Oracle® Database

Oracle Machine Learning Overview

Release Latest
G17030-01
April 2025

Oracle Database Oracle Machine Learning Overview, Release Latest

G17030-01

Primary Author: Sarika Surampudi

Contributors: Mark Hornick, Sherry LaMonica

Contents

5 Components

6 Use Cases Using Oracle Machine Learning

Glossary

Preface

Oracle Machine Learning Overview Guide introduces the foundational concepts and capabilities of machine learning within the Oracle ecosystem. This guide provides a high-level view of how Oracle Machine Learning (OML) integrates machine learning into the data management lifecycle.

This document is part of a phased documentation effort. In this first phase, we present key concepts, supported techniques and algorithms, Oracle components that support machine learning, and a representative industry use case to illustrate practical application.

Audience

Oracle Machine Learning Overview guide is curated for readers seeking a broad understanding of what machine learning is, how Oracle supports it, and where it fits in solving real-world business problems.

This guide is intended for a general audience, including:

- Professionals, developers, data engineers, and DBAs new to Oracle Machine Learning
- Business stakeholders and analysts seeking to understand how machine learning can solve industry-specific problems
- Students and educators interested in the foundational concepts and enterprise use of machine learning

The content focuses on building familiarity with the basic concepts, terminology, capabilities, and components of Oracle Machine Learning.

Documentation Accessibility

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc.

Access to Oracle Support

Oracle customer access to and use of Oracle support services will be pursuant to the terms and conditions specified in their Oracle order for the applicable services.

Diversity and Inclusion

Oracle is fully committed to diversity and inclusion. Oracle respects and values having a diverse workforce that increases thought leadership and innovation. As part of our initiative to build a more inclusive culture that positively impacts our employees, customers, and partners, we are working to remove insensitive terms from our products and documentation. We are also mindful of the necessity to maintain compatibility with our customers' existing technologies and the need to ensure continuity of service as Oracle's offerings and industry standards evolve. Because of these technical constraints, our effort to remove insensitive terms is ongoing and will take time and external cooperation.

Conventions

The following text conventions are used in this document:

Convention	Meaning
boldface	Boldface type indicates graphical user interface elements associated with an action, or terms defined in text or the glossary.
italic	Italic type indicates book titles, emphasis, or placeholder variables for which you supply particular values.
`monospace`	Monospace type indicates commands within a paragraph, URLs, code in examples, text that appears on the screen, or text that you enter.

1

Machine Learning Overview

Machine Learning (ML) is a technique of data analysis that lets computers learn from and base decisions on data without direct programming. It uses algorithms to find patterns, get better over time, and automate tasks making it necessary to solve hard data-driven problem

Machine learning helps businesses make faster, smarter decisions by uncovering insights from large data sets. It enables automation, enhances customer experiences, optimizes operations, and supports predictive capabilities, leading to cost savings, efficiency, and competitive advantage across various industries.

- What Is Machine Learning, AI, and Generative AI?
  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.
- Benefits of Machine Learning
  Machine learning is a powerful technology that can help you find patterns and relationships within your data.
- 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.

What Is Machine Learning, AI, and Generative AI?

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.

Machine learning is a technique that discovers previously unknown relationships in data. Some relationships may be known, but the algorithm learns those patterns, example, for inferencing. 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. 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. 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.

To compare machine learning with Generative AI (GenAI), GenAI is specifically designed for producing new content like text, code, or images by generative AI models, which are trained on vast data sets to create original outputs based on patterns they learn; essentially, one is about making predictions based on data, while the other is about creating new data based on learned patterns. Oracle Machine Learning concentrates on traditional machine learning tasks like prediction and classification using established algorithms.

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 plays a pivotal role in credit risk assessments by helping financial institutions predict the likelihood of a borrower defaulting on a loan or credit. By analyzing historical data, machine learning models can identify patterns and relationships between various factors (for example, income, credit history, debt levels, and employment status) that contribute to a borrower's creditworthiness.

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.

Provide 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.

Transform 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 on 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

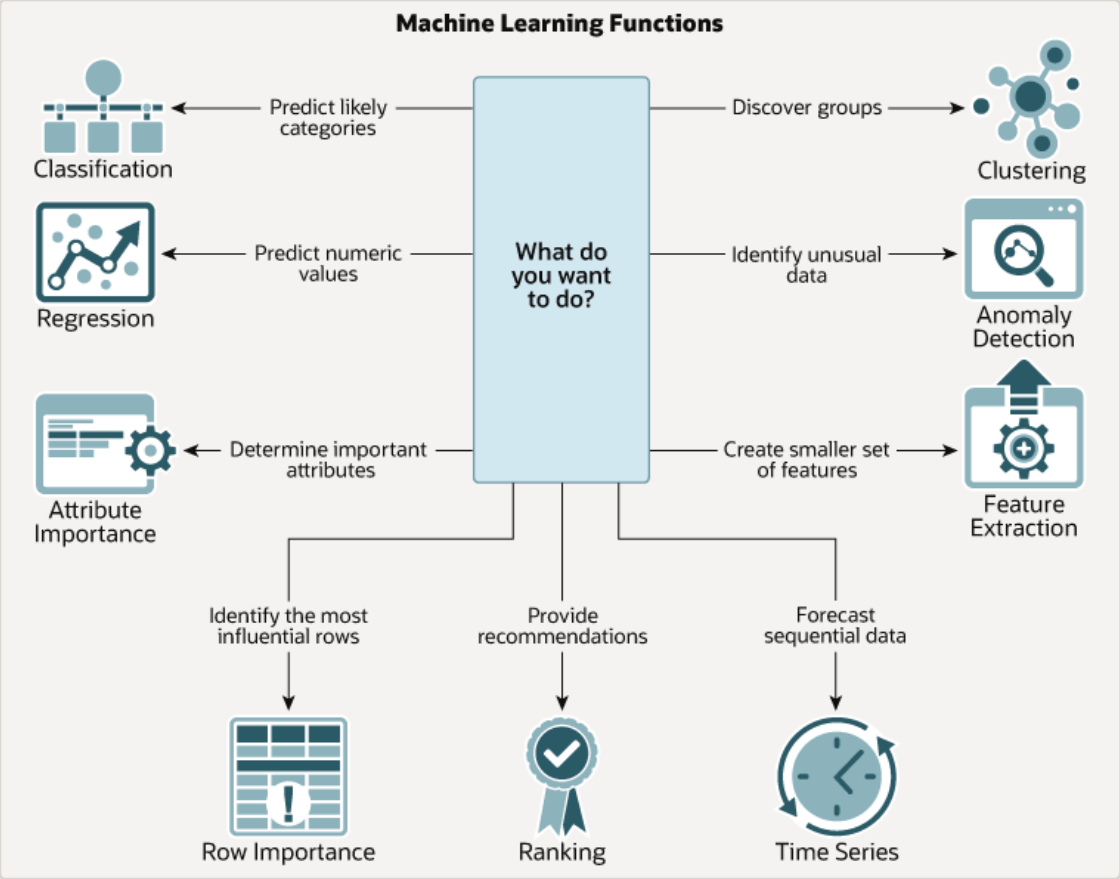
What Do You Want to Do?

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.

2

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.

- Oracle Machine Learning Process
  The machine learning process illustration is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Each stage is illustrated with points that summarize the key tasks. The CRISP-DM methodology is the most a commonly used methodology for machine learning.
- 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.
- Understand Data
  The data understanding phase involves data collection and exploration which includes loading the data and analyzing the data for your business problem.
- 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.
- Develop Models
  In this phase, you select and apply various modeling techniques and tune the algorithm parameters, called *hyperparameters*, to desired values.
- Evaluate
  At this stage of the project, it is time to evaluate how well the model satisfies the originally-stated business goal.
- 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.

Oracle Machine Learning Process

The machine learning process illustration is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Each stage is illustrated with points that summarize the key tasks. The CRISP-DM methodology is the most a commonly used methodology for machine learning.

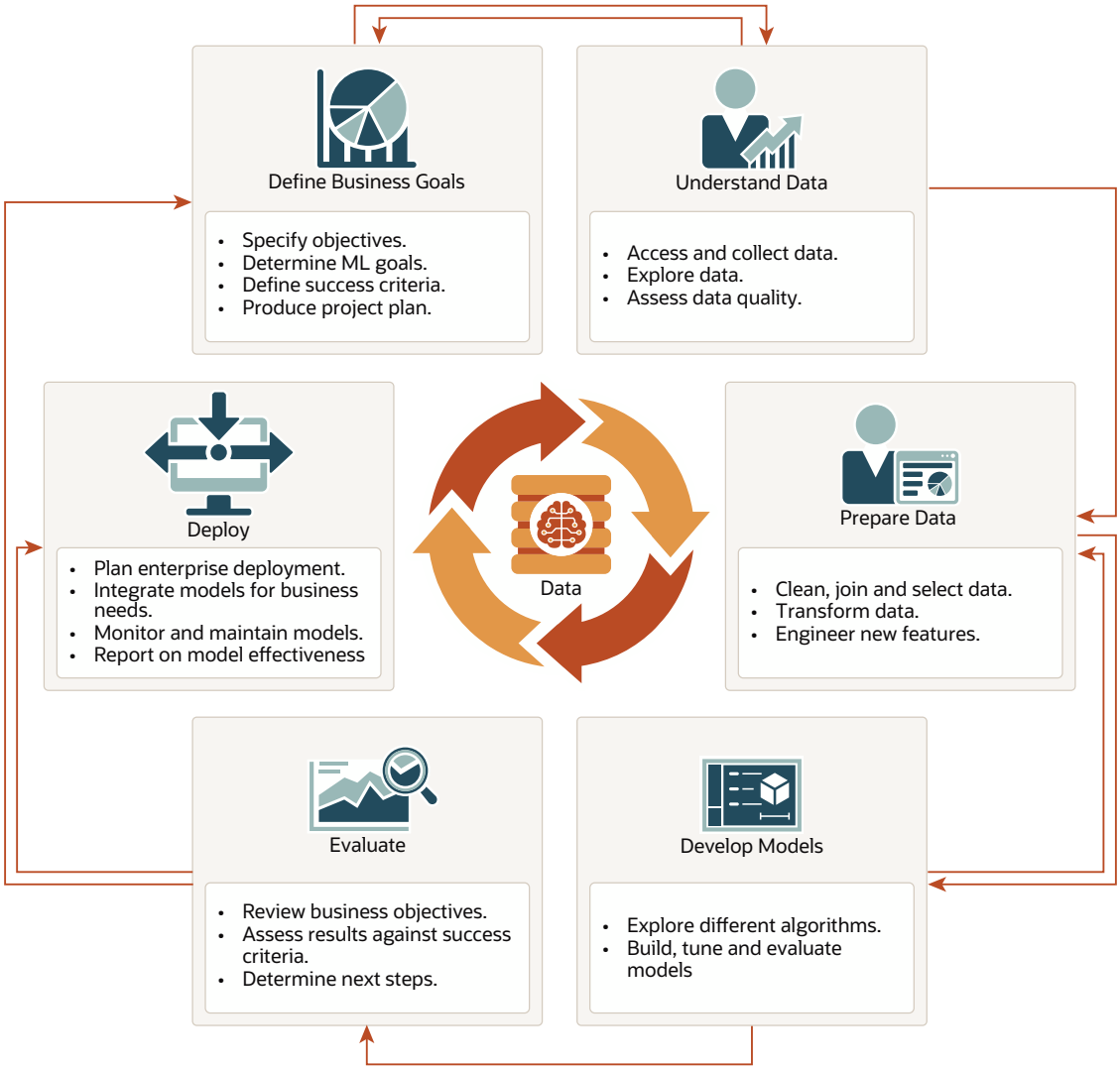
The following are the phases of the machine learning process:

- 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.

Figure 2-1 Machine Learning Process



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. Machine learning can help solve 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. 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.

Another 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.

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.

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*

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*

3

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

Machine Learning Techniques Overview

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 functions 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.

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 techniques:

Table 3-1 Supervised Machine Learning Techniques

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

- 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.

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.

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 techniques:

Table 3-2 Unsupervised Machine Learning Techniques

Technique	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.

4

What is In-Database Machine Learning

Oracle Machine Learning (OML) provides scalable in-database machine learning algorithms through PL/SQL, SQL, Python, and R APIs. OML has over 30 scalable machine learning algorithms directly in the database, which helps you develop and deploy solutions quickly for applications and dashboards.

OML eliminates costly and risky data movement for database data. By avoiding separate analytical engines, you simplify your solution architecture, as there's no need to manage and test workflows involving remote third-party engines. OML algorithms support algorithm-specific automatic data preparation and individual prediction details, with scalable batch and real-time scoring (inferencing). OML is included with Oracle Autonomous Database instances and Oracle Database licenses.

- Overview of In-Database Machine Learning
  OML 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.
- Benefits of In-Database Machine Learning
  Oracle Machine Learning in Oracle Database securely enables data scientists and non-experts to easily build accurate models without moving data, automating data preparation, leveraging no-code interfaces, APIs, and integrated analytics features.
- Features of In-Database Algorithms
  Oracle Machine Learning offers a suite of tools enhancing productivity for data scientists, developers, and data engineers. This suite streamlines machine learning model development, evaluation, and deployment, catering to both experts and non-experts in the field.
- Optimization features of Oracle Exadata and Oracle RAC
  Oracle Exadata and Oracle RAC offer advanced optimization features for machine learning by leveraging distributed parallelism, storage-tier processing, and dynamic resource allocation to enable scalability, high-performance model building, and real-time scoring.

Overview of In-Database Machine Learning

OML 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 OML 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. OML provides explanatory prediction details when scoring data, so you can understand why an individual prediction is made.

OML 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. OML 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, are components of Oracle Machine Learning that provide powerful APIs for in-database machine learning, among other features.

Benefits of In-Database Machine Learning

Oracle Machine Learning in Oracle Database securely enables data scientists and non-experts to easily build accurate models without moving data, automating data preparation, leveraging no-code interfaces, APIs, and integrated analytics features.

OML within Oracle Database offers the following 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 OML, 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, and get access to in-database and third-party models, and R and Python objects and scripts. In-database machine learning models are fully integrated objects within Oracle Database. They are created directly within the database and can be used immediately within the database environment.
- **Data Preparation**: 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. OML can automatically manage key steps in the data preparation process.
- **No-code User Interfaces** : No-code AutoML user interfaces, Data Monitor, Model Monitor, and the model's UI for model deployment, improve data scientist productivity and give non-experts access to powerful in-database classification and regression techniques.
- **Ease of Data Refresh**: Machine learning processes within Oracle Database have ready access to refreshed data. OML can easily deliver machine learning results based on current data, thereby maximizing its timeliness and relevance.
- **Platform for advanced analytics**: Oracle Database provides powerful features for advanced analytics and business intelligence, enabling seamless integration of machine learning with other analytics capabilities, such as statistical analysis, graph processing, spatial analysis, and analytic views: all within the same environment. This converged setup allows for more efficient, in-depth insights without the need to move data across different systems, enhancing performance and simplifying data management.
- **Oracle Technology Stack**: You can take advantage of the broader Oracle technology stack to integrate machine learning within a larger framework for business intelligence or scientific inquiry.
- **Application Programming Interfaces**: The APIs for SQL, R, Python, and REST along with SQL language operators, provide direct access to OML functionality in Oracle Database.

Features of In-Database Algorithms

Oracle Machine Learning offers a suite of tools enhancing productivity for data scientists, developers, and data engineers. This suite streamlines machine learning model development, evaluation, and deployment, catering to both experts and non-experts in the field.

The following summarize the features of In-Database algorithms:

- **In-database Machine Learning**:
  - Perform ML operations directly within Oracle Database without exporting data to separate ML engines. This approach eliminates data movement, ensuring efficiency and data security.
  - Oracle uses parallelized and distributed algorithms, scaling seamlessly across cluster nodes for faster processing.
  - It optimizes memory usage and leverages Exadata's storage-tier function push-down for high-speed scoring.
- **Scalability and Deployment**:
  - Perform batch and real-time predictions using OML's scalable architecture.
  - Use prediction operators in SQL queries or use them directly with programming languages like Python and R.
  - OML on Autonomous Database Serverless supports no-code deployment through REST interfaces, making deployment accessible to users with varying technical skills.
- **Machine Learning Models as First-Class Database Objects**:
  - Manage models with database-level access control, ensuring secure handling.
  - Track user actions through auditing, providing insights into usage and changes.
  - Export and import models between databases for efficient sharing and reuse.
  - Leverage database tools for backup, recovery, and secure storage of ML models.
- **Data Preparation**:

  Automate key steps like cleansing, filtering, normalizing, and sampling. 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.
- **Text Processing**:
  - Extract useful information from unstructured text, transforming it into structured data using machine learning techniques.
  - Text tokens or features allow querying and deriving insights from text data to address business challenges effectively.
- **Partitioned Models**:
  - Divide data into subsets based on characteristics to organize multiple models efficiently.
  - Use partitioning to manage diverse data sets while maintaining clarity and improving model management.
- **Faster Time-to-Market Solutions:**
  - Deploy ML models instantly with SQL prediction operators and REST interfaces.

  - Run predictions directly from R or Python environments without additional tools.
  - Simplify deployment on Autonomous Database Serverless to deliver actionable insights quickly, streamlining workflows for both experts and non-experts.

Topics:

Automatic Data Preparation

Machine learning models often require data transformations before training. Oracle Machine Learning (OML) automates this process using Automatic Data Preparation (ADP). ADP applies to OML4SQL, OML4Py, and OML4R in-database models, making data transformation easier.

When ADP is enabled, Oracle Machine Learning applies transformations based on the algorithm's needs. These transformations include:

- **Binning:** Grouping numerical values into ranges.
- **Normalization:** Scaling values to a common range.
- **Handling missing or sparse data:** Managing gaps in data sets.

ADP embeds these transformations in the model along with any user-specified transformation instruction, ensuring they are applied whenever new data is processed. Oracle Machine Learning follows consistent heuristics to determine the best transformations for an algorithm. This approach helps achieve reasonable model quality in most cases.

You can:

- Use automatic transformations provided by ADP.
- Define custom transformations to fit your data needs.
- Manually handle transformations using database functions.

You can customize data preparation for:

- OML4SQL: Use the `DBMS_DATA_MINING_TRANSFORM` PL/SQL package.
- OML4Py: Specify transformations using a model settings list (`params`).
- OML4R: Use the `odm.settings` list or enable ADP directly (`auto.data.prep=TRUE`).

OML 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 management of missing values and sparse data: Oracle Machine Learning uses consistent methodology across machine learning algorithms to handle sparsity and missing values.
- Transparency: Oracle Machine Learning 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 provides many common transformation routines, for example, in OML4SQL, the `DBMS_DATA_MINING_TRANSFORM` PL/SQL package. You can use these routines, or develop your own routines in SQL, or perform both. You can use custom transformation instructions instead of ADP or use it with ADP.

Integrated Text Mining

Integrated text mining in OML allows you to perform text analysis directly within the Oracle Database using SQL and PL/SQL. This integration enables you to extract meaningful insights from unstructured text data without the need to move data outside the database environment.

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. Oracle interprets columns of `VARCHAR2` (>4000), and `CLOB` as text. You can also identify columns of `CHAR`, `VARCHAR2` (<=4000), `BFILE`, and `BLOB` as text attributes (unstructured 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. These are transformed into a format the algorithms can analyze. 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.

Key features include:

- In-Database processing: Perform text mining operations within the database, leveraging Oracle's scalability and performance.
- Text preprocessing functions: Includes functions to clean and tokenize text data, converting it into a structured format suitable for analysis.
- Feature Extraction: Convert unstructured text into structured numerical data suitable for machine learning algorithms.
- Machine learning algorithms: Apply algorithms such as classification, clustering, and anomaly detection to text data.
- SQL and PL/SQL integration: Text mining tasks can be run using SQL and PL/SQL procedures, allowing seamless integration with existing data and workflows.

About Partitioned Models

Partitioned models allow you to divide your data set into multiple partitions based on specific attributes and build a model for each partition. The system automates the creation and management of these models, reducing manual effort.

When you build a model on a data set, a single generic model may not perform well on new or evolving data. To address this, you can specify partition columns to create separate models for each partition based on some characteristics. For example, if your data set includes a "REGION" attribute with four values, four models will be created automatically, each corresponding to a region. The system manages these sub-models as part of a single partitioned model.

The partitioned model resides as a first-class, persistent database object, with all sub-models stored on disk. This structure improves performance, especially for models with a large number of partitions, and allows for efficient management, such as dropping individual sub-models when needed.

When building or scoring partitioned models, not all sub-models need to be loaded into memory simultaneously. This approach optimizes memory usage and enhances processing efficiency. The system provides a single model for scoring, while users can still access individual component models if needed.

Related Topics

- *Oracle Database Reference*

> **See Also:**
>
> *Oracle Machine Learning for SQL User's Guide* to understand more about partitioned models.

Optimization features of Oracle Exadata and Oracle RAC

Oracle Exadata and Oracle RAC offer advanced optimization features for machine learning by leveraging distributed parallelism, storage-tier processing, and dynamic resource allocation to enable scalability, high-performance model building, and real-time scoring.

Distributed Parallelism and Scalability:

- Oracle RAC enables distributed parallelism across cluster nodes, improving efficiency for data processing and machine learning tasks.
- Exadata's architecture supports scalable performance through in-memory processing and autoscaling, ensuring consistent performance even during peak workloads.

Storage-Tier Processing with Smart Scan:

Exadata Smart Scan technology processes SQL predicates and machine learning model scoring directly at the storage tier. This reduces data movement and speeds up query execution by 2-5 times compared to non-Smart Scan in-database scoring.

Data Loading and Model Caching:

- Oracle Machine Learning (OML) loads data incrementally into memory, eliminating the need for the entire data set to fit in memory.
- Models are efficiently cached and shared across queries, minimizing memory overhead and improving multi-user performance.

High-Performance Scoring:

- Machine learning models are integrated as SQL functions, enabling high-performance scoring in both batch and online transactional processing (OLTP) environments.
- OML leverages Exadata's storage-tier optimization for real-time scoring using current table data.

Autoscaling on Autonomous Database:

The Autonomous Database dynamically adjusts computing power to handle multiple users and simultaneous queries.

Multi-User Support:

- Disk-Aware Structures optimize memory allocation using the database memory manager, thereby efficient performance in multi-user environments.
- For partitioned models, only necessary component models are loaded, reducing memory usage and increasing speed.

Security and Auditing:

In-database machine learning models follow database security schemes, including access control, privilege management, and audit tracking. This enables compliance and secure use of machine learning models across different environments.

5
Components

OML consists of a set of APIs and user interface components. Some components and features are API specific.

Topics:

APIs

OML supports OML4Py, OML4R, OML4SQL, and OML Services.

OML4Py

Oracle Machine Learning for Python (OML4Py) enables you to run Python code for data transformations and for statistical, machine learning, and graphical analysis on data stored in or accessible through Oracle Autonomous Database and Oracle Database instances using a Python API. OML4Py is a proprietary Python library that enables Python users to manipulate data in database tables and views using Python syntax. OML4Py provides a select set of Python functions and methods that transparently generate SQL and PL/SQL to perform requested functionality with in-database processing.

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 automated model tuning and selection. Users can use Embedded Python Execution to run user-defined Python functions in Python engines spawned by the Oracle Autonomous Database and Oracle Database instances.

OML4Py is included with Oracle Database on-premises, Base Database Service (BDBS), and Oracle Autonomous Database. To learn more, see Machine Learning System Requirements.

OML4R

Oracle Machine Learning for R (OML4R) enables you to run R code for data transformations, statistical analysis, machine learning, and graphical analysis on data stored in or accessible through Oracle Autonomous Database and Oracle Database instances using an R API. OML4R is a proprietary set of R packages that allows R users to manipulate data in database tables and views using R syntax and run user-defined R functions. OML4R functions and methods translate a select set of R functions that transparently generate SQL and PL/SQL to perform requested functionality with in-database processing.

OML4R supports in-database machine learning model building, scoring, and evaluation using Oracle Machine Learning algorithms. However, users can leverage Embedded R Execution to

run user-defined R functions inside R engines managed by Oracle Autonomous Database and Oracle Database instances.

OML4R is included with Oracle Database on-premises, Base Database Service (BDBS), and Oracle Autonomous Database.

OML4SQL

Oracle Machine Learning for SQL (OML4SQL) provides PL/SQL access to powerful, in-database machine learning algorithms and SQL access to corresponding models. You can use OML4SQL to build and deploy predictive and descriptive machine learning models that can be used to add intelligent capabilities to applications and dashboards. OML4SQL is included across Oracle Autonomous Database and Oracle Database instances.

To learn more, see Machine Learning System Requirements.

OML Services

Oracle Machine Learning Services (OML Services) provides MLOps support on Autonomous Database Serverless and Dedicated Region, where you can manage and use machine learning models from REST endpoints.

OML Services supports model management, deployment, data and model monitoring, as well as data bias detection. OML Services provides lightweight scoring using REST endpoints, making it appropriate for real-time and streaming applications.

Users have the option to deploy in-database models or "bring your own model" in Open Neural Networks Exchange (ONNX) format. These ONNX-format models can be created in third-party environments and imported into the database for use through the same REST API. OML Services supports various types of models, including classification, regression, clustering, and feature extraction.

Unlike other machine learning model deployment methods, which require you to provision, configure, manage, and pay for a Virtual Machine around the clock, OML Services only charges for the actual scoring. The Autonomous Database takes care of provisioning and managing the Virtual Machine environment.

To learn more, see What is OML Services.

User Interfaces

Oracle Machine Learning offers User Interfaces (UIs) catering to a broad range of users, from data scientists with advanced coding skills to users with limited technical background.

OML Home Page

The Oracle Machine Learning User Interface home page provides you quick links to important interfaces, help links, and the log of your high-level recent activities.

To learn more, see Oracle Machine Learning User Interface Home Page .

Notebooks

Oracle Machine Learning Notebooks is a collaborative web-based interface that supports notebook creation, scheduled run, and versioning. You can document work using Markdown and run SQL, R, and Python code for data exploration, visualization, and preparation, and machine learning model building, evaluation, and deployment. Oracle Machine Learning Notebooks also provide the Conda interpreter to create a custom Conda environment that includes user-specified third-party Python and R libraries.

To learn more, see About Oracle Machine Learning Notebooks.

AutoML User Interface

The AutoML no-code UI simplifies model building for users with limited technical expertise. When you create and run an experiment in AutoML UI, it performs automated algorithm selection, feature selection, and model tuning, thereby enhancing productivity as well as potentially increasing model accuracy and performance.

To learn more, see Get Started with AutoML UI.

OML Models

The Models page displays the user models and the list of deployed models. User Model lists the models in a user's schema, and Deployments lists the models deployed to Oracle Machine Learning Services.

You can access the Models page through the OML user interface. To learn more, see Get Started with Models.

- **Data Monitoring**

  Data Monitoring evaluates how your data evolves over time. It helps you with insights on trends and multivariate dependencies in the data. It also gives you an early warning about data drift.

  Data drift occurs when data diverges from the original baseline data over time. Data drift can happen for a variety of reasons, such as a changing business environment, evolving user behavior and interest, data modifications from third-party sources, data quality issues, or issues with upstream data processing pipelines.

  To learn more, see Get Started with Data Monitoring.

- **Model Monitoring**

  Model monitoring allows you to monitor the quality of model predictions over time and helps you with insights on the causes of model quality issues.

  You can access Model Monitor through the OML user interface. To learn more, see Get Started with Model Monitoring.

Oracle Data Miner

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. To learn more, see Oracle SQL Developer. Select your release from the drop-down and click Books to access Oracle Data Miner documents.

Platform Availability

OML offers flexibility by supporting deployment across various platforms.

The key deployments are:

- Oracle Autonomous Database (ADB)
- Base Database Service (BDBS)
- On-Premises Oracle Database

Oracle Machine Learning Family of Components

The following table describes the availability of OML components on different platforms.

OML Component	Autonomous Database Serverless | Dedicated Region	Autonomous Database on Dedicated Exadata Infrastructure | Cloud @ Customer	Oracle Database On premises, Base Database Service, Cloud Service, Cloud Infrastructure, Cloud@Customer
OML4SQL API Build machine learning models and score data with no data movement using SQL and PL/SQL	✓	✓	✓
OML4Py API Leverage the database as a high performance compute engine from Python with in-database machine learning	✓		✓
OML4R API Leverage the database as a high performance compute engine from R with in-database machine learning	✓		✓
OML Notebooks SQL, PL/SQL, Python, R, Conda, and markdown interpreters	✓		
OML AutoML UI No code automated modelling interface	✓		
OML Monitoring No-code user interface for monitoring changes in data and in-database machine learning model quality	✓		
OML Services RESTful model management, deployment, monitoring	✓		
Oracle Data Miner SQL Developer extension with a drag-n-drop interface for creating machine learning methodologies	✓	✓	✓

Supported operating systems includes Linux operating system and Windows Server versions. See Machine Learning System Requirements.

6

Use Cases Using Oracle Machine Learning

You can find Oracle Machine Learning (OML) use cases as example templates in OML Notebooks covering OML4SQL, OML4R, and OML4Py APIs and as scripts on GitHub.

See Oracle Machine Learning GitHub Examples. In addition, see Oracle Machine Learning Use Cases for use cases including house price predictions, targeted marketing, customer segmentation, and movie stream data analysis.

Topics:

Examples Supported by OML

Oracle Machine Learning business problems span across key industry specific verticals and common broadly applicable cross-industry machine learning use cases.

Broadly, the horizontal common use cases that cross industries are categorized as follows:

- Customers
- Products
- Equipment
- Employees

Machine learning has significant applications in industry-specific scenarios, where it drives innovation, safety, and efficiency. A few key industries are as follows:

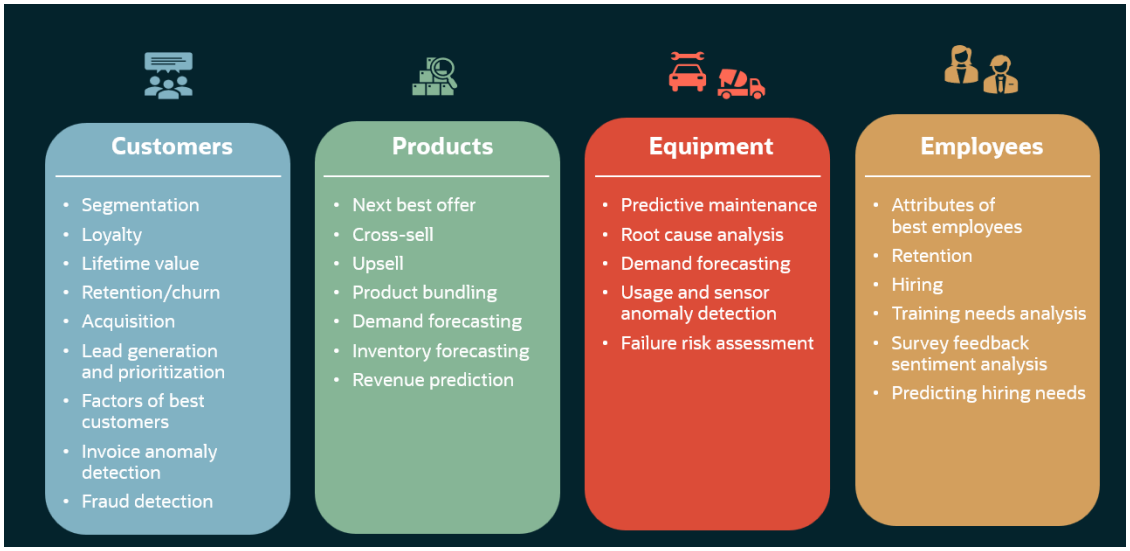
- Retail
- Finance
- Transport
- Entertainment
- Healthcare

> **Note:**
>
> This document provides insights on the common cross-industry use cases.

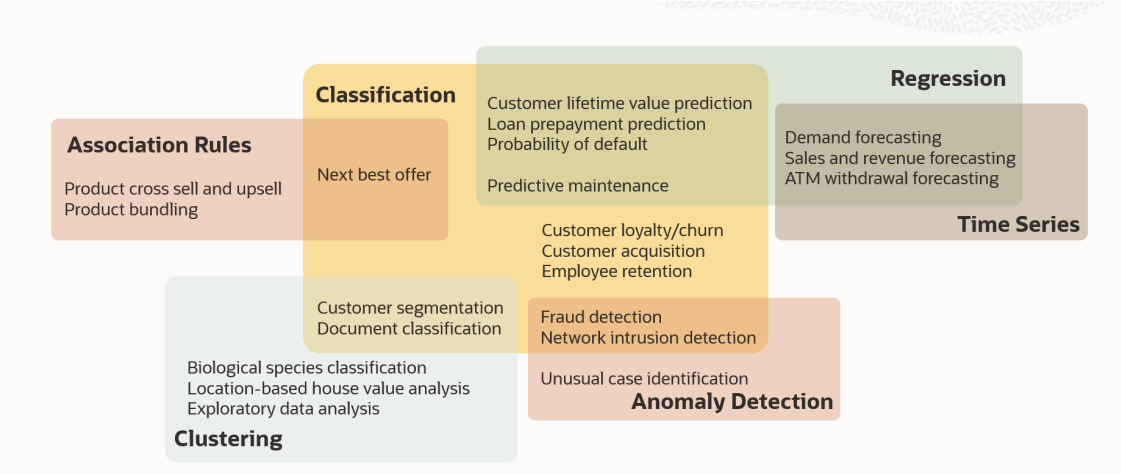
Common Use Cases that Cross Industries

Most enterprises, independent of industry, have customers, products, equipment, and employees. These are examples of use cases in each of these areas and can likely be applied in your enterprise.

> **Note:**
>
> This document highlights one use case from each of these focus areas.



The following image illustrates the techniques that can be applied for the sample use cases.



The following table summarizes one use case from common cross industry areas and applicable OML techniques and algorithms.

Area	Use Case	Problem Description	Data	Applicable Techniques and Algorithms	Related Resources
Customers	Customer Lifetime Value	Calculating the customer lifetime value (LTV) helps one determine the potential earnings for a company from a customer over their complete buying lifetime. It enables businesses to concentrate on long-term bonds with consumers rather than only short-term gains. Businesses anticipate future client expenditure and then modify that figure to consider time, therefore determining LTV. This enables companies to choose which consumers are most important and how best to allocate their money toward marketing and maintaining those relationships. See Wikipedia Customer Lifetime Value.	SH schema	• Classification • Regression	Blogs: • An Introduction to Predictive Customer Lifetime Value Modeling • Importing Data from Oracle Cloud Object Storage OML Notebook Example: • *Run-me-first* OML notebook example prepares customer data that can be used in various use cases.
Products	Cross-sell	Cross-selling is a sales approach that involves businesses suggesting extra products or services to their current customers. The goal is to increase sales and improve relationships with those customers. Different businesses have their own definitions of cross-selling, which can be shaped by things such as, size of the company, the industry it operates in, and its financial objectives. The main aim of cross-selling is to either boost how much customers spend or improve their loyalty by providing them with useful and relevant products. Companies can use this strategy by having different teams work on it within the organization or by teaming up with partner organizations to explore co-selling opportunities. Companies need to make sure that the new product or service adds value for customers. Big companies often use cross-selling and up-selling strategies together to increase their revenue. They encourage customers to consider buying more expensive items along with related products. See Wikipedia Cross-Selling.	SH schema	Association Rules	Blogs: Real Time Association Rules Recommendation Engine

Area	Use Case	Problem Description	Data	Applicable Techniques and Algorithms	Related Resources
Equipment	Predictive Maintenance	Manufacturers want to plan equipment maintenance at the right time to save replacement costs, prevent failures, and minimize downtime. They use time, usage, and sensor data including climate and vibration to guide decisions. Predicting failure with enough time to act meaningfully presents a major difficulty. For example, if scheduling service takes a week, a 24-hour failure warning is not useful.	Predictors typically include equipment details (make, manufacturer, in-service date, last maintenance, location, and so on), equipment-specific sensor data (For example, vibration, temperature, pressure), and environmental data (for example, ambient humidity, temperature, indoor or outdoor). The target definition (what we want to predict) must be correct for the current situation. For example, one such target could be an indication (0/1) variable labeled "failed within 5 to 7 days".	Classification Typical algorithms include: Support Vector Machines (SVM), Generalized Linear Model (GLM), Naive Bayes, Decision Tree, Neural Network, XGBoost.	Blogs: • Predictive Maintenance with Machine Learning on Oracle Database 20c • Plant predictive maintenance with Oracle Autonomous Data Warehouse

Area	Use Case	Problem Description	Data	Applicable Techniques and Algorithms	Related Resources
Employees	Retention	When employees leave a company, it costs a lot to find and train new ones. It's hard to know who might quit or why, which can affect how well the company works. This use case aims to predict churn risk and suggest retention strategies. It's like how businesses try to keep customers by offering them more; here, it's about keeping employees. The goal is to help companies act early to stop employees from quitting.	HR schema data with employee records. Typically a machine learning model looks at information about employees, such as, how long they've worked, their pay, how well they do their job, and if they're happy. It also uses details like their age, if they've been promoted, how often they're absent, and what jobs are available outside. Text from exit interviews or surveys can be included too.	Classification Typical algorithms include: XGBoost, Random Forest, Decision Tree, Neural Network.	Blogs: • 5 Ways Technology Supports Improved Employee Retention • AI as an Enabler in People Management

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 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.

confidence

Confidence in a rule is calculated by dividing the probability of the items occurring together by the probability of the occurrence of the antecedent.

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 model. An algorithm uses a specific data representation and a specific machine learning technique.

The algorithms supported by Oracle Machine Learning 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 that shares common high level characteristics. The Oracle Machine Learning 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.

model tuning

Model tuning is the process of optimizing a machine learning model's performance by systematically adjusting its hyperparameters - the settings that govern the training process but are not learned from the data itself.

The goal of model tuning is to improve model accuracy, generalization, or efficiency on unseen data by finding the optimal combination of hyperparameters. After selecting a model and training it with default settings, tuning is used to enhance its predictive power.

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.

precision

A metric used to evaluate the accuracy of a classification model by measuring the proportion of true positive predictions out of all positive predictions.

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.

recall

A metric used to evaluate the accuracy of a classification model by measuring the proportion of true positive predictions out of all actual positive instances.

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.

Root Mean Squared Error (RMSE)

RMSE is a commonly used metric for evaluating the accuracy of a regression model. It measures the average magnitude of the errors between predicted values and actual values, with greater weight given to larger errors. RMSE is the square root of the average of the squared differences between prediction and actual observation.
RMSE is ideal when large errors are particularly undesirable and is often used to compare different regression models' predictive accuracy.

Interpretation:

- Lower RMSE indicates better model performance.
- RMSE is expressed in the same units as the target variable.

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.

tokenize

Tokenization is the process of breaking down text into smaller units called tokens, which can be words, subwords, characters, or symbols, depending on the application. It prepares raw text for machine learning or natural language processing (NLP) tasks by converting unstructured language into a structured form that models can understand and process. It can be used for text classification, sentiment analysis, language translation, and search indexing.

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