Session 94 L, Detecting Fraudulent Claims in General Insurance

Moderator:
Anthony E. Cappelletti, FSA, FCAS, FCIA

Presenters:
Richard A. Derrig
Louise A. Francis, FCAS, MAAA
Using Predictive Analytics to Uncover Questionable Claims

Richard A. Derrig, PhD, CFE
President Opal Consulting, LLC
Visiting Professor of Risk, Insurance, and Healthcare Management
Temple University  Philadelphia PA

SOA Annual Meeting
October 28, 2014
Orlando, FL
Your Issues?

- Prosecuting Fraud?
- Detecting Fraud (which kind?)?
- Statutory Framework?
- Data Mining for Fraud?
- Investigators/Analysts?
- Medical Provider Fraud?
- Deterring Fraud (which kind?)?
- Catastrophe Fraud?
- World View of Fraud?
- Who Can Help?
Your Issues?

- Prosecuting Fraud?
  - Fraud Bureaus, Prosecutors (Funded), Felony, Runners, Sentencing (COIF)

- Detecting Fraud (which kind?)?
  - Expert Systems, Fraud Indicators, Investigations, Analysis

- Statutory Framework?
  - Reporting Immunity, Data Access, Investigation Option, Regulatory Oversight

- Data Mining for Fraud?
  - Requires Data First, Statewide Detail Data, Include Abuse, Technology

- Investigators/Analysts?
  - 2/3 police and 1/3 insurance for criminal, Analysts for connecting dots criminal and civil

- Medical Provider Fraud?
  - Detection by Billing Patterns, Strong Regulatory Response

- Deterring Fraud (which kind?)?
  - Criminal (sentencing, publicity); Civil (auditing by risk, investigation, profiling)

- Catastrophe Fraud?
  - Old damage, Clear rules, Public Adjusters, Material Shortages

- World View of Fraud?
  - Culturally dependent, Ethics, Increasing elsewhere, Interest in US

- Who Can Help?
  - OPAL, Coalition, IRC (auto), Fraud Bureaus, Studies, Ratepayers, Legislators
AGENDA

- What Exactly is Claim & Premium Fraud?
- What do Companies do about Fraud?
- How Can Fraud be Detected/Deterred?
- Who are the Fraudsters?
- How does “Data Mining” fit in?
- How to Find “Questionable Claims”? 
玲珑日语

- 硬欺诈，软欺诈，搭建
- 合理怀疑，压倒性证据
- 欺诈指标，模型特征
- 监督，无监督建模
- 预测模型，解释性模型
REAL PROBLEM-CLAIM CLASSIFICATION

- Classify all claims
- Identify valid classes
  - Pay the claim
  - No hassle
  - Visa Example
- Identify (possible) fraud
  - Investigation needed
- Identify “gray” classes
  - Minimize with “learning” algorithms
“Unfortunately, you have what we call ‘no insurance.’”
The treatment you need is extremely expensive & your policy doesn’t cover it. Accordingly, I’m going to prescribe an intensive program of insurance fraud.
Fraud Definition

PRINCIPLES

- Clear and willful act
- Proscribed by law
- Obtaining money or value
- Under false pretenses

Abuse: Fails one or more Principles
ABUSE DEFINITION

PRINCIPLES

- Not (Criminal) Fraud
- Unwanted, Unintended, Unnecessary Claims
- Disputable Damages
- Civil Matter
- Company’s Problem
- Regulator’s Problem
WELL YES, CERTAINLY... A HANGNAIL CAN BE A SERIOUS DISABILITY! IF YOU CAN MAKE IT "WORK RELATED", I DON'T SEE A PROBLEM!
Fraud Types

- Insurer Fraud
  - Fraudulent Company
  - Fraudulent Management
- Agent Fraud
  - No Policy
  - False Premium
- Company Fraud
  - Embezzlement
  - Inside/Outside Arrangements
- Claim Fraud
  - Claimant/Insured
  - Providers/Rings
What Companies Do About Fraud

- Segregate Suspicious/Questionable Claims
- Investigate Suspicious/Questionable Claims
- Cooperate with Fraud Bureaus, NICB
- Legislate by Lobbying for Insurance Fraud Laws
- Negotiate Lower Settlements Using Investigations
- Litigate by Civil Suit for Abusive Professionals
DERRIG TOP TEN IDEAS

1. “FRAUD” is ambiguous, ill-defined.
2. “FRAUD” should be reserved for criminal behavior.
3. “FRAUD” ambiguity inappropriately allows players to seek solutions in law enforcement rather than civil systems.
4. Criminal Fraud is several orders of magnitude less than popular estimates.
5. Fraud and Systematic Abuse can and should be mitigated by computer-assisted trained adjusters and special investigators.
6. The Questionable Claim Detection problem is of claim classification not specific claim identification.

7. Smart Claim Classification allows smart allocation of scarce loss adjusting and investigation expense.

8. Fraud Deterrence is optimal when positive audit ratios are constant across fraud and abuse risk classes.

9. When prosecuted, insurance fraud penalties are comparable to other fraud felony penalties.

10. About 50 percent of insurance fraud perpetrators have non-insurance criminal backgrounds.
HOW MUCH CLAIM FRAUD?

(CRIMINAL or CIVIL?)
10% Fraud
Massachusetts IFB 1991-2000
Mandatory Fraud Referrals Auto and WC Insurers

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait</td>
<td>450</td>
<td>13%</td>
</tr>
<tr>
<td>Assigned</td>
<td>222</td>
<td>7%</td>
</tr>
<tr>
<td>Closed</td>
<td>2,084</td>
<td>62%</td>
</tr>
<tr>
<td>Transferred</td>
<td>41</td>
<td>1%</td>
</tr>
<tr>
<td>At Prosecutor</td>
<td>139</td>
<td>4%</td>
</tr>
<tr>
<td>Declined</td>
<td>120</td>
<td>4%</td>
</tr>
<tr>
<td>Complete</td>
<td>293</td>
<td>9%</td>
</tr>
<tr>
<td>Total</td>
<td>3,349</td>
<td>100%</td>
</tr>
</tbody>
</table>
FRAUD AND ABUSE
THE TOP TEN DEFENSES

1. Adjusters
2. Computer Technology
3. Criminal Investigators
4. Data and Information
5. Experts
6. Judges
7. Lawyers
8. Legislators
9. Prosecutors
10. Special Investigators
Scientists from the RAND Corporation have created this model to illustrate how a "home computer" could look like in the year 2004. However, the needed technology will not be economically feasible for the average home. Also, the scientists readily admit that the computer will require not-yet-invented technology to actually work, but 50 years from now, scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.
FRAUD IDENTIFICATION

- Experience and Judgment
- Artificial Intelligence Systems
  - Regression & Tree Models
  - Fuzzy Clusters
  - Neural Networks
  - Expert Systems
  - Genetic Algorithms
  - All of the Above
DM
Databases

Scoring Functions

Graded Output

Non-Suspicious Claims
Routine Claims

Suspicious Claims
Complicated Claims
DM

Databases

Scoring Functions

Graded Output

Non-Suspicious Risks
Routine Underwriting

Suspicious Risks
Non-Routine Underwriting
POTENTIAL VALUE OF A PREDICTIVE MODELING SCORING SYSTEM

- Screening to Detect Fraud Early
- Auditing of Closed Claims to Measure Fraud, Both Kinds
- Sorting to Select Efficiently among Special Investigative Unit Referrals
- Providing Evidence to Support a Denial
- Protecting against Bad-Faith
PREDICTIVE MODELING
SOME PUBLIC TECHNIQUES

- Fuzzy Logic and Controllers
- Regression Scoring Systems
- Unsupervised Techniques: Kohonen and PRIDIT
- EM Algorithm (Medical Bills)
- Tree-based Methods
Fuzzy Clustering of Fraud Study Claims by Assessment Data
Membership Value Cut at 0.2
Supervised Models
Regression: Fraud Indicators

- Fraud Indicators Serve as Independent Dummy Variables
- Expert Evaluation Categories Serve as Dependent Target
- Regression Scoring Systems
- REF1: Weisberg-Derrig, 1998
- REF2: Viaene et al., 2002
AIB FRAUD INDICATORS

- **Accident Characteristics (19)**
  - No report by police officer at scene
  - No witnesses to accident

- **Claimant Characteristics (11)**
  - Retained an attorney very quickly
  - Had a history of previous claims

- **Insured Driver Characteristics (8)**
  - Had a history of previous claims
  - Gave address as hotel or P.O. Box
AIB FRAUD INDICATORS

- Injury Characteristics (12)
  - Injury consisted of strain/sprain only
  - No objective evidence of injury

- Treatment Characteristics (9)
  - Large number of visits to a chiropractor
  - DC provided 3 or more modalities on most visits

- Lost Wages Characteristics (6)
  - Claimant worked for self or family member
  - Employer wage differs from claimed wage loss
This is a More General Issue of Supervised Learning vs. Unsupervised Learning Methodologies

- Different techniques are needed when one cannot “teach” or “train” a technique to correctly classify object by reference to some known set of classified objects (supervised learning).

- Parameter fitting of models in regression, etc. assume you have data on the dependent variable with which to fit the unknown parameters.
DATA MODELING EXAMPLE: CLUSTERING

- Data on 16,000 Medicaid providers analyzed by unsupervised neural net
- Neural network clustered Medicaid providers based on 100+ features
- Investigators validated a small set of known fraudulent providers
- Visualization tool displays clustering, showing known fraud and abuse
- Subset of 100 providers with similar patterns investigated: Hit rate > 70%

Cube size proportional to annual Medicaid revenues

© 1999 Intelligent Technologies Corporation
Wrongful Death

“Good morning George, I didn’t expect to see you up for breakfast.”
Specifics of Motivation to Fraud Application

- Insurance investigators, adjusters, and insurance claim managers are faced with incomplete information for decision-making concerning the validity or possible fraudulent status of a particular filed claim.

- Strategic and tactical decisions must be made:
  - whether or not to pay the claim,
  - refer the file to a special Investigative Unit (SIU),
  - refer the case to the Fraud Bureau, or Attorney General's Office.
Claim Fraud Detection Plan

- **STEP 1:** **SAMPLE:** Systematic benchmark of a random sample of claims.
- **STEP 2:** **FEATURES:** Isolate red flags and other sorting characteristics.
- **STEP 3:** **FEATURE SELECTION:** Separate features into objective and subjective, early, middle and late arriving, acquisition cost levels, and other practical considerations.
- **STEP 4:** **CLUSTER:** Apply unsupervised algorithms (Kohonen, PRIDIT, Fuzzy) to cluster claims, examine for needed homogeneity.
Claim Fraud Detection Plan

- **STEP 5: ASSESSMENT:** Externally classify claims according to objectives for sorting.
- **STEP 6: MODEL:** Supervised models relating selected features to objectives (logistic regression, Naïve Bayes, Neural Networks, CART, MARS)
- **STEP 7: STATIC TESTING:** Model output versus expert assessment, model output versus cluster homogeneity (PRIDIT scores) on one or more samples.
- **STEP 8: DYNAMIC TESTING:** Real time operation of acceptable model, record outcomes, repeat steps 1-7 as needed to fine tune model and parameters. Use PRIDIT to show gain or loss of feature power and changing data patterns, tune investigative proportions to optimize detection and deterrence of fraud and abuse.
<table>
<thead>
<tr>
<th>PRIDIT</th>
<th>Adj. Reg. Score</th>
<th>Inv. Reg. Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC3</td>
<td>ACC1</td>
<td>ACC11</td>
</tr>
<tr>
<td>ACC4</td>
<td>ACC9</td>
<td>CLT4</td>
</tr>
<tr>
<td>ACC15</td>
<td>ACC10</td>
<td>CLT7</td>
</tr>
<tr>
<td>CLT11</td>
<td>ACC19</td>
<td>CLT11</td>
</tr>
<tr>
<td>INJ1</td>
<td>CLT11</td>
<td>INJ1</td>
</tr>
<tr>
<td>INJ2</td>
<td>INS6</td>
<td>INJ3</td>
</tr>
<tr>
<td>INJ5</td>
<td>INJ2</td>
<td>INJ8</td>
</tr>
<tr>
<td>INJ6</td>
<td>INJ9</td>
<td>INJ11</td>
</tr>
<tr>
<td>INS8</td>
<td>TRT1</td>
<td>TRT1</td>
</tr>
<tr>
<td>TRT1</td>
<td>LW6</td>
<td>TRT9</td>
</tr>
</tbody>
</table>
Oppportunistic Fraud

“...I TOLD THE POLICE I WAS NOT INJURED, BUT ON REMOVING MY HAT I FOUND I HAD A FRACTURED SKULL...”
The Problem: Nonlinear Functions
## Nonlinear Example Data

<table>
<thead>
<tr>
<th>Provider 2 Bill (Bonded)</th>
<th>Avg Provider 2 Bill</th>
<th>Avg Total Paid</th>
<th>Percent IME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>-</td>
<td>9,063</td>
<td>6%</td>
</tr>
<tr>
<td>1 – 250</td>
<td>154</td>
<td>8,761</td>
<td>8%</td>
</tr>
<tr>
<td>251 – 500</td>
<td>375</td>
<td>9,726</td>
<td>9%</td>
</tr>
<tr>
<td>501 – 1,000</td>
<td>731</td>
<td>11,469</td>
<td>10%</td>
</tr>
<tr>
<td>1,001 – 1,500</td>
<td>1,243</td>
<td>14,998</td>
<td>13%</td>
</tr>
<tr>
<td>1,501 – 2,500</td>
<td>1,915</td>
<td>17,289</td>
<td>14%</td>
</tr>
<tr>
<td>2,501 – 5,000</td>
<td>3,300</td>
<td>23,994</td>
<td>15%</td>
</tr>
<tr>
<td>5,001 – 10,000</td>
<td>6,720</td>
<td>47,728</td>
<td>15%</td>
</tr>
<tr>
<td>10,001 +</td>
<td>21,350</td>
<td>83,261</td>
<td>15%</td>
</tr>
<tr>
<td>All Claims</td>
<td>545</td>
<td>11,224</td>
<td>8%</td>
</tr>
</tbody>
</table>
Desirable Features of a Data Mining Method:

- Any nonlinear relationship can be approximated
- A method that works when the form of the nonlinearity is unknown
- The effect of interactions can be easily determined and incorporated into the model
- The method generalizes well on out-of-sample data
REFERENCES

Predictive Modeling for Fraud (Questionable Claims)

Society of Actuaries Annual Meeting
October 2014
Louise Francis, FCAS, MAAA
www.data-mines.com
Topics

- Fraud vs Questionable Claims
- Predictive modeling for Questionable Claims
- Supervised Learning
- Unsupervised Learning
  - Most emphasis on unsupervised learning
  - Chapter on Unsupervised Learning from Predictive Modeling Book
  - Some Methods
Unsupervised Learning Methods

- PRIDIT
- Clustering
- Random Forest
- Kohonen neural networks
Major Kinds of Modeling

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

- **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
    - Association rules
    - K-means clustering
    - Kohonen neural networks
Book Project

- Predictive Modeling 2 Volume Book Project
- A joint project leading to a two volume pair of books on Predictive Modeling in Actuarial Science.
- Volume 1 would be on Theory and Methods and
- Volume 2 would be on Property and Casualty Applications.
- The first volume will be introductory with basic concepts and a wide range of techniques designed to acquaint actuaries with this sector of problem solving techniques. The second volume would be a collection of applications to P&C problems, written by authors who are well aware of the advantages and disadvantages of the first volume techniques but who can explore relevant applications in detail with positive results.
Focus on Using R for Applications

```r
# Code to calculate RIDITs and PRIDITs on Questionable Claims Data
# Read in questionable claims data
# this version of data has dependent var and is 1000 lines
mydata1<-read.csv("C:/ClusterData/Sim_PIPPrdData.csv",header=TRUE)
# this version of data has dependent var and is 1500 lines
mydata2<-read.csv("C:/ClusterData/SimPIP.csv",header=TRUE)
names(mydata1)
nrow(mydata1)
ncol(mydata1)
table(Suspicion,legalrep)
mydata=mydata1[,3:27]
mydata[1:5,]

> rpart(Suspicion~Comp.1+Comp.2)
```
The Fraud Study Data

- 1993 AIB closed PIP claims
- Dependent Variables
  - Suspicion Score
  - Expert assessment of likelihood of fraud or abuse
- Predictor Variables
  - Red flag indicators
  - Claim file variables
- Simulated data modeled on 1993 study used for book chapters and will be on book website
The Fraud Red Flags

- Binary variables that capture characteristics of claims associated with fraud and abuse
- Accident variables (acc01 - acc19)
- Injury variables (inj01 – inj12)
- Claimant variables (ch01 – ch11)
- Insured variables (ins01 – ins06)
- Treatment variables (trt01 – trt09)
- Lost wages variables (lw01 – lw07)
## The Red Flag Variables

<table>
<thead>
<tr>
<th>Subject</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident</td>
<td>ACCO1</td>
<td>No report by police officer at scene</td>
</tr>
<tr>
<td></td>
<td>A0004</td>
<td>Single vehicle accident</td>
</tr>
<tr>
<td></td>
<td>A0009</td>
<td>No plausible explanation for accident</td>
</tr>
<tr>
<td></td>
<td>ACC10</td>
<td>Claimant in old, low valued vehicle</td>
</tr>
<tr>
<td></td>
<td>ACC11</td>
<td>Rental vehicle involved in accident</td>
</tr>
<tr>
<td></td>
<td>ACC14</td>
<td>Property Damage was inconsistent with accident</td>
</tr>
<tr>
<td></td>
<td>ACC15</td>
<td>Very minor impact collision</td>
</tr>
<tr>
<td></td>
<td>ACC16</td>
<td>Claimant vehicle stopped short</td>
</tr>
<tr>
<td></td>
<td>ACC19</td>
<td>Insured felt set up, denied fault</td>
</tr>
<tr>
<td>Claimant</td>
<td>CLT02</td>
<td>Had a history of previous claims</td>
</tr>
<tr>
<td></td>
<td>CLT04</td>
<td>Was an out of state accident</td>
</tr>
<tr>
<td></td>
<td>CLT07</td>
<td>Was one of three or more claimants in vehicle</td>
</tr>
<tr>
<td>Injury</td>
<td>INJ01</td>
<td>Injury consisted of strain or sprain only</td>
</tr>
<tr>
<td></td>
<td>INJ02</td>
<td>No objective evidence of injury</td>
</tr>
<tr>
<td></td>
<td>INJ03</td>
<td>Police report showed no injury or pain</td>
</tr>
<tr>
<td></td>
<td>INJ05</td>
<td>No emergency treatment was given</td>
</tr>
<tr>
<td></td>
<td>INJ06</td>
<td>Non-emergency treatment was delayed</td>
</tr>
<tr>
<td></td>
<td>INJ11</td>
<td>Unusual injury for auto accident</td>
</tr>
<tr>
<td>Insured</td>
<td>INSO1</td>
<td>Had history of previous claims</td>
</tr>
<tr>
<td></td>
<td>INSO3</td>
<td>Readily accepted fault for accident</td>
</tr>
<tr>
<td></td>
<td>INSO6</td>
<td>Was difficult to contact/uncooperative</td>
</tr>
<tr>
<td></td>
<td>INSO7</td>
<td>Accident occurred soon after effective date</td>
</tr>
<tr>
<td>Lost Wages</td>
<td>LWO1</td>
<td>Claimant worked for self or a family member</td>
</tr>
<tr>
<td></td>
<td>LW03</td>
<td>Claimant recently started employment</td>
</tr>
</tbody>
</table>
Dependent Variable Problem

- Insurance companies frequently do not collect information as to whether a claim is suspected of fraud or abuse.
- Even when claims are referred for special investigation.
- Solution: unsupervised learning.
R Libraries/Functions Needed

- reshape library
- Cluster library
- princomp, prcomp
- kohonen library
- cluster library
R Cluster Library

- 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.
Dissimilarity

- Euclidian Distance: the record by record squared difference between the value of each the variables for a record and the values for the record it is being compared to.

\[ d_{ij} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \]
RF Similarity

- Varies between 0 and 1
- Proximity matrix is an output of RF
  - After a tree is fit, all records run through model
  - If 2 records in same terminal node, their proximity increased by 1
  - 1-proximity forms distance
- Can be used as an input to clustering and other unsupervised learning procedures
- See “Unsupervised Learning with Random Forest Predictors” by Shi and Horvath
Clustering

- Hierarchical clustering
- K-Means clustering
- This analysis uses k-means
Hierarchical Clustering

```
Indicates(x = AutoBIVars, metric = "manhattan", stand = TRUE,

AutoBIVars
Divisive Coefficient = 0.98
```
K-means Clustering

- An iterative procedure is used to assign each record in the data to one of the k clusters.
- The iteration begins with the initial centers or mediods for k groups.
- Uses a dissimilarity measure to assign records to a group and to iterate to a final grouping. An iterative procedure is used to assign each record to one of the k clusters by the user.
R Cluster Output

```r
> BICluster1
Call:  clara(x = ClusterDat1, k = 2, metric = "euclidean")
Medoids:
  BIFrequency BISeverity
[1,] 11.39769  8202.802
[2,] 13.28089 10749.593
Objective function:  577.0351
Clustering vector:  [1:100] 1 1 2 1 1 1 1 1 1 1 1 2 1
Cluster sizes:  63 37
```
Cluster Plot

plot(clara(x = ClusterDat1, k = 2, metric = "manhattan", stand = TRUE))
clusplot(keep.data = TRUE))

Component 2

Component 1

These two components explain 100% of the point variability.
Random Forest

- A Tree based data mining technique
- An ensemble method: weighted average of many single models
- Can be run in “unsupervised mode”
  - Create measure of similarity between records
  - Use measure to create dissimilarity measure
  - Cluster with dissimilarity

\[ d_y = \sqrt{1 - p_y}, d_y = \text{dissimilarity}, p_y = \text{proximity} \]
Testing using Suspicion Indicator: Fit a Tree and Use for Importance Ranking

Clusters Rank Statistic

- Group 1: ClusterslabelRF2 = 1, ClusterslabelRF4 = 1,3,4
  - 1.1: n=675, 45%
  - 1.2: n=1147, 76%
  - 1.3: n=1500, 100%

- Group 2: ClusterlabelRF2 = 2
  - 1.8: n=353, 24%

- Group 3: ClusterlabelRF2 = 3
  - 1.3: n=472, 31%

# Clusters | Rank | Statistic
--- | --- | ---
4 Group | 1 | 115.0
2 Group | 2 | 5.3
3 Group | 3 | 3.7

Francis Analytics and Actuarial Data Mining, Inc.

10/13/2014
## Importance Ranking of the Clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Rank</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF 4 Group</td>
<td>1</td>
<td>115.0</td>
</tr>
<tr>
<td>RF 2 Group</td>
<td>2</td>
<td>5.3</td>
</tr>
<tr>
<td>RF 3 Group</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>Euclid 4 Group</td>
<td>4</td>
<td>2.9</td>
</tr>
<tr>
<td>Euclid 3 Group</td>
<td>5</td>
<td>2.7</td>
</tr>
</tbody>
</table>
RF Ranking of the “Predictors”: Top 10 of 44

<table>
<thead>
<tr>
<th>Variable</th>
<th>MeanDecreaseGini</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc10</td>
<td>10.50</td>
<td>Claimant in old low value vehicle</td>
</tr>
<tr>
<td>trt01</td>
<td>9.05</td>
<td>large # visits to chiro</td>
</tr>
<tr>
<td>inj01</td>
<td>8.64</td>
<td>strain or sprain</td>
</tr>
<tr>
<td>inj02</td>
<td>8.64</td>
<td>readily accepted fault</td>
</tr>
<tr>
<td>inj05</td>
<td>8.62</td>
<td>non emergency treatment given for injury</td>
</tr>
<tr>
<td>acc01</td>
<td>8.55</td>
<td>no police report</td>
</tr>
<tr>
<td>clt07</td>
<td>7.47</td>
<td>one of 3 or more claimants in vehicle</td>
</tr>
<tr>
<td>inj06</td>
<td>7.44</td>
<td>non emergency trt delayed</td>
</tr>
<tr>
<td>acc15</td>
<td>7.36</td>
<td>very minor collision</td>
</tr>
<tr>
<td>trt03</td>
<td>6.82</td>
<td>large # visits to PT</td>
</tr>
</tbody>
</table>
RIDIT

- Boss, “How to Use RIDIT Analysis”, Biometrics, 1958
- Addresses “borderland variables”
- Some variables a subjective scale
- May not be suitable for Chi-square or t-test analysis
- Distribution free
- Boss’s RIDIT = cum proportion less than level plus $\frac{1}{2}$ proportion in level
RIDIT

- Theory: variables are ordered so that lowest value is associated with highest probability of suspicion of questionable claim
- Use Cumulative distribution of claims at each value, i, to create RIDIT statistic for claim t, value i

\[ \text{RIDIT} (X_i) = P(X < X_{i-1}) + \frac{1}{2} P(X = X_i) \]
**Example: RIDIT for Legal Representation**

<table>
<thead>
<tr>
<th>Injury Severity</th>
<th>Count</th>
<th>Cumulative Count</th>
<th>Probability</th>
<th>Cumulative Probability</th>
<th>RIDIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>300</td>
<td>300</td>
<td>0.3</td>
<td>0.3</td>
<td>0.15</td>
</tr>
<tr>
<td>Medium</td>
<td>600</td>
<td>900</td>
<td>0.6</td>
<td>0.9</td>
<td>0.6</td>
</tr>
<tr>
<td>High</td>
<td>100</td>
<td>1000</td>
<td>0.1</td>
<td>1</td>
<td>0.95</td>
</tr>
</tbody>
</table>
PRIDIT

- PRIDIT = Principal Component of RIDITs
- Use RIDIT statistics in Principal Components Analysis
- The PRIDIT is often defined as the first component
### 1st 10 PRIDIT Loadings

<table>
<thead>
<tr>
<th>Variable</th>
<th>Component 1 Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>sprain</td>
<td>0.342</td>
</tr>
<tr>
<td>Inj01 (strain/sprain only)</td>
<td>(0.323)</td>
</tr>
<tr>
<td>Inj02 (No evidence inj)</td>
<td>(0.316)</td>
</tr>
<tr>
<td>legalrep</td>
<td>0.303</td>
</tr>
<tr>
<td>NumProv</td>
<td>(0.301)</td>
</tr>
<tr>
<td>NumTreat</td>
<td>(0.291)</td>
</tr>
<tr>
<td>chiro/pt</td>
<td>0.290</td>
</tr>
<tr>
<td>ambulcost</td>
<td>(0.219)</td>
</tr>
<tr>
<td>emtreat</td>
<td>0.212</td>
</tr>
</tbody>
</table>

**Absolute Load**

![Bar chart showing absolute loadings for different variables](chart.png)
Fit Tree with PRIDITs
ROC Curve using PRIDIT to predict Suspicion
Relation Between PRIDIT Factor and Suspicion

![Graph showing the relation between PRIDIT Factor and Mean Suspicion across percentile groups. The graph indicates a decreasing trend in suspicion as the PRIDIT Factor increases.](image-url)
Add RF and Euclid Clusters to PRIDIT Factors and Rank in Importance

Normalized Importance

- 4 Group RF Cluster
- PridtScore
- 3 Group RF Cluster
- Euclid4.Group
- Euclid3.Group
- 2 Group RF Cluster
- Euclid2.Group

Importance

Growing Method: CRT
Dependent Variable: Suspicion
Labeling Clusters

<table>
<thead>
<tr>
<th>RF Four Group Cluster</th>
<th>Sprain</th>
<th>Chiro/pt</th>
<th>NumProv</th>
<th>NumTreat</th>
<th>TrtLag</th>
<th>Thresh</th>
<th>ambulcos t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.83</td>
<td>.91</td>
<td>2.11</td>
<td>2.92</td>
<td>1.79</td>
<td>.84</td>
<td>62.81</td>
</tr>
<tr>
<td>2</td>
<td>.07</td>
<td>.31</td>
<td>3.32</td>
<td>4.52</td>
<td>1.61</td>
<td>.81</td>
<td>270.54</td>
</tr>
<tr>
<td>3</td>
<td>.88</td>
<td>.93</td>
<td>2.62</td>
<td>3.62</td>
<td>14.95</td>
<td>.31</td>
<td>203.58</td>
</tr>
<tr>
<td>4</td>
<td>.09</td>
<td>.30</td>
<td>5.98</td>
<td>8.72</td>
<td>4.55</td>
<td>.32</td>
<td>273.78</td>
</tr>
<tr>
<td>Total</td>
<td>.50</td>
<td>.64</td>
<td>3.29</td>
<td>4.62</td>
<td>5.06</td>
<td>.62</td>
<td>187.64</td>
</tr>
</tbody>
</table>
### Original Data Importance

**Ranking of Some PRIDITs**

#### Independent Variable Importance

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Importance</th>
<th>Normalized Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt factor score 1 for analysis 1</td>
<td>.139</td>
<td>100.0%</td>
</tr>
<tr>
<td>Clt factor score 1 for analysis 1</td>
<td>.087</td>
<td>63.0%</td>
</tr>
<tr>
<td>Ins factor score 1 for analysis 1</td>
<td>.048</td>
<td>34.4%</td>
</tr>
<tr>
<td>Inj factor score 1 for analysis 2</td>
<td>.024</td>
<td>17.2%</td>
</tr>
<tr>
<td>LW factor score 1 for analysis 1</td>
<td>.021</td>
<td>15.5%</td>
</tr>
<tr>
<td>Acc factor score 1 for analysis 2</td>
<td>.012</td>
<td>8.7%</td>
</tr>
</tbody>
</table>

Growing Method: CRT  
Dependent Variable: susp7
R Self Organized Feature Map Libraries

- kohonen
- som
- somplot
Plots from kohonen Package
The Data

- 1993 AIB fraud data
- 36 fraud indicators
- Indicator whether legitimate/suspicious
Nonlinear Unsupervised Method

- Data organized (mapped) to 2 dimensional grid of nodes
- Influence of Nearby Nodes

\[ i(t) = \exp\left(\frac{-\text{dist}^2}{2 \times s(t)}\right) \]
Map of Codes

PIP Claims Data Feature Map
Plot of Counts

PIP Claims Data Feature Map Counts

[Color-coded map with various shades indicating data counts]
Quality of Node

PIP Claims Data Feature Map Quality
Distance to neighbors

PIP Claims Data Feature Map Distance

[Diagram showing a grid with colors indicating distance to neighbors]
Amount of Legal Representation

PIP Claims Data Feature Map Legal
Some Conclusions

- Both RandomForest Clustering and PRIDITS show promise in unsupervised learning applications.
- Have potential to be very useful when dependent variable is missing from data, as in many fraud (questionable claims) applications.
- Data and code will be provided with book for testing methods.
- Kohonen neural networks can be used to visualize data in exploratory data analysis.