Oracle’s Machine Learning and Advanced Analytics 12.2 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.
Dilbert on Big Data

SO, WHAT DO YOU DO FOR A LIVING?

I'M WORKING ON A FRAMEWORK TO ALLOW CONSTRUCTION OF LARGE-SCALE ANALYTICAL QUERIES ON UNSTRUCTURED DATA.

I'M A LITTLE TURNED ON BY THAT.

SETTLE DOWN. IT'S JUST A FRAMEWORK.
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
Traditional vs. Oracle Machine Learning/Predictive Analytics

- **Traditional** — “Move the data”

- **Oracle** — “Don’t move the data!”

The Predictive Modeling Process: Data Flow

- Data Warehouse
- Statistical Software
- Data Extract
- Model Formula
- $y_i = \beta_0 + \beta_1 x_i$
Traditional vs. Oracle Machine Learning/Predictive Analytics

- **Traditional** — “Move the data”

  - **ORACLE** — “Move the algorithms”

Simpler, Smarter Data Management + Analytics / Machine Learning Architecture
Oracle’s Machine Learning/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

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

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

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

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

• **Association Rules**
  - A priori

• **Attribute Importance**
  - Minimum Description Length
  - Principal Component Analysis (PCA)
  - Unsupervised Pair-wise KL Divergence

• **Predictive Queries**

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

• **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

+ Ability to Mine Unstructured, Structured, & Transactional data
+ Support for SQL “Partition-By” Models

Copyright © 2016, Oracle and/or its affiliates. All rights reserved.
Oracle’s Machine Learning/Advanced Analytics Platforms

Machine Learning Algorithms Embedded in the Data Management Platforms

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

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

Oracle BDA Hadoop
“Oracle ML” Big Data Edition
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed Execution

Oracle Database EE
“Oracle Machine Learning” Database Edition
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed, in-DB Execution

Oracle Cloud
Oracle’s Machine Learning/Advanced Analytics Platforms

Machine Learning Algorithms Embedded in the Data Management Platforms

“Information Producers”

Data Scientists, R Users, Citizen Data Scientists

New Zeppelin notebook based UI for data scientists collaborating and sharing ML analytical methodologies in Clouds

“Oracle ML” Big Data Edition
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed Execution

“Oracle Machine Learning” Database Edition
Machine Learning Algorithms, Statistical Functions + R Integration for Scalable, Parallel, Distributed, in-DB Execution

Oracle BDA Hadoop

Oracle Database EE

Oracle Cloud

Copyright © 2016, Oracle and/or its affiliates. All rights reserved.
Oracle Advanced Analytics 12.2
Model Build Time Performance

<table>
<thead>
<tr>
<th>OAA 12.2 Algorithms</th>
<th>Rows (Ms)</th>
<th>T7-4 (Sparc &amp; Solaris) Model Build Time (Secs / Degree of Parallelism)</th>
<th>X5-4 (Intel and Linux)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes Importance</td>
<td>640</td>
<td>28s / 512</td>
<td>44s / 72</td>
</tr>
<tr>
<td>K Means Clustering</td>
<td>640</td>
<td>161s / 256</td>
<td>268s / 144</td>
</tr>
<tr>
<td>Expectation Maximization</td>
<td>159</td>
<td>455s / 512</td>
<td>588s / 144</td>
</tr>
<tr>
<td>Naive Bayes Classification</td>
<td>320</td>
<td>17s / 256</td>
<td>23s / 72</td>
</tr>
<tr>
<td>GLM Classification</td>
<td>640</td>
<td>154s / 512</td>
<td>363s / 144</td>
</tr>
<tr>
<td>GLM Regression</td>
<td>640</td>
<td>55s / 512</td>
<td>93s / 144</td>
</tr>
<tr>
<td>Support Vector Machine (IPM solver)</td>
<td>640</td>
<td>404s / 512</td>
<td>1411s / 144</td>
</tr>
<tr>
<td>Support Vector Machine (SGD solver)</td>
<td>640</td>
<td>84s / 256</td>
<td>188s / 72</td>
</tr>
</tbody>
</table>

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 ok 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!
Machine Learning & Advanced Analytical Methodologies

*Data Preparation & Adv. Analytical Process Runs In-Database*

Additional relevant data and “engineered features”

Oracle Database 12c

Historical data

Assembled historical data

Build Predictive Model

Make Predictions

Predictions & Insights

Historical or Current Data to be “scored” for predictions

Sensor data, Text, unstructured data, transactional data, spatial data, etc.
Example Predictive Appl: HCM Cloud—Workforce Predictions
Complete, Integrated, Embedded, Automated and Interactive “Predictive HCM” Solution

• Integrated data management + embedded predictive analytics
• Full 360 degree employee view
• Single source of HCM data data
• Interactive dashboards and “What if” analysis
• Customizable if desired to add input variables to predictive models
• Mobile + Oracle Cloud solutions
Oracle’s Machine Learning/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

Traditional Analytics

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

Oracle Advanced Analytics

- Model “Scoring”
- Embedded Data Prep
- Model Building
- Data Preparation

Secs, Mins or Hours

Hours, Days or Weeks

Savings

Copyright © 2016, Oracle and/or its affiliates. All rights reserved.
Oracle’s Machine Learning/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 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 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.
### Objectives
- Prevent $200M in losses every year using data to monitor, understand and anticipate fraud

### Solution
- We installed OAA analytics for model development during 2014
- When choosing the tools for fraud management, speed is a critical factor
- OAA provided a fast and flexible solution for model building, visualization and integration with production processes

> “When choosing the tools for fraud management, speed is a critical factor. Oracle Advance Analytics provided a fast and flexible solution for model building, visualization and integration with production processes.”

> – Miguel Barrera, Director of Risk Analytics, Fiserv Inc.
> – Julia Minkowski, Risk Analytics Manager, Fiserv Inc.
Ease of Deployment

Data Miner Survey 2016 by Rexer Analytics

While 6 out 10 data miners report the data is available for analysis within days of capture, the time to deploy the models takes substantially longer. For 60% of the respondents, the deployment time will range between 3 weeks and 1 year.

Everyone forgets about deployment – but is most important component!
UK National Health Service
Combating Healthcare Fraud

Objectives
- Use new insight to help identify cost savings and meet goals
- Identify and prevent healthcare fraud and benefit eligibility errors to save costs
- Leverage existing data to transform business and productivity

Solution
- Identified up to GBP100 million (US$156 million) potentially saved through benefit fraud and error reduction
- Used anomaly detection to uncover fraudulent activity where some dentists split a single course of treatment into multiple parts and presented claims for multiple treatments
- Analyzed billions of records at one time to measure longer-term patient journeys and to analyze drug prescribing patterns to improve patient care

"Oracle Advanced Analytics' data mining capabilities and Oracle Exalytics' performance really impressed us. The overall solution is very fast, and our investment very quickly provided value. We can now do so much more with our data, resulting in significant savings for the NHS as a whole"
– Nina Monckton, Head of Information Services, NHS Business Services Authority

Update: £300M confirmed fraud
£400+M additional potential identified

Now moving to Cloud

Oracle Exadata Database Machine
Oracle Advanced Analytics
Oracle Exalytics In-Memory Machine
Oracle Endeca Information Discovery
Oracle Business Intelligence EE
Rapidly Build, Evaluate & Deploy Analytical Methodologies
Leveraging a Variety of Data Sources and Types

SQL Joins and arbitrary SQL transforms & queries – power of SQL

Consider:
• Demographics
• Past purchases
• Recent purchases
• Comments & tweets

Unstructured data also mined by algorithms

Transactional POS data

Modeling Approaches

Generates SQL scripts and workflow API for deployment

Advanced Analytics

Inline predictive model to augment input data

ORACLE
More Data Variety—Better Predictive Models

- Increasing sources of relevant data can boost model accuracy

Model with “Big Data” and hundreds -- thousands of input variables including:
- Demographic data
- Purchase POS transactional data
- “Unstructured data”, text & comments
- Spatial location data
- Long term vs. recent historical behavior
- Web visits
- Sensor data
- etc.

**Engineered Features** – Derived attributes/variable that reflect domain knowledge—key to best models

- Naïve Guess or Random
- Model with 20 variables
- Model with 75 variables
- Model with 250 variables
Big Data Analytics using w Graph
Oracle Advanced Analytics/Machine Learning with Enhanced Graph & Spatial Data Sources

• Add new engineered features
  – Percentage time spent in zones
  – Amount time/encounters with persons of interest

• Better predictions using available data
  – At risk customers
  – Government approval processes
  – Medical claims
  – IoT predictive analytics

Better data and “engineered features”; better predictive models and predictive insights
DX Marketing
Cloud Based Predictive Analytics/Database Marketing

Objectives

- Cloud-based solution
- Increase revenue
- Reduce time-to-market

Solution

- The company considered only two solution vendors --SAS and Oracle to host its consumer data. SAS offered to help build the IT infrastructure from scratch and help develop a one-year plan. But when they looked at the number of personnel needed to manage the infrastructure including administrators, security specialists and analysts as well as Security & HIPAA compliance needed, Oracle’s DBCS solution looked far more attractive. Hence, they decided to go with Oracle. Oracle’s solution offered:
  - Scalability
  - Built in analytical tools including data mining,
  - Built in HIPPA compliance and security features.
  - Required fewer resources —only two analysts —Data Engineer and an expert in Predictive Analytics who now manage the entire ecosystem.

“Time to market has significantly improved from 4-6 weeks to less than a week with the result the company can bring new clients on board faster. This has helped boost revenues by 25% in the six months since using Oracle’s DBCS..”

– DX Marketing

DX Marketing Expands Customer Acquisition with Oracle Cloud – YouTube video
Zagrebačka Bank (biggest bank in Croatia)

Increases Cash Loans by 15% Within 18 Months of Deployment

Objectives

- Needed to speed up entire advanced analytics process; data prep was taking 3 days; model building 24 hours
- Faster time to “actionable analytics” for Credit Risk Modeling and Targeted Customer Campaigns

Solution

- Zaba migrated from SAS to the Oracle Advanced Analytics platform for statistical modeling and predictive analytics
- Increased prediction performance by leveraging the security, reliability, performance, and scalability of Oracle Database and Oracle Advanced Analytics for predictive analytics—running data preparation, transformation, model building, and model scoring within the database

“With Oracle Advanced Analytics we execute computations on thousands of attributes in parallel—impossible with open-source R. Analyzing in Oracle Database without moving data increases our agility. Oracle Advanced Analytics enables us to make quality decisions on time, increasing our cash loans business 15%.”

– Jadranka Novoselovic, Head of BI Dev., Zagrebačka Bank

“We chose Oracle because our entire data modeling process runs on the same machine with the highest performance and level of integration. With Oracle Database we simply switched on the Oracle Advanced Analytics option and needed no new tools,”

– Sinisa Behin, ICT coordinator at BI Dev. Zagrebačka Bank

ZabaBank Oracle Customer Snapshot on OTN
Oracle Data Miner GUI
Easy to Use for “Citizen Data Scientist”

• Easy to use to define analytical methodologies that can be shared
• SQL Developer Extension
• Workflow API and generates SQL code for immediate deployment
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Output</th>
<th>Rank</th>
<th>Importance</th>
<th>% Null</th>
<th>% Unique</th>
<th>% Constant</th>
<th>Hints</th>
</tr>
</thead>
<tbody>
<tr>
<td>BANK_FUNDS</td>
<td>NUMBER</td>
<td>1</td>
<td>1</td>
<td>0.2039</td>
<td>0</td>
<td>21.3511</td>
<td>38.5016</td>
<td></td>
</tr>
<tr>
<td>N_TRANS_ATM</td>
<td>NUMBER</td>
<td>2</td>
<td>2</td>
<td>0.1217</td>
<td>0</td>
<td>21.499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_TRANS_TELLER</td>
<td>NUMBER</td>
<td>3</td>
<td>3</td>
<td>0.4911</td>
<td>0</td>
<td>32.5471</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONEY_MONTHLY_OVERDRAWN</td>
<td>NUMBER</td>
<td>4</td>
<td>4</td>
<td>0.1207</td>
<td>0</td>
<td>18.787</td>
<td>16.3215</td>
<td></td>
</tr>
<tr>
<td>T_AMOUNT_AUTOM_PAYMENTS</td>
<td>NUMBER</td>
<td>5</td>
<td>5</td>
<td>0.1054</td>
<td>0</td>
<td>21.5483</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONTHLY_CHECKS_WRITTEN</td>
<td>NUMBER</td>
<td>6</td>
<td>6</td>
<td>0.0848</td>
<td>0</td>
<td>9.3699</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_OF_DEPENDENTS</td>
<td>NUMBER</td>
<td>7</td>
<td>7</td>
<td>0.0315</td>
<td>0</td>
<td>3.452</td>
<td>34.4108</td>
<td></td>
</tr>
<tr>
<td>TIME_AS_CUSTOMER</td>
<td>NUMBER</td>
<td>8</td>
<td>8</td>
<td>0.0221</td>
<td>0</td>
<td>2.465</td>
<td>31.3116</td>
<td></td>
</tr>
<tr>
<td>CHECKING_AMOUNT</td>
<td>NUMBER</td>
<td>9</td>
<td>9</td>
<td>0.0183</td>
<td>0</td>
<td>30.3254</td>
<td>62.7272</td>
<td></td>
</tr>
<tr>
<td>CREDIT_BALANCE</td>
<td>NUMBER</td>
<td>10</td>
<td>10</td>
<td>0.0135</td>
<td>0</td>
<td>9.4181</td>
<td>90.5819</td>
<td></td>
</tr>
<tr>
<td>N_TRANS_KIOSK</td>
<td>NUMBER</td>
<td>11</td>
<td>11</td>
<td>0.0004</td>
<td>0</td>
<td>5.2424</td>
<td>43.0014</td>
<td></td>
</tr>
<tr>
<td>MORTGAGE_AMOUNT</td>
<td>NUMBER</td>
<td>12</td>
<td>12</td>
<td>0.0073</td>
<td>0</td>
<td>21.2525</td>
<td>23.1262</td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td>VARCHAR2</td>
<td>13</td>
<td>13</td>
<td>0.0064</td>
<td>0</td>
<td>0.9896</td>
<td>0.1656</td>
<td></td>
</tr>
<tr>
<td>MARITAL_STATUS</td>
<td>VARCHAR2</td>
<td>14</td>
<td>14</td>
<td>0.0058</td>
<td>0</td>
<td>0.2486</td>
<td>34.5168</td>
<td></td>
</tr>
<tr>
<td>HOUSE_OWNERSHIP</td>
<td>VARCHAR2</td>
<td>15</td>
<td>15</td>
<td>0.0004</td>
<td>0</td>
<td>0.1497</td>
<td>71.1536</td>
<td></td>
</tr>
<tr>
<td>N_MORTGAGES</td>
<td>NUMBER</td>
<td>16</td>
<td>16</td>
<td>0.0014</td>
<td>0</td>
<td>0.147</td>
<td>74.4592</td>
<td></td>
</tr>
</tbody>
</table>
ORACLE®
DATA VISUALIZATION

Copyright © 2015, 2018 Oracle and/or its affiliates. All rights reserved.
PREDICTION_COUNT by PREDICTION, FULL_SIMPLE_RULE

PREDICTION_COUNT 751.00
FULL_SIMPLE_RULE (BANK_FUNDS > 248) AND (CHECKING_AMOUNT <= 282) AND (CREDIT_BALANCE <= 2445) AND (MONEY_MONTHLY_OVERDRAWN >= 54.095) AND (T_AMOUNT_AUTOM_PAYMENTS <= 14993)
Sharing, Automation and Deployment
Immediately Go to “Productionization” of Analytical Methodologies

• Share ODMr workflows
• Workflow API for 100% automation
  • Immediate deployment of data analyst’s methodologies
• SQL Script Generation
  • Deploy methodology as SQL scripts
Fraud Prediction Demo

Automated In-DB Analytical Methodology

drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values ('ALGO_NAME','ALGO_SUPPORT_VECTOR_MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;
begin
  dbms_data_mining.create_model('CLAIMSMODEL', 'CLASSIFICATION', 'CLAIMS', 'POLICYNUMBER', null, 'CLAIMS_SET');
end;
/

Top 5 most suspicious fraud policy holder claims
select * from
(select POLICYNUMBER, round(prob_fraud*100,2) percent_fraud,
   rank() over (order by prob_fraud desc) rnk from
(select POLICYNUMBER, prediction_probability(CLAIMSMODEL, '0' using *) prob_fraud from CLAIMS
where PASTNUMBEROFCLAIMS in ('2to4', 'morethan4'))
where rnk <= 5
order by percent_fraud desc;

Automated Monthly “Application”! Just add:
Create View CLAIMS2_30 As
Select * from CLAIMS2
Where mydate > SYSDATE – 30

Time measure: set timing on;
Oracle Advanced Analytics
Real-Time Scoring, Predictions and Recommendations

• On-the-fly, single record apply with new data (e.g. from call center)

Select `prediction_probability(CLAS_DT_1_5, 'Yes'
    USING 7800 as bank_funds, 125 as checking_amount, 20 as
credit_balance, 55 as age, 'Married' as marital_status,
250 as MONEY_MONLY_OVERDRAWN, 1 as house_ownership)
from dual;

Likelihood to respond:

<table>
<thead>
<tr>
<th>Query Result</th>
<th>All Rows Fetched: 1 in 0 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICTION_PROB...</td>
<td>0.8382936507936...</td>
</tr>
</tbody>
</table>
R: Transparency via function overloading

Invoke in-database aggregation function

```
> aggdata <- aggregate(ONTIME_S$DEST,
+             by = list(ONTIME_S$DEST),
+             FUN = length)
> class(aggdata)
[1] "ore.frame"
attr(,"package")
[1] "OREbase"
> head(aggdata)
  Group.1 x
  1 ABE 237
  2 ABI 34
  3 ABQ 1357
  4 ABY 10
  5 ACK 3
  6 ACT 33
```
R: Transparency via function overloading

Invoke in-database Data Mining model (Support Vector Machine)

```r
> svm_mod <- ore.odmSVM(BUY~INCOME+YRS_CUST+MARITAL_STATUS, data=CUST,
                         "classification", kernel="linear")
> summary(svm_mod)

Call:
ore.odmSVM(formula = BUY ~ INCOME + YRS_CUST + MARITAL_STATUS, data = CUST,
            type = "classification", kernel.function = "linear")

Settings:

Value
prep.auto on
active.learning al.enable
complexity.factor 46.044899
conv.tolerance 1e-04
kernel.function linear

Coefficients:

                   value     estimate
class variable     1       0         INCOME 5.204561e-05
2 MARITAL_STATUS     M -4.531359e-05
3 MARITAL_STATUS     S  4.531359e-05
4       YRS_CUST        1.264948e-04
5 (Intercept)      9.999269e-01
6       1         INCOME 2.032340e-05
7 MARITAL_STATUS     M  2.636552e-06
8 MARITAL_STATUS     S -2.636555e-06
9       YRS_CUST -1.588211e-04
10      1 (Intercept) -9.999324e-01
```

Oracle Database

Oracle PL/SQL

BEGIN
DBMS_DATA_MINING.CREATE_MODEL(
model_name => 'SVM_MOD',
mining_function => dbms_data_mining.classification
...
data(pistormings)
head(pistormings)
occ(diameter < 125, type="wbar")
hist(CARSTATS$MPG, col="red", breaks=25)
plot(CARSTATS, col="red")
hist(CARSTATS$MPG, col="red", breaks=25)
plot(CARSTATS, col="red")
hist(CARSTATS$MPG, col="red", breaks=25)
plot(CARSTATS, col="red")
hist(CARSTATS$MPG, col="red", breaks=25)
plot(CARSTATS, col="red")
hist(CARSTATS$MPG, col="red", breaks=25)
plot(CARSTATS, col="red")

library(OCR)

ore.connect("dmuser", "ora12c", "localhost", "dmuser", all=TRUE)

ore.ls()

Help.start()

names(CARSTATS)

summary(CARSTATS)

hist(CARSTATS$MPG, col="red", breaks=25)

plot(CARSTATS)

plot(CARSTATS, col="red")

library(OCR)

ore.connect("dmuser", "ora12c", "localhost", "dmuser", all=TRUE)

hist(CARSTATS$MPG, col="red", breaks=25)

ore.ls()

Help.start()

summary(CARSTATS)

plot(CARSTATS)

demo(oem_svm)

demo(oem_kmeans)

names(DM)
Invoke ORAAH custom parallel distributed GLM Model using Spark Caching

ORAAH: Machine Learning in Spark against HDFS data

Connects to Spark
spark.connect("yarn-client", memory="24g")

# Attaches the HDFS file for use within R
ont1bi <- hdfs.attach("/user/oracle/ontime_lbi")

# Formula definition: Cancelled flights (0 or 1) based on other attributes
form_oraah_glm2 <- CANCELLED ~ DISTANCE + ORIGIN + DEST + F(YEAR) + F(MONTH) +
+ F(DAYOFMONTH) + F(DAYOFWEEK)

system.time(m_spark_glm <- orch glm2(formula=form_oraah_glm2, ont1bi))

ORCH GLM: processed 6 factor variables, 25.806 sec
ORCH GLM: created model matrix, 100128 partitions, 32.871 sec
ORCH GLM: iter 1, deviance 1.38433414089348300E+09, elapsed time 9.582 sec
ORCH GLM: iter 2, deviance 3.39315388583931150E+08, elapsed time 9.213 sec
ORCH GLM: iter 3, deviance 2.06855738812683250E+08, elapsed time 9.218 sec
ORCH GLM: iter 4, deviance 1.75868100359263200E+08, elapsed time 9.104 sec
ORCH GLM: iter 5, deviance 1.70023181759611580E+08, elapsed time 9.124 sec
ORCH GLM: iter 6, deviance 1.69476890425481350E+08, elapsed time 9.124 sec
ORCH GLM: iter 7, deviance 1.69467574351380850E+08, elapsed time 9.077 sec
ORCH GLM: iter 8, deviance 1.69467574351380850E+08, elapsed time 9.164 sec

user system elapsed
84.107 5.606 143.591
Oracle Advanced Analytics 12.2, Oracle Data Miner 4.2 and ORAAH 2.7

New Features + Road Map
Oracle Advanced Analytics 12.2
New Oracle Database Features

• 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

• 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

• 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
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**
New Oracle Database Features

• 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
New Oracle Database Features

• 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

ESA vs. LDA (Latent Dirichlet Allocation); ESA is more interpretable than LDA

Topics discovered by LDA are latent, meaning difficult to interpret

- Topics are defined by their keywords, i.e., they have no names, no abstract descriptions
- To give meaning to topics, keywords can be extracted by LDA
- Definitions solely based on keywords are fuzzy, and keywords for different topics usually overlap
- Extracted keywords can be just generic words
- Set of automatically extracted keywords for a topic does not map to a convenient English topic name

Biggest LDA problem; set of topics is fluid

- Topic set changes with any changes to the training data
- Modification of training data changes topic boundaries
- Training data is almost never static

ESA discovers topics from a given set of topics in a knowledge base

- Topics are defined by humans → topics are well understood.
- Topic set of interest can be selected and augmented if necessary → full control of the selection of topics
- Set of topics can be geared toward a specific task, e.g., knowledge base for topic modeling of online messages possibly related to terrorist activities, which is different than one for topic modeling of technical reports from academia
- Can combine multiple knowledge bases, each with its own topic set, which may or may not overlap
- Topic overlapping does not affect ESA’s capability to detect relevant topics
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’
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
  – Supports data with “nested” attributes, handling text and aggregated transactional data for open source R packages
  – Extends ease of advanced analytics development from R to Oracle Database
  – 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
Oracle R Advanced Analytics for Hadoop

New Features in ORAAH 2.7

• Updated ORAAH GLM and LM algorithms which are much faster, stable and light on memory than comparable GLM and LM methods from Spark Mllib

• Both methods also bring a new summary feature that makes them comparable to solutions from open-source R glm and lm, but capable of handling Big Data at enterprise scale

• The Neural Networks algorithm has been enhanced to support the full formula processing and a full build and scoring in Spark

• The new Gaussian Mixture Models is an addition to the set of algorithms supported in Spark Mllib

• ORAAH's Spark-based LM with full formula support and summary - orch.lm2()
• ORAAH's Spark-based GLM with full formula support and summary - orch.glm2()
Oracle Advanced Analytics Strategy & Road Map

• One server side product, with a single analytic library, supporting multiple data platforms, analytical engines, UIs and deployment strategies.
What is a “notebook”
Oracle Machine Learning
Multi-Language, Multi-Server Engine Oracle Machine Learning for Clouds

Key Features

• Collaborative ML environment for data scientists
  – Shared Zeppelin notebooks, templates, and permissions

• Language—SQL ML algorithms API (ODM)

• DWCS server—Oracle Database

• Supports deployment of ML analytics solutions
  – Enables publishing libraries, templates, use cases

• Road map
  • Multi-Language support
    – R language
  • Multi-Server Engines
    – R, ORE, ORAAH, Spark
Oracle’s Advanced Analytics
Predictive Applications + OBIEE Integration
Enabling “Predictive” Enterprise Applications
Oracle Applications Using Oracle Advanced Analytics—Partial List

• **Oracle HCM Cloud**
  – Employee turnover and performance prediction and “What if?” analysis

• **Oracle Sales Cloud**
  – Prediction of sales opportunities, what to sell, amount, timing, etc.

• **Oracle Industry Data Models**
  – **Communications Data Model** churn prediction, segmentation, profiling, etc.
  – **Retail Data Model** loyalty and market basket analysis
  – **Airline Data Model** analysis frequent flyers, loyalty, etc.
  – **Utilities Data Model** customer churn, cross-sell, loyalty, etc.

• **Oracle Retail GBU Cloud Services**
  – Market Basket Analysis Insights
  – Customer Insights & Clustering

• **Oracle Customer Support**
  – Predictive Incident Monitoring (PIM)

• **Oracle Spend Classification**
  – Real-time and batch flagging of noncompliance and anomalies in expense submissions

• **Oracle FinServ Analytic Applications**
  – Customer Insight, Enterprise Risk Management, Enterprise Performance, Financial Crime and Compliance

• **Oracle Adaptive Access Manager**
  – Real-time security and fraud analytics
Human Capital Management Powered by OAA

- Oracle Advanced Analytics factory-installed predictive analytics
- Employees likely to leave and predicted performance
- Top reasons, expected behavior
- Real-time "What if?" analysis

Link to Oracle HCM on O.com
HCM Predictive Workforce demo
Performance and Voluntary Termination Predictions

Let’s Walk Through Again But Go More Slowly...

- Predicted Voluntary Termination for Worker
- Predicted Performance for Individual Worker
Predicted Performance for Individual Worker

**Prediction Details: Gloria Daas**

- **Name**: Gloria Daas
- **Manager**: Mitch Blum
- **Current Performance Rating**: 2-Inconsistent
- **Predicted Voluntary Termination**: 63%
- **Predicted Performance**: 79%

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Predicted Performance Reason</th>
<th>Current Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current performance rating</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Latest salary change</td>
<td></td>
<td>2.9 %</td>
</tr>
<tr>
<td>Time with current manager</td>
<td></td>
<td>146.2 Months</td>
</tr>
<tr>
<td>Current grade</td>
<td></td>
<td>Admin03</td>
</tr>
<tr>
<td>Time since last sickness</td>
<td></td>
<td>1.42 Months</td>
</tr>
<tr>
<td>Worker's performance compared to peers</td>
<td></td>
<td>58.05</td>
</tr>
<tr>
<td>Number of manager changes in the last 12 months</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Current or most recent manager</td>
<td></td>
<td>Mitch Blum</td>
</tr>
<tr>
<td>Time in current grade</td>
<td></td>
<td>56.52 Months</td>
</tr>
</tbody>
</table>
HCM Predictive Workforce
Predictive Analytics Applications

Fusion Human Capital Management
Powered by OAA

- Oracle Advanced Analytics factory-installed predictive analytics
- Employees likely to leave and predicted performance
- Top reasons, expected behavior
- Real-time "What if?" analysis

Link to Oracle HCM on O.com
HCM Predictive Workforce demo
Sales Predictor helps sales reps answer critical sales questions:

- Which products should be offered to a customer?
- Who are the customers buying products?
- What are the reasons a product is being bought?

Sales Predictor offers product recommendations that have a higher likelihood of being converted to a win.
Pre-Built Market Basket Analysis

- Gain actionable insight into your shoppers' behavior.
- Pre-built market-basket analysis identifies product affinities

Oracle Retail Market Basket Insights Cloud Service

[Diagram showing data flow from Sales and Products to Customer Count and Net Sales Amt]

Link to Oracle Retail MBA on O.com
Oracle Retail Customer Insights Cloud Service
Customer Segmentation/Clustering Analysis

Pre-Built Customer Clustering Models

• Gain actionable insight into your customer’s behavior.
• Pre-built clustering models identify hidden customer segments

Link to Oracle Retail CI Cloud on O.com
Oracle Communications Industry Data Model

Example Predictive Analytics Application

Pre-Built Predictive Models

- Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics
- OAA’s clustering and predictions available in-DB for OBIEE
- Automatic Customer Segmentation, Churn Predictions, and Sentiment Analysis

Link to OCDM on OTN
Getting started
Getting started: Oracle’s AA/ML Links and Resources

Oracle Advanced Analytics Overview Information
- Oracle's Machine Learning and Advanced Analytics 12.2c and Oracle Data Miner 4.2 New Features presentation
- Oracle Advanced Analytics Public Customer References
- Big Data Analytics with Oracle Advanced Analytics: Making Big Data and Analytics Simple white paper on OTN
- Oracle INTERNAL ONLY OAA Product Management Wiki and Beehive Workspace

YouTube recorded Oracle Advanced Analytics Presentations and Demos, White Papers
- Oracle's Machine Learning & Advanced Analytics 12.2 & Oracle Data Miner 4.2 New Features YouTube video
- Library of YouTube Movies on Oracle Advanced Analytics, Data Mining, Machine Learning (7+ “live” Demos e.g. Oracle Data Miner 4.0 New Features, Retail, Fraud, Loyalty, Overview, etc.)
- Overview YouTube video of Oracle’s Advanced Analytics and Machine Learning

Getting Started/Training/Tutorials
- Link to OAA/Oracle Data Miner Workflow GUI Online (free) Tutorial Series on OTN
- Link to OAA/Oracle R Enterprise (free) Tutorial Series on OTN
- Link to Try the Oracle Cloud Now!
- Link to Getting Started w/ ODM blog entry
- Link to New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course.
- Oracle Data Mining Sample Code Examples

Additional Resources, Documentation & OTN Discussion Forums
- Oracle Advanced Analytics Option on OTN page
- OAA/Oracle Data Mining on OTN page, ODM Documentation & ODM Blog
- OAA/Oracle R Enterprise page on OTN page, ORE Documentation & ORE Blog
- Oracle SQL based Basic Statistical functions on OTN
- Oracle R Advanced Analytics for Hadoop (ORAAH) on OTN

BIWA SIG User Community www.biwasummit.org
- Business Intelligence, Warehousing & Analytics—BIWA SUMMIT 2018 WITH SPATIAL SUMMIT
January 30 - February 1, 2018, Redwood Shores, CA (contains links to past presentations)
BIWA SUMMIT 2018 WITH SPATIAL SUMMIT

January 30 - February 1, 2018

THE Big Data + Analytics + Spatial + Cloud + IoT + Everything Cool User Conference

www.biwasummit.org