Oracle Advanced Analytics for Anomaly Detection and Fraud

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

Information Services and the Data Lab

Nina Monckton – Head of Information Services
Abigail Haigh – Senior Data Scientist

Business Services Authority
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.
About Consumer Fraud

- **Conservatively**, fraud steals $80 billion a year across all lines of insurance. (Coalition Against Insurance Fraud est.)
- **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;
- **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.”
- **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))
- **Detecting fraud before claims are paid, and upgrading analytics** were mentioned most often as the insurers’ main fraud-fighting priorities;
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 the funds; 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.
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
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
Oracle Advanced Analytics Database Option
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Scalable in-database 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 Advanced Analytics Database Architecture
Multi-lingual Component of Oracle Database—SQL, SQL Dev/ODMr GUI, R

Users
Data & Business Analysts
R programmers
Business Analysts/Mgrs
Domain End Users

Platform
SQL Developer/Oracle Data Miner
R Client
OBIEE
Applications

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 Advanced Analytics Database Evolution

1998
- Oracle acquires Thinking Machine Corp’s dev. team + “Darwin” data mining software

1999
- 7 Data Mining “Partners”

2002
- Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API
- Oracle Data Mining 10g & 10gR2 introduces SQL dm functions, 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI

2004
- ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
- SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
- OAA/ORE 1.3 + 1.4
- Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API

2005
- Integration with “R” and introduction/addition of Oracle R Enterprise
- Product renamed “Oracle Advanced Analytics (ODM + ORE)"

2008
- Oracle Data Mining 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements

2011
- New algorithms (EM, PCA, SVD)
- Predictive Queries
- SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
- OAA/ORE 1.3 + 1.4

2014
- New algorithms (EM, PCA, SVD)
- Predictive Queries
- SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
- OAA/ORE 1.3 + 1.4
- Oracle Advanced Analytics for Hadoop Connector launched with scalable BDA algorithms
- Integration with “R” and introduction/addition of Oracle R Enterprise
- Product renamed “Oracle Advanced Analytics (ODM + ORE)"
You Can Think of Oracle 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
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
Oracle Advanced Analytics Database Option
Fastest way to deliver enterprise-wide predictive analytics

**Key Features**

**Data remains in the Database**
- Scalable, parallel Data Mining algorithms in SQL kernel
- Fast parallelized native SQL data mining functions, SQL data preparation and efficient execution of R open-source packages
- High-performance parallel scoring of SQL data mining functions and R open-source models

---

<table>
<thead>
<tr>
<th>Traditional Analytics</th>
<th>Oracle Advanced Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Import</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td></td>
</tr>
<tr>
<td>Model “Scoring”</td>
<td></td>
</tr>
<tr>
<td>Data Prep. &amp;</td>
<td></td>
</tr>
<tr>
<td>Transformation</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td></td>
</tr>
<tr>
<td>Model Building</td>
<td></td>
</tr>
<tr>
<td>Data Prep &amp;</td>
<td></td>
</tr>
<tr>
<td>Transformation</td>
<td></td>
</tr>
<tr>
<td>Data Extraction</td>
<td></td>
</tr>
</tbody>
</table>

---

Data Import

Data Mining

Model “Scoring”

Data Prep. & Transformation

Data Mining Model Building

Data Prep & Transformation

Data Extraction

Hours, Days or Weeks

Secs, Mins or Hours

Model “Scoring”
Embedded Data Prep
Model Building
Data Preparation

Savings
Oracle Advanced Analytics Database Option
Fastest way to deliver enterprise-wide predictive analytics

Key Features

Lowest Total Cost of Ownership
- Eliminate data duplication
- Eliminate separate analytical servers
- Leverage investment in Oracle IT

Fastest way to deliver enterprise-wide predictive analytics
- Integrated GUI for Predictive Analytics
- Database scoring engine

Cost Savings

- Eliminate data duplication
- Eliminate separate analytical servers
- Leverage investment in Oracle IT

Integrated GUI for Predictive Analytics
- Database scoring engine

Oracle Advanced Analytics Database Option
- Fastest way to deliver enterprise-wide predictive analytics

Data Import
Data Mining Model “Scoring”
Data Prep. & Transformation
Data Mining Model Building
Data Prep & Transformation
Data Extraction

Traditional Analytics
- Hours, Days or Weeks

Oracle Advanced Analytics
- Secs, Mins or Hours
Why In-DBMS Analytics Deserves a Fresh Look

13 May 2015  G00273535
Analyst(s): Lakshmi Randall | Alexander Linden

Summary
Given the phenomenal growth in multistructured data and the facility with which complex analytics now can be implemented, information and analytics leaders must exploit in-DBMS analytics (aka in-database analytics) to reduce data movement, and push data-intensive processing closer to the data.

In-DBMS Analytics (aka In-database analytics) demands your scrutiny!

by Lakshmi Randall | May 13, 2015 | Submit a Comment

Is In-DBMS analytics a viable choice for democratizing your large-scale business analytics?

Please read the Gartner research “Why In-DBMS Analytics Deserves a Fresh Look”, which will provide guidance on the following:

- Reducing data movement, and pushing data-intensive processing closer to the data.
- Exploring the range of features that DBMSs and Analytics Platforms can offer to facilitate the integration of analytics in the DBMS platform.
- Investigating In-DBMS Analytics as a suitable approach for “fast analytics” in which analysis can be continuous or on-demand, and might require combining historical data with recent (or incoming) data.

In-DBMS analytics deserves a fresh look! Its ability to effectively support rapid analysis of large-scale data coupled with mature vendor offerings provides the opportunity to exploit advanced analytics in your organization.
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

Copyright © 2015 Oracle and/or its affiliates. All rights reserved.
Predictive Analytics & Data Mining
Finding Rare and Unusual Records in Large Datasets

• Finding needles in haystacks.
• Look for what’s different....

https://masonresearch.gmu.edu/2013/03/researcher-unlocks-the-big-potential-of-big-data/
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>Description</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>

All same $75 amount?

Monaco?

Pairs of $75?
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
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”
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.

Naïve Guess or Random
Model with 20 variables
Model with 75 variables
Model with 250 variables
Predicting Behavior
Identify “Likely Behavior” and their Profiles

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

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

Unstructured data also mined by algorithms

Transactional POS data

Generates SQL scripts for deployment

Inline predictive model to augment input data

POS Sales data

Aggregated POS data

Demographics and comments

Customer sentiment data

Credit score model

Predicted credit rating

Join

360 degree view of customers

Filter Columns

Multiple Predictive models

Most Likely customers

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

<table>
<thead>
<tr>
<th>Timeframe</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 months</td>
<td>Run &amp; deploy Logistic Regression (using SAS)</td>
</tr>
<tr>
<td>1 month</td>
<td>Estimate and deploy Trees and GLM</td>
</tr>
<tr>
<td>1 week</td>
<td>Estimate, 1 week to install rules</td>
</tr>
<tr>
<td>1 day</td>
<td>Estimate and deploy Trees + GLM models (using OAA)</td>
</tr>
<tr>
<td></td>
<td>Install rules in online application</td>
</tr>
</tbody>
</table>
Data Miner Survey 2015 by Rexer Analytics
While 6 out 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.

Time to Data Analysis

**Time to Deployment**

Everyone forgets about deployment – but is most important component!
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>Name</th>
<th>Statistics</th>
<th>Data Type</th>
<th>Percent NULL</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>
<tbody>
<tr>
<td>ADJUSTED_TAX_CREDIT</td>
<td>Histogram</td>
<td>NUMBER</td>
<td>0</td>
<td>261</td>
<td>93.0249</td>
<td></td>
<td>1,348.9596</td>
<td>0</td>
<td>0</td>
<td>008,951</td>
<td>21,602,8880</td>
<td>406,847,582,913</td>
</tr>
<tr>
<td>ADJUSTED_DEDUCTIONS</td>
<td>Histogram</td>
<td>NUMBER</td>
<td>0</td>
<td>626</td>
<td>20.7222</td>
<td></td>
<td>1,110,2786</td>
<td>0</td>
<td>0</td>
<td>24,720</td>
<td>2,287,6356</td>
<td>20,003,666,048</td>
</tr>
<tr>
<td>ADJUSTED_TAX_INC</td>
<td>NUMBER</td>
<td>0</td>
<td>2,882</td>
<td>91.3691</td>
<td>22.276,7042</td>
<td>22.004</td>
<td>29,923</td>
<td>47,002</td>
<td>19,500</td>
<td>47,462,812,410</td>
<td>54,8802</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>NUMBER</td>
<td>0</td>
<td>65</td>
<td>3.2419</td>
<td></td>
<td></td>
<td>36.9985</td>
<td>36</td>
<td>0</td>
<td>84</td>
<td>14,2732</td>
<td>203.7231</td>
</tr>
<tr>
<td>ATT_EIN</td>
<td>VARCATEGOR</td>
<td>0</td>
<td>4</td>
<td>0.1995</td>
<td></td>
<td></td>
<td>0.0002</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.2391</td>
<td>0.0572</td>
</tr>
<tr>
<td>CAR_OWNERSHIP</td>
<td>NUMBER</td>
<td>0</td>
<td>2</td>
<td>0.0998</td>
<td></td>
<td></td>
<td>0.0998</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.2391</td>
<td>0.0572</td>
</tr>
<tr>
<td>CLAIM</td>
<td>NUMBER</td>
<td>0</td>
<td>23</td>
<td>1.1472</td>
<td></td>
<td></td>
<td>1,234.243</td>
<td>1,000</td>
<td>500</td>
<td>5,000</td>
<td>806,8432</td>
<td>450,906,0271</td>
</tr>
<tr>
<td>FIRST</td>
<td>VARCATEGOR</td>
<td>0</td>
<td>1,391</td>
<td>69.3798</td>
<td></td>
<td></td>
<td>&lt;Other &gt;</td>
<td></td>
<td></td>
<td></td>
<td>54,8802</td>
<td></td>
</tr>
<tr>
<td>HAS_CHILDREN</td>
<td>NUMBER</td>
<td>0</td>
<td>2</td>
<td>0.0998</td>
<td></td>
<td></td>
<td>0.4663</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.5001</td>
<td>0.2961</td>
</tr>
<tr>
<td>HOUSE_OWNERSHIP</td>
<td>NUMBER</td>
<td>0</td>
<td>3</td>
<td>0.1466</td>
<td></td>
<td></td>
<td>0.8065</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0.5178</td>
<td>0.2681</td>
</tr>
</tbody>
</table>

**Adjusted_Tax_Income by Marital_Status**

![Bar chart](chart.png)
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
  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;

<table>
<thead>
<tr>
<th>POLICYNUMBER</th>
<th>PERCENT_FRAUD</th>
<th>RANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>6532</td>
<td>64.78</td>
<td>1</td>
</tr>
<tr>
<td>2749</td>
<td>64.17</td>
<td>2</td>
</tr>
<tr>
<td>3440</td>
<td>63.22</td>
<td>3</td>
</tr>
<tr>
<td>654</td>
<td>63.1</td>
<td>4</td>
</tr>
<tr>
<td>12650</td>
<td>62.36</td>
<td>5</td>
</tr>
</tbody>
</table>

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

- Explore Data
- CLAIMS
- Anomaly Build
- Most suspicious claims
- Clustering
- Cluster Assignments
- Infrequent HI Amount Claims
- Anomalous claims
- Unusual claims
- Anomalous claims 2
- Repeat claimants
- More Sophisticated Fraud Detection Methodology—Clustering + 1-Class SVM
- Anomalous claims 3

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
“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.)
**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

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

SQL

Oracle Big Data Appliance

Oracle Database 12c

Store JSON data unconverted in Hadoop

Store business-critical data in Oracle

Data analyzed via SQL or R

Store business-critical data in Oracle
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 white paper on OTN
  – 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 on OTN
  – 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 on OTN

• 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 on OTN
  – 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
Information Services and the Data Lab

Nina Monckton – Head of Information Services
Abigail Haigh – Senior Data Scientist
The NHS Business Services Authority is a Special Health Authority and an Arms Length Body of the Department of Health which provides a range of critical central services to NHS organisations, NHS contractors, patients and the public.
2.6 million
prescription prepayment
certificates distributed in
2014/15

3.5 million
EHIC applications
processed each year

We assess and award
bursaries to around
91,000
students (80,000 NHS Bursary
students and 11,000 Social Work
Bursary students)

44,305,178
FP17 dental claim forms
processed in 2014/15

1,441,735
Total Reward Statements
made available to active
NHS employees this year

£19 million
reduction in our gross
annual operating costs
compared to 2005/06
(In real terms)

We deal with
85
Higher Education
Institutions

We reimburse pharmacy
contractors for drugs
dispensed equating to
£9.4 billion
in 2014

1,055,729,706
prescription Items
processed in 2014/15

18,877
satisfaction survey
responses from
customers and
stakeholders
(July 14 - February 15)

30%
the reduction in unit cost for
administering the NHS Tax
Credit Exemption Certificate

We deal with
11,500
Chemists
who dispense

130 million

Our contact
centre handled
3,096,617
calls in 2014/15

There are currently
27 million
EHIC cards in circulation

The engagement
level of our staff
in 2014/15 was
81%

Over
86,000
monthly hits
on our ‘Ask Us’ online
knowledge base

43k
followers on social media sites

We have a proven track record of successfully transferring work from stakeholders across the BSA, having transferred in 10 work streams since 2006

Patient Services process

7.4 million
applications per year. This includes Low Income Scheme, Maternity Exemption, Prescription Pre-payment Certificate, Medical Exemption, Low Income Scheme and Tax Credit applications.

£32 billion
the amount of money that flows through our books
every year

Over
9 million
cards and certificates were
issued to members of the public.
All enable free or reduced
cost health charges.

Total amount of savings our
Pacific Programme has
realised so far for the NHS
and its patients
Data Analytics Learning Laboratory (DALL)

- The Data Analytics Learning Laboratory, or DALL, was established to take forward the NHSBSA’s ability to drive insight through analytics.

One of our key strategic goals for 2015/16 is……
  - ‘We will derive insight from data to drive change’

Our longer term vision is……
  - Over time, we intend to become an organisation known for its abilities to create insight that delivers improvements that matter to our service users and to patients across the NHS.

The DALL will……
  - Place analytics at the heart of our business
  - Appreciate the value of data as a product
DALL operations

- The DALL operating model must be flexible to respond to the needs of our customers
- Anyone from the NHSBSA can contact us for help
- All requests are assessed and logged within a tracker
- All work is monitored for progress
- Each project has a business sponsor
- We use an industry standard methodology (CRISP-DM) for managing our data mining projects
Examples of DALL work

<table>
<thead>
<tr>
<th>Area</th>
<th>Initiative</th>
<th>Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraud, Error &amp; Debt</td>
<td>• Pharmacy Contractor Analysis&lt;br&gt;• Dental Services</td>
<td>• Big data&lt;br&gt;• Enterprise R&lt;br&gt;• Anomaly detection</td>
</tr>
<tr>
<td>Patient care insight</td>
<td>• Care Homes&lt;br&gt;• Polypharmacy&lt;br&gt;• Cancer Registry</td>
<td>• Pseudonymisation&lt;br&gt;• Survival Analysis&lt;br&gt;• R code</td>
</tr>
<tr>
<td>Forecasting</td>
<td>• EPS Uptake</td>
<td>• Regression analysis&lt;br&gt;• R code</td>
</tr>
</tbody>
</table>

- DALL: Data Analytics for Learning
- EPS: Electronic Patient Services
Pharmacy contractor analysis (1)

Why?
• In 2013 NHS Protect produced an anti-crime threat assessment paper which estimated that £47.6m per annum is fraudulently claimed by pharmaceutical contractors
• £5.3m claimed in Out of Pocket Expenses (OoPE) and £89m for conducting Medicine Use Reviews (MURs) during 2014/15. Can we use our data to find who might be claiming payments inappropriately?

What?
• We have worked with NHS England to develop metrics to help quickly spot pharmacy contractors who might be claiming inappropriately. These metrics have been guided by DALL analysis.
Pharmacy contractor analysis - OoPE (2)

- Nutrition and Blood – more than half of OoPE claims relate to Vitamins, valued at £1.6m in FY 2014/15.

- Pro D3 (vitamin D) is the most common claim and accounts for more than half of the number and value of OoPE claims for vitamins and accounts for 18% of total claims of OoPE.
Pharmacy contractor analysis – MUR Outlier Activity (3)

The chart shows the monthly variation in the number of MURs conducted for a number of contractors in financial year 2014/15. These contractors were identified as anomalous.

Data Source: Pharmacy Payment Database

Analysis reveals that some contractors conduct zero (or a very small number of) MURs in eleven months of the year, and then conduct all 400 in the remaining month. Although no rules are broken, the quality of such MURs could be questioned.
Pharmacy contractor analysis (4)

Next steps – moving into provider management

- Our outlier analysis has been incorporated into a toolkit which was circulated to NHS England local offices on 21 August 2015. Focusing on payments for out of pocket expenses, medicine use reviews and new medicines service.

- We will seek feedback from NHS England on the toolkit’s success in 6 months

- £110 million pa spend on out of pocket expenses and reviews on meds use, new meds and appliances. Even a 1% reduction in spend through the provision of better info and targeted support is a saving of £1.1m pa.
Dental – Provision of Inlays (1)

Why?
• A risk assessment in 2010 highlighted considerable variation in inlay provision, with a minority of dental performers providing the bulk of activity. London was a regional outlier, and provision was heavily skewed towards charge-exempt adults

What?
• To investigate the risk associated with inlays in dentistry.
Dental – Provision of Inlays (2)

- National rate has decreased by a fifth from 10 to 8 (inclusion of inlay treatment per 100 Band 3 FP17s) over the six years from 2009/10 to 2014/15.

- Inlay provision in London has fallen by 27%, from above 18 per 100 Band 3 courses of treatment (FP17s) in 2009-10 to below 14 in 2014-15.

- Reduction across London means that £6 million available for other treatment in 2014/15 compared with 2009/10.

- Further £8 million a year would be available for other treatment if inlay rates in London matched those outside the capital.

Trend analysis 2009/10 to 2014/15
Regional analysis shows that rates can be reduced even with a lower starting point: Midlands and East of England was below the national average in 2009/10 but its rate decreased in step with the national trend falling some two percentage points by 2014/15.
Dental – Provision of Inlays 2014/15 (4)

• In 2014/15, 10 per cent of contracts provided half of all inlays.

• It was estimated that outlier contracts provided an excess equivalent to 34,800 courses of Band 3 treatment nominally equivalent to £10.5 million in 2014/15.

• Around 100 contracts were identified with unusually high volumes of inlay treatments but where there was an unusually low radiograph rate.
Dental - Orthodontic Treatment Completions (1)

Why?
• Primary Care NHS orthodontic treatment in England and Wales is accounted for by Units of Orthodontic Activity (UOA).
• Previous analysis highlighted systemic under-reporting of treatment outcomes which is thought to be a result of the current remuneration system, whereby contractors receive the full cost of a completed course of orthodontic treatment, averaging £1,200 per case, through crediting all Units of Orthodontic Activity (UOA) at the assessment and fitting of an appliance and none at termination.

What?
• Analysis was undertaken of orthodontic courses of treatment started in 2010, covering England & Wales, to assess how, when and if the treatments were completed over a five year period.
Dental - Orthodontic Treatment Completions (2)

- Cost of orthodontic treatment started and not completed – estimated at £20 million per year across England & Wales

- Analysis of 150 thousand courses of treatment for patients aged 11-17 at treatment start in 2010 where the treatment outcome is known

- Systemic under-reporting of treatment outcomes – one in five outcomes unknown four years after treatment start (a further 40,000 courses of treatment)

- Able to model treatment completion based on patient gender, age and assessment of treatment need

Data Source: NHSBSA Dental Services data warehouse
Dental - Orthodontic Treatment Duration (3)

- Survival analysis of patients with a known outcome where the patient completed or failed to complete (abandoned or discontinued) their treatment.

Analysis based on around two hundred thousand courses of orthodontic treatment which nominally represents a quarter of a billion pounds and 4.2 million UOAs.

Thirty six percent of cases where treatments are either incomplete, short in duration (likely to be misclassified) or still on-going, are categorised as inappropriate.
Dental - Orthodontic Treatments Summary (4)

Analysis based on around two hundred thousand courses of orthodontic treatment which nominally represents a quarter of a billion pounds and 4.2 million UOAs.

- Treatments which failed to complete - £20 million or 16,000 additional courses of treatment or 8% of total activity

- Short-lived completion cases likely to misclassified - £26 million or 21,000 additional courses of treatment or 10% of total activity

- Treatments that were still either on-going or the conclusion not reported after 48 months - £44 million or 35,000 additional courses of treatment or 17% of total activity.

In total an estimated £90 million at risk per year!
Dental - Treatment Splitting (1)

Why?
• NHS protect carried out an exercise “to measure the occurrence of dental claims where the dentist claims for more than one course of treatment within a 28 day period”.
• The exercise claims to have found “2.2 million claims that have commenced within 28 days for the payment period January to December 2013 involving 7.4 million Units of Dental Activity’ (UDA)s”.
• “Using an average UDA value of £25 this amounts to £185.7 million spend to the NHS”.

What?
• Levels of “irregular claims” were calculated by disallowing the lowest band claim when it occurred within 28 days of another (this could be the first or second claim).
Dental - Treatment Splitting (2)

- Data analysed for January 2013 to December 2013
- 38 million FP17s (dental claim forms) processed from dentists in England and Wales
- 908 thousand FP17s were submitted within 28 days of a previous FP17 involving 2 million Units of Dental Activity
- Estimation of £52m at risk
- Estimated amounts at risk for individual dental contracts range from £25 to £120k
- Twenty five contracts have an estimated risk of over £50k
- Dental Services have started a project to realise this risk.
Prescribing in Care Homes (1)

Why?
• Working collaboratively with internal and external services to reduce medicines waste within care homes. Baselining the current position so that Vanguard and other meds optimisation initiatives can be tested for their effectiveness. Providing patients with better care.
• An estimated £240m prescribing spend for care home users in England with an estimated £50m of pharmaceutical waste per annum.
• Many causes of waste are due to residents having multiple conditions that necessitate complex treatments (polypharmacy), together with a high turnover of staff, lack of staff education and a lack of continuity of care.
Care Homes - National Figures (2)

What?

• Using our prescription data to understand the current level of prescribing for patients in Care homes.

• Advanced Analytics has allowed us to combine publically available data to identify care home locations and help to identify patients in Carehomes so we can look at trends and waste within the wider NHS.

<table>
<thead>
<tr>
<th>Number of different care homes</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>177,006</td>
</tr>
<tr>
<td>2</td>
<td>3,516</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Number of distinct patients using EPS

- Female
- Male
Prescribing in Care Homes (3)

Why is this interesting?

• Growth opportunities for NHSBSA within provider management through providing data and helping monitor patients in care homes; identifying patients in receipt of 10 or more items (polypharmacy).

• Analyses could be extended to all prescriptions as NHS number is available on 9 in 10 paper forms (data sharing required).

• Information developed through this pilot could be made available to other Vanguards, CCGs and Care Homes.
Prescribing for cancer patients (1)

Why?
• Improve PHE understanding of a cancer patient’s journey and interactions with the NHS which in turn will improve the information and care provided to patients.

What?
• Prescription data transferred from an NHSBSA dataset covering data processed by the NHSBSA between the 1st February 2014 and the 30th April 2014 to the National Cancer Registration Service (NCRS).
  ✓ There were 9.9 million rows of data in the pilot prescription dataset
  ✓ Of which, 2.2 million rows of prescriptions data linked to PHE’s cancer register representing 22% of the total prescription dataset.
• This has importance in terms of public health by giving us a better understanding of:
  ✓ With which drugs cancer patients are being treated for in the community
  ✓ Other conditions patients may be receiving treatment for (co-morbidity)
  ✓ Geographical variations and differences according to age, sex, and cancer type.
  ✓ Which prescriptions cancer patients received prior to diagnosis and whether this can help identify at risk groups.
Uptake of Electronic Prescribing (1)

Why?
- The NHSBSA is working with the HSCIC to increase the uptake of electronic prescribing:
  - efficiencies for the NHS
  - operational benefits.

What?
- Identifying any patient groups and drugs which could be targeted to help increase the uptake.
## Uptake of electronic prescribing (2)

### Findings

#### Patients
- Younger patients (under 50 years old) are under-represented in electronic prescribing, particularly those aged under 30 years old.
- Males are less likely than females to receive electronic prescribing.

#### Drugs
- Electronic prescribing does not differ too much from paper prescribing in terms of the most common drugs.
- The main drug type where electronic prescribing is lower is antibacterial drugs - often prescribed on an acute, short-term basis.

#### Repeat Dispensing
Repeat dispensing is an area for opportunity to increase electronic prescribing:
- Top 10 drugs closely mirror each other for paper and electronic repeat dispensing.
- The age profile fits.
- Patients aren’t visiting for acute medication reasons.

#### Controlled drugs
- One-fifth of patients who receive controlled drugs via paper also receive items electronically
- Virtually all (98%) of prescription forms which contained controlled drugs only contained one schedule type on the same prescription.
Using knowledge of dental electronic uptake and the current rate of electronic prescriptions it is predicted that May 2016 will be the first month in which EPS processing will match paper processing.
Uptake of electronic prescribing (4)

Next steps

• Prioritise repeat dispensing patients as a patient group to target.
• Prioritise electronic prescribing patients who receive controlled drugs via paper prescriptions.
• Use the NHSBSA regular survey programme to gather insight from patients on awareness levels and perceptions of electronic prescribing.
• Use the above research to help inform a multi-strategy approach for different patient groups. There is not necessarily a lot of ground to be gained by targeting specific drugs / drugs groups.
• Develop a range of metrics in respect of electronic prescribing at CCG level, where comparisons can be made.
• Undertake further analysis to understand why patients who have received electronic prescriptions are also receiving paper prescriptions.
Potential Savings – FY2015/16

• Our target for 2015/16 is to highlight at least £200 million of potential savings for the NHS through the DALL. The thermometer below shows what we have achieved so far towards that target:

• The DALL, however, isn’t purely about saving money as we can also provide valuable insight into patient care, safety, probity and quality within the NHSBSA and wider NHS.
The DALL Story so far...

- **2013/14**
  - QTR4: LT approval for DALL Pilot to commence
  - QTR2: Workshop held to brainstorm potential initiatives

- **2014/15**
  - QTR4: Oracle hardware and software installed
  - QTR2: System and Process Goals agreed
  - QTR3: Quality Assurance Process established
  - QTR1: Pilot Team established
  - CRISP - DM Established
  - Operating Model Defined
  - POC with NHSBSA/Oracle

- **2014/16**
  - QTR4: NHSBSA potential savings identified = £1.1m
  - QTR3: NHS potential savings identified = £133m
  - QTR2: A total of 33 recommendations made with 9 being implemented
  - QTR1: DALL Pilot evaluation completed
  - QTR1: Decision to launch DALL as BAU

- **2015/16**
  - DALL adopts SCRUM methodology
  - DALL ensoanced in the NHSBSA Strategic Goals
  - Aligning and tracking benefits realisation with PACIFIC and Finance

  *In QTRs 1 & 2; 19 initiatives have been completed and potential savings of £146m identified*
If you have further questions or comments please contact:

Abigail.Haigh@NHS.net