Predictive Analytics and Fraud Prevention

James E. Parry
Client Technical Specialist, Predictive Analytics
IBM
Agenda

- The Nature of Fraud
- Fraud Detection Improvement with Predictive Analytics
  - Enforcement Funnels
- Data Mining Classes - Under the Hood
  - Supervised vs. Unsupervised Learning
    - Supervised: Prediction and Classification
    - Unsupervised: Clustering, Association and Anomaly Detection
- Example: Running a Decision Tree
- Deployment Technology: Making Findings Matter
- Customer Stories
- Why IBM SPSS?
- Q&A
The Nature of Fraud

“As fraud schemes become more sophisticated and migratory, access to real time data and the use of advanced data analysis to monitor claims and provider characteristics are critically important.”

-Daniel R. Levinson,
Inspector General
Office of the Inspector General
U.S. Department of Health and Human Services
The Nature of Fraud

“As fraud schemes become more sophisticated and migratory, access to real time data and the use of advanced data analysis to monitor claims and provider characteristics are critically important.”

-Daniel R. Levinson, Inspector General
Office of the Inspector General
U.S. Department of Health and Human Services
The Nature of Fraud

“As fraud schemes become more sophisticated and migratory, access to real time data and the use of advanced data analysis to monitor claims and provider characteristics are critically important.”

-Daniel R. Levinson,
Inspector General
Office of the Inspector General
U.S. Department of Health and Human Services
The Nature of Fraud

“As fraud schemes become more sophisticated and migratory, access to real time data and the use of advanced data analysis to monitor claims and provider characteristics are critically important.”

The Nature of Fraud

“As fraud schemes become more sophisticated and migratory, access to real time data and the use of advanced data analysis to monitor claims and provider characteristics are critically important.”

-Daniel R. Levinson, Inspector General
Office of the Inspector General
U.S. Department of Health and Human Services
Fraud Detection Improvement with Analytics

Fraud potential

Optimized Detection

Automatic Detection

Training

Intuition & Luck

Traditional

Analytics
**Behavior modeling** uses analytics to select providers for investigation based on the behavior they exhibit.

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is the behavior you are trying to identify?</strong></td>
<td><strong>Who is likely to behave in that manner?</strong></td>
</tr>
<tr>
<td><strong>What data can you use to measure “behavior”?</strong></td>
<td><strong>What data can you use to identify “who”?</strong></td>
</tr>
</tbody>
</table>

---

© 2010 IBM Corporation
Fraud Detection Improvement with Predictive Analytics

All Transactions

Real Fraud

Detected Fraud

Investigations based on expert rules & intuition
Fraud Detection Improvement with Predictive Analytics

All Transactions

- New groups for investigation
- Discover low fraud risk transactions
- Discover high fraud risk transactions
Fraud Detection Improvement with Predictive Analytics

- Investigate high risk transactions
- Do NOT investigate low risk transactions

Double or triple the detected fraud with the same resources
Fraud Detection Improvement with Predictive Analytics

All Transactions

Improved identification leads to detection of new forms of fraud

Similar frauds are detected and new fraud profiles are created
Enforcement funnels

- **Identify**
  - Inefficient Identification
    - Good Investigation
  - Good Identification
    - Problematic Investigation

- **Investigate**
  - "Martini Glass"
  - "Margarita Glass"

- **Prosecute**

- **Convict**

- **Identify**
  - Good Identification/Investigation
    - Problematic Prosecution
  - Streamlined process

- **Investigate**
  - "Wine Glass"
  - "Shot Glass"

- **Prosecute**

- **Convict**
Data Mining Classes
Data Mining

- Three classes of data mining algorithms
- Supervised vs. unsupervised
- Complementary

Supervised vs. unsupervised

Cluster “Differences”

Group cases that exhibit similar characteristics.

Predict who is likely to exhibit specific behavior in the future.

Predict “Relationships”

What events occur together? Given a series of actions; what action is likely to occur next?

Associate “Patterns”

Data Mining
What is Unsupervised Learning?

- A data mining technique when we do not know the output or outputs
- Can be thought of as finding ‘useful’ patterns above and beyond noise…or “fishing” for information
- Looks for natural groupings in the data
- Can be used for data reduction, preparation and simplification
What is Supervised Learning?

- A technique when we know the output or outputs
- We will “Supervise” the algorithm and tell it what we want to predict.
- Often uses the results of unsupervised learning as predictors
- Used to predict usually an outcome or a quantity.
Unsupervised Learning: Association, Clustering and Anomaly Detection

- **Associations**
  - Finds things that occur together – ex: events in a crime incident such as fraud
  - Associations can exist between any of the attributes (no single outcome like Decision Trees)

- **Clustering/Anomaly**
  - An exploratory data analysis technique
  - Reveals natural groups within a data set
  - No prior knowledge about groups or characteristics
  - Encompasses Anomaly Detection

- **Sequential Associations**
  - Discovers association rules in time-oriented data
  - Find the sequence or order of the events
Predictive Analytics: Association

- For Example: Patient Sharing

- Customized web plot shows linking between any two providers that shared 5 or more patients
- Link between each of the providers sharing patients displayed by a connecting line of various gradations of width and structure
The results show that 859 patients have been billed for services by At Your Doorstep HHA (28% of the total patient population analyzed in the LA area)

- Of these 859 patients, 4.3% (37 patients) were also billed for services by Help For the Homebound HHA
Unsupervised Learning Technique: Anomaly Detection

- **Anomalies**
  - Anomaly detection is an exploratory method
  - Designed for quick detection of unusual cases or records that should be candidates for further analysis
  - These should be regarded as suspected anomalies

- Unlike other modeling methods that store rules about unusual cases, anomaly detection models store information on what normal behavior looks like. This makes it possible to identify outliers even if they do not conform to any known pattern.
Anomalous Records

- Each record is assigned an anomaly index, ($O$-AnomalyIndex) which is the ratio of the group deviation index to its average over the cluster that the case belongs to.
- The larger the value of this index, the more deviation the case has than the average.
- Under the usual circumstance, cases with anomaly index values less than 1 or even 1.5 would not be considered as anomalies, because the deviation is just about the same or a bit more than the average.
- However, cases with an index value greater than 2 could be good anomaly candidates because the deviation is at least twice the average.
Supervised Learning: Profile and Predict Decision Trees

- Excellent at uncovering and modeling complex relationships
- Very accurate on even small data sets to inform decision making.
- Can handle nonlinear relationships with complex interactions.
- Very easy to understand and describe to others.
- Time to insight in minutes.
Running a Decision Tree
Running a Decision Tree
Deployment
Deployment: Tactical Environment

- In Data Mining, time to insight is half the battle. **Time to production** is the other half (and much more repetitive).
- Must be able to ‘deploy’ model into operations:
  - Quickly
  - In a standards-based, repeatable fashion
- Must be able to monitor model performance for ‘drift’.
  - Automating model performance monitoring and model refresh decreases errors because it’s a ‘hands off’ operation- no user intervention required.
- **Automating model refresh guarantees the most accurate models in the shortest amount of time (time to production).**
Deployment Steps

- Production considerations . . .
  - Models should be easily deployed and managed
  - No SQL programming necessary
  - No DBA intervention
  - Done in a standards-based, replicable fashion (not one-off)
Deployment
From Analyst to Production in Minutes

Real Time Prediction
Customer Story

**Analytics for Improving Collections & Compliance**

- **Business Challenge**
  Identify potential cases from the non-filer group that gets filtered out as low potential during the annual load process:
  - Identify those who failed to file tax returns
  - Determine which ones have positive tax potential
  - Take enforcement actions to obtain the missing returns and related payments

- **Solution**
  Use of IBM Solution to
  - Improve workload selection, enhance workflow, detect anomalies, predict client behavior to determine tax strategies, and increase collection rate
  - Improve the ability to assess the value of non-compliance issues – specifically missing returns – and focus on the “best ROI” workload
  - Identify accounts with good tax potential that are rejecting as “low potential” thru the current legacy business rules system

- **Results**
  - Based on a pilot estimated results for full production year
    - Approx $100 million in revenue for a single tax year
Customer Story

Analytics for Disability Claims

- **Business Challenge**
  Two primary problems with processes for reviewing and approving disability benefits for citizens:
  - Expensive process for performing doctor reviews for Continuing Disability Reviews
  - Huge backlog of new disability applications

- **Solution**
  - Developed a suite of analytical models that included advanced text analytics to infer meaning from unstructured data in disability applications
  - Deployed models to automatically score applications and continuing disability reviews to identify potential quick decisions

- **Results**
  - In use for the last 8 years, the solution has saved over $2 billion in reduced doctor review costs
  - The models for reviewing new disability applications has reduced the average cycle time for approving an application from 90 days to 20 days
  - Spared thousands of disabled citizens from the burden of medical reviews
Summary

• Fraud is complex, sophisticated and real-time.
• Majority of investigations based upon expert rules and intuition
• Predictive analytics algorithms ‘learn’ from the data and can quickly identify suspicious cases
• Applying predictive analytics to cases discovers more fraud, quickly and simply
• Time to model deployment is critical
• Inserting predictive analytics into the claims process leads to streamlined results, higher conviction rates as well as higher constituent satisfaction
Why Institutions Choose IBM Predictive Analytics

- **Total Cost of Ownership**: IBM SPSS leading competitor’s products in general have a significantly higher cost of ownership than IBM SPSS.

- **Flexible licensing** - IBM SPSS offers perpetual licenses as opposed to annual usage licenses which results in a lower TCO over the long term and more flexible licensing schemas.

- **Ease of Use** - IBM SPSS’ technology is much easier and faster to get analysts trained up and ready to use than leading competitors- weeks as opposed to months.

- **Integration**: IBM SPSS’ data mining solution with its open architecture make it easy to integrate with other applications and data sources with minimal coding necessary.

- **Openness of Data Sources**: Because no single database repository is mandatory to conduct data mining the organization is not ‘locked in’ to any one solution.

- **Sophisticated and Open Model Management**: Model Management with IBM SPSS Collaboration and Deployment Services is well ahead of any offering by our leading competitors. IBM SPSS Collaboration and Deployment Services includes features for collection of metadata to manage predictive assets.

- **Integration w/ powerful 3rd party open source languages** - IBM SPSS natively integrates with Python, .Net and the R-Project making it much easier to find qualified programmers and developers as well as creating powerful customizable statistical and data mining algorithms at no charge, in house.

- **Comprehensive Technical Support**: IBM SPSS has 24/7 technical support and a world-class professional services organization that is fully staffed by predictive analytics experts.
Questions?