Modeling in P&C
Actuarial Science

NAIC Book Club
Louise Francis, FCAS, MAAA, Francis Analytics and Actuarial Data Mining

www.data-mines.com
Louise_francis@msn.com
Why Predictive Modeling?

- Better use of data than traditional methods
- Advanced methods for dealing with messy data now available
- New ways to test and validate models
Real Life Insurance Application – The “Boris Gang”

New York Fraud Ring No Surprise to Russian Drivers
By SABRINA TAVERNAISE
New Yorkers may have been shocked by news of an insurance scheme that involved fake car crashes. But in Russia, fraud is a rule of the road.
August 16, 2003 | WORLD | NEWS
MORE ON ORGANIZED CRIME AND: FRAUDS AND SWINDLING, FOREIGN BANK ACCOUNTS, AUTOMOBILE INSURANCE AND LIABILITY, STATE FARM INSURANCE COS, NEW YORK CITY, RUSSIA, LONG ISLAND (NY)

Investigators Say Fraud Ring Staged Thousands of Crashes
By PATRICK HEALY
The ring used Russian immigrants to stage car accidents and then employed its own network of doctors and fake clinics in New York State to bilk an insurance company out of $48 million.
August 13, 2003 | FRONT PAGE | NEWS
MORE ON ORGANIZED CRIME AND: ACCIDENTS AND SAFETY, FRAUDS AND SWINDLING, FOREIGN BANK ACCOUNTS, CHILDREN AND YOUTH, AGED, WOMEN, AUTOMOBILE INSURANCE AND LIABILITY, SPOTA, THOMAS J, STATE FARM INSURANCE COS, NEW YORK CITY, RUSSIA, WESTCHESTER COUNTY (NY), LONG ISLAND (NY), SWITZERLAND
Overview

- Brief overview of Predictive Modeling
- Some key methods and applications
  - Trends in Methods
- Model Validation
- Data/Big Data
- Software
- Educational and educational resources
- Example of where it could have made a difference
Kinds of Applications

- Classification
  - Target variable is categorical
- Prediction
  - Target variable is numeric
Data Complexities: Nonlinearities
MARS Prediction of Primary Paid Severity

Scatterplot of \( \text{InPaidPred} \) against \( \text{LnReserve} \)
Predictions 1 5^v1844c
Data Complexities: Missing Data

- It is not uncommon for one third of the possible predictors to contain records with missing values.

Possible solutions:

- A data mining method such as CART that uses a statistical algorithm to find an alternative parameterization in the presence of missing data.
- A statistical method such as expectation maximization or data imputation to fill in a value.
Major Kinds of Data Mining

- **Supervised learning**
  - Most common situation
  - A dependent variable
    - Frequency
    - Loss ratio
    - Fraud/no fraud
  - Some methods
    - Regression
    - CART
    - Some neural networks
    - MARS

- **Unsupervised learning**
  - No dependent variable
  - Group like records together
    - A group of claims with similar characteristics might be more likely to be fraudulent
    - Ex: Territory assignment, Text Mining
  - Some methods
    - Principal Components
    - K-means clustering
    - Kohonen neural networks
Methods

- Classical
- Decision Trees
- Neural Networks
- Unsupervised learning
  - Clustering
  - Principal Components
- Newer Methods
  - Ensemble
  - SVM
  - Deep learning
  - Text Mining
Predictive Modeling

Classical

Predictive Modeling

GLMs

Data Mining/Machine Learning
Linear Modeling Tools Widely Available: Excel Analysis Toolpak

- Install Data Analysis Tool Pak (Add In) that comes with Excel
- Click Data, Data Analysis, Regression
Classical Model: Discriminant Analysis

**Canonical Discriminant Function 1**

- **Suspicion = 1**
  - Mean = -0.2
  - Std. Dev. =
  - N = 1,016

- **Suspicion = 2**
  - Mean = 1.49
  - Std. Dev. = 0.979
  - N = 130
Similarities with GLMs

**Linear Models**
- Transformation of Variables
- Use dummy coding for categorical variables
- Residual
- Test significance of coefficients

**GLMs**
- Link functions
- Use dummy coding for categorical variables
- Deviance
- Test significance of coefficients
Neural Networks

- Theoretically based on how neurons function
- Can be viewed as a complex non-linear regression

Deep Learning – Neural Networks on Steroids
Regression Trees

- Tree-based modeling for **continuous target variable**
  - most intuitively appropriate method for loss ratio analysis
- Find split that produces greatest separation in \[ \sum (y - E(y))^2 \]
- i.e.: find nodes with minimal **within variance**
  - and therefore greatest **between variance**
  - like credibility theory i.e.: find nodes with minimal **within variance**

- Every record in a node is assigned the same expectation ➔ model is a **step function**
C&RT

- Binary splits
- Gini Index (categorical), minimize squared error (numeric)
CHAID

- Minimize Chi-Square Statistic
Ensemble Models

- Single Trees (CART, CHAID)
- Ensemble Trees, a more recent development (TREENET, RANDOM FOREST)
  - A composite or weighted average of many trees (perhaps 100 or more)
  - There are many methods to fit the trees and prevent overfitting
    - Boosting: Iminer Ensemble and Treenet
    - Bagging: Random Forest
Data

- Data Management
- Data quality
  - Francis, “Dancing With Dirty Data”, CAS forum, [www.casact.org](http://www.casact.org)
- Big Data

% of Time
BIG DATA – WHAT IT IS, AND WHAT IT MEANS FOR THE INSURANCE INDUSTRY

“The term itself is vague, but it is getting at something that is real... Big Data is a tagline for a process to transform everything.” — Jon Kleinberg, Cornell University

Big data is one of the signature opportunities of the day, but also one of the most poorly defined. There is growing awareness that ever-expanding data sources can enable scientific advances, new product development, and business model innovations. Yet as the recent Google Flu Trends episode illustrates, sloppy methodological thinking about big data can lead to scientific errors. Analogously, sloppy business thinking about big data can lead to expensive strategic errors.

This session will begin by offering a variety of definitions and examples of big data in action. It will then sketch both the technologies and analytical methodologies that enable big data to be converted from a raw material to a strategic asset. With these prerequisites in hand, the session will focus on the strategic aspects of the topic: current uses of big data in insurance and beyond; promising future applications; ideas for product innovation; and ethical issues involving privacy and fairness.

**Moderator:** Ben Carrier, Managing Director, Aon Benfield
**Panelists:** Stephen Mildenhall, CEO of Analytics for Aon, Aon Benfield
Jim Guszcza, National Predictive Analytics Lead, Deloitte LLP
Examples of Applications

- Claim Frequency, Claim Severity
  - Use features of data to predict
  - Chapter in Predictive Modeling book
  - [www.casact.org](http://www.casact.org), “Intro to GLMs”
- Insurance Fraud
  - Derrig and Francis “Distinguishing the Forest from the Trees”, Variance, 2008
Examples of Applications (2)

- Reserving

- Financial Crisis
  - Could the defaulting mortgages have been predicted?
  - Francis and Prevosto, “Data and Disaster: The Role of Data in the Financial Crisis”
Validation of the Model

- Confusion Matrix
- ROC Curve
- Training/Test Data
- How good are the indications of accuracy?

### Classification Table

<table>
<thead>
<tr>
<th>Sample</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Training</td>
<td></td>
<td>608</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>278</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>63.9%</td>
<td>36.1%</td>
</tr>
<tr>
<td>Testing</td>
<td></td>
<td>271</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>63.2%</td>
<td>36.8%</td>
</tr>
</tbody>
</table>

Dependent Variable: Suspicion
Education

- How do we prepare insurance professionals to understand and participate in Predictive Modeling projects?
- Exam Syllabus
- Continuing Education
- iCAS CSPA credentials
- Many opportunities for self-study
Volumes 1 and 2, Book Project

- Predictive Modeling Applications in Actuarial Science Volume 1
  - The first volume contains an introduction to predictive modeling methods used by actuaries
  - It was published in 2014
- Predictive Modeling Applications in Actuarial Science Volume 2
  - The second volume is a collection of applications to P&C problems,
Roger Peng’s Books

- *The Art of Data Science*, by Roger Peng and Elizabeth Matsui
- Exploratory Data Analysis with R
Peng’s EDA with R

- On syllabus for Exam 2 (Data, EDA, Visualization)
- Pdf file version of books are available
- Provides a chapter on data management and data preprocessing
- Access to Youtube videos for each chapter
- Many worked out examples
Use of R for actuarial analytic applications is well established

Python is also growing in popularity

KDNuggets article “R vs Python for Data Science: and the Winner is” www.kdnuggets.com
R Libraries

- Code is provided with book
- The “cluster” library from R used
  - Many of the functions in the library are described in the Kaufman and Rousseeuw’s (1990) classic book on clustering, *Finding Groups in Data*.
- randomForest R library used to get dissimilarity matrix
- prcomp, princomp and factanal used for PRIDITs
- Some custom coding needed
Examples of Applications

- **Claim Severity**
  - Use features of data to predict
  - Chapter in Predictive Modeling book
  - [www.casact.org](http://www.casact.org), “Intro to GLMs”

- **Insurance Fraud**
  - Derrig and Francis “Distinguishing the Forest from the Trees”, Variance, 2008

- **Financial Crisis**
  - Could the defaulting mortgages have been predicted?
  - Francis and Prevosto, “Data and Disaster: The Role of Data in the Financial Crisis”
Case Study
Questionable Claims
The Fraud Problem

from: www.agentinsure.com

In Florida, Cops are Cracking Down on Car Insurance Fraud Rings

Recent raids by police in the Tampa Bay, Florida area are shedding light on a serious problem that’s plaguing the car insurance industry and having a negative impact on the ability of law abiding citizens to buy auto insurance at low prices.

According to statistics, incidents of staged car accidents in Tampa Bay are much higher than they are anywhere else. One of the reasons for this is said to be Florida’s existing PIP law, which stands for Personal Injury Protection, and guarantees as much as $10,000 in medical payments for every person injured in a car accident, no matter who’s to blame for it.
The Fraud Surrogates used as Dependent Variables

- Independent Medical Exam (IME) requested
- Special Investigation Unit (SIU) referral
- IME successful
- SIU successful
- DATA: Detailed Auto Injury Claim Database for Massachusetts
- Accident Years (1995-1997)
Fraud and Abuse

- Planned fraud
  - Staged accidents
- Abuse
  - Opportunistic
  - Exaggerate claim
- Both are referred to as questionable claims
Different Kinds of Decision Trees

- Single Trees (CART, CHAID)
- Ensemble Trees, a more recent development (TREENET, RANDOM FOREST)
  - A composite or weighted average of many trees (perhaps 100 or more)
  - There are many methods to fit the trees and prevent overfitting
    - Boosting: Iminer Ensemble and Treenet
    - Bagging: Random Forest
The Methods and Software Evaluated

1) TREENET  
2) Iminer Tree  
3) SPLUS Tree  
4) CART  
5) Iminer Ensemble  
6) Random Forest  
7) Naïve Bayes (Baseline)  
8) Logistic (Baseline)
The Fraud Surrogates used as Dependent Variables

- Independent Medical Exam (IME) requested
- Special Investigation Unit (SIU) referral
- IME successful
- SIU successful
- DATA: Detailed Auto Injury Claim Database for Massachusetts
- Accident Years (1995-1997)
Ensemble Prediction of IME Requested
## Results for IME Requested

<table>
<thead>
<tr>
<th></th>
<th>CART Tree</th>
<th>S-PLUS Tree</th>
<th>Iminer Tree</th>
<th>TREENET</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.669</td>
<td>0.688</td>
<td>0.629</td>
<td>0.701</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>0.661</td>
<td>0.680</td>
<td>0.620</td>
<td>0.693</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>0.678</td>
<td>0.696</td>
<td>0.637</td>
<td>0.708</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Iminer Ensemble</th>
<th>Random Forest</th>
<th>Iminer Naïve Bayes</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>0.649</td>
<td>703</td>
<td>0.676</td>
<td>0.677</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>0.641</td>
<td>695</td>
<td>0.669</td>
<td>0.669</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>0.657</td>
<td>711</td>
<td>0.684</td>
<td>0.685</td>
</tr>
</tbody>
</table>
Library for Getting Started

- Dahr, V, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall, 1997
- Derrig and Francis, “Distinguishing the Forest from the Trees”, Variance, 2008
- If you use R, get a book on doing analysis in R. See www.r-project.org
- Frees, Derrig and Francis, Predictive Modeling Applications in Actuarial Science, vol 1, Cambridge, 2014
- James, Witten, Hastie and Tibshirani, An Introduction to Statistical Learning with applications in R, Springer