroid value

The following table describes Global Name-Value Pairs view (DM$VG) for *k*-Means.

Table 6-229 *k*–Means Global Name-Value Pairs View

Name	Description
CONVERGED	Indicates whether the model build process has converged to specified tolerance. The following are the possible values: • YES • NO
NUM_ROWS	Number of rows used in the build
REMOVED_ROWS_ZERO_NORM	Number of rows removed due to 0 norm. This applies only to models using cosine distance.

Related Topics

- Model Detail Views for Clustering Algorithms
  Oracle Machine Learning for SQL supports these clustering algorithms: Expectation Maximization (EM), *k*-Means (KM), and orthogonal partitioning clustering (O-Cluster, OC).
- Model Detail Views for Global Information
  Model detail views for global information contain information about global statistics, alerts, and computed settings.

Model Detail Views for O-Cluster

Model detail views specific to O-Cluster (OC) contain information about description view, histograms view, and global view.

These are the available model views for O-Cluster:

Model Views	Description
DM$VA*model_name*	Clustering Attribute Statistics
DM$VB*model_name*	Automatic Data Preparation Binning
DM$VD*model_name*	Clustering Description
DM$VG*model_name*	Global Name-Value Pairs

Model Views	Description
DM$VH*model_name*	Clustering Histograms
DM$VR*model_name*	Clustering Rules
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

The following views contain information that is specific to an O-Cluster model. For the clustering views, refer to "Model Detail Views for Clustering Algorithms". The OC algorithm uses the same descriptive statistics views as Expectation Maximization (EM) and *k*-Means (KM). The following are the statistics views:

The Cluster Description view (DM$VD*model_name*) describes the O-Cluster components. The Cluster Description view has additional fields that specify the split predicate. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
OPERATOR                              VARCHAR2(2)
VALUE                                 SYS.XMLTYPE
```

Table 6-230   Cluster Description View for O-Cluster

Column Name	Description
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
OPERATOR	Split operator
VALUE	List of split values

The structure of the SYS.XMLTYPE is as follows:

```
<Element>splitval1</Element>
```

The OC algorithm uses a Clustering Histograms view (DM$VH*model_name*) with different columns than EM and KM. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITON_NAME                         VARCHAR2(128)
CLUSTER_ID                            NUMBER
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
BIN_ID                                NUMBER
LABEL                                 VARCHAR2(4000)
COUNT                                 NUMBER
```

Table 6-231 Clustering Histograms View for O-Cluster

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`CLUSTER_ID`	Unique identifier of a component
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_SUBNAME`	Nested column subname. The value is null for non-nested columns.
`BIN_ID`	Unique identifier
`LABEL`	Bin label
`COUNT`	Bin histogram count

The following table describes the Global Name-Value Pairs (`DM$VG`*model_name*) view specific to O-Cluster.

Table 6-232 O-Cluster Statistics Information In Model Global View

Name	Description
`NUM_ROWS`	The total number of rows used in the build

Related Topics

- Model Detail Views for Clustering Algorithms
  Oracle Machine Learning for SQL supports these clustering algorithms: Expectation Maximization (EM), *k*-Means (KM), and orthogonal partitioning clustering (O-Cluster, OC).

Model Detail Views for Explicit Semantic Analysis

Model detail views specific to Explicit Semantic Analysis (ESA) contain information about attribute statistics and features.

These are the available model views:

Model Views	Description
`DM$VA`*model_name*	Explicit Semantic Analysis Matrix
`DM$VF`*model_name*	Explicit Semantic Analysis Features
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VN`*model_name*	Normalization and Missing Value Handling
`DM$VS`*model_name*	Computed Settings
`DM$VW`*model_name*	Model Build Alerts
`DM$VX`*model_name*	Text Features

- Explicit Semantic Analysis Matrix (`DM$VA`*model_name*): This view has different columns for feature extraction and classification. For feature extraction, this view contains model attribute coefficients per feature. For classification, this view contains model attribute coefficients per target class.
- Explicit Semantic Analysis Features (`DM$VF`*model_name*): This view is applicable only for feature extraction.

The Explicit Semantic Analysis Matrix view (DM$VA*model_name*) has the following columns for feature extraction:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
FEATURE_ID                            NUMBER/VARHCAR2, DATE, TIMESTAMP,
                                      TIMESTAMP WITH TIME ZONE,
                                      TIMESTAMP WITH LOCAL TIME ZONE
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
ATTRIBUTE_VALUE                       VARCHAR2(4000)
COEFFICIENT                           BINARY_DOUBLE
```

Table 6-233 Explicit Semantic Analysis Matrix for Feature Extraction

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
FEATURE_ID	Unique identifier of a feature as it appears in the training data
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Categorical attribute value
COEFFICIENT	A measure of the weight of the attribute with respect to the feature

The (DM$VA*model_name*) view comprises of attribute coefficients for all target classes.

The view Explicit Semantic Analysis Matrix (DM$VA*model_name*) has the following columns for classification:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
TARGET_VALUE                          NUMBER/VARCHAR2
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
ATTRIBUTE_VALUE                       VARCHAR2(4000)
COEFFICIENT                           BINARY_DOUBLE
```

Table 6-234 Explicit Semantic Analysis Matrix for Classification

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
TARGET_VALUE	Value of the target
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Categorical attribute value

Table 6-234 (Cont.) Explicit Semantic Analysis Matrix for Classification

Column Name	Description
COEFFICIENT	A measure of the weight of the attribute with respect to the feature

The Explicit Semantic Analysis Features view (DM$VF*model_name*) has a unique row for every feature in one view. This feature is helpful if the model was pre-built and the source training data are not available. The view has the following columns:

```
Name                                  Type
------------------------------------ ----------------------------
PARTITION_NAME                        VARCHAR2(128)
FEATURE_ID                            NUMBER/VARHCAR2, DATE, TIMESTAMP,
                                      TIMESTAMP WITH TIME ZONE,
                                      TIMESTAMP WITH LOCAL TIME ZONE
```

Table 6-235 Explicit Semantic Analysis Features for Explicit Semantic Analysis

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
FEATURE_ID	Unique identifier of a feature as it appears in the training data

The following table describes the Global Name-Value Pairs view (DM$VG*model_name)* specific to ESA.

Table 6-236 Explicit Semantic Analysis Statistics Information In Model Global View

Name	Description
NUM_ROWS	The total number of input rows
REMOVED_ROWS_BY_FILTERS	Number of rows removed by filters

Model Detail Views for Non-Negative Matrix Factorization

Model detail views specific to Non-Negative Matrix Factorization (NMF) contain information about the encoding H matrix and H inverse matrix.

These are the available model views for NMF:

Model Views	Description
DM$VE*model_name*	Non-Negative Matrix Factorization H Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VI*model_name*	Non-Negative Matrix Factorization Inverse H Matrix
DM$VN*model_name*	Normalization andMissing Value Handling
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts

The views specific to NMF are:

- Non-Negative Matrix Factorization H Matrix view (DM$VE*model_name*)
- Non-Negative Matrix Factorization Inverse H Matrix view (DM$VI*model_name*)

The view DM$VE*model_name* describes the encoding (H) matrix of an NMF model. The FEATURE_NAME column type may be either NUMBER or VARCHAR2. The view has the following columns.

```
Name                   Type
-------------------    --------------------------
PARTITION_NAME         VARCHAR2(128)
FEATURE_ID             NUMBER
FEATURE_NAME           NUMBER/VARCHAR2
ATTRIBUTE_NAME         VARCHAR2(128)
ATTRIBUTE_SUBNAME      VARCHAR2(4000)
ATTRIBUTE_VALUE        VARCHAR2(4000)
COEFFICIENT            BINARY_DOUBLE
```

Table 6-237 Non-Negative Matrix Factorization H Matrix View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
FEATURE_ID	The ID of a feature in the model
FEATURE_NAME	The name of a feature in the model
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Specifies the value of attribute
COEFFICIENT	The attribute encoding that represents its contribution to the feature

The view DM$VI*model_view* describes the inverse H matrix of an NMF model. The FEATURE_NAME column type may be either NUMBER or VARCHAR2. The view has the following schema:

```
Name                 Type
-----------------    ------------------------
PARTITION_NAME       VARCHAR2(128)
FEATURE_ID           NUMBER
FEATURE_NAME         NUMBER/VARCHAR2
ATTRIBUTE_NAME       VARCHAR2(128)
ATTRIBUTE_SUBNAME    VARCHAR2(4000)
ATTRIBUTE_VALUE      VARCHAR2(4000)
COEFFICIENT          BINARY_DOUBLE
```

Table 6-238 Non-Negative Matrix Factorization Inverse H Matrix View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model

Table 6-238 (Cont.) Non-Negative Matrix Factorization Inverse H Matrix View

Column Name	Description
FEATURE_ID	The ID of a feature in the model
FEATURE_NAME	The name of a feature in the model
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
ATTRIBUTE_VALUE	Specifies the value of attribute
COEFFICIENT	The attribute encoding that represents its contribution to the feature

The following table describes the Global Name-Value Pairs view (DM$VG*model_name*) specific to NMF.

Table 6-239 Global Name-Value Pairs View for NMF

Name	Description
CONV_ERROR	Convergence error
CONVERGED	Indicates whether the model build process has converged to specified tolerance. The following are the possible values: • YES • NO
ITERATIONS	Number of iterations performed during build
NUM_ROWS	Number of rows used in the build input data set
SAMPLE_SIZE	Number of rows used by the build

Model Detail Views for Singular Value Decomposition

Model detail views specific to Singular Value Decomposition (SVD) contain information about the S matrix, right-singular vectors, and left-singular vectors.

These are the available model views for SVD:

Model Views	Description
DM$VE*model_name*	Singular Value Decomposition S Matrix
DM$VG*model_name*	Global Name-Value Pairs
DM$VN*model_name*	Normalization and Missing Value Handling
DM$VS*model_name*	Computed Settings
DM$VU*model_name*	Singular Value Decomposition U Matrix
DM$VV*model_name*	Singular Value Decomposition V Matrix
DM$VW*model_name*	Model Build Alerts

The Singular Value Decomposition S Matrix view (DM$VE*model_name*) leverages the fact that each singular value in the SVD model has a corresponding principal component in the associated Principal Components Analysis (PCA) model to relate a common set of information for both classes of models. For an SVD model, it describes the content of the S matrix. When PCA scoring is selected as a build setting, the variance and percentage cumulative variance for the corresponding principal components are shown as well. The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
FEATURE_ID                            NUMBER
FEATURE_NAME                          NUMBER/VARCHAR2
VALUE                                 BINARY_DOUBLE
VARIANCE                              BINARY_DOUBLE
PCT_CUM_VARIANCE                      BINARY_DOUBLE
```

Table 6-240 Singular Value Decomposition S Matrix View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
FEATURE_ID	The ID of a feature in the model
FEATURE_NAME	The name of a feature in the model
VALUE	The matrix entry value
VARIANCE	The variance explained by a component. This column is only present for SVD models with setting `dbms_data_mining.svds_scoring_mode` set to `dbms_data_mining.svds_scoring_pca` This column is non-null only if the build data is centered, either manually or because of the following setting:`dbms_data_mining.prep_auto` is set to `dbms_data_mining.prep_auto_on`.
PCT_CUM_VARIANCE	The percent cumulative variance explained by the components thus far. The components are ranked by the explained variance in descending order. This column is only present for SVD models with setting `dbms_data_mining.svds_scoring_mode` set to `dbms_data_mining.svds_scoring_pca` This column is non-null only if the build data is centered, either manually or because of the following setting:`dbms_data_mining.prep_auto` is set to `dbms_data_mining.prep_auto_on`.

The Singular Value Decomposition V Matrix view (`DM$VV`*model_view*) describes the right-singular vectors of an SVD model. For a PCA model it describes the principal components (eigenvectors). The view has the following columns:

```
Name                                  Type
----------------------------------    ----------------------------
PARTITION_NAME                        VARCHAR2(128)
FEATURE_ID                            NUMBER
FEATURE_NAME                          NUMBER/VARCHAR2
ATTRIBUTE_NAME                        VARCHAR2(128)
ATTRIBUTE_SUBNAME                     VARCHAR2(4000)
```

```
ATTRIBUTE_VALUE                    VARCHAR2(4000)
VALUE                              BINARY_DOUBLE
```

Table 6-241 Singular Value Decomposition V Matrix View

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`FEATURE_ID`	The ID of a feature in the model
`FEATURE_NAME`	The name of a feature in the model
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_SUBNAME`	Nested column subname. The value is null for non-nested columns.
`ATTRIBUTE_VALUE`	Categorical attribute value. For numerical attributes, `ATTRIBUTE_VALUE` is null.
`VALUE`	The matrix entry value

The Singular Value Decomposition U Matrix view (`DM$VU`*model_name*) describes the left-singular vectors of an SVD model. For a PCA model, it describes the projection of the data in the principal components. This view does not exist unless the settings `dbms_data_mining.svds_u_matrix_output` is set to `dbms_data_mining.svds_u_matrix_enable`. The view has the following columns:

```
Name                               Type
---------------------------------- ----------------------------
PARTITION_NAME                     VARCHAR2(128)
CASE_ID                            NUMBER/VARHCAR2, DATE, TIMESTAMP,
                                   TIMESTAMP WITH TIME ZONE,
                                   TIMESTAMP WITH LOCAL TIME ZONE
FEATURE_ID                         NUMBER
FEATURE_NAME                       NUMBER/VARCHAR2
VALUE                              BINARY_DOUBLE
```

Table 6-242 Singular Value Decomposition U Matrix View or Projection Data in Principal Components

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`CASE_ID`	Unique identifier of the row in the build data described by the **U** matrix projection.
`FEATURE_ID`	The ID of a feature in the model
`FEATURE_NAME`	The name of a feature in the model
`VALUE`	The matrix entry value

Global Details for Singular Value Decomposition

The following table describes the Global Name-Value Pairs view (`DM$VG`*model_name*) specific to a SVD model.

Table 6-243 Global Name-Value Pairs View for Singular Value Decomposition

Name	Description
`NUM_COMPONENTS`	Number of features (components) produced by the model
`NUM_ROWS`	The total number of rows used in the build
`SUGGESTED_CUTOFF`	Suggested cutoff that indicates how many of the top computed features capture most of the variance in the model. Using only the features below this cutoff would be a reasonable strategy for dimensionality reduction.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Model Detail Views for Minimum Description Length

Model detail views specific to Minimum Description Length (MDL) (for calculating attribute importance) contain information about attribute importance models.

These are the available model views for MDL:

Model Views	Description
`DM$VA`*model_name*	Attribute Importance
`DM$VB`*model_name*	Automatic Data Preparation Binning
`DM$VG`*model_name*	Global Name-Value Pairs
`DM$VS`*model_name*	Computed Settings
`DM$VW`*model_name*	Model Build Alerts

The Attribute Importance view (`DM$VA`*model_name*) describes the attribute importance as well as the attribute importance rank. The view has the following columns:

```
Name                                      Type
----------------------------------------- ----------------------------
PARTITION_NAME                            VARCHAR2(128)
ATTRIBUTE_NAME                            VARCHAR2(128)
ATTRIBUTE_SUBNAME                         VARCHAR2(4000)
ATTRIBUTE_IMPORTANCE_VALUE                BINARY_DOUBLE
ATTRIBUTE_RANK                            NUMBER
```

Table 6-244 Attribute Importance View for Minimum Description Length

Column Name	Description
`PARTITION_NAME`	Partition name in a partitioned model
`ATTRIBUTE_NAME`	Column name
`ATTRIBUTE_SUBNAME`	Nested column subname. The value is null for non-nested columns.
`ATTRIBUTE_IMPORTANCE_VALUE`	Importance value
`ATTRIBUTE_RANK`	Rank based on importance

The following table describes the Global Name-Value Pairs view (DM$VG*model_name*) specific to MDL.

Table 6-245 Global Name-Value Pairs View for MDL

Name	Description
NUM_ROWS	The total number of rows used in the build

Model Detail Views for Binning

The binning view DM$VB describes the bin boundaries used in automatic data preparation.

The view has the following columns:

```
Name                        Type
--------------------        --------------------
PARTITION_NAME              VARCHAR2(128)
ATTRIBUTE_NAME              VARCHAR2(128)
ATTRIBUTE_SUBNAME           VARCHAR2(4000)
BIN_ID                      NUMBER
LOWER_BIN_BOUNDARY          BINARY_DOUBLE
UPPER_BIN_BOUNDARY          BINARY_DOUBLE
ATTRIBUTE_VALUE             VARCHAR2(4000)
```

Table 6-246 Model Details View for Binning

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
ATTRIBUTE_NAME	Specifies the attribute name
ATTRIBUTE_SUBNAME	Specifies the attribute subname
BIN_ID	Bin ID (or bin identifier)
LOWER_BIN_BOUNDARY	Numeric lower bin boundary
UPPER_BIN_BOUNDARY	Numeric upper bin boundary
ATTRIBUTE_VALUE	Categorical value

Model Detail Views for Global Information

Model detail views for global information contain information about global statistics, alerts, and computed settings.

The Global Name-Value Pairs view (DM$VG*model_name*) describes global statistics related to the model build. Examples include the number of rows used in the build, the convergence status, and the model quality metrics. The view has the following columns:

```
Name                         Type
-------------------          --------------------
PARTITION_NAME               VARCHAR2(128)
NAME                         VARCHAR2(30)
```

```
NUMERIC_VALUE                NUMBER
STRING_VALUE                 VARCHAR2(4000)
```

Table 6-247 Global Name-Value Pairs View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
NAME	Name of the statistic
NUMERIC_VALUE	Numeric value of the statistic
STRING_VALUE	Categorical value of the statistic

The Model Build Alerts view (DM$VW*model_name*) lists alerts issued during the model build. The view has the following columns:

```
Name                       Type
-------------------        ----------------------
PARTITION_NAME             VARCHAR2(128)
ERROR_NUMBER               BINARY_DOUBLE
ERROR_TEXT                 VARCHAR2(4000)
```

Table 6-248 Model Build Alerts View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
ERROR_NUMBER	Error number (valid when event is Error)
ERROR_TEXT	Error message

The Computed Settings view (DM$VS*model_name*) lists the algorithm computed settings. The view has the following columns:

```
Name                        Type
-----------------           --------------------
PARTITION_NAME              VARCHAR2(128)
SETTING_NAME                VARCHAR2(30)
SETTING_VALUE               VARCHAR2(4000)
```

Table 6-249 Computed Settings View

Column Name	Description
PARTITION_NAME	Partition name in a partitioned model
SETTING_NAME	Name of the setting
SETTING_VALUE	Value of the setting

Model Detail Views for Normalization and Missing Value Handling

The Normalization and Missing Value Handling view DM$VN describes the normalization parameters used in Automatic Data Preparation (ADP) and the missing value replacement when a NULL value is encountered. Missing value replacement applies only to the two-dimensional columns and does not apply to the nested columns.

The view has the following columns:

```
Name                          Type
----------------------        -----------------------
PARTITION_NAME                VARCHAR2(128)
ATTRIBUTE_NAME                VARCHAR2(128)
ATTRIBUTE_SUBNAME             VARCHAR2(4000)
NUMERIC_MISSING_VALUE         BINARY_DOUBLE
CATEGORICAL_MISSING_VALUE     VARCHAR2(4000)
NORMALIZATION_SHIFT           BINARY_DOUBLE
NORMALIZATION_SCALE           BINARY_DOUBLE
```

Table 6-250 Normalization and Missing Value Handling View

Column Name	Description
PARTITION_NAME	A partition in a partitioned model
ATTRIBUTE_NAME	Column name
ATTRIBUTE_SUBNAME	Nested column subname. The value is null for non-nested columns.
NUMERIC_MISSING_VALUE	Numeric missing value replacement
CATEGORICAL_MISSING_VALUE	Categorical missing value replacement
NORMALIZATION_SHIFT	Normalization shift value
NORMALIZATION_SCALE	Normalization scale value

Model Detail Views for Exponential Smoothing

Model detail views specific to Exponential Smoothing (ESM) include information about the model output, global information about the model, and views that support time series regression.

These are the available model views for ESM:

Model Details	Description
DM$VG*model_name*	Global Name-Value Pairs
DM$VP*model_name*	Exponential Smoothing Forecast
DM$VS*model_name*	Computed Settings
DM$VW*model_name*	Model Build Alerts
DM$VR*model_name*	Time Series Regression Build
DM$VT*model_name*	Time Series Regression Score

Exponential Smoothing Forecast view (DM$VP*model_name*) displays the outcome of an ESM model. The output contains a set of records, ordered by partition and CASE_ID, that include the columns given in the *Exponential Smoothing Model Output* table. CASE_ID identifies the value's position in the time series. The user-specified CASE_ID can be a type that represents a numerical or datetime value. For each unique value of PARTITION, a distinct exponential smoothing model is built. The VALUE column for each PARTITION represents the observed or accumulated value of the target at that point in the sequence. The PREDICTION column is the forecast one step ahead at that point in the sequence. Backcasts are predictions that fall inside the range of the input data. The sequence also includes a user-specified number of values beyond the range of the input data. The VALUE column is *NULL* for any sequence value outside the range of input, and PREDICTION column is the model forecast for that sequence value. Lower and upper boundaries of the forecasts are denoted by the LOWER and UPPER columns. For backcasts, LOWER and UPPER are *NULL*. The bounds are based on a confidence interval that the user sets for the prediction.

Table 6-251 Exponential Smoothing Forecast View

Name	Description
PARTITION	Partition name in a partitioned model
CASE_ID	Sequence identifier (datetime or number type)
VALUE	Observed or accumulated value
PREDICTION	Backcast or Forecast value
UPPER	Upper bound of the forecast
LOWER	Lower bound of the forecast

Global Name-Value Pairs view (DM$VG*model_name*) includes the model's global information as well as the estimated smoothing constants, estimated initial state, and global diagnostic measures.

Depending on the type of model, the global diagnostics include some or all of the following for Exponential Smoothing.

Table 6-252 Global Name-Value Pairs View for ESM

Name	Description
-2 LOG-LIKELIHOOD	Negative log-likelihood of model
ALPHA	Smoothing constant
AIC	Akaike information criterion
AICC	Corrected Akaike information criterion
AMSE	Average mean square error over user-specified time window
BETA	Trend smoothing constant
BIC	Bayesian information criterion
GAMMA	Seasonal smoothing constant
INITIAL LEVEL	Model estimate of value one time interval prior to start of observed series
INITIAL SEASON i	Model estimate of seasonal effect for season *i* one time interval prior to start of observed series
INITIAL TREND	Model estimate of trend one time interval prior to start of observed series
MAE	Model mean absolute error
MSE	Model mean square error

Table 6-252 (Cont.) Global Name-Value Pairs View for ESM

Name	Description
`PHI`	Damping parameter
`STD`	Model standard error
`SIGMA`	Model standard deviation of residuals

Time series regression expands the features that can be included in a time series model and, possibly, increases forecast accuracy. Backcasts and forecasts of time series correlated to the "target" series of interest are included in the build and score views. The build and score views can be fed into a regression technique like Generalized Linear Model.

The Time Series Regression Build view (`DM$VR`*model_name*) depicts the schema for the build view. Each predictor series will have its own column. There can be a maximum of 20 predictor series in the build and score views. The names of the columns are obtained from the `EXSM_SERIES_LIST` setting.

Table 6-253 Time Series Regression Build View

Name	Description
`PARTITION`	Partition name in a partitioned model
`CASE_ID`	Sequence identifier (datetime or number type)
`target series name`	Observed or accumulated value of target series
`DM$`*`target series`*	Backcasted value of target series
`DM$`*`predictor series column name`*	Backcasted value of predictor series column. A maximum of *20* predictor series columns can be used.

The Time Series Regression Score view (`DM$VT`*model_name*) shows the schema for the score view. The schema is the same as in the build view, but the values in the *target series name* column are `NULL` because the future has not yet been observed.

Table 6-254 Time Series Regression Score View

Name	Description
`PARTITION`	Partition name in a partitioned model
`CASE_ID`	Sequence identifier (datetime or number type)
`target series name`	`NULL`s, because the future values of the target series have not been observed
`DM$`*`target series`*	Forecasted value of target series
`DM$`*`predictor series column name`*	Forecasted value of predictor series column name. A maximum of *20* predictor series columns can be used.

Related Topics

- About Exponential Smoothing
- About Generalized Linear Models

Model Detail Views for Text Features

The model details view for text features is DM$VX*model_name*.

The text feature view DM$VX*model_name* describes the extracted text features if there are text attributes present. The view has the following schema:

```
Name                                    Type
 ---------------                    -----------------------
 PARTITION_NAME                         VARCHAR2(128)
 COLUMN_NAME                            VARCHAR2(128)
 TOKEN                                  VARCHAR2(4000)
 DOCUMENT_FREQUENCY                     NUMBER
```

Table 6-255 Text Feature View for Extracted Text Features

Column Name	Description
PARTITION_NAME	A partition in a partitioned model to retrieve details
COLUMN_NAME	Name of the identifier column
TOKEN	Text token which is usually a word or stemmed word
DOCUMENT_FREQUENCY	A measure of token frequency in the entire training set

Model Detail Views for ONNX Models

You can view the details of an embedding model using the model detail views. The names of the views begin with DM$V.

This section lists the model detail views for embedding models.

DM$VJ Model Detail View

The DM$VJ<*model-name*> returns a single row containing a JSON object in one column that contains user-specified metadata of the model.

The view has the following columns:

```
Name                          Null?     Type
-------------------------- -------- ------------------------------
 METADATA                               CLOB
```

Column Name	Description
METADATA	It is a CLOB containing the user-specified metadata of the embedding model in JSON format.

The following table describes the output of the DM$VJ<*modle_name*> view of an embedding model.

Name	Value
`METADATA`	The JSON that was specified to the `IMPORT_ONNX_MODEL` call for importing the model.

The following example displays the output of an embedding model. The name of the model is *`doc_model`*:

```
select * from DM$VJdoc_model;
```

The output is as follows:

```
METADATA
------------------------------------------------------------------------------
--
{"function":"embedding","embeddingOutput":"embedding","input":{"input":
["DATA"]}}
```

DM$VM Model Detail View

The `DM$VM`*`<model-name>`* view reports information extracted from the metadata of the imported ONNX model and its input or output tensors.

The view has the following columns:

```
Name                                   Type
---------------------------------- ---------------------------
NAME                                VARCHAR2(4000)
VALUE                                  VARCHAR2(4000)
```

Table 6-256

Column Name	Description
`NAME`	The name of the metadata extracted from the ONNX model.
`VALUE`	Indicates a value for the metadata name

The following table describes the output of the `DM$VM`*`<model_name>`* view of an embedding model.

Name	Value
Producer Name	Name of the tools that generated the ONNX files
Graph Name	Name of the ONNX graph
Graph Description	Description given to the model
Version	Version of the model
Input	Describes the model input mapping
Output	Reports the vector information with dimension and value type

The following example displays the output of an embedding model. The name of the model is *DOC_MODEL*:

```
select * from DM$VMdoc_model;
```

The following is the output:

```
NAME                                    VALUE
----------------------------------------
----------------------------------------
Producer Name                           onnx.compose.merge_models
Graph Name                              g_8_main_graph_main_graph
Graph Description                       Graph combining g_8_main_graph and
main_

                                        graph
                                        g_8_main_graph



                                        main_graph



Version                                 1
Input[0]                                input:string[1]
Output[0]                               embedding:float32[?,384]

6 rows selected.
```

Related Topics

- https://github.com/onnx/onnx/blob/main/docs/IR.md

DM$VP Model Detail View

The DM$VP*<model-name>* view displays information extracted from parsing the JSON metadata. The view presents the JSON metadata of the model, including both explicitly declared properties and system-assigned default values for undeclared ones.

The reported properties are specific to the machine learning model and match the mandatory and optional fields of the JSON metadata.

The view has the following columns:

```
Name                                   Type
----------------------------------  ---------------------------
NAME                                VARCHAR2(4000)
VALUE                                  VARCHAR2(4000)
```

Column Name	Description
NAME	Displays the JSON parameters
VALUE	Indicates the value corresponding to the JSON parameter name value pair

Note that this information is already available in the `ALL_MINING_MODEL_ATTRIBUTES` view. The following example displays all the columns available to you in the `DM$VP`*`doc_model`* view of an embedding model. In this example, *doc_model* is the name of the model.

```
select * from DM$VPdoc_model;
```

The output is as follows:

```
NAME                                      VALUE
----------------------------------------
----------------------------------------
batching                                  False
embeddingOutput                           embedding
```

7

Concepts

Provides an overview of concepts related to Oracle Machine Learning and Oracle Machine Learning for SQL.

- Machine Learning Techniques
- Machine Learning Algorithms
- Integration of ONNX Runtime
- Machine Learning and Statistics
- Overview of Database Analytics
- Oracle Machine Learning and Analytic Views
- Oracle Machine Learning and Data Warehousing

Machine Learning Techniques

Provides an overview of concepts related to Oracle Machine Learning techniques.

Topics:

- About Anomaly Detection
- About Association
- About Classification
- About Clustering
- Embedding
- About Feature Extraction
- Finding the Attributes
- Integration of ONNX Runtime
- About Ranking
- About Regression
- About Row Importance
- About Time Series

About Anomaly Detection

Identify unusual items or events in seemingly normal data to detect fraud, network intrusions, and other rare, significant occurrences through Anomaly Detection.

The goal of anomaly detection is to identify items, events, or observations that are unusual within data that is seemingly 'normal'. This data may consist of traditional enterprise data or Internet of Things (IoT) sensor data. Anomaly detection is an important tool for detecting, for example, fraud, network intrusions, enterprise computing service interruptions, sensor time series prognostics, and other rare events that can have great significance but are hard to find. Anomaly detection can be used to solve problems like the following:

- A law enforcement agency compiles data about unpermitted activities, but nothing about legitimate activities. How can a suspicious activity be flagged?

  The law enforcement data is all of one class. There are no counter-examples.
- An insurance agency processes millions of insurance claims, knowing that a very small number are fraudulent. How can the fraudulent claims be identified?

  The claims data contains very few counter-examples. They are outliers.
- An IT department encounters compute resource performance anomalies. How can such anomalies be detected along with their source causes, such as resource-contention issues and complex memory leaks?

  The data contains sensor output from thousands of sensors.
- An oil and gas enterprise or utility company requires proactive maintenance of business-critical assets, such as oil rigs or smart meters, to reduce operations and maintenance costs, improve up-time of revenue-generating assets, and improve safety margins for life-critical systems.

Anomaly Detection as a form of One-Class Classification

Anomaly detection predicts whether a data point is typical for a given distribution or not. Atypical data points can be outliers or new classes. Traditional data should only have one class, hence anomaly detection is a one-class classification.

Normally, a classification model must be trained on data that includes both examples and counterexamples for each class so that the model can learn to distinguish between them. For example, a model that predicts the side effects of a medication must be trained on data that includes a wide range of responses to the medication.

A one-class classifier develops a profile that generally describes a typical case in the training data. Deviation from the profile is identified as an anomaly. One-class classifiers are sometimes referred to as positive security models, because they seek to identify "good" behaviors and assume that all other behaviors are bad.

In single-class data, all the cases have the same classification. Counterexamples, instances of another class, are hard to specify or expensive to collect. For instance, in text document classification, it is easy to classify a document under a given topic. However, the universe of documents outside of this topic can be very large and diverse. Thus, it is not feasible to specify other types of documents as counterexamples. Anomaly detection can be used to find unusual instances of a particular type of document.

> **Note:**
>
> Solving a one-class classification problem can be difficult. The accuracy of one-class classifiers cannot usually match the accuracy of standard classifiers built with meaningful counter examples.
>
> The goal of this type of anomaly detection is to provide some useful information where no information was previously attainable. However, if there are enough of the "rare" cases so that stratified sampling produces a training set with enough counterexamples for a standard classification model, then the classification may be a better solution.

Related Topics

- About Classification
  Classification is a machine learning technique that assigns items to target categories or classes to predict outcomes.

Anomaly Detection for Time Series Data

Identify anomalies in time series data from numerous sensors, essential for early detection in critical enterprise systems.

With the growing number of sensors in the internet of things, the ability to identify anomalous events among potentially thousands of sensors is essential. For example, in the early detection of anomalies in business-critical enterprise computing servers and software systems, storage systems, and networks. Enterprises require high anomaly detection accuracy, which implies lower false-alarm probabilities, lower missed-alarm probabilities, and lower overhead compute cost. The ability to distinguish between a real problem and sensor malfunction can significantly reduce costs in problem solution.

Building a model involves supplying historical, error-free operating data from, for example, monitored equipment. The resulting model is used to score new sensor data, also referred to as the *monitoring phase*, to estimate the expected sensor values.

Anomaly Detection Algorithms

For anomaly detection, Oracle Machine Learning for SQL has the following algorithms.

- Multivariate state Estimation Technique - Sequential Probability Ratio Test (MSET-SPRT)
- One-Class Support Vector Machine (SVM)
- Expectation Maximization (EM) Anomaly

Anomaly detection is a form of classification. When you create a model using the MSET-SPRT and One-Class SVM and EM Anomaly algorithms, specify the classification machine learning technique. These algorithms do not use a target.

About Association

Identify the probability of co-occurring items in a collection using Association.

The relationships between co-occurring items are expressed as **Association Rules**.

Association Rules

Identify the probability of co-occurring items in a collection within the data.

The results of an association model are the rules that identify patterns of association within the data. Oracle Machine Learning for SQL does not support the scoring operation for association modeling.

Association rules can be applied as follows:

Support: How often do these items occur together in the data?
Confidence: How frequently the consequent occurs in transactions that contain the antecedent.
Value: How much business value is connected to item associations

Market-Basket Analysis

Use association rules to analyze sales transactions, such as customers frequently buying cereal and milk together.

Association rules are often used to analyze sales transactions. For example, it is noted that customers who buy cereal at the grocery store often buy milk at the same time. In fact, association analysis find that 85% of the checkout sessions that include cereal also include milk. This relationship can be formulated as the following rule:

```
Cereal implies milk with 85% confidence
```

This application of association modeling is called **market-basket analysis**. It is valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalog design, and cross-sell.

Use Market Basket Data

Understand the use of association and Apriori for market basket analysis.

Market basket data identifies the items sold in a set of baskets or transactions. Oracle Machine Learning for SQL provides the association machine learning function for market basket analysis.

Association models use the Apriori algorithm to generate association rules that describe how items tend to be purchased in groups. For example, an association rule can assert that people who buy peanut butter are 80% likely to also buy jelly.

Market basket data is usually **transactional**. In transactional data, a case is a transaction and the data for a transaction is stored in multiple rows. OML4SQL association models can be built on transactional data or on single-record case data. The `ODMS_ITEM_ID_COLUMN_NAME` and `ODMS_ITEM_VALUE_COLUMN_NAME` settings specify whether the data for association rules is in transactional format.

> **Note:**
>
> Association models are the only type of model that can be built on native transactional data. For all other types of models, OML4SQL requires that the data be presented in single-record case format.

The Apriori algorithm assumes that the data is transactional and that it has many missing values. Apriori interprets all missing values as sparse data, and it has its own native mechanisms for handling sparse data.

> **See Also:**
>
> *Oracle Database PL/SQL Packages and Types Reference* for information on the `ODMS_ITEM_ID_COLUMN_NAME` and `ODMS_ITEM_VALUE_COLUMN_NAME` settings.

Example: Creating a Nested Column for Market Basket Analysis

The example shows how to define a nested column for market basket analysis.

Association models can be built on native transactional data or on nested data. The following example shows how to define a nested column for market basket analysis.

The following SQL statement transforms this data to a column of type `DM_NESTED_NUMERICALS` in a view called `SALES_TRANS_CUST_NESTED`. This view can be used as a case table for machine learning.

```
CREATE VIEW sales_trans_cust_nested AS
             SELECT trans_id,
                     CAST(COLLECT(DM_NESTED_NUMERICAL(
                     prod_name, 1))
                     AS DM_NESTED_NUMERICALS) custprods
                  FROM sales_trans_cust
             GROUP BY trans_id;
```

This query returns two rows from the transformed data.

```
SELECT * FROM sales_trans_cust_nested
               WHERE trans_id < 101000
               AND trans_id > 100997;
```

The output is as follows:

```
TRANS_ID  CUSTPRODS(ATTRIBUTE_NAME, VALUE)
-------  ------------------------------------------------
100998   DM_NESTED_NUMERICALS
          (DM_NESTED_NUMERICAL('O/S Documentation Set - English', 1)
100999   DM_NESTED_NUMERICALS
          (DM_NESTED_NUMERICAL('CD-RW, High Speed Pack of 5', 1),
           DM_NESTED_NUMERICAL('External 8X CD-ROM', 1),
           DM_NESTED_NUMERICAL('SIMM- 16MB PCMCIAII card', 1))
```

Example 7-1 Convert to a Nested Column

The view `SALES_TRANS_CUST` provides a list of transaction IDs to identify each market basket and a list of the products in each basket.

```
describe sales_trans_cust
```

The output is as follows:

```
 Name                                                  Null?    Type
 ----------------------------------------------------- --------
----------------
 TRANS_ID                                              NOT NULL NUMBER
 PROD_NAME                                             NOT NULL VARCHAR2(50)
 QUANTITY                                                       NUMBER
```

Related Topics

- Handle Missing Values
  Understand sparse data and missing values.

Association Rules and eCommerce

Apply association rules in eCommerce to personalize web pages by predicting user behavior based on page visits.

Association modeling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for Web page personalization. An association model might find that a user who visits pages A and B is 70% likely to also visit page C in the same session. Based on this rule, a dynamic link can be created for users who are likely to be interested in page C. The association rule is expressed as follows:

```
A and B imply C with 70% confidence
```

Related Topics

- Confidence
  The confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction.

Use Retail Data for Analysis

Retail analysis often makes use of association rules and association models.

The association rules are enhanced to calculate aggregates along with rules or itemsets.

Related Topics

- *Oracle Machine Learning for SQL Concepts*

Example: Calculating Aggregates

This example shows how to calculate aggregates using the customer grocery purchase and profit data.

Calculating Aggregates for Grocery Store Data

Assume a grocery store has the following data:

Table 7-1 Grocery Store Data

Customer	Item A	Item B	Item C	Item D
Customer 1	Buys (Profit $5.00)	Buys (Profit $3.20)	Buys (Profit $12.00)	NA
Customer 2	Buys (Profit $4.00)	NA	Buys (Profit $4.20)	NA
Customer 3	Buys (Profit $3.00)	Buys (Profit $10.00)	Buys (Profit $14.00)	Buys (Profit $8.00)
Customer 4	Buys (Profit $2.00)	NA	NA	Buys (Profit $1.00)

The basket of each customer can be viewed as a transaction. The manager of the store is interested in not only the existence of certain association rules, but also in the aggregated profit if such rules exist.

In this example, one of the association rules can be (A, B)=>C for customer 1 and customer 3. Together with this rule, the store manager may want to know the following:

- The total profit of item A appearing in this rule
- The total profit of item B appearing in this rule
- The total profit for consequent C appearing in this rule
- The total profit of all items appearing in the rule

For this rule, the profit for item A is $5.00 + $3.00 = $8.00, for item B the profit is $3.20 + $10.00 = $13.20, for consequent C, the profit is $12.00 + $14.00 = $26.00, for the antecedent itemset (A, B) is $8.00 + $13.20 = $21.20. For the whole rule, the profit is $21.20 + $26.00 = $47.40.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

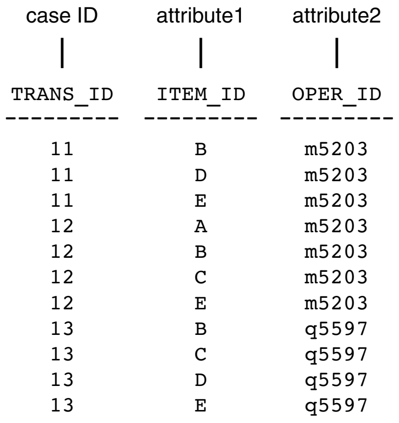
Transactional Data

Understand transactional data, where a case includes a collection of items like a market basket at checkout.

Unlike other machine learning functions, association is transaction-based. In transaction processing, a case includes a collection of items such as the contents of a market basket at the checkout counter. The collection of items in the transaction is an attribute of the transaction. Other attributes might be a timestamp or user ID associated with the transaction.

Transactional data, also known as **market-basket data**, is said to be in **multi-record case** format because a set of records (rows) constitute a case. For example, in the following figure, case 11 is made up of three rows while cases 12 and 13 are each made up of four rows.

Figure 7-1   Transactional Data


case ID
attribute1
attribute2

```
TRANS_ID    ITEM_ID     OPER_ID
---------   ---------   ---------
   11          B         m5203
   11          D         m5203
   11          E         m5203
   12          A         m5203
   12          B         m5203
   12          C         m5203
   12          E         m5203
   13          B         q5597
   13          C         q5597
   13          D         q5597
   13          E         q5597
```

Non transactional data is said to be in a **single-record case** format because a single record (row) constitutes a case. In Oracle Machine Learning, association models can be built using either transactional or non transactional or two-dimensional data formats. If the data is non transactional, it is possible to transform to a nested column to make it transactional before association machine learning activities can be performed. Transactional format is the usual format but, the association rules model does accept two-dimensional input format. For non transactional input format, each distinct combination of the content in all columns other than the case ID column is treated as a unique item.

Related Topics

- *Oracle Machine Learning for SQL User's Guide*
- Data Preparation for Apriori
  Prepare transactional data for Apriori by organizing it into case identifiers and associated values, for model processing.

Association Algorithm

Oracle Machine Learning for SQL uses the Apriori algorithm to calculate association rules for items in frequent itemsets.

Related Topics

- About Apriori
  Learn how to find associations involving rare events in a large number of items using Apriori.

About Classification

Classification is a machine learning technique that assigns items to target categories or classes to predict outcomes.

The goal of **classification** is to accurately predict the target class for each case in the data. For example, a classification model can be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk can be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating is the target, the other attributes are the predictors, and the data for each customer constitutes a case.

Classification are discrete and do not imply order. Continuous, floating-point values indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.

Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

Applying a classification model results in class assignments and probabilities for each case. For example, a model that classifies customers as low, medium, or high value also predicts the probability of each classification for each customer.

Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

Testing a Classification Model

Apply the classification model to test data, compare predictions with actual outcomes, and evaluate accuracy using test metrics.

A classification model is tested by applying it to test data with known target values and comparing the predicted values with the known values.

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the **build data** and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model.

Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

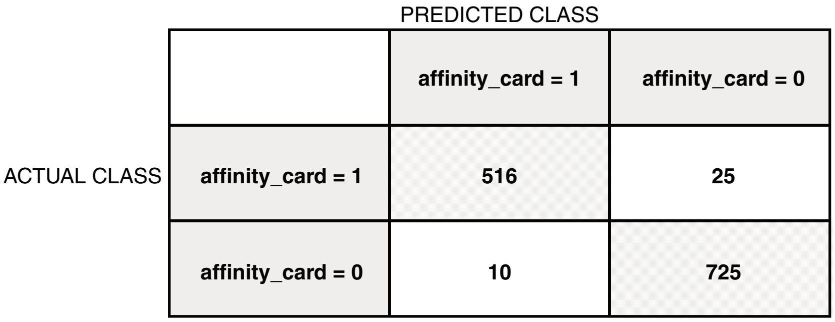
Confusion Matrix

Display the number of correct and incorrect predictions compared to actual classifications using a confusion matrix.

A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is *n*-by-*n*, where *n* is the number of classes.

The following figure shows a confusion matrix for a binary classification model. The rows present the number of actual classifications in the test data. The columns present the number of predicted classifications made by the model.

Figure 7-2 Confusion Matrix for a Binary Classification Model



		PREDICTED CLASS	
		affinity_card = 1	**affinity_card = 0**
ACTUAL CLASS	**affinity_card = 1**	**516**	**25**
	affinity_card = 0	**10**	**725**

In this example, the model correctly predicted the positive class (also called true positive (TP)) for `affinity_card` 516 times and incorrectly predicted (also called false negative (FN)) it 25 times. The model correctly predicted the negative class (also called true negative (TN)) for `affinity_card` 725 times and incorrectly predicted (also called false positive (FP)) it 10 times. The following can be computed from this confusion matrix:

- The model made 1241 correct predictions, that is, TP + TN, (516 + 725).
- The model made 35 incorrect predictions, that is, FN + FP, (25 + 10).
- There are 1276 total scored cases, (516 + 25 + 10 + 725).
- The error rate is 35/1276 = 0.0274. (FN+FP/Total)
- The overall accuracy rate is 1241/1276 = 0.9725 (TP+TN)/Total).

Precision and Recall

Consider the same example, the accuracy rate shows 0.97. However, there are cases where the model has incorrectly predicted. **Precision** (positive predicted value) is the ability of a classification model to return only relevant cases. Precision can be calculated as TP/TP+FP. **Recall** (sensitivity or true positive rate) is the ability of a classification model to return relevant cases. Recall can be calculated as TP/TP+FN. The precision in this example is 516/526 = 0.98. The recall in this example is 516/541 = 0.95. Ideally, the model is good when both precision and recall are 1. This can happen when the numerator and the denominator are equal. That means, for precision, FP is zero and for recall, FN is zero.

Lift

Lift measures the degree to which the predictions of a classification model are better than randomly-generated predictions.

Lift applies to binary classification only, and it requires the designation of a positive class. If the model itself does not have a binary target, you can compute lift by designating one class as positive and combining all the other classes together as one negative class.

Numerous statistics can be calculated to support the notion of lift. Basically, lift can be understood as a ratio of two percentages: the percentage of correct positive classifications made by the model to the percentage of actual positive classifications in the test data. For example, if 40% of the customers in a marketing survey have responded favorably (the positive classification) to a promotional campaign in the past and the model accurately predicts 75% of them, the lift is obtained by dividing .75 by .40. The resulting lift is 1.875.

Lift is computed against quantiles that each contain the same number of cases. The data is divided into quantiles after it is scored. It is ranked by probability of the positive class from highest to lowest, so that the highest concentration of positive predictions is in the top quantiles. A typical number of quantiles is 10.

Lift is commonly used to measure the performance of response models in marketing applications. The purpose of a response model is to identify segments of the population with potentially high concentrations of positive responders to a marketing campaign. Lift reveals how much of the population must be solicited to obtain the highest percentage of potential responders.

Related Topics

- Positive and Negative Classes
  Identify and prioritize positive and negative classes in a confusion matrix, crucial for computing Lift and ROC metrics.

Lift Statistics

Oracle Machine Learning computes various lift statistics to evaluate the effectiveness of classification models in predicting positive outcomes.

Oracle Machine Learning computes the following lift statistics:

- **Probability threshold** for a quantile *n* is the minimum probability for the positive target to be included in this quantile or any preceding quantiles (quantiles *n*-1, *n*-2,..., 1). If a cost matrix is used, a cost threshold is reported instead. The cost threshold is the maximum cost for the positive target to be included in this quantile or any of the preceding quantiles.
- **Cumulative gain** is the ratio of the cumulative number of positive targets to the total number of positive targets.
- **Target density** of a quantile is the number of true positive instances in that quantile divided by the total number of instances in the quantile.
- **Cumulative target density** for quantile *n* is the target density computed over the first *n* quantiles.
- **Quantile lift** is the ratio of the target density for the quantile to the target density over all the test data.
- **Cumulative percentage of records** for a quantile is the percentage of all cases represented by the first *n* quantiles, starting at the end that is most confidently positive, up to and including the given quantile.
- **Cumulative number of targets** for quantile *n* is the number of true positive instances in the first *n* quantiles.
- **Cumulative number of nontargets** is the number of actually negative instances in the first *n* quantiles.
- **Cumulative lift** for a quantile is the ratio of the cumulative target density to the target density over all the test data.

Related Topics

- Costs
  Influence model decisions by specifying a cost matrix to minimize costly misclassifications.

Receiver Operating Characteristic (ROC)

ROC is a metric for comparing predicted and actual target values in a classification model.

ROC, like Lift, applies to binary classification and requires the designation of a positive class.

You can use ROC to gain insight into the decision-making ability of the model. How likely is the model to accurately predict the negative or the positive class?

ROC measures the impact of changes in the **probability threshold**. The probability threshold is the decision point used by the model for classification. The default probability threshold for binary classification is 0.5. When the probability of a prediction is 50% or more, the model predicts that class. When the probability is less than 50%, the other class is predicted. (In multiclass classification, the predicted class is the one predicted with the highest probability.)

Related Topics

- Positive and Negative Classes
  Identify and prioritize positive and negative classes in a confusion matrix, crucial for computing Lift and ROC metrics.

The ROC Curve

Plot ROC as a curve on an X-Y axis, showing true positive rate versus false positive rate.

ROC can be plotted as a curve on an X-Y axis. The **false positive rate** is placed on the X axis. The **true positive rate** is placed on the Y axis.

The top left corner is the optimal location on an ROC graph, indicating a high true positive rate and a low false positive rate.

Area Under the Curve

Measure model discrimination ability with the area under the ROC curve (AUC); higher AUC indicates better performance.

The area under the ROC curve (AUC) measures the discriminating ability of a binary classification model. The larger the AUC, the higher the likelihood that an actual positive case is assigned, and a higher probability of being positive than an actual negative case. The AUC measure is especially useful for data sets with unbalanced target distribution (one target class dominates the other).

ROC and Model Bias

Adjust probability thresholds to optimize accuracy or favor specific classes, using ROC to understand model bias.

The ROC curve for a model represents all the possible combinations of values in its confusion matrix. Changes in the probability threshold affect the predictions made by the model. For instance, if the threshold for predicting the positive class is changed from 0.5 to 0.6, then fewer positive predictions are made. This affects the distribution of values in the confusion matrix: the number of true and false positives and true and false negatives differ.

You can use ROC to find the probability thresholds that yield the highest overall accuracy or the highest per-class accuracy. For example, if it is important to you to accurately predict the positive class, but you don't care about prediction errors for the negative class, then you can lower the threshold for the positive class. This can bias the model in favor of the positive class.

A cost matrix is a convenient mechanism for changing the probability thresholds for model scoring.

Related Topics

- Costs
  Influence model decisions by specifying a cost matrix to minimize costly misclassifications.

ROC Statistics

Calculate ROC statistics to evaluate classification model performance, including true positives, false positives, and probability thresholds.

Oracle Machine Learning computes the following ROC statistics:

- **Probability threshold:** The minimum predicted positive class probability resulting in a positive class prediction. Different threshold values result in different hit rates and different false alarm rates.
- **True negatives:** Negative cases in the test data with predicted probabilities strictly less than the probability threshold (correctly predicted).
- **True positives:** Positive cases in the test data with predicted probabilities greater than or equal to the probability threshold (correctly predicted).
- **False negatives:** Positive cases in the test data with predicted probabilities strictly less than the probability threshold (incorrectly predicted).
- **False positives:** Negative cases in the test data with predicted probabilities greater than or equal to the probability threshold (incorrectly predicted).
- **True positive fraction**: Hit rate. (true positives/(true positives + false negatives))

- **False positive fraction:** False alarm rate. (false positives**/**(false positives + true negatives))

Biasing a Classification Model

Adjust decision criteria using costs, prior probabilities, and class weights to influence model predictions.

Costs, prior probabilities, and class weights are methods for biasing classification models.

Costs

Influence model decisions by specifying a cost matrix to minimize costly misclassifications.

A cost matrix is a mechanism for influencing the decision making of a model. A cost matrix can cause the model to minimize costly misclassifications. It can also cause the model to maximize beneficial accurate classifications.

For example, if a model classifies a customer with poor credit as low risk, this error is costly. A cost matrix can bias the model to avoid this type of error. The cost matrix can also be used to bias the model in favor of the correct classification of customers who have the worst credit history.

ROC is a useful metric for evaluating how a model behaves with different probability thresholds. You can use ROC to help you find optimal costs for a given classifier given different usage scenarios. You can use this information to create cost matrices to influence the deployment of the model.

Costs Versus Accuracy

Compare cost and confusion matrices to evaluate model quality, considering both prediction accuracy and the relative importance of different predictions.

Like a confusion matrix, a cost matrix is an *n*-by-*n* matrix, where *n* is the number of classes. Both confusion matrices and cost matrices include each possible combination of actual and predicted results based on a given set of test data.

A confusion matrix is used to measure accuracy, the ratio of correct predictions to the total number of predictions. A cost matrix is used to specify the relative importance of accuracy for different predictions. In most business applications, it is important to consider costs in addition to accuracy when evaluating model quality.

Related Topics

- Confusion Matrix
  Display the number of correct and incorrect predictions compared to actual classifications using a confusion matrix.

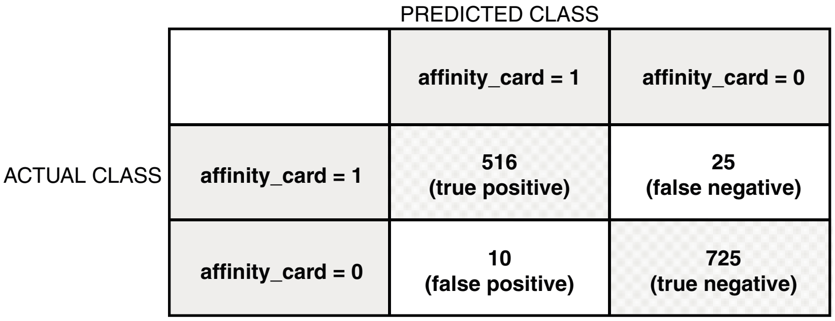
Positive and Negative Classes

Identify and prioritize positive and negative classes in a confusion matrix, crucial for computing Lift and ROC metrics.

The positive class is the class that you care the most about. Designation of a positive class is required for computing Lift and ROC.

In the confusion matrix, in the following figure, the value `1` is designated as the positive class. This means that the creator of the model has determined that it is more important to accurately predict customers who increase spending with an affinity card (`affinity_card`=1) than to accurately predict non-responders (`affinity_card`=0). If you give affinity cards to some customers who are not likely to use them, there is little loss to the company since the cost of the cards is low. However, if you overlook the customers who are likely to respond, you miss the opportunity to increase your revenue.

Figure 7-3 Positive and Negative Predictions



		PREDICTED CLASS	
		affinity_card = 1	affinity_card = 0
ACTUAL CLASS	affinity_card = 1	516 (true positive)	25 (false negative)
	affinity_card = 0	10 (false positive)	725 (true negative)

The true and false positive rates in this confusion matrix are:

- False positive rate — 10/(10 + 725) =.01
- True positive rate — 516/(516 + 25) =.95

Related Topics

- Lift
  Lift measures the degree to which the predictions of a classification model are better than randomly-generated predictions.
- Receiver Operating Characteristic (ROC)
  ROC is a metric for comparing predicted and actual target values in a classification model.

Assigning Costs and Benefits

Use a cost matrix to influence model decisions by assigning costs to negative outcomes and benefits to positive outcomes.

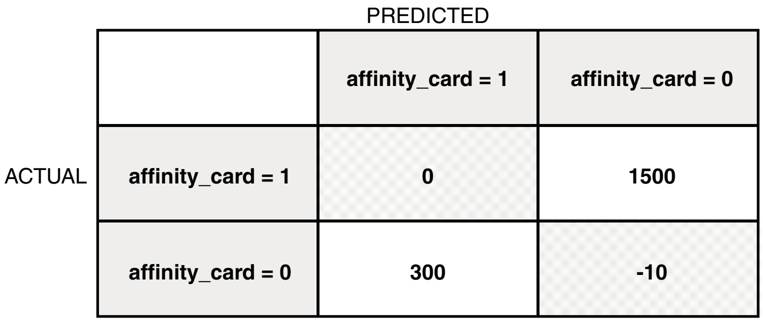
In a cost matrix, positive numbers (costs) can be used to influence negative outcomes. Since negative costs are interpreted as benefits, negative numbers (benefits) can be used to influence positive outcomes.

Suppose you have calculated that it costs your business $1500 when you do not give an affinity card to a customer who can increase spending. Using the model with the confusion matrix shown in Figure 7-3, each false negative (misclassification of a responder) costs $1500. Misclassifying a non-responder is less expensive to your business. You estimate that each false positive (misclassification of a non-responder) only costs $300.

You want to keep these costs in mind when you design a promotion campaign. You estimate that it costs $10 to include a customer in the promotion. For this reason, you associate a benefit of $10 with each true negative prediction, because you can eliminate those customers from your promotion. Each customer that you eliminate represents a savings of $10. In your cost matrix, you specify this benefit as -10, a negative cost.

The following figure shows how you would represent these costs and benefits in a cost matrix:

Figure 7-4   Cost Matrix Representing Costs and Benefits



		PREDICTED	
		affinity_card = 1	affinity_card = 0
ACTUAL	affinity_card = 1	0	1500
	affinity_card = 0	300	-10

With Oracle Machine Learning you can specify costs to influence the scoring of any classification model. Decision Tree models can also use a cost matrix to influence the model build.

Specify Costs

Specify a cost matrix table to build a Decision Tree model.

The `CLAS_COST_TABLE_NAME` setting specifies the name of a cost matrix table to be used in building a Decision Tree model. A cost matrix biases a classification model to minimize costly misclassifications. The cost matrix table must have the columns shown in the following table:

Table 7-2   Cost Matrix Table Required Columns

Column Name	Data Type
`actual_target_value`	valid target data type
`predicted_target_value`	valid target data type
`cost`	`NUMBER`

Decision Tree is the only algorithm that supports a cost matrix at build time. However, you can create a cost matrix and associate it with any classification model for scoring.

If you want to use costs for scoring, create a table with the columns shown in Table 7-2, and use the `DBMS_DATA_MINING.ADD_COST_MATRIX` procedure to add the cost matrix table to the model. You can also specify a cost matrix inline when invoking a `PREDICTION` function. Table 1-5 has details for valid target data types.

Related Topics

- *Oracle Machine Learning for SQL Concepts*

Cost-Sensitive Decision Making

Costs are user-specified numbers that bias classification. The algorithm uses positive numbers to penalize more expensive outcomes over less expensive outcomes. Higher numbers indicate higher costs.

The algorithm uses negative numbers to favor more beneficial outcomes over less beneficial outcomes. Lower negative numbers indicate higher benefits.

All classification algorithms can use costs for scoring. You can specify the costs in a cost matrix table, or you can specify the costs inline when scoring. If you specify costs inline and the model also has an associated cost matrix, only the inline costs are used. The `PREDICTION`, `PREDICTION_SET`, and `PREDICTION_COST` functions support costs.

Only the Decision Tree algorithm can use costs to bias the model build. If you want to create a Decision Tree model with costs, create a cost matrix table and provide its name in the `CLAS_COST_TABLE_NAME` setting for the model. If you specify costs when building the model, the cost matrix used to create the model is used when scoring. If you want to use a different cost matrix table for scoring, first remove the existing cost matrix table then add the new one.

A sample cost matrix table is shown in the following table. The cost matrix specifies costs for a binary target. The matrix indicates that the algorithm must treat a misclassified 0 as twice as costly as a misclassified 1.

Table 7-3 Sample Cost Matrix

ACTUAL_TARGET_VALUE	PREDICTED_TARGET_VALUE	COST
0	0	0
0	1	2
1	0	1
1	1	0

Example 7-2 Sample Queries With Costs

The table `nbmodel_costs` contains the cost matrix described in Table 7-3.

```
SELECT * from nbmodel_costs;
```

The output is as follows:

```
ACTUAL_TARGET_VALUE PREDICTED_TARGET_VALUE       COST
------------------- ---------------------- ----------
                  0                      0          0
                  0                      1          2
                  1                      0          1
                  1                      1          0
```

The following statement associates the cost matrix with a Naive Bayes model called `nbmodel`.

```
BEGIN
  dbms_data_mining.add_cost_matrix('nbmodel', 'nbmodel_costs');
END;
/
```

The following query takes the cost matrix into account when scoring `mining_data_apply_v`. The output is restricted to those rows where a prediction of 1 is less costly then a prediction of 0.

```
SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
        FROM mining_data_apply_v
        WHERE PREDICTION (nbmodel COST MODEL
       USING cust_marital_status, education, household_size) = 1
        GROUP BY cust_gender
        ORDER BY cust_gender;
```

The output is as follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         25         38
M        208         43
```

You can specify costs inline when you invoke the scoring function. If you specify costs inline and the model also has an associated cost matrix, only the inline costs are used. The same query is shown below with different costs specified inline. Instead of the "2" shown in the cost matrix table (Table 7-3), "10" is specified in the inline costs.

```
SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
      FROM mining_data_apply_v
      WHERE PREDICTION (nbmodel
                COST (0,1) values ((0, 10),
                           (1, 0))
                USING cust_marital_status, education, household_size) = 1
      GROUP BY cust_gender
      ORDER BY cust_gender;
```

The output is as follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         74         39
M        581         43
```

The same query based on probability instead of costs is shown below.

```
SELECT cust_gender, COUNT(*) AS cnt, ROUND(AVG(age)) AS avg_age
        FROM mining_data_apply_v
        WHERE PREDICTION (nbmodel
           USING cust_marital_status, education, household_size) = 1
        GROUP BY cust_gender
        ORDER BY cust_gender;
```

The output is as follows:

```
C        CNT    AVG_AGE
- ---------- ----------
F         73         39
M        577         44
```

Related Topics

- Example 5-2

Priors and Class Weights

Offset differences in data distribution with prior probabilities and class weights to produce useful classification results.

With Bayesian models, you can specify **Prior** probabilities to offset differences in distribution between the build data and the real population (scoring data). With other forms of classification, you are able to specify **Class Weights**, which have the same biasing effect as priors.

In many problems, one target value dominates in frequency. For example, the positive responses for a telephone marketing campaign is 2% or less, and the occurrence of fraud in credit card transactions is less than 1%. A classification model built on historic data of this type cannot observe enough of the rare class to be able to distinguish the characteristics of the two classes; the result can be a model that when applied to new data predicts the frequent class for every case. While such a model can be highly accurate, it is not be very useful. This illustrates that it is not a good idea to rely solely on accuracy when judging the quality of a classification model.

To correct for unrealistic distributions in the training data, you can specify priors for the model build process. Other approaches to compensating for data distribution issues include stratified sampling and anomaly detection.

Related Topics

- About Anomaly Detection
  Identify unusual items or events in seemingly normal data to detect fraud, network intrusions, and other rare, significant occurrences through Anomaly Detection.

Specify Prior Probabilities

Prior probabilities can be used to offset differences in distribution between the build data and the actual population.

The `CLAS_PRIORS_TABLE_NAME` setting specifies the name of a table of prior probabilities to be used in building a Naive Bayes model. The priors table must have the columns shown in the following table.

Table 7-4 Priors Table Required Columns

Column Name	Data Type
`target_value`	valid target data type
`prior_probability`	`NUMBER`

Related Topics

- Target Attribute
  Understand what a **target** means in machine learning and understand the different target data types.
- *Oracle Machine Learning for SQL Concepts*

Specify Class Weights

Specify class weights table settings in logistic regression or Support Vector Machine (SVM) classification to favor higher weighted classes.

The `CLAS_WEIGHTS_TABLE_NAME` setting specifies the name of a table of class weights to be used to bias a logistic regression (Generalized Linear Model classification) or SVM classification model to favor higher weighted classes. The weights table must have the columns shown in the following table.

Table 7-5 Class Weights Table Required Columns

Column Name	Data Type
`target_value`	Valid target data type
`class_weight`	`NUMBER`

Related Topics

- Target Attribute
  Understand what a **target** means in machine learning and understand the different target data types.
- *Oracle Machine Learning for SQL Concepts*

Classification Algorithms

Learn the different classification algorithms used in Oracle Machine Learning.

Oracle Machine Learning for SQL provides the following algorithms for classification:

- **Decision Tree**

  Decision trees automatically generate rules, which are conditional statements that reveal the logic used to build the tree.

- **Explicit Semantic Analysis**

  Explicit Semantic Analysis (ESA) is designed to make predictions for text data. This algorithm can address use cases with hundreds of thousands of classes.

- **Generalized Linear Model**

  Generalized Linear Model (GLM) is a popular statistical technique for linear modeling. OML4SQL implements GLM for binary classification and for regression. GLM provides extensive coefficient statistics and model statistics, as well as row diagnostics. GLM also supports confidence bounds.

- **Naive Bayes**

  Naive Bayes uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

- **Random Forest**

  Random Forest is a powerful and popular machine learning algorithm that brings significant performance and scalability benefits.

- **Support Vector Machine**

  Support Vector Machine (SVM) is a powerful, state-of-the-art algorithm based on linear and nonlinear regression. OML4SQL implements SVM for binary and multiclass classification.

- **XGBoost**
  XGBoost is machine learning algorithm for regression and classification that makes available the XGBoost open source package. Oracle Machine Learning for SQL XGBoost prepares training data, invokes XGBoost, builds and persists a model, and applies the model for prediction.

> **Note:**
>
> OML4SQL uses Naive Bayes as the default classification algorithm.

Related Topics

- About Decision Tree
  Decision Tree classifies data using a tree structure of rules, making predictions clear and easy to interpret.
- About Explicit Semantic Analysis
  , Explicit Semantic Analysis (ESA) was introduced as an unsupervised algorithm for feature extraction and is enhanced as a supervised algorithm for classification.
- About Generalized Linear Model
  The Generalized Linear Model (GLM) includes and extends the class of linear models which address and accommodate some restrictive assumptions of the linear models.
- About Multivariate State Estimation Technique - Sequential Probability Ratio Test
  Multivariate state Estimation Technique - Sequential Probability Ratio Test (MSET-SPRT) is an algorithm for anomaly detection and statistical testing.
- About Naive Bayes
  Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.
- About Random Forest
  Random Forest is a classification algorithm that builds an **ensemble** (also called **forest**) of trees.
- About Support Vector Machine
  Support Vector Machine (SVM) is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory.
- About XGBoost
  Oracle's XGBoost prepares training data, builds and persists a model, and applies the model for classification and regression.

About Clustering

Identify clusters of similar data objects, useful for exploring and preprocessing data without predefined categories.

The members of a cluster are more like each other than they are like members of other clusters. Different clusters can have members in common. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high.

Clustering, like classification, is used to segment the data. Unlike classification, clustering models segment data into groups that were not previously defined. Classification models segment data by assigning it to previously-defined classes, which are specified in a target. Clustering models do not use a target.

Clustering is useful for exploring data. You can use clustering algorithms to find natural groupings when there are many cases and no obvious groupings.

Clustering can serve as a useful data-preprocessing step to identify homogeneous groups on which you can build supervised models.

You can also use clustering for anomaly detection. Once you segment the data into clusters, you find that some cases do not fit well into any clusters. These cases are anomalies or outliers.

How are Clusters Computed?

Compute clusters using density-based, distance-based, or grid-based methods to identify high-density areas, measure similarity, and form clusters.

There are several different approaches to the computation of clusters. Oracle Machine Learning supports the methods listed here.

- **Density-based**: This type of clustering finds the underlying distribution of the data and estimates how areas of high density in the data correspond to peaks in the distribution. High-density areas are interpreted as clusters. Density-based cluster estimation is probabilistic.
- **Distance-based**: This type of clustering uses a distance metric to determine similarity between data objects. The distance metric measures the distance between actual cases in the cluster and the prototypical case for the cluster. The prototypical case is known as the **centroid**.
- **Grid-based**: This type of clustering divides the input space into hyper-rectangular cells and identifies adjacent high-density cells to form clusters.

Scoring New Data

Score new data probabilistically to predict cluster assignments for new cases.

Although clustering is an unsupervised machine learning technique, Oracle Machine Learning supports the scoring operation for clustering. New data is scored probabilistically.

Hierarchical Clustering

Perform hierarchical clustering to generate final leaf clusters and intermediate clusters in the hierarchy.

Oracle Machine Learning supports clustering algorithms that perform hierarchical clustering. The leaf clusters are the final clusters generated by the algorithm. Clusters higher up in the hierarchy are intermediate clusters.

Rules

Describe data in clusters using conditional statements that capture the logic for cluster assignments.

Rules describe the data in each cluster. A rule is a conditional statement that captures the logic used to split a parent cluster into child clusters. A rule describes the conditions for a case to be assigned with some probability to a cluster.

Support and Confidence

Evaluate clustering rules using support (percentage of applicable cases) and confidence (probability of correct cluster assignment).

Support and confidence are metrics that describe the relationships between clustering rules and cases. **Support** is the percentage of cases for which the rule holds. **Confidence** is the probability that a case described by this rule is actually assigned to the cluster.

Evaluating a Clustering Model

Assess clustering models by examining generated information, such as centroids and hierarchical rules, to ensure reliability for business decisions.

Since known classes are not used in clustering, the interpretation of clusters can present difficulties. How do you know if the clusters can reliably be used for business decision making? Oracle Machine Learning clustering models support a high degree of model transparency. You can evaluate the model by examining information generated by the clustering algorithm: for example, the centroid of a distance-based cluster. Moreover, because the clustering process is hierarchical, you can evaluate the rules and other information related to each cluster's position in the hierarchy.

Clustering Algorithms

Learn different clustering algorithms used in Oracle Machine Learning for SQL.

Oracle Machine Learning for SQL supports these clustering algorithms:

- **Expectation Maximization**

  Expectation Maximization is a probabilistic, density-estimation clustering algorithm.
- **k-Means**

  k-Means is a distance-based clustering algorithm. Oracle Machine Learning for SQL supports an enhanced version of *k*-Means.
- **Orthogonal Partitioning Clustering (O-Cluster)**

  O-Cluster is a proprietary, grid-based clustering algorithm.

> **See Also:**
>
> Campos, M.M., Milenova, B.L., "O-Cluster: Scalable Clustering of Large High Dimensional Data Sets", Oracle Data Mining Technologies, 10 Van De Graaff Drive, Burlington, MA 01803.

The main characteristics of the two algorithms are compared in the following table.

Table 7-6 Clustering Algorithms Compared

Feature	k-Means	O-Cluster	Expectation Maximization
Clustering methodolgy	Distance-based	Grid-based	Distribution-based

Table 7-6 (Cont.) Clustering Algorithms Compared

Feature	k-Means	O-Cluster	Expectation Maximization
Number of cases	Handles data sets of any size	More appropriate for data sets that have more than 500 cases. Handles large tables through active sampling	Handles data sets of any size
Number of attributes	More appropriate for data sets with a low number of attributes	More appropriate for data sets with a high number of attributes	Appropriate for data sets with many or few attributes
Number of clusters	User-specified	Automatically determined	Automatically determined
Hierarchical clustering	Yes	Yes	Yes
Probabilistic cluster assignment	Yes	Yes	Yes

> **Note:**
>
> OML4SQL uses *k*-Means as the default clustering algorithm.

Related Topics

- Oracle Machine Learning for SQL
- About Expectation Maximization
  Expectation Maximization (EM) estimates mixture models for variety of applications, enhancing clustering and anomaly detection.
- About *k*-Means
  The *k*-Means algorithm is a distance-based clustering algorithm that partitions the data into a specified number of clusters.
- About O-Cluster
  O-Cluster is a fast, scalable grid-based clustering algorithm well-suited for analysing large, high-dimensional data sets. The algorithm can produce high quality clusters without relying on user-defined parameters.

Embedding

Explore embedding as a machine learning technique that transforms data in numeric dimensions that are represented as vectors to enable content similarity search and other applications.

About Vector Embeddings

Transformer models, also known as embedding models, are used to convert various types of data, such as words, sentences, documents, images, and more, into numerical vectors that capture their semantics.

These vectors are represented as points in a multidimensional space, where the proximity of points reflects the semantic similarity of the data they represent. Put differently, vector embeddings are a way of representing various types of data, like text, images, videos, or music, as points in a multidimensional space. The locations of these points and their proximity to others are semantically meaningful. This transformation enables machine learning algorithms to process and analyze data more effectively, and compute various distance metrics to find similar content. Creating vector embeddings involves training machine learning models, often neural networks, on large data sets to learn patterns and relationships within the data. This process transforms the data into numerical vectors, each uniquely representing a data point in a high-dimensional vector space. Applications of vector embeddings span a wide range of fields, particularly in natural language processing (NLP), search engines, and recommendation systems to name a few.

Pretrained Models for Generating Embeddings

Many pretrained models exist that generate embeddings for various data types, such as words, text sentences, images, and so on. These pretrained models often require pre-processing or post-processing operations or both.

As an example, most models for creating sentence embeddings from text require a pre-processing step called tokenization. **Tokenization** is a process to convert a sequence of text into smaller parts, called tokens. The embedding model then processes the tokens as input. Further post-processing might also be necessary for the output of these pretrained sentence transformers. One such post-processing operation is pooling. **Pooling** in text embeddings is a technique used to aggregate and reduce the dimensionality of individual word or token embeddings within a text sequence. This process involves combining the features of multiple embeddings to form a single, fixed-size representation of the entire text. For example, pooling methods can be employed to perform aggregation functions such as mean, max, or others. Another post-processing operation is normalization. **Normalization** in text embeddings is a process that adjusts the individual embeddings to have a uniform scale or distribution. This step involves transforming the embeddings so that they conform to a specific structure, often aiming to have a consistent length or scale across the data set.

Therefore, you need to use pretrained models that are augmented with pre-processing and post-processing operations to generate embeddings. This document illustrates examples that use the `my_embedding_model.onnx` model as an augmented ONNX format model. If you want to download and convert a pretrained model to an ONNX format model and augment the model with pre-processing and post-processing steps, see Import ONNX Models and Generate Embeddings.

Data Types for ONNX Embedding Models

ONNX defines its own data types. When you import ONNX models into Oracle Database, their data types are automatically mapped to SQL data types.

Attribute Data Type for ONNX Embedding Models

For a text embedding model, the input is a string. Therefore, the supported data type are `VARCHAR2`, `CLOB`, `NVARCHAR2`, and `NCLOB`. This means that there is a limit on the size of input strings to 4000 bytes (32767 bytes if the maximum string size is set to extended).

The `USER_MINING_MODEL_ATTRIBUTES` view reports the SQL data types for the input of a model.

```
USER_MINING_MODEL_ATTRIBUTESVARCHAR2
```

```
SELECT model_name, attribute_name, attribute_type, data_type, vector_info
FROM user_mining_model_attributes
```

```
WHERE model_name = 'DOC_MODEL'
ORDER BY ATTRIBUTE_NAME;


MODEL_NAME           ATTRIBUTE_NAME       ATTRIBUTE_TYPE       DATA_TYPE
VECTOR_INFO
-------------------- -------------------- -------------------- ----------
---------------
DOC_MODEL                INPUT_STRING         TEXT                 VARCHAR2
DOC_MODEL                ORA$ONNXTARGET       VECTOR               VECTOR
VECTOR(128,FLOA
                                                                          T32)
```

Target Data Type for ONNX Embedding Models

The output of a text embedding model is an embedding vector. Therefore, the target data type is `VECTOR`. Use the `VECTOR_EMBEDDING` SQL scoring function to generate vectors from an embedding model.

For more detail on `VECTOR` data type, see Create Tables Using the VECTOR Data Type. To learn more about `VECTOR_EMBEDDING` SQL operator, see *Oracle Database SQL Language Reference*.

Examples: Static Data Dictionary Views

You can use the Oracle Machine Learning static data dictionary views to view information such as available models, attributes of an ONNX embedding model and others. Values to support ONNX embedding models have been added.

Database administrator (DBA) and USER versions of the views are also available.

This section lists the examples of the impacted data dictionary views of ONNX embedding model.

Example: ALL_MINING_MODEL_ATTRIBUTES

You, as a current user, can view the attributes of a machine learning model by querying the `ALL_MINING_MODEL_ATTRIBUTES` view.

Here is an example of the model attributes of an embedding model. The name of the ONNX embedding model is `DOC_MODEL`:

```
SELECT model_name, attribute_name, attribute_type, data_type, vector_info
FROM user_mining_model_attributes
WHERE model_name = 'DOC_MODEL'
ORDER BY ATTRIBUTE_NAME;
```

The output is as follows:

```
MODEL_NAME           ATTRIBUTE_NAME       ATTRIBUTE_TYPE       DATA_TYPE
VECTOR_INFO
-------------------- -------------------- -------------------- ----------
---------------
```

```
DOC_MODEL                INPUT_STRING         TEXT                 VARCHAR2
DOC_MODEL                ORA$ONNXTARGET       VECTOR               VECTOR
VECTOR(128,FLOA
                                                                          T32)
```

See Also:

ALL_MINING_MODEL_ATTRIBUTES in *Oracle Database Reference*

Example: ALL_MINING_MODELS

You can check machine learning models available to you as a current user by querying the ALL_MINING_MODELS view.

Here is an example of model details of an embedding model. The name of the ONNX embedding model is *DOC_MODEL*:

```
SELECT MODEL_NAME, MINING_FUNCTION, ALGORITHM,
ALGORITHM_TYPE, MODEL_SIZE
FROM user_mining_models
WHERE model_name = 'DOC_MODEL'
ORDER BY MODEL_NAME;
```

The output is as follows:

```
MODEL_NAME           MINING_FUNCTION                ALGORITHM
ALGORITHM_ MODEL_SIZE
-------------------- ------------------------------ --------------------
---------- ----------
DOC_MODEL                EMBEDDING                      ONNX
NATIVE       17762137
```

See Also:

ALL_MINING_MODELS in *Oracle Database Reference*

Scoring: Generate Vector Embeddings

After importing the ONNX embedding model into the Database, you can generate embedding vectors using the VECTOR_EMBEDDING SQL scoring function.

The VECTOR_EMBEDDING SQL scoring function returns VECTOR(dimension, type). The embedding models define the number of dimensions of the output vector of the VECTOR_EMBEDDING operator. To learn about the VECTOR_EMBEDDING SQL scoring operator, see VECTOR_EMBEDDING.

Example

The following example generates vector embeddings with "hello" as the input, utilizing the pretrained ONNX format model `my_embedding_model.onnx` imported into the Database. For complete example, see Import ONNX Models into Oracle Database End-to-End Example.

```
SELECT TO_VECTOR(VECTOR_EMBEDDING(doc_model USING 'hello' as data)) AS
embedding;
------------------------------------------------------------------------------
--
[-9.76553112E-002,-9.89954844E-002,7.69771636E-003,-4.16760892E-003,-9.6930563
4E-002,
-3.01141385E-002,-2.63396613E-002,-2.98553891E-002,5.96499592E-002,4.13885899E
-002,
5.32859489E-002,6.57707453E-002,-1.47056757E-002,-4.18472625E-002,4.1588001E-0
02,
-2.86354572E-002,-7.56499246E-002,-4.16395674E-003,-1.52879998E-001,6.60010576
E-002,
-3.9013084E-002,3.15719917E-002,1.2428958E-002,-2.47651711E-002,-1.16851285E-0
01,
-7.82847106E-002,3.34323719E-002,8.03267583E-002,1.70483496E-002,-5.42407483E-
002,
6.54291287E-002,-4.81935125E-003,6.11041225E-002,6.64106477E-003,-5.47
```

Note:

You can define the outputs explicitly in the metadata or implicitly. The system assumes a single output for a model if you don't specify the output in the metadata.

If a scoring function does not comply as per the description in Supported SQL Scoring Functions, you will receive an ORA-40290 error when performing the scoring operation on your data. Additionally, any unsupported scoring functions will raise the ORA-40290 error.

See Also:

A complete list of SQL scoring functions supported for ONNX models, in *Oracle Machine Learning for SQL User's Guide*.

Treatment of Missing Data During Scoring

ONNX does not support representation for non-existent values; that is, there is no equivalent to `NULL` for SQL.

Further, if the input values are not specified, then the ONNX embedding models fail to run.

- Absent attribute: If fewer attributes are used for scoring than were specified during model import (input), then you receive an error when you perform scoring. That is, if at least one of the input value is not specified in the `USING` clause of a scoring operator with ONNX model, then the query will not compile.

- NULL attribute: If any of the attributes has a NULL value, then the scoring operator does not perform inference of the model with the ONNX Runtime and returns a NULL result immediately. If you want to change this behavior, then provide an appropriate replacement to the NULL value, either by using an NVL expression as input attribute (for example, `NVL(input_attribute, default_value) AS input_attribute);`) or by specifying a default value for this input attribute using the JSON metadata when importing the model.

Import ONNX Models into Oracle Database End-to-End Example

Learn to import a pretrained embedding model that is in ONNX format and generate vector embeddings.

Follow the steps below to import a pretrained ONNX formatted embedding model into the Oracle Database.

Prepare Your Data Dump Directory

Prepare your data dump directory and provide the necessary access and privileges to dmuser.

1. Choose from:
   a. If you already have a pretrained ONNX embedding model, store it in your working folder.
   b. If you do not have pretrained embedding model in ONNX format, perform the steps listed in Convert Pretrained Models to ONNX Format.
2. Login to SQL*Plus as SYSDBA in your PDB.

   ```
   CONN sys/<password>@pdb as sysdba;
   ```

3. Grant the DB_DEVELOPER_ROLE to dmuser.

   ```
   GRANT DB_DEVELOPER_ROLE TO dmuser identified by <password>;
   ```

4. Grant CREATE MINING MODEL privilege to dmuser.

   ```
   GRANT create mining model TO dmuser;
   ```

5. Set your working folder as the data dump directory (DM_DUMP) to load the ONNX embedding model.

   ```
   CREATE OR REPLACE DIRECTORY DM_DUMP as '<work directory path>';
   ```

6. Grant READ permissions on the DM_DUMP directory to dmuser.

   ```
   GRANT READ ON DIRECTORY dm_dump TO dmuser;
   ```

7. Grant WRITE permissions on the DM_DUMP directory to dmuser.

   ```
   GRANT WRITE ON DIRECTORY dm_dump TO dmuser;
   ```

8. Drop the model if it already exits.

   ```
   exec DBMS_VECTOR.DROP_ONNX_MODEL(model_name => 'doc_model', force => true);
   ```

Import ONNX Model Into the Database

You created a data dump directory and now you load the ONNX model into the Database. Use the `DBMS_VECTOR.LOAD_ONNX_MODEL` procedure to load the model. The `DBMS_VECTOR.LOAD_ONNX_MODEL` procedure facilitates the process of importing ONNX format model into the Oracle Database. In this example, the procedure loads an ONNX model file, named `my_embedding_model.onnx` from the `DM_DUMP` directory, into the Database as `doc_model`, specifying its use for embedding tasks.

1. Connect as `dmuser`.

   ```
   CONN dmuser/<password>@<pdbname>;
   ```

2. Load the ONNX model into the Database.

   If the ONNX model to be imported already includes an output tensor named `embeddingOutput` and an input string tensor named `data`, JSON metadata is unnecessary. Embedding models converted from OML4Py follow this convention and can be imported without the JSON metadata.

   ```
   EXECUTE DBMS_VECTOR.LOAD_ONNX_MODEL(
     'DM_DUMP',
     'my_embedding_model.onnx',
     'doc_model');
   ```

   Alternately, you can load the ONNX embedding model by specifying the JSON metadata.

```
EXECUTE DBMS_VECTOR.LOAD_ONNX_MODEL(
  'DM_DUMP',
  'my_embedding_model.onnx',
  'doc_model',
  JSON('{"function" : "embedding", "embeddingOutput" : "embedding", "input": {"input":
["DATA"]}}'));
```

The procedure `LOAD_ONNX_MODEL` declares these parameters:

- `DM_DUMP`: specifies the directory name of the data dump.

  > **Note:**
  >
  > Ensure that the `DM_DUMP` directory is defined.

- `my_embedding_model`: is a `VARCHAR2` type parameter that specifies the name of the ONNX model.
- `doc_model`: This parameter is a user-specified name under which the model is stored in the Oracle Database.
- The JSON metadata associated with the ONNX model is declared as:

  `"function" : "embedding"`: Indicates the function name for text embedding model.

  `"embeddingOutput" : "embedding"`: Specifies the output variable which contains the embedding results.

- `"input": {"input": ["DATA"]}`: Specifies a JSON object (`"input"`) that describes the input expected by the model. It specifies that there is an input named `"input"`, and its value should be an array with one element, `"DATA"`. This indicates that the model expects a single string input to generate embeddings.

For more information about the `LOAD_ONNX_MODEL` procedure, see *Oracle Database PL/SQL Packages and Types Reference*.

Alternatively, if your ONNX embedding model is loaded on cloud object storage, the `LOAD_ONNX_MODEL_CLOUD` procedure can be used. For more information, see *Oracle Database PL/SQL Packages and Types Reference*.

Query Model Statistics

You can view model attributes and learn about the model by querying machine learning dictionary views and model detail views.

> **Note:**
>
> *DOC_MODEL* is the user-specified name of the embedding text model.

1. Query `USER_MINING_MODEL_ATTRIBUTES` view.

   ```
   SELECT model_name, attribute_name, attribute_type, data_type, vector_info
   FROM user_mining_model_attributes
   WHERE model_name = 'DOC_MODEL'
   ORDER BY ATTRIBUTE_NAME;
   ```

   To learn about `USER_MINING_MODEL_ATTRIBUTES` view, see USER_MINING_MODEL_ATTRIBUTES.

2. Query `USER_MINING_MODELS` view.

   ```
   SELECT MODEL_NAME, MINING_FUNCTION, ALGORITHM,
   ALGORITHM_TYPE, MODEL_SIZE
   FROM user_mining_models
   WHERE model_name = 'DOC_MODEL'
   ORDER BY MODEL_NAME;
   ```

   To learn about `USER_MINING_MODELS` view, see USER_MINING_MODELS.

3. Check model statistics by viewing the model detail views. Query the `DM$VMDOC_MODEL` view.

   ```
   SELECT * FROM DM$VMDOC_MODEL ORDER BY NAME;
   ```

   To learn about model details views for ONNX embedding models, see Model Details Views for ONNX Models.

4. Query the `DM$VPDOC_MODEL` model detail view.

   ```
   SELECT * FROM DM$VPDOC_MODEL ORDER BY NAME;
   ```

5. Query the `DM$VJDOC_MODEL` model detail view.

    ```
    SELECT * FROM DM$VJDOC_MODEL;
    ```

Generate Embeddings

Apply the model and generate vector embeddings for your input. Here, the input is *hello*.

Generate vector embeddings using the `VECTOR_EMBEDDING` function.

```
SELECT TO_VECTOR(VECTOR_EMBEDDING(doc_model USING 'hello' as data)) AS
embedding;
```

To learn about the `VECTOR_EMBEDDING` SQL function, see VECTOR_EMBEDDING. You can use the `UTL_TO_EMBEDDING` function in the `DBMS_VECTOR_CHAIN` PL/SQL package to generate vector embeddings generically through REST endpoints. To explore these functions, see the example Convert Text String to Embedding.

Example: Importing a Pretrained ONNX Model to Oracle Database

The following presents a comprehensive step-by-step example of importing ONNX embedding and generating vector embeddings.

```
conn sys/<password>@pdbname as sysdba
grant db_developer_role to dmuser identified by <password>;
grant create mining model to dmuser;

create or replace directory DM_DUMP as '<work directory path>';
grant read on directory dm_dump to dmuser;
grant write on directory dm_dump to dmuser;
>conn dmuser/<password>@<pdbname>;

-- Drop the model if it exits
exec DBMS_VECTOR.DROP_ONNX_MODEL(model_name => 'doc_model', force => true);

-- Load Model
EXECUTE DBMS_VECTOR.LOAD_ONNX_MODEL(
    'DM_DUMP',
    'my_embedding_model.onnx',
    'doc_model',
    JSON('{"function" : "embedding", "embeddingOutput" : "embedding"}'));
/

--check the attributes view
set linesize 120
col model_name format a20
col algorithm_name format a20
col algorithm format a20
col attribute_name format a20
col attribute_type format a20
col data_type format a20

SQL> SELECT model_name, attribute_name, attribute_type, data_type, vector_info
FROM user_mining_model_attributes
WHERE model_name = 'DOC_MODEL'
```

```
ORDER BY ATTRIBUTE_NAME;


OUTPUT:

MODEL_NAME           ATTRIBUTE_NAME       ATTRIBUTE_TYPE       DATA_TYPE
VECTOR_INFO
-------------------- -------------------- -------------------- ----------
---------------
DOC_MODEL                INPUT_STRING         TEXT                 VARCHAR2
DOC_MODEL                ORA$ONNXTARGET       VECTOR               VECTOR
VECTOR(128,FLOA
                                                                          T32)



SQL> SELECT MODEL_NAME, MINING_FUNCTION, ALGORITHM,
ALGORITHM_TYPE, MODEL_SIZE
FROM user_mining_models
WHERE model_name = 'DOC_MODEL'
ORDER BY MODEL_NAME;

OUTPUT:
MODEL_NAME           MINING_FUNCTION                ALGORITHM
ALGORITHM_ MODEL_SIZE
-------------------- ------------------------------ --------------------
---------- ----------
DOC_MODEL                EMBEDDING                      ONNX
NATIVE       17762137



SQL> select * from DM$VMDOC_MODEL ORDER BY NAME;

OUTPUT:
NAME                                     VALUE
----------------------------------------
----------------------------------------
Graph Description                        Graph combining g_8_torch_jit and
torch_
                                         jit
                                         g_8_torch_jit



                                         torch_jit


Graph Name                               g_8_torch_jit_torch_jit
Input[0]                                 input:string[1]
Output[0]                                embedding:float32[?,128]
Producer Name                            onnx.compose.merge_models
Version                                  1

6 rows selected.
```

```
SQL> select * from DM$VPDOC_MODEL ORDER BY NAME;

OUTPUT:
NAME                                     VALUE
----------------------------------------
----------------------------------------
batching                                 False
embeddingOutput                          embedding


SQL> select * from DM$VJDOC_MODEL;

OUTPUT:
METADATA
------------------------------------------------------------------------------
--
{"function":"embedding","embeddingOutput":"embedding","input":{"input":
["DATA"]}}



--apply the model
SQL> SELECT TO_VECTOR(VECTOR_EMBEDDING(doc_model USING 'hello' as data)) AS
embedding;

------------------------------------------------------------------------------
--
[-9.76553112E-002,-9.89954844E-002,7.69771636E-003,-4.16760892E-003,-9.6930563
4E-002,
-3.01141385E-002,-2.63396613E-002,-2.98553891E-002,5.96499592E-002,4.13885899E
-002,
5.32859489E-002,6.57707453E-002,-1.47056757E-002,-4.18472625E-002,4.1588001E-0
02,
-2.86354572E-002,-7.56499246E-002,-4.16395674E-003,-1.52879998E-001,6.60010576
E-002,
-3.9013084E-002,3.15719917E-002,1.2428958E-002,-2.47651711E-002,-1.16851285E-0
01,
-7.82847106E-002,3.34323719E-002,8.03267583E-002,1.70483496E-002,-5.42407483E-
002,
6.54291287E-002,-4.81935125E-003,6.11041225E-002,6.64106477E-003,-5.47
```

Oracle AI Vector Search SQL Scenario

To learn how you can chunk *database-concepts23ai.pdf* and *oracle-ai-vector-search-users-guide.pdf*, generate vector embeddings, and perform similarity search using vector indexes, see Quick Start SQL.

Alternate Method to Import ONNX Models

Use the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure to import the model and declare the input name. A PL/SQL helper block is used to facilitate the process of importing the ONNX format model into the Oracle Database in the included example.

Perform the following steps to import ONNX model into the Database using `DBMS_DATA_MINING` PL/SQL package.

- Connect as `dmuser`.

  ```
  CONN dmuser/<password>@<pdbname>;
  ```

- Run the following helper PL/SQL block:

  ```
  DECLARE
      m_blob BLOB default empty_blob();
      m_src_loc BFILE ;
      BEGIN
      DBMS_LOB.createtemporary (m_blob, FALSE);
      m_src_loc := BFILENAME('DM_DUMP', 'my_embedding_model.onnx');
      DBMS_LOB.fileopen (m_src_loc, DBMS_LOB.file_readonly);
      DBMS_LOB.loadfromfile (m_blob, m_src_loc, DBMS_LOB.getlength
  (m_src_loc));
      DBMS_LOB.CLOSE(m_src_loc);
      DBMS_DATA_MINING.import_onnx_model ('doc_model', m_blob,
  JSON('{"function" : "embedding", "embeddingOutput" : "embedding", "input":
  {"input": ["DATA"]}}'));
      DBMS_LOB.freetemporary (m_blob);
      END;
      /
  ```

  The code sets up a `BLOB` object and a `BFILE` locator, creates a temporary `BLOB` for storing the `my_embedding_model.onnx` file from the `DM_DUMP` directory, and reads its contents into the `BLOB`. It then closes the file and uses the content to import an ONNX model into the database with specified metadata, before releasing the temporary `BLOB` resources.

The schema of the `IMPORT_ONNX_MODEL` procedure is as follows: `DBMS_DATA_MINING.IMPORT_ONNX_MODEL(model_data, model_name, metadata)`. This procedure loads `IMPORT_ONNX_MODEL` from the `DBMS_DATA_MINING` package to import the ONNX model into the Database using the name provided in `model_name`, the BLOB content in `m_blob`, and the associated `metadata`.

- `doc_model`: This parameter is a user-specified name under which the imported model is stored in the Oracle Database.
- `m_blob`: This is a model data in `BLOB` that holds the ONNX representation of the model.
- `"function" : "embedding"`: Indicates the function name for text embedding model.
- `"embeddingOutput" : "embedding"`: Specifies the output variable which contains the embedding results.
- `"input": {"input": ["DATA"]}`: Specifies a JSON object (`"input"`) that describes the input expected by the model. It specifies that there is an input named `"input"`, and its value should be an array with one element, `"DATA"`. This indicates that the model expects a single string input to generate embeddings.

Alternately, the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure can also accept a `BLOB` argument representing an ONNX file stored and loaded from OCI Object Storage. The following is an example to load an ONNX model stored in an OCI Object Storage.

```
DECLARE
  model_source BLOB := NULL;
BEGIN
  -- get BLOB holding onnx model
```

```
  model_source := DBMS_CLOUD.GET_OBJECT(
    credential_name => 'myCredential',
    object_uri => 'https://objectstorage.us-phoenix -1.oraclecloud.com/' ||
      'n/namespace -string/b/bucketname/o/myONNXmodel.onnx');

  DBMS_DATA_MINING.IMPORT_ONNX_MODEL(
    "myonnxmodel",
    model_source,
    JSON('{ function : "embedding" })
  );
END;
/
```

This PL/SQL block starts by initializing a `model_source` variable as a `BLOB` type, initially set to NULL. It then retrieves an ONNX model from Oracle Cloud Object Storage using the `DBMS_CLOUD.GET_OBJECT` procedure, specifying the credentials `(OBJ_STORE_CRED`) and the URI of the model. The ONNX model resides in a specific bucket named `bucketname` in this case, and is accessible through the provided URL. Then, the script loads the ONNX model into the `model_source` BLOB. The `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure then imports this model into the Oracle Database as `myonnxmodel`. During the import, a JSON metadata specifies the model's function as `embedding`, for embedding operations.

See IMPORT_ONNX_MODEL Procedure and GET_OBJECT Procedure and Function to learn about the PL/SQL procedure.

Example: Importing a Pretrained ONNX Model to Oracle Database

The following presents a comprehensive step-by-step example of importing ONNX embedding and generating vector embeddings.

```
conn sys/<password>@pdb as sysdba
grant db_developer_role to dmuser identified by dmuser;
grant create mining model to dmuser;

create or replace directory DM_DUMP as '<work directory path>';
grant read on directory dm_dump to dmuser;
grant write on directory dm_dump to dmuser;
>conn dmuser/<password>@<pdbname>;

-- Drop the model if it exits
exec DBMS_VECTOR.DROP_ONNX_MODEL(model_name => 'doc_model', force => true);

-- Load Model
EXECUTE DBMS_VECTOR.LOAD_ONNX_MODEL(
    'DM_DUMP',
    'my_embedding_model.onnx',
    'doc_model',
    JSON('{"function" : "embedding", "embeddingOutput" : "embedding"}'));
/
--Alternately, load the model
EXECUTE DBMS_DATA_MINING.IMPORT_ONNX_MODEL(
       'my_embedding_model.onnx',
    'doc_model',
    JSON('{"function" : "embedding",
    "embeddingOutput" : "embedding",
    "input": {"input": ["DATA"]}}')
    );

--check the attributes view
set linesize 120
col model_name format a20
col algorithm_name format a20
col algorithm format a20
col attribute_name format a20
col attribute_type format a20
col data_type format a20

SQL> SELECT model_name, attribute_name, attribute_type, data_type, vector_info
FROM user_mining_model_attributes
WHERE model_name = 'DOC_MODEL'
ORDER BY ATTRIBUTE_NAME;


OUTPUT:

MODEL_NAME           ATTRIBUTE_NAME       ATTRIBUTE_TYPE       DATA_TYPE
VECTOR_INFO
-------------------- -------------------- -------------------- ----------
---------------
DOC_MODEL                INPUT_STRING         TEXT                 VARCHAR2
DOC_MODEL                ORA$ONNXTARGET       VECTOR               VECTOR
VECTOR(128,FLOA
                                                                          T32)



SQL> SELECT MODEL_NAME, MINING_FUNCTION, ALGORITHM,
ALGORITHM_TYPE, MODEL_SIZE
FROM user_mining_models
WHERE model_name = 'DOC_MODEL'
ORDER BY MODEL_NAME;

OUTPUT:
MODEL_NAME           MINING_FUNCTION                ALGORITHM
ALGORITHM_ MODEL_SIZE
-------------------- ------------------------------ --------------------
---------- ----------
DOC_MODEL                EMBEDDING                      ONNX
NATIVE       17762137



SQL> select * from DM$VMDOC_MODEL ORDER BY NAME;

OUTPUT:
NAME                                     VALUE
----------------------------------------
----------------------------------------
Graph Description                        Graph combining g_8_torch_jit and
torch_
                                         jit
                                         g_8_torch_jit



                                         torch_jit


Graph Name                               g_8_torch_jit_torch_jit
Input[0]                                 input:string[1]
Output[0]                                embedding:float32[?,128]
Producer Name                            onnx.compose.merge_models
Version                                  1

6 rows selected.


SQL> select * from DM$VPDOC_MODEL ORDER BY NAME;

OUTPUT:
NAME                                     VALUE
----------------------------------------
----------------------------------------
batching                                 False
embeddingOutput                          embedding


SQL> select * from DM$VJDOC_MODEL;

OUTPUT:
METADATA
------------------------------------------------------------------------------
--
{"function":"embedding","embeddingOutput":"embedding","input":{"input":
["DATA"]}}



--apply the model
SQL> SELECT TO_VECTOR(VECTOR_EMBEDDING(doc_model USING 'hello' as data)) AS
embedding;

------------------------------------------------------------------------------
--
[-9.76553112E-002,-9.89954844E-002,7.69771636E-003,-4.16760892E-003,-9.6930563
4E-002,
-3.01141385E-002,-2.63396613E-002,-2.98553891E-002,5.96499592E-002,4.13885899E
-002,
5.32859489E-002,6.57707453E-002,-1.47056757E-002,-4.18472625E-002,4.1588001E-0
02,
-2.86354572E-002,-7.56499246E-002,-4.16395674E-003,-1.52879998E-001,6.60010576
E-002,
-3.9013084E-002,3.15719917E-002,1.2428958E-002,-2.47651711E-002,-1.16851285E-0
01,
-7.82847106E-002,3.34323719E-002,8.03267583E-002,1.70483496E-002,-5.42407483E-
002,
6.54291287E-002,-4.81935125E-003,6.11041225E-002,6.64106477E-003,-5.47
```

About Feature Extraction

Feature extraction supports transforming original attributes into linear combinations to reduce dimensionality and enhance model quality.

Feature extraction is a dimensionality reduction technique. Unlike feature selection, which selects and retains the most significant attributes, feature extraction actually transforms the attributes. In Oracle Machine Learning, the transformed attributes, or **features**, are linear combinations of the original attributes.

The feature extraction technique results in a much smaller and richer set of attributes. The maximum number of features can be user-specified or determined by the algorithm. By default, the algorithm determines it.

Models built on extracted features can be of higher quality, 'because the attributes concentrate the signal found in weaker attributes in fewer attributes that describe the data. However, interpreting the resulting features and models becomes more challenging.

We can think of each feature or attribute as one such dimension. Feature extraction projects a data set with higher dimensionality onto a smaller number of dimensions. As such it is useful for data visualization, since a complex data set can be effectively visualized when it is reduced to two or three dimensions.

Some applications of feature extraction are latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. Feature extraction can also be used to enhance the speed and effectiveness of machine learning algorithms.

Feature extraction can be used to extract the themes of a document collection, where documents are represented by a set of key words and their frequencies. Each theme (feature) is represented by a combination of keywords. The documents in the collection can then be expressed in terms of the discovered themes.

Feature Extraction and Scoring

Transform input data into features without a target, using feature extraction for improved data representation.

Oracle Machine Learning (OML) provides two primary methods for using in-database feature extraction algorithms:

- Producing individual features or attributes as columns that can be used as input to other algorithms.
- Creating a vector output consisting of multiple dimensions, where each dimension corresponds to a feature or attribute.

Both approaches improve data representation and enable downstream analytic tasks, however, the vector output offers additional advantages for handling large, dense data.

Feature Extraction algorithms in OML transform input data into a set of features or dimensions improving the data's representation. As an unsupervised machine learning technique, feature extraction does not involve a target. This allows models to extract meaningful attributes from the input, optimizing the data for subsequent analysis.

OML supports scoring operations for feature extraction using the following operators:

- `FEATURE_ID` and `FEATURE_VALUE` operators extract individual features.
- `FEATURE_SET` returns feature ID and value pair sets.

- VECTOR_EMBEDDING operator enables VECTOR data type output for OML feature extraction models, facilitating a unified approach for vectorization.

The FEATURE_SET operator retrieves results of the full projections. That is, transforming high-dimensional data into a lower-dimensional space while preserving as much of the data's structure and information as possible. You can use this operation to query feature values across all feature IDs. The output representation is not ideal for large, dense data, and requires extra processing for use on any vector-based operations, like similarity search. OML supports using the VECTOR_EMBEDDING operator that enables VECTOR data type as an output representation for feature extraction models such as, SVD, PCA, NMF, and ESA (with **random projections**), which enhances the usability of those algorithms. The benefits of vector output include:

- Vector output optimizes data representation and provides a more compact format, reducing computational requirements for subsequent analytic tasks.
- Vector output enables vector-based operations, including similarity search on relational data.

The following cases present how the vector outputs are determined:

- **Determining the dimension of the output vector**: The data dictionary views, USER/ALL/DBA_MINING_MODEL_ATTRIBUTES and USER/ALL/DBA_MINING_MODEL_XFORMS for feature extraction models have a new attribute, ORA$VECTOR, of the DTYVEC data type. Its dimension and storage type are detailed in the VECTOR_INFO column.

  The output vector dimension corresponds to the number of features you require and specify. If you do not specify this, algorithms determine an optimal dimension based on the data, looking for natural cut-off points such as significant drops in explained variance.
- **Handling partitioned models with FLEX dimension vector**: For partitioned models, each partition of data might have different characteristics or levels of complexity, which can result in projections with different dimensions for each partition. For example, in one partition, there might be fewer meaningful features, leading to a lower-dimensional projection. In another partition, the data might have more complexity, resulting in a higher-dimensional projection. In these cases, the system utilizes a FLEX dimension vector to accommodate the varying dimensionality. The OML partitioned models that consider partitioned sets, will use each partition’s vector in isolation, leveraging the specific data characteristics of that partition. The FLEX dimension type is stored in VECTOR_INFO using the vector format. See ALL_MINING_MODEL_ATTRIBUTES.
- **Special case: Zero features**: When a model has zero features, the system outputs an empty entry, maintaining consistency with the current behavior of the FEATURE_VALUE operator.

For a step-by-step example on how you can use Feature Extraction in conjunction with the VECTOR_EMBEDDING operator, see Vectorize Relational Tables Using OML Feature Extraction Algorithms.

> **Note:**
>
> The VECTOR data type and VECTOR_EMBEDDING operator applies only to newly built models. Older models lack the necessary vector output metadata, and the system raises a 40290 error to show the operator is incompatible with the model if the VECTOR_EMBEDDING operator is used with them.

Related Topics

- Oracle Machine Learning for SQL Functions

Algorithms for Feature Extraction

Understand the algorithms used for feature extraction.

OML4SQL supports these feature extraction algorithms:

- **Explicit Semantic Analysis** (ESA).
- **Non-Negative Matrix Factorization** (NMF).
- **Singular Value Decomposition** (SVD) and **Principal Component Analysis** (PCA).

> **Note:**
>
> OML4SQL uses NMF as the default feature extraction algorithm.

Related Topics

- About Explicit Semantic Analysis
  , Explicit Semantic Analysis (ESA) was introduced as an unsupervised algorithm for feature extraction and is enhanced as a supervised algorithm for classification.
- About NMF
  Non-Negative Matrix Factorization is useful when there are many attributes and the attributes are ambiguous or have weak predictability. By combining attributes, NMF can produce meaningful patterns, topics, or themes. NMF is a feature extraction algorithm.
- About Singular Value Decomposition
  SVD and the closely-related PCA are well established feature extraction methods that have a wide range of applications. Oracle Machine Learning for SQL implements Singular Value Decomposition (SVD) as a feature extraction algorithm and Principal Component Analysis (PCA) as a special scoring method for SVD models.
- PCA scoring
  Learn about configuring Singular Value Decomposition (SVD) to perform Principal Component Analysis (PCA) projections.

Finding the Attributes

Identify important attributes by reducing noise, correlation, and high dimensionality through preprocessing steps.

Sometimes too much information can reduce the effectiveness of machine learning. Some of the columns of data attributes assembled for building and testing a model in a supervised learning do not contribute meaningful information to the model. Some actually detract from the quality and accuracy of the model.

For example, you want to collect a great deal of data about a given population because you want to predict the likelihood of a certain illness within this group. Some of this information, perhaps much of it, has little or no effect on susceptibility to the illness. It is possible that attributes such as the number of cars per household do not have effect whatsoever.

Irrelevant attributes add noise to the data and can affect model accuracy. Noise increases the size of the model and the time and system resources needed for model building and scoring.

Data sets with many attributes can contain groups of attributes that are correlated. These attributes actually measure the same underlying feature. Their presence together in the build data can skew the patterns found by algorithm and affect the accuracy of the model.

Wide data (many attributes) typically results in more processing by machine learning algorithms. Model attributes are the dimensions of the processing space used by the algorithm. The higher the dimensionality of the processing space, the higher the computation cost involved in algorithmic processing.

To minimize the effects of noise, correlation, and high dimensionality, some form of dimension reduction is often a desirable preprocessing step. Feature selection involves identifying those attributes that are most predictive and selecting among those to provide the algorithm for model building. By removing attributes that add little or no value to a model has these benefits: potentially increasing model accuracy while reducing compute time since fewer attributes need to be processed. Informative and representative samples are best suited in feature selection. Sometimes you can represent the variables that are important than to represent the linear combination of variables. You can single-out and measure the "importance" of a column or a row in a data matrix in an unsupervised manner (a low-rank matrix decomposition).

Feature selection optimization is performed in the Decision Tree algorithm and within Naive Bayes as an algorithm behavior. The Generalized Linear Model (GLM) algorithm can be configured to perform feature selection through model setting.

About Feature Selection and Attribute Importance

Rank attributes based on their significance to improve computational efficiency and predictive accuracy.

Finding the most significant predictors is the goal of some machine learning projects. For example, a model might seek to find the principal characteristics of clients who pose a high credit risk. Oracle Machine Learning supports the **attribute importance** machine learning technique, which ranks attributes according to their importance. Attribute importance does not actually select the features, but ranks them as to their relevance to predicting the result. It is up to the user to review the ranked features and create a data set to include the desired features.

Feature selection is useful as a preprocessing step to improve computational efficiency in predictive modeling.

Attribute Importance and Scoring

Rank attributes by their influence for better training data selection in classification and regression models.

The results of attribute importance are the attributes of the build data ranked according to their influence. The ranking and the measure of importance can be used in selecting training data for classification and regression models. Also, used for selecting data for unsupervised algorithm like CUR matrix decomposition. Oracle Machine Learning does not support the scoring operation for attribute importance.

Algorithms for Attribute Importance

Understand the algorithms used for attribute importance.

Oracle Machine Learning for SQL supports the following algorithms for attribute importance:

- Minimum Description Length
- CUR Matrix Decomposition

Related Topics

- About CUR Matrix Decomposition
  CUR Matrix Decomposition is a low-rank matrix decomposition algorithm that is explicitly expressed in a small number of actual columns and/or actual rows of data matrix.
- About MDL
  Minimum Description Length (MDL) is an information theoretic model selection principle that assumes the simplest representation of data is the most probable explanation.

About Ranking

Rank items to improve applications like e-commerce, social networks, and recommendation systems.

Ranking is useful for many applications in information retrieval such as e-commerce, social networks, recommendation systems, and so on. For example, a user searches for an article or an item to buy online. To build a recommendation system, it becomes important that similar articles or items of relevance appear to the user such that the user clicks or purchases the item. A simple regression model can predict the probability of a user to click an article or buy an item. However, it is more practical to use ranking technique and be able to order or rank the articles or items to maximize the chances of getting a click or purchase. The prioritization of the articles or the items influence the decision of the users.

The ranking technique directly ranks items by training a model to predict the ranking of one item over another item. In the training model, it is possible to have items, ranking one over the other by having a "score" for each item. Higher ranked items have higher scores and lower ranked items have lower scores. Using these scores, a model is built to predict which item ranks higher than the other.

Ranking Methods

Oracle Machine Learningsupports pairwise and listwise ranking methods through XGBoost.

For a training data set, in a number of sets, each set consists of objects and labels representing their ranking. A ranking function is constructed by minimizing a certain loss function on the training data. Using test data, the ranking function is applied to get a ranked list of objects. Ranking is enabled for XGBoost using the regression function. OML4SQL supports pairwise and listwise ranking methods through XGBoost.

Pairwise ranking: This approach regards a pair of objects as the learning instance. The pairs and lists are defined by supplying the same `case_id` value. Given a pair of objects, this approach gives an optimal ordering for that pair. Pairwise losses are defined by the order of the two objects. In OML4SQL, the algorithm uses LambdaMART to perform pairwise ranking with the goal of minimizing the average number of inversions in ranking.

Listwise ranking: This approach takes multiple lists of ranked objects as learning instance. The items in a list must have the same `case_id`. The algorithm uses LambdaMART to perform list-wise ranking.

See Also:

- "Ranking Measures and Loss Functions in Learning to Rank" a research paper presentation on the internet.
- *Oracle Database PL/SQL Packages and Types Reference* for a listing and explanation of the available model settings for XGBoost.

Note:

The term hyperparameter is also interchangeably used for model setting.

Related Topics

- About XGBoost
  Oracle's XGBoost prepares training data, builds and persists a model, and applies the model for classification and regression.
- DBMS_DATA_MINING — Algorithm Settings: XGBoost

Ranking Algorithms

Employ the XGBoost algorithm for ranking items, a regression function

OML4SQL supports XGBoost algorithm for ranking.

Related Topics

- About XGBoost
  Oracle's XGBoost prepares training data, builds and persists a model, and applies the model for classification and regression.

About Regression

Regression is a machine learning technique that predicts numeric values along a continuum.

Profit, sales, mortgage rates, house values, square footage, temperature, or distance can be predicted using Regression techniques. For example, a regression model can be used to predict the value of a house based on location, number of rooms, lot size, and other factors.

A regression task begins with a data set in which the target values are known. For example, a regression model that predicts house values can be developed based on observed data for many houses over a period of time. In addition to the value, the data can track the age of the house, square footage, number of rooms, taxes, school district, proximity to shopping centers, and so on. House value can be the target, the other attributes are the predictors, and the data for each house constitutes a case.

In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two data sets: one for building the model, the other for testing the model.

Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.

How Does Regression Work?

Estimate target values as a function of predictors, minimizing error to fit a set of data observations.

You do not need to understand the mathematics used in regression analysis to develop and use quality regression models for machine learning. However, it is helpful to understand a few basic concepts.

Regression analysis seeks to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide. The following equation expresses these relationships in symbols. It shows that regression is the process of estimating the value of a continuous target (*y*) as a function (*F*) of one or more predictors ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (*e*).

$$y = F(\mathbf{x},\theta) + e$$

The predictors can be understood as independent variables and the target as a dependent variable. The error, also called the **residual**, is the difference between the expected and predicted value of the dependent variable. The regression parameters are also known as **regression coefficients**.

The process of training a regression model involves finding the parameter values that minimize a measure of the error, for example, the sum of squared errors.

There are different families of regression functions and different ways of measuring the error.

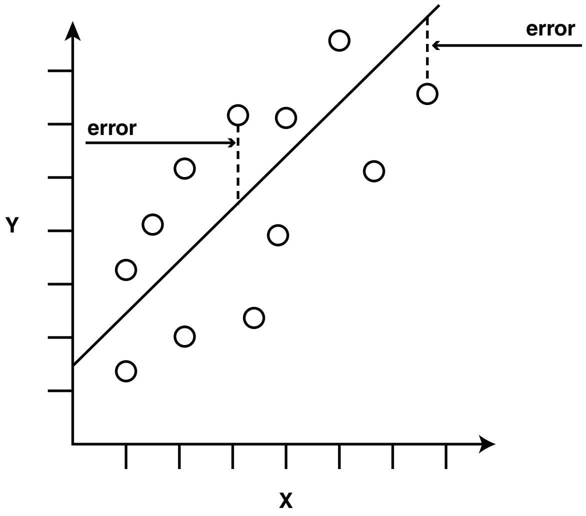
Linear Regression

Use linear regression to model relationships with a straight line, predicting outcomes based on one or more predictors.

A linear regression technique can be used if the relationship between the predictors and the target can be approximated with a straight line.

Regression with a single predictor is the easiest to visualize. Simple linear regression with a single predictor is shown in the following figure:

Figure 7-5 Linear Regression With a Single Predictor



Linear regression with a single predictor can be expressed with the following equation.

$$y = \theta_2\mathbf{x} + \theta_1 + e$$

The regression parameters in simple linear regression are:

- The **slope** of the line (${}_2$) — the angle between a data point and the regression line
- The ***y* intercept** (${}_1$) — the point where **x** crosses the *y* axis (**x** = 0)

Multivariate Linear Regression

Apply linear regression with multiple predictors, expanding the equation to include all relevant parameters.

The term **multivariate linear regression** refers to linear regression with two or more predictors (**x**$_1$, **x**$_2$, …, **x**$_n$). When multiple predictors are used, the regression line cannot be visualized in two-dimensional space. However, the line can be computed by expanding the equation for single-predictor linear regression to include the parameters for each of the predictors.

$$y = \theta_1 + \theta_2\mathbf{x}_1 + \theta_3\mathbf{x}_2 + ... \theta_n \mathbf{x}_{n-1} + e$$

Regression Coefficients

Evaluate the impact of predictors using regression coefficients in multivariate linear regression.

In multivariate linear regression, the regression parameters are often referred to as coefficients. When you build a multivariate linear regression model, the algorithm computes a coefficient for each of the predictors used by the model. The coefficient is a measure of the impact of the predictor **x** on the target *y*. Numerous statistics are available for analyzing the regression coefficients to evaluate how well the regression line fits the data.

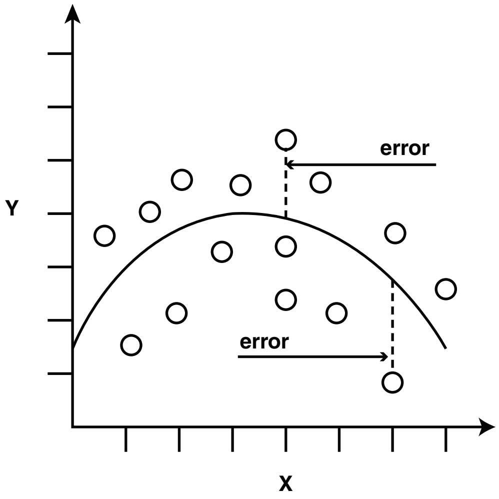
Nonlinear Regression

Represent complex relationships between predictors and targets using nonlinear regression techniques.

Often the relationship between **x** and ***y*** cannot be approximated with a straight line. In this case, a nonlinear regression technique can be used. Alternatively, the data can be preprocessed to make the relationship linear.

Nonlinear regression models define ***y*** as a function of **x** using an equation that is more complicated than the linear regression equation. In the following figure, **x** and ***y*** have a nonlinear relationship.

Figure 7-6 Nonlinear Regression With a Single Predictor



Multivariate Nonlinear Regression

Perform nonlinear regression with multiple predictors to capture data relationships.

The term **multivariate nonlinear regression** refers to nonlinear regression with two or more predictors (**x**$_1$, **x**$_2$, …, **x**$_n$). When multiple predictors are used, the nonlinear relationship cannot be visualized in two-dimensional space.

Confidence Bounds

Identify the range in which predicted values are likely to lie, enhancing prediction reliability.

A regression model predicts a numeric target value for each case in the scoring data. In addition to the predictions, some regression algorithms can identify confidence bounds, which are the upper and lower boundaries of an interval in which the predicted value is likely to lie.

When a model is built to make predictions with a given confidence, the confidence interval is produced along with the predictions. For example, a model predicts the value of a house to be $500,000 with a 95% confidence that the value is between $475,000 and $525,000.

Testing a Regression Model

Apply a regression model to test data, compare predicted values with actual ones, and use metrics to evaluate accuracy.

A regression model is tested by applying it to test data with known target values and comparing the predicted values with the known values.

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model.

Test metrics are used to assess how accurately the model predicts these known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

Regression Statistics

Use Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to assess the quality of regression models.

The Root Mean Squared Error and the Mean Absolute Error are commonly used statistics for evaluating the overall quality of a regression model. Different statistics may also be available depending on the regression methods used by the algorithm.

Root Mean Squared Error

Calculate the Root Mean Squared Error (RMSE) to determine the average squared distance of data points from the fitted line.

The following SQL expression calculates the RMSE:

```
SQRT(AVG((predicted_value - actual_value) * (predicted_value - actual_value)))
```

This formula shows the RMSE in mathematical symbols. The large sigma character represents summation; *j* represents the current predictor, and *n* represents the number of predictors.

Figure 7-7   Room Mean Squared Error

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Mean Absolute Error

Compute the Mean Absolute Error (MAE) to find the average of the absolute residuals (errors), indicating prediction accuracy.

The MAE is very similar to the RMSE but is less sensitive to large errors.

This SQL expression calculates the MAE.

```
AVG(ABS(predicted_value - actual_value))
```

This formula shows the MAE in mathematical symbols. The large sigma character represents summation; *j* represents the current predictor, and *n* represents the number of predictors.

Figure 7-8 Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{n} \sum_{j=1}^{n} | y_j - \hat{y}_j |$$

Regression Algorithms

Oracle Machine Learning supports these algorithms for regression: Generalized Linear Model (GLM), Neural Network (NN), Support Vector Machines (SVM), and XGBoost.

GLM and SVM algorithms are particularly suited for analysing data sets that have very high dimensionality (many attributes), including transactional and unstructured data.

- **Generalized Linear Model**

  GLM is a popular statistical technique for linear modeling. Oracle Machine Learning for SQL implements GLM for regression and for binary classification. GLM provides extensive coefficient statistics and model statistics, as well as row diagnostics. GLM also supports confidence bounds.

- **Neural Network**

  Neural Network is a powerful algorithm that can learn arbitrary nonlinear regression functions.

- **Support Vector Machine**

  SVM is a powerful, state-of-the-art algorithm for linear and nonlinear regression. OML4SQL implements SVM for regression, classification, and anomaly detection. SVM regression supports two kernels: the Gaussian kernel for nonlinear regression and the linear kernel for linear regression.

  > **Note:**
  >
  > OML4SQL uses the linear kernel SVM as the default regression algorithm.

- **XGBoost**

  XGBoost is machine learning algorithm for regression and classification that makes available the XGBoost open source package. Oracle Machine Learning for SQL XGBoost prepares training data, invokes XGBoost, builds and persists a model, and applies the model for prediction.

About Row Importance

Identify and rank influential rows in a data set using statistical leverage scores for dimensionality reduction.

Row importance technique is used in dimensionality reduction of large data sets. Row importance identifies the most influential rows of the data matrix. The rows with high importance are ranked by their importance scores. The "importance" of a row is determined by high statistical leverage scores. In CUR matrix decomposition, row importance is often combined with column (attribute) importance. Row importance can serve as a data preprocessing step prior to model building using regression, classification, and clustering.

Related Topics

- About CUR Matrix Decomposition
  CUR Matrix Decomposition is a low-rank matrix decomposition algorithm that is explicitly expressed in a small number of actual columns and/or actual rows of data matrix.
- Statistical Leverage Score
  Statistical leverage scores highlight the most representative columns or rows, aiding in the selection of important data points.
- CUR Matrix Decomposition Algorithm Configuration
  Configure the CUR Matrix Decomposition algorithm setting to build your model.

Row Importance Algorithms

Oracle Machine Learning supports CUR matrix decomposition algorithm to determine row and column (attribute) importance.

Popular algorithms for dimensionality reduction are Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and CUR Matrix Decomposition. All these algorithms apply low-rank matrix decomposition.

In CUR matrix decomposition, the attributes include 2-Dimensional numerical columns, levels of exploded 2D categorical columns, and attribute name or subname or value pairs for nested columns. To arrive at row importance or selection, the algorithm computes singular vectors, calculates leverage scores, and then selects rows. Row importance is performed when users specify `CURS_ROW_IMP_ENABLE` for the `CURS_ROW_IMPORTANCE` parameter in the settings table and the `case_id` column is present. Unless users explicitly specify, row importance is not performed.

Related Topics

- About Singular Value Decomposition
  SVD and the closely-related PCA are well established feature extraction methods that have a wide range of applications. Oracle Machine Learning for SQL implements Singular Value Decomposition (SVD) as a feature extraction algorithm and Principal Component Analysis (PCA) as a special scoring method for SVD models.
- About CUR Matrix Decomposition
  CUR Matrix Decomposition is a low-rank matrix decomposition algorithm that is explicitly expressed in a small number of actual columns and/or actual rows of data matrix.

About Time Series

Time series is a machine learning technique that forecasts target value based solely on a known history of target values. It is a specialized form of regression, known in the literature as auto-regressive modeling.

The input to time series analysis is a sequence of target values. A case id column specifies the order of the sequence. The case id can be of type `NUMBER` or a date type (date, datetime, timestamp with timezone, or timestamp with local timezone). Regardless of case id type, the user can request that the model include trend, seasonal effects or both in its forecast computation. When the case id is a date type, the user must specify a time interval (for example, month) over which the target values are to be aggregated, along with an aggregation procedure (for example, sum). Aggregation is performed by the algorithm prior to constructing the model.

The time series model provide estimates of the target value for each step of a time window that can include up to 30 steps beyond the historical data. Like other regression models, time series models compute various statistics that measure the goodness of fit to historical data.

Forecasting is a critical component of business and governmental decision making. It has applications at the strategic, tactical and operation level. The following are the applications of forecasting:

- Projecting return on investment, including growth and the strategic effect of innovations
- Addressing tactical issues such as projecting costs, inventory requirements and customer satisfaction
- Setting operational targets and predicting quality and conformance with standards

Related Topics

- About Regression
  Regression is a machine learning technique that predicts numeric values along a continuum.

Choosing a Time Series Model

Selecting a model depends on recognizing the patterns in the time series data. Consider trend, seasonality, or both that affect the data.

Time series data may contain patterns that can affect predictive accuracy. For example, during a period of economic growth, there may be an upward trend in sales. Sales may increase in specific seasons (bathing suits in summer). To accommodate such series, it can be useful to choose a model that incorporates trend, seasonal effects, or both.

Trend can be difficult to estimate, when you must represent trend by a single constant. For example, if there is a grow rate of 10%, then after 7 steps, the value doubles. Local growth rates, appropriate to a few time steps can easily approach such levels, but thereafter drop. **Damped trend** models can more accurately represent such data, by reducing cumulative trend effects. Damped trend models can better represent variability in trend effects over the historical data. Damped trend models are a good choice when the data have significant, but variable trend.

Since modeling attempts to reduce error, how error is measured can affect model predictions. For example, data that exhibit a wide range of values may be better represented by error as fraction of level. An error of a few hundred feet in the measurement of the height of a mountain may be equivalent to an error of an inch or two in the measurement of the height of a child. Errors that are measured relative to value are called **multiplicative errors**. Errors that are the same across values are called **additive errors**. If there are multiplicative effects in the model, then the error type is multiplicative. If there are no explicit multiplicative effects, error type is left to user specification. The type need not be the same across individual effects. For example, trend can be additive while seasonality is multiplicative. This particular mixed type effect combination defines the popular Holt-Winters model.

> **Note:**
>
> Multiplicative error is not an appropriate choice for data that contain zeros or negative values. Thus, when the data contains such values, it is best not to choose a model with multiplicative effects or to set error type to be multiplicative.

Automated Time Series Model Search

Automatically determine the best model type for time series forecasting if no specific model is defined.

If you do not specify a model type (EXSM_MODEL) the default behavior is for the algorithm to automatically determine the model type. The ESM settings are listed in DBMS_DATA_MINING — Algorithm Settings: Exponential Smoothing. Time Series model search considers a variety of models and selects the best one. For seasonal models, the seasonality is automatically determined.

The following example displays a sample code snippet that you can use for creating a model that automatically selects the best ESM model. In this example, EXSM_MODEL setting is not defined thereby allowing the algorithm to select the best model.

```
BEGIN DBMS_DATA_MINING.DROP_MODEL('ESM_SALES_FORECAST_1');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN

    v_setlst('ALGO_NAME')            := 'ALGO_EXPONENTIAL_SMOOTHING';
    v_setlst('EXSM_INTERVAL')        := 'EXSM_INTERVAL_QTR';
    v_setlst('EXSM_PREDICTION_STEP') := '4';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'ESM_SALES_FORECAST_1',
        MINING_FUNCTION     => 'TIME_SERIES',
        DATA_QUERY          => 'select * from ESM_SH_DATA',
        SET_LIST            => v_setlst,
        CASE_ID_COLUMN_NAME => 'TIME_ID',
        TARGET_COLUMN_NAME  => 'AMOUNT_SOLD');
END;
/
```

Time Series Statistics

Learn to evaluate model quality by applying commonly used statistics.

As with other regression functions, there are commonly used statistics for evaluating the overall model quality. An expert user can also specify one of these figures of merit as criterion to optimize by the model build process. Choosing an optimization criterion is not required because model-specific defaults are available.

Conditional Log-Likelihood

Log-likelihood is a figure of merit often used as an optimization criterion for models that provide probability estimates for predictions which depend on the values of the model's parameters.

The model probability estimates for the actual values in the training data then yields an estimate of the likelihood of the parameter values. Parameter values that yield high probabilities for the observed target values have high likelihood, and therefore indicate a good model. The calculation of log-likelihood depends on the form of the model.

Conditional log-likelihood breaks the parameters into two groups. One group is assumed to be correct and the other is assumed the source of any errors. Conditional log-likelihood is the log-likelihood of the latter group conditioned on the former group. For example, Exponential Smoothing (ESM) models make an estimate of the initial model state. The conditional log-likelihood of an ESM model is conditional on that initial model state (assumed to be correct). The ESM conditional log-likelihood is as follows:

$$L^{*}(\theta, X_0) = n\ln\left(\sum_{t=1}^{n} e_t^2/k^2(x_{t-1})\right) + 2\sum_{t=1}^{n}\ln|k(x_{t-1})|$$

where e_t is the error at time `t` and `k(x(t-1) )` is `1` for ESM models with additive errors and is the estimated level at the previous time step in models with multiplicative error.

Mean Square Error (MSE) and Other Error Measures

Compute Mean Square Error (MSE) to evaluate forecast accuracy. Use others metrics for additional error assessment.

The mean square error used as an optimization criterion, is computed as:

$$MSE = \sum_{t=1}^{n} {e_t}^2/n$$

where the error at time *t* is the difference between the actual and model one step ahead forecast value at time *t* for models with additive error and that difference divided by the one-step ahead forecast for models with multiplicative error.

> **Note:**
>
> These "forecasts" are for over periods already observed and part of the input time series.

Since time series models can forecast for each of multiple steps ahead, time series can measure the error associated with such forecasts. Average Mean Square Error (AMSE), another figure of merit, does exactly that. For each period in the input time series, it computes a multi-step forecast, computes the error of those forecasts and averages the errors. AMSE computes the individual errors exactly as MSE does taking cognizance of error type (additive or multiplicative). The number of steps, *k*, is determined by the user (default 3). The formula is as follows:

$$AMSE = \sum_{t=1}^{n}\left(\sum_{i=0}^{k-1} e_{t+i}^2/k\right)\Big/n$$

Other figure of merit relatives of MSE include the Residual Standard Error (RMSE), which is the square root of MSE, and the Mean Absolute Error (MAE) which is the average of the absolute value of the errors.

Irregular Time Series

Irregular time series are time series data where the time intervals between observed values are not equally spaced.

One common practice is for the time intervals between adjacent steps to be equally spaced. However, it is not always convenient or realistic to force such spacing on time series. Irregular time series do not make the assumption that time series are equally spaced, but instead use the case id’s date and time values to compute the intervals between observed values. Models are constructed directly on the observed values with their observed spacing. Oracle time series analysis handles irregular time series.

Build and Apply

Build a new time series model when new data arrives, producing statistics and forecasts during the build process.

Many of the Oracle Machine Learning for SQL functions have separate build and apply operations, because you can construct and potentially apply a model to many different sets of input data. However, time series input consists of the target value history only. Thus, there is only one set of appropriate input data. When new data arrive, good practice dictates that a new model be built. Since the model is only intended to be used once, the model statistics and forecasts are produced during model build and are available through the model views.

Time Series Algorithm

Oracle Machine Learning uses Exponential Smoothing to forecast from time series data.

Related Topics

- About Exponential Smoothing
  Exponential smoothing is a forecasting method for time series data. It is a moving average method where exponentially decreasing weights are assigned to past observations.

Machine Learning Algorithms

Provides an overview of Oracle Machine Learning algorithm concepts.

Topics:

- About Apriori
- About CUR Matrix Decomposition
- About Decision Tree
- About Expectation Maximization
- About Explicit Semantic Analysis
- About Exponential Smoothing
- About Generalized Linear Model
- About *k*-Means
- About MDL

- About Multivariate State Estimation Technique - Sequential Probability Ratio Test
- About Naive Bayes
- About Neural Network
- About NMF
- About O-Cluster
- Oracle Machine Learning for SQL with R Extensibility
- About Random Forest
- About Singular Value Decomposition
- About Support Vector Machine
- About XGBoost

About Apriori

Learn how to find associations involving rare events in a large number of items using Apriori.

An association machine learning problem can be decomposed into the following subproblems:

- Find all combinations of items in a set of transactions that occur with a specified minimum frequency. These combinations are called **frequent itemsets**.
- Calculate rules that express the probable co-occurrence of items within frequent itemsets.

Apriori calculates the probability of an item being present in a frequent itemset, given that another item or items is present.

Association rule machine learning is not recommended for finding associations involving rare events in problem domains with a large number of items. Apriori discovers patterns with frequencies above the minimum support threshold. Therefore, to find associations involving rare events, the algorithm must run with very low minimum support values. However, doing so potentially explodes the number of enumerated itemsets, especially in cases with a large number of items. This increases the execution time significantly. Classification or anomaly detection is more suitable for discovering rare events when the data has a high number of attributes.

The build process for Apriori supports parallel execution.

Related Topics

- Example: Calculating Rules from Frequent Itemsets
  Calculate association rules from frequent itemsets, using examples to illustrate rule generation and confidence calculation.
- *Oracle Database VLDB and Partitioning Guide*

Association Rules and Frequent Itemsets

Apriori calculates rules expressing probabilistic relationships between items in frequent itemsets, indicating item co-occurrence probabilities.

For example, a rule derived from frequent itemsets containing A, B, and C might state that if A and B are included in a transaction, then C is likely to also be included.

An association rule states that an item or group of items implies the presence of another item with some probability. Unlike decision tree rules, which predict a target, association rules express correlation.

Antecedent and Consequent

Defines antecedent and consequent in an Apriori algorithm.

The IF component of an association rule is known as the **antecedent**. The THEN component is known as the **consequent**. The antecedent and the consequent are disjoint; they have no items in common.

Oracle Machine Learning for SQL supports association rules that have one or more items in the antecedent and a single item in the consequent.

Confidence

Specify the minimum confidence for rules, representing the conditional probability of the consequent given the antecedent.

Rules have an associated confidence, which is the conditional probability that the consequent occurs given the occurrence of the antecedent. You can specify the minimum confidence for rules.

Data Preparation for Apriori

Prepare transactional data for Apriori by organizing it into case identifiers and associated values, for model processing.

Association models are designed to use transactional data. In transactional data, there is a one-to-many relationship between the case identifier and the values for each case. Each case ID/value pair is specified in a separate record (row).

Native Transactional Data and Star Schemas

Store transactional data in native or star schema formats, transforming non-native formats for Apriori processing.

Transactional data may be stored in native transactional format, with a non-unique case ID column and a values column, or it may be stored in some other configuration, such as a star schema. If the data is not stored in native transactional format, it must be transformed to a nested column for processing by the Apriori algorithm.

Related Topics

- Transactional Data
  Understand transactional data, where a case includes a collection of items like a market basket at checkout.
- *Oracle Machine Learning for SQL User’s Guide*

Items and Collections

Understand that transactional data associates a subset of possible items with each case, reflecting purchase patterns in a store.

In transactional data, a collection of items is associated with each case. The collection theoretically includes all possible members of the collection. For example, all products can theoretically be purchased in a single market-basket transaction. However, in actuality, only a tiny subset of all possible items are present in a given transaction; the items in the market-basket represent only a small fraction of the items available for sale in the store.

Sparse Data

Transactional data is typically sparse, with missing items indicating absence rather than null values.

Missing items in a collection indicate **sparsity**. Missing items may be present with a null value, or they may be missing.

Nulls in transactional data are assumed to represent values that are known but not present in the transaction. For example, three items out of hundreds of possible items might be purchased in a single transaction. The items that were not purchased are known but not present in the transaction.

Oracle Machine Learning assumes sparsity in transactional data. The Apriori algorithm is optimized for processing sparse data.

> **Note:**
>
> Apriori is not affected by Automatic Data Preparation.

Related Topics

- *Oracle Machine Learning for SQL User's Guide*

Improved Sampling

Use improved sampling techniques to determine appropriate sample sizes for association rule generation with performance guarantees.

Association rules (AR) can use a good sample size with performance guarantee, based on the work of Riondato and Upfal.The AR algorithm computes the sample size by the following inputs:

- *d*-index of the dataset
- Absolute error *ε*
- Confidence level *γ*

d-index is defined as the maximum integer *d* such that the dataset contains at least *d* transactions of length *d* at the minimum. It is the upper bound of Vapnik-Chervonenkis (VC) dimension. The AR algorithm computes *d*-index of the dataset by scanning the length of all transactions in the dataset.

Users specify absolute error *ε* and confidence level *γ* parameters. A large *d*-index, small AR support, small *ε* or large *γ* can cause a large sample size. The sample size theoretically guarantees that the absolute error of both the support and confidence of the approximated AR (from sampling) is less than *ε* compared to the exact AR with probability (or confidence level) at least *γ*. In this document this sample size is called AR-specific sample size.

Sampling Implementation

Specify sampling settings to determine sample sizes or rely on algorithm-calculated sample sizes for efficient rule generation.

Usage Notes

1. If `ODMS_SAMPLING` is unspecified or set as `ODMS_SAMPLING_DISABLE`, the sampling is not performed for AR and the exact AR is obtained.
2. If `ODMS_SAMPLING` is set as `ODMS_SAMPLING_ENABLE` and if `ODMS_SAMPLE_SIZE` is specified as positive integer number then the user-specified sample size (`ODMS_SAMPLE_SIZE`) is utilized. The sampling is performed in the general data preparation stage before the AR algorithm. The AR-specific sample size is not computed. The approximated AR is obtained.
3. If `ODMS_SAMPLING` is set as `ODMS_SAMPLING_ENABLE` and `ODMS_SAMPLE_SIZE` is not specified, the AR-specified sample size is computed and then sampling is performed in the AR algorithm. The approximated AR is obtained.

> **Note:**
>
> If the computed AR-specific sample size is larger than or equal to the total transaction size in the data set, the sampling is not performed and the exact AR is obtained.

If users do not have a good idea on the choice of sample size for AR, it is suggested to leave `ODMS_SAMPLE_SIZE` unspecified, only specify proper values for sampling parameters and let AR algorithm compute the suitable AR-specific sample size.

> **See Also:**
>
> DBMS_DATA_MINING — Machine Learning Function Settings for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Calculating Association Rules

Enumerate itemsets from transactions and calculate association rules.

The first step in association analysis is the enumeration of **itemsets**. An itemset is any combination of two or more items in a transaction.

Itemsets

Define itemsets as combinations of items within transactions, specifying the maximum number of items per itemset.

The maximum number of items in an itemset is user-specified. If the maximum is two, then all the item pairs are counted. If the maximum is greater than two, then all the item pairs, all the item triples, and all the item combinations up to the specified maximum are counted.

The following table shows the itemsets derived from the transactions shown in the following example, assuming that maximum number of items in an itemset is set to 3.

Table 7-7 Itemsets

Transaction	Itemsets
11	(B,D) (B,E) (D,E) (B,D,E)
12	(A,B) (A,C) (A,E) (B,C) (B,E) (C,E) (A,B,C) (A,B,E) (A,C,E) (B,C,E)
13	(B,C) (B,D) (B,E) (C,D) (C,E) (D,E) (B,C,D) (B,C,E) (B,D,E) (C,D,E)

Example 7-3 Sample Transactional Data

```
TRANS_ID   ITEM_ID
---------  -------------------
11         B
11         D
11         E
12         A
12         B
12         C
12         E
13         B
13         C
13         D
13         E
```

Frequent Itemsets

Identify frequently bought items (itemsets), filtered based on a minimum user-specified limits, to create rules.

Association rules are calculated from itemsets. If rules are generated from all possible itemsets, there can be a very high number of rules and the rules may not be very meaningful. Also, the model can take a long time to build. Typically it is desirable to only generate rules from itemsets that are well-represented in the data. **Frequent itemsets** are those that occur with a minimum frequency specified by the user.

The minimum frequent itemset **support** is a user-specified percentage that limits the number of itemsets used for association rules. An itemset must appear in at least this percentage of all the transactions if it is to be used as a basis for rules.

The following table shows the itemsets from Table 7-7 that are frequent itemsets with support > 66%.

Table 7-8 Frequent Itemsets

Frequent Itemset	Transactions	Support
(B,C)	2 of 3	67%
(B,D)	2 of 3	67%
(B,E)	3 of 3	100%
(C,E)	2 of 3	67%
(D,E)	2 of 3	67%
(B,C,E)	2 of 3	67%

Table 7-8 (Cont.) Frequent Itemsets

Frequent Itemset	Transactions	Support
(B,D,E)	2 of 3	67%

Related Topics

- About Apriori
  Learn how to find associations involving rare events in a large number of items using Apriori.

Example: Calculating Rules from Frequent Itemsets

Calculate association rules from frequent itemsets, using examples to illustrate rule generation and confidence calculation.

The following tables show the itemsets and frequent itemsets that were calculated in "Association". The frequent itemsets are the itemsets that occur with a minimum support of 67%; at least 2 of the 3 transactions must include the itemset.

Table 7-9 Itemsets

Transaction	Itemsets
11	(B,D) (B,E) (D,E) (B,D,E)
12	(A,B) (A,C) (A,E) (B,C) (B,E) (C,E) (A,B,C) (A,B,E) (A,C,E) (B,C,E)
13	(B,C) (B,D) (B,E) (C,D) (C,E) (D,E) (B,C,D) (B,C,E) (B,D,E) (C,D,E)

Table 7-10 Frequent Itemsets with Minimum Support 67%

Itemset	Transactions	Support
(B,C)	12 and 13	67%
(B,D)	11 and 13	67%
(B,E)	11, 12, and 13	100%
(C,E)	12 and 13	67%
(D,E)	11 and 13	67%
(B,C,E)	12 and 13	67%
(B,D,E)	11 and 13	67%

A rule expresses a conditional probability. Confidence in a rule is calculated by dividing the probability of the items occurring together by the probability of the occurrence of the antecedent.

For example, if B (antecedent) is present, what is the chance that C (consequent) is also present? What is the confidence for the rule "IF B, THEN C"?

As shown in Table 7-9:

- All 3 transactions include B (3/3 or 100%)
- Only 2 transactions include both B and C (2/3 or 67%)
- Therefore, the confidence of the rule "IF B, THEN C" is 67/100 or 67%.

The following table the rules that can be derived from the frequent itemsets in Table 7-10.

Table 7-11 Frequent Itemsets and Rules

Frequent Itemset	Rules	prob(antecedent and consequent) / prob(antecedent)	Confidence
(B,C)	(If B then C)	67/100	67%
	(If C then B)	67/67	100%
(B,D)	(If B then D)	67/100	67%
	(If D then B)	67/67	100%
(B,E)	(If B then E)	100/100	100%
	(If E then B)	100/100	100%
(C,E)	(If C then E)	67/67	100%
	(If E then C)	67/100	67%
(D,E)	(If D then E)	67/67	100%
	I(f E then D)	67/100	67%
(B,C,E)	(If B and C then E)	67/67	100%
	(If B and E then C)	67/100	67%
	(If C and E then B)	67/67	100%
(B,D,E)	(If B and D then E)	67/67	100%
	(If B and E then D)	67/100	67%
	(If D and E then B)	67/67	100%

If the minimum confidence is 70%, ten rules are generated for these frequent itemsets. If the minimum confidence is 60%, sixteen rules are generated.

Tip:

Increase the minimum confidence if you want to decrease the build time for the model and generate fewer rules.

Related Topics

- About Association
  Identify the probability of co-occurring items in a collection using Association.

Aggregates

Aggregates refer to the quantities associated with each item that the user opts for association rules model to aggregate.

There can be more than one aggregate. For example, the user can specify the model to aggregate both profit and quantity.

Example: Calculating Aggregates

This example shows how to calculate aggregates using the customer grocery purchase and profit data.

Calculating Aggregates for Grocery Store Data

Assume a grocery store has the following data:

Table 7-12 Grocery Store Data

Customer	Item A	Item B	Item C	Item D
Customer 1	Buys (Profit $5.00)	Buys (Profit $3.20)	Buys (Profit $12.00)	NA
Customer 2	Buys (Profit $4.00)	NA	Buys (Profit $4.20)	NA
Customer 3	Buys (Profit $3.00)	Buys (Profit $10.00)	Buys (Profit $14.00)	Buys (Profit $8.00)
Customer 4	Buys (Profit $2.00)	NA	NA	Buys (Profit $1.00)

The basket of each customer can be viewed as a transaction. The manager of the store is interested in not only the existence of certain association rules, but also in the aggregated profit if such rules exist.

In this example, one of the association rules can be (A, B)=>C for customer 1 and customer 3. Together with this rule, the store manager may want to know the following:

- The total profit of item A appearing in this rule
- The total profit of item B appearing in this rule
- The total profit for consequent C appearing in this rule
- The total profit of all items appearing in the rule

For this rule, the profit for item A is $5.00 + $3.00 = $8.00, for item B the profit is $3.20 + $10.00 = $13.20, for consequent C, the profit is $12.00 + $14.00 = $26.00, for the antecedent itemset (A, B) is $8.00 + $13.20 = $21.20. For the whole rule, the profit is $21.20 + $26.00 = $47.40.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Including and Excluding Rules

Explains including rules and excluding rules used in association.

Including rules enables a user to provide a list of items such that at least one item from the list must appear in the rules that are returned. Excluding rules enables a user to provide a list of items such that no item from the list can appear in the rules that are returned.

Note:

Since each association rule includes both antecedent and consequent, a set of including or excluding rules can be specified for antecedent while another set of including or excluding rules can be specified for consequent. Including or excluding rules can also be defined for the association rule.

Related Topics

- *Oracle Machine Learning for SQL User's Guide*
- *Oracle Database PL/SQL Packages and Types Reference*

Performance Impact for Aggregates

Aggregating data for association rules necessitates increased memory and processing power to ensure smooth performance.

For each item, the user may supply several columns to aggregate. It requires more memory to buffer the extra data and more time to compute the aggregate values.

Evaluating Association Rules

Evaluate association rules by using support and confidence.

Minimum support and confidence are used to influence the build of an association model. Support and confidence are also the primary metrics for evaluating the quality of the rules generated by the model. Additionally, Oracle Machine Learning for SQL supports lift for association rules. These statistical measures can be used to rank the rules and hence the usefulness of the predictions.

Support

Measure support to indicate the frequency of item co-occurrence, helping identify significant itemsets in transactions.

The support of a rule indicates how frequently the items in the rule occur together. For example, cereal and milk might appear together in 40% of the transactions. If so, the following rules each have a support of 40%:

```
cereal implies milk
milk implies cereal
```

Support is the ratio of transactions that include all the items in the antecedent and consequent to the number of total transactions.

Support can be expressed in probability notation as follows:

```
support(A implies B) = P(A, B)
```

Minimum Support Count

Define a minimum support count to ensure itemsets appear frequently enough in transactions to be considered significant.

Minimum support count defines minimum threshold in transactions that each rule must satisfy. When the number of transactions is unknown, the support percentage threshold parameter can be tricky to set appropriately. For this reason, support can also be expressed as a count of transactions, with the greater of the two thresholds being used to filter out infrequent itemsets. The default is `1` indicating that this criterion is not applied.

Related Topics

- Association Rules
  Identify the probability of co-occurring items in a collection within the data.
- *Oracle Machine Learning for SQL User's Guide*
- Frequent Itemsets
  Identify frequently bought items (itemsets), filtered based on a minimum user-specified limits, to create rules.

Confidence

The confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction.

Confidence is the conditional probability of the consequent given the antecedent. For example, cereal appears in 50 transactions; 40 of the 50 might also include milk. The rule confidence is:

```
cereal implies milk with 80% confidence
```

Confidence is the ratio of the rule support to the number of transactions that include the antecedent.

Confidence can be expressed in probability notation as follows.

```
confidence (A implies B) = P (B/A), which is equal to P(A, B) / P(A)
```

Related Topics

- Confidence
  Specify the minimum confidence for rules, representing the conditional probability of the consequent given the antecedent.
- Frequent Itemsets
  Identify frequently bought items (itemsets), filtered based on a minimum user-specified limits, to create rules.

Reverse Confidence

The reverse confidence of a rule is defined as the number of transactions in which the rule occurs divided by the number of transactions in which the consequent occurs.

Reverse confidence eliminates rules that occur because the consequent is frequent. The default is `0`.

Related Topics

- Confidence
  Specify the minimum confidence for rules, representing the conditional probability of the consequent given the antecedent.
- Example: Calculating Rules from Frequent Itemsets
  Calculate association rules from frequent itemsets, using examples to illustrate rule generation and confidence calculation.
- *Oracle Machine Learning for SQL User's Guide*
- *Oracle Database PL/SQL Packages and Types Reference*

Lift

Measure lift to evaluate the strength of a rule over random co-occurrence, ensuring the rule's predictive value.

Both support and confidence must be used to determine if a rule is valid. However, there are times when both of these measures may be high, and yet still produce a rule that is not useful. For example:

```
Convenience store customers who buy orange juice also buy milk with
a 75% confidence.
The combination of milk and orange juice has a support of 30%.
```

This at first sounds like an excellent rule, and in most cases, it would be. It has high confidence and high support. However, what if convenience store customers in general buy milk 90% of the time? In that case, orange juice customers are actually *less* likely to buy milk than customers in general.

A third measure is needed to evaluate the quality of the rule. Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support. It provides information about the improvement, the increase in probability of the consequent given the antecedent. Lift is defined as follows.

```
(Rule Support) /(Support(Antecedent) * Support(Consequent))
```

This can also be defined as the confidence of the combination of items divided by the support of the consequent. So in our milk example, assuming that 40% of the customers buy orange juice, the improvement would be:

```
30% / (40% * 90%)
```

which is 0.83 – an improvement of less than 1.

Any rule with an improvement of less than 1 does not indicate a real cross-selling opportunity, no matter how high its support and confidence, because it actually offers less ability to predict a purchase than does random chance.

> **Tip:**
>
> - Decrease the maximum rule length if you want to decrease the build time for the model and generate simpler rules.
> - Increase the minimum support if you want to decrease the build time for the model and generate fewer rules.

About CUR Matrix Decomposition

CUR Matrix Decomposition is a low-rank matrix decomposition algorithm that is explicitly expressed in a small number of actual columns and/or actual rows of data matrix.

CUR Matrix Decomposition was developed as an alternative to Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). CUR Matrix Decomposition selects columns and rows that exhibit high **statistical leverage** or large **influence** from the data matrix. By implementing the CUR Matrix Decomposition algorithm, a small number of most important attributes and/or rows can be identified from the original data matrix. Therefore, CUR Matrix

Decomposition is an important tool for exploratory data analysis. CUR Matrix Decomposition can be applied to a variety of areas and facilitates regression, classification, and clustering.

Related Topics

- Data Preparation for SVD
  Prepare data for Singular Value Decomposition using Automatic Data Preparation for numerical and categorical attributes.

Singular Vectors

Singular Value Decomposition (SVD) initiates CUR Matrix Decomposition by providing singular vectors essential for calculating leverage scores.

SVD returns left and right singular vectors for calculating column and row leverage scores. Perform SVD on the following matrix:

$$A \ \varepsilon \ \mathbf{R}^{mxn}$$

The matrix is factorized as follows:

$$A=U\Sigma V^{T}$$

where $U = [u^1 \ u^2 ... u^m]$ and $V = [v^1 \ v^2 ... v^n]$ are orthogonal matrices.

Σ is a diagonal m × n matrix with non-negative real numbers $\sigma 1,...,\sigma_\rho$ on the diagonal, where $\rho = \min \{m,n\}$ and σ_ξ is the ξ^{th} singular value of *A*.

Let u^ξ and v^ξ be the ξ^{th} left and right singular vector of *A*, the j^{th} column of A can thus be approximated by the top *k* singular vectors and corresponding singular values as:

$$A^j \approx \sum_{\xi=1}^{k} \left(\sigma_\xi u^\xi\right) v_j^\xi$$

where v^ξ_j is the j^{th} coordinate of the ξ^{th} right singular vector.

Statistical Leverage Score

Statistical leverage scores highlight the most representative columns or rows, aiding in the selection of important data points.

Leverage scores are statistics that determine which column (or rows) are most representative with respect to a rank subspace of a matrix. The statistical leverage scores represent the column (or attribute) and row importance. The normalized statistical leverage scores for all columns are computed from the top *k* right singular vectors as follows:

$$\pi_j = \frac{1}{k}\sum_{\xi=1}^{k} (v_j^\xi)^2$$

where *k* is called rank parameter and $j = 1,...,n$. Given that π_j>=0 and

$$\sum_{j=1}^{n} \pi_j = 1$$

, these scores form a probability distribution over the *n* columns.

Similarly, the normalized statistical leverage scores for all rows are computed from the top *k* left singular vectors as:

$$\pi'_i = \frac{1}{k}\sum_{\xi=1}^{k}(u_i^{\xi})^2$$

where $i = 1, ..., m$.

Column (Attribute) Selection and Row Selection

CUR Matrix Decomposition identifies and ranks attributes and rows by their leverage scores, ensuring high importance in analysis.

The CUR matrix decomposition in Oracle Machine Learning is designed for attribute and or row importance. It returns attributes and rows with high importance that are ranked by their leverage (importance) scores. Column (attribute) selection and row selection is the final stage in CUR. Attribute selection: Selects attributes with high leverage scores and reports their names, scores (as importance) and ranks (by importance).

Row selection: Selects rows with high leverage scores and reports their names, scores (as importance) and ranks (by importance).

1. CUR first selects the *j*th column (or attribute) of *A* with probability p_j= min $\{1, c\pi_j\}$ for all j ε $\{1, ..., n\}$
2. If users enable row selection, select *i*th row of *A* with probability p'_i = min $\{1, r\pi'_i\}$ for all i ε $\{1, ..., m\}$
3. Report the name (or ID) and leverage score (as importance) for all selected attributes (if row importance is disabled) or for all selected attributes and rows (if row importance is enabled).

c is the approximated (or expected) number of columns that users want to select, and *r* is the approximated (or expected) number of rows that users want to select.

To realize column and row selections, you need to calculate the probability to select each column and row.

Calculate the probability for each column as follows:

p_j = min $\{1, c\pi_j\}$

Calculate the probability for each row as follows:

p'_i = min$\{1,\ c\pi'_i\}$.

A column or row is selected if the probability is greater than some threshold.

CUR Matrix Decomposition Algorithm Configuration

Configure the CUR Matrix Decomposition algorithm setting to build your model.

Create a model with the algorithm specific settings. Define the algorithm name as `ALGO_CUR_DECOMPOSITION` and mining function as `ATTRIBUTE_IMPORTANCE`.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: CUR Matrix Decomposition for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Row Selection

To use this feature, specify the row importance setting `CURS_ROW_IMPORTANCE` to `CURS_ROW_IMP_ENABLE`.

> **Note:**
>
> The row selection is performed only when users specify that row importance is enabled and the `CASE_ID` column is present.

About Decision Tree

Decision Tree classifies data using a tree structure of rules, making predictions clear and easy to interpret.

Decision tree is a supervised machine learning algorithm used for classifying data. Decision tree has a tree structure built top-down that has a root node, branches, and leaf nodes. In some applications of machine learning, the reason for predicting one outcome or another may not be important in evaluating the overall quality of a model. In others, the ability to explain the reason for a decision can be crucial. You can use decision tree rules to validate models in such problems. The Decision Tree algorithm, like Naive Bayes, is based on conditional probabilities. Unlike Naive Bayes, decision trees generate **rules**. A rule is a conditional statement that can be understood by humans and used within a database to identify a set of records.

For example, a Marketing professional requires complete descriptions of customer segments to launch a successful marketing campaign. The Decision Tree algorithm is ideal for this type of application.

Use decision tree rules to validate models. If the rules make sense to a subject matter expert, then this validates the model.

Decision Tree Rules

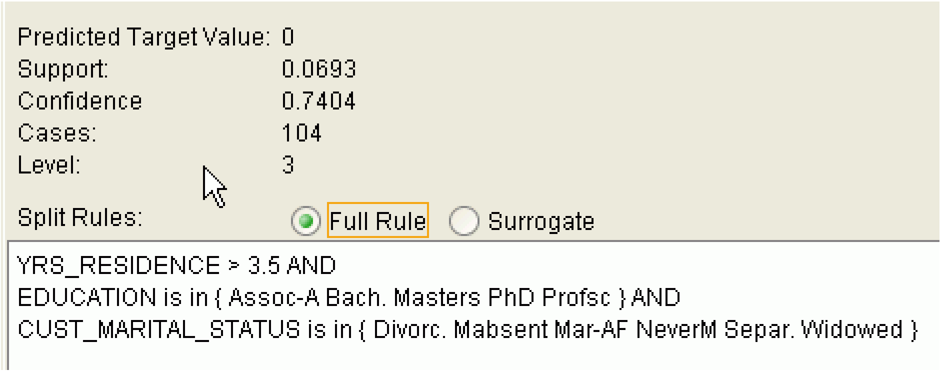
Decision Tree generates rules that provide transparency, helping validate models by showing the basis for predictions.

Oracle Machine Learning supports several algorithms that provide rules. In addition to decision trees, clustering algorithms provide rules that describe the conditions shared by the members of a cluster, and association rules provide rules that describe associations between attributes.

Rules provide **model transparency**, a window on the inner workings of the model. Rules show the basis for the model's predictions. Oracle Machine Learning supports a high level of model transparency. While some algorithms provide rules, *all* algorithms provide **model details**. You can examine model details to determine how the algorithm handles the attributes internally, including transformations and reverse transformations. Transparency is discussed in the context of data preparation and in the context of model building in *Oracle Machine Learning for SQL User’s Guide*.

The following figure shows a rule generated by a Decision Tree model. This rule comes from a decision tree that predicts the probability that customers increase spending if given a loyalty card. A target value of 0 means not likely to increase spending; 1 means likely to increase spending.

Figure 7-9 Sample Decision Tree Rule



Predicted Target Value:	0
Support:	0.0693
Confidence	0.7404
Cases:	104
Level:	3

Split Rules: ● Full Rule ○ Surrogate

YRS_RESIDENCE > 3.5 AND
EDUCATION is in { Assoc-A Bach. Masters PhD Profsc } AND
CUST_MARITAL_STATUS is in { Divorc. Mabsent Mar-AF NeverM Separ. Widowed }

The rule shown in the figure represents the conditional statement:

```
IF
          (current residence > 3.5 and has college degree and is single)
THEN
          predicted target value = 0
```

This rule is a full rule. A surrogate rule is a related attribute that can be used at apply time if the attribute needed for the split is missing.

Related Topics

- Understanding Reverse Transformations
- Model Detail Views for Decision Tree
- About Clustering
  Identify clusters of similar data objects, useful for exploring and preprocessing data without predefined categories.

- About Association
  Identify the probability of co-occurring items in a collection using Association.

Confidence and Support

Confidence and support are properties of rules. These statistical measures can be used to rank the rules and hence the predictions.

Support: The number of records in the training data set that satisfy the rule.

Confidence: The likelihood of the predicted outcome, given that the rule has been satisfied.

For example, consider a list of 1000 customers (1000 cases). Out of all the customers, 100 satisfy a given rule. Of these 100, 75 are likely to increase spending, and 25 are not likely to increase spending. The **support of the rule** is 100/1000 (10%). The **confidence of the prediction** (likely to increase spending) for the cases that satisfy the rule is 75/100 (75%).

Advantages of Decision Trees

Decision Tree is fast, accurate, and interpretable, suitable for binary and multiclass classification with minimal intervention.

The Decision Tree algorithm produces accurate and interpretable models with relatively little user intervention. The algorithm can be used for both binary and multiclass classification problems.

The algorithm is fast, both at build time and apply time. The build process for Decision Tree supports parallel execution. (Scoring supports parallel execution irrespective of the algorithm.)

Decision Tree scoring is especially fast. The tree structure, created in the model build, is used for a series of simple tests, (typically 2-7). Each test is based on a single predictor. It is a membership test: either IN or NOT IN a list of values (categorical predictor); or LESS THAN or EQUAL TO some value (numeric predictor).

Related Topics

- *Oracle Database VLDB and Partitioning Guide*

XML for Decision Tree Models

Learn about generating XML representation of Decision Tree models.

You can generate XML representing a Decision Tree model; the generated XML satisfies the definition specified in the Predictive Model Markup Language (PMML) version 2.1 specification.

Related Topics

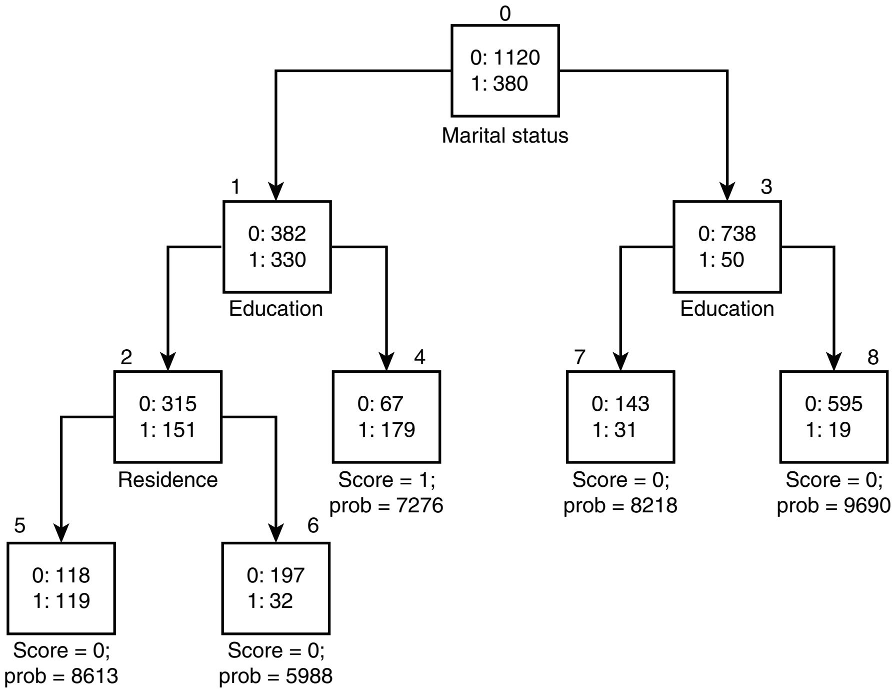
- https://dmg.org

Growing a Decision Tree

Predict a target value by a sequence of questions to form or grow a decision tree. A sample here shows how to grow a decision tree.

A decision tree predicts a target value by asking a sequence of questions. At a given stage in the sequence, the question that is asked depends upon the answers to the previous questions. The goal is to ask questions that, taken together, uniquely identify specific target values. Graphically, this process forms a tree structure.

Figure 7-10 Sample Decision Tree



The figure is a decision tree with nine nodes (and nine corresponding rules). The target attribute is binary: 1 if the customer increases spending, 0 if the customer does not increase spending. The first split in the tree is based on the `CUST_MARITAL_STATUS` attribute. The root of the tree (node 0) is split into nodes 1 and 3. Married customers are in node 1; single customers are in node 3.

The rule associated with node 1 is:

```
Node 1 recordCount=712,0 Count=382, 1 Count=330
CUST_MARITAL_STATUS isIN  "Married",surrogate:HOUSEHOLD_SIZE isIn "3""4-5"
```

Node 1 has 712 records (cases). In all 712 cases, the `CUST_MARITAL_STATUS` attribute indicates that the customer is married. Of these, 382 have a target of 0 (not likely to increase spending), and 330 have a target of 1 (likely to increase spending).

Splitting

Decision Tree uses homogeneity metrics like gini and entropy to create the most homogeneous child nodes.

During the training process, the Decision Tree algorithm must repeatedly find the most efficient way to split a set of cases (records) into two child nodes. Oracle Machine Learning offers two homogeneity metrics, **gini** and **entropy**, for calculating the splits. The default metric is gini.

Homogeneity metrics asses the quality of alternative split conditions and select the one that results in the most homogeneous child nodes. Homogeneity is also called **purity**; it refers to the degree to which the resulting child nodes are made up of cases with the same target value. The objective is to maximize the purity in the child nodes. For example, if the target can be either yes or no (does or does not increase spending), the objective is to produce nodes where most of the cases either increase spending or most of the cases do not increase spending.

Cost Matrix

Use a cost matrix to optimize Decision Tree scoring,

All classification algorithms, including Decision Tree, support a cost-benefit matrix at apply time. You can use the same cost matrix for building and scoring a decision tree model, or you can specify a different cost or benefit matrix for scoring.

Related Topics

- Costs
  Influence model decisions by specifying a cost matrix to minimize costly misclassifications.
- Priors and Class Weights
  Offset differences in data distribution with prior probabilities and class weights to produce useful classification results.

Preventing Over-Fitting

Prevent over-fitting with automatic pruning and configurable limits.

In principle, the Decision Tree algorithm can grow each branch of the tree deeply enough to perfectly classify the training examples. While this is sometimes a reasonable strategy, in fact it can lead to difficulties when there is noise in the data, or when the number of training examples is too small to produce a representative sample of the true target function. In either of these cases, this simple algorithm can produce trees that over-fit the training examples. Over-fit is a condition where a model is able to accurately predict the data used to create the model, but does poorly on new data presented to it.

To prevent over-fitting, Oracle Machine Learning supports automatic **pruning** and configurable **limit conditions** that control tree growth. Limit conditions prevent further splits once the conditions have been satisfied. Pruning removes branches that have insignificant predictive power.

Tuning the Decision Tree Algorithm

Fine tune the Decision Tree algorithm with various parameters.

The Decision Tree algorithm is implemented with reasonable defaults for splitting and termination criteria. However several build settings are available for fine tuning.

You can specify a homogeneity metric for finding the optimal split condition for a tree. The default metric is gini. The entropy metric is also available.

Settings for controlling the growth of the tree are also available. You can specify the maximum depth of the tree, the minimum number of cases required in a child node, the minimum number of cases required in a node in order for a further split to be possible, the minimum number of cases in a child node, and the minimum number of cases required in a node in order for a further split to be possible.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

The training data attributes are binned as part of the algorithm's data preparation. You can alter the number of bins used by the binning step. There is a trade-off between the number of bins used and the time required for the build.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Decision Tree for a listing and description of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Data Preparation for Decision Tree

The Decision Tree algorithm manages its own data preparation internally. It does not require pretreatment of the data.

Decision Tree is not affected by Automatic Data Preparation (ADP).

Related Topics

- Prepare the Data

About Expectation Maximization

Expectation Maximization (EM) estimates mixture models for variety of applications, enhancing clustering and anomaly detection.

Oracle Machine Learning uses EM to implement a distribution-based clustering algorithm (EM-clustering) and a distribution-based anomaly detection algorithm (EM Anomaly).

Expectation Step and Maximization Step

EM iteratively computes and maximizes likelihood to improve model accuracy, ensuring reliable clustering results.

Expectation maximization is an iterative method. It starts with an initial parameter guess. The parameter values are used to compute the likelihood of the current model. This is the Expectation step. The parameter values are then recomputed to maximize the likelihood. This is the Maximization step. The new parameter estimates are used to compute a new expectation and then they are optimized again to maximize the likelihood. This iterative process continues until model convergence.

Probability Density Estimation

You can compute reliable cluster assignment using probability density.

In density estimation, the goal is to construct a density function that captures how a given population is distributed. In probability density estimation, the density estimate is based on observed data that represents a sample of the population. Areas of high data density in the model correspond to the peaks of the underlying distribution.

Density-based clustering is conceptually different from distance-based clustering (for example *k*-Means) where emphasis is placed on minimizing inter-cluster and maximizing the intra-cluster distances. Due to its probabilistic nature, density-based clustering can compute reliable probabilities in cluster assignment. It can also handle missing values automatically.

A distribution-based anomaly detection algorithm identifies an object as an outlier if its probability density is lower than the density of other data records in a data set. The EM Anomaly algorithm can capture the underlying data distribution and thus flag records that do not fit the learned data distribution well.

Algorithm Enhancements

Expectation Maximization (EM) is enhanced to resolve some challenges in its standard form.

Although EM is well established as a distribution-based algorithm, it presents some challenges in its standard form. The Oracle Machine Learning for SQL implementation includes significant enhancements, such as scalable processing of large volumes of data and automatic parameter initialization. The strategies that OML4SQL uses to address the inherent limitations of EM clustering and EM Anomaly are described further.

> **Note:**
>
> The EM abbreviation is used here to refer to general EM technique for probability density estimation that is common for both EM Clustering and EM Anomaly.

Limitations of Standard Expectation Maximization:

- Scalability: EM has linear scalability with the number of records and attributes. The number of iterations to convergence tends to increase with growing data size (both rows and columns). EM convergence can be slow for complex problems and can place a significant load on computational resources.
- High dimensionality: EM has limited capacity for modeling high dimensional (wide) data. The presence of many attributes slows down model convergence, and the algorithm becomes less able to distinguish between meaningful attributes and noise. The algorithm is thus compromised in its ability to find correlations.
- Number of components: EM typically requires the user to specify the number of components. In most cases, this is not information that the user can know in advance.
- Parameter initialization: The choice of appropriate initial parameter values can have a significant effect on the quality of the model. Initialization strategies that have been used for EM have generally been computationally expensive.
- From components to clusters: In EM Clustering model, components are often treated as clusters. This approach can be misleading since cohesive clusters are often modeled by multiple components. Clusters that have a complex shape need to be modeled by multiple components. To accomplish this, the Oracle Machine Learning for SQL implementation of EM Clustering creates a component hierarchy based on the overlap of the distributions of the individual components. The OML4SQL EM Clustering algorithm employs agglomerative hierarchical clustering. The OML4SQL implementation of EM Custering produces an assignment of the model components to high-level clusters.
- Anomaly Detection: In EM Anomaly detection, an anomaly probability is used to classify whether an object is normal or anomalous. The EM algorithm estimates the probability density of a data record which is mapped to a probability of an anomaly.

Scalability

Expectation Maximization (EM) uses database parallel processing to achieve excellent scalability, efficiently handling large data sets.

The Oracle Machine Learning for SQL implementation of Expectation Maximization uses database parallel processing to achieve excellent scalability. EM computations naturally lend themselves to row parallel processing, and the partial results are easily aggregated. The parallel implementation efficiently distributes the computationally intensive work across secondary processes and then combines the partial results to produce the final solution.

Related Topics

- *Oracle Database VLDB and Partitioning Guide*

High Dimensionality

Process high dimensional data through Expectation Maximization.

The Oracle Machine Learning for SQL implementation of Expectation Maximization (EM) can efficiently process high-dimensional data with thousands of attributes. This is achieved through a two-fold process:

- The data space of single-column (not nested) attributes is analyzed for pair-wise correlations. Only attributes that are significantly correlated with other attributes are included in the EM mixture model. The algorithm can also be configured to restrict the dimensionality to the *M* most correlated attributes.
- High-dimensional (nested) numerical data that measures events of similar type is projected into a set of low-dimensional features that are modeled by EM. Some examples of high-dimensional, numerical data are: text, recommendations, gene expressions, and market basket data.

Number of Components

EM automatically determines the optimal number of components, improving model accuracy and avoiding overfitting.

Typical implementations of Expectation Maximization (EM) require the user to specify the number of model components. This is problematic because users do not generally know the correct number of components. Choosing too many or too few components can lead to over-fitting or under-fitting, respectively.

When model search is enabled, the number of EM components is automatically determined. The algorithm uses a held-aside sample to determine the correct number of components, except in the cases of very small data sets when Bayesian Information Criterion (BIC) regularization is used.

Parameter Initialization

Choosing appropriate initial parameter values can have a significant effect on the quality of the solution.

Expectation maximization (EM) is not guaranteed to converge to the global maximum of the likelihood function but may instead converge to a local maximum. Therefore different initial parameter values can lead to different model parameters and different model quality.

In the process of model search, the EM model is grown independently. As new components are added, their parameters are initialized to areas with poor distribution fit.

From Components to Clusters

Expectation Maximization produces assignment of model components to high-level clusters.

Expectation Maximization (EM) model components are often treated as clusters. However, this approach can be misleading. Cohesive clusters are often modeled by multiple components. The shape of the probability density function used in EM effectively predetermines the shape of the identified clusters. For example, Gaussian density functions can identify single peak symmetric clusters. Clusters of more complex shape need to be modeled by multiple components.

Ideally, high density areas of arbitrary shape must be interpreted as single clusters. To accomplish this, the Oracle Machine Learning for SQL implementation of EM builds a component hierarchy that is based on the overlap of the individual components' distributions. OML4SQL EM uses agglomerative hierarchical clustering. Component distribution overlap is measured using the Bhattacharyya distance function. Choosing an appropriate cutoff level in the hierarchy automatically determines the number of high-level clusters.

The OML4SQL implementation of EM produces an assignment of the model components to high-level clusters. Statistics like means, variances, modes, histograms, and rules additionally describe the high-level clusters. The algorithm can be configured to either produce clustering assignments at the component level or at the cluster level.

Expectation Maximization for Anomaly Detection

EM identifies anomalies based on probability density, ensuring accurate anomaly detection for better data integrity.

An object is identified as an outlier in an EM Anomaly model if its anomaly probability is greater than 0.5. A label of 1 denotes normal, while a label of 0 denotes anomaly. The EM technique models the underlying data distribution of a data set, and the probability density of a data record is translated into an anomaly probability.

The following example displays the code snippet used for anomaly detection using the Expectation Maximization algorithm. Specify the `EMCS_OUTLIER_RATE` setting to capture the desired rate of outliers in the training data set.

```
-- SET OUTLIER RATE IN SETTINGS TABLE - DEFAULT IS 0.05
--

BEGIN DBMS_DATA_MINING.DROP_MODEL('CUSTOMERS360MODEL_AD');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
  v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
  v_setlst('ALGO_NAME')         := 'ALGO_EXPECTATION_MAXIMIZATION';
  v_setlst('PREP_AUTO')         := 'ON';
  v_setlst('EMCS_OUTLIER_RATE') := '0.1';

  DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'CUSTOMERS360MODEL_AD',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'SELECT * FROM CUSTOMERS360_V',
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        SET_LIST            => v_setlst,
        TARGET_COLUMN_NAME  => NULL); -- NULL target indicates anomaly
detection
END;
/
```

To view the complete example, see https://github.com/oracle-samples/oracle-db-examples/blob/main/machine-learning/sql/23ai/oml4sql-anomaly-detection-em.sql.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Expectation Maximization

Configuring the Algorithm

Configure Expectation Maximization (EM).

In Oracle Machine Learning for SQL, EM can effectively model very large data sets (both rows and columns) without requiring the user to supply initialization parameters or specify the number of model components. While the algorithm offers reasonable defaults, it also offers flexibility.

The following list describes some of the configurable aspects of EM:

- Whether or not independent non-nested column attributes are included in the model. For EM Clustering, it is system-determined by default. For EM Anomaly, extreme values in each column attribute can indicate a potential outlier, even when the attribute itself has low dependency on other columns. Therefore, by default the algorithm disables attribute removal in EM Anomaly.
- Whether to use Bernoulli or Gaussian distribution for numerical attributes. By default, the algorithm chooses the most appropriate distribution, and individual attributes may use different distributions. When the distribution is user-specified, it is used for all numerical attributes.
- Whether the convergence criterion is based on a held-aside data set or on Bayesian Information Criterion (BIC). The convergence criterion is system-determined by default.
- The percentage improvement in the value of the log likelihood function that is required to add a new component to the model. The default percentage is 0.001.
- For EM Clustering, whether to define clusters as individual components or groups of components. Clusters are associated to groups of components by default.
- The maximum number of components in the model. If model search is enabled, the algorithm determines the number of components based on improvements in the likelihood function or based on regularization (BIC), up to the specified maximum.
- For EM Clustering, whether the linkage function for the agglomerative clustering step uses the nearest distance within the branch (single linkage), the average distance within the branch (average linkage), or the maximum distance within the branch (complete linkage). By default, the algorithm uses single linkage.
- For EM Anomaly, whether to specify the percentage of the data that is expected to be anomalous. If it is known in advance that the number of "suspicious" cases is a certain percentage of the data, then the outlier rate can be set to that percentage. The algorithm's default value is 0.05.

See Also:

DBMS_DATA_MINING —Algorithm Settings: Expectation Maximization for a listing and explanation of the available model settings.

Note:

The term hyperparameter is also interchangeably used for model setting.

Related Topics

- DBMS_DATA_MINING - Global Settings

Data Preparation for Expectation Maximization

Learn how to prepare data for Expectation Maximization (EM).

If you use Automatic Data Preparation (ADP), you do not need to specify additional data preparation for Expectation Maximization. ADP normalizes numerical attributes (in non-nested columns) when they are modeled with Gaussian distributions. ADP applies a topN binning transformation to categorical attributes.

Missing value treatment is not needed since Oracle Machine Learning for SQL algorithms handle missing values automatically. The EM algorithm replaces missing values with the mean in single-column numerical attributes that are modeled with Gaussian distributions. In other single-column attributes (categoricals and numericals modeled with Bernoulli distributions), NULLs are not replaced; they are treated as a distinct value with its own frequency count. In nested columns, missing values are treated as zeros.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

About Explicit Semantic Analysis

, Explicit Semantic Analysis (ESA) was introduced as an unsupervised algorithm for feature extraction and is enhanced as a supervised algorithm for classification.

As a feature extraction algorithm, ESA does not discover latent features but instead uses explicit features represented in an existing knowledge base. As a feature extraction algorithm, ESA is mainly used for calculating semantic similarity of text documents and for explicit topic modeling. As a classification algorithm, ESA is primarily used for categorizing text documents. Both the feature extraction and classification versions of ESA can be applied to numeric and categorical input data as well.

The input to ESA is a set of attributes vectors. Every attribute vector is associated with a concept. The concept is a feature in the case of feature extraction or a target class in the case of classification. For feature extraction, only one attribute vector may be associated with any feature. For classification, the training set may contain multiple attribute vectors associated with any given target class. These rows related to one target class are aggregated into one by the ESA algorithm.

The output of ESA is a sparse attribute-concept matrix that contains the most important attribute-concept associations. The strength of the association is captured by the weight value of each attribute-concept pair. The attribute-concept matrix is stored as a reverse index that lists the most important concepts for each attribute.

> **Note:**
>
> For feature extraction the ESA algorithm does not project the original feature space and does not reduce its dimensionality. ESA algorithm filters out features with limited or uninformative set of attributes.

The scope of classification tasks that ESA handles is different than the classification algorithms such as Naive Bayes and Support Vector Machine. ESA can perform large scale classification with the number of distinct classes up to hundreds of thousands. The large scale classification requires gigantic training data sets with some classes having significant number of training samples whereas others are sparsely represented in the training data set.

While projecting a document to the ESA topic space produces a high-dimensional sparse vector, it is unsuitable as an input to other machine learning algorithms. Embeddings are added to address this issue. In natural language processing embeddings refer to a set of language modeling and feature learning techniques in which words, phrases, or documents are mapped to vectors of real numbers. It entails a mathematical transformation from a multi-dimensional space to a continuous vector space with a considerably smaller dimension. Embeddings are usually built on top of an existing knowledge base to gather context data. This method is used to map sparse high-dimensional vectors to dense lower-dimensional vectors while keeping the ESA context available to other machine learning algorithms. The output is a doc2vec (document to vector) mapping, which can be used instead of "bag of words" approach. ESA embeddings allow you to utilize ESA models to generate embeddings for any text or other ESA input. This includes, but is not limited to, embeddings for single words.

To lower the dimensionality of a set of points, a sparse version of the random projection algorithm is utilized. In random projections, the original data is projected into a suitable lower-dimensional space in such a way that the distances between the points are roughly preserved. When compared to other approaches, random projection methods are noted for their power, simplicity, and low error rates. Many natural language tasks apply random projection methods.

```
mining_build_textmining_datadmsh.sqlCREATE_MODEL2


BEGIN DBMS_DATA_MINING.DROP_MODEL('ESA_text_sample_dense');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
  xformlist dbms_data_mining_transform.TRANSFORM_LIST;

  v_setlst DBMS_DATA_MINING.SETTING_LIST;

BEGIN
  v_setlst('PREP_AUTO')                  := 'ON';
  v_setlst('ALGO_NAME')                  := 'ALGO_EXPLICIT_SEMANTIC_ANALYS';
  v_setlst('ODMS_TEXT_POLICY_NAME')      := 'DMDEMO_ESA_POLICY';
  v_setlst('ESAS_MIN_ITEMS')             := '5';
  v_setlst('ODMS_TEXT_MIN_DOCUMENTS')    := '2';
  v_setlst('ESAS_EMBEDDINGS')            := 'ESAS_EMBEDDINGS_ENABLE';
  v_setlst('ESAS_EMBEDDING_SIZE')        := '1024';

  dbms_data_mining_transform.SET_TRANSFORM(
    xformlist, 'comments', null, 'comments', 'comments',
      'TEXT(POLICY_NAME:DMDEMO_ESA_POLICY)(TOKEN_TYPE:STEM)');

  DBMS_DATA_MINING.CREATE_MODEL2(
    model_name          => 'ESA_text_sample_dense',
    mining_function     => 'FEATURE_EXTRACTION',
    data_query          => 'SELECT * FROM mining_build_text',
    case_id_column_name => 'cust_id',
    set_list            => v_setlst,
    xform_list          => xformlist);
END;
/
```

To view the complete example, see https://github.com/oracle-samples/oracle-db-examples/blob/main/machine-learning/sql/23ai/oml4sql-feature-extraction-text-mining-esa.sql.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Explicit Semantic Analysis

ESA for Text Analysis

Learn how Explicit Semantic Analysis (ESA) can be used for machine learning operations on text.

Explicit knowledge often exists in text form. Multiple knowledge bases are available as collections of text documents. These knowledge bases can be generic, for example, Wikipedia, or domain-specific. Data preparation transforms the text into vectors that capture attribute-concept associations. ESA is able to quantify semantic relatedness of documents even if they do not have any words in common. The function FEATURE_COMPARE can be used to compute semantic relatedness.

Related Topics

- *Oracle Database SQL Language Reference*

Data Preparation for ESA

Automatic Data Preparation normalizes input vectors to a unit length for Explicit Semantic Analysis (ESA).

When there are missing values in columns with simple data types (not nested), ESA replaces missing categorical values with the mode and missing numerical values with the mean. When there are missing values in nested columns, ESA interprets them as sparse. The algorithm replaces sparse numeric data with zeros and sparse categorical data with zero vectors. The Oracle Machine Learning for SQL data preparation transforms the input text into a vector of real numbers. These numbers represent the importance of the respective words in the text.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Explicit Semantic Analysis for a listing and explanation of the available model settings.

Note:

The term hyperparameter is also interchangeably used for model setting.

Scoring with ESA

A typical feature extraction application of Explicit Semantic Analysis (ESA) is to identify the most relevant features of a given input and score their relevance. Scoring an ESA model produces data projections in the concept feature space.

If an ESA model is built from an arbitrary collection of documents, then each one is treated as a feature. You can then identify the most relevant documents in the collection. The feature extraction functions are: `FEATURE_DETAILS`, `FEATURE_ID`, `FEATURE_SET`, `FEATURE_VALUE`, and `FEATURE_COMPARE`. The same functions are utilized in the implementation of ESA embeddings, but the space of the features is different. The names of features for ESA embeddings are successive integers starting with 1. The output of `FEATURE_ID` is numeric. Feature IDs in the output of `FEATURE_SET` and `FEATURE_DETAILS` are also numeric.

A typical classification application of ESA is to predict classes of a given document and estimate the probabilities of the predictions. As a classification algorithm, ESA implements the following scoring functions: `PREDICTION`, `PREDICTION_PROBABILITY`, `PREDICTION_SET`, `PREDICTION_DETAILS`, `PREDICTION_COST`.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*
- *Oracle Database SQL Language Reference*

Scoring Large ESA Models

Optimize performance by adjusting the System Global Area (SGA) to accommodate large ESA models, ensuring efficient model scoring.

Building an Explicit Semantic Analysis (ESA) model on a large collection of text documents can result in a model with many features or titles. The model information for scoring is loaded into SGA as a shared (shared pool size) library cache object. Different SQL predictive queries can reference this object. When the model size is large, it is necessary to set the SGA parameter in the database to a sufficient size that accommodates large objects. If the SGA is too small, the model may need to be re-loaded every time it is referenced which is likely to lead to performance degradation.

Terminologies in Explicit Semantic Analysis

Discusses the terms associated with Explicit Semantic Analysis (ESA).

Multi-target Classification

The training items in these large scale classifications belong to several classes. The goal of classification in such case is to detect possible multiple target classes for one item. This kind of classification is called multi-target classification. The target column for ESA-based classification is extended. Collections are allowed as target column values. The collection type for the target in ESA-based classification is `ORA_MINING_VARCHAR2_NT`.

Large-scale classification

Large-scale classification applies to ontologies that contain gigantic numbers of categories, usually ranging in tens or hundreds of thousands. This large-scale classification also requires gigantic training datasets which are usually unbalanced, that is, some classes may have significant number of training samples whereas others may be sparsely represented in the training dataset. Large-scale classification normally results in multiple target class assignments for a given test case.

Topic modeling

Topic modelling refers to derivation of the most important topics of a document. Topic modeling can be explicit or latent. Explicit topic modeling results in the selection of the most relevant topics from a pre-defined set, for a given document. Explicit topics have names and can be verbalized. Latent topic modeling identifies a set of latent topics characteristic for a collection of documents. A subset of these latent topics is associated with every document under examination. Latent topics do not have verbal descriptions or meaningful interpretation.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

About Exponential Smoothing

Exponential smoothing is a forecasting method for time series data. It is a moving average method where exponentially decreasing weights are assigned to past observations.

Exponential smoothing methods have been widely used in forecasting for over half a century. A forecast is a prediction based on historical data and patterns. preIt has applications at the strategic, tactical, and operation level. For example, at a strategic level, forecasting is used for projecting return on investment, growth and the effect of innovations. At a tactical level, forecasting is used for projecting costs, inventory requirements, and customer satisfaction. At an operational level, forecasting is used for setting targets and predicting quality and conformance with standards.

In its simplest form, exponential smoothing is a moving average method with a single parameter which models an exponentially decreasing effect of past levels on future values. With a variety of extensions, exponential smoothing covers a broader class of models than other well-known approaches, such as the Box-Jenkins auto-regressive integrated moving average (ARIMA) approach. Oracle Machine Learning for SQL implements exponential smoothing using a state of the art state space method that incorporates a single source of error (SSOE) assumption which provides theoretical and performance advantages.

Exponential smoothing is extended to the following:

- A matrix of models that mix and match error type (additive or multiplicative), trend (additive, multiplicative, or none), and seasonality (additive, multiplicative, or none)
- Models with damped trends.
- Models that directly handle irregular time series and time series with missing values.
- Multiple time series models

> **See Also:**
>
> Ord, J.K., et al, *Time Series Forecasting: The Case for the Single Source of Error State Space Approach, Working Paper*, Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia, April 2, 2005.

Exponential Smoothing Models

Exponential Smoothing models are a broad class of forecasting models that are intuitive, flexible, and extensible.

Members of this class include simple, single parameter models that predict the future as a linear combination of a previous level and a current shock. Extensions can include parameters for linear or non-linear trend, trend damping, simple or complex seasonality, related series, various forms of non-linearity in the forecasting equations, and handling of irregular time series.

Exponential smoothing assumes that a series extends infinitely into the past, but that influence of past on future, decays smoothly and exponentially fast. The smooth rate of decay is expressed by one or more smoothing constants. The **smoothing constants** are parameters that the model estimates. The assumption is made practical for modeling real world data by using an equivalent recursive formulation that is only expressed in terms of an estimate of the current level based on prior history and a shock to that estimate dependent on current conditions only.The procedure requires an estimate for the time period just prior to the first observation, that encapsulates all prior history. This initial observation is an additional model parameter whose value is estimated by the modeling procedure.

Components of ESM such as trend and seasonality extensions, can have an additive or multiplicative form. The simpler additive models assume that shock, trend, and seasonality are linear effects within the recursive formulation.

Simple Exponential Smoothing

Simple exponential smoothing assumes the data fluctuates around a stationary mean, with no trend or seasonal pattern.

In a simple Exponential Smoothing model, each forecast (smoothed value) is computed as the weighted average of the previous observations, where the weights decrease exponentially depending on the value of smoothing constant α. Values of the smoothing constant, α, near one, put almost all weight on the most recent observations. Values of α near zero allows the distant past observations to have a large influence.

Models with Trend but No Seasonality

The preferred form of additive (linear) trend is sometimes called Holt’s method or double exponential smoothing.

Models with trend add a smoothing parameter γ and optionally a damping parameter φ. The damping parameter smoothly dampens the influence of past linear trend on future estimates of level, often improving accuracy.

Models with Seasonality but No Trend

When the time series average does not change over time (stationary), but is subject to seasonal fluctuations, the appropriate model has seasonal parameters but no trend.

Seasonal fluctuations are assumed to balance out over periods of length *m*, where *m* is the number of seasons, For example, *m*=4 might be used when the input data are aggregated quarterly. For models with additive errors, the seasonal parameters must sum to zero. For models with multiplicative errors, the product of seasonal parameters must be one.

Models with Trend and Seasonality

Holt and Winters introduced both trend and seasonality in an Exponential Smoothing model.

The original model, also known as Holt-Winters or triple exponential smoothing, considered an additive trend and multiplicative seasonality. Extensions include models with various combinations of additive and multiplicative trend, seasonality and error, with and without trend damping.

Multiple Time Series Models

Multiple time series is a convenience operation for constructing multiple time series models with a common time interval for use as input to a time series regression.

One of the time series models is identified as the target time series of interest. All of the time series output is produced for the target. The other time series are assumed to be correlated with the target. This operation produces backcasts and forecasts on each time series and computes upper and lower confidence bounds for the identified target series. This operation can be used to forecast a wide variety of events, such as rainfall, sales, and customer satisfaction.

In the example of weather forecasting, the temperature and humidity attributes can be considered as the dependent or correlated time series and rainfall can be identified as the target time series.

Related Topics

- Model Detail Views for Exponential Smoothing

Prediction Intervals

To compute prediction intervals, an Exponential Smoothing (ESM) model is divided into three classes.

The simplest class is the class of linear models, which include, among others, simple ESM, Holt's method, and additive Holt-Winters. Class 2 models (multiplicative error, additive components) make an approximate correction for violations of the Normality assumption. Class 3 modes use a simple simulation approach to calculate prediction intervals.

Data Preparation for Exponential Smoothing Models

Prepare your data for exponential smoothing by providing input data, aggregation methods, and model build parameters.

To build an ESM model, you must supply the following :

- Input data
- An aggregation level and method, if the case id is a date type
- Partitioning column, if the data are partitioned

In addition, for a greater control over the build process, the user may optionally specify model build parameters, all of which have defaults:

- Model
- Error type
- Optimization criterion

- Forecast Window
- Confidence level for forecast bounds
- Missing value handling
- Whether the input series is evenly spaced

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Exponential Smoothing Models for a listing and explanation of the available model settings.
>
> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Input Data

Time series analysis requires ordered input data. Hence, each data row must consist of an [index, value] pair, where the index specifies the ordering.

When you create an Exponential Smoothing (ESM) model using the `CREATE_MODEL` or the `CREATE_MODEL2` procedure, the `CASE_ID_COLUMN_NAME` and the `TARGET_COLUMN_NAME` parameters are used to specify the columns used to compute the input indices and the observed time series values, respectively. The time column bears Oracle number, or Oracle date, timestamp, timestamp with time zone, or timestamp with local time zone. When the case id column is of type Oracle `NUMBER`, the model considers the input time series to be equally spaced. Only the ordinal position matters, with a lower number indicating a later time. In particular, the input time series is sorted based on the value of `case_id` (time label). The case_id column cannot contain missing values. To indicate a gap, the value column can contain missing values as `NULL`. The magnitude of the difference between adjacent time labels is irrelevant and is not used to calculate the spacing or gap size. Integer numbers passed as `CASE_ID` are assumed to be non-negative.

ESM also supports partitioned models and in such cases, the input table contains an extra column specifying the partition. All [index, value] pairs with the same partition ID form one complete time series. The Exponential Smoothing algorithm constructs models for each partition independently, although all models use the same model settings.

Data properties may result in a warning notice, or settings may be disregarded. If the user sets a model with a multiplicative trend, multiplicative seasonality, or both, and the data contains values Y_t<= 0, the model type is set to default. If the series contains fewer values than the number of seasons given by the user, then the seasonality specifications are ignored and a warning is issued.

If the user has selected a list of predictor series using the parameter `EXSM_SERIES_LIST`, the input data can also include up to twenty additional time series columns.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings:Exponential Smoothing

Accumulation

Use accumulation procedures for date-type columns to generate equally spaced time series data.

For the Exponential Smoothing algorithm, the accumulation procedure is applied when the column is a date type (`date`, `datetime`, `timestamp`, `timestamp with timezone`, or `timestamp with local timezone`). The case id can be a `NUMBER` column whose sort index represents the position of the value in the time series sequence of values. The case id column can also be a date type. A date type is accumulated in accordance with a user specified accumulation window. Regardless of type, the case id is used to transform the column into an equally spaced time series. No accumulation is applied for a case id of type `NUMBER`. As an example, consider a time series about promotion events. The time column contains the date of each event, and the dates can be unequally spaced. The user must specify the spacing interval, which is the spacing of the accumulated or transformed equally spaced time series. In the example, if the user specifies the interval to be month, then an equally spaced time series with profit for each calendar month is generated from the original time series. Setting `EXSM_INTERVAL` is used to specify the spacing interval. The user must also specify a value for `EXSM_ACCUMULATE`, for example, `EXSM_ACCU_MAX`, in which case the equally spaced monthly series would contain the maximum profit over all events that month as the observed time series value.

Missing Value

Handle missing values effectively in your time series data for reliable exponential smoothing models.

Input time series can contain missing values. A `NULL` entry in the target column indicates a missing value. When the time column is of the type datetime, the accumulation procedure can also introduce missing values. The setting `EXSM_SETMISSING` can be used to specify how to handle missing values. The special value `EXSM_MISS_AUTO` indicates that, if the series contains missing values it is to be treated as an irregular time series.

> **Note:**
>
> Missing value handling setting must be compatible with model setting, otherwise an error is thrown.

Prediction

Specify the prediction window for your exponential smoothing model to generate accurate forecasts.

Setting `EXSM_PREDICTION_STEP` can be used to specify the prediction window. The prediction window is expressed in terms of number of intervals (setting `EXSM_INTERVAL`), when the time column is of the type datetime. If the time column is a number then the prediction window is the number of steps to forecast. Regardless of whether the time series is regular or irregular, `EXSM_PREDICTION_STEP` specifies the prediction window.

See Also:

Oracle Database PL/SQL Packages and Types Reference for a listing and explanation of the available model settings.

Note:

The term hyperparameter is also interchangeably used for model setting.

Parallellism by Partition

Enhance performance by processing time series data in parallel, using partitioning for efficient model building.

For example, a user can choose `PRODUCT_ID` as one partition column and can generate forecasts for different products in a model build. Although a distinct smoothing model is built for each partition, all partitions share the same model settings. For example, if setting `EXSM_MODEL` is set to `EXSM_SIMPLE`, all partition models will be simple Exponential Smoothing models. Time series from different partitions can be distributed to different processes and processed in parallel. The model for each time series is built serially.

About Generalized Linear Model

The Generalized Linear Model (GLM) includes and extends the class of linear models which address and accommodate some restrictive assumptions of the linear models.

Linear models make a set of restrictive assumptions, most importantly, that the target (dependent variable *y*) is normally distributed conditioned on the value of predictors with a constant variance regardless of the predicted response value. The advantage of linear models and their restrictions include computational simplicity, an interpretable model form, and the ability to compute certain diagnostic information about the quality of the fit.

GLM relaxes these restrictions, which are often violated in practice. For example, binary (yes/no or 0/1) responses do not have same variance across classes. Furthermore, the sum of terms in a linear model typically can have very large ranges encompassing very negative and very positive values. For the binary response example, we would like the response to be a probability in the range [0,1].

GLM accommodates responses that violate the linear model assumptions through two mechanisms: a link function and a variance function. The link function transforms the target range to potentially -infinity to +infinity so that the simple form of linear models can be maintained. The variance function expresses the variance as a function of the predicted response, thereby accommodating responses with non-constant variances (such as the binary responses).

Oracle Machine Learning for SQL includes two of the most popular members of the GLM family of models with their most popular link and variance functions:

- **Linear regression** with the identity link and variance function equal to the constant 1 (constant variance over the range of response values).
- **Logistic regression**

In other words, the methods of linear regression assume that the target value ranges from minus infinity to infinity and that the target variance is constant over the range. The logistic regression target is either 0 or 1. A logistic regression model estimate is a probability. The job of the link function in logistic regression is to transform the target value into the required range, minus infinity to infinity.

GLM Function	Default Link Function	Other Supported Link Functions
Linear regression (gaussian)	identity	none
Logistic regression (binomial)	logit	probit, cloglog, cauchit, and binomial variance

Related Topics

- Linear Regression
  Use linear regression to model relationships with a straight line, predicting outcomes based on one or more predictors.
- Linear Regression
  GLM supports linear regression, assuming no target transformation and constant variance over target values.
- Logistic Regression
  GLM implements binary logistic regression, transforming target values into a probability scale for classification.

Logit Link Function

The logit link transforms a probability into the log of the odds ratio. The odds ratio is the ratio of the predicted probability of the positive to the predicted probability of the negative class. The log of the odds ratio has the appropriate range.

The odds ratio is a measure of the evidence for or against the positive target class. Odds ratios can be associated with particular predictor value. Odds ratios are naturally multiplicative, which makes the log of odds ratios additive. The log-odds ratio interprets the influence of a predictor as additive evidence for or against the positive class.

An advantage of the logit link is that the training data can be sampled independently from the two classes. This can be very significant in cases in which one class is rare or costly, such as the instances of a disease. Analysis of disease factors can be done directly from a sample of healthy people and a sample of people with the disease. This type of sampling is known as retrospective sampling.

For logistic regression, the logit link is the default. For technical reasons, this link is called the canonical link.

Probit Link Function

Probit link uses standard normal distribution to transform target values, ideal for normally distributed targets.

One approach to transforming the range of a probability to the range minus infinity to infinity is to choose a probability distribution that is defined on that range and assign the distribution value that corresponds to the probability as the target value. For example, the probabilities, 0, 0.5 and 1.0 corresponds to the value -infinity, 0 and infinity in a standard normal distribution. An inverse cumulative distribution function is a function that determines the value that corresponds to a probability. In this approach, a user matches the particular probability distribution to assumptions regarding the distribution of the target. Users often find transformation of a target using the target's known associated distribution as natural. The probit link takes this approach, using the standard normal distribution. An example use case is an analysis of high blood pressure. Blood pressure is assumed to have a normal distribution.

Cloglog Link Function

Cloglog link models extreme events effectively, transforming target values using Gumbel distribution.

The Complimentary Log-Log (cloglog) link is another example of using an inverse cumulative distribution function to transform the target. It differs from logit and probit function because it is asymmetric. It works best when the chance of an event is extremely low or extremely high. Gumbel described these extreme value distributions. The cloglog model is closely related to continuous-time models for event occurrence. The cloglog link function corresponds to Gumbel CDF. The precipitation from the worst rainstorm in 100 years is an example of data that follows an extreme value distribution (the hundred year rain).

Cauchit Link Function

Cauchit link uses the Cauchy distribution to transform target values, suitable for data with infinite variance.

The Cauchit link is another application of an inverse cumulative distribution function to transform the target. In this case, the distribution is the Cauchy distribution. The Cauchy distribution is symmetric, however, it has infinite variance. An infinite variance means the probability decays slowly as the values become more extreme. Such distributions are called fat-tailed. The Cauchit link is often used where fewer assumptions are justified with respect to the distribution of the target. The Cauchit link is used to measure data in binomial form when the variance is not considered to be finite.

GLM in Oracle Machine Learning

Learn how Oracle Machine Learning implements the Generalized Linear Model (GLM) algorithm.

GLM is a parametric modeling technique. Parametric models make assumptions about the distribution of the data. When the assumptions are met, parametric models can be more efficient than non-parametric models.

The challenge in developing models of this type involves assessing the extent to which the assumptions are met. For this reason, quality diagnostics are key to developing quality parametric models.

Interpretability and Transparency

You can interpret and understand key characteristics of Generalized Linear Model (GLM) model through model details and global details.

You can interpret Oracle Machine Learnings' GLM with ease. Each model build generates many statistics and diagnostics. Transparency is also a key feature: model details describe key characteristics of the coefficients, and global details provide high-level statistics.

Related Topics

- Tuning and Diagnostics for GLM
  Tuning and diagnostics in GLM help optimize model performance and quality through detailed evaluations.

Wide Data

Generalized Linear Model(GLM) handles wide data efficiently, building quality models with numerous predictors.

GLM in Oracle Machine Learning is uniquely suited for handling wide data. The algorithm can build and score quality models that use a virtually limitless number of predictors (attributes). The only constraints are those imposed by system resources.

Confidence Bounds

Predict confidence bounds through the Generalized Linear Model (GLM) algorithm.

GLM have the ability to predict confidence bounds. In addition to predicting a best estimate and a probability (classification only) for each row, GLM identifies an interval wherein the prediction (regression) or probability (classification) lies. The width of the interval depends upon the precision of the model and a user-specified confidence level.

The confidence level is a measure of how sure the model is that the true value lies within a confidence interval computed by the model. A popular choice for confidence level is 95%. For example, a model might predict that an employee's income is $125K, and that you can be 95% sure that it lies between $90K and $160K. Oracle Machine Learning for SQL supports 95% confidence by default, but that value can be configured.

> **Note:**
>
> Confidence bounds are returned with the coefficient statistics. You can also use the `PREDICTION_BOUNDS` SQL function to obtain the confidence bounds of a model prediction.

Related Topics

- *Oracle Database SQL Language Reference*

Ridge Regression

Understand the use of ridge regression for singularity (exact multicollinearity) in data.

The best regression models are those in which the predictors correlate highly with the target, but there is very little correlation between the predictors themselves. **Multicollinearity** is the term used to describe multivariate regression with correlated predictors.

Ridge regression is a technique that compensates for multicollinearity. Oracle Machine Learning for SQL supports ridge regression for both regression and classification machine learning techniques. The algorithm automatically uses ridge if it detects singularity (exact multicollinearity) in the data.

Information about singularity is returned in the global model details.

Related Topics

- Global Model Statistics for Linear Regression
  Generalized Linear Model regression models generate the following statistics.
- Global Model Statistics for Logistic Regression
  GLM generates global statistics for logistic regression, supporting model assessment.

Configuring Ridge Regression

Configure ridge regression through build settings.

You can choose to explicitly enable ridge regression by specifying a build setting for the model. If you explicitly enable ridge, you can use the system-generated ridge parameter or you can supply your own. If ridge is used automatically, the ridge parameter is also calculated automatically.

The configuration choices are summarized as follows:

- Whether or not to override the automatic choice made by the algorithm regarding ridge regression
- The value of the ridge parameter, used only if you specifically enable ridge regression.

> **See Also:**
>
> *Oracle Database PL/SQL Packages and Types Reference* for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Ridge and Confidence Bounds

Models built with ridge regression do not support confidence bounds.

Related Topics

- Confidence Bounds
  Predict confidence bounds through the Generalized Linear Model (GLM) algorithm.

Ridge and Data Preparation

Learn about preparing data for ridge regression.

When ridge regression is enabled, different data preparation is likely to produce different results in terms of model coefficients and diagnostics. Oracle recommends that you enable Automatic Data Preparation for Generalized Linear Model models, especially when ridge regression is used.

Related Topics

- Data Preparation for GLM
  Learn about preparing data for the Generalized Linear Model (GLM) algorithm.

Scalable Feature Selection

Oracle Machine Learning supports a highly scalable and automated version of feature selection and generation for the Generalized Linear Model algorithm.

This scalable and automated capability can enhance the performance of the algorithm and improve accuracy and interpretability. Feature selection and generation are available for both linear regression and binary logistic regression.

Feature Selection

Feature selection in GLM simplifies models, enhancing interpretability and accuracy by removing irrelevant predictors.

Feature selection is the process of choosing the terms to be included in the model. The fewer terms in the model, the easier it is for human beings to interpret its meaning. In addition, some columns may not be relevant to the value that the model is trying to predict. Removing such columns can enhance model accuracy.

Configuring Feature Selection

GLM configured for feature selection automatically determines the default behavior of the model.

Feature selection is a build setting for Generalized Linear Model models. It is not enabled by default. When configured for feature selection, the algorithm automatically determines appropriate default behavior, but the following configuration options are available:

- The feature selection criteria can be AIC, SBIC, RIC, or α-investing. When the feature selection criteria is α-investing, feature acceptance can be either strict or relaxed.
- The maximum number of features can be specified.
- Features can be pruned in the final model. Pruning is based on t-statistics for linear regression or wald statistics for logistic regression.

Feature Selection and Ridge Regression

Choose between feature selection and ridge regression to configure GLM models.

Feature selection and ridge regression are mutually exclusive. When feature selection is enabled, the algorithm can not use ridge.

> **Note:**
>
> If you configure the model to use both feature selection and ridge regression, then you get an error.

Feature Generation

Feature generation in GLM adds transformed terms, fitting complex relationships between target and predictors.

Feature generation is the process of adding transformations of terms into the model. Feature generation enhances the power of models to fit more complex relationships between target and predictors.

Configuring Feature Generation

Learn about configuring feature generation.

Feature generation is only possible when feature selection is enabled. Feature generation is a build setting. By default, feature generation is not enabled.

The feature generation method can be either quadratic or cubic. By default, the algorithm chooses the appropriate method. You can also explicitly specify the feature generation method.

The following options for feature selection also affect feature generation:

- Maximum number of features
- Model pruning

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Tuning and Diagnostics for GLM

Tuning and diagnostics in GLM help optimize model performance and quality through detailed evaluations.

The process of developing a Generalized Linear Model machine learning model typically involves a number of model builds. Each build generates many statistics that you can evaluate to determine the quality of your model. Depending on these diagnostics, you may want to try changing the model settings or making other modifications.

Build Settings

Specify the build settings for Generalized Linear Model (GLM).

You can use specify build settings.

Additional build settings are available to:

- Control the use of ridge regression.
- Specify the handling of missing values in the training data.
- Specify the target value to be used as a reference in a logistic regression model.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Generalized Linear Models for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Related Topics

- Ridge Regression
  Understand the use of ridge regression for singularity (exact multicollinearity) in data.
- Data Preparation for GLM
  Learn about preparing data for the Generalized Linear Model (GLM) algorithm.

- Logistic Regression
  GLM implements binary logistic regression, transforming target values into a probability scale for classification.

Diagnostics

A Generalized Linear Model model generates many metrics to help you evaluate the quality of the model.

Coefficient Statistics

Learn about coeffficient statistics for linear and logistic regression.

The same set of statistics is returned for both linear and logistic regression, but statistics that do not apply to the machine learning technique are returned as NULL.

Coefficient statistics are returned by the model detail views for a Generalized Linear Model (GLM) model.

Related Topics

- Coefficient Statistics for Linear Regression
  Lists coefficient statistics for linear regression.
- Coefficient Statistics for Logistic Regression
  GLM provides detailed coefficient statistics for logistic regression, aiding in model evaluation.
- *Oracle Machine Learning for SQL User’s Guide*

Global Model Statistics

Learn about high-level statistics describing the model.

Separate high-level statistics describing the model as a whole, are returned for linear and logistic regression. When ridge regression is enabled, fewer global details are returned.

Global statistics are returned by the model detail views for a Generalized Linear Model model.

Related Topics

- Global Model Statistics for Linear Regression
  Generalized Linear Model regression models generate the following statistics.
- Global Model Statistics for Logistic Regression
  GLM generates global statistics for logistic regression, supporting model assessment.
- Ridge Regression
  Understand the use of ridge regression for singularity (exact multicollinearity) in data.
- *Oracle Machine Learning for SQL User’s Guide*

Row Diagnostics

Generate row-statistics by configuring the Generalized Linear Model (GLM) algorithm.

GLM generates per-row statistics if you specify the name of a diagnostics table in the build setting `GLMS_DIAGNOSTICS_TABLE_NAME`.

GLM requires a case ID to generate row diagnostics. If you provide the name of a diagnostic table but the data does not include a case ID column, an exception is raised.

Related Topics

- Row Diagnostics for Linear Regression
  The diagnostics table for GLM regression models provides detailed row-level insights.
- Row Diagnostics for Logistic Regression
  GLM provides detailed row diagnostics for logistic regression, offering insights into individual predictions.

GLM Solvers

Generalized Linear Model (GLM) algorithm applies different solvers. These solvers employ different approaches for optimization.

The GLM algorithm supports four different solvers: Cholesky, QR, Stochastic Gradient Descent (SGD),and Alternating Direction Method of Multipliers (ADMM) (on top of L-BFGS). The Cholesky and QR solvers employ classical decomposition approaches. The Cholesky solver is faster compared to the QR solver but less stable numerically. The QR solver handles better rank deficient problems without the help of regularization.

The SGD and ADMM (on top of L-BFGS) solvers are best suited for large scale data. The SGD solver employs the stochastic gradient descent optimization algorithm while ADMM (on top of L-BFGS) uses the Broyden-Fletcher-Goldfarb-Shanno optimization algorithm within an Alternating Direction Method of Multipliers framework. The SGD solver is fast but is sensitive to parameters and requires suitable scaled data to achieve good convergence. The L-BFGS algorithm solves unconstrained optimization problems and is more stable and robust than SGD. Also, L-BFGS uses ADMM in conjunction, which, results in an efficient distributed optimization approach with low communication cost.

Related Topics

- DBMS_DATA_MINING - Algorithm Settings: Neural Network
- DBMS_DATA_MINING — Algorithm Settings: Generalized Linear Models
- DBMS_DATA_MINING — Algorithm Settings: ADMM
- DBMS_DATA_MINING — Algorithm Settings: LBFGS

Data Preparation for GLM

Learn about preparing data for the Generalized Linear Model (GLM) algorithm.

Automatic Data Preparation (ADP) implements suitable data transformations for both linear and logistic regression.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Generalized Linear Models for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.
> Oracle recommends that you use ADP with GLM.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

Data Preparation for Linear Regression

ADP ensures optimal data transformations for linear regression, enhancing model accuracy.

When ADP is enabled, the algorithm chooses a transformation based on input data properties and other settings. The transformation can include one or more of the following for numerical data: subtracting the mean, scaling by the standard deviation, or performing a correlation transformation (Neter, et. al, 1990). If the correlation transformation is applied to numeric data, it is also applied to categorical attributes.

Prior to standardization, categorical attributes are exploded into N-1 columns where N is the attribute cardinality. The most frequent value (mode) is omitted during the explosion transformation. In the case of highest frequency ties, the attribute values are sorted alpha-numerically in ascending order, and the first value on the list is omitted during the explosion. This explosion transformation occurs whether or not ADP is enabled.

In the case of high cardinality categorical attributes, the described transformations (explosion followed by standardization) can increase the build data size because the resulting data representation is dense. To reduce memory, disk space, and processing requirements, use an alternative approach. Under these circumstances, the VIF statistic must be used with caution.

Related Topics

- Ridge and Data Preparation
  Learn about preparing data for ridge regression.

> **See Also:**
>
> - Neter, J., Wasserman, W., and Kutner, M.H., "Applied Statistical Models", Richard D. Irwin, Inc., Burr Ridge, IL, 1990.

Data Preparation for Logistic Regression

ADP optimizes data for logistic regression, standardizing numerical attributes and exploding categorical attributes.

Categorical attributes are exploded into *N*-1 columns where *N* is the attribute cardinality. The most frequent value (mode) is omitted during the explosion transformation. In the case of highest frequency ties, the attribute values are sorted alpha-numerically in ascending order and the first value on the list is omitted during the explosion. This explosion transformation occurs whether or not Automatic Data Preparation (ADP) is enabled.

When ADP is enabled, numerical attributes are scaled by the standard deviation. This measure of variability is computed as the standard deviation per attribute with respect to the origin (not the mean) (Marquardt, 1980).

See Also:

Marquardt, D.W., "A Critique of Some Ridge Regression Methods: Comment", Journal of the American Statistical Association, Vol. 75, No. 369 , 1980, pp. 87-91.

Missing Values

GLM automatically replaces missing values.

When building or applying a model, Oracle Machine Learning automatically replaces missing values of numerical attributes with the mean and missing values of categorical attributes with the mode.

You can configure the Generalized Linear Model algorithm to override the default treatment of missing values. With the `ODMS_MISSING_VALUE_TREATMENT` setting, you can cause the algorithm to delete rows in the training data that have missing values instead of replacing them with the mean or the mode. However, when the model is applied, Oracle Machine Learning for SQL performs the usual mean/mode missing value replacement. As a result, it is possible that the statistics generated from scoring does not match the statistics generated from building the model.

If you want to delete rows with missing values in the scoring the model, you must perform the transformation explicitly. To make build and apply statistics match, you must remove the rows with NULLs from the scoring data before performing the apply operation. You can do this by creating a view.

```
CREATE VIEW viewname AS SELECT * from tablename
     WHERE column_name1 is NOT NULL
     AND   column_name2 is NOT NULL
     AND   column_name3 is NOT NULL .....
```

Note:

In Oracle Machine Learning for SQL, missing values in nested data indicate sparsity, not values missing at random.

The value `ODMS_MISSING_VALUE_DELETE_ROW` is only valid for tables without nested columns. If this value is used with nested data, an exception is raised.

Linear Regression

GLM supports linear regression, assuming no target transformation and constant variance over target values.

Oracle Machine Learning supports linear regression as the Generalized Linear Model regression algorithm. The algorithm assumes no target transformation and constant variance over the range of target values. The algorithm uses the identity link function.

Coefficient Statistics for Linear Regression

Lists coefficient statistics for linear regression.

Generalized Linear Model regression models generate the following coefficient statistics:

- Linear coefficient estimate
- Standard error of the coefficient estimate
- t-value of the coefficient estimate
- Probability of the t-value
- Variance Inflation Factor (VIF)
- Standardized estimate of the coefficient
- Lower and upper confidence bounds of the coefficient

Global Model Statistics for Linear Regression

Generalized Linear Model regression models generate the following statistics.

Generalized Linear Model regression models generate the following statistics that describe the model as a whole:

- Model degrees of freedom
- Model sum of squares
- Model mean square
- Model *F* statistic
- Model *F* value probability
- Error degrees of freedom
- Error sum of squares
- Error mean square
- Corrected total degrees of freedom
- Corrected total sum of squares
- Root mean square error
- Dependent mean
- Coefficient of variation
- R-Square
- Adjusted R-Square
- Akaike's information criterion
- Schwarz's Baysian information criterion
- Estimated mean square error of the prediction
- Hocking Sp statistic
- JP statistic (the final prediction error)
- Number of parameters (the number of coefficients, including the intercept)
- Number of rows
- Whether or not the model converged
- Whether or not a covariance matrix was computed

Row Diagnostics for Linear Regression

The diagnostics table for GLM regression models provides detailed row-level insights.

For linear regression, the diagnostics table has the columns described in the following table. All the columns are `NUMBER`, except the `CASE_ID` column, which preserves the type from the training data.

Table 7-13 Diagnostics Table for GLM Regression Models

Column	Description
`CASE_ID`	Value of the case ID column
`TARGET_VALUE`	Value of the target column
`PREDICTED_VALUE`	Value predicted by the model for the target
`HAT`	Value of the diagonal element of the hat matrix
`RESIDUAL`	Measure of error
`STD_ERR_RESIDUAL`	Standard error of the residual
`STUDENTIZED_RESIDUAL`	Studentized residual
`PRED_RES`	Predicted residual
`COOKS_D`	Cook's D influence statistic

Logistic Regression

GLM implements binary logistic regression, transforming target values into a probability scale for classification.

Oracle Machine Learning supports binary logistic regression as a Generalized Linear Model classification algorithm. Link and variance functions are the mechanism that allows GLM to handle targets of a regression that departs in known ways from normality. In logistic regression, a link function is used to relate the explanatory variables (covariates) and the expectation of the response variable. Binomial regression predicts the probability of a success by applying the inverse of a specified link function to a linear combination of covariates. The specified inverse link function can be any monotonically increasing function that maps values from the range (-∞, ∞) to [0,1]. The inverse link function is created from cumulative distribution functions (CDFs) of well-known random distributions. The variance has a known functional relationship with the probability, and a binary target probability varies between zero and one. For logistic regression, the variance function is fixed to its known functional relationship with probability. However, there are other options for the link function. The link function not only transforms the target range into a linear-methods-friendly format, but it also represents a target concept. The analyst can use the target concept to interpret a forecast on two scales: the link scale and the transformed scale. The transformed scale in logistic regression is probability.

Reference Class

Specify the reference class for binary logistic regression in GLM to improve prediction accuracy.

You can use the build setting `GLMS_REFERENCE_CLASS_NAME` to specify the target value to be used as a reference in a binary logistic regression model. Probabilities are produced for the other (non-reference) class. By default, the algorithm chooses the value with the highest prevalence. If there are ties, the attributes are sorted alpha-numerically in an ascending order.

Class Weights

Use class weights to influence target class weighting during model building in GLM.

You can use the build setting `CLAS_WEIGHTS_TABLE_NAME` to specify the name of a class weights table. Class weights influence the weighting of target classes during the model build.

Coefficient Statistics for Logistic Regression

GLM provides detailed coefficient statistics for logistic regression, aiding in model evaluation.

Generalized Linear Model classification models generate the following coefficient statistics:

- Name of the predictor
- Coefficient estimate
- Standard error of the coefficient estimate
- Wald chi-square value of the coefficient estimate
- Probability of the Wald chi-square value
- Standardized estimate of the coefficient
- Lower and upper confidence bounds of the coefficient
- Exponentiated coefficient
- Exponentiated coefficient for the upper and lower confidence bounds of the coefficient

Global Model Statistics for Logistic Regression

GLM generates global statistics for logistic regression, supporting model assessment.

Generalized Linear Model classification models generate the following statistics that describe the model as a whole:

- Akaike's criterion for the fit of the intercept only model
- Akaike's criterion for the fit of the intercept and the covariates (predictors) model
- Schwarz's criterion for the fit of the intercept only model
- Schwarz's criterion for the fit of the intercept and the covariates (predictors) model
- -2 log likelihood of the intercept only model
- -2 log likelihood of the model
- Likelihood ratio degrees of freedom
- Likelihood ratio chi-square probability value
- Pseudo R-square Cox an Snell
- Pseudo R-square Nagelkerke
- Dependent mean
- Percent of correct predictions
- Percent of incorrect predictions
- Percent of ties (probability for two cases is the same)
- Number of parameters (the number of coefficients, including the intercept)

- Number of rows
- Whether or not the model converged
- Whether or not a covariance matrix was computed.

Row Diagnostics for Logistic Regression

GLM provides detailed row diagnostics for logistic regression, offering insights into individual predictions.

For logistic regression, the diagnostics table has the columns described in the following table. All the columns are `NUMBER`, except the `CASE_ID` and `TARGET_VALUE` columns, which preserve the type from the training data.

Table 7-14 Row Diagnostics Table for Logistic Regression

Column	Description
`CASE_ID`	Value of the case ID column
`TARGET_VALUE`	Value of the target value
`TARGET_VALUE_PROB`	Probability associated with the target value
`HAT`	Value of the diagonal element of the hat matrix
`WORKING_RESIDUAL`	Residual with respect to the adjusted dependent variable
`PEARSON_RESIDUAL`	The raw residual scaled by the estimated standard deviation of the target
`DEVIANCE_RESIDUAL`	Contribution to the overall goodness of fit of the model
`C`	Confidence interval displacement diagnostic
`CBAR`	Confidence interval displacement diagnostic
`DIFDEV`	Change in the deviance due to deleting an individual observation
`DIFCHISQ`	Change in the Pearson chi-square

About *k*-Means

The *k*-Means algorithm is a distance-based clustering algorithm that partitions the data into a specified number of clusters.

Distance-based algorithms rely on a distance function to measure the similarity between cases. Cases are assigned to the nearest cluster according to the distance function used.

Oracle Machine Learning for SQL Enhanced *k*-Means

Oracle Machine Learning offers an enhanced *k*-Means algorithm with efficient initialization, scalable parallel model build, and detailed cluster properties.

Oracle Machine Learning for SQL implements an enhanced version of the *k*-Means algorithm with the following features:

- **Distance function**: The algorithm supports Euclidean and Cosine distance functions. The default is Euclidean.
- **Scalable Parallel Model build**: The algorithm uses a very efficient method of initialization based on *Bahmani, Bahman, et al. "Scalable k-means++." Proceedings of the VLDB Endowment 5.7 (2012): 622-633*.

- **Cluster properties**: For each cluster, the algorithm returns the centroid, a histogram for each attribute, and a rule describing the hyperbox that encloses the majority of the data assigned to the cluster. The centroid reports the mode for categorical attributes and the mean and variance for numerical attributes.

This approach to *k*-Means avoids the need for building multiple *k*-Means models and provides clustering results that are consistently superior to the traditional *k*-Means.

Centroid

A centroid represents the most typical case in a cluster, with mean values for numerical attributes and mode values for categorical attributes.

The **centroid** represents the most typical case in a cluster. For example, in a data set of customer ages and incomes, the centroid of each cluster would be a customer of average age and average income in that cluster. The centroid is a prototype. It does not necessarily describe any given case assigned to the cluster.

The attribute values for the centroid are the mean of the numerical attributes and the mode of the categorical attributes.

k-Means Algorithm Configuration

The Oracle Machine Learning enhanced *k*-Means algorithm supports several build-time settings.

All the settings have default values. There is no reason to override the defaults unless you want to influence the behavior of the algorithm in some specific way.

You can configure *k*-Means by specifying the following considerations:

- Number of clusters
- Distance Function. The default distance function is Euclidean.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: k-Means for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Data Preparation for *k*-Means

Learn about preparing data for *k*-Means algorithm.

Normalization is typically required by the *k*-Means algorithm. Automatic Data Preparation performs normalization for *k*-Means. If you do not use ADP, you must normalize numeric attributes before creating or applying the model.

When there are missing values in columns with simple data types (not nested), *k*-Means interprets them as missing at random. The algorithm replaces missing categorical values with the mode and missing numerical values with the mean.

When there are missing values in nested columns, *k*-Means interprets them as sparse. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors.

Data can be constrained in a window size of 6 standard-deviations around the mean value by using the `KMNS_WINSORIZE` parameter. The `KMNS_WINSORIZE` parameter can be used whether ADP is set to `ON` or `OFF`. Values outside the range are mapped to the range's ends. This parameter is applicable only when the Euclidean distance is used.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*
- Prepare the Data

About MDL

Minimum Description Length (MDL) is an information theoretic model selection principle that assumes the simplest representation of data is the most probable explanation.

Information theoretic model selection principle is an important concept in information theory (the study of the quantification of information) and in learning theory (the study of the capacity for generalization based on empirical data).

MDL assumes that the simplest, most compact representation of the data is the best and most probable explanation of the data. The MDL principle is used to build Oracle Machine Learning attribute importance models.

The build process for attribute importance supports parallel processing.

Related Topics

- *Oracle Database VLDB and Partitioning Guide*

Compression and Entropy

Data compression is the process of encoding information using fewer **bits** than what the original representation uses. The MDL Principle is based on the notion that the shortest description of the data is the most probable. In typical instantiations of this principle, a model is used to compress the data by reducing the uncertainty (entropy) as discussed below. The description of the data includes a description of the model and the data as described by the model.

Entropy is a measure of uncertainty. It quantifies the uncertainty in a random variable as the information required to specify its value. **Information** in this sense is defined as the number of yes/no questions known as **bits** (encoded as 0 or 1) that must be answered for a complete specification. Thus, the information depends upon the number of values that variable can assume.

For example, if the variable represents the sex of an individual, then the number of possible values is two: female and male. If the variable represents the salary of individuals expressed in whole dollar amounts, then the values can be in the range $0-$10B, or billions of unique values. Clearly it takes more information to specify an exact salary than to specify an individual's sex.

Values of a Random Variable: Statistical Distribution

Information (the number of bits) depends on the statistical distribution of the values of the variable as well as the number of values of the variable. If we are judicious in the choice of

Yes/No questions, then the amount of information for salary specification cannot be as much as it first appears. Most people do not have billion dollar salaries. If most people have salaries in the range $32000-$64000, then most of the time, it requires only 15 questions to discover their salary, rather than the 30 required, if every salary from $0-$1000000000 were equally likely. In the former example, if the persons were known to be pregnant, then their sex is known to be female. There is no uncertainty, no Yes/No questions need be asked. The entropy is 0.

Values of a Random Variable: Significant Predictors

Suppose that for some random variable there is a predictor that when its values are known reduces the uncertainty of the random variable. For example, knowing whether a person is pregnant or not, reduces the uncertainty of the random variable sex-of-individual. This predictor seems like a valuable feature to include in a model. How about name? Imagine that if you knew the name of the person, you would also know the person's sex. If so, the name predictor would seemingly reduce the uncertainty to zero. However, if names are unique, then what was gained? Is the person named Sally? Is the person named George?... We would have as many Yes/No predictors in the name model as there are people. Therefore, specifying the name model would require as many bits as specifying the sex of each person.

Total Entropy

For a random variable, X, the **total entropy** is defined as minus the Probability(X) multiplied by the log to the base 2 of the Probability(X). This can be shown to be the variable's most efficient encoding.

Model Size

A Minimum Description Length (MDL) model takes into consideration the size of the model as well as the reduction in uncertainty due to using the model. Both model size and entropy are measured in bits. For our purposes, both numeric and categorical predictors are binned. Thus the size of each single predictor model is the number of predictor bins. The uncertainty is reduced to the within-bin target distribution.

Model Selection

Minimum Description Length (MDL) considers each attribute as a simple predictive model of the target class. **Model selection** refers to the process of comparing and ranking the single-predictor models.

MDL uses a communication model for solving the model selection problem. In the communication model there is a sender, a receiver, and data to be transmitted.

These single predictor models are compared and ranked with respect to the MDL metric, which is the relative compression in bits. MDL penalizes model complexity to avoid over-fit. It is a principled approach that takes into account the complexity of the predictors (as models) to make the comparisons fair.

The MDL Metric

Attribute importance uses a two-part code as the metric for transmitting each unit of data. The first part (preamble) transmits the model. The parameters of the model are the target probabilities associated with each value of the prediction.

For a target with *j* values and a predictor with *k* values, n_i (i= 1,..., k) rows per value, there are C_i, the combination of *j*-1 things taken n_i-1 at a time possible conditional probabilities. The size of the preamble in bits can be shown to be Sum($\log_2(C_i)$), where the sum is taken over *k*.

Computations like this represent the penalties associated with each single prediction model. The second part of the code transmits the target values using the model.

It is well known that the most compact encoding of a sequence is the encoding that best matches the probability of the symbols (target class values). Thus, the model that assigns the highest probability to the sequence has the smallest target class value transmission cost. In bits, this is the Sum($\log_2(p_i)$), where the p_i are the predicted probabilities for row $_i$ associated with the model.

The predictor rank is the position in the list of associated description lengths, smallest first.

Data Preparation for MDL

Learn about preparing data for Minimum Description Length (MDL).

Automatic Data Preparation performs supervised binning for MDL. Supervised binning uses decision trees to create the optimal bin boundaries. Both categorical and numerical attributes are binned.

MDL handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing values in columns with simple data types are interpreted as missing at random.

If you choose to manage your own data preparation, keep in mind that MDL usually benefits from binning. However, the discriminating power of an attribute importance model can be significantly reduced when there are outliers in the data and external equal-width binning is used. This technique can cause most of the data to concentrate in a few bins (a single bin in extreme cases). In this case, quantile binning is a better solution.

> **See Also:**
>
> DBMS_DATA_MINING — Automatic Data Preparation for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Related Topics

- Prepare the Data

About Multivariate State Estimation Technique - Sequential Probability Ratio Test

Multivariate state Estimation Technique - Sequential Probability Ratio Test (MSET-SPRT) is an algorithm for anomaly detection and statistical testing.

MSET is a nonlinear, nonparametric anomaly detection machine learning technique that calibrates the expected behavior of a system based on historical data from the normal operational sequence of monitored signals. It incorporates the learned behavior of a system into a persistent model that represents the normal estimated behavior. You can deploy the model to evaluate a subsequent stream of live signal vectors using Oracle Machine Learning for SQL scoring functions. To form a hypothesis as to the overall health of the system, these functions calculate the difference between the estimated and the actual signal values (residuals) and use SPRT calculations to determine whether any of the signals have become degraded.

To build a good model, MSET requires sufficient historical data that adequately captures all normal modes of behavior of the system. Incomplete data results in false alerts when the system enters a mode of operation that was poorly represented in the historical data. MSET assumes that the characteristics of the data being monitored do not change over time. Once deployed, MSET is a stationary model and does not evolve as it monitors a data stream.

Both MSET and SPRT operate on continuous time-ordered sensor data. If the raw data stream needs to be pre-processed or sampled, you must do that before you pass the data to the MSET-SPRT model.

The `ALGO_MSET_SPRT` algorithm is designated as a classification machine learning technique. It generates a model in which each data row is labeled as either normal or anomalous. For anomalous predictions, the prediction details provide a list of the sensors that show the anomaly and a weight.

When creating an MSET-SPRT model with the `DBMS_DATA_MINING.CREATE_MODEL` function, use the `case_id` argument to provide a unique row identifier for the time-ordered data that the algorithm requires. The build is then able to sort the training data and create windows for sampling and variance estimation. If you do not provide a `case_id`, then an exception occurs.

MSET-SPRT supports only numeric data. An exception occurs if other column types are in the build data.

When the number of sensors is very high, MSET-SPRT leverages random projections to improve the scalability and robustness of the algorithm. Random projections is a technique that reduces dimensionality while preserving pairwise distances. By randomly projecting the sensor data, the problem is solved in a distance-preserving, lower-dimension space. The MSET hypothesis testing approach is applied on the projected data where each random projection can be viewed as a Monte Carlo simulation of system health. The overall probability of an anomaly follows a binomial distribution with the number of projections as the number of trials and the number of alerting projections as the number of successes.

Note:

An MSET-SPRT model with random projections does not produce prediction details. When random projections are employed, the nature of the prediction output changes. The prediction captures the global health of the system and it is not possible to attribute the cause to individual attributes. Therefore, `PREDICTION_DETAILS` returns an empty list.

See Also:

DBMS_DATA_MINING - Algorithm Settings: Multivariate State Estimation Technique - Sequential Probability Ratio Test for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Score an MSET-SPRT Model

Scoring data with MSET-SPRT models is similar to scoring with classification algorithms, except that the SPRT methodology relies on ordered data because it tracks gradual shifts over multiple MSET predictions.

This is different than the typical usage of Oracle Database SQL prediction functions, which do not keep state information between rows.

The following functions are supported: `PREDICTION`, `PREDICTION_COST`, `PREDICTION_DETAILS`, `PREDICTION_PROBABILITY`, and `PREDICTION_SET`. These functions have syntax new in Oracle Database 21c for scoring MSET-SPRT models. That syntax has an `ORDER BY` clause to order and window the historical data.

The prediction functions return the following information:

- `PREDICTION` indicates whether the record is flagged as anomalous. It uses the same automatically generated labels as one-class SVM models: 1 for normal and 0 for anomalous.
- `PREDICTION_COST` performs an auto-cost analysis or a user-specified cost. A user-specified cost typically assigns a higher cost to false positives than to false negatives.
- `PREDICTION_DETAILS` specify the signals that support the prediction along with a weight.
- `PREDICTION_PROBABILITY` conveys a measure of certainty based on the consolidation logic.
- `PREDICTION_SET` returns the set of predictions (0, 1) and the corresponding prediction probabilities for each observation.

> **Note:**
>
> If the values in one or more of the columns specified in the `ORDER BY` clause are not unique, or do not represent a true chronology of data sample values, the SPRT predictions are not guaranteed to be meaningful or consistent between query executions.

Unlike other classification models, an MSET-SPRT model has no obvious probability measure associated with the anomalous label for the record as a whole. However, the consolidation logic can produce a measure of uncertainty in place of probability. For example, if an alert is raised for 2 anomalies over a window of 5 observations, a certainty of 0.5 is reported when 2 anomalies are seen within the 5 observation window. The certainty increases if more than 3 anomalies are seen and decreases if no anomalies are seen.

The `PREDICTION_DETAILS` function accommodates output of varying forms and can convey the required information regarding the individual signals that triggered an alarm. When random projections are engaged, only the overall `PREDICTION` and `PREDICTION_PROBABILITY` are computed and `PREDICTION_DETAILS` are not reported.

You must score the historical data in order to tune the SPRT parameters, such as false alerts and miss rates or consolidation logic, before you deploy the MSET model. The SPRT parameters are embedded in the model object to facilitate deployment. While scoring in the database is needed for parameter tuning and forensic analysis on historical data, monitoring a stream of sensor data is more easily done outside of the database in an IoT service or on the edge device itself.

You can build and score an MSET-SPRT model as a partitioned model if the same columns that you use to build the model are present in the input scoring data set. If those columns are not present, the query results in an error.

Related Topics

- MSET_SPRT example on GitHub

About Naive Bayes

Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

Bayes' theorem finds the probability of an event occurring given the probability of another event that has already occurred. If `B` represents the dependent event and `A` represents the prior event, Bayes' theorem can be stated as follows.

> **Note:**
>
> Prob(B given A) = Prob(A and B)/Prob(A)

To calculate the probability of `B` given `A`, the algorithm counts the number of cases where `A` and `B` occur together and divides it by the number of cases where `A` occurs alone.

Example 7-4 Use Bayes' Theorem to Predict an Increase in Spending

Suppose you want to determine the likelihood that a customer under 21 increases spending. In this case, the prior condition (`A`) is "under 21," and the dependent condition (`B`) is "increase spending."

If there are 100 customers in the training data and 25 of them are customers under 21 who have increased spending, then:

Prob(A and B) = 25%

If 75 of the 100 customers are under 21, then:

Prob(A) = 75%

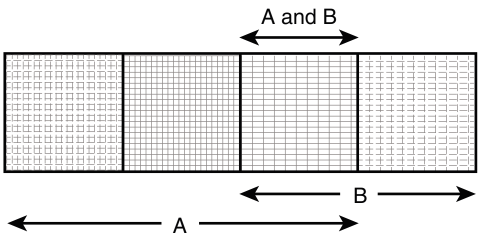
Bayes' theorem predicts that 33% of customers under 21 are likely to increase spending (25/75).

The cases where both conditions occur together are referred to as **pairwise**. In Example 7-4, 25% of all cases are pairwise.

The cases where only the prior event occurs are referred to as **singleton**. In Example 7-4, 75% of all cases are singleton.

A visual representation of the conditional relationships used in Bayes' theorem is shown in the following figure.

Figure 7-11 Conditional Probabilities in Bayes' Theorem



P(A) = 3/4
P(B) = 2/4
P(A and B) = P(AB) = 1/4
P(A|B) = P(AB) / P(B) = (1/4) / (2/4) = 1/2
P(B|A) = P(AB) / P(A) = (1/4) / (3/4) = 1/3

For purposes of illustration, Example 7-4 and Figure 7-11 show a dependent event based on a single independent event. In reality, the Naive Bayes algorithm must usually take many independent events into account. In Example 7-4, factors such as income, education, gender, and store location might be considered in addition to age.

Naive Bayes makes the assumption that each predictor is conditionally independent of the others. For a given target value, the distribution of each predictor is independent of the other predictors. In practice, this assumption of independence, even when violated, does not degrade the model's predictive accuracy significantly, and makes the difference between a fast, computationally feasible algorithm and an intractable one.

Sometimes the distribution of a given predictor is clearly not representative of the larger population. For example, there might be only a few customers under 21 in the training data, but in fact there are many customers in this age group in the wider customer base. To compensate for this, you can specify **prior probabilities** when training the model.

Related Topics

- Priors and Class Weights
  Offset differences in data distribution with prior probabilities and class weights to produce useful classification results.

Advantages of Naive Bayes

Learn about the advantages of Naive Bayes.

The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows.

The build process for Naive Bayes supports parallel execution. (Scoring supports parallel execution irrespective of the algorithm.)

Naive Bayes can be used for both binary and multiclass classification problems.

Related Topics

- *Oracle Database VLDB and Partitioning Guide*

Tuning a Naive Bayes Model

Naive Bayes calculates probabilities by dividing pairwise occurrence percentages by singleton occurrence percentages, improving model performance with threshold adjustments.

If these percentages are very small for a given predictor, they probably do not contribute to the effectiveness of the model. Occurrences below a certain threshold can usually be ignored.

The following build settings are available for adjusting the probability thresholds. You can specify:

- The minimum percentage of pairwise occurrences required for including a predictor in the model.
- The minimum percentage of singleton occurrences required for including a predictor in the model .

The default thresholds work well for most models, so you need not adjust these settings.

See Also:

DBMS_DATA_MINING — Algorithm Settings: Naive Bayes for a listing and explanation of the available model settings.

Note:

The term hyperparameter is also interchangeably used for model setting.

Data Preparation for Naive Bayes

Automatic Data Preparation performs supervised binning, handling missing values effectively for accurate probability calculations.

Automatic Data Preparation (ADP) performs supervised binning for Naive Bayes. Supervised binning uses decision trees to create the optimal bin boundaries. Both categorical and numeric attributes are binned.

Naive Bayes handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing values in columns with simple data types are interpreted as missing at random.

If you choose to manage your own data preparation, keep in mind that Naive Bayes usually requires binning. Naive Bayes relies on counting techniques to calculate probabilities. Columns must be binned to reduce the cardinality as appropriate. Numerical data can be binned into ranges of values (for example, low, medium, and high), and categorical data can be binned into meta-classes (for example, regions instead of cities). Equi-width binning is not recommended, since outliers cause most of the data to concentrate in a few bins, sometimes a single bin. As a result, the discriminating power of the algorithms is significantly reduced

Related Topics

- Prepare the Data

About Neural Network

Neural Networks in Oracle Machine Learning are designed for complex tasks like classification and regression, inspired by biological neural networks.

In machine learning, an artificial neural network is an algorithm inspired from biological neural network and is used to estimate or approximate functions that depend on a large number of generally unknown inputs. An artificial neural network is composed of a large number of interconnected neurons which exchange messages between each other to solve specific problems. They learn by examples and tune the weights of the connections among the neurons during the learning process. The Neural Network algorithm is capable of solving a wide variety of tasks such as computer vision, speech recognition, and various complex business problems.

Related Topics

- About Regression
  Regression is a machine learning technique that predicts numeric values along a continuum.
- About Classification
  Classification is a machine learning technique that assigns items to target categories or classes to predict outcomes.

Neurons and Activation Functions

Neurons process inputs through weighted sums and activation functions like Sigmoid and Rectified Linear Units (ReLU).

A neuron takes one or more inputs having different weights and has an output which depends on the inputs. The output is achieved by adding up inputs of each neuron with weights and feeding the sum into the activation function.

A Sigmoid function is usually the most common choice for activation function but other non-linear functions, piecewise linear functions or step functions are also used. The Rectified Linear Units function `NNET_ACTIVATIONS_RELU` is a commonly used activation function that addresses the vanishing gradient problem for larger neural networks.

The following are some examples of activation functions:

- Logistic Sigmoid function
- Linear function
- Tanh function
- Arctan function
- Bipolar sigmoid function
- Rectified Linear Units

Loss or Cost function

A loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event.

An optimization problem seeks to minimize a loss function. The form of loss function is chosen based on the nature of the problem and mathematical needs.

The following are the different loss functions for different scenarios:

- Binary classification: binary cross entropy loss function.
- Multi-class classification: multi cross entropy loss function.
- Regression: squared error function.

Forward-Backward Propagation

Understand forward-backward propagation.

Forward propagation computes the loss function value by weighted summing the previous layer neuron values and applying activation functions. Backward propagation calculates the gradient of a loss function with respect to all the weights in the network. The weights are initialized with a set of random numbers uniformly distributed within a region specified by user (by setting weights boundaries), or region defined by the number of nodes in the adjacent layers (data driven). The gradients are fed to an optimization method which in turn uses them to update the weights, in an attempt to minimize the loss function.

Optimization Solvers

An optimization solver is a function that searches for the optimal solution of the loss function to find the extreme value (maximum or minimum) of the loss (cost) function. Neural Networks use L-BFGS and Adam solvers for efficient and effective optimization.

Oracle Machine Learning implements Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) together with line search and the Adam solver.

Limited-memory Broyden–Fletcher–Goldfarb–Shanno Solver

L-BFGS is a Quasi-Newton method. This method uses rank-one updates specified by gradient evaluations to approximate a Hessian matrix. This method only needs a limited amount of memory. L-BFGS is used to find the descent direction and line search is used to find the appropriate step size. The number of historical copies kept in the L-BFGS solver is defined by the `LBFGS_HISTORY_DEPTH` solver setting. When the number of iterations is smaller than the history depth, the Hessian computed by L-BFGS is accurate. When the number of iterations is larger than the history depth, the Hessian computed by L-BFGS is an approximation. Therefore, the history depth should not be too small or too large to avoid making the computation too slow. Typically, the value is between `3` and `10`.

Adam Solver

Adam is an extension to stochastic gradient descent that uses mini-batch optimization. The L-BFGS solver may be a more stable solver whereas the Adam solver can make progress faster by seeing less data. Adam is computationally efficient, with little memory requirements, and is well-suited for problems that are large in terms of data or parameters or both.

Regularization

Regularization techniques, such as L1 norms, L2 norms, and held-aside prevent overfitting and improve model generalization.

Regularization refers to a process of introducing additional information to solve an ill-posed problem or to prevent over-fitting. Ill-posed or over-fitting can occur when a statistical model describes random errors or noise instead of the underlying relationship. Typical regularization techniques include L1-norm regularization, L2-norm regularization, and held-aside.

Held-aside is usually used for large training date sets whereas L1-norm regularization and L2-norm regularization are mostly used for small training date sets.

Convergence Check

Convergence checks ensure optimization processes reach optimal solutions, stopping when performance criteria are met.

In L-BFGS solver, the convergence criteria includes maximum number of iterations, infinity norm of gradient, and relative error tolerance. For held-aside regularization, the convergence criteria checks the loss function value of the test data set, as well as the best model learned so far. The training is terminated when the model becomes worse for a specific number of iterations (specified by `NNET_HELDASIDE_MAX_FAIL`), or the loss function is close to zero, or the relative error on test data is less than the tolerance.

LBFGS_SCALE_HESSIAN

`LBFGS_SCALE_HESSIAN` setting improves optimization by adjusting initial inverse Hessian approximations.

The setting adjusts the initial approximation of the inverse Hessian at the beginning of each iteration. If the value is `LBFGS_SCALE_HESSIAN_ENABLE`, then the initial inverse Hessian is approximated with Oren-Luenberger scaling. If it is `LBFGS_SCALE_HESSIAN_DISABLE`, identity is used as the initial approximation of the inverse Hessian at the beginning of each iteration.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

NNET_HELDASIDE_MAX_FAIL

`NNET_HELDASIDE_MAX_FAIL` setting uses validation data (held-aside) to halt training if network performance fails to improve after a set number of epochs.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Data Preparation for Neural Network

Neural Network algorithms normalize numeric data, convert categorical data into binary attributes, and handle missing values automatically.

The algorithm automatically "explodes" categorical data into a set of binary attributes, one per category value. Oracle Machine Learning algorithms automatically handle missing values and therefore, missing value treatment is not necessary.

The algorithm automatically replaces missing categorical values with the mode and missing numerical values with the mean. The algorithm requires the normalization of numeric input and it uses z-score normalization. The normalization occurs only for two-dimensional numeric columns (not nested). Normalization places the values of numeric attributes on the same scale and prevents attributes with a large original scale from biasing the solution. Neural Network scales the numeric values in nested columns by the maximum absolute value seen in the corresponding columns.

Related Topics

- Prepare the Data

Neural Network Algorithm Configuration

Configure Neural Network algorithms by specifying nodes per layer and activation functions to optimize performance.

Specify Nodes Per Layer

```
INSERT INTO SETTINGS_TABLE (setting_name, setting_value) VALUES
                   ('NNET_NODES_PER_LAYER', '2,3');
```

Specify Activation Functions Per Layer

`NNET_ACTIVATIONS` setting specifies the activation functions or hidden layers.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Neural Network for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Scoring with Neural Network

Score data with Neural Networks using standard regression and classification scoring functions.

Scoring with Neural Network is the same as any other classification or regression algorithm. The following functions are supported: `PREDICTION`, `PREDICTION_PROBABILITY`, `PREDICTION_COST`, `PREDICTION_SET`, and `PREDICTION_DETAILS`.

Related Topics

- *Oracle Database SQL Language Reference*

About NMF

Non-Negative Matrix Factorization is useful when there are many attributes and the attributes are ambiguous or have weak predictability. By combining attributes, NMF can produce meaningful patterns, topics, or themes. NMF is a feature extraction algorithm.

Each feature created by NMF is a linear combination of the original attribute set. Each feature has a set of coefficients, which are a measure of the weight of each attribute on the feature. There is a separate coefficient for each numerical attribute and for each distinct value of each categorical attribute. The coefficients are all non-negative.

Matrix Factorization

Non-Negative Matrix Factorization uses techniques from multivariate analysis and linear algebra. It decomposes the data as a matrix *M* into the product of two lower ranking matrices

W and *H*. The sub-matrix *W* contains the NMF basis; the sub-matrix *H* contains the associated coefficients (weights).

The algorithm iteratively modifies of the values of *W* and *H* so that their product approaches *M*. The technique preserves much of the structure of the original data and guarantees that both basis and weights are non-negative. The algorithm terminates when the approximation error converges or a specified number of iterations is reached.

The NMF algorithm must be initialized with a seed to indicate the starting point for the iterations. Because of the high dimensionality of the processing space and the fact that there is no global minimization algorithm, the appropriate initialization can be critical in obtaining meaningful results. Oracle Machine Learning for SQL uses a random seed that initializes the values of W and H based on a uniform distribution. This approach works well in most cases.

Scoring with NMF

Non-Negative Matrix Factorization (NMF) can be used as a pre-processing step for dimensionality reduction in classification, regression, clustering, and other machine learning tasks. Scoring an NMF model produces data projections in the new feature space. The magnitude of a projection indicates how strongly a record maps to a feature.

The SQL scoring functions for feature extraction support NMF models. When the functions are invoked with the analytical syntax, the functions build and apply a transient NMF model. The feature extraction functions are: `FEATURE_DETAILS`, `FEATURE_ID`, `FEATURE_SET`, and `FEATURE_VALUE`.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

Text Analysis with NMF

NMF analyzes text effectively by introducing context through combining attributes, enhancing explanatory power.

NMF is especially well-suited for analyzing text. In a text document, the same word can occur in different places with different meanings. For example, "hike" can be applied to the outdoors or to interest rates. By combining attributes, NMF introduces context, which is essential for explanatory power:

- "hike" + "mountain" -> "outdoor sports"
- "hike" + "interest" -> "interest rates"

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

Tuning the NMF Algorithm

Learn about configuring parameters for Non-Negative Matrix Factorization (NMF).

Oracle Machine Learning for SQL supports five configurable parameters for NMF. All of them have default values which are appropriate for most applications of the algorithm. The NMF settings are:

- Number of features. By default, the number of features is determined by the algorithm.
- Convergence tolerance. The default is .05.
- Number of iterations. The default is 50.

- Random seed. The default is -1.
- Non-negative scoring. You can specify whether negative numbers must be allowed in scoring results. By default they are allowed.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Non-Negative Matrix Factorization for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Data Preparation for NMF

You can use Automatic Data Preparation (ADP) or supply your transformation like binning or normalization to prepare the data for Non-Negative Matrix Factorization (NMF).

ADP normalizes numerical attributes for NMF.

When there are missing values in columns with simple data types (not nested), NMF interprets them as missing at random. The algorithm replaces missing categorical values with the mode and missing numerical values with the mean.

When there are missing values in nested columns, NMF interprets them as sparse. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors.

If you choose to manage your own data preparation, keep in mind that outliers can significantly impact NMF. Use a clipping transformation before binning or normalizing. NMF typically benefits from normalization. However, outliers with min-max normalization cause poor matrix factorization. To improve the matrix factorization, you need to decrease the error tolerance. This in turn leads to longer build times.

Related Topics

- Prepare the Data

About O-Cluster

O-Cluster is a fast, scalable grid-based clustering algorithm well-suited for analysing large, high-dimensional data sets. The algorithm can produce high quality clusters without relying on user-defined parameters.

The objective of O-Cluster is to identify areas of high density in the data and separate the dense areas into clusters. It uses axis-parallel uni-dimensional (orthogonal) data projections to identify the areas of density. The algorithm looks for splitting points that result in distinct clusters that do not overlap and are balanced in size.

O-Cluster operates recursively by creating a binary tree hierarchy. The number of leaf clusters is determined automatically. The algorithm can be configured to limit the maximum number of clusters.

Partitioning Strategy

O-Cluster identifies dense regions using histograms, balancing well-separated clusters and size.

Partitioning strategy refers to the process of discovering areas of density in the attribute histograms. The process differs for numerical and categorical data. When both are present in the data, the algorithm performs the searches separately and then compares the results.

In choosing a partition, the algorithm balances two objectives: finding well separated clusters, and creating clusters that are balanced in size. The following paragraphs detail how partitions for numerical and categorical attributes are identified.

Partitioning Numerical Attributes

To find the best valid cutting plane, O-Cluster searches the attribute histograms for bins of low density (valleys) between bins of high density (peaks).

O-Cluster attempts to find a pair of peaks with a valley between them where the difference between the peak and valley histogram counts is statistically significant.

A **sensitivity** level parameter specifies the lowest density that may be considered a peak. Sensitivity is an optional parameter for numeric data. It may be used to filter the splitting point candidates.

Partitioning Categorical Attributes

Categorical values do not have an intrinsic order associated with them. Therefore it is impossible to apply the notion of histogram peaks and valleys that is used to partition numerical values. Instead the counts of individual values form a histogram.

Bins with large counts are interpreted as regions with high density. The clustering objective is to separate these high-density areas and effectively decrease the entropy (randomness) of the data.

O-Cluster identifies the histogram with highest entropy along the individual projections. Entropy is measured as the number of bins above **sensitivity** level. O-Cluster places the two largest bins into separate partitions, thereby creating a splitting predicate. The remainder of the bins are assigned randomly to the two resulting partitions.

Active Sampling

The O-Cluster algorithm operates on a data buffer of a limited size. It uses an active sampling mechanism to handle data sets that do not fit into memory.

After processing an initial random sample, O-Cluster identifies cases that are of no further interest. Such cases belong to *frozen* partitions where further splitting is highly unlikely. These cases are replaced with examples from *ambiguous* regions where further information (additional cases) is needed to find good splitting planes and continue partitioning. A partition is considered ambiguous if a valid split can only be found at a lower confidence level.

Cases associated with frozen partitions are marked for deletion from the buffer. They are replaced with cases belonging to ambiguous partitions. The histograms of the ambiguous partitions are updated and splitting points are reevaluated.

Process Flow

At a high level, O-Cluster algorithm evaluates, splits the data into new partition, and searches for cutting planes inside the new partitions.

The O-Cluster algorithm evaluates possible splitting points for all projections in a partition, selects the best one, and splits the data into two new partitions. The algorithm proceeds by searching for good cutting planes inside the newly created partitions. Thus, O-Cluster creates a binary tree structure that divides the input space into rectangular regions with no overlaps or gaps.

The main processing stages are:

1. Load the buffer. Assign all cases from the initial buffer to a single active root partition.
2. Compute histograms along the orthogonal uni-dimensional projections for each active partition.
3. Find the best splitting points for active partitions.
4. Flag ambiguous and frozen partitions.
5. When a valid separator exists, split the active partition into two new active partitions and start over at step 2.
6. Reload the buffer after all recursive partitioning on the current buffer is completed. Continue loading the buffer until either the buffer is filled again, or the end of the data set is reached, or until the number of cases is equal to the data buffer size.

> **Note:**
>
> O-Cluster requires at most one pass through the data

Scoring

O-Cluster generates a Bayesian probability model for scoring new data based on discovered clusters.

The generated probability model is a mixture model where the mixture components are represented by a product of independent normal distributions for numerical attributes and multinomial distributions for categorical attributes.

Tuning the O-Cluster Algorithm

You can configure build-time settings for O-Cluster.

The O-Cluster algorithm supports two build-time settings. Both settings have default values. There is no reason to override the defaults unless you want to influence the behavior of the algorithm in some specific way.

You can configure O-Cluster by specifying the following:

Sensitivity factor — A fraction that specifies the peak density required for separating a new cluster.

See Also:

DBMS_DATA_MINING — Algorithm Settings: O-Cluster for a listing and explanation of the available model settings.

Note:

The term hyperparameter is also interchangeably used for model setting.

Related Topics

- Active Sampling
  The O-Cluster algorithm operates on a data buffer of a limited size. It uses an active sampling mechanism to handle data sets that do not fit into memory.
- Partitioning Strategy
  O-Cluster identifies dense regions using histograms, balancing well-separated clusters and size.

Data Preparation for O-Cluster

Use Automatic Data Preparation (ADP) for binning and handling missing values, ensuring optimal clustering performance.

ADP bins numerical attributes for O-Cluster. It uses a specialized form of equi-width binning that computes the number of bins per attribute automatically. Numerical columns with all nulls or a single value are removed. O-Cluster handles missing values naturally as missing at random.

Note:

O-Cluster does not support nested columns, sparse data, or unstructured text.

Related Topics

- Prepare the Data

User-Specified Data Preparation for O-Cluster

You can prepare the data for O-Cluster by considering equi-width binning and managing outliers.

Keep the following in mind if you choose to prepare the data for O-Cluster:

- O-Cluster does not necessarily use all the input data when it builds a model. It reads the data in batches (the default batch size is 50000). It only reads another batch if it believes, based on statistical tests, that uncovered clusters can still exist.
- Binary attributes must be declared as categorical.
- Automatic equi-width binning is highly recommended. The bin identifiers are expected to be positive consecutive integers starting at 1.

- The presence of outliers can significantly impact clustering algorithms. Use a clipping transformation before binning or normalizing. Outliers with equi-width binning can prevent O-Cluster from detecting clusters. As a result, the whole population appears to fall within a single cluster.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Oracle Machine Learning for SQL with R Extensibility

Learn how you can use Oracle Machine Learning for SQL to build, score, and view OML4SQL models as well as R models.

The Oracle Machine Learning for SQL framework is enhanced extending the Oracle Machine Learning for SQL algorithm set with algorithms from the open source R ecosystem. Oracle Machine Learning for SQL is implemented in the Oracle Database kernel. The Oracle Machine Learning for SQL models are Database schema objects. With the extensibility enhancement, the Oracle Machine Learning for SQL framework can build, score, and view both Oracle Machine Learning for SQL models and R models.

Registration of R scripts

The R engine on the database server runs the R scripts to build, score, and view R models. These R scripts must be registered with the database beforehand by a privileged user with `rqAdmin` role. You must first install Oracle Machine Learning for R to register the R scripts.

Functions of Oracle Machine Learning for SQL with R Model

The following functions are supported for an R model:

- Oracle Machine Learning for SQL `DBMS_DATA_MINING` package is enhanced to support R model. For example, `CREATE_MODEL` and `DROP_MODEL`.
- `MODEL VIEW` to get the R model details about a single model and a partitioned model.
- Oracle Machine Learning for SQL SQL functions are enhanced to operate with the R model functions. For example, `PREDICTION` and `CLUSTER_ID`.

R model extensibility supports the following Oracle Machine Learning for SQL functions:

- Association
- Attribute Importance
- Regression
- Classification
- Clustering
- Feature Extraction

About Algorithm Metadata Registration

Algorithm metadata registration allows for a uniform and consistent approach of registering new algorithm functions and their settings.

Users have the ability to add new R-based algorithms through the registration process. The new algorithms appear as available within Oracle Machine Learning for R and within the appropriate machine learning techniques. Based on the registration metadata, the settings page is dynamically rendered. The advantages are as follows:

- Manage R-based algorithms more easily
- Specify R-based algorithm for model build
- Clean individual properties in JSON structure
- Share R-based algorithm across user

Algorithm metadata registration extends the machine learning model capability of Oracle Machine Learning for SQL.

> **See Also:**
>
> DBMS_DATA_MINING — Algorithm Settings: ALGO_EXTENSIBLE_LANG for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Related Topics

- Create Model Using Registration Information
- FETCH_JSON_SCHEMA Procedure
- REGISTER_ALGORITHM Procedure
- JSON Schema for R Extensible Algorithm

Scoring with R

Oracle Machine Learning for SQL supports R models, enabling scoring and predictions using registered R scripts.

For more information, see *Oracle Machine Learning for SQL User’s Guide*

About Random Forest

Random Forest is a classification algorithm that builds an **ensemble** (also called **forest**) of trees.

The algorithm builds a number of Decision Tree models and predicts using the ensemble. An individual decision tree is built by choosing a random sample from the training data set as the input. At each node of the tree, only a random sample of predictors is chosen for computing the split point. This introduces variation in the data used by the different trees in the forest. The parameters `RFOR_SAMPLING_RATIO` and `RFOR_MTRY` are used to specify the sample size and number of predictors chosen at each node. Users can use `ODMS_RANDOM_SEED` to set the random seed value before running the algorithm.

Related Topics

- About Decision Tree
  Decision Tree classifies data using a tree structure of rules, making predictions clear and easy to interpret.

- Splitting
  Decision Tree uses homogeneity metrics like gini and entropy to create the most homogeneous child nodes.
- Data Preparation for Decision Tree
  The Decision Tree algorithm manages its own data preparation internally. It does not require pretreatment of the data.

Building a Random Forest

Random Forest models provide attribute importance ranking and are built using existing Oracle Machine Learning for SQL APIs.

Random forest models provide attribute importance ranking of predictors. The model is built by specifying parameters in the existing APIs. The scoring is performed using the same SQL queries and APIs as the existing classification algorithms. Oracle Machine Learning for SQL implements a variant of classical Random Forest algorithm. This implementation supports big data sets. The implementation of the algorithm differs in the following ways:

- Oracle Machine Learning for SQL does not support bagging and instead provides sampling without replacement
- Users have the ability to specify the depth of the tree. Trees are not built to maximum depth.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Related Topics

- DBMS_DATA_MINING — Algorithm Settings: Random Forest

Scoring with Random Forest

Scoring with Random Forest uses standard classification functions.

Scoring with Random Forest is the same as any other classification algorithm. The following functions are supported: `PREDICTION`, `PREDICTION_PROBABILITY`, `PREDICTION_COST`, `PREDICTION_SET`, and `PREDICTION_DETAILS`.

Related Topics

- *Oracle Database SQL Language Reference*

About Singular Value Decomposition

SVD and the closely-related PCA are well established feature extraction methods that have a wide range of applications. Oracle Machine Learning for SQL implements Singular Value Decomposition (SVD) as a feature extraction algorithm and Principal Component Analysis (PCA) as a special scoring method for SVD models.

SVD and PCA are orthogonal linear transformations that are optimal at capturing the underlying variance of the data. This property is very useful for reducing the dimensionality of high-dimensional data and for supporting meaningful data visualization.

SVD and PCA have a number of important applications in addition to dimensionality reduction. These include matrix inversion, data compression, and the imputation of unknown data values.

Matrix Manipulation

SVD decomposes a matrix into orthonormal bases, capturing data variance and aligning with maximum variance directions.

Singular Value Decomposition (SVD) is a factorization method that decomposes a rectangular matrix **X** into the product of three matrices: **U**, **S**, and **V**.

Figure 7-12 Matrix Manipulation

$$\mathbf{X = USV'}$$

The **U** matrix consists of a set of 'left' orthonormal bases
The **S** matrix is a diagonal matrix
The **V** matrix consists of set of 'right' orthonormal bases

The values in **S** are called singular values. They are non-negative, and their magnitudes indicate the importance of the corresponding bases (components). The singular values reflect the amount of data variance captured by the bases. The first basis (the one with largest singular value) lies in the direction of the greatest data variance. The second basis captures the orthogonal direction with the second greatest variance, and so on.

SVD essentially performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance in the data. This is a useful procedure under the assumption that the observed data has a high signal-to-noise ratio and that a large variance corresponds to interesting data content while a lower variance corresponds to noise.

SVD makes the assumption that the underlying data is Gaussian distributed and can be well described in terms of means and covariances.

Low Rank Decomposition

Singular Value Decomposition (SVD) keeps lower-order bases (the ones with the largest singular values) and ignores higher-order bases (the ones with the smallest singular values) to capture the most important aspects of the data.

To reduce dimensionality, SVD keeps lower-order bases and ignores higher-order bases. The rationale behind this strategy is that the low-order bases retain the characteristics of the data that contribute most to its variance and are likely to capture the most important aspects of the data.

Given a data set **X** (*nxm*), where *n* is the number of rows and *m* is the number of attributes, a low-rank SVD uses only *k* components (*k* <= **min**(*m*, *n*)). In typical implementations of SVD, the value of *k* requires a visual inspection of the ranked singular values associated with the individual components. In Oracle Machine Learning for SQL, SVD automatically estimates the cutoff point, which corresponds to a significant drop in the explained variance.

SVD produces two sets of orthonormal bases (**U** and **V**). Either of these bases can be used as a new coordinate system. In Oracle Machine Learning for SQL, SVD, **V** is the new coordinate system, and **U** represents the projection of **X** in this coordinate system. The algorithm computes the projection of new data as follows:

Figure 7-13 Computing Projection of New Data

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_k\mathbf{S}_k^{-1}$$

where **X** (*n*x*k*) is the projected data in the reduced data space, defined by the first *k* components, and $\mathbf{V}_k$ and $\mathbf{S}_k$ define the reduced component set.

Scalability

SVD processes large data sets efficiently, recommending appropriate feature numbers for dimensionality reduction.

Singular Value Decomposition (SVD) can process data sets with millions of rows and thousands of attributes. Oracle Machine Learning automatically recommends an appropriate number of features, based on the data, for dimensionality reduction.

SVD has linear scalability with the number of rows and cubic scalability with the number of attributes when a full decomposition is computed. A low-rank decomposition is typically linear with the number of rows and linear with the number of columns. The scalability with the reduced rank depends on how the rank compares to the number of rows and columns. It can be linear when the rank is significantly smaller or cubic when it is on the same scale.

Configuring the Algorithm

Configure SVD for optimal performance, model size, and projection methods, including PCA.

Several options are available for configuring the Singular Value Decomposition (SVD) algorithm. Among several options are: settings to control model size and performance, and whether to score with SVD projections or Principal Component Analysis (PCA) projections.

> **See Also:**
>
> DBMS_DATA_MINING — Algorithm Constants and Settings: Singular Value Decomposition for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Model Size

Learn how a model size is decided based on the rows in the build data and algorithm-specific setting.

The **U** matrix in Singular Value Decomposition has as many rows as the number of rows in the build data. To avoid creating a large model, the **U** matrix persists only when an algorithm-specific setting is enabled. By default the **U** matrix does not persist.

Performance

Singular Value Decomposition can use approximate computations to improve performance.

Approximation may be appropriate for data sets with many columns. An approximate low-rank decomposition provides good solutions at a reasonable computational cost. The quality of the approximation is dependent on the characteristics of the data.

PCA scoring

Learn about configuring Singular Value Decomposition (SVD) to perform Principal Component Analysis (PCA) projections.

SVD models can be configured to perform PCA projections. PCA is closely related to SVD. PCA computes a set of orthonormal bases (principal components) that are ranked by their corresponding explained variance. The main difference between SVD and PCA is that the PCA projection is not scaled by the singular values. The PCA projection to the new coordinate system is given by:

Figure 7-14 PCA Projection Calculation

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_k$$

where

$$\overline{X}$$

(*n*x*k*) is the projected data in the reduced data space, defined by the first *k* components, and $\mathbf{V}_k$ defines the reduced component set.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Data Preparation for SVD

Prepare data for Singular Value Decomposition using Automatic Data Preparation for numerical and categorical attributes.

When the build data is scored with SVD, Automatic Data Preparation does nothing. When the build data is scored with Principal Component Analysis (PCA), Automatic Data Preparation shifts the numerical data by mean.

Missing value treatment is not needed, because Oracle Machine Learning algorithms handle missing values automatically. SVD replaces numerical missing values with the mean and categorical missing values with the mode. For sparse data (missing values in nested columns), SVD replaces missing values with zeros.

Related Topics

- Prepare the Data

About Support Vector Machine

Support Vector Machine (SVM) is a powerful, state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory.

SVM has strong **regularization** properties. Regularization refers to the generalization of the model to new data.

Advantages of SVM

Support Vector Machine (SVM) implements solvers for scalability and handling large volumes of data.

Oracle Machine Learning for SQL SVM implementation includes two types of solvers, an Interior Point Method (IPM) solver and a Sub-Gradient Descent (SGD) solver. The IPM solver provides stable and accurate solutions, however, it may not be able to handle data of high dimensionality. For high-dimensional and/or large data, for example, text, ratings, and so on, the SGD solver is a better choice. Both solvers have highly scalable parallel implementations and can handle large volumes of data.

Advantages of SVM in Oracle Machine Learning for SQL

Oracle's SVM implementation provides scalability, usability, and enhanced performance.

Oracle Machine Learning has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. SVM provides the scalability and usability that are needed in a production quality Oracle Machine Learning for SQL system.

Usability

Oracle's SVM minimizes data preparation and tuning, making it accessible for non-experts.

Usability is a major enhancement, because SVM has often been viewed as a tool for experts. The algorithm typically requires data preparation, tuning, and optimization. Oracle Machine Learning minimizes these requirements. You do not need to be an expert to build a quality SVM model. For example:

- Data preparation is not required in most cases.
- Default tuning parameters are generally adequate.

Related Topics

- Data Preparation for SVM
  Support Vector Machine (SVM) uses normalization and missing value treatment for data preparation.
- Tuning an SVM Model
  The Support Vector Machine (SVM) algorithm has built-in mechanisms that automatically choose appropriate settings based on the data.

Scalability

Oracle's SVM uses stratified sampling and incremental model building for large data sets.

When dealing with very large data sets, sampling is often required. However, sampling is not required with Oracle Machine Learning for SQL SVM, because the algorithm itself uses stratified sampling to reduce the size of the training data as needed.

Oracle's SVM is highly optimized. It builds a model incrementally by optimizing small working sets toward a global solution. The model is trained until convergence on the current working set, then the model adapts to the new data. The process continues iteratively until the convergence conditions are met. The Gaussian kernel uses caching techniques to manage the working sets.

Related Topics

- Kernel-Based Learning
  Learn about kernal-based functions to transform the input data for Support Vector Machine (SVM).

Kernel-Based Learning

Learn about kernal-based functions to transform the input data for Support Vector Machine (SVM).

SVM is a kernel-based algorithm. A **kernel** is a function that transforms the input data to a high-dimensional space where the problem is solved. Kernel functions can be linear or nonlinear.

Oracle Machine Learning for SQL supports linear and Gaussian (nonlinear) kernels.

In Oracle Machine Learning for SQL, the **linear kernel** function reduces to a linear equation on the original attributes in the training data. A linear kernel works well when there are many attributes in the training data.

The **Gaussian kernel** transforms each case in the training data to a point in an *n*-dimensional space, where *n* is the number of cases. The algorithm attempts to separate the points into subsets with homogeneous target values. The Gaussian kernel uses nonlinear separators, but within the kernel space it constructs a linear equation.

> **Note:**
>
> Active Learning is not relevant in Oracle Database 12*c* Release 2 and later. A setting similar to Active Learning is `ODMS_SAMPLING`.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Tuning an SVM Model

The Support Vector Machine (SVM) algorithm has built-in mechanisms that automatically choose appropriate settings based on the data.

You may need to override the system-determined settings for some domains.

Settings pertain to regression, classification, and anomaly detection unless otherwise specified.

> **See Also:**
>
> DBMS_DATA_MINING —Algorithm Settings: Support Vector Machine for a listing and explanation of the available model settings.

> **Note:**
>
> The term hyperparameter is also interchangeably used for model setting.

Data Preparation for SVM

Support Vector Machine (SVM) uses normalization and missing value treatment for data preparation.

The SVM algorithm operates natively on numeric attributes. SVM uses z-score normalization on numeric attributes. The normalization occurs only for two-dimensional numeric columns (not nested). The algorithm automatically "explodes" categorical data into a set of binary attributes, typically one per category value. For example, a character column for marital status with values `married` or `single` is transformed to two numeric attributes: `married` and `single`. The new attributes can have the value `1` (true) or `0` (false).

When there are missing values in columns with simple data types (not nested), SVM interprets them as missing at random. The algorithm automatically replaces missing categorical values with the mode and missing numerical values with the mean.

When there are missing values in the nested columns, SVM interprets them as sparse. The algorithm automatically replaces sparse numerical data with zeros and sparse categorical data with zero vectors.

Normalization

Normalization ensures numeric attributes are on the same scale, preventing bias in the SVM model.

SVM require the normalization of numeric input. Normalization places the values of numeric attributes on the same scale and prevents attributes with a large original scale from biasing the solution. Normalization also minimizes the likelihood of overflows and underflows.

SVM and Automatic Data Preparation

You can prepare data by treating and transforming data manually or through Automatic Data Preparation (ADP) for Support Vector Machine (SVM).

The SVM algorithm automatically handles missing value treatment and the transformation of categorical data, but normalization and outlier detection must be handled by ADP or prepared manually. ADP performs min-max normalization for SVM.

> **Note:**
>
> Oracle recommends that you use ADP with SVM. The transformations performed by ADP are appropriate for most models.

Related Topics

- *Oracle Machine Learning for SQL User’s Guide*

SVM Classification

Support Vector Machine (SVM) classification is based on the concept of decision planes that define decision boundaries.

A decision plane is one that separates between a set of objects having different class memberships. SVM finds the vectors ("support vectors") that define the separators giving the widest separation of classes.

SVM classification supports both binary, multiclass, and multitarget classification. Multitarget alllows multiple class labels to be associated with a single row. The target type is a collection of type `ORA_MINING_VARCHAR2_NT`.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Class Weights

Implement class weights in SVM to bias the model towards under-represented classes.

In SVM classification, weights are a biasing mechanism for specifying the relative importance of target values (classes).

SVM models are automatically initialized to achieve the best average prediction across all classes. However, if the training data does not represent a realistic distribution, you can bias the model to compensate for class values that are under-represented. If you increase the weight for a class, then the percent of correct predictions for that class must increase.

Related Topics

- Priors and Class Weights
  Offset differences in data distribution with prior probabilities and class weights to produce useful classification results.

One-Class SVM

One-Class SVM detects anomalies by identifying cases that deviate from normal data patterns.

Oracle Machine Learning uses SVM as the one-class classifier for anomaly detection. When SVM is used for anomaly detection, it has the classification machine learning technique but no target.

One-class SVM models, when applied, produce a prediction and a probability for each case in the scoring data. If the prediction is 1, the case is considered typical. If the prediction is 0, the case is considered anomalous. This behavior reflects the fact that the model is trained with normal data.

You can specify the percentage of the data that you expect to be anomalous with the `SVMS_OUTLIER_RATE` build setting. If you have some knowledge that the number of "suspicious" cases is a certain percentage of your population, then you can set the outlier rate to that percentage. The model approximately identifies that many "rare" cases when applied to the general population.

SVM Regression

SVM regression uses epsilon-insensitivity loss function to achieve generalization and minimal error.

SVM uses an epsilon-insensitive loss function to solve regression problems.

SVM regression tries to find a continuous function such that the maximum number of data points lie within the epsilon-wide insensitivity tube. Predictions falling within epsilon distance of the true target value are not interpreted as errors.

The epsilon factor is a regularization setting for SVM regression. It balances the margin of error with model robustness to achieve the best generalization to new data.

Related Topics

- SVM Model Settings

About XGBoost

Oracle's XGBoost prepares training data, builds and persists a model, and applies the model for classification and regression.

Oracle Machine Learning for SQL XGBoost is a scalable gradient tree boosting system that supports both classification and regression. It makes available the open source gradient boosting framework.

You can use XGBoost as a stand-alone predictor or incorporate it into real-world production pipelines for a wide range of problems such as ad click-through rate prediction, hazard risk prediction, web text classification, and so on.

The Oracle Machine Learning for SQL XGBoost algorithm takes three types of parameters: general parameters, booster parameters, and task parameters. You set the parameters through the model settings table. The algorithm supports most of the settings of the open source project.

Through XGBoost, Oracle Machine Learning for SQL supports a number of different classification and regression specifications, ranking models, and survival models. Binary and multiclass models are supported under the classification machine learning technique while regression, ranking, count, and survival are supported under the regression machine learning technique.

XGBoost also supports partitioned models and internalizes the data preparation. Currently, XGBoost is available only on Oracle Database Linux platform.

XGBoost Feature Constraints

Feature interaction constraints allow users to specify which variables can and cannot interact. By focusing on key interactions and eliminating noise, it aids in improving predicting performance. This, in turn, may lead to more generalized predictions.

The feature interaction constraints are described in terms of groupings of features that are allowed to interact. Variables that appear together in a traversal path in decision trees interact with one another because the condition of a child node is dependent on the condition of the parent node. These additional controls on model fit are beneficial to users who have a good understanding of the modeling task, including domain knowledge. Oracle Machine Learning for SQL supports more of the available XGBoost capabilities once these constraints are applied.

Monotonic constraints allow you to impose monotonicity constraints on the features in your boosted model. There may be a strong prior assumption that the genuine relationship is constrained in some way in many circumstances. This could be owing to commercial factors (just specific feature interactions are of interest) or the type of scientific subject under investigation. A typical form of constraint is that some features have a monotonic connection to the predicted response. In these situations, monotonic constraints may be employed to improve the model's prediction performance. For example, let *X* be the feature vector with features *[x1,…, xi , …, xn]* and *ƒ(X)* be the prediction response. Then `ƒ(X) ≤ ƒ(X′)` whenever `xi ≤ xi′` is an increasing constraint; `ƒ(X) ≥ ƒ(X′)` whenever `xi ≤ xi′` is a decreasing constraint. These feature constraints are listed in DBMS_DATA_MINING — Algorithm Settings: XGBoost.

The following example displays the code snippet for defining feature constraints using the XGBoost algorithm. XGBoost `interaction_constraints` setting is used to specify the interaction constraints. The example predicts customers most likely to respond positively for an affinity card loyalty program.

```
-----------------------------------------------------------------------
--   Build a Classification Model using Interaction Contraints
-----------------------------------------------------------------------
-- The interaction constraints setting can be used to specify permitted
-- interactions in the model. The constraints must be specified
-- in the form of nested list, where each inner list is a group of
-- features (column names) that are allowed to interact with each other.
-- For example, assume x0, x1, x2, x3, x4, x5 and x6 are
-- the feature names (column names) of interest.
-- Then setting value [[x0,x1,x2],[x0,x4],[x5,x6]] specifies that:
--   * Features x0, x1 and x2 are allowed to interact with each other
--   but with no other feature.
--   * Features x0 & x4 are allowed to interact with one another
--   but with no other feature.
--   * Features x5 and x6 are allowed to interact with each other
--   but with no other feature.
-------------------------------------------------------------------------

BEGIN DBMS_DATA_MINING.DROP_MODEL('XGB_CLASS_MODEL_INTERACTIONS');
EXCEPTION WHEN OTHERS THEN NULL; END;
/

DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME')   := 'ALGO_XGBOOST';
    v_setlst('PREP_AUTO')   := 'ON';
    v_setlst('max_depth')   := '2';
    v_setlst('eta')         := '1';
    v_setlst('num_round')   := '100';
    v_setlst('interaction_constraints')  := '[[YRS_RESIDENCE, OCCUPATION],
                                              [OCCUPATION, Y_BOX_GAMES],
                                              [BULK_PACK_DISKETTES,
                                              BOOKKEEPING_APPLICATION]]';

    DBMS_DATA_MINING.CREATE_MODEL2(
            MODEL_NAME          => 'XGB_CLASS_MODEL_INTERACTIONS',
            MINING_FUNCTION     => 'CLASSIFICATION',
            DATA_QUERY          => 'SELECT * FROM TRAIN_DATA_CLAS',
            SET_LIST            =>  v_setlst,
            CASE_ID_COLUMN_NAME => 'CUST_ID',
            TARGET_COLUMN_NAME  => 'AFFINITY_CARD');

    DBMS_OUTPUT.PUT_LINE('Created model: XGB_CLASS_MODEL_INTERACTIONS');
```

```
END;
/
```

To view the complete example, see https://github.com/oracle-samples/oracle-db-examples/tree/main/machine-learning/sql/23ai.

XGBoost AFT Model

Survival analysis is a field of statistics that examines the time elapsed between one or more occurrences, such as death in biological organisms and failure in mechanical systems.

The goals of survival analysis include evaluating patterns of event times, comparing distributions of survival times in different groups of people, and determining if and how much certain factors affect the likelihood of an event of interest. The existence of censored data is an important feature of survival analysis. If a person does not experience an event within the observation period, they are labeled as censored. **Censoring** is a type of missing data problem in which the time to event is not recorded for a variety of reasons, such as the study being terminated before all enrolled subjects have demonstrated the event of interest, or the subject leaving the study before experiencing an event. Right censoring is defined as knowing only the lower limit *l* for the genuine event time *T* such that $T > l$. Right censoring will take place, for example, for those subjects whose birth date is known but who are still living when they are lost to follow-up or when the study concludes. We frequently come upon data that has been right-censored. The data is said to be left-censored if the event of interest occurred before the subject was included in the study but the exact date is unknown. Interval censoring occurs when an occurrence can only be described as occurring between two observations or examinations.

The Cox proportional hazards model and the Accelerated Failure Time (AFT) model are two major survival analysis methods. Oracle Machine Learning for SQL supports both these models.

Cox regression works for right censored survival time data. The hazard rate is the risk of failure (that is, the risk or likelihood of suffering the event of interest) in a Cox proportional hazards regression model, assuming that the subject has lived up to a particular time. The Cox predictions are returned on a hazard ratio scale. A Cox proportional hazards model has the following form:

$h\ (t,x) = h_0(t)e^{\beta x}$

Where *h(t)* is the baseline hazard, *x* is a covariate, and *β* is an estimated parameter that represents the covariate's effect on the outcome. A Cox proportional hazards model's estimated amount is understood as relative risk rather than absolute risk.

The AFT model fits models to data that can be censored to the left, right, or interval. The AFT model, which models time to an event of interest, is one of the most often used models in survival analysis. AFT is a parametric (it assumes the distribution of response data) survival model. The outcome of AFT models has a physical interpretation that is intuitive. The model has the following form:

$ln\ Y = < W, X> + \sigma Z$

Where *X* is the vector in R^d representing the features. W is a vector consisting of *d* coefficients, each corresponding to a feature. *<W, X>* is the usual dot product in R^d. *Y* is the random variable modeling the output label. *Z* is a random variable of a known probability distribution. Common choices are the normal distribution, the logistic distribution, and the extreme distribution. It represents the “noise”. σ is a parameter that scales the size of noise.

AFT model that works with XGBoost or gradient boosting has the following form:

$ln Y = T(x) + \sigma Z$

Where *T(x)* represents the output of a decision tree ensemble, using the supplied input *x*. Since *Z* is a random variable, you have a likelihood defined for the expression $lnY=T(x)+\sigma Z$. As a result, XGBoost's purpose is to maximize (log) likelihood by fitting a suitable tree ensemble *T(x)*.

The AFT parameters are listed in DBMS_DATA_MINING — Algorithm Settings: XGBoost.

The following example displays code snippet of survival analysis using the XGBoost algorithm. In this example, a `SURVIVAL_DATA` table is created that contains data for survival analysis. XGBoost AFT settings `aft_right_bound_column_name`, `aft_loss_distribution`, and `aft_loss_distribution_scale` are illustrated in this example.

```
-----------------------------------------------------------------------------
--         Create a data table with left and right bound columns
-----------------------------------------------------------------------------

-- The data table 'SURVIVAL_DATA' contains both exact data point and
-- right-censored data point. The left bound column is set by
-- parameter target_column_name. The right bound column is set
-- by setting aft_right_bound_column_name.

-- For right censored data point, the right bound is infinity,
-- which is represented as NULL in the right bound column.

BEGIN EXECUTE IMMEDIATE 'DROP TABLE SURVIVAL_DATA';
EXCEPTION WHEN OTHERS THEN NULL; END;
/
CREATE TABLE SURVIVAL_DATA (INST NUMBER, LBOUND NUMBER, AGE NUMBER,
                            SEX NUMBER, PHECOG NUMBER, PHKARNO NUMBER,
                            PATKARNO NUMBER, MEALCAL NUMBER, WTLOSS NUMBER,
                            RBOUND NUMBER);
INSERT INTO SURVIVAL_DATA VALUES(26, 235, 63, 2, 0, 100,  90,  413,  0,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(22, 444, 75, 2, 2,  70,  70,  438,  8,
444);
INSERT INTO SURVIVAL_DATA VALUES(16, 806, 44, 1, 1,  80,  80, 1025,  1,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(16, 551, 77, 2, 2,  80,  60,  750, 28,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(3,  202, 50, 2, 0, 100, 100,  635,  1,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(7,  583, 68, 1, 1,  60,  70, 1025,  7,
583);
INSERT INTO SURVIVAL_DATA VALUES(32, 135, 60, 1, 1,  90,  70, 1275,  0,
135);
INSERT INTO SURVIVAL_DATA VALUES(21, 237, 69, 1, 1,  80,  70, NULL, NULL,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(26, 356, 53, 2, 1,  90,  90, NULL,   2,
NULL);
INSERT INTO SURVIVAL_DATA VALUES(13, 387, 56, 1, 2,  80,  60, 1075, NULL,
387);

-----------------------------------------------------------------------------
--             Build an XGBoost survival model with survival:aft
-----------------------------------------------------------------------------
```

```
BEGIN DBMS_DATA_MINING.DROP_MODEL('XGB_SURVIVAL_MODEL');
EXCEPTION WHEN OTHERS THEN NULL; END;
/
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME')                    := 'ALGO_XGBOOST';
    v_setlst('max_depth')                    := '6';
    v_setlst('eval_metric')                  := 'aft-nloglik';
    v_setlst('num_round')                    := '100';
    v_setlst('objective')                    := 'survival:aft';
    v_setlst('aft_right_bound_column_name')  := 'rbound';
    v_setlst('aft_loss_distribution')        := 'normal';
    v_setlst('aft_loss_distribution_scale')  := '1.20';
    v_setlst('eta')                          := '0.05';
    v_setlst('lambda')                       := '0.01';
    v_setlst('alpha')                        := '0.02';
    v_setlst('tree_method')                  := 'hist';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'XGB_SURVIVAL_MODEL',
        MINING_FUNCTION     => 'REGRESSION',
        DATA_QUERY          => 'SELECT * FROM SURVIVAL_DATA',
        TARGET_COLUMN_NAME  => 'LBOUND',
        CASE_ID_COLUMN_NAME =>  NULL,
        SET_LIST            =>  v_setlst);
END;
/
```

To view the complete example, see https://github.com/oracle-samples/oracle-db-examples/blob/main/machine-learning/sql/23ai/oml4sql-survival-analysis-xgboost.sql.

Scoring with XGBoost

Score with XGBoost using the supported SQL functions to predict values.

The SQL scoring functions supported for a classification XGBoost model are `PREDICTION`, `PREDICTION_COST`, `PREDICTION_DETAILS`, `PREDICTION_PROBABILITY`, and `PREDICTION_SET`.

The scoring functions supported for a regression XGBoost model are `PREDICTION` and `PREDICTION_DETAILS`.

The prediction functions return the following information:

- `PREDICTION` returns the predicted value.
- `PREDICTION_COST` returns a measure of cost for a given prediction as an Oracle NUMBER. (classification only)
- `PREDICTION_DETAILS` returns the SHAP (SHapley Additive exPlanation) contributions.
- `PREDICTION_PROBABILITY` returns the probability for a given prediction. (classification only)
- `PREDICTION_SET` returns the prediction and the corresponding prediction probability for each observation. (classification only)

Integration of ONNX Runtime

Learn about ONNX Runtime that enables you to use ONNX models for machine learning tasks within your Oracle Database instance.

- About ONNX
- Examples of Using ONNX Models
- Traditional Machine Learning ONNX Format Models
- Text Transformer ONNX Format Models
- Image Transformer ONNX Format Models

About ONNX

ONNX is an open-source format designed for machine learning models. It ensures cross-platform compatibility. This format also supports major languages and frameworks, facilitating efficient model exchange.

The ONNX format allows for model serialization. It simplifies the exchange of models across various platforms. These platforms include cloud, web, edge, and mobile experiences on Microsoft Windows, Linux, Mac, iOS, and Android. ONNX models also offer flexibility to export and import model in many languages such as Python, C++, C#, and Java to name a few. The ONNX format is useful for compute-heavy tasks such as training machine learning models and data processing that often uses trained models. Many leading machine learning development frameworks such as TensorFlow, Pytorch, and Scikit-learn, offer the capability to convert models into the ONNX format.

Once you represent the models in the ONNX format, you can run them with the ONNX Runtime. The architecture of the ONNX Runtime is adaptable, enabling providers to modify or enhance how some operations are implemented to make better use of particular hardware, such as, Graphical Processing Units (GPUs), Single Instruction Multiple Data (SIMD) instruction sets or specialized libraries. To learn more on ONNX Runtime, see https://onnxruntime.ai/docs/.

The ONNX Runtime integration with Oracle Database allows for the import of ONNX-formatted models, including embedding models. To support embedding models, Oracle Machine Learning has introduced a new machine learning technique called *embedding*. If you do not have a pretrained model in ONNX format, Oracle offers a Python utility package that automates the conversion for the user. It downloads a pretrained model, converts the model to ONNX format augmented with pre-processsing and post-processing operations and imports the ONNX format model to Oracle Database. For more information on the Python utility tool, see Convert Pretrained Models to ONNX Format.

Oracle supports ONNX Runtime version 1.15.1.

Supported Machine Learning Functions for ONNX Runtime

Describes the supported machine learning functions to import pretrained models and perform scoring.

The following are the supported machine learning functions:

- Classification
- Clustering

- Embedding
- Regression

Supported Attribute Data Types

Discover the supported ONNX input data types mapped to SQL data types.

Data Type	SQL Type	Supported ONNX Data Type
Numerical	BINARY_DOUBLE NUMBER	float, int8, int16, int32, int64, uint8, uint16, uint32, uint64
Categorical	VARCHAR	For VARCHAR type: string
Text	VARCHAR2 CLOB	string
Vectors	VECTOR(float32,*<dimension>*)	float

The following data types are not supported:

- complex64, complex128
- float16, bfloat16
- fp8
- int4, uint4

Supported Target Data Types

Discover the supported ONNX target data types mapped to SQL data types.

Depending on the machine learning function, different scoring functions are used. Different scoring function for same machine learning function can produce different data types. A few points to note:

- Classification models have different rules to determine the type of PREDICTION function to be used. If you are using PREDICTION_PROBABILITY, then BINARY_DOUBLE is returned. See labels in JSON Metadata Parameters for ONNX Models.
- For an embedding model, the VECTOR_EMBEDDING function returns a VECTOR type.
- For a regression model, VARCHAR is not a valid target type and BINARY_DOUBLE is returned.
- For a clustering model, if you are using CLUSTERING_PROBABILITY and CLUSTER_DISTANCE, then BINARY_DOUBLE is returned.

To learn more, see JSON Metadata Parameters for ONNX Models

Machine Learning Function	SQL Function	SQL Type	Supported ONNX Target Output
Regression	PREDICTION	BINARY_DOUBLE	regressionOutput
Classification	PREDICTION	VARCHAR2	classificationLabelOutput
Classification	PREDICTION	NUMBER	classificationLabelOutput

Machine Learning Function	SQL Function	SQL Type	Supported ONNX Target Output
Classification	PREDICTION_PROBABILITY	BINARY_DOUBLE	classificationProbOutput
Classification	PREDICTION_SET	set of (NUMBER , BINARY_DOUBLE) set of (target_type, BINARY_DOUBLE)	NA
Clustering	CLUSTER_PROBABILITY	BINARY_DOUBLE	clusteringProbOutput
Clustering	CLUSTER_DISTANCE	BINARY_DOUBLE	clusteringDistanceOutput
Clustering	CLUSTER_SET	set of (NUMBER , BINARY_DOUBLE)	NA
Embedding	VECTOR_EMBEDDING	VECTOR(float32, n)	embeddingOutput

Custom ONNX Runtime Operations

If you are looking to customize a pretrained embedding model by augmenting with pre-processing and post-processing operations, Oracle supports tokenization of an embedding model as a pre-processing operation and pooling and normalization as post-processing custom ONNX Runtime operations for version 1.15.1.

Oracle offers a Python utility that provides a mechanism to augment a pretrained model with tokenization, pooling and normalization. The Python utility can augment the model with pre-processing and post-processing operations and convert a pretrained model to an ONNX format. Models using any other custom operations will fail on import. For details on how to use the Python utility, see Convert Pretrained Models to ONNX Format.

Use PL/SQL Packages to Import Models

Use the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure or the `DBMS_VECTOR.LOAD_ONNX_MODEL` procedure to import ONNX format models. You can then use the imported ONNX format models through a scoring function run by the in-database ONNX Runtime.

- To import a pretrained ONNX format model, use IMPORT_ONNX_MODEL Procedure or LOAD_ONNX_MODEL Procedure.
- To drop an ONNX model, use DROP_ONNX_MODEL. See also DROP_MODEL procedure.
- A complete step-by-step example that illustrates these procedures is in Import ONNX Models and Generate Embeddings.

> **Note:**
>
> In-database embedding models must include tokenization and postprocessing. Providing only the core ONNX model is insufficient, as users would need to handle tokenization externally, pass tensors into the SQL operator, and convert output tensors into vectors.

The `DBMS_DATA_MINING.RENAME_MODEL` procedure is also supported.

Most of the existing Oracle Machine Learning for SQL APIs are available to the ONNX models. As partitioning is not applicable for external pretrained models, ONNX models do not support the following procedures:

- `ADD_PARTITION`
- `DROP_PARTITION`
- `ADD_COST_MATRIX`
- `REMOVE_COST_MATRIX`

Related Topics

- Summary of DBMS_DATA_MINING Subprograms

Supported SQL Scoring Functions

Supported scoring functions for in-database scoring of machine learning models imported in the ONNX format are listed.

Machine Learning Technique	Operator	Supported	Return Type
Embedding	`VECTOR_EMBEDDING`	always	`VECTOR(<dimensions, FLOAT32>)` The number of dimensions of the output vector of a `VECTOR_EMBEDDING` operator is defined by the embedding models.
Regression	`PREDICTION`	always	Data type of the target. For regression, the data type is converted to `BINARY_DOUBLE` SQL type.
Classification	`PREDICTION`	always	Data type of the target.
Classification	`PREDICTION_PROBABILITY`	always	`BINARY_DOUBLE`
Classification	`PREDICTION_SET`	always	Set of `( t, NUMBER , BINARY_DOUBLE )` where `t` is the data type of the target.
Clustering	`CLUSTER_ID`	only if `clusteringProbOutput` is specified	`NUMBER`

Machine Learning Technique	Operator	Supported	Return Type
Clustering	CLUSTER_PROBABILITY	only if clusteringProbOutput is specified	BINARY_DOUBLE
Clustering	CLUSTER_SET	only if clusteringProbOutput is specified	Set of (NUMBER, BINARY_DOUBLE)
Clustering	CLUSTER_DISTANCE	only if clusteringDistanceOutput is specified	BINARY_DOUBLE

Note:

You can define the outputs explicitly in the metadata or implicitly.

- The metadata must explicitly specify how to find the result in the model output for some SQL scoring functions. For example, CLUSTER_PROBABILITY is supported only if clusteringProbOutput is specified in the metadata.
- The system automatically assumes the output for a model with only one output if you don't specify it in the metadata.
- If a scoring function does not comply according to the description provided, you will receive an ORA-40290 error when performing the scoring operation on your data. Additionally, any unsupported scoring functions will raise the ORA-40290 error.

To learn more about classification data types that are returned, see labels and classificationLabelOutput in JSON Metadata Parameters for ONNX Models.

Cost Matrix Clause

Specify a cost matrix directly within the PREDICTION and PREDICTION_SET scoring functions. To learn more about Cost Matrix, see *Oracle Machine Learning for SQL Concepts*.

Examples of Using ONNX Models

The following examples use the Iris data set to showcase loading and inference from ONNX format machine learning models for machine learning techniques such as Classification, Regression, and Clustering in your Oracle Database instance.

Iris is a flower and this data set has information such as petal length, sepal length, petal width, and sepal width collected from three types of Iris flowers: sentosa, versicolour, and virginica.

These examples assume that the data set is available to the user.

ONNX Classification Examples

The following examples showcase various JSON metadata parameters that can be defined for ONNX models.

Example: Specifying JSON Metadata for Classification Models

The following example illustrates JSON metadata parameters with Classification as the function. Assume the model has an output named `probabilities` for the probability of the prediction. To use the `PREDICTION_PROBABILITY` scoring function, you must set the field `classificationProbOutput` to the name of the model output that holds the probability.

```
BEGIN
DBMS_VECTOR.LOAD_ONNX_MODEL('classification_model.onnx', 'doc_model',
JSON('{"function" : "classification",
       "classificationProbOutput": "probabilities"}'));
END;
/
```

Example: Specifying labels in JSON Metadata for Classification Models

The following example illustrates how you can specify custom labels in the JSON metadata.

```
BEGIN
DBMS_VECTOR.LOAD_ONNX_MODEL('classification_model.onnx', 'doc_model',
JSON('{"function" : "classification",
       "classificationProbOutput": "probabilities",
       "labels": ["Setosa", "Versicolour", "Virginica"]}'));
END;
/
```

You can use the `PREDICTION` and `PREDICTION_PROBABILITY` functions for inference or scoring:

```
SELECT
    iris.*,
    PREDICTION(doc_model USING *) as predicted_species_id,
    PREDICTION_PROBABILITY(doc_model, 'setosa' USING *) as setosa_probability
FROM iris;
```

The query predicts `iris` species and the probability of *setosa* species using the `iris` data set. The data from `iris` table is used in a `SELECT` query to predict a species ID and the probability that the species is *setosa* using a machine learning model named `doc_model`. The `PREDICTION` function predicts the species based on the attributes in the table, and the `PREDICTION_PROBABILITY` function computes the probability that the predicted species is *setosa*. The result includes all columns from the `iris` view along with the predicted species ID and the probability of the species being *setosa*.

Example: Specifying input in JSON Metadata for Classification Models

The following example illustrates how you can specify input attribute names that map to the actual ONNX model input names. This example assumes a model with four inputs named `SEPAL_LENGTH`, `SEPAL_WIDTH`, `PETAL_LENGTH`, and `PETAL_WIDTH`. You can specify alternative input attribute names using the JSON metadata as shown in this example. Here, each input is assumed to be a tensor with a dimension of 1. The `input` field must be a JSON object where each field is a model input name (For example, `SEPAL_LENGTH`), and its value is a JSON array sized according to the tensor's dimension (here, 1) with one attribute name per element in the array.

```
BEGIN DBMS_VECTOR.LOAD_ONNX_MODEL('classification_model.onnx', 'doc_model',
JSON('{"function" : "classification",
       "classificationProbOutput": "probabilities",
       "input": { "SEPAL_LENGTH": ["SEPAL_LENGTH_CM"],
                  "SEPAL_WIDTH": ["SEPAL_WIDTH_CM"],
                  "PETAL_LENGTH": ["PETAL_LENGTH_CM"],
                  "PETAL_WIDTH": ["PETAL_WIDTH_CM"] } }'));
END;
/
```

You can also have a different order of the columns as input.

```
BEGIN DBMS_VECTOR.LOAD_ONNX_MODEL('classification_model.onnx', 'doc_model',
    JSON('{"function" : "classification",
           "classificationProbOutput": "probabilities",
           "input": { "SEPAL_WIDTH": ["SEPAL_WIDTH_CM"],
                   "PETAL_LENGTH": ["PETAL_LENGTH_CM"],
                   "PETAL_WIDTH": ["PETAL_WIDTH_CM"],
                 "SEPAL_LENGTH": ["SEPAL_LENGTH_CM"] } }'));
END;
/
```

Example: Specifying a Single input With Four Dimensions

Here is an example where the model has a single input tensor named `x` with four dimensions. The corresponding JSON metadata for this scenario is:

```
JSON('{"function" : "classification",
       "classificationProbOutput": "probabilities",
       "input": { "x": ["SEPAL_LENGTH_CM",
                        "SEPAL_WIDTH_CM",
                        "PETAL_LENGTH_CM",
                        "PETAL_WIDTH_CM"]
                      }'));
```

You can use `PREDICTION` and `PREDICTION_PROBABILITY` functions for inference or scoring.

```
WITH
dummy_iris AS (
    SELECT
    4.5 as petal_length_cm,
    1.5 as petal_width_cm,
    4.3 as sepal_length_cm,
    2.9 as sepal_width_cm
    FROM iris
)
SELECT
    dummy_iris.*,
    PREDICTION(doc_model USING *) as predicted_species_id,
    PREDICTION_PROBABILITY(doc_model 'setosa' USING *) as setosa_probability
FROM dummy_iris;
```

The query predicts `iris` species and the probability of *setosa* species using specified attributes in a temporary data set. The query creates a temporary `dummy_iris` view with attributes values set. This temporary view is then used in a `SELECT` query to predict a species ID and the probability that the species is *setosa* using a machine learning model named `doc_model`. The `PREDICTION` function predicts the species based on the attributes provided, and the `PREDICTION_PROBABILITY` function computes the probability that the predicted species is *setosa*. The result includes all columns from the `dummy_iris` view along with the predicted species ID and the probability of the species being *setosa*.

Example: Specifying defaultOnNull in JSON Metadata for Classification Models

The following examples illustrates how you can specify `defaulOnNull` provides default values to be used for specific attributes when their values are NULL in the data set. Use the names `SEPAL_LENGTH`, `SEPAL_WIDTH`, `PETAL_LENGTH`, and `PETAL_WIDTH` as fields in the `defaultOnNull` object, which are the assumed input attribute names for a ONNX model with four inputs. These names serve as the default input attribute names, so you can use them as fields in the `defaultOnNull`.

```
BEGIN DBMS_VECTOR.LOAD_ONNX_MODEL('classification_model.onnx', 'doc_model',
    JSON('{"function" : "classification",
           "classificationProbOutput": "probabilities",
           "defaultOnNull": {"SEPAL_LENGTH": "5.1",
           "SEPAL_WIDTH": "3.5",
           "PETAL_LENGTH": "1.4",
           "PETAL_WIDTH": "0.2"}}'));
END;
/
```

- `"SEPAL_LENGTH": "5.1"`: If the sepal length is null, use 5.1 as the default value.
- `"SEPAL_WIDTH": "3.5"`: If the sepal width is null, use 3.5 as the default value.
- `"PETAL_LENGTH": "1.4"`: If the petal length is null, use 1.4 as the default value.
- `"PETAL_WIDTH": "0.2"`: If the petal width is null, use 0.2 as the default value.

Example: Specifying input and defaultOnNull JSON Metadata for Classification Models

Here is a combined example of specifying `input` and `defaultOnNull` values. This example uses the values that were illustrated in the earlier examples where `input` and `defaultOnNull` values are specified:

```
JSON('{"function" : "classification",
               "classificationProbOutput": "probabilities",
               "input": { "SEPAL_WIDTH": ["SEPAL_WIDTH_CM"],
                          "PETAL_LENGTH": ["PETAL_LENGTH_CM"],
                          "PETAL_WIDTH": ["PETAL_WIDTH_CM"],
                          "SEPAL_LENGTH": ["SEPAL_LENGTH_CM"] },
                    "defaultOnNull": {"SEPAL_LENGTH_CM": "5.1",
                                      "SEPAL_WIDTH_CM": "3.5"}}')
```

ONNX Clustering Examples

The following examples showcase various JSON metadata parameters that can be defined for ONNX models.

Example: Specifying JSON Metadata for Clustering Models

The following example illustrates JSON metadata parameters with Clustering as the function. Assume the model has an output named `probabilities` for the probability of the prediction. To use the `CLUSTER_PROBABILITY` scoring function, you must set the field `clusteringProbOutput` to the name of the model output that holds the probability.

```
BEGIN
DBMS_VECTOR.LOAD_ONNX_MODEL('clustering_model.onnx','doc_model',
    JSON('{"function": "clustering",
           "clusteringProbOutput": "probabilities"
        }
    ')
);
END;
/
```

You can use `CLUSTER_ID` and `CLUSTER_PROBABILITY` functions for inference or scoring.

```
SELECT
    iris.*,
    CLUSTER_ID(doc_model USING *) as cluster_id,
    CLUSTER_PROBABILITY(doc_model, 1 USING *) as cluster_1_probability
FROM iris;
```

This query predicts the cluster assignments and the probabilities of belonging to a specific cluster for each record of the `iris` data set. The query retrieves all columns of each record (`iris.*`) and applies the clustering model named `doc_model` to each record of the `iris` data set and predicts the cluster ID. The `USING *` clause tells the model to use all available columns in the `iris` table for this prediction. The `CLUSTER_PROBABILITY(doc_model, 1 USING *) as cluster_1_probability` part of the query calculates the probability that each record belongs to cluster 1, according to the `doc_model` from the `iris` data set. This provides insights into how likely each record is to be part of cluster 1, giving a quantitative measure of membership strength.

Example: Specifying clusteringDistanceOutput in JSON Metadata for Clustering Models

The following example illustrates how you can specify `clusteringDistanceOutput` and for ONNX Clustering models.

In this model, an output tensor named `distances` provides distances for the input, which is a single tensor named `float_input` with a dimension of 4. The JSON metadata `input` field must map attribute names to entries of the tensor, such as `"SEPAL_LENGTH", "SEPAL_WIDTH", "PETAL_LENGTH", "PETAL_WIDTH"`.

```
BEGIN
DBMS_VECTOR.LOAD_ONNX_MODEL('clustering_model.onnx', 'doc_model',
JSON('{"function" : "clustering",
"clusteringDistanceOutput": "distances",
"normalizeProb": "softmax",
"input": { "float_input": ["SEPAL_LENGTH", "SEPAL_WIDTH", "PETAL_LENGTH",
"PETAL_WIDTH"] }
      }')
);
```

```
END;
/
```

You can use CLUSTER_DISTANCE function for inference or scoring. These SQL queries utilize clustering models to predict cluster distances from the IRIS data set.

```
SELECT CLUSTER_DISTANCE(doc_model USING *) AS predicted_target_value,
CLUSTER_DISTANCE (doc_model,1 USING *) AS dist1,
CLUSTER_DISTANCE (doc_model,2 USING *) AS dist2,
CLUSTER_DISTANCE (doc_model,3 USING *) AS dist3
FROM IRIS
ORDER BY ID
FETCH NEXT 10 ROWS ONLY;
```

Here, the query focuses on understanding the physical distance of data points from cluster centroids, which is particularly useful for identifying outliers or for performing detailed cluster analysis. The query calculates the distance of each record in the IRIS data set from the centroids of different clusters using the doc_model. The USING * syntax indicates that the model must use all available columns of the IRIS data set for making the prediction. CLUSTER_DISTANCE(doc_model, n USING *) computes the distance from cluster n (n being 1, 2, and 3 in this query). Each distance is selected as a separate column (dist1, dist2, dist3).

The output is limited to the first 10 rows of the result set ordered by the ID column of the IRIS table.

Example: Specifying clusteringProbOutput and normalizeProb in JSON Metadata for Clustering Models

The following example illustrates how you can specify clusteringProbOutput and normalizeProb for ONNX Clustering models.

```
BEGIN
DBMS_VECTOR.LOAD_ONNX_MODEL('clustering_model.onnx', 'doc_model',
               JSON('{"function" :
               "clustering",
               "clusteringProbOutput": "probabilities",
                        "normalizeProb" : "softmax",
               "input": { "float_input": ["SEPAL_LENGTH", "SEPAL_WIDTH",
"PETAL_LENGTH", "PETAL_WIDTH"] } }')
      );
END;
/
```

You can use CLUSTER_PROBABILITY and CLUSTER_SET functions for inference or scoring:

```
SELECT CLUSTER_ID (doc_model USING *) AS predicted_target_value,
                    CLUSTER_PROBABILITY (doc_model,1 USING *) AS prob1,
                    CLUSTER_PROBABILITY (doc_model,2 USING *) AS prob2,
                    CLUSTER_PROBABILITY (doc_model,3 USING *) AS prob3
                    FROM IRIS
                    ORDER BY ID
                    FETCH NEXT 10 ROWS ONLY;
```

In this case, a clustering model is used to predict the cluster IDs and associated probabilities for records from the IRIS data set. Because the JSON metadata specifies softmax for the normalizeProb field, the model applies softmax normalization to the probabilities before returning them as the result of the CLUSTER_PROBABILITY scoring operator.

The SQL query selects CLUSTER_ID column from the IRIS table and adds a new column, predicted_target_value, which contains predictions made by the doc_model. The USING * syntax means that all columns of the current row are used as input features for the doc_model model to predict the value as predicted_target_value. The result of this prediction is then included as a new column in the output of the query.

CLUSTER_PROBABILITY(model, n USING *): Computes the probability that the record belongs to cluster n (n being 1, 2, and 3 in this query). This is done for three different clusters, and each probability is selected as a separate column (prob1, prob2, prob3).

The output is limited to the first 10 rows of the result set ordered by the ID column of the IRIS table.

```
SELECT S.CLUSTER_ID, S.PROBABILITY
 FROM (SELECT CLUSTER_SET(doc_model USING *) pset
          FROM IRIS ORDER BY ID) T,
       TABLE(T.pset) S
FETCH NEXT 10 ROWS ONLY;
```

The CLUSTER_SET query generates a set of cluster data using the doc_model. The resultant column pset represents all possible cluster assignments for each record, which includes cluster IDs and their respective probabilities ordered by the ID column. The SELECT S.CLUSTER_ID, S.PROBABILITY part of the query selects the cluster ID and probability from the resultant column set. The output is limited to the first 10 rows of the result set.

ONNX Regression Examples

The following examples showcase various JSON metadata parameters that can be defined for ONNX Regression models. All examples assume an ONNX model that has one output named regressionOutput and four input tensors of dimension 1 whose name match exactly the name of the IRIS table columns, namely, SEPAL_LENGTH, SEPAL_WIDTH, PETAL_LENGTH, PETAL_WIDTH.

Example: Specifying JSON Metadata for Regression Models

The following is a simple example illustrating JSON metadata parameters with Regression as the function. Assume the ONNX model features one output named regressionOutput and four input tensors of dimension 1, whose names match exactly after the IRIS table columns ("SEPAL_LENGTH", "SEPAL_WIDTH", "PETAL_LENGTH", "PETAL_WIDTH"). The JSON metadata can be as simple as the following:

```
BEGIN DBMS_VECTOR.LOAD_ONNX_MODEL(
    'regression_model.onnx',
    'doc_model',
    JSON('{"function": "regression"}
    ')
);
END;
/
```

You can use the `PREDICTION` function for inference or scoring:

```
SELECT
    iris.*,
    PREDICTION(doc_model USING *) as predicted_petal_width_cm
FROM iris;
```

In this case, the SQL query selects all columns from the `iris` table and adds a new column, `predicted_petal_width_cm`, which contains predictions made by the `doc_model`. The `USING *` syntax means that all columns of the current row are used as input features for the `doc_model` model to predict the value of `PETAL_WIDTH` as `predicted_petal_width_cm`. The result of this prediction is then included as a new column in the output of the query.

Example: Specifying input and defaultOnNull in JSON Metadata for Regression Models

The following example illustrates how you can specify input attribute names that map to the actual ONNX model input names. The `defaulOnNull` providing default values to be used for specific attributes when their values are NULL in the data set.

```
BEGIN DBMS_VECTOR.LOAD_ONNX_MODEL('regression_model.onnx','doc_model',
    JSON('{"function": "regression",
            "input": {
                "SEPAL_LENGTH": ["dummy_sepal_length_cm"],
                "SEPAL_WIDTH": ["dummy_sepal_width_cm"]
                      },
            "defaultOnNull": {
                "dummy_sepal_length_cm": "5.1",
                "dummy_sepal_width_cm": "3.5",
                 }
           }
         ')
);
END;
/
```

You can use the `PREDICTION` function for inference or scoring.

```
WITH
dummy_iris AS (
    SELECT
    (CASE WHEN petal_length > 5 THEN 4.9 ELSE NULL END)
        as dummy_sepal_length_cm,
    (CASE WHEN petal_length < 4 THEN 2.5 ELSE NULL END)
        as dummy_sepal_width_cm,
    petal_length
    petal_width
    FROM iris
)
SELECT
    dummy_iris.*,
    PREDICTION(doc_model USING *) as predicted_petal_width_cm
FROM dummy_iris;
```

In this case, a temporary `dummy_iris` table is created with three columns: `dummy_sepal_length_cm`, `dummy_sepal_width_cm`, and `petal_length`. The values of the `dummy_sepal_length_cm` and `dummy_sepal_width_cm` are based on `petal_length` values of the `iris` table. If `petal_length` is greater than 5, `dummy_sepal_length_cm` is set to 4.9, otherwise it is NULL. If `petal_length` is less than 4, `dummy_sepal_width_cm` is set to 2.5, otherwise it remains NULL.

Then the `SELECT` query retrieves all columns from the `dummy_iris` table and uses the `doc_model` to predict `petal_width`, adding this prediction as a new column named `predicted_petal_width_cm`. The model uses the derived dummy columns, `petal_length` and `petal_width` for its predictions.

See Also:

- LOAD_ONNX_MODEL in *Oracle Database PL/SQL Packages and Types Reference*
- Supported SQL Scoring Functions

Traditional Machine Learning ONNX Format Models

Traditional machine learning models using algorithms such as decision trees, random forests, and support vector machines, among others, can be converted to ONNX format. Such models may be produced in other environments and deployed through Oracle Database.

Once such models are converted to ONNX format, they can be deployed directly in Oracle Database and use the ONNX Runtime for inference through the SQL prediction operators. These models are typically used for tasks such as Classification, Regression, and Clustering.

Related Topics

- Examples of Using ONNX Models
  The following examples use the Iris data set to showcase loading and inference from ONNX format machine learning models for machine learning techniques such as Classification, Regression, and Clustering in your Oracle Database instance.

Text Transformer ONNX Format Models

Text transformers have the ability to translate natural language text into a numerical vector representation also known as an embedding, you use such vectors for semantic similarity search or other Natural Language Processing (NLP) use cases.

Models such as BERT, sentence transformer models from Hugging Face, and other transformer-based models can be converted into ONNX format models. These models can be run within Oracle Database. These models can be used in AI vector search within Oracle Database, where documents are compared based on their mathematical distance between the vectors to determine the similarity.

Related Topics

- Examples of Using ONNX Models
  The following examples use the Iris data set to showcase loading and inference from ONNX format machine learning models for machine learning techniques such as Classification, Regression, and Clustering in your Oracle Database instance.

Image Transformer ONNX Format Models

Image transformer is a part of machine learning that helps computers interpret and analyze images and videos. It provides tools to perform tasks like creating image embeddings (using an image transformer), classifying objects, detecting anomalies, and identifying objects in pictures or videos.

Image transformers don’t directly use images as input. They need pre-processing to convert images into a form the model can understand. Common pre-processing steps include:

- Decoding images from formats like JPEG to a 3D numeric array.
- Resizing images to standard dimensions.
- Normalizing pixel values.
- Reducing noise in the image.
- Cropping parts of the image for focus.

Image transformer models can be converted into the ONNX format and used directly in Oracle Database. Each image transformer requires its own specific pre-processing pipeline and Oracle offers OML4Py pre-processing pipeline for such models.

Pretrained Image Transformer Models in Oracle Database

Oracle Database supports using pretrained image transformer models for generating vectors for semantic similarity search.

You can access image transformer models through machine learning platforms like Hugging Face that provide pretrained models for immediate use.

To use pretrained image transformer models in Oracle Database, here are the high-level steps:

- Download pretrained models: Download image transformer models into the database.
- Convert image transformer model to ONNX format: Use ONNX pipeline to convert the pretrained image transformer model to ONNX format. Add image pre-processing by implementing Oracle's custom ONNX operation for image decoding and create a model-specific ONNX pre-processing pipeline. See Import Pretrained Models in ONNX Format for Vector Generation Within the Database for more details.
- Import ONNX format image transformer model: Use the `DBMS_VECTOR.LOAD_ONNX_MODEL` procedure or `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` to import the ONNX model into your Oracle database. After importing, use the `VECTOR_EMBEDDING` operator to generate vector embeddings from JPEG images stored as BLOB in the database.

> **Note:**
>
> Only JPEG images are supported. Multiple ONNX models may have to be loaded for multi-modal model because each modality has a different pre-processing and post-processing pipeline.

The Oracle database supports popular pretrained models such as:

- ResNet-50: A widely used model for image classification.
- CLIP ViT-Base-Patch32: A multi-modal model for linking text and image content.

- ViT Base-Patch: A vision transformer model designed for image analysis and classification.

Example: Generate Embeddings from Image Transformer Models

The following examples illustrate generating embeddings from images with image transformer model using `DBMS_VECTOR` or `DBMS_DATA_MINING` packages and use the ONNX Runtime for inference through the SQL prediction operators.

These examples assume that:

- the data set is available to the user.
- the `DM_DUMP` directory exists and contains the ONNX model file for image transformer models augmented with image pre-processing. Follow the steps in ONNX Pipeline Models: Image Embedding and ONNX Pipeline Models: Multi-modal Embedding to generate the ONNX files for the ResNet-50 and Clip ViT models. See also Import ONNX Models into Oracle Database End-to-End Example.

Load File Contents into a BLOB
The following example loads the contents of a file stored in a directory object (`DM_DUMP`) into a BLOB in the database. The function returns the `BLOB` containing the file content.

```
create or replace
    function loader(p_filename varchar2) return blob is
      bf bfile := bfilename('DM_DUMP',p_filename);
      b blob;
    begin
      dbms_lob.createtemporary(b,true);
      dbms_lob.fileopen(bf, dbms_lob.file_readonly);
      dbms_lob.loadfromfile(b,bf,dbms_lob.getlength(bf));
      dbms_lob.fileclose(bf);
     return b;
   end;
   /
```

Create image_data Table
The following example creates the `image_data` table assuming image files are under the `DM_DUMP` directory. The `image_data` table is used further for generating vector embeddings.

```
SQL> CREATE TABLE image_data (
      ID NUMBER,
      NAME VARCHAR2(20),
      IMAGE BLOB
    );

Table created.


SQL> insert into image_data values (1,'cat.jpg',loader('cat.jpg'));

1 row created.

SQL> insert into image_data values (2,'cat2.jpg',loader('cat2.jpg'));

1 row created.
```

```
SQL> insert into image_data values (3,'chicken.jpg',loader('chicken.jpg'));

1 row created.

SQL> insert into image_data values (4,'horse.jpg',loader('horse.jpg'));

1 row created.

SQL> insert into image_data values (5,'dog.jpg',loader('dog.jpg'));

1 row created.

SQL> insert into image_data values (6,'cat.png',loader('cat.png'));

1 row created.

SQL> commit;

Commit complete.
```

Load a ResNet-50 Computer Image Transformer and Generate Vector Embeddings

The following example demonstrates loading an image tranformer model extended with image pre-processing pipeline and using it to generate vector embeddings from images stored in a BLOB. The example assumes that the `DM_DUMP` directory exists and contains the ONNX file for ResNet-50 model augmented with ONNX-based image pre-processing pipeline.

The example imports a pretrained ONNX-format transformer model (`pp_resnet_50.onnx`) into the database as `ppresnet50` using the `DBMS_VECTOR.LOAD_ONNX_MODEL` procedure. Alternately, you can load the model using the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL`. After checking the dictionary views, and examining the schema of the `image_data` table, the model runs a query that generates vector embeddings for each image stored in the `image_data` table using the `VECTOR_EMBEDDING` operator. The vector embeddings can be further used for image classification, similarity search, or feature extraction. The query returns the first 40 characters of each vector. For unsupported formats such as `cat.png` (a PNG file), the `VECTOR_EMBEDDING` operator returns a NULL value.

```
-- Metadata for an embedding model
SQL> define ppjsonmd = '{"function" : "embedding"}';

SQL> exec DBMS_VECTOR.LOAD_ONNX_MODEL('DM_DUMP', 'pp_resnet_50.onnx',
'ppresnet50');

PL/SQL procedure successfully completed.


SQL> SELECT mining_function, algorithm, model_size FROM user_mining_models
WHERE model_name = 'PPRESNET50';

MINING_FUNCTION                ALGORITHM            MODEL_SIZE
------------------------------ -------------------- ----------
EMBEDDING                      ONNX                   93979933


SQL> SELECT attribute_name, attribute_type, data_type, vector_info FROM
user_mining_model_attributes WHERE model_name = 'PPRESNET50' ORDER BY 1;
```

```
ATTRIBUTE_NAME              ATTRIBUTE_TYPE       DATA_TYPE  VECTOR_INFO
--------------------    --------------------  ----------
--------------------
DATA                         UNSTRUCTURED         BLOB
ORA$ONNXTARGET               VECTOR               VECTOR
VECTOR(2048,FLOAT32)


SQL> describe image_data

Name
                                          Null?    Type
 ----------------------------------------------------------------- --------
--------------------------------------------

ID
                                                                  NUMBER

NAME
                                                            VARCHAR2(20)

IMAGE
                                                            BLOB


SQL> SELECT name, substr(vector_embedding(ppresnet50 using image as data), 0, 
40) as vec FROM image_data;

NAME                       VEC
-------------------- ----------------------------------------
cat.jpg              [0,3.69947255E-002,1.727576E-002,0,6.437
cat2.jpg             [5.25364205E-002,0,0,2.8940714E-003,0,4.
chicken.jpg          [2.14146048E-001,7.94866239E-004,2.95593
horse.jpg            [1.63398478E-002,0,4.99145657E-001,0,0,1
dog.jpg              [0,0,7.96773005E-004,0,0,0,1.00504747E-0
cat.png

6 rows selected.
```

Alternately, use the `DBMS_DATA_MINING.IMPORT_ONNX_MODEL` procedure to import the `ppresnet50` model into the database and proceed with the rest of the steps as shown in the example. Here, the loader function loads the content of the file or ONNX files into a blob.

```
SQL> exec DBMS_DATA_MINING.IMPORT_ONNX_MODEL('ppresnet50', 
loader('pp_resnet_50.onnx'), JSON('&ppjsonmd'));

PL/SQL procedure successfully completed.
```

Load a CLIP ViT Model to Generate Vector Embeddings from Images (Image Modality) and Search Images by Generating Embedding from Text Description (Text Modality)
The following example uses CLIP ViT Base patch model (`ppclip`) to check pre-configured ONNX-based image embedding pipeline and generates vector embeddings. The example assumes that the `DM_DUMP` directory exists and contains the ONNX files for each modality of the CLIP ViT Base patch model. The `pp_clip_img.onnx` holds the model augmented with ONNX-

based image pre-processing and post-processing pipelines needed for image modality. The `pp_clip_txt.onnx` holds the model augmented with ONNX-based pre-processing and post-processing pipelines for text modality. Follows the steps in ONNX Pipeline Models: Multi-modal Embedding to get the ONNX files for each of the modality of the CLIP ViT Base patch model.

```
SQL> set echo on
SQL> -- Import clip model with image preprocessing (image modality)
SQL> exec DBMS_VECTOR.LOAD_ONNX_MODEL('DM_DUMP', 'pp_clip_img.onnx',
'clipimg');

PL/SQL procedure successfully completed.

SQL> -- Import clip model with text preprocessing (text modality)
SQL> exec DBMS_VECTOR.LOAD_ONNX_MODEL('DM_DUMP', 'pp_clip_txt.onnx',
'cliptxt');

PL/SQL procedure successfully completed.

SQL> -- Show difference between the two modality:
SQL> SELECT model_name, attribute_name, attribute_type, data_type,
vector_info FROM user_mining_model_attributes WHERE model_name LIKE 'CLIP%'
ORDER BY 1,2;


MODEL_NAME     ATTRIBUTE_NAME      ATTRIBUTE_TY        DATA_TYPE
VECTOR_INFO
----------     ---------------- ------------       ----------------
--------------------
CLIPIMG     DATA             UNSTRUCTURED           BLOB
CLIPIMG     ORA$ONNXTARGET         VECTOR                 VECTOR
VECTOR(512,FLOAT32)
CLIPTXT     DATA             TEXT            VARCHAR2
CLIPTXT     ORA$ONNXTARGET         VECTOR                 VECTOR
VECTOR(512,FLOAT32)


SQL> -- Create a table with vectors generated from image using clip
SQL> CREATE TABLE image_vectors as select name, vector_embedding(clipimg
using image as data) as embedding FROM image_data;

Table created.

SQL> -- Find top-3 similar image from text description
SQL> select name from image_vectors order by
vector_distance(vector_embedding(cliptxt using 'Cat picture' as data),
embedding) fetch first 2 rows only;

NAME
--------------------
cat.jpg
cat2.jpg
```

Alternately, use the DBMS_DATA_MINING.IMPORT_ONNX_MODEL procedure to load the clipimg and cliptxt models into the database.

```
-- Import CLIP model with image preprocessing (image modality)
SQL> exec DBMS_DATA_MINING.IMPORT_ONNX_MODEL('clipimg',
loader('pp_clip_img.onnx'), JSON('{"function" : "embedding"}'));

PL/SQL procedure successfully completed.

-- Import CLIP model with text preprocessing (text modality)
SQL> exec DBMS_DATA_MINING.IMPORT_ONNX_MODEL('cliptxt',
loader('pp_clip_txt.onnx'), JSON('{"function" : "embedding"}'));

PL/SQL procedure successfully completed.
```

Machine Learning and Statistics

Machine learning uses algorithms with fewer assumptions about data than traditional statistics, enabling automation and robust model creation with minimal user intervention.

There is a great deal of overlap between machine learning and statistics. In fact most of the techniques used in machine learning can be placed in a statistical framework. However, machine learning techniques are not the same as traditional statistical techniques.

Statistical models usually make strong assumptions about the data and, based on those assumptions, they make strong statements about the results. However, if the assumptions are flawed, the validity of the model becomes questionable. By contrast, the machine learning methods typically make weak assumptions about the data. As a result, machine learning cannot generally make such strong statements about the results. Yet machine learning can produce very good results regardless of the data.

Traditional statistical methods, in general, require a great deal of user interaction in order to validate the correctness of a model. As a result, statistical methods can be difficult to automate. Statistical methods rely on testing hypotheses or finding correlations based on smaller, representative samples of a larger population.

Less user interaction and less knowledge of the data is required for machine learning. The user does not need to massage the data to guarantee that a method is valid for a given data set. Oracle Machine Learning techniques are easier to automate than traditional statistical techniques.

Overview of Database Analytics

Oracle Database supports native analytical features. Since all these features are on a common server, they can be combined efficiently. Analytical results can be integrated with Oracle Business Intelligence Suite Enterprise Edition and other BI tools.

The possibilities for combining different analytics are virtually limitless. Example 7-5 shows Oracle Machine Learning for SQL and text processing within a single SQL query. The query selects all customers who have a high propensity to attrite (> 80% chance), are valuable customers (customer value rating > 90), and have had a recent conversation with customer services regarding a Checking Plus account. The propensity to attrite information is computed using a OML4SQL model called tree_model. The query uses the Oracle Text CONTAINS operator to search call center notes for references to Checking Plus accounts.

The following table shows some of the built-in analytics that Oracle Database can do:

Table 7-15 Oracle Database Native Analytics

Analytical Feature	Description	Documented In...
Complex data transformations	Data transformation is a key aspect of analytical applications and ETL (extract, transform, and load). You can use SQL expressions to implement data transformations, or you can use the `DBMS_DATA_MINING_TRANSFORM` package. `DBMS_DATA_MINING_TRANSFORM` is a flexible data transformation package that includes a variety of missing value and outlier treatments, as well as binning and normalization capabilities.	*Oracle Database PL/SQL Packages and Types Reference*
Statistical functions	Oracle Database provides a long list of SQL statistical functions with support for: hypothesis testing (such as t-test, F-test), correlation computation (such as pearson correlation), cross-tab statistics, and descriptive statistics (such as median and mode). The `DBMS_STAT_FUNCS` package adds distribution fitting procedures and a summary procedure that returns descriptive statistics for a column.	*Oracle Database SQL Language Reference* and *Oracle Database PL/SQL Packages and Types Reference*
Window and analytic SQL functions	Oracle Database supports analytic and windowing functions for computing cumulative, moving, and centered aggregates. With windowing aggregate functions, you can calculate moving and cumulative versions of `SUM`, `AVERAGE`, `COUNT`, `MAX`, `MIN`, and many more functions.	*Oracle Database Data Warehousing Guide*
Linear algebra	The `UTL_NLA` package exposes a subset of the popular `BLAS` and `LAPACK` (Version 3.0) libraries for operations on vectors and matrices represented as VARRAYs. This package includes procedures to solve systems of linear equations, invert matrices, and compute eigenvalues and eigenvectors.	*Oracle Database PL/SQL Packages and Types Reference*
Analytic views	Analytic views organize data using a dimensional model. They enable you to easily add aggregations and calculations to data sets and to present data in views that can be queried with relatively simple SQL.	*Oracle Database Data Warehousing Guide*
Spatial analytics	Oracle Spatial provides advanced spatial features to support high-end GIS and LBS solutions. Oracle Spatial's analysis and machine learning capabilities include functions for binning, detection of regional patterns, spatial correlation, colocation machine learning, and spatial clustering. Oracle Spatial also includes support for topology and network data models and analytics. The topology data model of Oracle Spatial allows one to work with data about nodes, edges, and faces in a topology. It includes network analysis functions for computing shortest path, minimum cost spanning tree, nearest-neighbors analysis, traveling salesman problem, among others.	*Oracle Spatial Developer's Guide*
Graph	The Property Graph delivers advanced graph query and analytics capabilities in Oracle Database. The in-memory graph server (PGX) provides a machine learning library, which supports graph-empowered machine learning algorithms. The machine learning library supports DeepWalk, supervised GraphWise, and Pg2vec algorithms.	*Oracle Database Graph Developer's Guide for Property Graph*
Text Analysis	Oracle Text uses standard SQL to index, search, and analyze text and documents stored in the Oracle database, in files, and on the web. Oracle Text also supports automatic classification and clustering of document collections. Many of the analytical features of Oracle Text are layered on top of Oracle Machine Learning functionality.	*Oracle Text Application Developer's Guide*

Example 7-5   SQL Query Combining Oracle Machine Learning for SQL and Oracle Text

```
SELECT A.cust_name, A.contact_info
  FROM customers A
 WHERE PREDICTION_PROBABILITY(tree_model,
            'attrite' USING A.*) > 0.8
   AND A.cust_value > 90
   AND A.cust_id IN
       (SELECT B.cust_id
          FROM call_center B
         WHERE B.call_date BETWEEN '01-Jan-2005'
                               AND '30-Jun-2005'
         AND CONTAINS(B.notes, 'Checking Plus', 1) > 0);
```

Oracle Machine Learning and Analytic Views

Analytic views and Oracle Machine Learning complement each other by combining dimensional data organization with inductive inference for comprehensive data analysis.

Analytic views organize data using a dimensional model. Analytic views provide a fast and efficient way to create analytic queries of data stored in existing database tables and views. Analytic Views and Oracle Machine Learning are complementary activities.

An analytic view includes navigation, join, aggregation, and calculation rules, thus eliminating the need to include these rules in queries. However, analytic views do not have inductive inference capabilities. Inductive inference, the process of reaching a general conclusion from specific examples, is a characteristic of machine learning. Inductive inference is also known as computational learning.

Analytic views provide a multidimensional view of the data, including support for hierarchies, and analytic view objects. From a business perspective, analytic views offer a way to present the data.

Oracle Machine Learning and analytic views can be used together in a number of ways. Analytic views can be used to analyze machine learning results at different levels of granularity. Machine learning can help you construct more interesting and useful analytic view. For example, the results of predictive machine learning can be added as custom measures to an analytic view or to suggest important attributes. Such measures can provide information such as "likely to default" or "likely to buy" for each customer. Analytic views can then aggregate and summarize the probabilities.

Oracle Machine Learning and Data Warehousing

Data warehousing supports machine learning by facilitating data cleansing and preparation, ensuring the data is suitable for solving specific problems.

Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved.

Proper data cleansing and preparation are very important for machine learning, and a data warehouse can facilitate these activities. However, a data warehouse is of no use if it does not contain the data you need to solve your problem.

8 Administer

Upgrade and Downgrade

Explains how to perform administrative tasks related to Oracle Machine Learning for SQL.

Upgrade or Downgrade Oracle Machine Learning for SQL

Upgrade and downgrade Oracle Machine Learning for SQL by following the steps listed.

Pre-Upgrade Steps

Pre-upgrade considerations.

Before upgrading, you must drop any machine learning models and machine learning activities that were created inOracle Data Miner.

Upgrade Oracle Machine Learning for SQL

You can upgrade your database by using the Database Upgrade Assistant (DBUA) or you can perform a manual upgrade using export/import utilities.

All models and machine learning metadata are fully integrated with the Oracle Database upgrade process whether you are upgrading from 19*c* or from earlier releases.

Upgraded models continue to work as they did in prior releases. Both upgraded models and new models that you create in the upgraded environment can make use of the new machine learning functionality introduced in the new release.

Related Topics

- Pre-Upgrade Steps
  Pre-upgrade considerations.
- *Oracle Database Upgrade Guide*

Use Database Upgrade Assistant to Upgrade Oracle Machine Learning for SQL

Oracle Database Upgrade Assistant provides a graphical user interface that guides you interactively through the upgrade process.

On Windows platforms, follow these steps to start the Upgrade Assistant:

1. Go to the Windows **Start** menu and choose the Oracle home directory.
2. Choose the **Configuration and Migration Tools** menu.
3. Launch the **Upgrade Assistant**.

On Linux platforms, run the `DBUA` utility to upgrade Oracle Database.

Related Topics

- *Oracle Database Upgrade Guide*

Export/Import Oracle Machine Learning for SQL Models

Use the export and import functions of the Oracle Database to export the previously created models and import the models in an instance of Oracle Database version.

If required, you can use a less automated approach to upgrading machine learning models. You can export the models created in a previous version of Oracle Database and import them into an instance of the Oracle Database version.

To export models from an instance of a previous release of Oracle Database to a dump file, follow the instructions in Export and Import Oracle Machine Learning for SQL Models.

Post Upgrade Steps

Perform steps to view the upgraded database.

After upgrading the database, check the `DBA_MINING_MODELS` view in the upgraded database. The newly upgraded machine learning models must be listed in this view.

After you have verified the upgrade and confirmed that there is no need to downgrade, you must set the initialization parameter `COMPATIBLE` to `23.0.0`. In Oracle Database 23ai, when the `COMPATIBLE` initialization parameter is not set in your parameter file, the `COMPATIBLE` parameter value defaults to `23.0.0`.

> **Note:**
>
> The `CREATE MINING MODEL` privilege must be granted to Oracle Machine Learning for SQL user accounts that are used to create machine learning models.

Related Topics

- Create an Oracle Machine Learning for SQL User
  An OML4SQL user is a database user account that has privileges for performing machine learning activities.
- Control Access to Oracle Machine Learning for SQL Models and Data
  You can create a Oracle Machine Learning for SQL user and grant necessary privileges by following the steps listed.

Downgrade Oracle Machine Learning for SQL

Before downgrading the Oracle database back to the previous version, ensure that no models are present.

Use the DBMS_DATA_MINING.DROP_MODEL routine to drop the models before downgrading. If you do not do this, the database downgrade process terminates.

Issue the following SQL statement in SYS to verify the downgrade:

```
SELECT o.name FROM sys.model$ m, sys.obj$ o
                  WHERE m.obj#=o.obj# AND m.version=2;
```

Export and Import Oracle Machine Learning for SQL Models

You can export machine learning models to move models to a different Oracle Database instance, such as from a development database to a production database.

The DBMS_DATA_MINING package includes procedures for migrating machine learning models between database instances.

EXPORT_MODEL exports a single model or list of models to a dump file so it can be imported, queried, and scored in a separate Oracle Machine Learning database instance.

IMPORT_MODEL takes the dump file and creates the model in the destination database.

EXPORT_SERMODEL exports a single model to a serialized BLOB so it can be imported and scored in a separate Oracle Machine Learning database instance or to OML Services.

IMPORT_SERMODEL takes the serialized BLOB and creates the model in the destination database.

Related Topics

- EXPORT_MODEL
- IMPORT_MODEL
- EXPORT_SERMODEL
- IMPORT_SERMODEL

About Oracle Data Pump

Use the command-line clients of Oracle Data Pump to export and import schemas or databases.

Oracle Data Pump consists of two command-line clients and two PL/SQL packages. The command-line clients, expdp and impdp, provide an easy-to-use interface to the Data Pump export and import utilities. You can use expdp and impdp to export and import entire schemas or databases respectively.

The Data Pump export utility writes the schema objects, including the tables and metadata that constitute machine learning models, to a dump file set. The Data Pump import utility retrieves the schema objects, including the model tables and metadata, from the dump file set and restores them in the target database.

expdp and impdp cannot be used to export/import individual machine learning models.

See Also:

Oracle Database Utilities for information about Oracle Data Pump and the expdp and impdp utilities

About Exporting Models

As a result of building models, each model has a set of model detail views that provide information about the model, such as model statistics for evaluation. The user can query these model detail views. With serialized models, only the model data and metadata required for scoring are available in the serialized model. This is more compact and transfers faster to the destination environment than dump files produced by the EXPORT_MODEL procedure.

To retain complete model details, use the DMBS_DATA_MINING.EXPORT_MODEL procedure and the DBMS_DATA_MINING.IMPORT_MODEL procedure. Serialized model export only works with models that produce scores. Specifically, it doesn't support Attribute Importance, Association Rules, Exponential Smoothing, or O-Cluster (although O-Cluster does allow scoring). Use EXPORT_MODEL to export these models and scenarios when full model details are needed.

Related Topics

- EXPORT_MODEL Procedure
- IMPORT_MODEL Procedure

Options for Exporting and Importing Oracle Machine Learning for SQL Models

Lists options for exporting and importing machine learning models.

Options for exporting and importing machine learning models are described in the following table.

Table 8-1 Export and Import Options for Oracle Machine Learning for SQL

Task	Description
Export or import a full database	(DBA only) Use expdp to export a full database and impdp to import a full database. All machine learning models in the database are included.
Export or import a schema	Use expdp to export a schema and impdp to import a schema. All machine learning models in the schema are included.
Export or import models within a database or between databases	Use DBMS_DATA_MINING.EXPORT_MODEL to export one or more models and DBMS_DATA_MINING.IMPORT_MODEL to import one or more models. These procedures can export and import a single machine learning model, all machine learning models, or machine learning models that match specific criteria. To import models, you must have the CREATE TABLE, CREATE VIEW, and CREATE MINING MODEL privileges.
Export or import individual models to or from a remote database	Use a database link to export individual models to a remote database or import individual models from a remote database. A database link is a schema object in one database that enables access to objects in a different database. The link must be created before you run EXPORT_MODEL or IMPORT_MODEL. To create a private database link, you must have the CREATE DATABASE LINK system privilege. To create a public database link, you must have the CREATE PUBLIC DATABASE LINK system privilege. Also, you must have the CREATE SESSION system privilege on the remote Oracle Database. Oracle Net must be installed on both the local and remote Oracle Databases.

Table 8-1 (Cont.) Export and Import Options for Oracle Machine Learning for SQL

Task	Description
Serialized model export and import	Starting from Oracle Database 18c, the serialized model format was introduced as a lightweight approach to support scoring. The `DBMS_DATA_MINING.EXPORT_SERMODEL` procedure exports a single model to a serialized `BLOB` so it can be imported and scored in a separate Oracle Machine Learning (OML) database instance or to OML Services. `DBMS_DATA_MINING.IMPORT_SERMODEL` takes the serialized `BLOB` and creates the model in the target database.

Related Topics

- IMPORT_MODEL Procedure
- EXPORT_MODEL Procedure
- *Oracle Database SQL Language Reference*

Directory Objects for EXPORT_MODEL and IMPORT_MODEL

Learn how to use directory objects to identify the location of the dump file set containing the models.

`EXPORT_MODEL` and `IMPORT_MODEL` use a directory object to identify the location of the dump file set. A directory object is a logical name in the database for a physical directory on the host computer.

To export machine learning models, you must have write access to the directory object and to the file system directory that it represents. To import machine learning models, you must have read access to the directory object and to the file system directory. Also, the database itself must have access to file system directory. You must have the `CREATE ANY DIRECTORY` privilege to create directory objects.

The following SQL command creates a directory object named `omldir`. The file system directory that it represents must already exist and have shared read/write access rights granted by the operating system. For example, if the directory path is `/home/omluser`, the command is:

```
CREATE OR REPLACE DIRECTORY omldir AS '/home/omluser';
```

The following SQL command gives user `omluser` both read and write access to `omldir`.

```
GRANT READ,WRITE ON DIRECTORY omldir TO OMLUSER;
```

Related Topics

- *Oracle Database SQL Language Reference*

Use EXPORT_MODEL and IMPORT_MODEL

The examples illustrate various export and import scenarios with `EXPORT_MODEL` and `IMPORT_MODEL`.

The examples use the directory object `OMLDIR` shown in Example 8-1 and two schemas, `DM1` and `DM2`. Both schemas have machine learning privileges. `DM1` has two models. `DM2` has one model.

The `DM1` schema has the following models:

- The EM_SH_CLUS_SAMPLE model: it is created by the oml4sql-clustering-expectation-maximization.sql example.
- The DT_SH_CLAS_SAMPLE model: it is created by the oml4sql-classification-decision-tree.sql example.

The DM2 schema has the SVD_SH_SAMPLE model and is created by the oml4sql-singular-value-decomposition.sql. In the following code, models in DM1 schema are displayed.

```
SELECT owner, model_name, mining_function, algorithm FROM all_mining_models where
OWNER='DM1';
```

The output is as follows:

```
OWNER      MODEL_NAME           MINING_FUNCTION      ALGORITHM
---------- -------------------- --------------------
--------------------------
DM1        EM_SH_CLUS_SAMPLE    CLUSTERING           EXPECTATION_MAXIMIZATION
DM1        DT_SH_CLAS_SAMPLE    CLASSIFICATION       DECISION_TREE
```

Example 8-1 Creating the Directory Object

```
-- connect as system user
CREATE OR REPLACE DIRECTORY OMLDIR AS '/home/oracle';
GRANT READ, WRITE ON DIRECTORY OMLDIR TO DM1;
GRANT READ, WRITE ON DIRECTORY OMLDIR TO DM2;
SELECT * FROM all_directories WHERE directory_name = 'OMLDIR';



OWNER      DIRECTORY_NAME            DIRECTORY_PATH
---------- ------------------------- ----------------------------------------
SYS        OMLDIR                     /home/omluser
```

Example 8-2 Exporting All Models From DM1

```
-- connect as DM1
BEGIN
  dbms_data_mining.export_model (
                   filename =>   'all_DM1',
                   directory =>  'OMLDIR');
END;
/
```

A log file and a dump file are created in /home/omluser, the physical directory associated with OMLDIR. The name of the log file is dm1_exp_11.log. The name of the dump file is all_dm101.dmp.

Example 8-3 Importing the Models Back Into DM1

The models that were exported in Example 8-2 still exist in DM1. Since an import does not overwrite models with the same name, you must drop the models before importing them back into the same schema.

```
BEGIN
  dbms_data_mining.drop_model('EM_SH_CLUS_SAMPLE');
  dbms_data_mining.drop_model('DT_SH_CLAS_SAMPLE');
  dbms_data_mining.import_model(
                   filename => 'all_dm101.dmp',
                   directory => 'OMLDIR');
END;
/
SELECT model_name FROM user_mining_models;


MODEL_NAME
------------------------------
DT_SH_CLAS_SAMPLE
EM_SH_CLUS_SAMPLE
```

Example 8-4 Importing Models Into a Different Schema

In this example, the models that were exported from `DM1` in Example 8-2 are imported into `DM2`. The `DM1` schema uses the `USER1` tablespace; the `DM2` schema uses the `USER2` tablespace.

```
-- CONNECT as sysdba
BEGIN
  dbms_data_mining.import_model (
                   filename => 'all_d101.dmp',
                   directory => 'OMLDIR',
                   schema_remap => 'DM1:DM2',
                   tablespace_remap => 'USER1:USER2');
END;
/
-- CONNECT as DM2
SELECT model_name from user_mining_models;


MODEL_NAME
------------------------------------------------------------------------------
--
SVD_SH_SAMPLE
EM_SH_CLUS_SAMPLE
DT_SH_CLAS_SAMPLE
```

Example 8-5 Exporting Specific Models

You can export a single model, a list of models, or a group of models that share certain characteristics.

```
-- Export the model named dt_sh_clas_sample
EXECUTE dbms_data_mining.export_model (
             filename => 'one_model',
             directory =>'OMLDIR',
             model_filter => 'name in (''DT_SH_CLAS_SAMPLE'')');
-- one_model01.dmp and dm1_exp_37.log are created in /home/omluser

-- Export Decision Tree models
EXECUTE dbms_data_mining.export_model(
             filename => 'algo_models',
             directory => 'OMLDIR',
             model_filter => 'ALGORITHM_NAME IN (''DECISION_TREE'')');
-- algo_model01.dmp and dm1_exp_410.log are created in /home/omluser

-- Export clustering models
```

```
EXECUTE dbms_data_mining.export_model(
             filename =>'func_models',
             directory => 'OMLDIR',
             model_filter => 'FUNCTION_NAME = ''CLUSTERING''');
-- func_model01.dmp and dm1_exp_513.log are created in /home/omluser
```

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

EXPORT and IMPORT Serialized Models

From Oracle Database Release 18c onwards, `EXPORT_SERMODEL` and `IMPORT_SERMODEL` procedures are available to export or import serialized models to or from a database.

The serialized format allows the models to be moved to another database instance or OML Services for scoring. The model is exported to a serialized `BLOB` . The import routine takes the serialized content in the `BLOB` and the name of the model to be created with the content.

Related Topics

- EXPORT_SERMODEL Procedure
- IMPORT_SERMODEL Procedure

Import From PMML

You can import regression models represented in Predictive Model Markup Language (PMML).

PMML is an XML-based standard specified by the Data Mining Group (`https://www.dmg.org`). Applications that are PMML-compliant can deploy PMML-compliant models that were created by any vendor. Oracle Machine Learning for SQL supports the core features of PMML 3.1 for regression models.

You can import regression models represented in PMML. The models must be of type `RegressionModel`, either linear regression or binary logistic regression.

Related Topics

- *Oracle Database PL/SQL Packages and Types Reference*

Control Access to Oracle Machine Learning for SQL Models and Data

You can create a Oracle Machine Learning for SQL user and grant necessary privileges by following the steps listed.

Create an Oracle Machine Learning for SQL User

An OML4SQL user is a database user account that has privileges for performing machine learning activities.

Example 8-6 shows how to create a database user. Example 8-7 shows how to assign machine learning privileges to the user.

Note:

To create a user for the OML4SQL examples, you must run two configuration scripts as described in Install the OML4SQL Examples.

Example 8-6 Creating a Database User in SQL*Plus

1. Log in to SQL*Plus with system privileges.

   ```
   Enter user-name: sys as sysdba
   Enter password: password
   ```

2. To create a user named `oml_user`, type these commands. Specify a password of your choosing.

   ```
   CREATE USER oml_user IDENTIFIED BY password
          DEFAULT TABLESPACE USERS
          TEMPORARY TABLESPACE TEMP
          QUOTA UNLIMITED ON USERS;
   Commit;
   ```

   The `USERS` and `TEMP` tablespaces are included in Oracle Database. `USERS` is used mostly by demo users; it is appropriate for running the examples described in About the OML4SQL Examples. `TEMP` is the temporary tablespace that is shared by most database users.

   Note:

   Tablespaces for OML4SQL users must be assigned according to standard DBA practices, depending on system load and system resources.

3. To log in as `oml_user`, enter the following.

   ```
   CONNECT oml_user
   Enter password: password
   ```

See Also:

Oracle Database SQL Language Reference for the complete syntax of the `CREATE USER` statement

Grant Privileges for Oracle Machine Learning for SQL

The CREATE MINING MODEL is a privilege that you must have to create and perform operations on your model. Some other machine learning privileges can be assigned by issuing GRANT statements.

You must have the CREATE MINING MODEL privilege to create models in your own schema. You can perform any operation on models that you own. This includes applying the model, adding a cost matrix, renaming the model, and dropping the model.

The GRANT statements in the following example assign a set of basic machine learning privileges to the oml_user account. Some of these privileges are not required for all machine learning activities, however it is prudent to grant them all as a group.

Additional system and object privileges are required for enabling or restricting specific machine learning activities.

The following table lists the system privileges required for running the OML4SQL examples.

Table 8-2 System Privileges Granted by dmshgrants.sql to the OML4SQL User

Privilege	Allows the OML4SQL User To
CREATE SESSION	Log in to a database session
CREATE TABLE	Create tables, such as the settings tables for CREATE_MODEL
CREATE VIEW	Create views, such as the views of tables in the SH schema
CREATE MINING MODEL	Create OML4SQL models
EXECUTE ON ctxsys.ctx_ddl	Run procedures in the ctxsys.ctx_ddl PL/SQL package; required for text mining

Example 8-7 Privileges Required for Machine Learning

This example grants the required privileges to the user oml_user.

```
GRANT CREATE SESSION TO oml_user;
GRANT CREATE TABLE TO oml_user;
GRANT CREATE VIEW TO oml_user;
GRANT CREATE MINING MODEL TO oml_user;
GRANT EXECUTE ON CTXSYS.CTX_DDL TO oml_user;
```

READ or SELECT privileges are required for data that is not in your schema. For example, the following statement grants SELECT access to the sh.customers table.

```
GRANT SELECT ON sh.customers TO oml_user;
```

System Privileges for Oracle Machine Learning for SQL

A system privilege confers the right to perform a particular action in the database or to perform an action on a type of schema objects. For example, the privileges to create tablespaces and to delete the rows of any table in a database are system privileges.

You can perform specific operations on machine learning models in other schemas if you have the appropriate system privileges. For example, CREATE ANY MINING MODEL enables you to create models in other schemas. SELECT ANY MINING MODEL enables you to apply models that reside in other schemas. You can add comments to models if you have the COMMENT ANY MINING MODEL privilege.

To grant a system privilege, you must either have been granted the system privilege with the ADMIN OPTION or have been granted the GRANT ANY PRIVILEGE system privilege.

The system privileges listed in the following table control operations on machine learning models.

Table 8-3 System Privileges for Oracle Machine Learning for SQL

System Privilege	Allows you to....
CREATE MINING MODEL	Create machine learning models in your own schema.
CREATE ANY MINING MODEL	Create machine learning models in any schema.
ALTER ANY MINING MODEL	Change the name or cost matrix of any machine learning model in any schema.
DROP ANY MINING MODEL	Drop any machine learning model in any schema.
SELECT ANY MINING MODEL	Apply a machine learning model in any schema, also view model details in any schema.
COMMENT ANY MINING MODEL	Add a comment to any machine learning model in any schema.
AUDIT_ADMIN role	Generate an audit trail for any machine learning model in any schema. (See *Oracle Database Security Guide* for details.)

Example 8-8 Grant System Privileges for Oracle Machine Learning for SQL

The following statements allow oml_user to score data and view model details in any schema as long as SELECT access has been granted to the data. However, oml_user can only create models in the oml_user schema.

```
GRANT CREATE MINING MODEL TO oml_user;
GRANT SELECT ANY MINING MODEL TO oml_user;
```

The following statement revokes the privilege of scoring or viewing model details in other schemas. When this statement is run, oml_user can only perform machine learning activities in the oml_user schema.

```
REVOKE SELECT ANY MINING MODEL FROM oml_user;
```

Related Topics

- Add a Comment to an Oracle Machine Learning for SQL Model
  You can add a comment to an OML4SQL model object using SQL COMMENT statement.
- *Oracle Database Security Guide*

Object Privileges for Oracle Machine Learning for SQL Models

Learn about machine learning object privileges.

An object privilege confers the right to perform a particular action on a specific schema object. For example, the privilege to delete rows from the SH.PRODUCTS table is an example of an object privilege.

You automatically have all object privileges for schema objects in your own schema. You can grant object privilege on objects in your own schema to other users or roles.

The object privileges listed in the following table control operations on specific machine learning models.

Table 8-4 Object Privileges for Oracle Machine Learning for SQL Models

Object Privilege	Allows you to....
`ALTER MINING MODEL`	Change the name or cost matrix of the specified machine learning model object.
`SELECT MINING MODEL`	Apply the specified machine learning model object and view its model details.

Example 8-9 Grant Object Privileges on Oracle Machine Learning for SQL Models

The following statements allow `oml_user` to apply the model `testmodel` to the `sales` table, specifying different cost matrixes with each apply. The user `oml_user` can also rename the model `testmodel`. The `testmodel` model and `sales` table are in the `sh` schema, not in the `oml_user` schema.

```
GRANT SELECT ON MINING MODEL sh.testmodel TO oml_user;
GRANT ALTER ON MINING MODEL sh.testmodel TO oml_user;
GRANT SELECT ON sh.sales TO oml_user;
```

The following statement prevents `oml_user` from renaming or changing the cost matrix of `testmodel`. However, `oml_user` can still apply `testmodel` to the `sales` table.

```
REVOKE ALTER ON MINING MODEL sh.testmodel FROM oml_user;
```

Audit and Add Comments to Oracle Machine Learning for SQL Models

Perform audit of Oracle Machine Learning for SQL model objects through SQL statements.

OML4SQL model objects support SQL `COMMENT` and `AUDIT` statements.

Add a Comment to an Oracle Machine Learning for SQL Model

You can add a comment to an OML4SQL model object using SQL `COMMENT` statement.

Comments can be used to associate descriptive information with a database object. You can associate a comment with a machine learning model using a SQL `COMMENT` statement.

```
COMMENT ON MINING MODEL schema_name.model_name IS string;
```

> **Note:**
>
> To add a comment to a model in another schema, you must have the `COMMENT ANY MINING MODEL` system privilege.

To drop a comment, set it to the empty `''` string.

The following statement adds a comment to the model `DT_SH_CLAS_SAMPLE` in your own schema.

```
COMMENT ON MINING MODEL dt_sh_clas_sample IS
           'Decision Tree model predicts promotion response';
```

You can view the comment by querying the catalog view `USER_MINING_MODELS`.

```
SELECT model_name, mining_function, algorithm, comments FROM user_mining_models;
```

The output is as follows:

```
MODEL_NAME          MINING_FUNCTION    ALGORITHM        COMMENTS
------------------- ------------------ ---------------
----------------------------------------------------
DT_SH_CLAS_SAMPLE   CLASSIFICATION     DECISION_TREE    Decision Tree model
predicts promotion response
```

To drop this comment from the database, issue the following statement:

```
COMMENT ON MINING MODEL dt_sh_clas_sample '';
```

See Also:

- Table 8-3
- *Oracle Database SQL Language Reference* for details about SQL `COMMENT` statements

Audit Oracle Machine Learning for SQL Models

Use Oracle Database auditing system to audit models to track operations on machine learning models.

The Oracle Database auditing system is a powerful, highly configurable tool for tracking operations on schema objects in a production environment. The auditing system can be used to track operations on machine learning models.

Note:

To audit machine learning models, you must have the `AUDIT_ADMIN` role.

Unified auditing is documented in *Oracle Database Security Guide*. However, the full unified auditing system is not enabled by default. Instructions for migrating to unified auditing are provided in *Oracle Database Upgrade Guide*.

See Also:

- "Auditing Oracle Machine Learning for SQL Events" in *Oracle Database Security Guide* for details about auditing machine learning models
- "Monitoring Database Activity with Auditing" in *Oracle Database Security Guide* for a comprehensive discussion of unified auditing in Oracle Database
- "About the Unified Auditing Migration Process for Oracle Database" in *Oracle Database Upgrade Guide* for information about migrating to unified auditing
- *Oracle Database Upgrade Guide*

Glossary

ADP

See Automatic Data Preparation.

aggregation

The process of consolidating data values into a smaller number of values. For example, sales data collected on a daily basis can be totaled to the week level.

algorithm

A sequence of steps for solving a problem. See Oracle Machine Learning for SQL algorithm. The Oracle Machine Learning for SQL API supports the following algorithms: Apriori, Decision Tree, k-Means, MDL, Naive Bayes, GLM, O-Cluster, Support Vector Machines, Expectation Maximization, and Singular Value Decomposition.

algorithm settings

The settings that specify algorithm-specific behavior for model building.

anomaly detection

The detection of outliers or atypical cases. Oracle Machine Learning for SQL implements anomaly detection as one-class SVM.

apply

The machine learning operation that scores data. Scoring is the process of applying a model to new data to predict results.

Apriori

The algorithm that uses frequent itemsets to calculate associations.

association

A machine learning technique that identifies relationships among items.

association rules

A machine learning technique that captures co-occurrence of items among transactions. A typical rule is an implication of the form A -> B, which means that the presence of itemset A implies the presence of itemset B with certain support and confidence. The support of the rule is the ratio of the number of transactions where the itemsets A and B are present to the total number of transactions. The confidence of the rule is the ratio of the number of transactions where the itemsets A and B are present to the number of transactions where itemset A is present. Oracle Machine Learning for SQL uses the Apriori algorithm for association models.

attribute

An attribute is a predictor in a predictive model or an item of descriptive information in a descriptive model. **Data attributes** are the columns of data that are used to build a model. Data attributes undergo transformations so that they can be used as categoricals or numericals by the model. Categoricals and numericals are **model attributes**. See also target.

attribute importance

A machine learning technique that provides a measure of the importance of an attribute and predicts a specified target. The measure of different attributes of a training data table enables users to select the attributes that are found to be most relevant to a machine learning model. A smaller set of attributes results in a faster model build; the resulting model could be more accurate. Oracle Machine Learning for SQL uses the Minimum Description Length to discover important attributes. Sometimes referred to as *feature selection* or *key fields*.

Automatic Data Preparation

machine learning models can be created with Automatic Data Preparation (ADP), which transforms the build data according to the requirements of the algorithm and embeds the transformation instructions in the model. The embedded transformations are executed whenever the model is applied to new data.

bagging

Combine independently trained models on bootstrap samples (bagging is bootstrap aggregating).

binning

See discretization.

build data

Data used to build (train) a model. Also called *training data*.

case

All the data collected about a specific transaction or related set of values. A data set is a collection of cases. Cases are also called *records* or *examples*. In the simplest situation, a case corresponds to a row in a table.

case table

A table or view in single-record case format. All the data for each case is contained in a single row. The case table may include a case ID column that holds a unique identifier for each row. Machine learning data must be presented as a case table.

categorical attribute

An attribute whose values correspond to discrete categories. For example, *state* is a categorical attribute with discrete values (CA, NY, MA). Categorical attributes are either non-ordered (nominal) like state or gender, or ordered (ordinal) such as high, medium, or low temperatures.

centroid

See cluster centroid.

classification

A machine learning technique for predicting categorical target values for new records using a model built from records with known target values. Oracle Machine Learning for SQL supports the following algorithms for classification: Naive Bayes, Decision Tree, Generalized Linear Model, Explicit Semantic Analysis, Random Forest, Support Vector Machine, and XGBoost.

clipping

See trimming.

cluster centroid

The vector that encodes, for each attribute, either the mean (if the attribute is numerical) or the mode (if the attribute is categorical) of the cases in the training data assigned to a cluster. A cluster centroid is often referred to as "the centroid."

clustering

A machine learning technique for finding naturally occurring groupings in data. More precisely, given a set of data points, each having a set of attributes, and a similarity measure among them, clustering is the process of grouping the data points into different clusters such that data points in the same cluster are more similar to one another and data points in different clusters are less similar to one another. Oracle Machine Learning for SQL supports three algorithms for clustering, k-Means, Orthogonal Partitioning Clustering, and Expectation Maximization.

confusion matrix

Measures the correctness of predictions made by a model from a test task. The row indexes of a confusion matrix correspond to *actual values* observed and provided in the test data. The column indexes correspond to *predicted values* produced by applying the model to the test data. For any pair of actual/predicted indexes, the value indicates the number of records classified in that pairing.

When predicted value equals actual value, the model produces correct predictions. All other entries indicate errors.

cost matrix

An *n* by *n* table that defines the cost associated with a prediction versus the actual value. A cost matrix is typically used in classification models, where *n* is the number of distinct values in the target, and the columns and rows are labeled with target values. The rows are the actual values; the columns are the predicted values.

counterexample

Negative instance of a target. Counterexamples are required for classification models, except for one-class Support Vector Machines.

machine learning

Machine learning is the practice of automatically searching large stores of data to discover patterns and trends from experience that go beyond simple analysis. Machine learning uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Machine learning is also known as *Knowledge Discovery in Data* (KDD).

A machine learning model implements a machine learning algorithm to solve a given type of problem for a given set of data.

Oracle Machine Learning for SQL algorithm

A specific technique or procedure for producing an Oracle Machine Learning for SQL model. An algorithm uses a specific data representation and a specific machine learning technique.

The algorithms supported by Oracle Machine Learning for SQL are Naive Bayes, Support Vector Machines, Generalized Linear Model, Decision Tree, and XGBoost for classification; Support Vector Machines , Generalized Linear Model, and XGBoost for regression; k-Means, O-Cluster and Expectation Maximization for clustering; Minimum Description Length for attribute importance; Non-Negative Matrix Factorization and Singular Value Decomposition for feature extraction; Apriori for associations, and one-class Support Vector Machines and Multivariate State Estimation Technique - Sequential Probability Ratio Test for anomaly detection.

machine learning server

The component of Oracle Database that implements the machine learning engine and persistent metadata repository. You must connect to a machine learning server before performing machine learning tasks.

data set

In general, a collection of data. A data set is a collection of cases.

descriptive model

A descriptive model helps in understanding underlying processes or behavior. For example, an association model may describe consumer buying patterns. See also machine learning model.

discretization

Discretization (binning) groups related values together under a single value (or bin). This reduces the number of distinct values in a column. Fewer bins result in models that build faster. Many Oracle Machine Learning for SQL algorithms (for example NB) may benefit from input data that is *discretized* prior to model building, testing, computing lift, and applying (scoring). Different algorithms may require different types of binning. Oracle Machine Learning for SQL supports supervised binning, top N frequency binning for categorical attributes and equi-width binning and quantile binning for numerical attributes.

distance-based (clustering algorithm)

Distance-based algorithms rely on a distance metric (function) to measure the similarity between data points. Data points are assigned to the nearest cluster according to the distance metric used.

Decision Tree

A decision tree is a representation of a classification system or supervised model. The tree is structured as a sequence of questions; the answers to the questions trace a path down the tree to a leaf, which yields the prediction.

Decision trees are a way of representing a series of questions that lead to a class or value. The top node of a decision tree is called the root node; terminal nodes are called leaf nodes. Decision trees are grown through an iterative splitting of data into discrete groups, where the goal is to maximize the distance between groups at each split.

An important characteristic of the Decision Tree models is that they are transparent; that is, there are rules that explain the classification.

See also rule .

equi-width binning

Equi-width binning determines bins for numerical attributes by dividing the range of values into a specified number of bins of equal size.

Expectation Maximization

Expectation Maximization is a probabilistic clustering algorithm that creates a density model of the data. The density model allows for an improved approach to combining data originating in different domains (for example, sales transactions and customer demographics, or structured data and text or other unstructured data).

exponential smoothing

Exponential Smoothing algorithms are widely used for forecasting and can be extended to damped trends and time series.

explode

For a categorical attribute, replace a multi-value categorical column with several binary categorical columns. To explode the attribute, create a new binary column for each distinct value that the attribute takes on. In the new columns, 1 indicates that the value of the attribute takes on the value of the column; 0, that it does not. For example, suppose that a categorical attribute takes on the values {1, 2, 3}. To explode this attribute, create three new columns, `col_1`, `col_2`, and `col_3`. If the attribute takes on the value 1, the value in `col_1` is 1; the values in the other two columns is 0.

feature

A combination of attributes in the data that is of special interest and that captures important characteristics of the data. See feature extraction.

See also text feature.

feature extraction

Creates a new set of features by decomposing the original data. Feature extraction lets you describe the data with a number of features that is usually far smaller than the number of original attributes. See also Non-Negative Matrix Factorization and Singular Value Decomposition.

Generalized Linear Model

A statistical technique for linear modeling. Generalized Linear Model (GLM) models include and extend the class of simple linear models. Oracle Machine Learning for SQL supports logistic regression for GLM classification and linear regression for GLM regression.

GLM

See Generalized Linear Model.

k-Means

A distance-based clustering algorithm that partitions the data into a predetermined number of clusters (provided there are enough distinct cases). Distance-based algorithms rely on a distance metric (function) to measure the similarity between data points. Data points are assigned to the nearest cluster according to the distance metric used. Oracle Machine Learning for SQL provides an enhanced version of *k*-Means.

lift

A measure of how much better prediction results are using a model than could be obtained by chance. For example, suppose that 2% of the customers mailed a catalog make a purchase; suppose also that when you use a model to select catalog recipients, 10% make a purchase. Then the lift for the model is 10/2 or 5. Lift may also be used as a measure to compare different machine learning models. Since lift is computed using a data table with actual outcomes, lift compares how well a model performs with respect to this data on predicted outcomes. Lift indicates how well the model improved the predictions over a random selection given actual results. Lift allows a user to infer how a model performs on new data.

lineage

The sequence of transformations performed on a data set during the data preparation phase of the model build process.

linear regression

The GLM regression algorithm supported by Oracle Machine Learning for SQL.

logistic regression

The GLM classification algorithm supported by Oracle Machine Learning for SQL.

MDL

See Minimum Description Length.

min-max normalization

Normalizes numerical attributes using this transformation:

```
x_new = (x_old-min) / (max-min)
```

Minimum Description Length

Given a sample of data and an effective enumeration of the appropriate alternative theories to explain the data, the best theory is the one that minimizes the sum of

- The length, in bits, of the description of the theory
- The length, in bits, of the data when encoded with the help of the theory

The Minimum Description Length principle is used to select the attributes that most influence target value discrimination in attribute importance.

machine learning technique

A major subdomain of Oracle Machine Learning for SQL that shares common high level characteristics. The Oracle Machine Learning for SQL API supports the following machine learning techniques: classification , regression, attribute importance, feature extraction, clustering, and anomaly detection.

machine learning model

A first-class schema object that specifies a machine learning model in Oracle Database.

missing value

A data value that is missing at random. The value could be missing because it is unavailable, unknown, or because it was lost. Oracle Machine Learning for SQL interprets missing values in columns with simple data types (not nested) as missing at random. Oracle Machine Learning for SQL interprets missing values in nested columns as sparsity.

Machine learning algorithms vary in the way they treat missing values. There are several typical ways to treat them: ignore them, omit any records containing missing values, replace missing values with the mode or mean, or infer missing values from existing values. See also sparse data.

model

A model uses an algorithm to implement a given machine learning technique. A model can be a supervised model or an unsupervised model. A model can be used for direct inspection, for example, to examine the rules produced from an association model, or to score data (predict an outcome). In Oracle Database, machine learning models are implemented as machine learning model schema objects.

multi-record case

Each case in the data table is stored in multiple rows. Also known as transactional data. See also single-record case.

Multivariate State Estimation Technique - Sequential Probability Ratio Test

MSET-SPRT (Multivariate State Estimation Technique - Sequential Probability Ratio Test) is an anomaly detection algorithm in Oracle Machine Learning. This algorithm analyzes historical sensor data to learn a system's normal behavior. Monitors live sensor data streams for deviations from the expected behavior (anomalies). It doesn't require specific assumptions about data distribution. It analyzes data points one by one, improving efficiency. Handles high-dimensional data (many sensors) using random projections.

Naive Bayes

An algorithm for classification that is based on Bayes's theorem. Naive Bayes makes the assumption that each attribute is conditionally independent of the others: given a particular value of the target, the distribution of each predictor is independent of the other predictors.

nested data

Oracle Machine Learning for SQL supports transactional data in nested columns of name/ value pairs. Multidimensional data that expresses a one-to-many relationship can be loaded into a nested column and mined along with single-record case data in a case table.

Neural Network

Neural Network is a machine learning algorithm that mimics the biological human brain neural network to recognize relationships in a data set that depend on large number of unknown inputs.

NMF

See Non-Negative Matrix Factorization.

Non-Negative Matrix Factorization

A feature extraction algorithm that decomposes multivariate data by creating a user-defined number of features, which results in a reduced representation of the original data.

normalization

Normalization consists of transforming numerical values into a specific range, such as [–1.0,1.0] or [0.0,1.0] such that `x_new = (x_old-shift)/scale`. Normalization applies only to numerical attributes. Oracle Machine Learning for SQL provides transformations that perform min-max normalization, scale normalization, and z-score normalization.

numerical attribute

An attribute whose values are numbers. The numeric value can be either an integer or a real number. Numerical attribute values can be manipulated as continuous values. See also categorical attribute.

O-Cluster

See Orthogonal Partitioning Clustering.

one-class Support Vector Machine

The version of Support Vector Machines used to solve anomaly detection problems. The algorithm performs classification without a target.

Orthogonal Partitioning Clustering

An Oracle proprietary clustering algorithm that creates a hierarchical grid-based clustering model, that is, it creates axis-parallel (orthogonal) partitions in the input attribute space. The algorithm operates recursively. The resulting hierarchical structure represents an irregular grid that tessellates the attribute space into clusters.

outlier

A data value that does not come from the typical population of data or extreme values. In a normal distribution, outliers are typically at least three standard deviations from the mean.

partitioned models

Partitioned models enable users to build an ensemble model for each data partition. The top-level model includes sub-models that are automatically generated based on specified attribute options. For example, if your data set has an attribute called `REGION` with four values, defining this as the partitioned attribute will create four sub-models for each region. These sub-models are managed and used as a single model. This approach automates a typical machine learning task and can achieve better accuracy through multiple targeted models.

positive target value

In binary classification problems, you may designate one of the two classes (target values) as positive, the other as negative. When Oracle Machine Learning for SQL computes a model's lift, it calculates the density of positive target values among a set of test instances for which the model predicts positive values with a given degree of confidence.

predictive model

A predictive model is an equation or set of rules that makes it possible to predict an unseen or unmeasured value (the dependent variable or output) from other, known values (independent variables or input). The form of the equation or rules is suggested by machine learning data collected from the process under study. Some training or estimation technique is used to estimate the parameters of the equation or rules. A predictive model is a supervised model.

predictor

An attribute used as input to a supervised algorithm to build a model.

prepared data

Data that is suitable for model building using a specified algorithm. Data preparation often accounts for much of the time spent in a machine learning project. Automatic Data Preparation greatly simplifies model development and deployment by automatically preparing the data for the algorithm.

Principal Component Analysis

Principal Component Analysis is implemented as a special scoring method for the Singular Value Decomposition algorithm.

prior probabilities

The set of prior probabilities specifies the distribution of examples of the various classes in the original source data. Also referred to as *priors*, these could be different from the distribution observed in the data set provided for model build.

priors

See prior probabilities.

quantile binning

A numerical attribute is divided into bins such that each bin contains approximately the same number of cases.

random projections

Random projections refer to a technique used in dimensionality reduction where the original high-dimensional data is projected onto a lower-dimensional subspace. This is achieved using a random matrix, preserving the structure of the data while significantly reducing its complexity. The goal is to approximate the data in a lower-dimensional space while retaining as much of the original information as possible. In the context of OML, random projections are used to create efficient, compact representations of the data for tasks like similarity searches or clustering, without having to manually identify the most important features. This approach is computationally efficient and scalable, making it well-suited for large data sets.

random sample

A sample in which every element of the data set has an equal chance of being selected.

recode

Literally "change or rearrange the code." Recoding can be useful in preparing data according to the requirements of a given business problem, for example:

- Missing values treatment: Missing values may be indicated by something other than `NULL`, such as "0000" or "9999" or "NA" or some other string. One way to treat the missing value is to recode, for example, "0000" to `NULL`. Then the Oracle Machine Learning for SQL algorithms and the database recognize the value as missing.
- Change data type of variable: For example, change "Y" or "Yes" to 1 and "N" or "No" to 0.

- Establish a cutoff value: For example, recode all incomes less than $20,000 to the same value.
- Group items: For example, group individual US states into regions. The "New England region" might consist of ME, VT, NH, MA, CT, and RI; to implement this, recode the five states to, say, NE (for New England).

record

See case.

regression

A machine learning technique for predicting continuous target values for new records using a model built from records with known target values. Oracle Machine Learning for SQL supports linear regression (GLM) and Support Vector Machines algorithms for regression.

rule

An expression of the general form *if X, then Y*. An output of certain algorithms, such as clustering, association, and Decision Tree. The predicate *X* may be a compound predicate.

sample

See random sample.

scale normalization

Normalize numerical attributes using this transformation:

```
x_new = (x_old - 0) / (max(abs(max),abs(min)))
```

schema

A collection of objects in an Oracle database, including logical structures such as tables, views, sequences, stored procedures, synonyms, indexes, clusters, and database links. A schema is associated with a specific database user.

score

Scoring data means applying a machine learning model to data to generate predictions.

settings

See algorithm settings.

single-record case

Each case in the data table is stored in one row. Contrast with multi-record case.

Singular Value Decomposition

A feature extraction algorithm that uses orthogonal linear projections to capture the underlying variance of the data. Singular Value Decomposition scales well to very large data sizes (both rows and attributes), and has a powerful data compression capability.

See Singular Value Decomposition.

sparse data

Data for which only a small fraction of the attributes are non-zero or non-null in any given case. Market basket data and unstructured text data are typically sparse. Oracle Machine Learning for SQL interprets nested data as sparse. See also missing value.

split

Divide a data set into several disjoint subsets. For example, in a classification problem, a data set is often divided in to a training data set and a test data set.

stratified sample

Divide the data set into disjoint subsets (strata) and then take a random sample from each of the subsets. This technique is used when the distribution of target values is skewed greatly. For example, response to a marketing campaign may have a positive target value 1% of the time or less. A stratified sample provides the machine learning algorithms with enough positive examples to learn the factors that differentiate positive from negative target values. See also random sample.

supervised binning

A form of intelligent binning wherein bin boundaries are derived from important characteristics of the data. Supervised binning builds a single-predictor decision tree to find the interesting bin boundaries with respect to a target. Supervised binning can be used for numerical or categorical attributes.

supervised learning

See supervised model.

supervised model

A data mining model that is built using a known dependent variable, also referred to as the target. Classification and regression techniques are examples of supervised mining. See unsupervised model. Also referred to as predictive model.

Support Vector Machine

An algorithm that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector Machine can make predictions with sparse data, that is, in domains that have a large number of predictor columns and relatively few rows, as is the case with bioinformatics data. Support Vector Machine can be used for classification, regression, and anomaly detection.

SVM

See Support Vector Machines.

target

In supervised learning, the identified attribute that is to be predicted. Sometimes called *target value* or *target attribute*. See also attribute.

text feature

A combination of words that captures important attributes of a document or class of documents. Text features are usually keywords, frequencies of words, or other document-derived features. A document typically contains a large number of words and a much smaller number of features.

text analysis

Conventional machine learning done using text features. Text features are usually keywords, frequencies of words, or other document-derived features. Once you derive text features, you mine them just as you would any other data. Both Oracle Machine Learning for SQL and Oracle Text support text analysis.

time series

Time Series is a machine learning function that forecasts target value based solely on a known history of target values. It is a specialized form of Regression, known in the literature as auto-regressive modeling. Time Series supports exponential smoothing.

top N frequency binning

This type of binning bins categorical attributes. The bin definition for each attribute is computed based on the occurrence frequency of values that are computed from the data. The user specifies a particular number of bins, say N. Each of the bins bin_1,..., bin_N corresponds to the values with top frequencies. The bin bin_N+1 corresponds to all remaining values.

training data

See build data.

transactional data

The data for one case is contained in several rows. An example is market basket data, in which a case represents one basket that contains multiple items. Oracle Machine Learning for SQL supports transactional data in nested columns of attribute name/value pairs. See also nested data, multi-record case, and single-record case.

transformation

A function applied to data resulting in a new representation of the data. For example, discretization and normalization are transformations on data.

trimming

A technique for minimizing the impact of outliers. Trimming removes values in the tails of a distribution in the sense that trimmed values are ignored in further computations. Trimming is achieved by setting the tails to `NULL`.

unstructured data

Images, audio, video, geospatial mapping data, and documents or text data are collectively known as unstructured data. Oracle Machine Learning for SQL supports the analysis of unstructured text data.

unsupervised learning

See unsupervised model.

unsupervised model

A machine learning model built without the guidance (supervision) of a known, correct result. In supervised learning, this correct result is provided in the target attribute. Unsupervised learning has no such target attribute. Clustering and association are examples of unsupervised machine learning techniques. See supervised model.

winsorizing

A technique for minimizing the impact of outliers. Winsorizing involves setting the tail values of an particular attribute to some specified value. For example, for a 90% Winsorization, the bottom 5% of values are set equal to the minimum value in the 6th percentile, while the upper 5% are set equal to the maximum value in the 95th percentile.

XGBoost

XGBoost (eXtreme Gradient Boosting) is a scalable machine learning system available within Oracle Machine Learning. It's based on the open-source XGBoost framework and provides functionalities for classification, regression, ranking, and survival analysis tasks.

z-score normalization

Normalize numerical attributes using this transformation:

```
x_new = (x_old-mean) / standard_deviation
```