Predicting Return to Work with Data Mining

Claim Analytics
Jonathan Polon  Barry Senensky
www.claimanalytics.com

IAAHS Colloquium
Dresden, Germany
April 27-29, 2004
Predicting Return to Work with Data Mining

- About Us
- Why Score Claims?
- Data Mining
- Model Benefits
- About the Model
- Building the Model
- Results
- Conclusion
About Us

- Founded by Barry Senensky and Jonathan Polon in 2001
- To use data mining tools to bring new insights and solutions to the insurance industry.

Recent Projects

1. Create a predictive scoring model for group LTD claimants.
2. Produce case study of creating predictive model for SOA Health Section.
Why Score Claims?

Every day, claims managers make many decisions and choices.

*Should they:*

- Order an independent medical examination for Derek T.?
- Provide extensive rehab to Pat B.?
- Call Jacob Z. again, to monitor his progress?
- Have an investigator check out that suspicious-sounding bad back of Brenda B.?
Imagine a system that:

- Scores each claim with a number from 1 to 10, predicting likelihood of recovery within a given time frame.

- Is a fast, objective, consistent method of ranking claims.

- Helps claims managers spend time where it is most productive, and optimize resource allocation.

*What benefits would accrue from such a system?*
Benefits
Benefits

Helps claims managers to quickly establish a starting point for a new claim.
Benefits

Helps claims managers to optimize allocation of time and resources.
Benefits

Can be used to balance the workload among claims personnel.
Benefits

Provides an early indication of changes in aggregate claim quality, allowing claims management to take appropriate financial measures.
Model Benefits

Does not tie up claims staff in new operational activities...

Entails no costly interference in established working procedures.
Benefits

Facilitates early intervention in claims management.
Benefits

Distinguishes between ‘gray area’ claims (those claims which are neither particularly promising nor particularly unpromising) by scoring them as low as ‘4’ or as high as ‘7.’
Data Mining

- How it works
- What it can do
- Data mining tools
Data Mining

- Uses sophisticated statistical tools to “mine” through databases to find hidden patterns and trends.

- Harnesses speed, capability, and capacity of modern computers.

Most statistical methods used in business:

- Predate the invention of electric light, cars and the telephone.

- Were developed under the constraints of what humans were able to calculate, in a reasonable timespan.
Our Data Mining Tools

We use three:

1. CART
   - Powerful filter. Identifies factors with greatest impact; reduces amount of ‘noise’ being introduced to the model from non-impacting factors

2. Neural Networks
   - Optimization tools

3. Genetic Algorithms
Neural Networks
How they learn

- Network is presented with data sample with known outcomes
- Network predicts result, and compares it to actual outcome
- Network parameters are changed to better approximate the sample...
- ...Over and over again.
Uses of Neural Networks

- Neural nets are a statistical tool for making predictions

Used in:
- Detection of credit card, tax, and securities fraud
- Bioinformatics
- Customer behaviour prediction
- Text analysis

But, as yet, rarely in the insurance industry.

Example: Design a neural networks model to predict the results of professional rugby matches.
Neural Networks: Example

Who’s going to win the footy game?

The neural network weights each variable as it sees fit.
Use of This Neural Network Analysis

A tipping model that outperforms all but one professional in its first year of use.

- Alan McCabe, a computer scientist from James Cook University, developed software to predict the results of Australian Rugby League matches.

- He used data from a number of different seasons of the Australian National Rugby League to develop his model.

- In its first year of use, the model achieved 67% accuracy, tying the top newspaper tipper and beating every one one of the rest. In the Final Series matches, having ‘learned’ from the season, the model achieved a 78% success rate.
Genetic Algorithms

- Inspired by Darwinian concept of **survival of the fittest**
- Multiple solutions considered in simultaneity
- Best of these solutions are most likely to “survive”
Genetic Algorithms

Process

- Solutions evolve in two manners:
  - Reproduction
  - Mutation

Solution A + Solution B = New Solution
Genetic Algorithms

Summary

- Solutions evolve over several generations
- When process stops, best surviving solution is chosen
About the Model

- Assigns each claim a score from one to ten, predicting recovery within a given time frame
- Incorporates predictive strengths of both neural networks and genetic algorithms
- Incorporates industry and other external data to enhance robustness and predictivity
- Uses a committee of experts approach: final score averages output from several hundred models.
How We Built the Model

State the Goal

Data Requirements

Split the Data

Filter the Factors

Prepare the Data

Train the Model

Neural Networks and Genetic Algorithms

Validate

The Completed Model
How We Built the Model

The Goal

Build a model to predict likeliness of recovery for LTD claims, producing a single comprehensible output, the score.

Define a benchmark for success:

- 75% or more of claims scored with an 8 - 10 return to work within 2 years.
- 5% or less of claims scored with a 1 - 3 return to work within 2 years.
How We Built the Model

Data Requirements

- Determine the factors that influence recovery
- Determine what data is needed to decide if a claim has recovered
- Determine how many records are required.
How We Built the Model

Split the Data

Validate the data, using a series of manual and automatic checks, and then split it into three parts:

(i) 80% for training the model
(ii) 10% for testing
(iii) 10% for final validation.
How We Built the Model

Filter the Factors

Use an initial filtering tool (we used Salford Systems CART) to key in on which data factors impact recovery most.
Considerable data manipulation goes into readying the data for modeling.

Example I: Diagnostic Category

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Muscular Dystrophy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostic category</strong></td>
<td>Brain &amp; Nervous System</td>
</tr>
<tr>
<td>Impact on vision</td>
<td>0</td>
</tr>
<tr>
<td>Impact on fine motor skills</td>
<td>9</td>
</tr>
<tr>
<td>Impact on gross motor skills</td>
<td>9</td>
</tr>
<tr>
<td>Likelihood of drug treatment</td>
<td>2</td>
</tr>
<tr>
<td>Likelihood of being fatal</td>
<td>3</td>
</tr>
</tbody>
</table>
How We Built the Model

Prepare the Data - II

Example II – Age

Age treated as 3 categorical (unordered) variables: 18-34, 35-49 and 50-65.

<table>
<thead>
<tr>
<th></th>
<th>18-34</th>
<th>35-49</th>
<th>50-65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashley</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bruce</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Claire</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The absolute differences found in the three age categories are much more meaningful to the neural network than what it would have found in just one category, comparing the relative values of the ages.
How We Built the Model

Train the Model

Make design decisions to maximize the ability of the model to learn.

Examples:
Set network size
Set training tolerance
How We Built the Model

Train the Model: Example I

Set network size

• Determines # of weights (parameters) in the model
• Probably the most critical setting.

There is a trade off between accuracy (more weights) and ability to generalize (less weights).
Set Training Tolerance

How accurate must the output be to be considered correct?

During training:

- Neural network cycles through data one record at a time.
- At each record the network compares predicted output to actual output, and adjusts its weights, if necessary, to better approximate the actual output.
- The network continues cycling through the data until there is a set of weights for which every record is within the training tolerance.
How We Built the Model

Neural Networks and Genetic Algorithms...

Neural Nets are:
Fast and efficient but carry risk of getting caught in local minima

Genetic Algorithms are:
Not fooled by local minima but may be slow and inefficient

Characteristics of a Hybrid Approach
Genetic Algorithm finds a good solution ...
Neural Network optimizes it
How We Built the Model

Validate

The completed model was validated by comparing its recovery predictions for the validation data (the 10% of set-aside historic data not seen by the model until this point) to real-world outcomes.
Model Results

Recovery Rate by Score
(real-life example)

<table>
<thead>
<tr>
<th>Model's Predictive Score</th>
<th>0%</th>
<th>20%</th>
<th>40%</th>
<th>60%</th>
<th>80%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Recover</td>
<td>0%</td>
<td>0%</td>
<td>11%</td>
<td>29%</td>
<td>44%</td>
<td>56%</td>
</tr>
<tr>
<td>Actual Recovery</td>
<td>70%</td>
<td>80%</td>
<td>75%</td>
<td>91%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We were delighted with the results of our project. Our model proved itself able to accurately score claimants and predict recovery, providing a valuable tool to help with claims management.

The Claim Analytics scoring model is a

Fast
Objective
Consistent

method of ranking claims, that

Integrates easily into the workplace, and

Helps claims managers optimize resource allocation.
Questions?