Oracle Data Mining —
In-Database Data Mining Made Easy!

Charlie Berger
Sr. Director Product Management, Data Mining and Advanced Analytics
Oracle Corporation
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine

Copyright 2010 Oracle Corporation
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.
Agenda

- Market Drivers
- Oracle Data Mining
- Exadata and Oracle Data Mining
- Oracle Data Miner 11g Release 2 New GUI
- Oracle Statistical Functions
- Ability to Import 3rd Party e.g. SAS models
- Applications Powered by Oracle Data Mining
- Getting Started with ODM
Analytics: Strategic and Mission Critical

- **Competing on Analytics**, by Tom Davenport
  - “Some companies have built their very businesses on their ability to collect, analyze, and act on data.”
  - “Although numerous organizations are embracing analytics, only a handful have achieved this level of proficiency. But analytics competitors are the leaders in their varied fields—consumer products, finance, retail, and travel and entertainment among them.”
  - “Organizations are moving beyond query and reporting” - IDC 2006

- **Super Crunchers**, by Ian Ayers
  - “In the past, one could get by on intuition and experience. Times have changed. Today, the name of the game is data.” —Steven D. Levitt, author of *Freakonomics*
  - “Data-mining and statistical analysis have suddenly become cool…. Dissecting marketing, politics, and even sports, stuff that complex and important shouldn't be this much fun to read.” —Wired
Competitive Advantage

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

What’s the best that can’t 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?

Source: Competing on Analytics, by T. Davenport & J. Harris
In-Database Data Mining
What is In-Database Analytics?

Move the data??  Move the algorithms?
What is In-Database Analytics?

Move the data??  Move the algorithms?
What is In-Database Analytics?

Move the algorithms!!!!!
Traditional Analytics Environment

*Move Data → Algorithms*

- **Source Data**
  - (Oracle, DB2, SQL Server, Teradata, Ext. Tables, etc.)

- **Analytics Work Area**
  - (Datasets)

- **Analytics Processing**
  - (Statistical functions/Data mining)

- **Process Output**
  - (Work Area)

- **Target**
  - (e.g. Oracle)

---

Hours, Days or Weeks

- Traditional analytics environment results in:
  - Data movement
  - Data duplication
  - Loss of security

*Copyright 2010 Oracle Corporation*
Oracle Architecture

**Move Data ↔ Algorithms**

- **Source Data**
  - (Oracle, DB2, SQL Server, TeraData, Ext. Tables, etc.)

- **Oracle architecture:**
  - Eliminates data movement
  - Eliminates data duplication
  - Preserves security

- **Secs, Mins or Hours**

**Data Mining**
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 Import

Data Mining Model “Scoring”

Data Preparation and Transformation

Data Mining Model Building

Data Prep & Transformation

Data Extraction

Results

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

Secs. Mins or Hours

Savings

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
• 11 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;
```
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 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></td>
<td>Hierarchical K-Means</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Text mining</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>Attribute Importance</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>Association Rules</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>NMF</td>
<td>Text analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
Oracle Data Miner 11g Release 2 GUI
Free SQL Developer Extension on OTN

- 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
Oracle Data Miner 11g Release 2 GUI
Free SQL Developer Extension on OTN
The Forrester Wave™: Predictive Analytics And Data Mining Solutions, Q1 2010

Oracle Data Mining Cited as a Leader; 2nd place in Current Offering

• Ranks 2nd place in Current Offering
• “Oracle focuses on in-database mining in the Oracle Database, on integration of Oracle Data Mining into the kernel of that database, and on leveraging that technology in Oracle’s branded applications.”
Exadata & ODM
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;
```
Exadata + Data Mining 11g Release 2

**Benefits**

- Eliminates data movement
  - 2X-5X+ faster scoring on Exadata
    - Depends on number of joins involved with data for scoring
- Preserves security
- Significant architecture and performance advantages over SAS Institute
  - Years ahead of SAS’s road map to move SAS analytics towards RDBMSs ([http://support.sas.com/resources/papers/InDatabase07.pdf](http://support.sas.com/resources/papers/InDatabase07.pdf))
- Netezza performance but using industry standard RDBMS + SQL-based in-database advanced analytics
- Best platform for building enterprise predictive analytics applications e.g. Fusion Applications ➔ “Analytical iPod for the Enterprise”
TurkCell Prepaid Churn Model

Oracle Data Mining on Exadata 11g Release 2

• Churn Problem
  • Churn prediction starts with turning an abundance of data into valuable information and continues as a cyclic process

• Approach
  • Initially we have used a large Solaris (100+ UltraSparc 7 cores and 640 GB memory) box to build our first SVM models:
  • It took 29 hours to complete model build & apply.

• Conclusion
  • On Exadata this reduces to a few hours mainly due to enormous improvement in data preparation stage
  • Churn prediction over various customer groups is and will be the focus of Turkcell
  • Embedded data mining with ODM is faster, more robust (due to stability of SVM algorithm), easier to automate, easier to manage

Excerpts from TurkCell presentation at OOW 2010, September 21, 2010
Necdet Deniz Halıcıoğlu deniz.halicioglu@turkcellteknoloji.com.tr

Copyright 2010 Oracle Corporation
Oracle Data Miner
11g Release 2

Easier
Oracle Data Miner 11g Release 2 GUI

- Predict customer behavior
- Identify key factors
- Predict next-likely product
- Customer profiling
- Detect fraud & anomalies
- Mine “text” and unstructured data
Explore Data

- Thumbnail distributions of every attribute
  - Grouped by another attribute

- Summary statistics for all attributes
  - Min, max, stdev, variance, median, mean, skewness, kurtosis, etc.
Build and Evaluate Models

• Comparative model performance results
• Adjust and tune predictive models
Understand Model Details

- Interactive model viewers
Analytical “Work Flow” Methodologies

- Build, share and automate predictive analytics methodologies
SQL Developer Active Query Builder

• New, easy to use, interactive query builder in SQL Developer for assembling and preparing data—for mining
Example: Simple, Predictive SQL

Select customers who are more than 85% likely to be HIGH VALUE customers & display their AGE & MORTGAGE_AMOUNT

```
SELECT * from(
    SELECT A.CUST_ID, A.AGE, MORTGAGE_AMOUNT, PREDICTION_PROBABILITY(CUST_INSUR_LT46939_DT, 'VERY HIGH' USING A.*) prob
    FROM CBERGER.CUST_INSUR_LTV A)
WHERE prob > 0.85;
```
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
Real-time Prediction

with
records as (select
   78000  SALARY,
   250000  MORTGAGE_AMOUNT,
   6   TIME_AS_CUSTOMER,
   12  MONTHLY_CHECKS_WRITTEN,
   55  AGE,
   423  BANK_FUNDS,
   'Married'  MARITAL_STATUS,
   'Nurse'  PROFESSION,
   'M'  SEX,
   4000  CREDIT_CARD_LIMITS,
   2  N_OF_DEPENDENTS,
   1  HOUSE_OWNERSHIP  from dual)
select s.prediction prediction, s.probability probability
from ( select PREDICTION_SET(CUST_INSUR_LT46939_DT, 1  USING *) pset
   from records) t, TABLE(t.pset) s;

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

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH</td>
<td>.65123504738232096</td>
</tr>
</tbody>
</table>
Oracle Statistical Functions (Free)
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

Copyright 2010 Oracle Corporation
Split Lot A/B Offer testing

- Offer “A” to one population and “B” to another
- Over time period “t” calculate median purchase amounts of customers receiving offer A & B
- Perform t-test to compare
- If statistically significantly better results achieved from one offer over another, offer everyone higher performing offer
Independent Samples T-Test (Pooled Variances)

- Query compares the mean of AMOUNT_SOLD between MEN and WOMEN within CUST_INCOME_LEVEL ranges

```sql
SELECT substr(cust_income_level,1,22) income_level,
       avg(decode(cust_gender,'M',amount_sold,null)) sold_to_men,
       avg(decode(cust_gender,'F',amount_sold,null)) sold_to_women,
       stats_t_test_indep(cust_gender, amount_sold, 'STATISTIC','F') t_observed,
       stats_t_test_indep(cust_gender, amount_sold) two_sided_p_value
FROM sh.customers c, sh.sales s
WHERE c.cust_id=s.cust_id
GROUP BY rollup(cust_income_level)
ORDER BY 1;
```
Ability to Import 3rd Party e.g. SAS Models
Ability to Import 3rd Party DM Models

• Capability to import 3rd party dm models, import, and convert to native ODM models

• Benefits
  • SAS, SPSS, R, etc. data mining models can be used for scoring inside the Database
  • Imported dm models become native ODM models and inherit all ODM benefits including scoring at Exadata storage layer, 1st class objects, security, etc.
In-Database SAS Scoring

*Score the SAS_ODM Model*

- SAS models become native ODM models
  - No loss of information
- Original source data for scoring remains in Database
- “Exadata scoring” of SAS models
In-Database SAS Scoring

Import the SAS Model

```
begin
  dbms_data_mining.import_model
  (    'SAS_Log_Reg_Model4',
       XMLType(bfilename('PMML_DIR',
' getSAS_Logistic_Regression_PMML_Model.xml'),
       nls_charset_id('AL32UTF8'))
  );
end;
/
```
In-Database SAS Scoring

Score the SAS_ODM Model

```sql
select
    prediction(SAS_Log_Reg_Model4 using *),
    prediction_probability(SAS_Log_Reg_Model using *)
from
    sas_dataset where id < 10;
```
Applications Powered by Oracle Data Mining
Integration with Oracle BI EE

Oracle BI EE defines results for end user presentation.

Oracle Data Mining results available to Oracle BI EE administrators.
### Example

**Better Information for OBI EE Reports and Dashboards**

ODM’s predictions & probabilities are available in the database for reporting using Oracle BI EE and other tools.
Predictive Analytics Applications

Powered by Oracle Data Mining

(Partial List as of March 2010)

CRM OnDemand—Sales Prospector

Oracle Communications Data Model

Oracle Retail Data Model

Spend Classification

Oracle Open World - Schedule Builder
Getting Started

- Oracle Data Miner Cue Cards—part of client install
- Oracle By Example Online Learning on OTN

Using Oracle Data Miner for Oracle Database 11g Release 2

Purpose

This tutorial covers the use of Oracle Data Miner to perform data mining against Oracle Database 11g Release 2. The Oracle Data Miner GUI is included as an extension of Oracle SQL Developer. Oracle SQL Developer is a free graphical tool for database development. With SQL Developer, you can browse database statements. Starting with SQL Developer, version 3.0, you can also access the Oracle Data Miner GUI, which provides additional functionality.

DISCLAIMER: This tutorial has been developed with pre-production software and is not available for external audit.

Time to Complete

Approximately 30 mins.

Overview

Data mining is the process of extracting useful information from masses of data by extracting patterns and trends.

- Predict individual behavior, for example, the customers likely to respond to a promotional offer or the customer
- Find profiles of targeted people or items (Classification using Decision Trees)
- Find natural segments or clusters (Clustering)
- Identify factors more associated with a target attribute (Attribute importance)
- Find co-occurring events or purchases (Associations, sometimes known as Market Basket Analysis)
- Find fraudulent or rare events (Anomaly Detection)

The phases of solving a business problem using Oracle Data Mining are as follows:

1. Problem Definition in Terms of Data Mining and Business Goals
2. Data Acquisition and Preparation
3. Building and Evaluation of Models
4. Deployment

Compare the Models

After you build the selected models, you can view and evaluate the results for all of the models in a comparative format.

Follow these steps:

1. Right-click the build node and select Compare from the menu.

Results: A Class Build display tab opens, showing a graphical comparison of the four models, as shown here.
“This presentation is for informational purposes only and may not be incorporated into a contract or agreement.”