Starting Smart with Oracle Advanced Analytics

ODTUG Kscope 15

Tim Vlamis
Michael Caskey
Vlamis Software Solutions
816-781-2880
http://www.vlamis.com

Copyright © 2015, Vlamis Software Solutions, Inc.
• Vlamis Software founded in 1992 in Kansas City, Missouri
• Developed more than 200 Oracle BI systems
• Specializes in ORACLE-based:
  • Data Warehousing
  • Business Intelligence
  • Data Mining and Predictive Analytics
  • Data Visualization
• Expert presenter at major Oracle conferences
• Authors of 2014 book “Data Visualization for Oracle BI 11g”
• Co-author of book “Oracle Essbase & Oracle OLAP”
• www.vlamis.com (blog, papers, newsletters, services)
• Developer for IRI (former owners of Oracle OLAP)
• Beta tester for OBIEE 11g
• Conference chair for BIWA Summit 2014 - 2016
Tim Vlamis and Michael Caskey

• Tim (business analyst and academic guy)
  • 25+ years in business modeling, valuation, and scenario analysis
  • Oracle ACE
  • Professional Certified Marketer (PCM) from AMA
  • Adjunct Professor of Business, Benedictine College
  • MBA Kellogg School of Management (Northwestern University)
  • BA Economics, Yale University

• Mike (IT Architect and hands-on expert)
  • 19+ years in data warehousing, software engineer and OLAP
  • 10+ years of this time in Healthcare BI as co-founder and lead architect of a software company, developing 6 product solutions
Presentation Agenda

• Background on Analytic Options to the Oracle DB
• Oracle Advanced Analytics
  • Oracle Data Mining
  • Oracle R Enterprise
• How to start with OAA – comparison of options
• Demo of OAA with Oracle BI Front end - SampleApp
• Bonus info on Oracle In Memory
Analytical Options to Oracle Database

• Oracle OLAP
  • Defines a multi-dimensional data structure that allows information for highly complex calculations to done quickly.
  • Fast query performance and incremental update
  • Simplified access to analytic calculations

• Oracle Advanced Analytics (Data Mining & R)
  • Refers to the process of automatically sifting through data to find hidden patterns and make predictions.
  • Series of highly advanced algorithms and procedures.
  • Extends the “R” language to the Oracle Database

• Oracle Spatial & Graph
  • Provides the capability of relating data to geo positional coordinates, objects, and constructs.
  • Allows the construction and analysis of network topologies.

• Oracle In-memory
  • Provides fast summarization of columnar data
<table>
<thead>
<tr>
<th>OLAP</th>
<th>Data Mining &amp; R</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summaries, hierarchies and dimensional data</td>
<td>Knowledge discovery of hidden patterns</td>
<td>Spatial relationships between data</td>
</tr>
<tr>
<td>&quot;Analysis&quot;</td>
<td>&quot;Insight &amp; Prediction&quot;</td>
<td>&quot;Location&quot;</td>
</tr>
<tr>
<td>What is the average income of mutual fund buyers, by region, by year?</td>
<td>Who is likely to purchase a mutual fund in the next 6 months and why?</td>
<td>Where were mutual funds purchased in the last 3 years?</td>
</tr>
</tbody>
</table>
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 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

Results

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

Secs. Mins or Hours

Data remains in the Database

Model “Scoring”

Data remains in the Database

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 © 2015, Vlamis Software Solutions, Inc.
Oracle Data Mining

- Oracle Data Mining is an option for the Enterprise Edition of the Oracle Database.
- A collection of APIs and specialized SQL functions.
- Includes a large number of specialized algorithms and built-in procedures.
- Makes use of many built-in capabilities of the Oracle Database.
- ODM typically refers to "Oracle Data Mining"
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
### 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></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Linear 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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Unknown fraud cases or anomalies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Minimum Description Length</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td></td>
<td>Principal Component Analysis</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>Next Best Offer</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>Nonnegative Matrix Factorization (NMF)</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td>Singular Value Decomposition (SVD)</td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
• 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
Oracle Data Miner

- Easy to Use
  - Oracle Data Miner GUI for data analysts
  - “Work flow” paradigm
- Powerful
  - Multiple algorithms & data transformations
  - Runs 100% in-DB
  - Build, evaluate and apply models
- Automate and Deploy
  - Save and share analytical workflows
  - Generate SQL scripts for deployment
Understand Model Details

- Interactive model viewers

Copyright © 2015, Vlamis Software Solutions, Inc.
ODM’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools.
Dynamically Using ODM From Oracle BI

Model Attributes Significance
Time run: 5/15/2014 7:37:48 AM

Most Significant Attributes in the Model

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.117</td>
</tr>
<tr>
<td>MARRITAL_ST</td>
<td>0.009</td>
</tr>
<tr>
<td>CREDIT_RATE</td>
<td>0.028</td>
</tr>
<tr>
<td>INCOME_LVL</td>
<td>0.027</td>
</tr>
<tr>
<td>MONTHLY_CONTACT</td>
<td>0.005</td>
</tr>
</tbody>
</table>

LTV Probability by Ages

- Probab Very High LTV
- M13 Cust Probab Low LTV

LTV Probability by Income Level

<table>
<thead>
<tr>
<th>Income Level</th>
<th>Probab Very High LTV</th>
<th>Probab High LTV</th>
<th>Probab Low LTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 - 49</td>
<td>39.0%</td>
<td>52.0%</td>
<td>3.0%</td>
</tr>
<tr>
<td>30 - 39</td>
<td>45.0%</td>
<td>43.0%</td>
<td>12.0%</td>
</tr>
<tr>
<td>20 - 29</td>
<td>48.0%</td>
<td>41.0%</td>
<td>11.0%</td>
</tr>
<tr>
<td>10 - 19</td>
<td>50.0%</td>
<td>39.0%</td>
<td>11.0%</td>
</tr>
<tr>
<td>0 - 9</td>
<td>52.0%</td>
<td>37.0%</td>
<td>11.0%</td>
</tr>
</tbody>
</table>

SQL Code:

```
FROM ODM_SAMP_CUSTOMERS_V C;
```
What is R?

- Popular alternative to SAS, SPSS & other proprietary statistical environments
- Around 2 million R users worldwide
- Thousands of R packages available
2013 Rexer Analytics Data Miner Survey

- 6\textsuperscript{th} survey since 2007
- 68 questions
- 10,000+ invitations emailed, plus promoted by newsgroups, vendors, and bloggers
- Respondents: 1,259 data miners from 75 countries
- Data collected in first half of 2013

*Data from software vendors is excluded from analyses in this presentation unless otherwise noted.
The proportion of data miners using R is rapidly growing, and since 2010, R has been the most-used data mining tool. While R is frequently used along with other tools, an increasing number of data miners also select R as their primary tool. Among data miners who say they are likely to switch their primary package in the coming year, R is frequently identified as the tool they are plan to switch to – more than 2.5 times more often that any other tool.
Oracle R Enterprise

• Part of the Advanced Analytics Option to the Oracle Database Enterprise Edition
• Provides transparent access to database-resident data from R
• Embedded R script execution through database managed R engines with SQL language integration
• Provides data and task parallelism and full power of Oracle database for R
• Enables advanced statistics for in-database execution
• Integrates R into the IT software stack
• Extends and enhances open source R
Oracle R Distribution

- Improved scalability at client and database for embedded R execution
- Enhanced linear algebra performance using Intel’s MKL, AMD’s ACML, and Solaris Sun Performance Library
- Enterprise support for customers of Oracle Advanced Analytics option, Big Data Appliance, and Oracle Linux
- Available as a free download from Oracle
- Oracle to contribute bug fixes and enhancements to open source R

ability to dynamically load:
InteMath Kernel Library (MKL)
AMDCore Math Library
Solaris Sun Performance Library

Oracle Support
Other R Offerings

- **ROracle**
  - Open source Oracle database interface driver for R based on OCI
  - Maintained by Oracle, optimizations and bug fixes released to open source community

- **Oracle R Advanced Analytics for Hadoop**
  - R interface to Oracle Hadoop Cluster on Big Data Appliance
  - Access and manipulate data in HDFS, database, and file system
  - Write MapReduce functions using R and execute through R

- **Rstudio**
  - Popular open source user interface for R
  - Integrated Development Environment

- **Rcommander**
  - Extended GUI for R
  - Used for plugins (e.g. text mining)
  - Often used with Rstudio
R now integrated into OBIEE 11.1.1.7

QCC-1 Graphics Result

xbar Chart for arrival delay - years 1987 - 1999

S Chart for arrival delay - years 1987 - 1999

new data 2000 - 2008

new data 2000 - 2008

Calibration data in set1
New data in set2

Calibration data in set1
New data in set2

Number of groups = 22
Comar = 7.209899
StdDev = 29.19689
LCL is variable
UCL is variable
Number beyond limits = 16
Number violating limits = 0

Group summary statistics

Copyright © 2015, Vlamis Software Solutions, Inc.
require(gplots)

set.seed(120)
# simulate an AR(1) process
coeffs <- 0.95
data <- arima.sim(list(order=coeffs), n=250)

# fit AR(1) with the 200 first data
model <- arima(data[1:200], c(1, 0, 0))

# make forecast from the model
forecast <- predict(model, 80)

# compute the limits of the graph
ylim <- c(min(data[1:200], forecast$pred - 1.96 * forecast$se),
          max(data[1:200], forecast$pred + 1.96 * forecast$se))

# prepare the space where to plot
opar <- par(mar=c(4,4,2,2), las=1)

plot(data, ylim=ylim, type="n", xlab="Time")
usr <- par("usr")

# split the figure in two parts
# - the part used to fit the model
rect(usr[1],usr[3],201 ,usr[4],border="NA",col="lemonchiffon")

# - the part used to make the forecast
rect(201 ,usr[3],usr[2],usr[4],border="NA",col="lavender")

abline(h= c(-3:3)*2 , col = "gray" , lty =3)

# draw a 95% confidence band
polygon( c(201:280,280:201),
         c(forecast$pred - 1.96*forecast$se,rev(forecast$pred + 1.96*f
         col = "orange",
         lty=2,border="NA")

lines(201:280 , forecast$pred - 1.96*forecast$se , lty=2)
lines(201:280 , forecast$pred + 1.96*forecast$se , lty=2)
Customer “most likely” be HIGH and VERY HIGH value customer in the future.
What is Spatial Data?

• Business data that contains or describes location
  • Street and postal address (customers, stores, factory, etc.)
  • Sales data (sales territory, customer registration, etc.)
  • Assets (cell towers, pipe lines, electrical transformers, etc.)
  • Geographic features (roads, rivers, parks, etc.)

• Anything connected to a physical location

• Any data sets that contain “link and node” relationships between data objects. Can be directional or non-directional.
How to Get Started

- Lots of internal experts and lots of people who would like to be involved and learn
- Lots of people intimidated by what they don’t know
- Start by “level setting” and establishing a strong foundation with basic training (2 days)
  - Bring people along on the journey, establish culture
  - Everyone shares a minimum common knowledge base
- Immediately conduct a workshop (JAD style session) investigation of possibilities
  - Evaluation of data sources and data sets
  - Recognition of major business issues
  - Review of basic algorithms
  - Identification of potential PoC projects (plusses and minuses)
- Decide on pilot projects and who works on it
- Start simple and return value quickly
Need Representatives from Across Organization

- Get different perspectives
- Data Mining requires synergistic insights
- Builds broad support
- Non-siloed
- Takes Data mining out of IT and into Line of Business
- Participants learn from each other
- Formal statistics background not required
- Only prerequisite: analytical mind set / love data
Methodology

• Implement a highly scalable infrastructure
• Establish a common foundational understanding of data mining
• Demonstrate the Value of Analytics by Completing a Market Basket Project Immediately
ODM Machine Bundle Overview

- **Hardware**
  - Oracle Database Appliance/Oracle Database Cloud Service

- **Software**
  - Oracle Database 12c (with options)
  - Oracle Advanced Analytics Option including Oracle Data Mining
  - Oracle SQL Developer: Data Miner Add-in (free download)

- **Services**
  - Implementation and configuration from Vlamis Software Solutions (Oracle Gold Partner)
  - Oracle University Oracle Data Mining Techniques course (taught by Vlamis Software Solutions)
  - Market Basket Analysis Project performed on company data

- **Time frame:** 9 business days (less than 2 weeks)
Compressed Schedule

• Day 1:
  • Two consultants meet with client team to review project plan, review data sources, identification of best data to start with, set technical objectives for project (basic market basket analysis deliverable)

• Day 2:
  • Consultant One: Install ODA and configure to network (need support from client tech staff)
  • Consultant Two: Conduct first day of ODM class with client team

• Day 3:
  • Consultant One: Install new pluggable Database, SQL Developer
  • Consultant Two: Conduct second day of ODM class with client team

• Day 4:
  • Two consultants establish data plan for project with client and import data

• Day 5:
  • Consultant One: Prepare tables for mining (add keys, new tables, transforms, etc.)
  • Consultant Two: Document data plan

• Day 6:
  • Consultant Two: Build market basket workflow

• Day 7:
  • Consultant Two: Conduct market basket analyses

• Day 8:
  • Consultant Two: Prepare presentation of findings from market basket analyses

• Day 9:
  • Consultant Two: Deliver presentation with client
• Introduction
• Data Mining Concepts and Terminology
• The Data Mining Process
• Introducing Oracle Data Miner 11g Release 2
• Using Classification Models
• Using Regression Models
• Using Clustering Models
• Performing Market Basket Analysis
• Performing Anomaly Detection
• Deploying Data Mining Results
Oracle R Enterprise Training (2 Days)

- Oracle R Enterprise technologies introduction
- Introduction to R hands-on
- ORE transparency layer with hands-on exercises
- ORE embedded R execution with hands-on exercises
- ORE predictive analytics with hands-on exercises
- Using ROracle
- Overview of ORE with OBIEE
Comparison of Training Courses

Oracle Data Mining
- Organized by algorithm
- Intro to data mining
- MBAs, BI Admin, DBAs
- Focused on business issues
- Uses GUI
- Approachable for new users

Oracle R Enterprise
- Organized by process
- Intro to Oracle R Enterprise
- Data Scientists, BI Admin, DBAs
- Focused on executing R in Oracle Database
- Uses R scripts
- Technical
Oracle In Memory

• Announced at Oracle OpenWorld 2013
• Launched by Larry Ellison June 2014
• Allows Oracle tables to be loaded into memory
• Super-fast scans of data using SIMD and vector instruction sets
• Works with OLTP
• First release likely limited in scope
• Allows for more flexible aggregation
Oracle In Memory Predictions

• In Memory capabilities will start overlapping with Oracle OLAP capabilities
• Allows for more flexible deployments
• No need to pre-define hierarchies
• Does not have metadata for pre-defined hierarchies yet
• Does not have capabilities for OLAP calculations yet
• Does not involve misunderstood cube technology – simple for DBAs to understand
Oracle Test Drive

• Free to try Oracle BI, Advanced Analytics and Big Data
• Go to www.vlamis.com/td
• Runs off of Amazon AWS
• Test Drives for:
  • Oracle BI
  • Oracle Advanced Analytics
  • Big Data
• Once sign up, you have private instance for 5 hours
• Available now
<table>
<thead>
<tr>
<th>Presenter</th>
<th>Session</th>
<th>Time</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tim and Dan Vlamis</td>
<td>Session 1</td>
<td>Monday 8:30 - 9:30 AM</td>
<td>Forecasting, Prediction Models, and Time Series Analysis with Database Analytics and OBIEE</td>
</tr>
<tr>
<td>Dan and Tim Vlamis</td>
<td>Session 4</td>
<td>Monday 2:00 – 3:00 PM</td>
<td>Data Visualization for Oracle Business Intelligence 11g</td>
</tr>
<tr>
<td>Tim Vlamis and Michael Caskey</td>
<td>HOT-EPM</td>
<td>Tuesday 3:30 – 5:45 PM</td>
<td>Hands-on Training: Integrating Oracle Advanced Analytics into OBIEE Dashboards</td>
</tr>
<tr>
<td>Tim Vlamis and Michael Caskey</td>
<td>Session 11</td>
<td>Wednesday 8:30 - 9:30 AM</td>
<td>Starting Smart in Oracle Advanced Analytics</td>
</tr>
<tr>
<td>Mark Rittman, Alex Gorbachev and Tim Vlamis</td>
<td>Deep Dive</td>
<td>Thursday</td>
<td>Bringing Oracle Tools to Big Data</td>
</tr>
</tbody>
</table>
Thank You for Attending Session **Starting Smart with Oracle Advanced Analytics**

**Presenter Information**
Tim Vlamis  
Michael Caskey  
Vlamis Software Solutions, Inc.  
816-781-2880  
tvlamis@vlamis.com  
mcaskey@vlamis.com

For more information go to [www.vlamis.com](http://www.vlamis.com)