Oracle Advanced Analytics for Fraud and Anomaly Detection

Make Big Data + Analytics Simple

Charlie Berger, MS Engineering, MBA
Sr. Director Product Management, Data Mining and Advanced Analytics
charlie.berger@oracle.com    www.twitter.com/CharlieDataMine
Safe Harbor Statement

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.
Fraud Statistics
By The Numbers

• Overall
  – Conservatively, fraud steals $80 billion a year across all lines of insurance. (Coalition Against Insurance Fraud est.).
  – Fraud comprises about 10 percent of property-casualty insurance losses and loss adjustment expenses each year.

• Fraud costs for insurers
  – Fraud accounts for 5-10 percent of claims costs for U.S. and Canadian insurers. Nearly one-third of insurers (32 percent) say fraud was as high as 20 percent of claims costs;
  – About 35 percent say fraud costs their companies 5-10 percent of claim volume. More than 30 percent say fraud losses cost 10-20 percent of claim volume;
  – Detecting fraud before claims are paid, and upgrading analytics, were mentioned most often as the insurers’ main fraud-fighting priorities;

http://www.insurancefraud.org/statistics.htm#VfreKRFVhBc
Fraud Statistics
By The Numbers

• Medicare & Medicaid
  – Nearly **$80 billion of improper Medicare and Medicaid payments** were made in FY 2014;
  – Anti-fraud efforts **recovered $3.3 billion in taxpayer dollars** in FY 2014; and
  – $7.70 was returned for every anti-fraud dollar invested. This is about $2 higher than the average ROI since 1997. It’s also the third-highest ROI. (U.S. Department of HHS, March 2015)

• Automobile - Bodily injury claims
  – Staged-crash rings **fleece auto insurers out of billions of dollars a year by billing for unneeded treatment of phantom injuries**. Usually these are bogus soft-tissue injuries such as sore backs or whiplash, which are difficult to medically identify and dispute.

• Hotspot states
  – Drivers in Lawrence, MA— the “worst hotbed of fraudulent claims” — have saved more than $68 million; Larger chiropractors in Lawrence have decreased in both clinic counts and billings by up to 90 percent. High-volume physical therapy clinics (billings exceeding $100,000 annually) have been eliminated, and attorney involvement in PIP claims has dropped;
People's Attitudes About Fraud

Consumers

• Nearly one of four Americans say it’s ok to defraud insurers
  – Some 8 percent say it’s “quite acceptable” to bilk insurers, while 16 percent say it’s “somewhat acceptable.”
  – About one in 10 people agree it’s ok to submit claims for items that aren’t lost or damaged, or for personal injuries that didn’t occur.
  – Two of five people are “not very likely” or “not likely at all” to report someone who ripped off an insurer. Accenture Ltd.(2003)

• Nearly one of 10 Americans would commit insurance fraud if they knew they could get away with it.

• Nearly three of 10 Americans (29 percent) wouldn't report insurance scams committed by someone they know. Progressive Insurance (2001)
American Society of Certified Fraud Examiners

20 Ways to Detect Fraud

1. Unusual Behavior
   The perpetrator will often display unusual behavior, that when taken as a whole is a strong indicator of fraud. The fraudster may not ever take a vacation or call in sick in fear of being caught. He or she may not assign out work even when overloaded. Other symptoms may be changes in behavior such as increased drinking, smoking, defensiveness, and unusual irritability and suspiciousness.

2. Complaints
   Frequently tips or complaints will be received which indicate that a fraudulent action is going on. Complaints have been known to be some of the best sources of fraud and should be taken seriously. Although all too often the motives of the complainant may be suspect, the allegations usually have merit that warrant further investigation.

3. Stale Items in Reconciliations
   In bank reconciliations, deposits or checks not included in the reconciliation could be indicative of theft. Missing deposits could mean the perpetrator absconded with them; missing checks could indicate one made out to a bogus payee.

4. Excessive Voids
   Voided sales slips could mean that the sale was rung up, the payment diverted to the use of the perpetrator, and the sales slip subsequently voided to cover the theft.

5. Missing Documents
   Documents which are unable to be located can be a red flag for fraud. Although it is expected that some documents will be misplaced, the auditor should look for explanations as to why the documents are missing, and what steps were taken to locate the requested items. All too often, the auditors will select an alternate item or allow the auditee to select an alternate without determining whether or not problem exists.

6. Excessive Credit Memos
   Similar to excessive voids, this technique can be used to cover the theft of cash. A credit memo to a phony customer is written out, and the cash is taken to make total cash balance.
Data, data everywhere
Growth of Data Exponentially Greater than Growth of Data Analysts!

The Useful Data GAP

12%

Executives who feel they understand the impact data will have on their organizations

Produce

Data

Data Analysis platforms requirements:

- Be extremely powerful and handle large data volumes
- Be easy to learn
- Be highly automated & enable deployment

Use

Data

http://www.delphianalytics.net/more-data-than-analysts-the-real-big-data-problem/
Analytics + Data Warehouse + Hadoop

• Platform Sprawl
  – More Duplicated Data
  – More Data Movement Latency
  – More Security challenges
  – More Duplicated Storage
  – More Duplicated Backups
  – More Duplicated Systems
  – More Space and Power
Vision

• Big Data + Analytic Platform for the Era of Big Data and Cloud
  – Make Big Data + Analytics Model Discovery *Simple*
    • Any data size, on any computer infrastructure
    • Any variety of data (structured, unstructured, transactional, geospatial), in any combination
  – Make Big Data + Analytics Model Deployment *Simple*
    • As a service, as a platform, as an application
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Scalable in-Database + Hadoop data mining algorithms and R integration
- Powerful predictive analytics and deployment platform
- Drag and drop workflow, R and SQL APIs
- Data analysts, data scientists & developers
- Enables enterprise predictive analytics applications
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Major Benefits

- Data remains in Database & Hadoop
  - Model building and scoring occur in-database
  - Use R packages with data-parallel invocations
- Leverage investment in Oracle IT
  - Eliminate data duplication
  - Eliminate separate analytical servers
- Deliver enterprise-wide applications
  - GUI for Predictive Analytics & code gen
  - R interface leverages database as HPC engine

- Eliminate data duplication
- Eliminate separate analytical servers
- Deliver enterprise-wide applications
  - GUI for Predictive Analytics & code gen
  - R interface leverages database as HPC engine
Fiserv
Risk Analytics in Electronic Payments

Objectives
- Prevent $200M in losses every year using data to monitor, understand and anticipate fraud

Solution
- We installed OAA analytics for model development during 2014
- When choosing the tools for fraud management, speed is a critical factor
- OAA provided a fast and flexible solution for model building, visualization and integration with production processes

“When choosing the tools for fraud management, speed is a critical factor. Oracle Advance Analytics provided a fast and flexible solution for model building, visualization and integration with production processes.”

– Miguel Barrera, Director of Risk Analytics, Fiserv Inc.
– Julia Minkowski, Risk Analytics Manager, Fiserv Inc.
Data Miner Survey 2016 by Rexer Analytics

While 6 out of 10 data miners report the data is available for analysis within days of capture, the time to deploy the models takes substantially longer. For 60% of the respondents, the deployment time will range between 3 weeks and 1 year.

Ease of Deployment

Everyone forgets about deployment – but is most important component!
UK National Health Service
Combating Healthcare Fraud

Objectives
- Use new insight to help identify cost savings and meet goals
- Identify and prevent healthcare fraud and benefit eligibility errors to save costs
- Leverage existing data to transform business and productivity

Solution
- Identified up to GBP100 million (US$156 million) potentially saved through benefit fraud and error reduction
- Used anomaly detection to uncover fraudulent activity where some dentists split a single course of treatment into multiple parts and presented claims for multiple treatments
- Analyzed billions of records at one time to measure longer-term patient journeys and to analyze drug prescribing patterns to improve patient care

“Oracle Advanced Analytics’ data mining capabilities and Oracle Exalytics' performance really impressed us. The overall solution is very fast, and our investment very quickly provided value. We can now do so much more with our data, resulting in significant savings for the NHS as a whole”
– Nina Monckton, Head of Information Services, NHS Business Services Authority

Oracle Exadata Database Machine
Oracle Advanced Analytics
Oracle Exalytics In-Memory Machine
Oracle Endeca Information Discovery
Oracle Business Intelligence EE
Oracle’s Advanced Analytics
Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

Users

R programmers

Data & Business Analysts

Business Analysts/Mgrs

Domain End Users

Platform

R Client

SQL Developer/
Oracle Data Miner

OBIEE

Applications

Platform

Hadoop

ORAAH
Parallel, distributed algorithms

Oracle Database Enterprise Edition

Oracle Advanced Analytics - Database Option
SQL Data Mining & Analytic Functions + R Integration for Scalable, Distributed, Parallel in-Database ML Execution

Oracle Cloud
Oracle Advanced Analytics Database Evolution

- **1998**
  - 7 Data Mining “Partners”

- **1999**
  - Oracle acquires Thinking Machine Corp’s dev. team + "Darwin" data mining software

- **2002**
  - Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API

- **2004**
  - Oracle Data Mining 10gR2 SQL - 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI
  - SQL statistical functions introduced

- **2005**
  - Oracle Data Mining 10gR2 SQL

- **2008**
  - ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements

- **2011**
  - Integration with “R” and introduction/addition of Oracle R Enterprise
  - Product renamed “Oracle Advanced Analytics (ODM + ORE)”

- **2014**
  - New algorithms (EM, PCA, SVD)
  - Predictive Queries

---

**ORACLE DATABASE**

- **12c**
  - SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
  - OAA/ORE 1.3 + 1.4

- **11g**
  - SQLDEV/Oracle Data Miner adds NN, Stepwise, scalable R algorithms

- **10g**
  - Integration with “R” and introduction/addition of Oracle R Enterprise

- **9i**
  - Oracle Data Mining

- **8i**
  - Oracle acquires Thinking Machine Corp’s dev. team + "Darwin" data mining software

---

Oracle Advanced Analytics for Hadoop Connector launched with scalable BDA algorithms
You Can Think of Oracle’s Advanced Analytics Like This...

**Traditional SQL**
- “Human-driven” queries
- Domain expertise
- Any “rules” must be defined and managed

**SQL Queries**
- SELECT
- DISTINCT
- AGGREGATE
- WHERE
- AND OR
- GROUP BY
- ORDER BY
- RANK

**Oracle Advanced Analytics - SQL & R**
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine/analyze

**Analytical SQL “Verbs”**
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Oracle Advanced Analytics—On Premise or Cloud

100% Compatibility Enables Easy Coexistence and Migration

CoExistence and Migration

Same Architecture
Same Analytics
Same Standards

Transparently move workloads and analytical methodologies between On-premise and public cloud
Oracle’s Advanced Analytics

In-Database Data Mining Algorithms*—SQL & R & GUI Access

**Classification**
- Decision Tree
- Logistic Regression (GLM)
- Naïve Bayes
- Support Vector Machine (SVM)
- Random Forest

**Regression**
- Multiple Regression (GLM)
- Support Vector Machine (SVM)
- Linear Model
- Generalized Linear Model
- Multi-Layer Neural Networks
- Stepwise Linear Regression

**Clustering**
- Hierarchical k-Means
- Orthogonal Partitioning Clustering
- Expectation-Maximization

**Attribute Importance**
- Minimum Description Length
- Unsupervised pair-wise KL div.

**Anomaly Detection**
- 1 Class Support Vector Machine

**Time Series**
- Single & Double Exp. Smoothing

**Predictive Queries**
- Clustering
- Regression
- Anomaly Detection
- Feature Extraction

**Feature Extraction & Creation**
- Nonnegative Matrix Factorization
- Principal Component Analysis
- Singular Value Decomposition

**Market Basket Analysis**
- Apriori – Association Rules

**Open Source R Algorithms**
- Ability to run any R package via Embedded R mode

* supports partitioned models, text mining
Oracle Advanced Analytics
How Oracle R Enterprise Compute Engines Work

1. R-> SQL Transparency “Push-Down”
   - R language for interaction with the database
   - R-SQL Transparency Framework overloads R functions for scalable in-database execution
   - Function overload for data selection, manipulation and transforms
   - Interactive display of graphical results and flow control as in standard R
   - Submit user-defined R functions for execution at database server under control of Oracle Database

2. In-Database Adv Analytical SQL Functions
   - 15+ Powerful data mining algorithms (regression, clustering, AR, DT, etc.)
   - Run Oracle Data Mining SQL data mining functioning (ORE.odmSVM, ORE.odmDT, etc.)
   - Speak “R” but executes as proprietary in-database SQL functions—machine learning algorithms and statistical functions
   - Leverage database strengths: SQL parallelism, scale to large datasets, security
   - Access big data in Database and Hadoop via SQL, R, and Big Data SQL

3. Embedded R Package Callouts
   - R Engine(s) spawned by Oracle DB for database-managed parallelism
   - ore.groupApply high performance scoring
   - Efficient data transfer to spawned R engines
   - Emulate map-reduce style algorithms and applications
   - Enables production deployment and automated execution of R scripts
Data Mining & Anomaly Detection Concepts
What is Data Mining & Predictive Analytics?

*Automatically* sifting through large amounts of data to create models that find previously hidden patterns, discover valuable new insights and make predictions

- Identify most important factor (*Attribute Importance*)
- Predict customer behavior (*Classification*)
- Predict or estimate a value (*Regression*)
- Find profiles of targeted people or items (*Decision Trees*)
- Segment a population (*Clustering*)
- Find fraudulent or “rare events” (*Anomaly Detection*)
- Determine co-occurring items in a “baskets” (*Associations*)
Data Mining Provides Better Information, Valuable Insights and Predictions

Segment #1
IF CUST_MO > 14 AND INCOME < $90K, THEN Prediction = Lease Churner
Confidence = 100%
Support = 8/39

Segment #3
IF CUST_MO > 7 AND INCOME < $175K, THEN Prediction = Lease 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
Data Mining When Lack Examples
Better Information, Valuable Insights and Predictions

Cell Phone Fraud vs. Loyal Customers

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff
Finding Rare and Unusual Records in Large Datasets

- Finding needles in haystacks.
- Look for what’s different....
Challenge: Finding Anomalies

• Considering multiple attributes
• Taken alone, may seem “normal”
• Taken collectively, a record may appear to be anomalous
• Look for what is “different”
### A Real Fraud Example

**My credit card statement—**Can you see the fraud?**

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Category</th>
<th>Location</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 22</td>
<td>1:14 PM</td>
<td>FOOD</td>
<td>Monaco Café</td>
<td>$127.38</td>
</tr>
<tr>
<td>May 22</td>
<td>7:32 PM</td>
<td>WINE</td>
<td>Wine Bistro</td>
<td>$28.00</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>Gas Station?</td>
<td></td>
</tr>
<tr>
<td>June 14</td>
<td>2:05 PM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
<tr>
<td>June 14</td>
<td>2:06 PM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
<tr>
<td>June 15</td>
<td>11:48 AM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
<tr>
<td>June 15</td>
<td>11:49 AM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
<tr>
<td>May 28</td>
<td>6:31 PM</td>
<td>WINE</td>
<td>Acton Shop</td>
<td>$31.00</td>
</tr>
<tr>
<td>May 29</td>
<td>8:39 PM</td>
<td>FOOD</td>
<td>Crossroads</td>
<td>$128.14</td>
</tr>
<tr>
<td>June 16</td>
<td>11:48 AM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
<tr>
<td>June 16</td>
<td>11:49 AM</td>
<td>MISC</td>
<td>Mobil Mart</td>
<td>$75.00</td>
</tr>
</tbody>
</table>

**Total purchases exceed** time period average

**All same $75 amount?**

Monaco?

Pairs of $75?
“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.)
Start with a Business Problem Statement

Clearly Define Problem

“If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions.”

— Albert Einstein
More Data Variety—Better Predictive Models

• Increasing sources of relevant data can boost model accuracy

Model with “Big Data” and hundreds -- thousands of input variables including:
• Demographic data
• Purchase POS transactional data
• “Unstructured data”, text & comments
• Spatial location data
• Long term vs. recent historical behavior
• Web visits
• Sensor data
• etc.

Population Size

Responders

Copyright © 2015 Oracle and/or its affiliates. All rights reserved.
Multiple Data Sources/Types with Predictive Modeling
Ease of Deployment through SQL Script Generation

SQL Joins and arbitrary SQL transforms & queries – power of SQL

Generates SQL scripts for deployment

Consider:
- Demographics
- Past purchases
- Recent purchases
- Customer comments & tweets

Inline predictive model to augment input data

Unstructured data also mined by algorithms
Tax Noncompliance Audit Selection
Two Example Approaches - There are many possible more!

• Anomaly Detection
  – Build 1-Class Support Vector Machine (SVM) models on “normal or compliant” tax submissions
    • Unsupervised machine learning when few know examples on which to train e.g. < 2%
  – Build Decision Tree models for classification of Noncompliant tax submissions (yes/no) based on historical 2011 data
    • Supervised machine learning approach when many known examples of target classes are available oh which to train
SQL Developer/Oracle Data Miner GUI

Anomaly Detection—Simple Conceptual Workflow

Train on “normal” records
Apply model and sort on likelihood to be “different”
<table>
<thead>
<tr>
<th>TAX_PERSON_ID</th>
<th>COMPLAINEE</th>
<th>SALARY</th>
<th>MARRITAL_STATUS</th>
<th>MEDICAL_DEDUCTIONS</th>
<th>STATE</th>
<th>PALM_LATE_PAYMENTS</th>
<th>SEX</th>
<th>ADJUSTED_TAX_INE</th>
<th>ADJ_TA</th>
<th>ADJUSTED_DEDUCTIONS</th>
<th>IRS_RESIDENCE</th>
<th>HOUSE_OWNERSHIP</th>
<th>MONTHLY_CHECK_WRITTEN</th>
<th>IN_MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>31,130</td>
<td>6,419</td>
<td>DIVORCED</td>
<td>5,000</td>
<td>NY</td>
<td>0</td>
<td>F</td>
<td>17,800</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3,160</td>
<td>1,000</td>
<td>MARRIED</td>
<td>4,000</td>
<td>CA</td>
<td>0</td>
<td>M</td>
<td>23,413</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7,612</td>
<td>2,616</td>
<td>MARRIED</td>
<td>5,000</td>
<td>NY</td>
<td>0</td>
<td>F</td>
<td>27,853</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>58,395</td>
<td>7,111</td>
<td>MARRIED</td>
<td>4,000</td>
<td>CA</td>
<td>0</td>
<td>M</td>
<td>21,840</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>58,475</td>
<td>1,500</td>
<td>MARRIED</td>
<td>4,000</td>
<td>NY</td>
<td>0</td>
<td>M</td>
<td>26,210</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>75,096</td>
<td>1,000</td>
<td>MARRIED</td>
<td>4,000</td>
<td>WA</td>
<td>0</td>
<td>M</td>
<td>24,772</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>62,924</td>
<td>3,000</td>
<td>MARRIED</td>
<td>4,000</td>
<td>MO</td>
<td>0</td>
<td>F</td>
<td>24,131</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>7,350</td>
<td>1,000</td>
<td>MARRIED</td>
<td>0</td>
<td>CA</td>
<td>0</td>
<td>M</td>
<td>86,400</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>66,407</td>
<td>1,000</td>
<td>MARRIED</td>
<td>4,000</td>
<td>NY</td>
<td>0</td>
<td>M</td>
<td>2,165</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>66,407</td>
<td>1,000</td>
<td>MARRIED</td>
<td>4,000</td>
<td>NY</td>
<td>0</td>
<td>M</td>
<td>2,165</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Adjusted Tax Inc by Marital Status

#### Statistics: 10 Columns from 2,048 Rows (Sample)

<table>
<thead>
<tr>
<th>Name</th>
<th>Data Type</th>
<th>Percent NULLS</th>
<th>Distinct Values</th>
<th>Distinct Percent</th>
<th>Mode</th>
<th>Average</th>
<th>Median</th>
<th>Min Value</th>
<th>Max Value</th>
<th>Standard Deviation</th>
<th>Variance</th>
</tr>
</thead>
</table>
| ADDL_TAX_CREDIT     | NUMBER    | 0             | 201             | 30.02%           |      | 1,248.956 | 25     | 0         | 0         | 508,911           | 408,847,592,13
| ADJUSTED_DEDUCTIONS| NUMBER    | 0             | 626             | 30.72%           |      | 1,110.276 | 25     | 25        | 24,720    | 2,227,435         | 20,903,666,428
| ADJUSTED_TAX_INC   | NUMBER    | 0             | 2,882           | 32.39%           |      | 22,276.7042 | 25      | 25        | 24,720    | 2,227,435         | 20,903,666,428
| AGE                 | NUMBER    | 0             | 65              | 2.47%            |      | 36.9985 | 36     | 0         | 94        | 21.2732           | 303.7231 |
| AT_LSN              | VARCHAR2  | 4             | 0.9998          | 100              | HIGH | 0.9998 | 1      | 1         | 1         | 0.2391           | 0.0572   |
| CAR_OWNERSHIP       | NUMBER    | 2             | 0.0988          | 100              | No   | 0.0988 | 1      | 1         | 1         | 0.2391           | 0.0572   |
| COMPLAINT           | VARCHAR2  | 2              | 0.0988          | 100              | No   | 0.0988 | 1      | 1         | 1         | 0.2391           | 0.0572   |
| CREDIT_CARD_LIMITS  | NUMBER    | 23             | 1.1472          | 100              | <Other> | 1,234.243 | 1,000 | 0         | 5,000     | 806.8432          | 650,904,0279
| FIRST               | VARCHAR2  | 1 191          | 69.3786         | 100              | <Other> | 1,191   | 1,000 | 0         | 3,000     | 986.3786          | 967,378,6279
| HAS_CHILDREN        | NUMBER    | 2              | 0.0988          | 100              | No   | 0.0988 | 1      | 1         | 1         | 0.2391           | 0.0572   |
| HOUSE_OWNERSHIP     | NUMBER    | 3              | 1.4066          | 100              | No   | 1.4066 | 1      | 1         | 3         | 0.5178           | 0.2681   |

#### Adjusted Tax Inc by Marital Status

- **Married**: [Bar Graph]
- **Widowed**: [Bar Graph]
- **Divorced**: [Bar Graph]
- **Single**: [Bar Graph]
## Edit Filter Columns Node

- **Show Attribute Importance**: Checked
- **Show Data Quality**: Checked

### Columns

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Output</th>
<th>Rank</th>
<th>Importance</th>
<th>% Null</th>
<th>% Unique</th>
<th>% Constant</th>
<th>Hints</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEMIZED_DEDUCTIONS</td>
<td>NUMBER</td>
<td>1</td>
<td>1</td>
<td>0.2085</td>
<td>0</td>
<td>21.6963</td>
<td>36.0451</td>
<td></td>
</tr>
<tr>
<td>NUM_LATE_PAYMENTS</td>
<td>NUMBER</td>
<td>2</td>
<td>1</td>
<td>0.1325</td>
<td>0</td>
<td>0.4408</td>
<td>20.099</td>
<td></td>
</tr>
<tr>
<td>PAST_RETURNS</td>
<td>NUMBER</td>
<td>3</td>
<td>1</td>
<td>0.1103</td>
<td>0</td>
<td>0.4921</td>
<td>33.2483</td>
<td></td>
</tr>
<tr>
<td>T_AMOUNT_AUTOM_PAYMENTS</td>
<td>NUMBER</td>
<td>4</td>
<td>1</td>
<td>0.0939</td>
<td>0</td>
<td>61.6864</td>
<td>19.9211</td>
<td></td>
</tr>
<tr>
<td>MONEY_MONLY_OVERDREW</td>
<td>NUMBER</td>
<td>5</td>
<td>1</td>
<td>0.0759</td>
<td>0</td>
<td>1.5286</td>
<td>60.365</td>
<td></td>
</tr>
<tr>
<td>MONTHLY_CHECKS_WRITTEN</td>
<td>NUMBER</td>
<td>6</td>
<td>1</td>
<td>0.0717</td>
<td>0</td>
<td>0.9369</td>
<td>18.1953</td>
<td></td>
</tr>
<tr>
<td>N_OP_DEPENDENTS</td>
<td>NUMBER</td>
<td>7</td>
<td>1</td>
<td>0.0280</td>
<td>0</td>
<td>3.4852</td>
<td>35.4043</td>
<td></td>
</tr>
<tr>
<td>YRS_RESIDENCE</td>
<td>NUMBER</td>
<td>8</td>
<td>1</td>
<td>0.0191</td>
<td>0</td>
<td>0.2455</td>
<td>32.000</td>
<td></td>
</tr>
<tr>
<td>MORTGAGE_AMOUNT</td>
<td>NUMBER</td>
<td>9</td>
<td>1</td>
<td>0.0114</td>
<td>0</td>
<td>21.3151</td>
<td>23.3152</td>
<td></td>
</tr>
<tr>
<td>ADJUSTED_DEDUCTIONS</td>
<td>NUMBER</td>
<td>10</td>
<td>1</td>
<td>0.0113</td>
<td>0</td>
<td>30.8679</td>
<td>61.9822</td>
<td></td>
</tr>
<tr>
<td>ADDL_TAX_CREDIT</td>
<td>NUMBER</td>
<td>11</td>
<td>1</td>
<td>0.0094</td>
<td>0</td>
<td>10.0999</td>
<td>89.9002</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>NUMBER</td>
<td>12</td>
<td>1</td>
<td>0.0066</td>
<td>3.3531</td>
<td>3.2051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td>VARCHAR2</td>
<td>13</td>
<td>1</td>
<td>0.0015</td>
<td>0.0935</td>
<td>66.9312</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARITAL_STATUS</td>
<td>VARCHAR2</td>
<td>14</td>
<td>1</td>
<td>0.0083</td>
<td>0.2485</td>
<td>66.6667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICAL_DEDUCTIONS</td>
<td>NUMBER</td>
<td>15</td>
<td>1</td>
<td>0.0063</td>
<td>0</td>
<td>20.6114</td>
<td>19.6252</td>
<td></td>
</tr>
<tr>
<td>N_MORTGAGE</td>
<td>NUMBER</td>
<td>16</td>
<td>1</td>
<td>0.0057</td>
<td>0</td>
<td>1.4234</td>
<td>20.9999</td>
<td></td>
</tr>
</tbody>
</table>

**Help**
<table>
<thead>
<tr>
<th>TAX_PERSON_ID</th>
<th>ANON_SIM_42_INDEX</th>
<th>PAST_RETURNS</th>
<th>ADJUSTED_TAX_INC</th>
<th>MEDICAL_DEDUCTIONS</th>
<th>ANON_SIM_2_41_INDEX</th>
<th>ADJUSTED_DEDUCTIONS</th>
<th>CAR_OWNERSHIP</th>
<th>MARITAL_STATUS</th>
<th>LAST_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.56836587153656</td>
<td>25,111</td>
<td>8,000</td>
<td>2,273</td>
<td>0</td>
<td>25</td>
<td>1 WIDOWED</td>
<td>KE...</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.56836587153656</td>
<td>2</td>
<td>2,273</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1 WIDOWED</td>
<td>KE...</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.56836587153656</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1 WIDOWED</td>
<td>KE...</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.56836587153656</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>1 WIDOWED</td>
<td>KE...</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.56836587153656</td>
<td>1</td>
<td>42,362</td>
<td>5,000</td>
<td>0</td>
<td>25</td>
<td>1 WIDOWED</td>
<td>KE...</td>
<td></td>
</tr>
</tbody>
</table>

View Value

```xml
<Details algorithm="Support Vector Machines" class="G">
  <Attribute name="MONTHLY_OVERDRAFT" actualValue="-123" weight=".077" rank="1"/>
  <Attribute name="MARITAL_STATUS" actualValue="WIDOWED" weight=".04" rank="2"/>
  <Attribute name="STATE" actualValue="NC" weight=".17" rank="3"/>
  <Attribute name="YRS_RESPIDENCE" actualValue="4" weight=".014" rank="4"/>
  <Attribute name="SEX" actualValue="F" weight=".009" rank="5"/>
</Details>
```
Fraud Prediction Demo

Automated In-DB Analytical Methodology

drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values ('ALGO_NAME','ALGO_SUPPORT_VECTOR_MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;

begin
begin
  dbms_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 ('2to4', 'morethan4')) 
where rnk <= 5
order by percent_fraud desc;

Automated Monthly “Application”!
Just add:
Create View CLAIMS2_30
  As
Select * from CLAIMS2
Where mydate > SYSDATE – 30

Time measure: set timing on;
Financial Sector/Accounting/Expenses

Anomaly Detection

Simple Fraud Detection Methodology—1-Class SVM

More Sophisticated Fraud Detection Methodology—Clustering + 1-Class SVM
Multiple Approaches To Detect Potential Fraud

1. **Anomaly Detection (1-Class SVM)**
   - Add feedback loop to purify the input training data over time and improve model performance

2. **Classification**
   - IF you have a lot of examples (25% or more) of fraud on which to train/learn

3. **Clustering**
   - Find records that don’t high very high probability to fit any particular cluster and/or lie in the outlier/edges of the clusters

4. **Hybrid of #3 and then #1**
   - Pre-cluster the records to create “similar” segments and then apply anomaly detection models for each cluster

5. **Panel of Experts**
   - i.e. 3 out of 5 models predict possibly anomalous above 40% or any 1 out of N models considers this record unusual
Turkcell
Combating Communications Fraud

Objectives
- Prepaid card fraud—millions of dollars/year
- Extremely fast sifting through huge data volumes; with fraud, time is money

Solution
- Monitor 10 billion daily call-data records
- Leveraged SQL for the preparation—1 PB
- Due to the slow process of moving data, Turkcell IT builds and deploys models in-DB
- Oracle Advanced Analytics on Exadata for extreme speed. Analysts can detect fraud patterns almost immediately

"Turkcell manages 100 terabytes of compressed data—or one petabyte of uncompressed raw data—on Oracle Exadata. With Oracle Data Mining, a component of the Oracle Advanced Analytics Option, we can analyze large volumes of customer data and call-data records easier and faster than with any other tool and rapidly detect and combat fraudulent phone use."

– Hasan Tonguç Yılmaz, Manager, Turkcell İletişim Hizmetleri A.Ş.
Big Data SQL
Push down SQL predicts to storage layers
Introducing **Oracle Big Data SQL**

**Massively Parallel SQL Query across Oracle, Hadoop and NoSQL**

- Offload Query to Data Nodes
- Offload Query to Exadata Storage Servers

Small data subset quickly returned

Hadoop & NoSQL  
Oracle Database 12c
Manage and **Analyze** All Data—SQL & Oracle Big Data SQL

Oracle’s Advanced Analytics

**Structured and Unstructured Data Reservoir**
- JSON data
- HDFS / Hive
- NoSQL
- Spatial and Graph data
- Image and Video data
- Social Media

**Oracle Big Data Appliance**

**Oracle Database 12c**

**Store business-critical data in Oracle**
- Customer data
- Transactional data
- Unstructured documents, comments
- Spatial and Graph data
- Image and Video data
- Social Media

**Data analyzed via SQL / R / GUI**
- R Clients
- SQL Clients
- Oracle Data Miner
Getting started
OAA Links and Resources

- **Oracle Advanced Analytics Overview:**
  - OAA presentation — [Big Data Analytics in Oracle Database 12c With Oracle Advanced Analytics & Big Data SQL](#)
  - [Big Data Analytics with Oracle Advanced Analytics: Making Big Data and Analytics Simple](#)
  - [Oracle Internal OAA Product Management Wiki and Workspace](#)

- **YouTube** recorded OAA Presentations and Demos:
  - [Oracle Advanced Analytics and Data Mining at the YouTube Movies](#)
    (6 + OAA “live” Demos on ODM’r 4.0 New Features, Retail, Fraud, Loyalty, Overview, etc.)

- **Getting Started:**
  - Link to [Getting Started w/ ODM blog entry](#)
  - Link to [New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course](#)
  - Link to [OAA/Oracle Data Mining 4.0 Oracle by Examples (free) Tutorials](#)
  - Take a [Free Test Drive of Oracle Advanced Analytics (Oracle Data Miner GUI)](#) on the Amazon Cloud
  - Link to [OAA/Oracle R Enterprise (free) Tutorial Series](#)

- **Additional Resources:**
  - [Oracle Advanced Analytics Option on OTN page](#)
  - [OAA/Oracle Data Mining on OTN page, ODM Documentation & ODM Blog](#)
  - [OAA/Oracle R Enterprise page on OTN page, ORE Documentation & ORE Blog](#)
  - [Oracle SQL based Basic Statistical functions](#)
  - [BIWA Summit’16, Jan 26-28, 2016 – Oracle Big Data & Analytics User Conference @ Oracle HQ Conference Center](#)
BIWA SUMMIT 2016

The Oracle Big Data + Analytics User Conference
January 26-28, 2016
Including Oracle Spatial Summit

• Hands-on-Labs
• Customer stories, told by the customers
• Educational sessions by Practitioners and Direct from Developers
• Oracle Keynote presentations
• Presentations covering: Advanced Analytics, Big Data, Business Intelligence, Cloud, Data Warehousing and Integration, Spatial and Graph, SQL
• Networking with product management and development professionals

Publicity

• Oracle Business Analytics Newsletter
• DB Insider Dec 2014
• Oracle Magazine
• Latest BIWA SIG Blog Entry
• Jeff Shauer Blog Entry
• Daily BIWA Newsletter
• Email to BIWA members
• Real Time BI Webcast
• Oracle Events Calendar
• Oracle ACE Newsletter
• DB Insider Jan 2015 with Spatial Summit
• Lots of other emails