In-Database Analytics: Predictive Analytics, Oracle Exadata and Oracle Business Intelligence

Charlie Berger
Sr. Director Product Management, Data Mining and Advanced Analytics
Oracle Corporation
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine
<table>
<thead>
<tr>
<th>Spectrum of BI &amp; Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Queries &amp; Reports</strong></td>
</tr>
<tr>
<td>Extraction of detailed and roll up data</td>
</tr>
<tr>
<td>Who purchased mutual funds in the last 3 years?</td>
</tr>
<tr>
<td><strong>OLAP</strong></td>
</tr>
<tr>
<td>Summaries, trends and forecasts</td>
</tr>
<tr>
<td>What is the average <strong>income</strong> of mutual fund buyers, by region, by year?</td>
</tr>
<tr>
<td><strong>Data Mining</strong></td>
</tr>
<tr>
<td>Knowledge discovery of hidden patterns</td>
</tr>
<tr>
<td>Who is likely to mutual fund in the next 6 months and why?</td>
</tr>
</tbody>
</table>

*Copyright 2011 Oracle Corporation*
Data Mining Provides
Better Information, Valuable Insights and Predictions

Cell Phone Churners vs. Loyal Customers

Insight & Prediction

Segment #1:
IF CUST_MO > 14 AND INCOME < $90K, THEN Prediction = Cell Phone Churner, Confidence = 100%, Support = 8/39

Segment #3:
IF CUST_MO > 7 AND INCOME < $175K, THEN Prediction = Cell Phone Churner, Confidence = 83%, Support = 6/39

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff

Copyright 2011 Oracle Corporation
Data Mining Provides
Better Information, Valuable Insights and Predictions

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff

Copyright 2011 Oracle Corporation
My Personal Experience

Purchases were made in pairs of $75.00 purchases

- May 22 1:14 PM  FOOD  Monaco Café  $127.00
- May 22 7:32 PM  WINE  Wine Bistro  $28.00

... Gas Station?

- June 14 2:05 PM  MISC  Mobil Mart  $75.00
- June 14 2:06 PM  MISC  Mobil Mart  $75.00
- June 15 11:48 AM  MISC  Mobil Mart  $75.00
- June 15 11:49 AM  MISC  Mobil Mart  $75.00
- May 22 7:32 PM  WINE  Wine Bistro  $28.00
- June 16 11:48 AM  MISC  Mobil Mart  $75.00
- June 16 11:49 AM  MISC  Mobil Mart  $75.00

All same $75 amount?

Total purchases exceed time period average

France

Pairs of $75?
Finding Needles in Haystacks

• Haystacks are usually BIG

• Needles are typically small and rare
Look for What is “Different”
Oracle Data Mining
Anomaly Detection

• “One-Class” SVM Models
  - Fraud, noncompliance
  - Outlier detection
  - Network intrusion detection
  - Disease outbreaks
  - Rare events, true novelty

Problem: Detect rare cases
# Oracle Data Mining Algorithms

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Logistic Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Popular / Rules / transparency</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>Embedded app</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td>Regression</td>
<td>Multiple Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Lack examples of target field</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td>Clustering</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td></td>
<td>Hierarchical K-Means</td>
<td>Link analysis</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text mining</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>NonNegative Matrix Factorization</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
Typical Data Mining Use Cases

- **Retail**
  - Customer segmentation
  - Response modeling
  - Recommend next likely product
  - Profile high value customers

- **Banking**
  - Credit scoring
  - Probability of default
  - Customer profitability
  - Customer targeting

- **Insurance**
  - Risk factor identification
  - Claims fraud
  - Policy bundling
  - Employee retention

- **Higher Education**
  - Alumni donations
  - Student acquisition
  - Student retention
  - At-risk student identification

- **Healthcare**
  - Patient procedure recommendation
  - Patient outcome prediction
  - Fraud detection
  - Doctor & nurse note analysis

- **Life Sciences**
  - Drug discovery & interaction
  - Common factors in (un)healthy patients
  - Cancer cell classification
  - Drug safety surveillance

- **Telecommunications**
  - Customer churn
  - Identify cross-sell opportunities
  - Network intrusion detection

- **Public Sector**
  - Taxation fraud & anomalies
  - Crime analysis
  - Pattern recognition in military surveillance

- **Manufacturing**
  - Root cause analysis of defects
  - Warranty analysis
  - Reliability analysis
  - Yield analysis

- **Automotive**
  - Feature bundling for customer segments
  - Supplier quality analysis
  - Problem diagnosis

- **Chemical**
  - New compound discovery
  - Molecule clustering
  - Product yield analysis

- **Utilities**
  - Predict power line / equipment failure
  - Product bundling
  - Consumer fraud detection
Competitive Advantage

Optimization
Predictive Modeling
Forecasting/Extrapolation
Statistical Analysis
Alerts
Query/drill down
Ad hoc reports
Standard Reports

What’s the best that can happen?
What will happen next?
What if these trends continue?
Why is this happening?
What actions are needed?
Where exactly is the problem?
How many, how often, where?
What happened?

Degree of Intelligence

Source: Competing on Analytics, by T. Davenport & J. Harris

Copyright 2011 Oracle Corporation
Targeting the Right Customers

• 1:1 Relationships
• Understand and predict individual customer behavior
• Offer products and services that anticipate customer needs
• Build loyalty and increase profitability

Wants a new cell phone
Has two daughters
Travels across state lines frequently

Offer her:
1. Wide area digital phone plan
2. Emergency use plan for daughters
Oracle—Hardware and Software Engineered to Work Together

- Oracle is the world's most complete, open, and integrated business software and hardware systems company
  - Data Warehousing, VLDB and ILM
    - Oracle Data Mining Option
      - 12-in-DB data mining algorithms
      - In-DB model build
      - In-DB model apply
      - In-DB text mining
      - 50+ in-DB statistical functions

Oracle has taught the Database how to do Advanced Math/Stats/Data Mining
What is Data Mining?

• Automatically sifts through data to find hidden patterns, discover new insights, and make predictions

• Data Mining can provide valuable results:
  • Predict customer behavior (Classification)
  • Predict or estimate a value (Regression)
  • Segment a population (Clustering)
  • Identify factors more associated with a business problem (Attribute Importance)
  • Find profiles of targeted people or items (Decision Trees)
  • Determine important relationships and “market baskets” within the population (Associations)
  • Find fraudulent or “rare events” (Anomaly Detection)
In-Database Analytics

...a growing number of enterprises are doing it. It's understood as a best practice [or a] target architecture towards which you evolve your data warehousing practices if you're big on data mining. You know, a great many data warehouses in the real world are for operational business intelligence and reporting and ad hoc queries and don't do any data mining. But the bigger you get, the more likely you are to be doing extensive data mining and the more likely you are to be implementing or moving towards in-database analytics. [The goal there is] both to accelerate and scale up your data mining initiatives but also to harmonize all of your data mining initiatives around a common pool of reference data that you maintain in the data warehouse.

—Jim Kobielus, a senior data management analyst with Cambridge, Mass.-based Forrester Research Inc.

quote from “Customary Data Warehouse Concepts vs. Hadoop: Forrester Makes the Call”,
Mark Brunelli, Senior News Editor
This RSS Reprints Published: 11 Aug 2011
SQL Developer 3.0/Oracle Data Miner 11g Release 2 GUI

- Graphical User Interface for data analyst
- SQL Developer Extension (OTN download)
- Explore data—discover new insights
- Build and evaluate data mining models
- Apply predictive models
- Share analytical workflows
- Deploy SQL Apply code/scripts
• 12 years “stem celling analytics” into Oracle
  • Designed advanced analytics into database kernel to leverage relational database strengths
  • Naïve Bayes and Association Rules—1st algorithms added
  • Leverages counting, conditional probabilities, and much more

• Now, analytical database platform
  • 12 cutting edge machine learning algorithms and 50+ statistical functions
  • A data mining model is a schema object in the database, built via a PL/SQL API and scored via built-in SQL functions.
  • When building models, leverage existing scalable technology
    • (e.g., parallel execution, bitmap indexes, aggregation techniques) and add new core database technology (e.g., recursion within the parallel infrastructure, IEEE float, etc.)
  • True power of embedding within the database is evident when scoring models using built-in SQL functions (incl. Exadata)

```sql
select cust_id
from customers
where region = 'US'
  and prediction_probability(churnmod, 'Y' using *) > 0.8;
```
In-Database Data Mining

Traditional Analytics

Data Import

Data Mining Model “Scoring”

Data Preparation and Transformation

Data Mining Model Building

Data Prep & Transformation

Data Extraction

Oracle Data Mining

Data Preparation

Model “Scoring”

Model Building

Savings

Results
• Faster time for “Data” to “Insights”
• Lower TCO—Eliminates
  • Data Movement
  • Data Duplication
• Maintains Security

Model “Scoring”
Data remains in the Database
Embedded data preparation

Cutting edge machine learning algorithms inside the SQL kernel of Database

SQL—Most powerful language for data preparation and transformation

Data remains in the Database

Source Data
Dataset s/Work Area...
Analytical Processing...
Process Output...
Target...

Copyright 2011 Oracle Corporation
You Can Think of It 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 Data Mining
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine

ODM “Verbs”
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Oracle Data Miner Nodes (Partial List)

Tables and Views
- CUSTOMERS
- SALES
- SUPPLEMENTARY_DEMOGRAPHICS
- CUST_INSUR_LTV
- OUTPUT_1_25
- Update Table

Transformations
- Sample
- Transform
- Aggregate
- Join

Explore Data
- Explore Data

Modeling
- Regress Build
- Class Build
- Clust Build
- Assoc Build
- Model
- Model Details
- Feature Build
- Apply Text 1
- Apply
- Text

Text
- CUSTOMER COMMENTS
- Build Text
- Apply Text
Oracle Data Mining and Unstructured Data

- Oracle Data Mining mines unstructured i.e. “text” data
- Include free text and comments in ODM models
- Cluster and Classify documents
- Oracle Text used to preprocess unstructured text
Oracle Data Miner 11g Release 2 GUI

Churn Demo—Simple Conceptual Workflow
Churn models to product and “profile” likely churners

Oracle Data Miner 11g Release 2 GUI
Churn Demo—Simple Conceptual Workflow

Churn models to product and “profile” likely churners

Node: 10  Prediction: 0  Support: 658 (53.37%)  Confidence: 98.19%
          0 : 546 (98.18%)  1 : 12 (1.82%)
          Split: TOT_MIN_3M_CH

Node: 4   Prediction: 0  Support: 73 (5.92%)  Confidence: 83.59%
          0 : 61 (83.56%)  1 : 12 (16.44%)
          Split: INCOME

Node: 14  Prediction: 0  Support: 48 (3.89%)  Confidence: 75%
          0 : 36 (75.00%)  1 : 12 (25.00%)

Node: 13  Prediction: 1  Support: 19 (1.54%)  Confidence: 100%
          0 : 0 (0.00%)  1 : 19 (100.00%)

Rule | Surrogates | Target Values
--- | --- | ---
IF MIN_CH3M <= .5 AND TOT_MIN_3MCH <= 70 AND MIN_PREV_MO > 688.5 AND INCOME <= 139500 THEN 1
Oracle Data Miner 11g Release 2 GUI

Churn Demo—Simple Conceptual Workflow

Market Basket Analysis to identify potential product bundles

Rule Details:

<table>
<thead>
<tr>
<th>ID</th>
<th>Lift</th>
<th>Confidence (%)</th>
<th>Support (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17313</td>
<td>29.1235</td>
<td>70.8661</td>
<td>1.0689</td>
</tr>
<tr>
<td>17673</td>
<td>28.7389</td>
<td>69.8304</td>
<td>1.0523</td>
</tr>
<tr>
<td>20557</td>
<td>27.8285</td>
<td>67.7793</td>
<td>1.0795</td>
</tr>
<tr>
<td>17314</td>
<td>27.6127</td>
<td>77.5469</td>
<td>1.0989</td>
</tr>
<tr>
<td>18071</td>
<td>27.5923</td>
<td>76.3947</td>
<td>1.0207</td>
</tr>
<tr>
<td>17942</td>
<td>27.4802</td>
<td>68.3697</td>
<td>1.0485</td>
</tr>
</tbody>
</table>

Copyright 2011 Oracle Corporation
Clustering analysis to discover customer segments based on behavior, demographics, plans, equipment, etc.
Fraud Prediction Demo

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', 'CLAIMS2', '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 CLAIMS2
where PASTNUMBEROFCLAIMS in ('2 to 4', 'more than 4'))
where rnk <= 5
order by percent_fraud desc;

<table>
<thead>
<tr>
<th>POLICYNUMBER</th>
<th>PERCENT_FRAUD</th>
<th>RNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>6532</td>
<td>64.78</td>
<td>1</td>
</tr>
<tr>
<td>2749</td>
<td>64.17</td>
<td>2</td>
</tr>
<tr>
<td>3440</td>
<td>63.22</td>
<td>3</td>
</tr>
<tr>
<td>654</td>
<td>63.1</td>
<td>4</td>
</tr>
<tr>
<td>12650</td>
<td>62.36</td>
<td>5</td>
</tr>
</tbody>
</table>

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

Oracle
In 11g Release 2, SQL predicates and Oracle Data Mining models are pushed to storage level for execution. For example, find the US customers likely to churn:

```sql
select cust_id
from customers
where region = 'US'
and prediction_probability(churnmod, 'Y' using *) > 0.8;
```
Real-time Prediction for a Customer

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

```sql
SELECT prediction_probability(CLAS_DT_5_2, '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;
```
Ability to Import/Export 3\textsuperscript{rd} Party DM Models

- ODM 11g Release 2 adds ability to import 3\textsuperscript{rd} party models (PMML), convert to native ODM models and score them in-DB
  - Supported models for ODM model export:
    - Decision Trees (PMML)
  - Supported algorithms for ODM model import:
    - Multiple regression models (PMML)
    - Logistic regression models (PMML)

- Benefits
  - SAS, SPSS, R, etc. data mining models can scored on Exadata
    - Imported dm models become native ODM models and inherit all ODM benefits including scoring at Exadata storage layer, 1st class objects, security, etc.
11g Statistics & SQL Analytics (Free)

- Ranking functions
  - rank, dense_rank, cume_dist, percent_rank, ntile

- Window Aggregate functions
  (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value

- LAG/LEAD functions
  - Direct inter-row reference using offsets

- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report

- Statistical Aggregates
  - Correlation, linear regression family, covariance

- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions

Descriptive Statistics
- DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, median, stats_mode, variance, standard deviation, quantile values, +/- n sigma values, top/bottom 5 values

Correlations
- Pearson’s correlation coefficients, Spearman's and Kendall's (both nonparametric).

Cross Tabs
- Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa

Hypothesis Testing
- Student t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA

Distribution Fitting
- Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition
• R’s rapid adoption has earned its reputation as a new statistical software standard

*While it is difficult to calculate exactly how many people use R, those most familiar with the software estimate that close to 250,000 people work with it regularly.*


http://www.r-project.org/
"Oracle R Enterprise" Architecture

"R for the Enterprise"
- Combines open source R statistical community with power & architecture of the Database
- Develop and immediately deploy R Scripts

Save money on SA$!
- Migrate functions into the Database and reduce SA$ Annual Usage Fees
- Private analytical sandboxes for LOB/data analyst

Oracle in-Database Analytics for Big Data
- Eliminate data movement and maximize performance and security
ODM’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools.
Exadata with Analytics and Business Intelligence—Better Together

- In-database data mining builds predictive models that predict customer behavior
- OBIEE’s integrated spatial mapping shows where

Customer “most likely” be be HIGH and VERY HIGH value customer in the future
Exadata with Analytics and Business Intelligence—Better Together

- Deliver advanced in-database analytics through OBIEE
- Ability to drill-through for detail
- Harness the power of Exadata for “Better BI & analytics”

Oracle Data Mining’s Predictions versus “Actuals” highlight areas for improvement and insights
Exadata with Analytics and Business Intelligence—Better Together

- Exadata power
- OBIEE ease-of-use

Drill-through for details about top factors that define HIGH and VERY HIGH value customers
Fusion HCM Predictive Analytics
Factory Installed PA/ODM Methodologies

The size of the person marker is based on the size of the team.
Learn More
Oracle Data Mining PL/SQL Sample Programs

- The PL/SQL Sample Programs provide examples of mini-solutions and use cases for Oracle Data Mining
- Excellent starting point when developing an ODM Application

<table>
<thead>
<tr>
<th>Mining Function</th>
<th>Algorithm</th>
<th>Sample Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly Detection</td>
<td>One-Class Support Vector Machine</td>
<td>dmsvcdem.sql</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Apriori</td>
<td>dmarndemo.sql</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Minimum Descriptor Length</td>
<td>dmaidemo.sql</td>
</tr>
<tr>
<td>Classification</td>
<td>Decision Tree</td>
<td>dmdtdemo.sql</td>
</tr>
<tr>
<td>Classification</td>
<td>Decision Tree (cross validation)</td>
<td>dmdtxvlddemo.sql</td>
</tr>
<tr>
<td>Classification</td>
<td>Logistic Regression</td>
<td>dmlcdem.sql</td>
</tr>
<tr>
<td>Classification</td>
<td>Naive Bayes</td>
<td>dmnbndemo.sql</td>
</tr>
<tr>
<td>Classification</td>
<td>Support Vector Machine</td>
<td>dmsvcdem.sql</td>
</tr>
<tr>
<td>Clustering</td>
<td>k-Means</td>
<td>dmkndemo.sql</td>
</tr>
<tr>
<td>Clustering</td>
<td>O-Cluster</td>
<td>dmocdemo.sql</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>Non-Negative Matrix Factorization</td>
<td>dmnmdemo.sql</td>
</tr>
<tr>
<td>Regression</td>
<td>Linear Regression</td>
<td>dmglrdem.sql</td>
</tr>
<tr>
<td>Text Mining</td>
<td>Text transformation using Oracle Text</td>
<td>dmtxtfe.sql</td>
</tr>
<tr>
<td>Text Mining</td>
<td>Non-Negative Matrix Factorization</td>
<td>dmtxtnmf.sql</td>
</tr>
<tr>
<td>Text Mining</td>
<td>Support Vector Machine (Classification)</td>
<td>dmtxtsvm.sql</td>
</tr>
</tbody>
</table>