Data Mining Techniques and Opportunities for Taxation Agencies

Louis Panebianco
Florida

Eric Wojcik
Consultant
Audit, Tax, Consulting, Financial Advisory.

In This Session ...

- You will learn the data mining techniques below and their application for Tax Agencies
  - ABC Analysis
  - Association Analysis
  - Clustering
  - Decision Trees
  - Score Carding Techniques

- You will be provided with descriptions of real-world usage and ideas for practical application in your agency
- You will hear examples of their usage from the State of Florida
Agenda

- What is Data Mining? Eric Wojcik
- Data Mining Techniques Eric Wojcik
- How can we use Data Mining in our Agency? Eric Wojcik
- Usage in Florida Louis Panebianco
- Recap/Closing Louis Panebianco

A Few Common Questions...

Data mining still holds a mythical reputation

- Is it very complicated?
  - Advanced techniques still are... on paper!
  - Conceptually, techniques are still a challenge. Application is usually not.

- What can data mining do?
  - Turn tomes of data into small, digestible, ideally actionable amounts of information

- What staff do I need to mine data?
  - In the past it was the actuarial staff, select financial staff, and PhDs. Tools of today allow the process Power Users to data mine effectively.
What is Data Mining? – The Answer!

Definition:
“…the science of extracting useful information from large data sets or databases.”

“…the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.”

“…the statistical and logical analysis of large sets of … data, looking for patterns that can aid decision making.”

“Intentionally using statistical techniques on a large amount of data to either discover new insight or to convert it into actionable information.”
- Eric Wojcik Reporting and Analytics 2008 Conference

What is data mining? – Vital Concepts

- Focused Ideas – Especially Goal
- Data Quality/Integrity
- Discovery vs. Predictive
- Training, Benchmarking, and Prediction
ABC Analysis - Concept

- ABC Analysis is similar to the Pareto Principle
  - 80/20 Rule
  - The Law of Vital Few
  - The Principle of Factor Sparsity
- ABC Analysis is primarily used as a discovery technique.

- Concept: 80% of the outcome come from 20% of the root causes
  - Call Center: 80% of all incoming calls are in reference to 20% of the potential reasons for calling in
  - Economics: 80% of a country’s wealth is owned by 20% of the population.*
ABC Analysis – Concept (cont.)

- ABC Analysis is a slight deviation of the Pareto Principle
  - Began as an inventory classification technique and commonly referenced in Supply Chain applications

  ABC Codes break down as following:
  - “A” Class contains 60% of total value
  - “B” Class contains 20% of total value
  - “C” Class accounts for the remaining 20% of total value

  - Classification codes are very process specific and likely have a different meaning across each process

Association Analysis - Concept

- Association Analysis unveils hidden interactions, correlations, and patterns in your data.
  - Affinity Analysis
  - Market Basket Analysis (MBA)

- Concept:
  - Uncover patterns in the data and then define rules to the interaction.
  - “If a person purchases onions and potatoes, they are 50% more likely to also purchase beef.”
    - {Onions, Potatoes} → {Beef}
Association Analysis – Concept (cont.)

- Transactions determine the rules
  - The rules should change over time - moving timeframe

- Association Analysis is typically augmented with other techniques if automated

Clustering - Concept

- Clustering is statistically separating records into smaller unique groups based on common similarities
  - Also known as a Classification Problem
- Clustering can be used for discovery or prediction.

- The two primary forms of clusters are Hierarchical and Partitional
Clustering – Concept (cont.)

- Hierarchical Clusters

- Partitional Clusters
Decision Trees

- An easy to understand method that classifies possible outcomes from previous decisions based on characteristics in the data and develops rules to predict the future outcomes.

- Once a tree is “trained” and created it can be used on new data to predict decisions.

- Trees can even be partially defined to meet regulatory requirements.

Decision Trees – Concept (cont.)

- Decision Trees can also be used in two ways
  - Discovery – to find new insight into the data
  - Prediction – to statistically guess what the outcome will be (requires a “trained” decision tree)

- A prediction tree can automate business processes, but, must be periodically retrained.

- Commonly used in
  - Insurance
  - Legal
  - Sales/Marketing
Score Carding Techniques

- Score Carding is a technique in which a valuation function is applied to value (score) a data set
  - Approximation

- The two most common techniques are weighted scoring and regression
  - Weighted Scoring is a manually defined score
  - Regression (either linear or non-linear) is an automated method for scoring

Score Carding Techniques (cont.)

Weighted Average

\[ \bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \]

Linear Regression

Non-Linear Regression
Agenda

What is Data Mining? Eric Wojcik

Data Mining Techniques Eric Wojcik

How can we use Data Mining in our Agency? Eric Wojcik

Usage in Florida Louis Panebianco

Recap/Closing Louis Panebianco

How can we use Data Mining?

ABC Analysis
- Audit Lead/Discovery
- Inventory Management
- Intake/Tax Form Processing
- Call Center Optimization

Association Analysis
- Audit Discovery / Campaigns
- Taxpayer Education
- Website/E-Service Optimization
- System User Interface

Clustering
- Taxpayer Education
- Audit Handling
- Call Center Optimization
- Macro-Economic Discovery

Copyright © 2008 Deloitte Development LLC. All rights reserved.
How can we use Data Mining?

**Decision Trees**
- Intake/Tax Form Processing
- Fund Distribution
- Call Centers/Customer Service
- Audit Selection/Leads
- Regulatory Compliance

**Score Carding/Approximation**
- Audit Selection/Leads
- Campaigns
- Performance Metrics

---

**Agenda**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is Data Mining?</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>Data Mining Techniques</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>How can we use Data Mining in our Agency?</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>Usage in Florida</td>
<td>Louis Panebianco</td>
</tr>
<tr>
<td>Recap/Closing</td>
<td>Louis Panebianco/Eric Wojcik</td>
</tr>
</tbody>
</table>
Florida’s Decision Tree Distributions

1. Extraction from core financials system into data warehouse
2. Legislative and business logic applied via decision tree
3. Results are reloaded into financial system
4. Summary sent to Comptroller’s Office for payment

Florida’s Lead Development

1. Data is extracted into data warehouse
2. Rules are applied based on tax payer characteristics in each industry to derive score
3. Data is loaded into the Case Management System (CRM) and rank ordered to select leads
Agenda

<table>
<thead>
<tr>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is Data Mining?</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>Data Mining Techniques</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>How can we use Data Mining in our Agency?</td>
<td>Eric Wojcik</td>
</tr>
<tr>
<td>Usage in Florida</td>
<td>Louis Panebianco</td>
</tr>
<tr>
<td>Recap/Closing</td>
<td>Louis Panebianco/Eric Wojcik</td>
</tr>
</tbody>
</table>

Questions