Oracle’s Machine Learning and Advanced Analytics
Release 12.2 and Oracle Data Miner 4.2 New Features

Move the Algorithms; Not the Data!

Charlie Berger, MS Engineering, MBA
Sr. Director Product Management,
Advanced Analytics and Machine Learning
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine
Safe Harbor Statement

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, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Machine Learning/Analytics + Data Warehouse + Hadoop

- Platform Sprawl
  - More Duplicated Data
  - More Data Movement Latency
  - More Security challenges
  - More Duplicated Storage
  - More Duplicated Backups
  - More Duplicated Systems
  - More Space and Power
Oracle’s Advanced Analytics and Machine Learning Platform

Machine Learning Algorithms Embedded in the Data Management Platform

“Information Producers”
Data Scientists, R Users, Citizen Data Scientists

R Studio Client

SQL Dev/Data Miner

“Information Consumers”
BI Analysts, Managers
Functional Users (HCM, CRM)

Oracle Data Viz

“Predictive” Applications

Oracle’s Advanced Analytics and Machine Learning Platform

Machine Learning Algorithms Embedded in the Data Management Platform

Oracle BDA Hadoop

ORAAH
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed Execution

Oracle Database EE

Oracle Advanced Analytics (ODM + ORE)
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed, in-DB Execution

Oracle Cloud

Copyright © 2017, Oracle and/or its affiliates. All rights reserved.
Oracle’s Adv. Analytics Machine Learning Algorithms

- **Classification**
  - Naïve Bayes
  - Logistic Regression (GLM)
  - Decision Tree
  - Random Forest
  - Neural Network
  - Support Vector Machine
  - Explicit Semantic Analysis
  - Gaussian Mixture Models

- **Regression**
  - Generalized Linear Model
  - Support Vector Machine (SVM)
  - Random Forest
  - Linear Model
  - Stepwise Linear regression
  - LASSO

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

- **Anomaly Detection**
  - One-Class Support Vector Machine (SVM)

- **Association Rules**
  - A priori

- **Algorithm Support for Text**
  - Algorithms support text type
  - Tokenization and theme extraction
  - Explicit Semantic Analysis (ESA) for document similarity

- **Feature Extraction**
  - Principal Component Analysis (PCA)
  - Non-negative Matrix Factorization
  - Singular Value Decomposition (SVD)

- **Time Series**
  - Single Exponential Smoothing
  - Double Exponential Smoothing

- **Open Source ML Algorithms**
  - CRAN R Algorithm Packages through Embedded R Execution
  - Spark MLlib algorithm integration

- **Predictive Queries**

- **Statistical Functions**
  - Basic statistics: median, stdev, t-test, F-test, Pearson’s, Chi-sq, Anova, etc.

- **Ability to Mine Unstructured, Structured, & Transactional data**
  - Support for SQL “Partition-By” Models
Machine Learning & Advanced Analytical Methodologies
Data Preparation & Adv. Analytical Process Runs In-Database

Oracle Database 12c

Additional relevant data and “engineered features”

Historical data, Text, unstructured data, transactional data, spatial data, etc.

Historical data

Assembled historical data

Build Predictive Model

Make Predictions

Predictions & Insights

Copyright © 2016, Oracle and/or its affiliates. All rights reserved.
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide ML/Predictive Analytics

Major Benefits

- Data remains in Database & Hadoop
  - Model building and scoring occur in-database
  - Use R packages with data-parallel invocations
- Leverage investment in Oracle IT
  - Eliminate data duplication
  - Eliminate separate analytical servers
- Deliver enterprise-wide applications
  - GUI for ML/Predictive Analytics & code gen
  - R interface leverages database as HPC engine

Oracle Advanced Analytics vs. Traditional Analytics:

- Data Import
- Data Prep. & Transformation
- Data Mining Model “Scoring”
- Data Mining Model Building

Oracle’s Advanced Analytics:
- Hours, Days or Weeks
- Secs, Mins or Hours

Savings
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Parallel, scalable data mining algorithms and R integration
- In-Database + Hadoop—Don’t move the data
- Data analysts, data scientists & developers
- Drag and drop workflow, R and SQL APIs
- Extends data management into powerful advanced/predictive analytics platform
- Enables enterprise predictive analytics deployment + applications
You Can Think of Oracle’s Advanced Analytics Like This...

**Traditional SQL**
- “Human-driven” queries
- Domain expertise
- Any “rules” must be defined and managed

**SQL Queries**
- SELECT
- DISTINCT
- AGGREGATE
- WHERE
- AND OR
- GROUP BY
- ORDER BY
- RANK

**Oracle Advanced Analytics - SQL &**
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine/analyze

**Analytical SQL “Verbs”**
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Oracle Text

Native Capability of every Oracle Database

• Oracle Text uses standard SQL to index, search, and analyze text and documents stored in the Oracle database, in files, and on the web.

• Oracle Text supports multiple languages and uses advanced relevance-ranking technology to improve search quality.

• Oracle Advanced Analytics leverages Oracle Text to pre-process ("tokenize") unstructured data for the OAA SQL ML/data mining functions.
Oracle Advanced Analytics
How Oracle R Enterprise Compute Engines Work

1. R-> SQL Transparency “Push-Down”
   - R language for interaction with the database
   - R-SQL Transparency Framework overloads R functions for scalable in-database execution
   - Function overload for data selection, manipulation and transforms
   - Interactive display of graphical results and flow control as in standard R
   - Submit user-defined R functions for execution at database server under control of Oracle Database

2. In-Database Adv Analytical SQL Functions
   - 30+ Powerful data mining algorithms (regression, clustering, AR, DT, etc.)
   - Run Oracle Data Mining SQL data mining functioning (ORE.odmSVM, ORE.odmDT, etc.)
   - Speak “R” but executes as proprietary in-database SQL functions—machine learning algorithms and statistical functions
   - Leverage database strengths: SQL parallelism, scale to large datasets, security
   - Access big data in Database and Hadoop via SQL, R, and Big Data SQL

3. Embedded R Package Callouts
   - R Engine(s) spawned by Oracle DB for database-managed parallelism
   - ore.groupApply high performance scoring
   - Efficient data transfer to spawned R engines
   - Emulate map-reduce style algorithms and applications
   - Enables production deployment and automated execution of R scripts
Oracle Advanced Analytics 12.2, Oracle Data Miner 4.2
New Features – Agenda

1. New Machine Learning Algorithms
2. Significant ML Model Build Performance Improvements
3. Partitioned Models
4. Oracle Data Miner Update of OAA Server Enhancements
5. Analytical Platform that Enables “Predictive” Applications
Oracle Advanced Analytics 12.2

New Oracle Database Features

• Unsupervised Feature Selection
  – Unsupervised algorithm for pair-wise correlations (Kullback-Leibler Divergence (KLD)) for numeric & categorical attributes to find highest “information containing” attributes

• Association Rules Enhancements
  – Adds calculation of values associated with AR rules such as sales amount to indicate the value of co-occurring items in baskets
  – Can filter input items prior to market basket analysis

• Significant Performance Improvements for all Algorithms
  – New parallel model build / apply redesigned infrastructure to enable faster new algorithm introduction
  – Scale to larger data volumes found in big data and cloud use cases

• Partitioned Models
  – Instead of building, naming and referencing 10s or 1000s of models, partitioned models organize and represent multiple models as partitions in a single model entity
## Oracle Advanced Analytics 12.2

### Model Build Time Performance

**OAA 12.2 Algorithms** | **Rows (Ms)** | **Model Build Time (Secs / Degree of Parallelism)**
--- | --- | ---
Attributes Importance | 640 | 28s / 512
K Means Clustering | 640 | 161s / 256
Expectation Maximization | 159 | 455s / 512
Naive Bayes Classification | 320 | 17s / 256
GLM Classification | 640 | 154s / 512
GLM Regression | 640 | 55s / 512
Support Vector Machine (IPM solver) | 640 | 404s / 512
Support Vector Machine (SGD solver) | 640 | 84s / 256

### Notes

The way to read their results is that they compare 2 chips: X5 (Intel and Linux) and T7 (Sparc and Solaris). They are measuring scalability (time in seconds) with increase degree of parallelism (dop). The data also has high cardinality categorical columns which translates in 9K mining attributes (when algorithms require explosion). There are no comparisons to 12.1 and it is fair to say that the 12.1 algorithms could not run on data of this size.

**Wow! That’s Fast!**
Oracle Advanced Analytics 12.2
New Oracle Database Features

• Explicit Semantic Analysis (ESA) algorithm
  – Useful technique for extracting meaningful, interpretable features; better than LDA
  – English Wikipedia is Text corpus default to equate tokens with human identifiable features and concepts
  – ESA improves text processing, classification, document similarity and topic identification
  – Compare documents that may not even mention same topics e.g. al-Qa ida or Osama bin Laden:

  **Document 1**
  – ‘Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa ida, including bin Laden. The money enabled al-Qa ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia.’

  **Document 2**
  – ‘The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed “and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court,” the official Saudi Press Agency quoted an unidentified ministry official as saying.’

  ESA Similarity Score = 0.62
Explicit Semantic Analysis (ESA) algorithm

"The more things change... Yes, I'm inclined to agree, especially with regards to the historical relationship between stock prices and bond yields. The two have generally traded together, rising during periods of economic growth and falling during periods of contraction. Consider the period from 1998 through 2010, during which the U.S. economy experienced two expansions as well as two recessions: Then central banks came to the rescue. Fed Chairman Ben Bernanke led from Washington with the help of the bank's current $3.6T balance sheet. He's accompanied by Mario Draghi at the European Central Bank and an equally forthright Shinzo Abe in Japan. Their coordinated monetary expansion has provided all the sugar needed for an equities moonshot, while they vowed to hold global borrowing costs at record lows"

Top topics (concepts, people, organizations, events) discovered by ESA using Wikipedia as model source data

- Recession, Ben Bernanke, Lost Decade Japan, Mario Draghi, Quantitative easing, Long Depression, Great Recession, Federal Open Market Committee, Bank of Canada, Monetary policy, Japanese asset price bubble, Money supply, Great Depression, Central bank, Federal Reserve System

If instead of using the entire Wikipedia, we limit ourselves to the source dataset comprised of concepts only, this result would translate to:

- Recession, Quantitative easing, Monetary policy, Money supply, Central bank, Federal Reserve System
Oracle Advanced Analytics 12.2
New Oracle Database Features

• Explicit Semantic Analysis (ESA) algorithm

SELECT FEATURE_COMPARE(feat_esa_1.1)
  USING 'Oracle database is the best available for managing your data. Text
  AND USING 'The SQL language is the one language that all databases have in common'.
FROM DUAL;

The result we get is 0.7629.
Oracle Advanced Analytics 12.2
New Oracle Database Features

• Extensibility for R Models
  – Register R models as in-database models for build, apply, settings, and viewing
  – Extends ease of advanced analytics development from R to Oracle Database
  – Supports data with “nested” attributes, handling text and aggregated transactional data for open source R packages
  – Oracle Data Miner R Build Node
    • The R Build Node allows you to register R models. It builds R models and generates R model test results for Classification and Regression mining function. R Build nodes supports Classification, Regression, Clustering, and Feature Extraction mining functions only.
    – Enables R users to roll out new analytics and more rapidly take advantage of existing R packages
Oracle Data Miner 4.2

New Features for OAA

- Add/Expose all 12.2 features in Oracle Data Miner UI
NEW IN 4.2
Workflow Scheduler
Previewing a 4.2 Feature