

ODTUG

Kscope16



CHICAGO, ILLINOIS • JUNE 26-30

PLEASE FILL OUT YOUR EVALUATIONS

Clustering Data with Oracle Data Mining: The Easiest Place to Start in Predictive Analytics

KScope16

Tim Vlamis

Tuesday, June 28, 2016

VlamiS Software Solutions

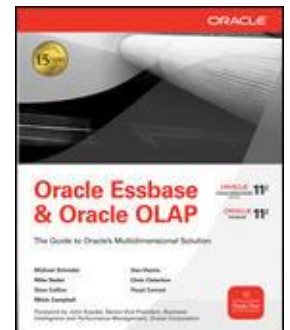
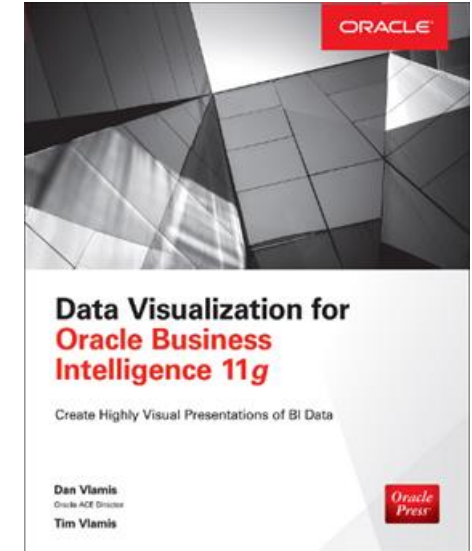
- VlamiS Software founded in 1992 in Kansas City, Missouri
- Developed 200+ Oracle BI and analytics systems
- Specializes in Oracle-based:
 - Enterprise Business Intelligence & Analytics
 - Analytic Warehousing
 - Data Mining and Predictive Analytics
 - Data Visualization
- Multiple Oracle ACEs, consultants average 15+ years
- www.vlamiS.com (blog, papers, newsletters, services)
- Co-authors of book “Data Visualization for OBI 11g”
- Co-author of book “Oracle Essbase & Oracle OLAP”
- Oracle University Partner
- Oracle Gold Partner

 EDUCATION RESELLER

 APPROVED
EDUCATION CENTER


 Gold
Partner

Specialized
Oracle Business Intelligence
Foundation Suite 11g





Vice President & Analytics Strategist

- 30+ years in business modeling and valuation, forecasting, and scenario analyses
- Oracle ACE 
- Instructor for Oracle University's Predictive Analytics, Data Mining Techniques and Oracle R Enterprise Essentials Courses
- Professional Certified Marketer (PCM) from AMA
- Adjunct Professor of Business Benedictine College
- MBA Kellogg School of Management (Northwestern University)
- BA Economics Yale University



Vlami Involvement in Presentations

Presenter	Time	Location	Title
Dan Vlami & Arthur Dayton	Mon 8:30 AM	Mayfair	Upgrading to Oracle Business Intelligence 12c
Dan Vlami & Tim Vlami	Mon 4:30 PM	Mayfair	Data Visualization for Oracle Business Intelligence
Tim Vlami	Tues 8:30 AM	Missouri	Clustering Data with Oracle Data Mining: The Easiest Place to Start in Predictive Analytics
Arthur Dayton	Tues 11:15 AM	Superior A	Data Discovery Best Practices with Visual Analyzer – Hands On Lab
Tim Vlami & Dan Vlami	Tues 2:00 PM	Mayfair	Visual Analyzer and Best Practices for Data Discovery through Data Visualization



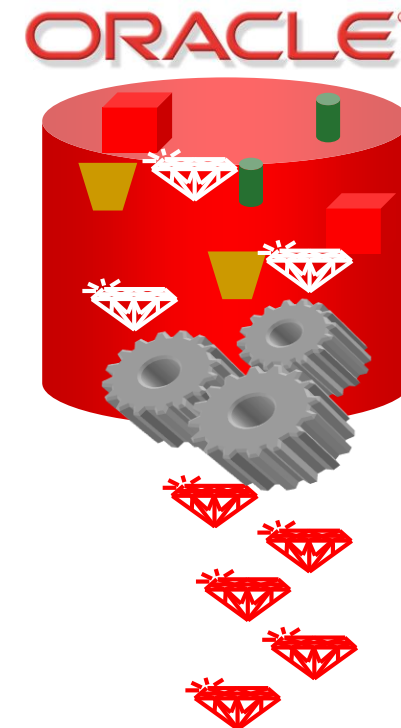
Presentation Agenda

- Background on Data Mining and Oracle Advanced Analytics
- What is clustering?
- Use cases for clustering
- Market Segmentation
- Three different algorithms
- Your questions and comments at all times!



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 co-occurrences and “market baskets” within an event set (*Associations*)
 - Find fraudulent or “rare events” (*Anomaly Detection*)

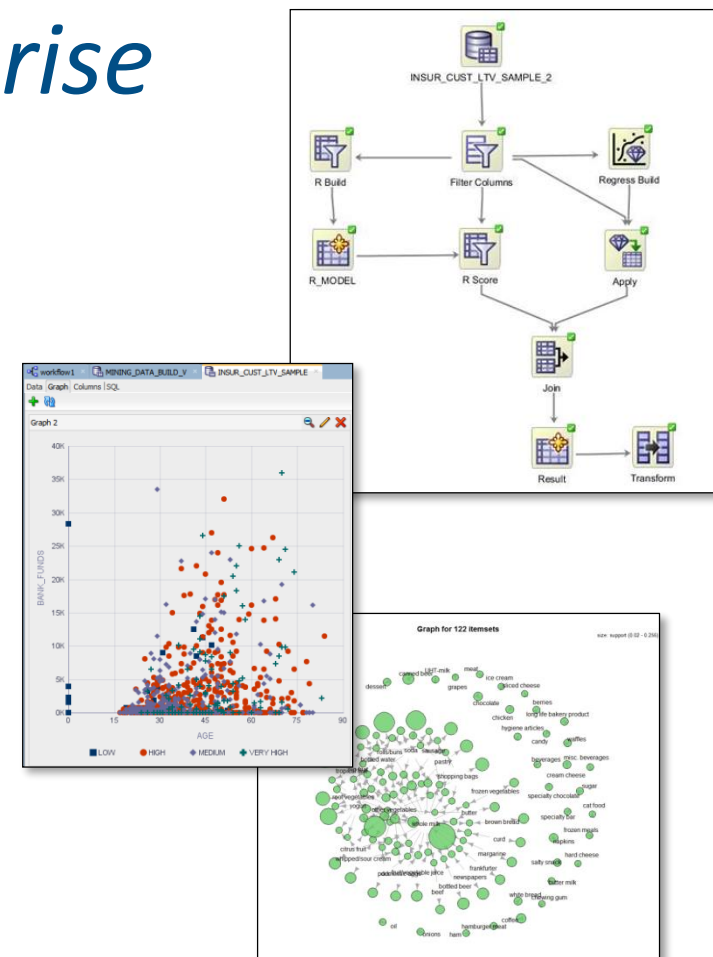




Oracle Advanced Analytics (OAA) DB Option

Oracle Data Mining + Oracle R Enterprise

- Powerful **in-database** algorithms for **Data Mining** and **Statistical Analysis**
- **Easy** to add **predictive analytics** to enterprise applications and BI
- **Fastest** way to deliver **scalable, enterprise-wide** predictive analytics
- ORE eliminates R's limitations (memory and speed) for **Enterprise-scale analytics**



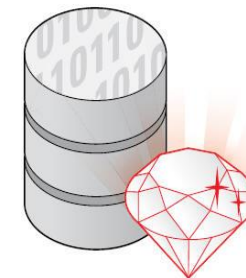
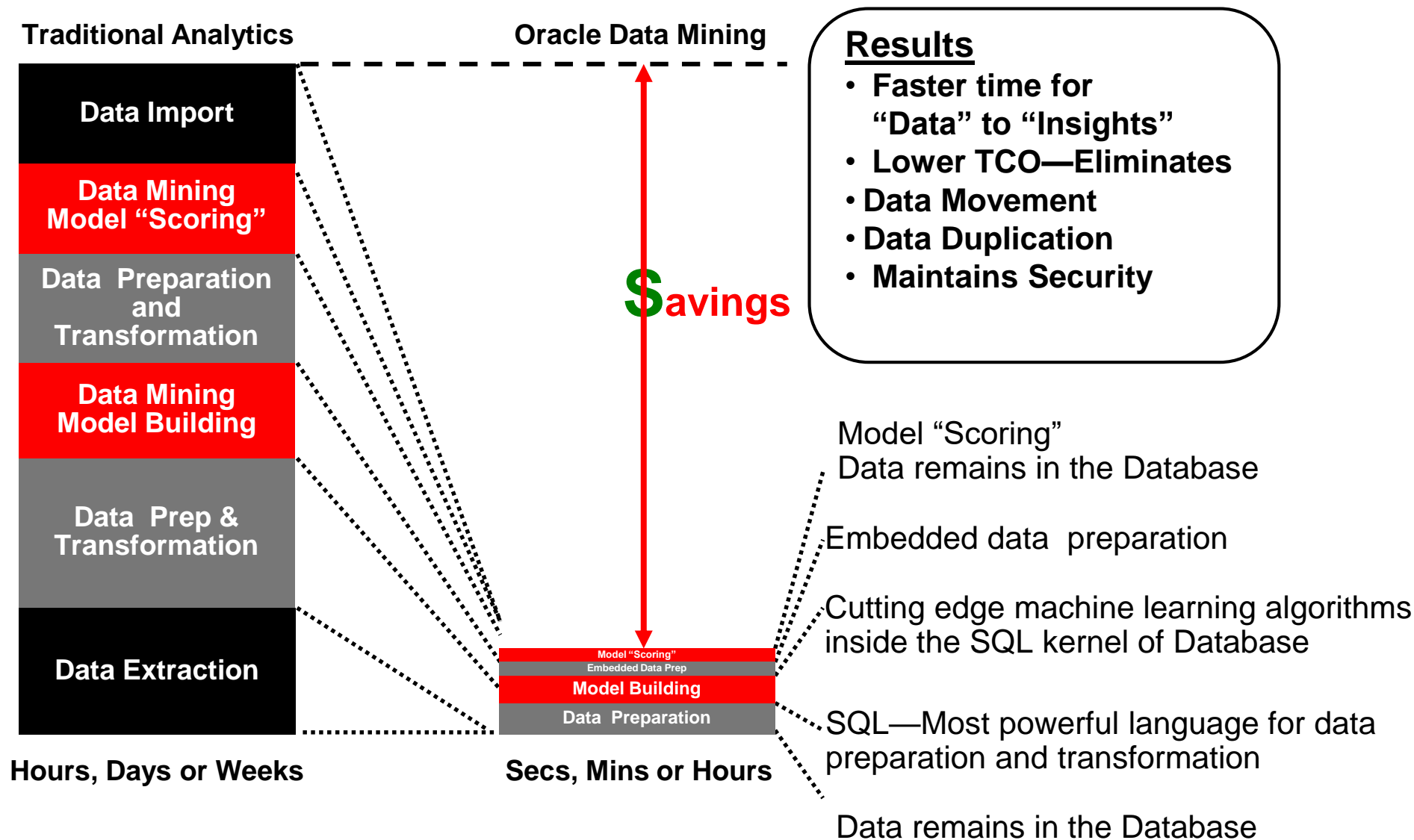


Oracle Data Mining

- 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.
- Automated data preparation
- Makes use of many built-in capabilities of the Oracle Database
- ODM typically refers to “Oracle Data Mining”

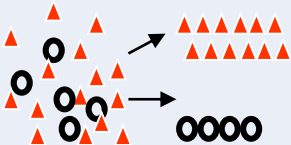

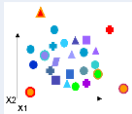
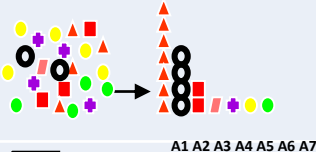
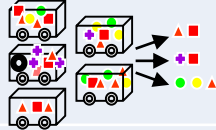
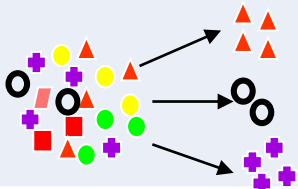



In-Database Data Mining



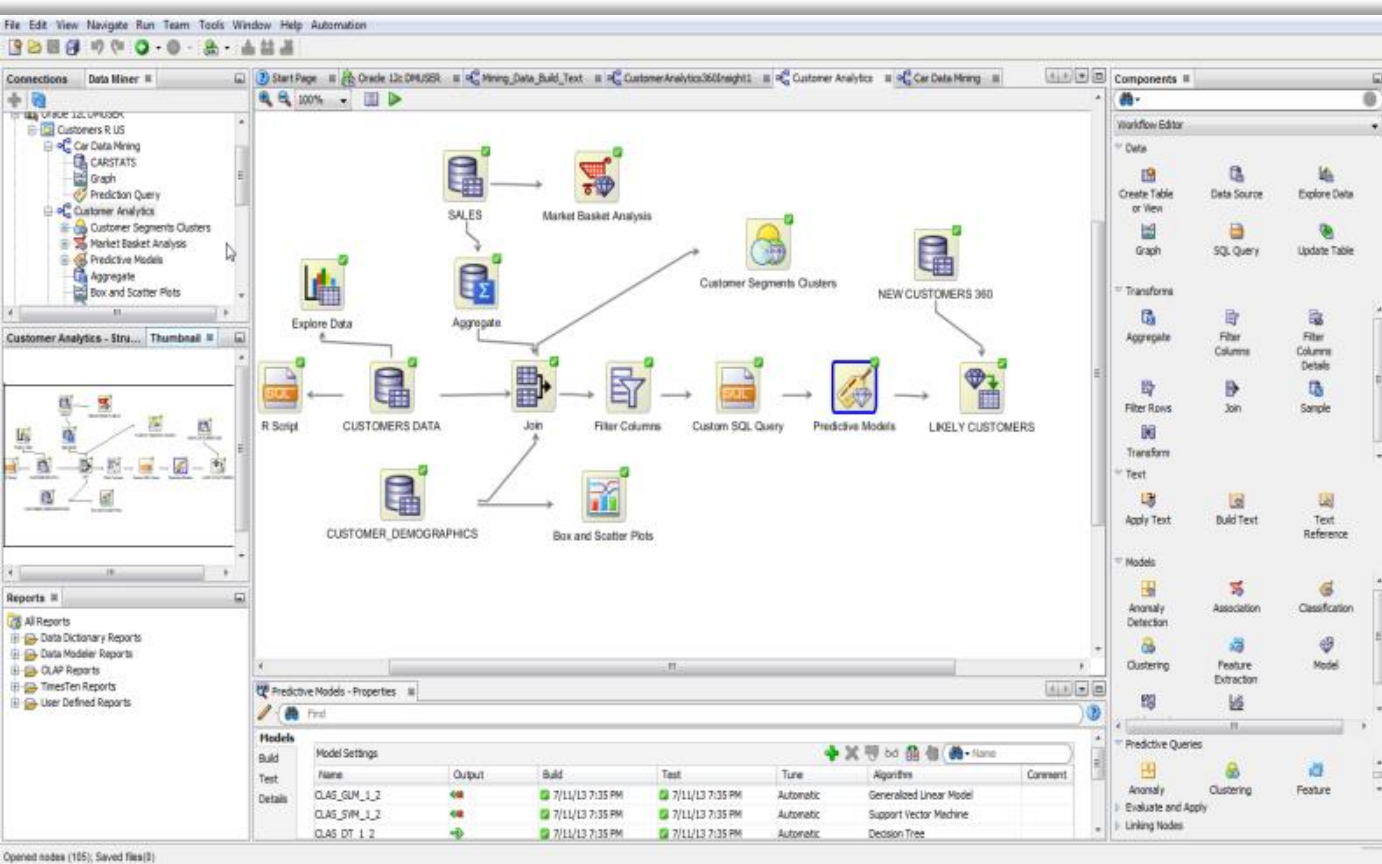


Oracle Data Mining Algorithms

Problem	Algorithm	Applicability
Classification 	Logistic Regression (GLM) Decision Trees Naïve Bayes Support Vector Machine	Classical Statistical Technique Popular/Rules/Transparency Embedded app Wide/Narrow Data or Text
Regression 	Linear Regression (GLM) Support Vector Machine	Classical Statistical Technique Wide/Narrow Data or Text
Anomaly Detection 	One Class SVM	Unknown fraud cases or anomalies
Attribute Importance 	Minimum Description Length Principal Component Analysis	Attribute reduction Reduce data noise
Association Rules 	Apriori	Market Basket Analysis
Clustering 	Hierarchical K-Means Hierarchical O-Cluster Expectation Maximization	Market Segmentation Product / Location Groupings Text analysis
Feature Extraction 	Non-negative Matrix Factorization Singular Value Decomposition	Feature Reduction Text Analysis



Oracle Data Miner

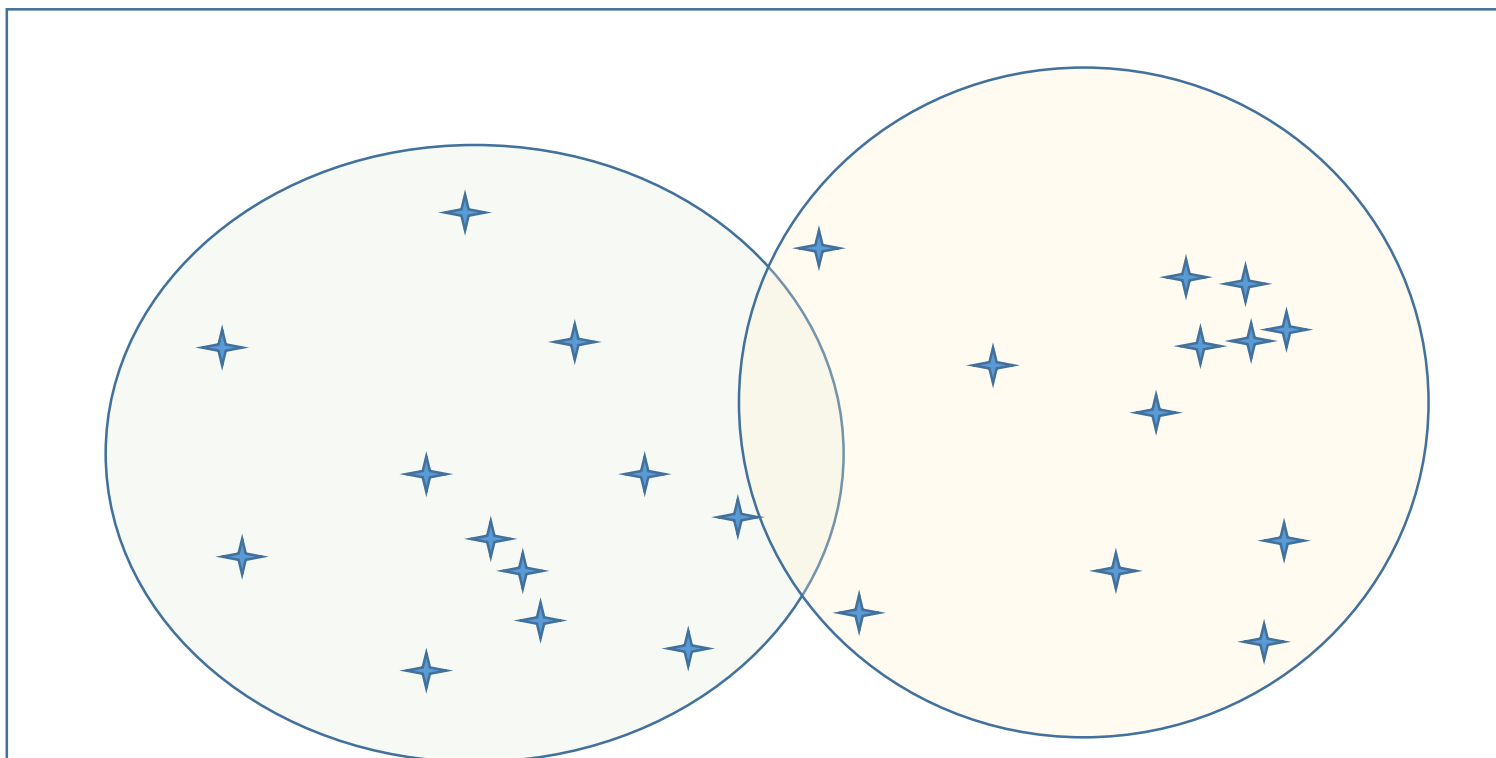


- Easy to Use
 - Oracle Data Miner GUI for data analysts
 - “Work flow” paradigm
- Powerful
 - Multiple algorithms & data transformations
 - Runs 100% in-DB
 - Build, evaluate and apply models
- Automate and Deploy
 - Save and share analytical workflows
 - Generate SQL scripts for deployment



What is Clustering?

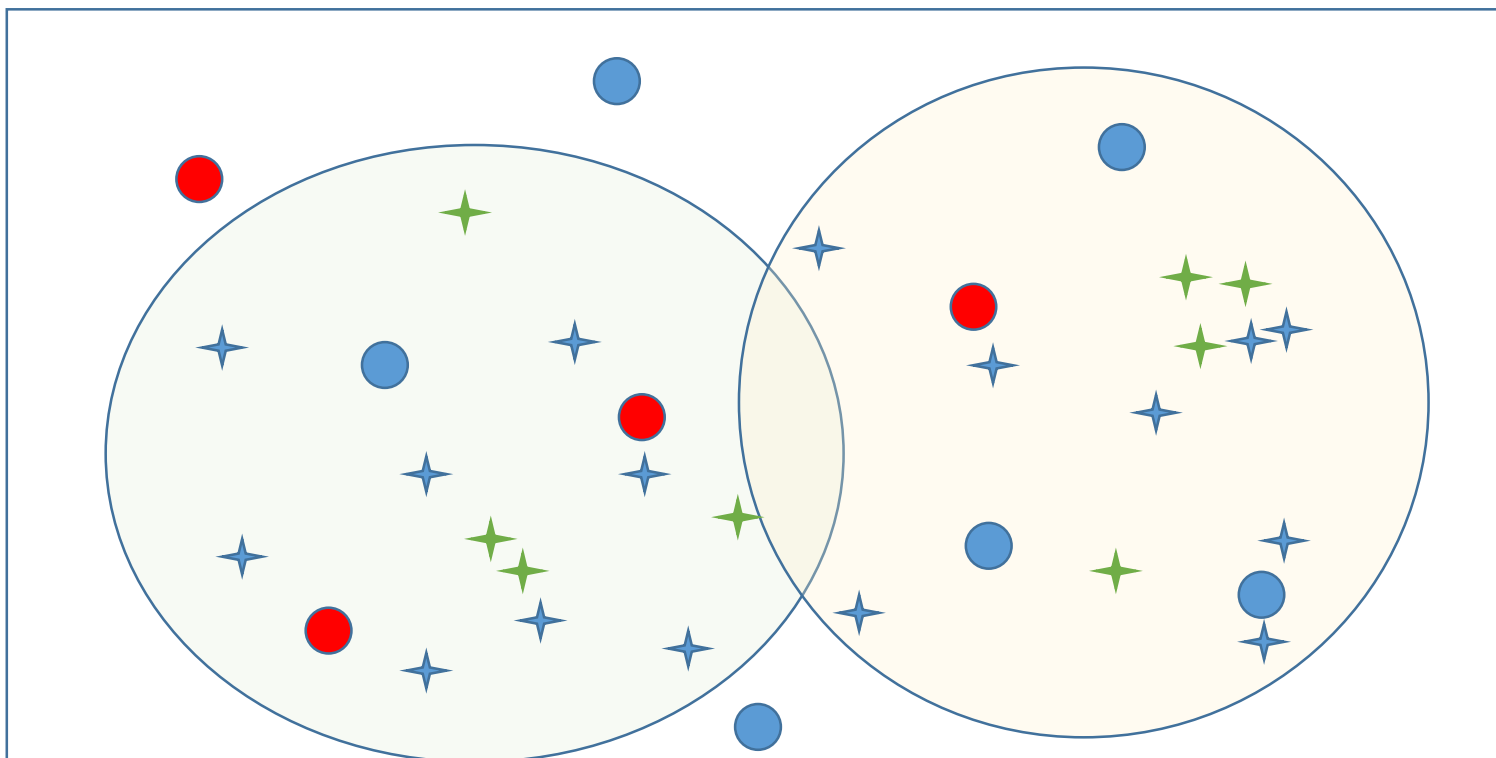
- Dividing a large set into smaller groups of similar and dissimilar members





More Dimensions Makes Clustering Harder

- It's hard to visualize clusters with high dimensionality





There is no “right” or “wrong” in clustering

- Clustering algorithms define a procedure and produce results, understanding that procedure is fundamental to explaining cluster assignments
- Large data sets can produce an uncountable number of clustering results
- Cluster assignments are also non-assignments
- Highly-dimensional data is harder to cluster than lower dimensioned data





Common use cases for clustering

- 1) Customer Segmentation using Clustering algorithms
 - Discovered patterns can be extremely meaningful
 - Able to include hundreds of dimensions
 - Great first project
- 2) Understand retail locations
 - Group business locations into similar groups
 - Discovers “like” locations
- 3) Understand website visits/sessions
 - Discover similar groups among highly dimensioned website data sets
 - Behavior findings are often surprising



Go to demo

- Let's cluster some data
- Show a tree
- Show some clusters
- Get people comfortable with what we are doing



Four Realms of Analytics

Probability Based

**Diagnostic
Analytics**

**Predictive
Analytics**

Rules Based

**Descriptive
Analytics**

**Prescriptive
Analytics**

Past

Future



Five Dimensions of Market Segmentation

- Demographics
 - Facts about people or businesses
- Geographics
 - Locations of people or businesses
- Psychographics
 - Attitudes, beliefs, preferences
- Behavior
 - Actions and activities
- Association/Affiliation
 - Groups that are joined through self-selection/choice

Demographics are Facts



AMERICAN
FactFinder



Feedback FA

MAIN

COMMUNITY FACTS

GUIDED SEARCH

ADVANCED SEARCH

DOWNLOAD CENTER

▼ Community Facts

Find popular facts (population, income, etc.) and frequently requested data about your community.

Enter a state, county, city, town, or zip code:

e.g., Atlanta, GA

GO

▶ Guided Search

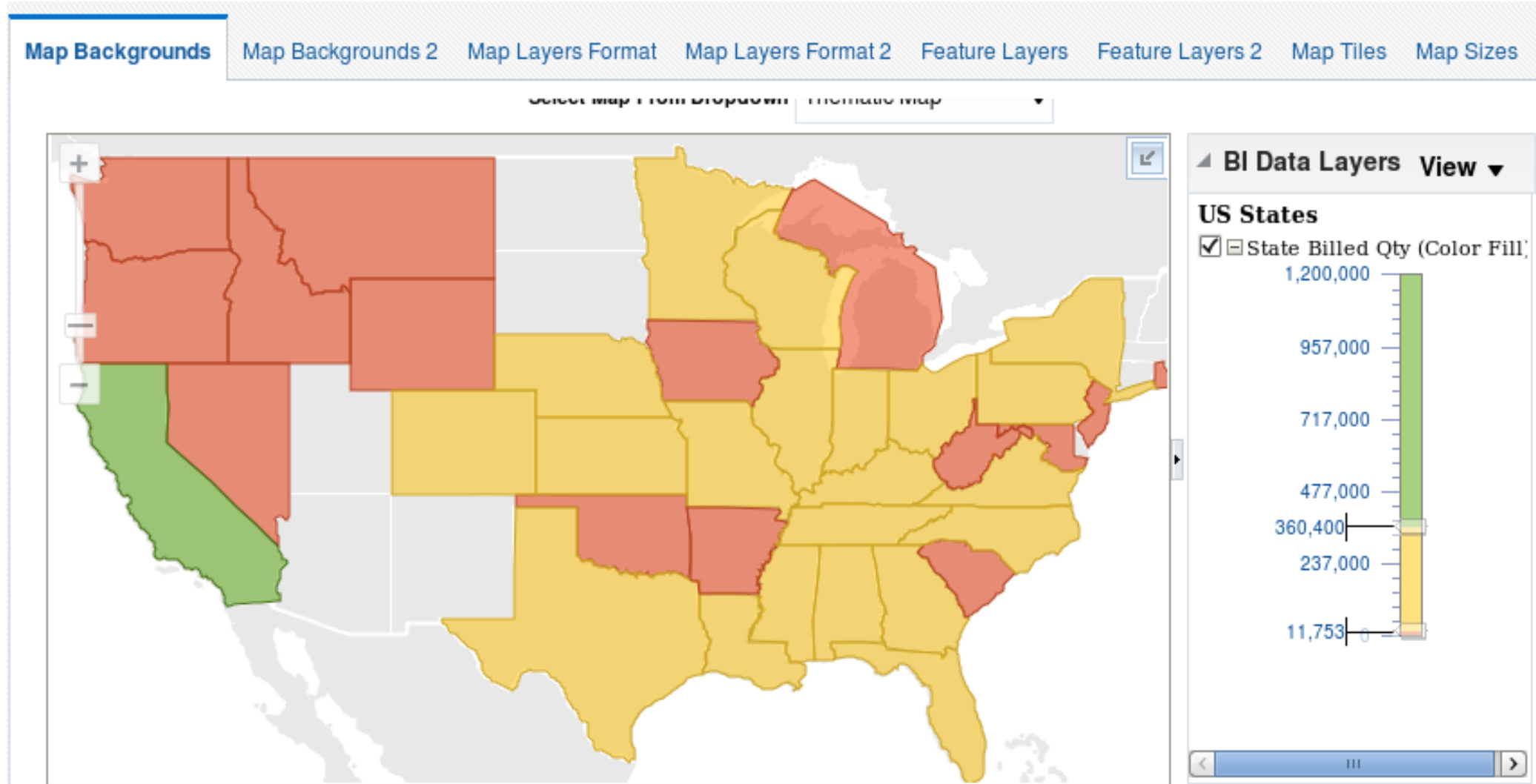
▶ Advanced Search

▶ Download Center



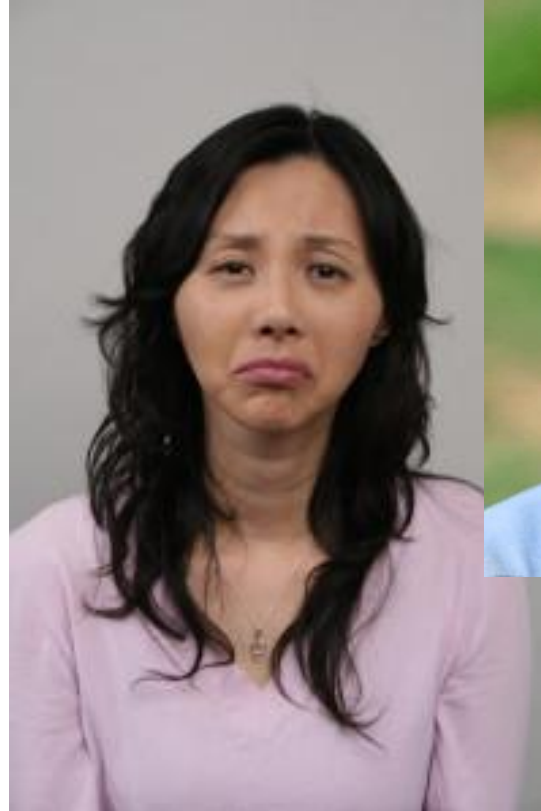
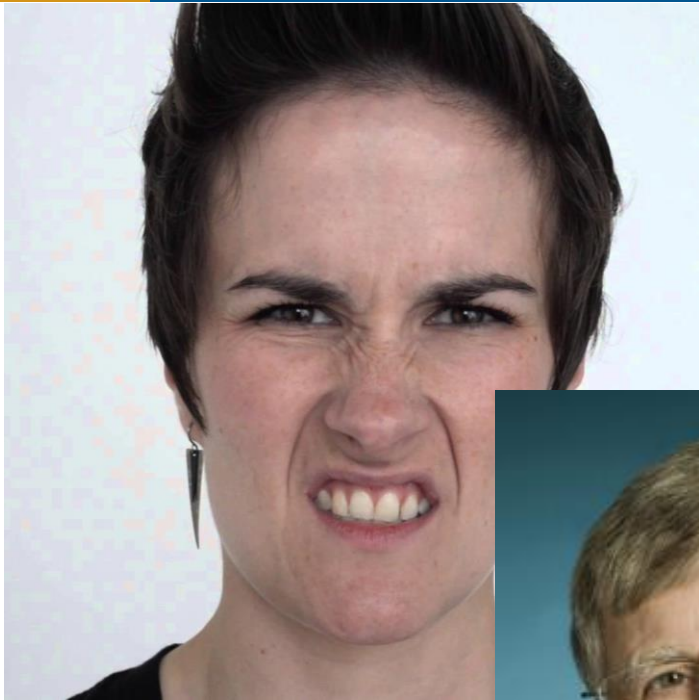


Geographics are about Location





Psychographics are attitudes, beliefs, likes





Behaviors are actions and activities





Association/Affiliation (subset of behavior)

- Self-selected groups that are joined





Marketing Segmentation

- Used for “target marketing”
- Has traditionally relied on demographics
- Mass marketing is expensive, target marketing has higher returns
- Behavior data is more available now
- Behavior data is usually much superior



Three Fundamental Methods to Cluster

- Distance (K-Means)
- Division (O-Cluster)
- Density (Expectation Maximization)



Distance – K Means

- The centroids of a specific number of clusters are placed so that they minimize the total distance between all data points and the centroids
- Imagine centroid dots moving around until they settle into position
- Most common clustering methodology
- Easy to explain mathematically (closely related to regression)
- User chooses number of (leaf) clusters
- Can be used with nested tables



Division O-Cluster

- Algorithm divides space with straight lines through areas of minimal density (orthogonal partitioning)
- Imagine lines slicing through and “tessellating” the data space
- User sets minimal level of density for finding clusters
- Excellent for extremely large data sets
- No predetermined number of clusters
- Oracle patented algorithm



Density – Expectation Maximization

- Algorithm finds areas of high density
- Imagine a population “heat map” that shows areas of high population densities with irregular shapes
- Excellent for data of diverse type (text, numeric, attribute)
- Clusters can be of irregular shape and size
- “state of the art” clustering



Rules in ODM Cluster Models

- Describe cluster centroids
- Not determinate (not applicable to every cluster member)
- Rule minimum support can be set
- Can be used for future classification models (cluster assignment)
- Use for guidance and understanding



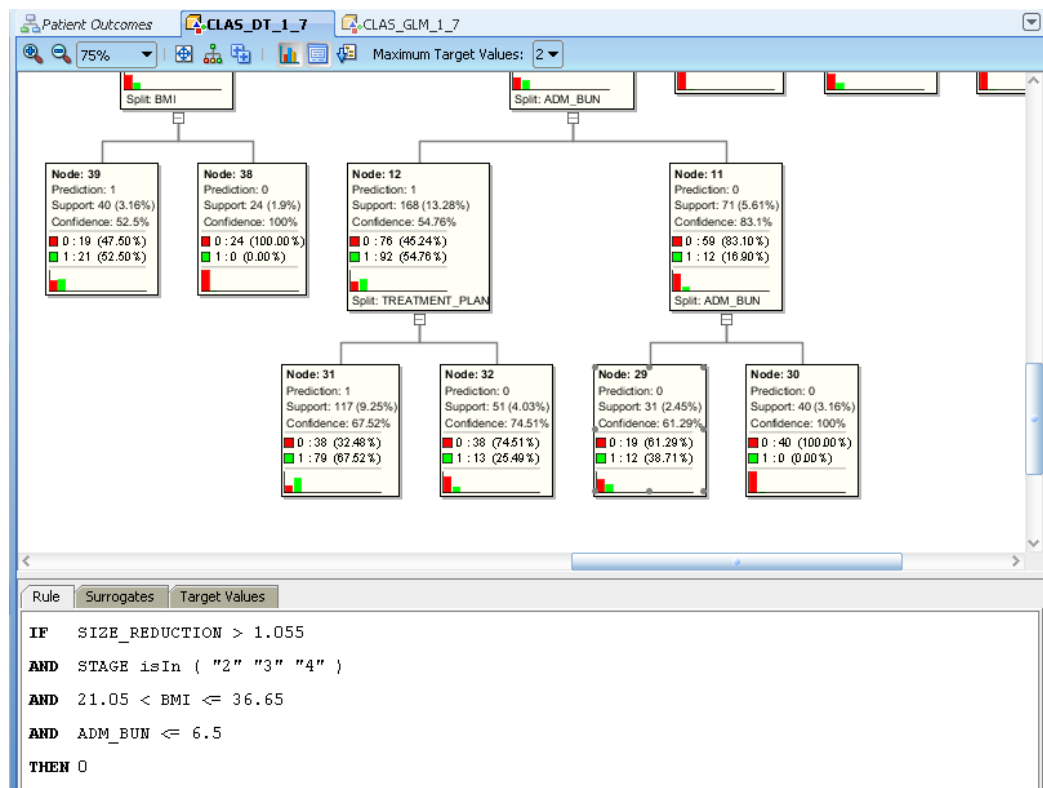
Hierarchical Clustering

- Computationally intensive
- Requires more iterations
- Hierarchical clusters are “grown” within ODM’s clustering algorithms
- Position within a “tree” carries important implications
- Two different strategies for normalizing leaf sets
 - Percentage of population splits (number of members)
 - Produces regularly-sized clusters with bigger leaves at the top and smaller leaves at the bottom of the tree
 - Use clustering to assign a predictable number of customers to different sales persons
 - Differences in population splits
 - Produces more meaningful clusters
- Number of clusters defines the number of leaves. Total number of clusters = # of leaves * 2 – 1 (10 clusters means 10 leaf clusters that are not split and 19 total clusters in tree)



Understand Model Details

■ Interactive model viewers



Target Value: 1
Sort by: absolute value
Fetch Size: 10,000

Coefficients: 297 out of 297

Attribute	Value	Coefficient	Standardized Coefficient	Exp(Coefficient)
<Intercept>	NULL	-1.83481346	0	6.26396556
TREATMENT_PLAN	Chemo_only	-0.46513283	0.11735002	1.59222567
WEEKDAY	W	-0.40697858	0.0869471	1.50227193
WEEKDAY	Th	-0.34941526	0.05883753	1.418238
RECURRENT	1	-0.33993936	0.07348783	1.4048624
STAGE	3	0.29916993	-0.06150948	0.74143341
FREQ_CHEMO	1	0.29378459	-0.06262496	0.74543705
FREQ_CHEMO	0	-0.26376819	0.05597178	1.30182638
IV_PAINMED	DEM	-0.26085980	0.036163	1.29804567
TREATMENT_PLAN	Chemo&Radiation	-0.25534174	0.03324906	1.2909027
TYPE_PROCD	closed	0.25466832	-0.01992872	0.77517356
PREOP_GI_MED	1	0.25194913	-0.06873117	0.77728428
MALIGNANCY	1	0.24061736	-0.05486614	0.78614238
QUARTER	A	0.23306129	-0.05746447	0.79210502
SIZE_REDUCTION	NULL	0.22915110	-0.15356344	0.79520837
TYPE_PROCD	1	-0.22759025	0.03846051	1.25557075
EPIDURAL	1	-0.22715954	0.05119796	1.25503009
INSURANCE	B	0.21168257	-0.05517357	0.80922152
OR_TRANSFUSIONS	1	0.20613024	-0.0550411	0.81372709
TYPE_ABX	Cipro	0.20248206	-0.02044382	0.81670114
EKG	SB	0.19228831	-0.02216336	0.82506896
IV_PAINMED	TORD	-0.19105185	0.01912802	1.21052222
INCISION	KNEE	-0.18882816	0.01878139	1.20783338
INSURANCE	C	0.18859100	-0.02710814	0.82812514
WT_LOSS_TIME	NULL	-0.17535293	0.11368976	1.19166672
WEEKDAY	Sa	0.17096336	-0.02674837	0.84285246



Oracle Data Mining & OBI EE

8.4 Oracle Datamining

LTV Prediction LTV Details **Classification Tree** LTV Probabilities What If Scoring

Classification Tree

Page Information (click to collapse or expand)

Classification Tree
Time run: 12/9/2011 1:03:03 PM

20 Actual Unit Price

		2008	2009	2010	Grand Total
0 - All Individuals	MEDIUM	9.302	9.302	9.382	9.331
1 - M_MARITAL_ST in 'DIVORCED', 'SINGLE'	MEDIUM	9.207	9.329	9.421	9.322
2 - M_CRDT_RATE <= 657.5	LOW	9.225	9.164	9.377	9.261
12 - M_INCOME_LVL in 'LEVEL 5', 'LEVEL 6', 'LEVEL 7', 'LEVEL 8', 'LEVEL 9'	MEDIUM	8.904	9.131	9.670	9.261
13 - M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	LOW	9.345	9.176	9.259	9.261
3 - M_CRDT_RATE > 657.5	MEDIUM	9.193	9.462	9.454	9.370
14 - M_MONTHS_CONTACT <= 12.5	VERY HIGH	8.815	9.418	8.690	8.951
4 - M_MONTHS_CONTACT > 12.5	MEDIUM	9.242	9.468	9.543	9.421
7 - M_MARITAL_ST in 'MARRIED', 'WIDOW'	HIGH	9.397	9.276	9.343	9.341

1- Revenue

		2008	2009	2010	Grand Total
0 - All Individuals	MEDIUM	16,500,000	15,000,000	18,500,000	50,000,000
1 - M_MARITAL_ST in 'DIVORCED', 'SINGLE'	MEDIUM	8,155,247	7,589,505	9,289,014	25,033,766
2 - M_CRDT_RATE <= 657.5	LOW	3,560,875	3,340,550	4,015,646	10,917,071
12 - M_INCOME_LVL in 'LEVEL 5', 'LEVEL 6', 'LEVEL 7', 'LEVEL 8', 'LEVEL 9'	MEDIUM	938,983	889,059	1,189,016	3,017,058
13 - M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	LOW	2,621,892	2,451,491	2,826,630	7,900,013

ODM's predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools

Classification Tree Details
Time run: 12/9/2011 1:03:03 PM

#	M23 Full Rule	Predicted LTV	# of Cust	1- Revenue	Trend
12	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE <= 657.5; AND M_INCOME_LVL in 'LEVEL 5', 'LEVEL 6', 'LEVEL 7', 'LEVEL 8', 'LEVEL 9'	MEDIUM	0		
13	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE <= 657.5; AND M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	LOW	0		
14	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE > 657.5; AND M_MONTHS_CONTACT <= 12.5	VERY HIGH	0		
15	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE > 657.5; AND M_MONTHS_CONTACT > 12.5	MEDIUM	0		
16	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE > 657.5; AND M_MONTHS_CONTACT > 12.5	LOW	0		
17	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_CRDT_RATE > 657.5; AND M_MONTHS_CONTACT > 12.5	MEDIUM	0		
18	M_MARITAL_ST in 'DIVORCED', 'SINGLE' ; AND M_MONTHS_CONTACT > 12.5; AND M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	HIGH	18	48,866	
19	M_MARITAL_ST in 'MARRIED', 'WIDOW' ; AND M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	MEDIUM	0		
20	M_MARITAL_ST in 'MARRIED', 'WIDOW' ; AND M_INCOME_LVL in 'LEVEL 1', 'LEVEL 2', 'LEVEL 3', 'LEVEL 4'	HIGH	0		



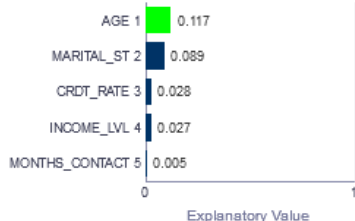
Dynamically Using ODM From Oracle BI



Model Attributes Significance

Time run: 5/15/2014 7:37:48 AM

Most Significant Attributes in the Model



Select Table Details By Credit Rate

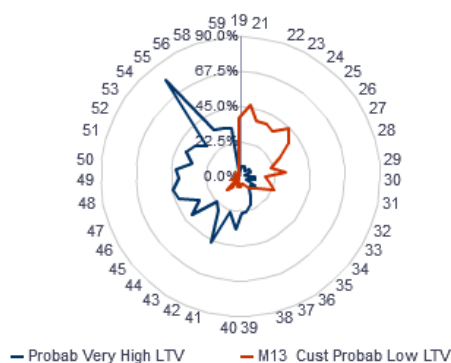
C6 Credit Rate	# of Custs	1- Revenue	Proba Low LTV	Probab Very High LTV
600	63	4,578,456	31.0%	11.0%
615	72	4,378,773	27.3%	8.0%
630	76	5,302,193	30.0%	7.6%
645	189	14,404,249	28.1%	9.7%
650	46	3,663,435	19.0%	12.9%
665	58	3,581,326	11.8%	15.3%
680	76	5,488,196	12.6%	14.3%
695	140	9,315,649	11.1%	22.7%
700	36	2,483,128	14.0%	23.9%
715	27	1,502,589	11.4%	33.1%
730	37	2,547,730	10.1%	28.3%
745	80	5,278,155	7.8%	21.7%
750	10	617,865	7.5%	11.5%
765	14	1,406,422	4.2%	39.6%



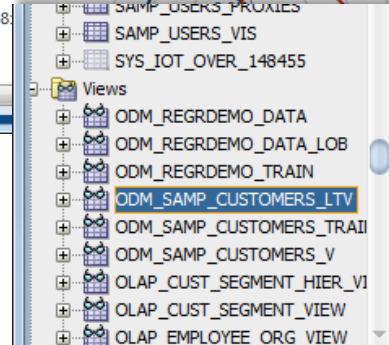
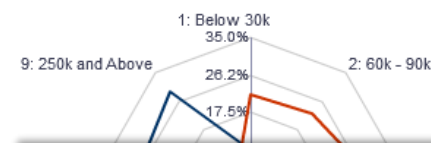
Predicted LTV

Time run: 5/15/2014 7:37:48 AM

LTV Probability by Ages



LTV Probability by Income Level



```
CREATE OR REPLACE FORCE VIEW "BISAMPLE"."ODM_SAMP_CUSTOMERS_LTV" ("CUST_KEY", "M_INCOME_LVL", "M_MARITAL_ST", "M_CRDT_RATE", "M_LTV", "M_AGE", "M_MONTHS_CONTACT", "M_PRED_LTV_NODE", "M_PRED_PROB_VH", "M_PRED_PROB_H", "M_PRED_PROB_M", "M_PRED_PROB_L", "M_PRED_BIN") AS
SELECT C."CUST_KEY",
       C."M_INCOME_LVL",
       C."M_MARITAL_ST",
       C."M_CRDT_RATE",
       C."M_LTV",
       C."M_AGE",
       C."M_MONTHS_CONTACT",
       to_number(extractValue(PREDICTION_DETAILS(ODM_LTV_BIN USING *), 'Details/@node')) M_PRED_LTV_NODE,
       PREDICTION_PROBABILITY(ODM_LTV_BIN, 'VERY HIGH' USING *) M_PRED_PROB_VH,
       PREDICTION_PROBABILITY(ODM_LTV_BIN, 'HIGH' USING *) M_PRED_PROB_H,
       PREDICTION_PROBABILITY(ODM_LTV_BIN, 'MEDIUM' USING *) M_PRED_PROB_M,
       PREDICTION_PROBABILITY(ODM_LTV_BIN, 'LOW' USING *) M_PRED_PROB_L,
       PREDICTION(ODM_LTV_BIN USING *) M_PRED_BIN
FROM ODM_SAMP_CUSTOMERS_V C;
```




Basic Ways to Get Started

- Do a POC project on your own
- Conduct a workshop for key stakeholders to build support
 - One hour to one day
 - Half-day works great
- Conduct ODM and ORE training classes with 1-day workshop
- Use a defined Quick Start program (2 weeks)



ODM Quick Start Overview

- Hardware or Cloud
 - 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)



Quick Start 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



Important Factors in Getting Started

- Lots of internal experts and 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
 - Bring people along on the journey, establish culture
 - Everyone shares a minimum common knowledge base
- Use workshops (JAD style session) for 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



Oracle Data Mining Training (2 days)

- 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 Test Drive

- Free to try Oracle BI, Advanced Analytics Go to www.vlamis.com/td
- Runs off of Oracle Cloud
- Test Drives for:
 - Oracle BI
 - Oracle Advanced Analytics
- Once sign up, you have private instance for one day
- Available now



Thank You!

Clustering Data with Oracle Data Mining

Tim Vlami

tvlamis@vlamis.com

www.vlamis.com

ODTUG

Kscope16



CHICAGO, ILLINOIS • JUNE 26-30

PLEASE FILL OUT YOUR EVALUATIONS