Take Your Analytics to the Next Level of Insight using Machine Learning in the Oracle Autonomous Data Warehouse

Charlie Berger, MS Engineering, MBA
Sr. Director Product Management, Machine Learning, AI and Cognitive Analytics
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.
Autonomous Data Warehouse Cloud

• Easy
  – Fully-managed, pre-configured and optimized for DW workloads
  – Simply load data and run
    • No need to define indexes, create partitions, etc.

• Fast
  – Based on Exadata technology

• Elastic
  – Instant scaling of compute or storage with no downtime

• Powerful: Machine Learning included
  – Library of ML algorithms implemented as fully parallelized SQL functions
ADW + Oracle Machine Learning Notebooks
Powerful In-Database Machine Learning and Analytics

1. Extract value using SQL analytics
   - Hierarchical Analytics
   - Summary & Descriptive Statistics
   - SQL Windowing Functions
   - Tests for Statistical Correlations
   - Approximate Analytics
   - Pattern Matching
   - SQL Models
   - Advanced Aggregations
   - Ranking
   - Pivoting
   - Used-Defined PTFs
   - Text Analytics

2. Gain insights, make predictions via ML
   - Classification / Prediction
   - Regression
   - Anomaly Detection
   - Attribute Importance
   - Association Rules / Market Basket Analysis
   - Clustering
   - Feature Extraction / Selection
   - Time Series / Forecasting
   - Ensemble Models
   - Predictive Queries
   - Text Mining
   - Cognitive Text Analytics
   - Ensemble Models
   - Predictive Queries
   - Text Mining

Copyright © 2018, Oracle and/or its affiliates. All rights reserved.
What is Machine Learning?

*Automatically* sift through *large amounts* of data to find hidden patterns, discover new insights and make predictions

- Identify most important factor *(Attribute Importance)*
- Predict customer behavior *(Classification)*
- Predict or estimate a value *(Regression)*
- Find profiles of targeted people or items *(Decision Trees)*
- Segment a population *(Clustering)*
- Find fraudulent or “rare events” *(Anomaly Detection)*
- Determine co-occurring items in a “baskets” *(Associations)*
ADWC Machine Learning Algorithms

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

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

**ANOMALY DETECTION**
- One-Class SVM

**TIME SERIES**
- Holt-Winters, Regular & Irregular, with and w/o trends & seasonal
- Single, Double Exp Smoothing

**REGRESSION**
- Linear Model
- Generalized Linear Model
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network

**ATTRIBUTE IMPORTANCE**
- Minimum Description Length
- Principal Comp Analysis (PCA)
- Unsupervised Pair-wise KL Div

**ASSOCIATION RULES**
- A priori/ market basket

**PREDICTIVE QUERIES**
- Predict, cluster, detect, features

**SQL ANALYTICS**
- SQL Windows, SQL Patterns, SQL Aggregates

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

**STATISTICAL FUNCTIONS**
- Basic statistics: min, max, median, stdev, t-test, F-test, Pearson’s, Chi-Sq, ANOVA, etc.

*OAA includes support for Partitioned Models, Transactional, data, etc.*
Oracle’s Data Management and Machine Learning

Architectural Strategy

• In-Database proprietary implementations of machine learning algorithms
• Leverage strengths of the Database and adds new ML tech
  – Counting, conditional probabilities, sort, rank, partition, group-by, collections, etc.
  – Parallel execution, bitmap indexes, partitioning, aggregations, recursion w/in parallel infrastructure, IEEE float, frequent itemsets, Automatic Data Preparation (ADP), Text processing, etc.
• Focus on intelligent ML defaults, simplification & automation to enable applications
  – ADP, xforms, binning, missing values, Prediction_Details, Predictive_Queries, Model_views
• Machine learning models built via PL/SQL script; scored via SQL functions (1st class DB objects)
  ```sql
  select cust_id
  from customers
  where region = 'US'
  and prediction_probability(churnmod, 'Y' using *) > 0.8;
  ```
  – “Smart scan” ML model scoring “push down”; Supports OLTP and ATPC environments
• Machine Learning and Advanced Analytics are peer to rest of Oracle Data Mgmt features
  – Best ML & analytical development and deployment platform

True power evident when scoring models using SQL functions, e.g.
Oracle’s Machine Learning & Advanced Analytics

Fastest Way to Deliver Enterprise-wide 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 ML

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

### Oracle’s in-DB Machine Learning

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

### Time Comparison

- Hours, Days or Weeks
- Secs, Mins or Hours

### Savings

- Fastest Way to Deliver Enterprise-wide Predictive Analytics
Multiple Data Scientist User Roles Supported

Oracle’s Machine Learning/Advanced Analytics

DBAs  Application Developers

New! RStudio  python

R, Python Users, Data Scientists  OML Notebook Users

Data Analysts, Citizen Data Scientists

Oracle’s Machine Learning/Advanced Analytics

Historical or Current Data to be “scored” for predictions

Historical data

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

Copyright © 2018, Oracle and/or its affiliates. All rights reserved.
OAA Model Build and Real-time SQL Apply Prediction

Simple SQL Syntax

ML Model Build (PL/SQL)

```plsql
begin
    dbms_data_mining.create_model('BUY_INSURL', 'CLASSIFICATION',
        'CUST_INSUR_LTV', 'CUST_ID', 'BUY_INSURANCE', CUST_INSUR_LTV_SET);
end;
/
```

Model Apply (SQL query)

```sql
Select prediction_probability('BUY_INSURL', 'Yes'
    USING 3500 as bank_funds, 825 as checking_amount, 400 as credit_balance, 22 as age, 'Married' as marital_status, 93 as MONEY_MONTLY_OVERDRAWN, 1 as house_ownership)
from dual;
```

Oracle Database

Copyright © 2018, Oracle and/or its affiliates. All rights reserved.
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;
ML Model Deployment for Real-Time Scoring

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_15, 'Yes'
    USING 7800 as bank_funds, 125 as checking_amount, 20 as credit_balance, 55 as age, 'Married' as marital_status,
    250 as MONEY_MONTLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```

Likelihood to respond:

<table>
<thead>
<tr>
<th>Query Result</th>
<th>SQL</th>
<th>All Rows Fetched: 1 in 0 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Query Result" /></td>
<td><img src="image" alt="SQL" /></td>
<td><img src="image" alt="All Rows Fetched" /></td>
</tr>
</tbody>
</table>

Oracle Cloud
Manage and Analyze All Your Data

Boil down the Data Lake

**Big Data SQL / R**

“Engineered Features”
– Derived attributes that reflect domain knowledge—key to best models e.g.:
  - Counts
  - Totals
  - Changes over time

Object Store
NoSQL
Oracle
kafka

Copyright © 2018, Oracle and/or its affiliates. All rights reserved.
Oracle’s Data Management and Machine Learning

Market Observations:

• Machine learning, predictive analytics & “AI” now *must-have* requirements
• Separate islands for data management and data science just don’t work
• Enterprises whose data science teams most rapidly extract insights and predictions win

Conclusions:

• Must “operationalize” ML insights and predictions throughout enterprise
• Multilingual Machine Learning: SQL, R, Python, Workflow UI, Notebooks, Embed ML in Apps
• Evolving towards *combined* Data Management + Machine Learning environment that can essentially to manage and “think” about data
Introducing:
Oracle Machine Learning SQL Notebooks
Oracle Machine Learning

Machine Learning Notebook for Autonomous Data Warehouse Cloud

Key Features

• Collaborative UI for data scientists
  – Packaged with Autonomous Data Warehouse
  – Easy access to shared notebooks, templates, permissions, scheduler, etc.

• Supports development of ML methodologies in-ADW
  – SQL ML algorithms API
  – Predictions, churn, cross-sell, fraud, associations, statistics, correlations, forecasting, identify clusters, etc.
Oracle Machine Learning Quick DEMO
Sign In

Tenant: TENANT1
Database: PDB1
Username: CBERGER
Password: **********

Sign In
| CUST_ID | CUST_FIRST_NAME | CUST_LAST_NAME | CUST_GENDER | CUST_YEAR_OF_BIRTH | CUST_MARITAL_STATUS | CUST_STREET_ADDRESS   | CUST_POSTAL_CODE | CUST_CITY       | CUST_CITY_ID | CUST_STATE_PROVINCE | CUST_STATE_PROVINCE |
|---------|----------------|----------------|-------------|-------------------|---------------------|----------------------|-------------------|----------------|---------------|----------------|---------------------|---------------------|
| 37,057  | Bernard        | Wright         | M           | 1,941             | married             | 107 East Catano Avenue | 66,361           | Vebert         | 52,436        | Noord-Brabant    | 52,682               |
| 40,612  | Bernard        | Wright         | M           | 1,941             | married             | 107 South Prentiss Avenue | 33,866           | Bergen op Zoom | 51,191        | Noord-Brabant    | 52,682               |
| 44,167  | Bernard        | Wright         | M           | 1,941             | single              | 117 North Door Avenue  | 83,601           | San Francisco  | 52,289        | CA              | 52,567               |
| 34,860  | Bernard        | Wright         | M           | 1,955             | married             | 17 North Lehigh Court  | 59,882           | Malaga         | 51,864        | Malaga          | 52,661               |
| 47,946  | Bernard        | Wright         | M           | 1,957             | married             | 27 West Baraga Boulevard | 46,864           | Lauret         | 51,768        | Languedoc-Roussil | 52,645               |
| 1,502   | Bernard        | Wright         | M           | 1,945             | single              | 37 Mountain View Street | 80,041           | Wolverhampton  | 52,514        | England - West Midlands | 52,503               |
| 3,057   | Bernard        | Wright         | M           | 1,947             | single              | 37 South Catano Street | 34,216           | Munau          | 51,934        | Bayern          | 52,561               |
| 8,612   | Bernard        | Wright         | M           | 1,939             |                     | 47 West Prentiss Road  | 72,059           | Los Angeles    | 51,806        | CA              | 52,567               |
### Notebooks

<table>
<thead>
<tr>
<th>Name</th>
<th>Last Update</th>
<th>Updated By</th>
<th>Connection Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly Detection</td>
<td>2/8/18 1:37 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>Association Rules</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>Classification Prediction Model</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>Clustering</td>
<td>2/8/18 12:59 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>My First Notebook</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>Regression _1</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>SQL Query Scratchpad</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>SQL Script Scratchpad</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
<tr>
<td>SQL Statistical Functions</td>
<td>2/8/18 1:00 PM</td>
<td>CBERGER</td>
<td>Global</td>
</tr>
</tbody>
</table>
Simple Oracle Machine Learning notebook example

Oracle Machine Learning example notebook for learning basic functions using SH schema data and highlights basic data selection and data viewing using the Oracle Autonomous Data Warehouse Cloud (ADWC).

By Charlie Berger


Show all tables

```
SELECT * FROM all_tables WHERE owner = 'SH';
```

<table>
<thead>
<tr>
<th>OWNER</th>
<th>TABLE_NAME</th>
<th>TABLESPACE_NAME</th>
<th>CLUSTER_NAME</th>
<th>OWNER_TOT_NAME</th>
<th>STATUS</th>
<th>PCT_FREE</th>
<th>PCT_USED</th>
<th>IN_TRANS</th>
<th>MAX_TRANS</th>
<th>INITIAL_EXTENT</th>
<th>NEXT_EXTENT</th>
<th>MIN_EXTENTS</th>
<th>MAX_EXTENTS</th>
<th>PCT_INCREASE</th>
<th>FREELISTS</th>
<th>FREELIST_GROUPS</th>
<th>LOGGING</th>
<th>BACK_UP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SH</td>
<td>SALES</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>TIMES</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>CHANNELS</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>PROMOTIONS</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>CUSTOMERS</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>COUNTRIES</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>SUPPLEMENTARY_DEMographics</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>SH</td>
<td>SALES_TRANSACTIONS_EXT</td>
<td>SYSTEM</td>
<td></td>
<td></td>
<td>VALID</td>
<td>10</td>
<td>40</td>
<td>1</td>
<td>255</td>
<td>65,536</td>
<td>1,048,576</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,147,483,645</td>
<td>1</td>
<td>1</td>
<td>YES</td>
</tr>
</tbody>
</table>
Example Templates

Anomaly Detection
This notebook shows how to detect...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘Anomaly Detection’ ‘Machine...

Association Rules
Notebook to show the use of Assoc...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘SQL’ ‘Associations’ ‘Rules’ ‘M...

Attribute Importance
Notebook to identify key attributes...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘SQL’ ‘Attribute Importance’ ‘K...

Classification Prediction M...
Example notebook to predict custo...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘Classification’ ‘Prediction’ ‘De...

Clustering
This notebook shows how to identi...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘Clustering’ ‘K-Means’ ‘Expect...

My First Notebook
Oracle Machine Learning example ...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘SQL’ ‘Data’ ‘Graph’

Regression
This notebook shows how to predic...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘Regression’ ‘SVM’ ‘GLM’ ‘Logi...

Statistical Function
Oracle Machine Learning example ...

Author:
Date Added: 5/4/18 6:59 AM
Tags: ‘Statistics’ ‘ANOVA’ ‘T-test’ ‘F-

Anomaly Detection to Detect Suspicious or Rare Occurrences

This notebook shows how to detect rare records, customers, or transactions using an unsupervised learning algorithm (1-Class Support Vector Machine). The notebook first builds a 1-Class SVM model and then applies the model to flag unusual or suspicious records. The anomaly detection model can also be applied to “score” new records. The entire machine learning methodology runs inside the Oracle Autonomous Data Warehouse Cloud (ADWC).

By Charlie Berger


Clean up and drop any table if previously exists for notebook reproducibility

```sql
BEGIN
EXECUTE IMMEDIATE 'DROP TABLE SUPPLEMENTARY_DEMOGRAPHICS2';
EXCEPTION
WHEN OTHERS THEN NULL;
END;

PL/SQL procedure successfully completed.
```

Create SUPPLEMENTARY_DEMOGRAPHICS2 table that remove COMMENTS unstructured data for simplicity.

```sql
CREATE TABLE SUPPLEMENTARY_DEMOGRAPHICS2 AS
(SELECT AFFINITY CARD, BOOKKEEPING_APPLICATION, BULK_PACK_DISKETTES, CUST ID, EDUCATION, FLAT PANEL MONITOR, HOME THEATER PACKAGE, HOUSEHOLD SIZE, OCCUPATION, OS_DOC_SET KANJI, PRINTER SUPPLIES, YRS RESIDENCE, Y BOX GAMES
FROM SH SUPPLEMENTARY_DEMOGRAPHICS);
```

Updated 4500 row(s).
Predicting Target Customers using Classification

Example notebook to predict customers most likely to be positive responders to an Affinity Card loyalty program. This notebook builds and applies classification models (decision tree) using the SH schema data and processed inside the Oracle Autonomous Data Warehouse Cloud (ADWC).

By Charlie Berger


Display the SH.SUPPLEMENTARY_DEMOGRAPHICS data

<table>
<thead>
<tr>
<th>CUST_ID</th>
<th>EDUCATION</th>
<th>OCCUPATION</th>
<th>HOUSEHOLD_SIZE</th>
<th>YRS_RESIDENCE</th>
<th>AFFINITY_CARD</th>
<th>BULK_PACK_DISKETTES</th>
<th>FLAT_PANEL_MONITOR</th>
<th>HOME_THEATER_PACKAGE</th>
<th>BOOKKEEPING_APPLICATION</th>
<th>PRINTER_SUPPLIES</th>
<th>Y_BOX_GAMES</th>
<th>OS_DOC_SET_KANJI</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.001</td>
<td>&lt; Bach</td>
<td>Crafts</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Sh</td>
</tr>
<tr>
<td>100.002</td>
<td>HS-grad</td>
<td>Machine</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Al</td>
</tr>
<tr>
<td>100.003</td>
<td>HS-grad</td>
<td>Sales</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Al</td>
</tr>
<tr>
<td>100.004</td>
<td>Bach</td>
<td>?</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Al</td>
</tr>
<tr>
<td>100.005</td>
<td>&lt; Bach</td>
<td>Other</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Gr</td>
</tr>
<tr>
<td>100.006</td>
<td>9th</td>
<td>Crafts</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Al</td>
</tr>
<tr>
<td>100.007</td>
<td>HS-grad</td>
<td>Crafts</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Al</td>
</tr>
<tr>
<td>100.008</td>
<td>HS-grad</td>
<td>Farming</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Al</td>
</tr>
</tbody>
</table>
Targeting Likely Good Credit Customers using Oracle Machine Learning’s (OML) Classification Models

Heather has spent most of her time over the past couple of years extracting and preparing data for analysis. The large volumes of data need extracting and processing mean she spends most of her time waiting for jobs to finish and very little of her time analyzing the data. Demands from marketing are forcing a new approach whereby the data remains in the data warehouse and is processed there. The alternative cloud solution is more complex, and has no direct out of the box processes to analyze the data in place. She started taking a look at Oracle, and found the simple SQL commands in ADWC are familiar, and execute extremely fast, leveraging all the performance features of the platform. Further once she is done can apply the learning models to incoming data on the fly, and allow end user analysts to immediately see mining results. This drastically reduces the cycle of data preparation, analysis, and publishing. It also means there is no change to analysis/reporting Data Visualization toolset that users are familiar with.

Scroll down this notebook and learn how to use OML to create predictive perspectives on data in ADWC, WITHOUT moving it. We will process a small 100k data set, but could use a 100M or billion row data set without worrying about processing time.

This is an extract of Alphaoffice customer information. We will first get acclimated to Apache Zeppelin, the open source interface for interactive collaboration in a team environment.

The Business Problem:
Increase Sales by Targeting our Best Customers; Good Credit Customers!

Heather has a hunch that weakening sales may be due to the company selling to non-optimal customers; customers who
STEP 6: Review Data by Occupation

```sql
-- This shows an alternative presentation style - a pie chart. Note that Zeppelin visualizations are limited. In lab 400 we will use Oracle Data Visualization to create more more interesting perspectives.

select customer_id, age, income, tenure, loan_type, loan_amount, occupation, marital_status
from credit_scoring_100k_v where rownum < 1000
```

Another Pie Chart Example

```sql
-- This shows an alternative presentation style - a pie chart. Note that Zeppelin visualizations are limited. In lab 400 we will use Oracle Data Visualization to create more more interesting perspectives.

select customer_id, age, income, tenure, loan_type, loan_amount, occupation, marital_status
from credit_scoring_100k_v where rownum < 1000
```
Credit Score Predictions Simplified ...

STEP 9: Enough with Simple Charting; Let's Run Some OML Machine Learning Algorithms!

Create Attribute Importance Machine Learning Model

```sql
-- Find the importance of attributes that independently impact the target attribute: CREDIT_SCORE_BIN

DECLARE
v_sql varchar2(100);

BEGIN
EXECUTE IMMEDIATE 'DROP TABLE ai_explain_output_credit_score_bin';
EXCEPTION WHEN OTHERS THEN NULL;
END;

BEGIN
DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
  data_table_name => 'CREDIT_SCORING_100K_V',
  explain_column_name => 'CREDIT_SCORE_BIN',
  result_table_name => 'AI_EXPLAIN_OUTPUT_CREDIT_SCORE_BIN');
END;
```

Display the Top N Attributes for Good Credit Customers

```sql
-- Display those attributes that most influence the target field (Good Credit customers)

SELECT * FROM ai_explain_output_CREDIT_SCORE_BIN WHERE rownum < 7;
```
STEP 11: Create Predictive Model to Target Good Credit Customers

Now that Heather has found the key attributes that most influence finding more Good Credit customers and also making better Maximum Credit Card Amount decisions, she wants to leverage Oracle Machine Learning’s powerful in-Database, parallelized algorithms to build predictive models that help her company to better target “the right customers” with the “right offers”.

Let’s quickly review the machine learning process:

Problem Definition: Target Good Credit Customers

Data Gathering and Preparation: We’ve assembled 100K records with 100+ variable about each customer and have created a target field (Good Customer/Other Customer) so we can use OML’s Supervised Algorithms, specifically let’s start by using a decision tree algorithm.

Model Building and Evaluation: We’ll create a randomly selected sample from our Credit_Scoring_100k historical data and use 60% as training data for the machine learning model building phase. Then, we’ll use the remaining 40% as a holdout sample to test our model’s accuracy using various model evaluation tools such as a “lift chart”.

Knowledge Deployment: Once we’re satisfied that we have a useful ML model that can predict with some accuracy which customers we should target (Good Credit customers), we want to apply our OML model to new customer data inside ADWC and then take a deeper look at them. Lastly, we’ll jump over to Oracle Analytics Cloud for a more interactive, exploratory data analysis experience but now focusing on our customers of interest (Good Credit customers).
### Credit Score Predictions Simplified ...

**STEP 17:** Review the CREDIT_SCORING_NEW_PREDICTIONS Table and rank Good Customers Based on Prediction Probability, and Other Factors

<table>
<thead>
<tr>
<th>CUSTOMER_ID</th>
<th>PROB_GOOD_CREDIT</th>
<th>AGE</th>
<th>INCOME</th>
<th>TENURE</th>
<th>LOAN_TYPE</th>
<th>LOAN_AMOUNT</th>
<th>OCCUPATION</th>
<th>EDUCATION_LEVEL</th>
<th>MARITAL_STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>41,434</td>
<td>98.94</td>
<td>54</td>
<td>4,250</td>
<td>Need</td>
<td>30,000</td>
<td>Manager</td>
<td>Bachelor's Degree</td>
<td>Married</td>
<td></td>
</tr>
<tr>
<td>49,459</td>
<td>98.94</td>
<td>27</td>
<td>5,250</td>
<td>Auto</td>
<td>40,000</td>
<td>Manager</td>
<td>Bachelor's Degree</td>
<td>Married</td>
<td></td>
</tr>
<tr>
<td>54,025</td>
<td>98.94</td>
<td>18</td>
<td>4,950</td>
<td>Auto</td>
<td>50,000</td>
<td>Manager</td>
<td>Bachelor's Degree</td>
<td>Married</td>
<td></td>
</tr>
</tbody>
</table>

(Took 1 sec. Last updated by CHARLIE at August 03 2018, 4:54:24 PM. (outdated))

**STEP 18:** Apply a ML Model to a Single Record in a Transactional Application

```sql
-- Try this out by running query and then change customer_value_segment from 'Gold' to 'Silver' to see lower probability of customer having Good Credit

select prediction_probability(ML_CLASS_MODEL, 'Good Credit' using 'Very Rich' as WEALTH, 2000 as income, 'Silver' as customer_value_segment, 'Owner' as residential_status) Prediction_Probability
from dual;
```

**PREDICTION_PROBABILITY**

0.98945
### Key Factors of Good Credit Customers

#### Importance Value, Rank by Attribute Name

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Importance Value</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSUMER_FINDEX_SCORE</td>
<td>0.74</td>
<td>1</td>
</tr>
<tr>
<td>CUSTOMER_DMG_SEGMENT</td>
<td>0.54</td>
<td>3</td>
</tr>
<tr>
<td>CUSTOMER_LIFETIME_VALUE</td>
<td>0.08</td>
<td>13</td>
</tr>
<tr>
<td>CUSTOMER_VALUE_SEGMENT</td>
<td>0.40</td>
<td>6</td>
</tr>
<tr>
<td>DELINQUENCY_STATUS</td>
<td>0.23</td>
<td>11</td>
</tr>
<tr>
<td>EDUCATION_LEVEL</td>
<td>0.17</td>
<td>12</td>
</tr>
<tr>
<td>FAMILY_SIZE</td>
<td>0.04</td>
<td>14</td>
</tr>
<tr>
<td>HIGHEST_CREDIT_CARD_LIMIT</td>
<td>0.42</td>
<td>4</td>
</tr>
<tr>
<td>INCOME</td>
<td>0.42</td>
<td>5</td>
</tr>
<tr>
<td>MAX_CC_SPENT_AMOUNT</td>
<td>0.25</td>
<td>9</td>
</tr>
<tr>
<td>MAX_CC_SPENT_AMOUNT_PREV</td>
<td>0.27</td>
<td>8</td>
</tr>
<tr>
<td>NEW_BANKRUPTCY</td>
<td>0.00</td>
<td>16</td>
</tr>
<tr>
<td>NUMBER_OF_COLLECTIONS</td>
<td>0.00</td>
<td>17</td>
</tr>
<tr>
<td>NUMBER_OF_LIABLES</td>
<td>0.04</td>
<td>15</td>
</tr>
<tr>
<td>OCCUPATION</td>
<td>0.24</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Attribute Name Importance Value

- CONSUMER_FINDEX_SCORE
- WEALTH
- HIGHEST_CREDIT_CARD_LIMIT
- INCOME
- MAX_CC_SPENT_AMOUNT
- MAX_CC_SPENT_AMOUNT_PREV
- OCCUPATION
- NUMBER_OF_LIABLES
- NUMBER_OF_COLLECTIONS
- NEW_BANKRUPTCY
- RESIDENTIAL_STATUS
- DELINQUENCY_STATUS
- EDUCATION_LEVEL
- FAMILY_SIZE
- HIGHEST_CREDIT_CARD_LIMIT
- INCOME
- MAX_CC_SPENT_AMOUNT
- MAX_CC_SPENT_AMOUNT_PREV
- OCCUPATION
- NUMBER_OF_LIABLES
- NUMBER_OF_COLLECTIONS
- NEW_BANKRUPTCY
- RESIDENTIAL_STATUS
- DELINQUENCY_STATUS
- EDUCATION_LEVEL
- FAMILY_SIZE
Targeting High Credit Customers - Project

Click here or drag data to add a filter

Exploratory Graphs of Credit Scores ▼ More Expl Graphs ▼ Key Factors of Good Credit Customers ▼ Profiles of Good Credit Customer
Targeting High Credit Customers - Project

Exploratory Graphs of Credit Scores ▼ More Expl Graphs ▼ Key Factors of Good Credit Customers ▼ Profiles of Good Credit Customer

Copyright © 2018, Oracle and/or its affiliates. All rights reserved.
Example Oracle ML Customer References
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
£700M additional potential identified
£1 Billion in savings...
...Moving to Cloud
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 helped 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 eco system.

“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
Get 3300 hours, 2 TB of Exadata Storage on Oracle Cloud for free*

No Manual Tuning or Administration Needed

cloud.oracle.com/try-autonomous-database

* Trial expires upon usage of 3300 CPU hours or trial has reached 30 days, whichever comes first.
Analytics and Data Summit
All Analytics. All Data. No Nonsense.
March 12 – 14, 2019

Formerly called the BIWA Summit with the Spatial and Graph Summit
Same great technical content...new name!
www.AnalyticsandDataSummit.org