Oracle Data Mining Case Study: Xerox

Session ID: S283051

Charlie Berger
Sr. Dir. Product Management, Life & Health Sciences Industry & Data Mining Technologies
Oracle Corporation
charlie.berger@oracle.com

Tracy E. Thieret
Principal Scientist, Imaging and Systems Technology Center
Xerox Innovation Group
Webster, New York

Copyright © 2006 Oracle Corporation
Agenda

1. Oracle Data Mining Overview

2. Xerox Case Study
What is Data Mining?

• Process of sifting through massive amounts of data to find hidden patterns and discover new insights

• Data Mining can provide valuable results:
  • Identify factors more associated with a target attribute (*Attribute Importance*)
  • Predict individual behavior (*Classification*)
  • Find profiles of targeted people or items (*Decision Trees*)
  • Segment a population (*Clustering*)
  • Determine important relationships with the population (*Associations*)
  • Find fraud or rare “events” (*Anomaly Detection*)
Data Mining: Find hidden Patterns

- Data Mining can find previously hidden patterns and relationships to help you:
  - Make informed predictions and…
  - Better understand customers

- Data Mining can help answer questions such as:
  - Which customers are likely to churn or attrite?
  - Which customers are likely to respond to this offer?
  - Which employees are likely to leave?
  - What “next product” should I recommend to this customer?
  - Which factors are most associated with a target attribute e.g. high value customers
  - Which customer or transactions are most “unnatural” or possibly suspicious?
Data Mining: Discover New Insights

- Data Mining uncover hidden patterns and relationships to help you:
  - Discover new segments, clusters, and subgroups and …
- Data Mining can help answer questions such as:
  - *What are the profiles subpopulations or items of interest e.g. churners, profitable customers, defective product, etc.*
  - *What natural segments or clusters exist in my data?*
  - *Which items are typically purchased together?*
  - *What items seems to fail together?*
  - *Which genes are most associated with this disease?*
Oracle Data Mining 10gR2
Oracle in-Database Mining Engine

• Oracle Data Miner (GUI)
  • Simplified, guided data mining
• Spreadsheet Add-In for Predictive Analytics
  • “1-click data mining” from a spreadsheet
• PL/SQL API & Java (JDM) API
  • Develop advanced analytical applications

• Wide range of algorithms
  • Anomaly detection
  • Attribute importance
  • Association rules
  • Clustering
  • Classification & regression
  • Nonnegative matrix factorization
  • Structured & unstructured data (text mining)
  • BLAST (life sciences similarity search algorithm)
10g Statistics & SQL Analytics
FREE (Included in Oracle SE & EE)

- 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.

Note: Statistics and SQL Analytics are included in Oracle

- Descriptive Statistics
  - average, standard deviation, variance, min, max, median
    (via percentile_count), mode, group-by & roll-up
  - 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
- Pareto Analysis (documented)
  - 80:20 rule, cumulative results table

Copyright © 2006 Oracle Corporation
In-Database Analytics

Advantages

• Data remains in the database at all times…with appropriate access security control mechanisms—fewer moving parts

• Straightforward inclusion within interesting and arbitrarily complex queries

• Real-world scalability—available for mission critical apps

• Enabling pipelining of results without costly materialization

• Scalable & Performant
  • Real-time scoring 2.5 million records scored in 6 seconds on a single CPU system
Oracle Data Mining 10g
DEMONSTRATION
Oracle Data Mining provides summary statistical information prior to data mining.
Oracle Data Mining provides model performance and evaluation viewers.

Oracle Data Mining’s Activity Guides simplify & automate data mining for business users.
Example #1:  
*Simple, Predictive SQL*

- Select customers who are more than 60% likely to purchase a 6 month CD and display their marital status

```
SELECT * from(
SELECT A.CUST_ID, A.MARITAL_STATUS,
      PREDICTION_PROBABILITY(CD_BUYERS76485_DT, 1
                             USING A.*) prob
FROM CBERGER.CD_BUYERS A)
WHERE prob > 0.6;
```
Oracle Data Mining 10g R2

**Decision Trees**

- Decision Trees
  - Classification
  - Prediction
  - Customer “profiling”

**Problem:** Find customers likely to buy a new car and their profiles

```
IF (Income >50K AND Gender=F AND Status >Single… ), THEN P(Buy Car=1)
Confidence= .77
Support = 250
```
Anomaly Detection

- “One-Class” SVM Models
  - Fraud, noncompliance
  - Outlier detection
  - Network intrusion detection
  - Disease outbreaks
  - Rare events, true novelty
## Oracle Data Mining
### Algorithm Summary 10gR2

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Decision Tree</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</td>
</tr>
<tr>
<td></td>
<td>Adaptive Bayes Network</td>
<td>Rules / transparency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Support Vector Machine</td>
<td>Wide / narrow data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></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></td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</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></td>
<td></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>NMF</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
Create Categories of Customers

Oracle Data Mining reveals important relationships, patterns, predictions & insights to the business users.
Spreadsheet Add-In for Predictive Analytics

- Enables Excel users to “mine” Oracle or Excel data using “one click” Predict and Explain predictive analytics features.

- Users select a table or view, or point to data in Excel, and select a target attribute.
Oracle Data Miner (gui)
10gR2 Summer OTN Release

• PL/SQL code generation for Mining Activities
Oracle Data Miner (gui)
10gR2 Summer OTN Release
### ORACLE Analytics vs. SAS

**1. In-Database Analytics Engine**
- Basic Statistics *(Free)*
- Data Mining
- Text Mining

**2. Development Platform**
- Java (standard)
- SQL (standard)
- J2EE (standard)

**3. Costs (ODM: $20K cpu)**
- Simplified environment
- Single server
- Security

---

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

**2. Development Platform**
- SAS Code (proprietary)

**3. Costs (SAS EM: $150K/5 users)**
- Annual Renewal Fee (~40% each year)

---

*Copyright © 2006 Oracle Corporation*
Partners
SAP Business Warehouse Connector (ODM-BW Connector)

- Seamless integration for SAP customers
- Secure
  - Data remains in database
  - Single version of truth
- Easy to use
SPSS Clementine

- NASDAQ-listed, top 25 software company
  - 35+ year heritage in analytic technologies
  - Operations in over 60 countries
- More than 95% of FORTUNE 1000 are SPSS customers
- Combine SPSS Clementine ease of use with ODM in-Database functionality & scalability
  - Build, store, browse and score models in the Database for optimal performance
- For more information:
  - SPSS – Roger Lonsberry, (312) 651-3475 or rlonsberry@spss.com
  - Oracle – Alan Manewitz, (925) 984-9910 or alan.manewitz@oracle.com
  - Oracle – Charlie Berger, (781) 744-0324 or charlie.berger@oracle.com
InforSense -- A Single Optimized Environment for Real Time Business Analytics within the Database

Oracle Functionalities:
- Data Mining
- Preprocess
- Statistics
- Text
- OLAP
- Scheduler

Oracle Decision Tree Model

Interact with (visualize) data at any step in the workflow

Oracle Data Sources

InforSenseService

Deploy the analytic workflow as a service embedding to BPEL, SFA, CRM

Deploy the analytic workflow as an Oracle Portal

SAS free analytics: leverage Oracle analytics
SQL free analytics: drag-drop application build
Visual analytics: interactive visualisation

Integrative analytics: unified analytical environment
Automated analytics: deploy to Oracle Portal and BPEL

Copyright © 2006 Oracle Corporation
Oracle Real-Time Decision Engine

For enabling Operational Business Intelligence

- Telco
- Fins
- Retail
- Health
- Travel
- Others

Contact Center

IVR

Web

ATM

Kiosk

Front Office

Oracle Real-Time Decision (RTD) Engine

Eligibility Engine

Prediction / Scoring Engine

Learning Engine

Campaign Management

Business Intelligence

Oracle Data Mining

Data Warehouse

Copyright © 2006 Oracle Corporation
## Benefits of Oracle’s Approach

<table>
<thead>
<tr>
<th>In-Database Analytics</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Platform for Analytical Applications</td>
<td>• Eliminates data movement and security exposure</td>
</tr>
<tr>
<td></td>
<td>• Fastest: Data → Information</td>
</tr>
<tr>
<td>• Wide range of data mining algorithms &amp; statistical</td>
<td>• Supports most analytical problems</td>
</tr>
<tr>
<td>functions</td>
<td></td>
</tr>
<tr>
<td>• Runs on multiple platforms</td>
<td>• Applications may be developed and deployed</td>
</tr>
<tr>
<td>• Built on Oracle Technology</td>
<td>• Grid, RAC, integrated BI,…</td>
</tr>
<tr>
<td></td>
<td>• SQL &amp; PL/SQL available</td>
</tr>
<tr>
<td></td>
<td>• Leverage existing skills</td>
</tr>
</tbody>
</table>
“This presentation is for informational purposes only and may not be incorporated into a contract or agreement.”
The role of Data Mining in Rules-based Remote Services Delivery

Tracy E. Thieret
Principal Scientist
Imaging and Systems Technology Center
Xerox Innovation Group
Webster, New York
Talk Track

• Introduction to Xerox
• Business Metrics and Requirements
• How do we get data from our devices?
• OK, we have data. Now what?
• Before you can do Data Mining…
• The Process and some Results
• The Rewards
Xerox Innovation Group Locations

PARC
El Segundo
XRCC
XRCE
WCR&T/ISTC
Stamford

Oracle OpenWorld: October 2006
Introduction to Xerox

It’s all about Documents
- Copying and Printing
- Format conversion – electronic to paper and back

How do we make money?
- Engineering design of marking products
- Chemistry and Physics of Materials
- Services around Marking and Scanning
Some Xerox Engineered Products
Full Range: Desktop to Production

Phaser 6250
Nuvera
WorkCenterPro 90
DocuColor iGen3
US Consumables Industry
$~37 Billion Annually

- Photoreceptors
  - 220,000/day
- Toners/Carriers
  - 25,000 Freight Cars/Year
- Paper & Transparencies
  - 4 Billion Sheets/Day
- Fuser Components
  - 35 Million Rolls/Year
- Inks
  - 490,000 Cartridges/Day
- Specialty Materials
  - Fuser Oil
  - Cleaner Blades

Copying
Printing
Faxing
Toner
A Highly Complex and Constrained Material

20 µm
Business Objectives:
Reducing Costs in each LoB

Engineering Design
- Providing Increased Functionality within Boundaries
  - Total Manufacturing Costs
  - Software Development

Toner Chemistry and Physics
- New Designs
- Improved Functionality

Services Delivery
- Xerox’s Internal Service Force
- Parts and Labor
- Accelerate Collective Learning

Convergence to Mature Metrics

Measure of "Goodness"

Product Launch

Time...

Product EoL
Data from Devices

Make use of external and internal standards to speed development and deployment of capabilities.

Customer Site

Active Device Agent with embedded intelligence

Network/Systems Mgmt App

On-site Solutions

standard web services technologies

Information Flow

Xerox & Partner Sites

Web Presence and Back-Office

Enhanced web access to tools and services

Device

Oracle OpenWorld: October 2006
OK – we have the data. Now what?

Deliver Data-Centric Services to Customers

- AMR: Automated Meter Reading
- ASR: Automated Supplies Replenishment
- Others in the Pipe

Feed-forward to Service Reps for Repair Hints

Knowledge Development in Engineering
Focus on Break-Fix Service

A host of Questions before you start…

- How to deploy Knowledge to Field Personnel?
- Knowledge Representation?
- Transparency?
- Ease of Knowledge Development?
- Decoupling of Cycle Times?
  - Machine Software Releases
  - Knowledge Discovery

Rules

- Out of Favor in the ’80s
- Back in with deployment of Business Rules
Where do the Rules come from?

From the knowledge of the experts

- Interviews with Engineering and Service Reps.
- Computational Capture and Analysis
- “Same problem, Different Machine”

Fast cycle time discovery of hidden Rules

- Data Mining
  - Many algorithms that deliver rule ready results
  - Triage the rules with the SMEs before deployment
  - Test for effectiveness in the field
- Using Oracle Data Mining in Xerox Research Group
Competitive Benchmarking

Data Mining Tools
META Spectrum℠ Evaluation

Our Choice
Detecting the Unexpected
Before you can do Data Mining...

- **Domain Analysis**
  - Business Hypotheses – What problem are you trying to address?
  - Cost/benefit modeling
  - Domain Knowledge Acquisition

- **Target Data Set(s)**
  - Assemble Relevant Data Sources
  - Business Processes
  - Numerical and Textual

- **Data Preprocessing**
  - SQL to summarize/aggregate data
  - Pre-computed fields

- **Data Reduction**
  - Find useful variables

- **Data Mining Task Selection**
  - Classification: Identify clusters that describe behaviors
  - Association: What variables describe the problem?

- **Algorithm Selection**
  - Statistical methods, decision trees, Bayesian nets, …

- **Data Mining**
  - Search for patterns
  - Discover knowledge

- **Interpretation of Results**
  - Explain mined patterns
  - Quantify correlations, create rules
  - Triage with SMEs

- **Deploy Knowledge**
  - Tools & documentation
  - Reports and proposals for Business Decisions and Implementation
  - Rollout & Feedback – Quantify Benefits

---

**Data Mining Expertise can be utilized for:**
- Product and Architecture Decisions
- Post Launch Product Improvement
- Expand revenue opportunities
- Post Sale Services Improvement
- Customer Relationship Management
Rewards of Data Mining

The Pleasure of Finding New Things that Matter

Corporate Financial Benefits

Personal Financial Benefits