Anomaly and Fraud Detection with Oracle Data Mining 11g Release 2

Charlie Berger
Sr. Director Product Management, Data Mining Technologies
Oracle Corporation
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine
20 Ways to Detect Fraud

http://www.auditnet.org/testing_20_ways.htm

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 a 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.

http://www.fraudiscovery.com/detect.html

Copyright © 2010 Oracle Corporation
What is Data Mining?

• Automatically sifts through data to find hidden patterns, discover new insights, and make predictions

• Data Mining can provide valuable results:
  • Predict customer behavior (*Classification*)
  • Predict or estimate a value (*Regression*)
  • Segment a population (*Clustering*)
  • Identify factors more associated with a business problem (*Attribute Importance*)
  • Find profiles of targeted people or items (*Decision Trees*)
  • Determine important relationships and “market baskets” within the population (*Associations*)
  • Find fraudulent or “rare events” (*Anomaly Detection*)
Data Mining Provides
Better Information, Valuable Insights and Predictions

Customer Months

Cell Phone Churners vs. Loyal Customers

Insight & Prediction

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

Segment #3:
IF CUST_MO > 7 AND INCOME < $175K, THEN Prediction = Cell Phone Churner, Confidence = 83%, Support = 6/39

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff

Copyright © 2010 Oracle Corporation
# Oracle Data Mining Algorithms

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Logistic Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Popular / Rules / transparency</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>Embedded app</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Multiple Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Lack examples</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>Hierarchical K-Means</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Text mining</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td>NMF</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
You Can Think of It 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 Data Mining
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine

ODM “Verbs”
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Fraud Prediction

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', 'CLAIMS2', '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 CLAIMS2
where PASTNUMBEROFCLAIMS in ('2 to 4', 'more than 4'))
where rnk <= 5
order by percent_fraud desc;

<table>
<thead>
<tr>
<th>POLICYNUMBER</th>
<th>PERCENT_FRAUD</th>
<th>RNK</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>
Building a Check Fraud Detection System Using Oracle 11g and Oracle Data Mining

Abstract

It is estimated that the nation's banks experience over $10 billion per year in attempted check fraud. In order to minimize this type of fraud loss, a data driven predictive model is assembled, tested and deployed within a single platform using Oracle 11g and Oracle Data Mining (ODM). Since all of the processes can be executed within a database, it provides a level of security crucial to banks. Oracle 11g and ODM allow users to perform a multitude of analysis using large volumes of transactional data. The features of Oracle 11g and ODM (e.g. analytical functions, various predictive models, ease of model deployment, materialized views, partitioning, etc.) result in substantially increased productivity, manageability and scalability.

About the presenter

Jiang Zhou
• 11 years “stem celling analytics” into Oracle
  • Designed advanced analytics into database kernel to leverage relational database strengths
  • Naïve Bayes and Association Rules—1st algorithms added
  • Leverages counting, conditional probabilities, and much more

• Now, analytical database platform
  • 12 cutting edge machine learning algorithms and 50+ statistical functions
  • A data mining model is a schema object in the database, built via a PL/SQL API and scored via built-in SQL functions.
  • When building models, leverage existing scalable technology
    • (e.g., parallel execution, bitmap indexes, aggregation techniques) and add new core database technology (e.g., recursion within the parallel infrastructure, IEEE float, etc.)
  • True power of embedding within the database is evident when scoring models using built-in SQL functions (incl. Exadata)

```sql
select cust_id
from customers
where region = 'US'
and prediction_probability(churnmod, 'Y' using *) > 0.8;
```
Oracle Data Miner 11g Release 2 GUI
Free product on OTN for Oracle Data Mining Option

- Graphical User Interface for data analyst
- SQL Developer Extension (OTN download)
- Explore data—discover new insights
- Build and evaluate data mining models
- Apply predictive models
- Share analytical workflows
- Deploy SQL Apply code/scripts
The Forrester Wave™: Predictive Analytics And Data Mining Solutions, Q1 2010

Oracle Data Mining Cited as a Leader; 2nd place in Current Offering

- Ranks 2nd place in Current Offering
- “Oracle focuses on in-database mining in the Oracle Database, on integration of Oracle Data Mining into the kernel of that database, and on leveraging that technology in Oracle’s branded applications.”
Data Mining Provides
Better Information, Valuable Insights and Predictions

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff
Finding Needles in Haystacks

• Haystacks are usually BIG

• Needles are typically small and rare
Look for What is “Different”
Challenge: Finding Anomalous Records

- Considering multiple attributes, taken collectively, a record may appear anomalous.
In-Database Data Mining

Traditional Analytics

Data Import

Data Mining Model “Scoring”

Data Preparation and Transformation

Data Mining Model Building

Data Prep & Transformation

Data Extraction

Oracle Data Mining

Results

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

Savings

Model “Scoring”
Data remains in the Database
Embedded data preparation
Cutting edge machine learning algorithms inside the SQL kernel of Database
SQL—Most powerful language for data preparation and transformation
Data remains in the Database

Copyright © 2010 Oracle Corporation
# Oracle Data Mining

## Algorithm Summary 11g Release 2

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Logistic Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Popular / Rules / transparency</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>Embedded app</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td>Regression</td>
<td>Multiple Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Anomaly Detection</strong></td>
<td>One Class SVM</td>
<td>Lack examples</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</td>
<td>Hierarchical K-Means</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Text mining</td>
</tr>
<tr>
<td></td>
<td>NMF</td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>

---

**Oracle Corporation**

Copyright © 2010 Oracle Corporation
Oracle Data Mining 11g Release 2
Anomaly Detection

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

Problem: Detect rare cases
Oracle Data Mining and Unstructured Data

- Oracle Data Mining mines unstructured i.e. "text" data
- Include free text and comments in ODM models
- Cluster and Classify documents
- Oracle Text used to preprocess unstructured text
Predictive Analytics Applications

**Powered by Oracle Data Mining**

(Partial List as of March 2010)

**Oracle Communications Data Model**

- CRM OnDemand—Sales Prospector
  - Oracle Open World - Schedule Builder

**Oracle Retail Data Model**

- Spend Classification

---

(Copyright © 2010 Oracle Corporation)
Oracle Data Miner 11g Release 2 GUI
Oracle Data Miner GUls
Free Products on OTN

11g Release 2 and previous releases—ODM’r “Classic”

11g Release 2 and future—ODM’r “New” or just ODM’r
IF BANK_FUNDS > 246
AND CHECKING_AMOUNT <= 232.5
AND CREDIT_BALANCE <= 24.5
AND MONEY_MONTHLY_OVERDRAWN > 54.145
AND T_AMOUNT_AUTOM_PAYMENTS <= 8783.5
THEN Yes
### Coefficients

**Target Value:** Yes

**Sort by absolute value:** Yes

**Fetch Size:** 10,000

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_TRANS_ATM</td>
<td>NULL</td>
<td>4.61320837</td>
</tr>
<tr>
<td>T_AMOUNT_AUTOM_PAYMENTS</td>
<td>NULL</td>
<td>-4.25073236</td>
</tr>
<tr>
<td>CREDIT_BALANCE</td>
<td>NULL</td>
<td>-3.93007442</td>
</tr>
<tr>
<td>CHECKING_AMOUNT</td>
<td>NULL</td>
<td>-3.73001909</td>
</tr>
<tr>
<td>N_TRANS_WEB_BANK</td>
<td>NULL</td>
<td>-3.63538508</td>
</tr>
<tr>
<td>&lt; Intercept&gt;</td>
<td>NULL</td>
<td>-2.76534902</td>
</tr>
<tr>
<td>N_TRANS_TELLER</td>
<td>NULL</td>
<td>2.20391478</td>
</tr>
<tr>
<td>LTV</td>
<td>NULL</td>
<td>2.14265801</td>
</tr>
<tr>
<td>AGE</td>
<td>NULL</td>
<td>-1.66463722</td>
</tr>
<tr>
<td>N_OF_DEPENDENTS</td>
<td>NULL</td>
<td>0.66557898</td>
</tr>
<tr>
<td>MORTGAGE_AMOUNT</td>
<td>NULL</td>
<td>-0.85111609</td>
</tr>
<tr>
<td>STATE</td>
<td>UT</td>
<td>-0.76673662</td>
</tr>
<tr>
<td>STATE</td>
<td>TX</td>
<td>0.68542101</td>
</tr>
<tr>
<td>SALARY</td>
<td>NULL</td>
<td>-0.55453918</td>
</tr>
<tr>
<td>STATE</td>
<td>AZ</td>
<td>-0.48900215</td>
</tr>
<tr>
<td>STATE</td>
<td>MO</td>
<td>-0.34647433</td>
</tr>
<tr>
<td>STATE</td>
<td>MO</td>
<td>-0.34306028</td>
</tr>
<tr>
<td>STATE</td>
<td>NV</td>
<td>0.32794196</td>
</tr>
<tr>
<td>STATE</td>
<td>WI</td>
<td>0.30801241</td>
</tr>
<tr>
<td>HOUSE_OWNERSHIP</td>
<td>2</td>
<td>-0.34647433</td>
</tr>
<tr>
<td>STATE</td>
<td>MO</td>
<td>-0.34306028</td>
</tr>
<tr>
<td>STATE</td>
<td>NV</td>
<td>0.32315484</td>
</tr>
<tr>
<td>STATE</td>
<td>WI</td>
<td>0.28906774</td>
</tr>
<tr>
<td>HOUSE_OWNERSHIP</td>
<td>0</td>
<td>-0.27872530</td>
</tr>
<tr>
<td>TIME_A5_CUSTOMER</td>
<td>5</td>
<td>0.24943188</td>
</tr>
<tr>
<td>TIME_A5_CUSTOMER</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
### Rule Content: Name,Subname

#### Rules: 306 out of 306

<table>
<thead>
<tr>
<th>ID</th>
<th>Antecedent</th>
<th>Consequent</th>
<th>Lift</th>
<th>Confidence(%)</th>
<th>Support(%)</th>
<th>Length</th>
<th>Antecedent Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>175</td>
<td>PROD_ID.139</td>
<td>PROD_ID.139</td>
<td>12.7801</td>
<td>68.9365</td>
<td>4.1163</td>
<td>1</td>
<td>5.3941</td>
</tr>
<tr>
<td>176</td>
<td>PROD_ID.137</td>
<td>PROD_ID.137</td>
<td>12.7801</td>
<td>76.3114</td>
<td>4.1163</td>
<td>1</td>
<td>5.9711</td>
</tr>
<tr>
<td>305</td>
<td>PROD_ID.140 AND PROD_ID.148</td>
<td>PROD_ID.146</td>
<td>11.1358</td>
<td>78.8665</td>
<td>4.8603</td>
<td>2</td>
<td>7.084</td>
</tr>
<tr>
<td>141</td>
<td>PROD_ID.126</td>
<td>PROD_ID.125</td>
<td>11.0495</td>
<td>83.4936</td>
<td>5.3396</td>
<td>1</td>
<td>7.5563</td>
</tr>
<tr>
<td>142</td>
<td>PROD_ID.125</td>
<td>PROD_ID.126</td>
<td>11.0495</td>
<td>70.6638</td>
<td>5.3396</td>
<td>1</td>
<td>6.3952</td>
</tr>
<tr>
<td>288</td>
<td>PROD_ID.123 AND PROD_ID.127</td>
<td>PROD_ID.128</td>
<td>10.6796</td>
<td>94.8814</td>
<td>4.0793</td>
<td>2</td>
<td>8.8844</td>
</tr>
<tr>
<td>253</td>
<td>PROD_ID.116 AND PROD_ID.119</td>
<td>PROD_ID.115</td>
<td>10.6282</td>
<td>58.6733</td>
<td>4.27</td>
<td>2</td>
<td>5.5205</td>
</tr>
<tr>
<td>306</td>
<td>PROD_ID.140 AND PROD_ID.146</td>
<td>PROD_ID.146</td>
<td>10.6153</td>
<td>86.724</td>
<td>4.8603</td>
<td>2</td>
<td>8.1697</td>
</tr>
<tr>
<td>304</td>
<td>PROD_ID.146 AND PROD_ID.148</td>
<td>PROD_ID.140</td>
<td>10.6014</td>
<td>85.6246</td>
<td>4.8603</td>
<td>2</td>
<td>8.0768</td>
</tr>
<tr>
<td>256</td>
<td>PROD_ID.117 AND PROD_ID.119</td>
<td>PROD_ID.115</td>
<td>10.5721</td>
<td>58.3635</td>
<td>4.1708</td>
<td>2</td>
<td>5.5205</td>
</tr>
<tr>
<td>301</td>
<td>PROD_ID.125 AND PROD_ID.117</td>
<td>PROD_ID.115</td>
<td>10.5047</td>
<td>91.7456</td>
<td>4.1931</td>
<td>2</td>
<td>8.8012</td>
</tr>
</tbody>
</table>

#### Rule Details:

**ID:** 305

**IF**

- PROD_ID.140 AND
- PROD_ID.148

**THEN**

- PROD_ID.146
Presentation of Results and Integration with Applications
Integration with Oracle BI EE

Oracle BI EE defines results for end user presentation.

Oracle Data Mining results available to Oracle BI EE administrators.
Example

Better Information for OBI EE Reports and Dashboards

ODM’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools.
Integration with Oracle BI EE

ODM provides likelihood of expense reporting fraud ....and other important questions.
Oracle Spend Classification
Classify Spend into Purchasing Categories

• Features
  • Hierarchical classification and scoring
  • Auto Spend Classification – Inline and Batch
  • Integration to OBIA Procurement & Spend Analytics 7.9.6

• Benefits
  • Classifies spend data from various sources into procurement category hierarchies
  • Category normalization aids strategic sourcing and contract negotiation
  • In-line mode integrated with EBS iProcurement
Additional Information

- Preview of the new Oracle Data Miner 11g R2 “work flow” New GUI
- Oracle Data Mining 11gR2 presentation at Oracle Open World 2009
- Oracle Data Mining Blog
- Funny YouTube video that features Oracle Data Mining
- Oracle Data Mining on the Amazon Cloud
- Oracle Data Mining 11gR2 data sheet
- Oracle Data Mining 11gR2 white paper
- New TechCast (audio and video recording): ODM overview and several demos
- Fraud and Anomaly Detection using Oracle Data Mining 11g presentation
- Algorithm technical summary with links to Documentation
- Getting Started w/ ODM page w/ instructions to download
  - Oracle Data Miner graphical user interface (GUI),
  - ODM Step-by-Step Tutorial
  - Demo datasets
- ODM Discussion Forum on OTN (great for posting questions/answers)
- ODM 11g Sample Code (examples of ODM SQL and Java APIs applied in several use cases; great for developers)
- Oracle’s 50+ SQL based statistical functions (t-test, ANOVA, Pearson’s, etc.)
“This presentation is for informational purposes only and may not be incorporated into a contract or agreement.”