Oracle Machine Learning Overview and Roadmap

Oracle Machine Learning Product Management
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle’s products may change and remains at the sole discretion of Oracle Corporation.
Oracle Machine Learning Key Attributes

**Automated**
Get better results faster with less effort – even non-expert users

**Scalable**
Handle big data volumes using parallel, distributed algorithms – no data movement

**Production-ready**
Deploy and update data science solutions faster with integrated ML platform

Increase productivity, Achieve enterprise goals, Innovate more
Using Oracle Machine Learning to achieve business goals

A few examples

- **Zagrebačka banka** (UniCredit Group)
  - Increase Cash Loans by 15% within 18 Months of Deployment

- **NHS Business Services Authority**
  - Combat healthcare fraud and address dependency-forming medicine abuse

- **StubHub**
  - Dramatically reduce online fraud, significantly improve conversions

- **DXM**
  - Reveal data insights on target customers and focus marketing dollars

https://www.oracle.com/search/customers
OML empowers Enterprise Users

- Data Scientists
- Executives
- Business and Data Analysts
- Application / Dashboard Developers
- DBA and IT Professionals
Data Scientists

- Popular data science languages: Python, R, SQL
- Augment with 3rd party packages
- Scalability and performance
- Automation-enhanced productivity
- Greater enterprise collaboration
- Integrate/analyze data across the enterprise
Business Analysts and Data Analysts

• Expand analytical tool set with ML
• Enable non-ML experts with AutoML
• Leverage domain knowledge for better results
• Collaborate with Data Scientists and IT
Even greater value from Oracle investment
Support scalability and performance
Simpler, streamlined infrastructure
Maintain data security, backup, recovery
Use SQL, expand to Python and R
Leverage Database and Big Data sources
Application and Dashboard Developers

- Realize intelligent solutions faster through Oracle stack integration
- Easily uptake data scientists’ R, Python, SQL scripts and rapidly deploy solutions
- Embed ML in applications and dashboards using, e.g., SQL and REST APIs

Copyright © 2020 Oracle and/or its affiliates.
Executives

• Benefit from world-class data management technology and support
• Democratize ML across the enterprise to enable better data-driven decisions
• Deploy solutions faster to realize ROI
Cross-Platform Machine Learning

Multiple user interfaces and APIs
Deployed in cloud and on-premises
From database to entire data management ecosystem

User Interfaces, e.g.
- OML Notebooks
- SQL Developer
- Popular R IDEs
- Popular Python IDEs
- OCI Data Science
- OAC

APIs
- OML4SQL
- OML4R
- OML4Py
- REST
- OML4Spark

Cloud or On-premises
Reach broader Data Sources
- Oracle Cloud SQL
- Oracle Big Data SQL

Data Lake
- Oracle Database
- Oracle Object Storage
- Amazon S3
- Azure Blob Storage
- NoSQL Databases
- Kafka Streams
- Big Data Service (HDFS)

Copyright © 2020 Oracle and/or its affiliates.
Oracle Machine Learning interfaces to Oracle Database

Tool
- Apache Zeppelin
- Python client, Jupyter Notebooks
- SQL Developer SQL*Plus
- R client, RStudio
- SQL Developer

Oracle Machine Learning Component
- OML Notebooks
- OML4Py*
- OML4SQL
- OML4R*

Data Management Platform
- Autonomous Database
- Oracle Database
- Database Cloud Service

*Coming soon
Oracle Machine Learning Algorithms and Analytics

**Classification**
Naïve Bayes
Logistic Regression (GLM)
Decision Tree
Random Forest
Neural Network
Support Vector Machine (SVM)
Explicit Semantic Analysis
XGBoost

**Regression**
Linear Model
Generalized Linear Model (GLM)
Support Vector Machine (SVM)
Stepwise Linear regression
Neural Network
LASSO
XGBoost

**Clustering**
Hierarchical K-Means
Hierarchical O-Cluster
Expectation Maximization (EM)

**Anomaly Detection**
One-Class SVM
MSET-SPRT

**Time Series**
Forecasting - Exponential Smoothing
Includes popular models
e.g. Holt-Winters with trends, seasonality, irregularity, missing data

**Feature Extraction**
Principal Comp Analysis (PCA)
Non-negative Matrix Factorization
Singular Value Decomposition (SVD)
Explicit Semantic Analysis (ESA)

**Text Mining Support**
Algorithms support text columns
Tokenization and theme extraction
Explicit Semantic Analysis (ESA) for document similarity

**Statistical Functions**
Basic statistics: min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

**R and Python Packages**
Third-party R and Python Packages through Embedded Execution
Spark MLlib algorithm integration

New in 20c
Oracle Machine Learning Notebooks
Autonomous Database as a Data Science Platform

Collaborative UI
- Based on Apache Zeppelin
- Supports data scientists, data analysts, application developers, DBAs
- Easy sharing of notebooks and templates
- Permissions, versioning, and execution scheduling

Included with Autonomous Database
- Automatically provisioned, managed, backed up
- In-database SQL algorithms and analytics functions
- Explore and prepare, build and evaluate models, score data, deploy solutions
- Soon to be augmented with Python and R

Credit Score Predictions

Review Data by Occupation

Copyright © 2020 Oracle and/or its affiliates.
Oracle Machine Learning for SQL (OML4SQL) Python (OML4Py) R (OML4R)

Empower SQL users with immediate access to ML in Oracle Database and Oracle Autonomous Database
Empower data scientists with open source environments
Oracle Machine Learning for SQL

Empower SQL users with immediate access to ML included with Oracle Database and Oracle Autonomous Database

In-database, parallel, distributed algorithms
- No extracting data to separate ML engine
- Fast and scalable
- Batch and real-time scoring
- Explanatory prediction details

ML models as first class database objects
- Access control via permissions
- Audit user actions
- Export / import models across databases

Leverage ML across Oracle stack

Copyright © 2020 Oracle and/or its affiliates.
OML4SQL: Model Build and Real-time Prediction

Simple SQL Syntax—Classification Model

Model build (PL/SQL)

```sql
BEGIN
    DBMS_DATA_MINING.CREATE_MODEL(
        model_name => 'BUY_INSUR1',
        mining_function => dbms_data_mining.classification,
        data_table_name => 'CUST_INSUR_LTV',
        case_id_column_name => 'CUST_ID',
        target_column_name => 'BUY_INSURANCE',
        settings_table_name => 'CUST_INSUR_LTV_SET');
END;
```

Real-time scoring (SQL query)

```sql
SELECT prediction_probability('BUY_INSUR1', 'Yes'
    USING 3500 as bank_funds, 825 as checking_amount, 400 as credit_balance, 22 as age,
    'Married' as marital_status, 93 as MONEY_MONTHLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```
Oracle Data Miner User Interface

Create analytical workflows – supports “Citizen Data Scientists”

SQL Developer Extension for Oracle Database
Automates typical data science steps
Easy to use drag-and-drop interface
Analytical workflows quickly defined and shared
Wide range of algorithms and data transformations
Generate SQL code for immediate deployment

Copyright © 2020 Oracle and/or its affiliates.
Traditional Analytics and Data Source Interaction

Access latency
Paradigm shift: R/Python \(\rightarrow\) \textit{Data Access Language} \(\rightarrow\) R/Python
Memory limitation – data size, in-memory processing
Single threaded
Issues for backup, recovery, security
Ad hoc production deployment
Oracle Machine Learning for R and Python

Included with Oracle Database – R today, Python coming soon. Both coming soon to Oracle Autonomous Database.

Oracle Database as HPC environment
In-database parallel and distributed machine learning algorithms
Manage scripts and objects in Oracle Database
Integrate results into applications and dashboards via SQL
OML4Py automated machine learning

Empower data scientists with open source environments

Copyright © 2020 Oracle and/or its affiliates.
Oracle Machine Learning for R and Python

Transparency layer
- Leverage proxy objects so data remain in database
- Overload native functions translating functionality to SQL
- Use familiar R / Python syntax on database data

Parallel, distributed algorithms
- Scalability and performance
- Exposes in-database algorithms available from OML4SQL

Embedded execution
- Manage and invoke R or Python scripts in Oracle Database
- Data-parallel, task-parallel, and non-parallel execution
- Use open source packages to augment functionality

OML4Py AutoML
- Algorithm selection, feature selection, hyper-parameter tuning

Copyright © 2020 Oracle and/or its affiliates.
Transparency Layer

In-database performance – indexes, query optimization, parallelism, partitioning

Leverages proxy objects for database data: `oml.DataFrame`

```python
# Create table from Pandas DataFrame data
DATA = oml.create(data, table = 'BOSTON')

# Get proxy object to DB table boston
DATA = oml.sync(table = 'BOSTON')
```

Uses familiar Python syntax to manipulate database data

Overloads Python functions translating functionality to SQL

```python
DATA.shape
DATA.head()
DATA.describe()
DATA.std()
DATA.skew()

TRAIN, TEST = DATA.split()
TRAIN.shape
TEST.shape
```
from oml import svm

# create proxy object
ONTIME_S = oml.sync(table='ONTIME_S')

# define model object
settings = {'svms_outlier_rate': 0.01}
svm_mod = svm('anomaly_detection',
              svms_kernel_function =
              'dbms_data_mining.svms_linear',
              **settings)

# build anomaly detection model
svm_mod = svm_mod.fit(x=ONTIME_S, y=None)

# view model object
svm_mod
Deploy user-defined functions in applications and dashboards
Example of parallel execution for partitioned data flow using third party package

```python
# user-defined function using sklearn
def build_lm(dat):
    from sklearn import linear_model
    lm = linear_model.LinearRegression()
    X = dat[['PETAL_WIDTH']]
    y = dat[['PETAL_LENGTH']]
    lm.fit(X, y)
    return lm

# select column(s) for partitioning data
index = oml.DataFrame(IRIS['SPECIES'])
# invoke function in parallel on IRIS table
mods = oml.group_apply(IRIS, index, func=build_lm, parallel=2)

mods.pull().items()
```
AutoML – *new* with OML4Py

Increase data scientist productivity – reduce overall compute time

**Auto Algorithm Selection**
- Identify in-database algorithm that achieves highest model quality
- Find best algorithm faster than with exhaustive search

**Auto Feature Selection**
- Reduce # of features by identifying most predictive
- Improve performance and accuracy

**AutoTune Hyperparameters**
- Significantly improve model accuracy
- Avoid manual or exhaustive search techniques

Enables non-expert users to leverage Machine Learning
Oracle Machine Learning for Python (OML4Py) - SQL API

Hands-On Lab

(c) 2019 Oracle Corporation

-- Random Red Dots
-- Generate images of random red dots and a simple data.frame result

BEGIN
    sys.pyScriptDrop('RandomRedDots_Py');
    sys.pyScriptCreate('RandomRedDots_Py',
        'def RandomRedDots_Py (num_dots_1=100, num_dots_2=10):
            import numpy as np
            import pandas as pd
            import matplotlib.pyplot as plt

            d = np.random.rand(num_dots_1, 1) * 100
            f = np.random.rand(num_dots_2, 1) * 100

            for x in range(10):
                plt.plot(d[:,0], f[0,:])
            plt.show()
    ');
Oracle Machine Learning for Spark (OML4Spark)
supported by Oracle R Advanced Analytics for Hadoop
Oracle Machine Learning for Spark

R Language API Component to Oracle Big Data Connectors

Leverage Spark 2 environment for powerful data preparation and machine learning
Use data across range of Data Lake sources
Achieve scalability and performance using full Hadoop cluster
Parallel and distributed ML algorithms from native and Spark MLlib implementations
Oracle Machine Learning for Spark

R Language API Component to Oracle Big Data Connectors

Transparency layer
- Proxy objects reference data from file system, HDFS, Hive, Impala, Spark DataFrame and JDBC sources
- Overloaded R functions translate functionality to native language, e.g., HiveQL for HIVE and Impala
- Users manipulate data via standard R syntax

Parallel, distributed machine learning algorithms
- Scalability and performance leveraging full Hadoop cluster
- Spark-based custom LM, GLM, NN, K-Means plus Spark MLlib
- Use expressive R Formula specification

Compute framework with custom R mappers/reducers
- Data-parallel and task-parallel execution
- Allows for open source CRAN packages run on Cluster Nodes

Copyright © 2020 Oracle and/or its affiliates.
OML4Spark Performance

Logistic Regression (GLM)
Data fits in memory
  Up to 7x faster than Spark MLlib
Data cannot fit memory
  Able to solve a 10B row model
Benchmark environment
  ORAAH 2.8.0
  Big Data Appliance X7-2
  6 Nodes, 256GB of RAM per Node

Formula: cancelled ~ distance + origin + dest + as.factor(month) + as.factor(year) + as.factor(dayofmonth) + as.factor(dayofweek) + as.factor(flightnum)

OML4Spark vs. Spark MLlib for GLM Logistic Regression

Dataset Size (# rows) vs. Execution Time (seconds)

- OML4Spark
- MLlib

Copyright © 2020 Oracle and/or its affiliates.
Oracle Applications

Integrations using Oracle Machine Learning
Applications integrating OML

HCM Cloud
Workforce Predictions

CRM Sales Cloud
Sales Prediction

Retail GBU
Customer Insights,
Customer Segmentation
Adaptive Intelligent Applications
for Manufacturing

Configure, Price, Quote Cloud

Content and Experience
Unstructured Data Analytics

Integration Cloud
Digital Process Automation

Industry Data Models
Communications, SNA, Utilities, Airlines, Retail, …

EBS Spend Classification
Organize spend into logical categories

EBS Depot Repair
Optimize speed, cost, quality of
product repair, reuse, recycling

Identity Management
Adaptive Access Management

FSGBU
Analytical Applications
Infrastructure
Application platforms using OML Services

Oracle Integration Cloud (OIC)

Digital Process Automation
- Help business users make better decisions by using recommendations from ML models
- Increase automation of human-centric approval workflows

Used by Oracle SaaS process-centric apps

PaaS service that exposes OML features
- Build models in ADB
- Deploy via OML Services

Oracle Content and Experience (OCE)

Improve Content Discoverability
- Search, organize content, reduce duplication
- Find relevant images/docs during content creation
- Automatic tagging and classification of images, videos, text
- Visual search

Cloud-based content hub to drive omni-channel content management and accelerate experience delivery

Leverages OML Cognitive Services

Copyright © 2020 Oracle and/or its affiliates.
Oracle Machine Learning Roadmap

Expanding Oracle’s investment in machine learning
Roadmap: OML Services

Currently integrated with *internal* Oracle Applications

Model Management Services
- Building and deploying OML models
- Model Monitoring of accuracy and prediction/predictor drift

Model repository
- Store, version, compare ML models

Cognitive Services
- Feature Extraction, Image and Text

User-defined scripts deployment
- Python and R user-defined functions invoked via REST API

REST APIs for application integration

Copyright © 2020 Oracle and/or its affiliates.
Roadmap: Sample of OML Services APIs

**Model Management**
- GET /models
- GET /{model name}
- GET /{model name}/{version}
- POST /{model name}
- POST /{model name}/{version}
- DELETE /{model name}/{version}

**Model Deployment**
- GET /models
- GET /{uri}
- GET /{uri}/api
- POST /{uri}
- POST /{uri}/score
- DELETE /{uri}

**Cognitive Image**
- POST /imageClassification
- POST /nsfw
- POST /objectDetection
- POST /faceDetection
- POST /imageSimilarity
- POST /faceSimilarity

**Cognitive Text**
- POST /ner
- POST /topics
- POST /keywords
- POST /sentiment
- POST /summary
- POST /similarity
Roadmap: Expand Autonomous Database with Python and R

Autonomous Database as a Data Science Platform

OML Notebooks add support for Python and R
Python and R scripts managed in-database
  - Invoke from OML Notebooks, and REST or SQL APIs
  - Deploy into SQL and Web applications easily
Scalable Python and R execution
  - Transparency layer-enabled database functionality
  - In-database machine learning algorithms
AutoML functionality via OML4Py
OML4Py integrated with OCI Data Science

Copyright © 2020 Oracle and/or its affiliates.
**Roadmap: OML AutoML User Interface**

“Code-free” user interface supporting automated end-to-end machine learning

Automate production and deployment of ML models
- Enhance Data Scientist productivity, user-experience
- Enable non-expert users to leverage ML
- Unify model deployment and monitoring
- Support model management

Features
- Minimal user input: data, target
- Model leaderboard
- Model deployment in applications via REST endpoint
- Model monitoring: accuracy, prediction/predictor drift
- Cognitive features for processing image and text

*Sample screen mock-up*

---

Copyright © 2020 Oracle and/or its affiliates.
Roadmap: OML4R and OML4Py

Expand support for open source languages and ecosystems

Expose additional OML4SQL algorithms to Python and R
Support for recent R and Python releases
Enable Oracle Database standard integrated installation, patching, upgrade/downgrade
Roadmap: OML4Spark

New cloud-based architecture with powerful Spark analytics

Support advanced machine learning activities on Big Data
  - Model management
  - Cognitive image and text processing
  - Model deployment and monitoring on Big Data (including Database models)

Cloud-oriented packaging (containers, REST APIs)

Expand set of natively-supported data formats and sources

Enable OML4Py and OML4R for uniform experience across platforms

Third-party open source library support, e.g., TensorFlow, MxNet, ResNet, Keras

Algorithms
  - Neural Network deep learning support, including GPU compute options
  - Neural Network gradient descent enhancements avoid over-fitting
  - Native Support Vector Machine with linear and non-linear kernels
  - Native k-Means and k-Modes clustering algorithms
Roadmap: Enabling OML on GPUs

Leverage GPUs for user-defined R and Python functions
- Include 3rd party packages leveraging GPUs, e.g., Tensorflow, Keras
- Support state-of-the-art ML processing, e.g., deep learning

Augment OML Services for GPU processing – key for images
Why Oracle for Machine Learning?

Oracle integrates ML across the Oracle Stack and the Enterprise

Empower data scientists and analysts, developers, and DBAs/IT with ML

Eliminate costly data movement and latency

Fast and scalable data exploration, data preparation, and ML algorithms

Over 30 algorithms supporting: regression, classification, time series, clustering, feature extraction, anomaly detection

R and Python integration supports data scientists

Ease of ML model and R/Python script deployment

Automate key ML process steps

Leverage existing backup, recovery, and security mechanisms and protocols

That’s where most enterprise data lives – bring the algorithms to the data!

*Oracle Database and Oracle Autonomous Database*
Oracle Machine Learning **Key Attributes**

**Automated**

**Scalable**

**Production-ready**

Increase productivity, Achieve enterprise goals, Innovate more
Oracle Machine Learning

The Oracle Machine Learning product family enables scalable data science projects. Data scientists, analysts, developers, and IT can achieve data science project goals faster while taking full advantage of the Oracle platform.

Oracle Machine Learning consists of complementary components supporting scalable machine learning algorithms for in-database and big data environments, notebook technology, SQL and R APIs, and Hadoop/Spark environments.

See also AskTOM OML Office Hours
Thank You

Oracle Machine Learning Product Management