Oracle Data Mining
Overview and Demo

Charlie Berger
Sr. Director Product Management, Data Mining Technologies
Oracle Corporation
charlie.berger@oracle.com
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.
Outline

• Today’s BI must go beyond simple reporting
• To succeed, companies must
  • Eliminate data movement
  • Collapse information latency
  • Deliver better BI through analytics
• ODM makes the Database an “Analytical Database”
  • Enables applications “Powered by Oracle Data Mining”
• Brief demonstrations
  1. Oracle Data Mining
  2. OBI EE Dashboards with ODM Results
  3. Oracle Sales Prospector with embedded 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

Degree of Intelligence

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?

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

Copyright © 2009 Oracle Corporation
Oracle Data Mining Option
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)*
## Oracle Data Mining Example Use Cases

<table>
<thead>
<tr>
<th>Sector</th>
<th>Use Cases</th>
</tr>
</thead>
</table>
| 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 |
Data Mining Provides Better Information, Valuable Insights and Predictions

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
Predicting High LTV Customers Using a Decision Tree Model

Simple model: Other ODM models can mine:
- unstructured data (e.g. text comments)
- transactions data (e.g. purchases), etc.

IF (Mortgage_Amount > $500K AND House_Own = 2 or more AND Age = >42)
THEN Probability(Lifetime Customer Value is “VERY HIGH” = 77%, Support = 15%
“Essentially, all models are wrong, but some are useful.”

- George Box

(one of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.)
## Oracle Data Mining

**Overview (Classification)**

### Cases

<table>
<thead>
<tr>
<th>Name</th>
<th>Income</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>30,000</td>
<td>30</td>
</tr>
<tr>
<td>Smith</td>
<td>55,000</td>
<td>67</td>
</tr>
<tr>
<td>Lee</td>
<td>25,000</td>
<td>23</td>
</tr>
<tr>
<td>Rogers</td>
<td>50,000</td>
<td>44</td>
</tr>
</tbody>
</table>

### Historic Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Income</th>
<th>Age</th>
<th>Respond?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>30,000</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Smith</td>
<td>55,000</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>Lee</td>
<td>25,000</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>Rogers</td>
<td>50,000</td>
<td>44</td>
<td>0</td>
</tr>
</tbody>
</table>

### New Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Income</th>
<th>Age</th>
<th>Predict?</th>
<th>Prediction Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campos</td>
<td>40,500</td>
<td>52</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Horn</td>
<td>37,000</td>
<td>73</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Habers</td>
<td>57,200</td>
<td>32</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Berger</td>
<td>95,600</td>
<td>34</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

### Functional Relationship:

\[ Y = F(X_1, X_2, \ldots, X_m) \]

**Oracle Database 11g**
# Oracle Data Mining

## Algorithm Summary 11g

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td><strong>Logistic Regression (GLM)</strong></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><strong>Regression</strong></td>
<td><strong>Multiple Regression (GLM)</strong></td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Anomaly Detection</strong></td>
<td>One Class SVM</td>
<td>Lack examples</td>
</tr>
<tr>
<td><strong>Attribute Importance</strong></td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction / Identify useful data / Reduce data noise</td>
</tr>
<tr>
<td><strong>Association Rules</strong></td>
<td><strong>Apriori</strong></td>
<td>Market basket analysis / Link analysis</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td><strong>Hierarchical K-Means</strong></td>
<td>Product grouping / Text mining</td>
</tr>
<tr>
<td></td>
<td><strong>Hierarchical O-Cluster</strong></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td><strong>Feature Extraction</strong></td>
<td><strong>NMF</strong></td>
<td>Text analysis / Feature reduction</td>
</tr>
</tbody>
</table>
Traditional Analytics (SAS) Environment

- SAS environment requires:
  - Data movement
  - Data duplication
  - Loss of security
Oracle Architecture

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

- 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

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

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

Copyright © 2009 Oracle Corporation
In-Database Data Mining

Advantages

• ODM architecture provides greater
  • Performance, scalability, and data security
• Data remains in the database
  • Fewer moving parts; shorter information latency
• Straightforward inclusion within interesting and arbitrarily complex queries
  • “SELECT Customers WHERE Income > 100K, AND Probability(Buy Product A) > .85;”
• Real-world scalability—available for mission critical appls
• Enables pipelining of results without costly materialization
• Performant and scalable:
  • Fast scoring: 2.5 million records scored in 6 seconds on a single CPU system
  • Real-time scoring: 100 models on a single CPU: 0.085 seconds
HP Oracle Database Machine & ODM

• Integrated data warehouse solution
• Extreme Performance
  • 10-100X faster than conventional DW systems
• Scalability to Petabytes
• Enterprise-Ready
  • Complete data warehouse functionality
  • Enterprise-level availability and security
• Scoring of Oracle Data Mining models
  • Blazingly fast performance
  • 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;
```
“If I had one hour to save the world, I would spend fifty-five minutes defining the problem and only five minutes finding the solution”

- Albert Einstein

(see also http://www.wikihow.com/Define-a-Problem)
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
Example: Simple, Predictive SQL

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

```sql
SELECT * from(
  SELECT A.CUSTOMER_ID, A.AGE, MORTGAGE_AMOUNT, PREDICTION_PROBABILITY(INSUR_CUST_LT27754_DT, 'VERY HIGH' USING A.*) prob 
  FROM CBERGER.INSUR_CUST_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;
begindbms_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 ('2 to 4', 'more than 4'))
where rnk <= 5
order by percent_fraud desc;
Oracle Data Mining 11g

- Data Mining Functions (Server)
  - PL/SQL & Java APIs
  - Develop & deploy predictive analytics applications
- Wide range of DM algorithms (12)
  - Classification & regression
  - Clustering
  - Anomaly detection
  - Attribute importance
  - Feature extraction (NMF)
  - Association rules (Market Basket analysis)
  - Structured & unstructured data (text mining)
- Oracle Data Miner (GUI)
  - Simplified, guided data mining using wizards
- Predictive Analytics
  - “1-click data mining” from a spreadsheet
It boils down to this:

**Less** data movement = **faster** analytics, and faster analytics = **better** BI throughout the enterprise
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

<table>
<thead>
<tr>
<th>Most Suspicious Claims</th>
<th>Number Of Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLICYNUMBER</td>
<td>PREDICTION</td>
</tr>
<tr>
<td>1345.00</td>
<td>0.00</td>
</tr>
<tr>
<td>14485.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8532.00</td>
<td>0.00</td>
</tr>
<tr>
<td>12631.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>15032.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11015.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11192.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4558.00</td>
<td>0.00</td>
</tr>
<tr>
<td>12850.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2749.00</td>
<td>0.00</td>
</tr>
<tr>
<td>12397.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3381.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6653.00</td>
<td>0.00</td>
</tr>
<tr>
<td>13677.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11192.00</td>
<td>0.00</td>
</tr>
<tr>
<td>13294.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6294.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2239.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3440.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8139.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11088.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11191.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11877.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4245.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

ODM’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools.
Oracle SQL Statistical Functions
(Free in Every Oracle Database)
11g Statistics & SQL Analytics

• 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, stats_mode, variance, standard deviation, median, 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 © 2009 Oracle Corporation
Descriptive Statistics

- **MEDIAN & MODE**
  - Median: takes numeric or datetype values and returns the middle value
  - Mode: returns the most common value

```sql
A. SELECT STATS_MODE(AGE) from LYMPHOMA;

B. SELECT MEDIAN(AGE) from LYMPHOMA;

C. SELECT TREATMENT_PLAN, STATS_MODE(LYMPH_TYPE)
   from lymphoma GROUP BY TREATMENT_PLAN;

D. SELECT LYMPH_TYPE, MEDIAN(SIZE_REDUCTION) from
   LYMPHOMA GROUP BY LYMPH_TYPE ORDER BY
   MEDIAN(SIZE_REDUCTION) ASC;
```
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

```
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;
```

SQL Worksheet
Correlation Functions

- The CORR_S and CORR_K functions support nonparametric or rank correlation (finding correlations between expressions that are ordinal scaled).

- Correlation coefficients take on a value ranging from –1 to 1, where:
  - 1 indicates a perfect relationship
  - –1 indicates a perfect inverse relationship
  - 0 indicates no relationship

- The following query determines whether there is a correlation between the AGE and WEIGHT of people, using Spearman's correlation:

```sql
select CORR_S(AGE, WEIGHT) coefficient, 
       CORR_S(AGE, WEIGHT, 'TWO_SIDED_SIG') p_value, 
       substr(TREATMENT_PLAN, 1,15) as TREATMENT_PLAN 
from CBERGER.LYMPHOMA 
GROUP BY TREATMENT_PLAN;
```
### In-Database Analytics Engine vs. External Analytical Engine

#### Basic Statistics
- **Free**

#### Data Mining

#### Text Mining

#### Costs (ODM: $23K cpu)
- Simplified environment
- Single server
- Security

#### IT Platform
- SQL (standard)
- Java (standard)

---

#### External Analytical Engine
- Basic Statistics
- Data Mining
- Text Mining (separate: SAS EM for Text)
- Advanced Statistics

#### Costs (SAS EM: $150K/5 users)
- Duplicates data
- Annual Renewal Fee (AUF) (~45% each year)

#### IT Platform
- SAS Code (proprietary)
### In-Database Analytics Engine

1. **Basic Statistics** *(Free)*
2. Data Mining
3. Text Mining

### Costs *(ODM: $23K cpu)*

- Simplified environment
- Single server
- Security

### IT Platform

- SQL *(standard)*
- Java *(standard)*

---

### External Analytical Engine

1. **Basic Statistics**
2. Data Mining
3. **Text Mining** *(separate: SAS EM for Text)*
4. **Advanced Statistics**

### Costs *(SAS EM: $150K/5 users)*

- Duplicates data
- Annual Renewal Fee *(AUF)* (~45% each year)

### IT Platform

- SAS Code *(proprietary)*

---

Copyright © 2009 Oracle Corporation
“The goal of the SAS In-Database initiative is … to achieve deeper technical integration with database providers..

..., the SAS engine often must load and extract data over a network to and from the DBMS. This presents a series of challenges:

- …Network bottlenecks between SAS and the DBMS constrain access to large volumes of data
- … the results of the SAS processing must be transferred back to the DBMS for final storage, which further increases the cost.
IDC Worldwide Business Analytics Software
Brief Demonstrations

1. Oracle Data Mining
2. Oracle Business Intelligence EE
3. CRM Sales Prospector
Oracle Data Mining + OBI EE
Quick Demo: Oracle Data Mining

- Scenario: Insurance Company
- Business problem(s):
  1. Better understand the business by looking at graphs of the data
  2. Identify the factors (attributes) most associated with Customer who BUY_INSURANCE
  3. Target Best Customers
     a. Build a predictive model to understand who will be a VERY_HIGH VALUE Customer …. And WHY (IF… THEN.. Rules that can describe them)
     b. Predict who is likely to be a VERY_HIGH VALUE Customer in the future
     c. View results in an OBI EE Dashboard
        • Including other business problems e.g. Fraud, Cross-Sell, etc.
        • (Entire process can be automated w/ PL/SQL and/or Java APIs)
Oracle Data Mining + OBI EE
Understand the Data

Oracle Data Mining helps to visualize the data.
Oracle Data Mining + OBI EE

Target the Right Customers

Oracle Data Miner guides the analyst through the data mining process.
Oracle Data Mining builds a model that differentiates HI_VALUE_CUSTOMERS from others.
Oracle Data Mining + OBI EE
Targeting High Value Customers

Oracle Data Mining creates a prioritized list of customer who are likely to be high value.
Integration with Oracle BI EE

Oracle Data Mining provides more information and better insight
### Oracle Data Mining

**Know More, Do More, Spend Less**

<table>
<thead>
<tr>
<th>Business Decision Makers</th>
<th>Data Analysts</th>
<th>Integrators and IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Make Better Decisions</td>
<td>• Get Results Faster</td>
<td>• Create More Value for Your Organization</td>
</tr>
<tr>
<td>• Extract More Value from Your Data</td>
<td>• Get More Results</td>
<td>• Make Your Work Easier</td>
</tr>
<tr>
<td>• Lower Your Total Cost of Ownership</td>
<td>• Easy to Use</td>
<td>• Transform IT from a Cost to a Profit Center</td>
</tr>
</tbody>
</table>

Copyright © 2009 Oracle Corporation
Oracle Data Mining (SQL & Java) APIs
drop table HCM_SET;
exec dbms_data_mining.drop_model('HCMMODEL');

create table HCM_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into HCM_SET values ('ALGO_NAME', 'ALGO_SUPPORT_VECTOR_MACHINES');
insert into HCM_SET values ('PREP_AUTO', 'ON');
commit;
begin
dbms_data_mining.create_model('HCMMODEL', 'CLASSIFICATION', 'EMPL_DATA', 'EMPL_ID', 'CURR_EMPL', 'HCM_SET');
end;
/

-- accuracy (per-class and overall)
col actual format a6
select actual, round(corr*100/total,2) percent, corr, total-corr incorr, total from
(select actual, sum(decode(actual,predicted,1,0)) corr, count(*) total from
(select CURR_EMPL actual, prediction(HCMMODEL using *) predicted
from EMPL_DATA_JUNE07)
group by rollup(actual));

-- top 5 very high value, current employees most likely to leave
select * from
(select empl_id, round(prob_leave*100,2) percent_leave,
 rank() over (order by prob_leave desc) rnk from
(select empl_id, prediction_probability(HCMMODEL, 'NO' using *) prob_leave
from EMPL_DATA_JUNE07
where CURR_EMPL = 'YES' and LTV_BIN = 'VERY HIGH')
where rnk <= 5
order by percent_leave desc;
Predictive Analytics Use Case

• The cast:
  • Peter: a data mining analyst
  • Sally: a marketing manager
• Peter builds a decision tree classification model, tree_model
• Peter grants the ability to view/score the tree model to Sally
  
  GRANT SELECT MODEL ON tree_model TO Sally;

• Sally inspects the model, likes it, and wants it deployed
• Sally scores the customer database using the new model and his understanding of the cost of contacting a customer and sends the new contact list to the head of the sales department

  CREATE TABLE AS SELECT cust_name, cust_phone FROM customers
  WHERE prediction(Peter.tree_model cost matrix (0,5,1,0) using *) = ‘responder’;
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(INSUR_CUST_LT68054_DT, 1 USING *) pset from records) t, TABLE(t.pset) s;

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH</td>
<td>.65123504736232096</td>
</tr>
</tbody>
</table>
Prediction Multiple Models/Optimization

with records as (select
    178255 ANNUAL_INCOME,
    30 AGE,
    'Bach.' EDUCATION,
    'Married' MARITAL_STATUS,
    'Male' SEX,
    70 HOURS_PER_WEEK,
    98 PAYROLL_DEDUCTION from dual)
select t.*
from (select
    'CAR_MODEL' MODEL, s1.prediction prediction, s1.probability probability, s1.probability*25000 as expected_revenue from (select PREDICTION_SET(NBMODEL_JDM, 1 USING *) pset from records ) t1, TABLE(t1.pset) s1
UNION
    select 'MOTOCYCLE_MODEL' MODEL, s2.prediction prediction, s2.probability probability, s1.probability*2000 as expected_revenue from (select PREDICTION_SET(ABNMODEL_JDM, 1 USING *) pset from records ) t2, TABLE(t2.pset) s2
UNION
    select 'TRICYCLE_MODEL' MODEL, s3.prediction prediction, s3.probability probability, s1.probability*50 as expected_revenue from (select PREDICTION_SET(TREEMODEL_JDM, 1 USING *) pset from records ) t3, TABLE(t3.pset) s3
UNION
    select 'BICYCLE_MODEL' MODEL, s4.prediction prediction, s4.probability probability, s1.probability*200 as expected_revenue from (select PREDICTION_SET(SVCMODEL_JDM, 1 USING *) pset from records ) t4, TABLE(t4.pset) s4
) t
order by t.expected_revenue desc;

On-the-fly, multiple models; then sort by expected revenues
Oracle Sales Prospector
Larry Ellison Oracle Open World Keynote
November 2007

- Announces Fusion Edge CRM On-Demand Hosted Application with integrated data mining to mine customer database
How Can I Sell More?

- Which types of customers are buying which products?
- Which prospects most resemble those customers?
- Which references can I use to help me close my deals?
Oracle Data Mining = the Science of Selling

Oracle Sales Prospector
ODM Predictions exposed via Social CRM Dashboards

Oracle Database 11G
Social CRM schema ships with Oracle Database EE 11g + Data Mining Option
Oracle Data Mining predicts likelihood of purchases.

Oracle Data Mining recommends products customer is likely to buy.

Oracle Data Mining suggests likely references.
Summary
Oracle Data Mining Summary

• Powers Next-Generation Predictive Applications
  • Rapidly Build Applications that Automatically Mine Data
  • Code Once, Run Anywhere
  • Parallel and Distributed Processing
  • Industry Standard SQL and Java APIs

• Industry Leader in In-Database Data Mining
  • Option to the Industry Leading RDBMS—Oracle Database
  • Classification, Regression, Attribute Importance
  • Clustering, Market Basket Analysis, Anomaly Detection, Feature Extraction
  • Cutting Edge Algorithms: SVM, One-Class SVM, NMF, Scalable GLM
Oracle Data Mining Summary

• More Information from More Data
  • Easy to use Oracle Data Miner Graphical User Interface
  • Wide Range of In-Database Data Mining Algorithms and Statistics
  • Mine Text, Transactional, and Star Schema Data
  • Mine XML, Semantic RDF, Spatial, and OLAP Data

• Eliminate Barriers Between Analysts and IT
  • Quickly Disseminate Analytical Results and Models Throughout the Organization
  • Include Real-Time Predictive Models and New Insights in SQL queries
  • Eliminate Data Movement, Maximize Security
Getting Started
Data Mining Projects

• “The vast majority of BI professionals are excited about the prospects of data mining, but are fully mystified about where to begin or even how to prepare”
• “Of those who did initiate a modeling initiative, …51% of data mining projects either never left the ground, did not realize value or the ultimate results were not measurable”
• “In most cases, those who attempted an implementation ended up building excellent predictive models that answer the wrong questions”
• “For any organization with annual revenues more than $50 million, employing data mining technology is not a matter of whether, but when”
Getting Started with Oracle Data Mining

• You can download a free evaluation copy of Oracle Data Mining and try it out on your own computer. See the Oracle Data Mining Administrators Guide, which tells how to install a database and set up a user account. Download the Oracle Database Enterprise Edition (10gR2 or 11g) from the Oracle Technology Network. The Oracle Data Mining Option is installed by default with Oracle Database EE. For data analysts or those new to data mining, you will also want to download and install Oracle Data Miner, the free, optional graphical user interface. A summary of algorithms supported by ODM with links to the documentation is posted here.

• To get started quickly, Part I of ODM Concepts introduces you to the features and terminology of Oracle Data Mining. Then, use the Oracle Data Mining Tutorial to provide step-by-step guidance for using the Oracle Data Miner graphical interface. ... You can use the Oracle Data Miner (Data --> Import...) to import your own data in .csv text files and begin mining.

• For application developers, the ODM Application Developer's Guide along with the Oracle Data Mining sample programs gets you started writing SQL- or Java-based data mining applications.

• Some additional datasets for learning Oracle Data Mining include:
  CUST_INSUR_LTV (dmp file), CD_BUYERS (dmp file), EMPL_DATA (dmp file), LYMPHOMA (dmp file)

• Application developers can integrate predictive analytics into any report or enterprise application using ODM's server-based PL/SQL or Java APIs. See ODM Sample Programs for demo sample code.

• Oracle Data Mining Education through Oracle University
  • Installing Data Miner (Oracle By Example)
  • Solving Business Problems with Data Mining (Oracle By Example)

Copyright © 2009 Oracle Corporation
More Information:

Oracle Data Mining 11g
- oracle.com/technology/products/bi/odm/index.html

Oracle Statistical Functions

Oracle Business Intelligence Solutions
- oracle.com/bi

http://search.oracle.com

Contact Information: Email: Charlie.berger@oracle.com
“This presentation is for informational purposes only and may not be incorporated into a contract or agreement.”