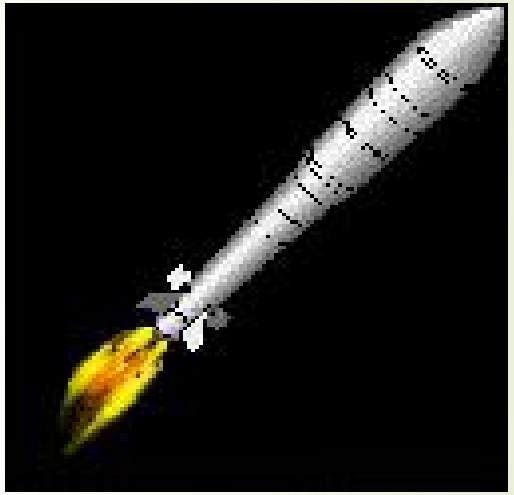
Overview of Predictive Modeling for Actuaries

Michigan Actuarial Society
March 18, 2015

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 TAVERNISE

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

Kinds of Applications

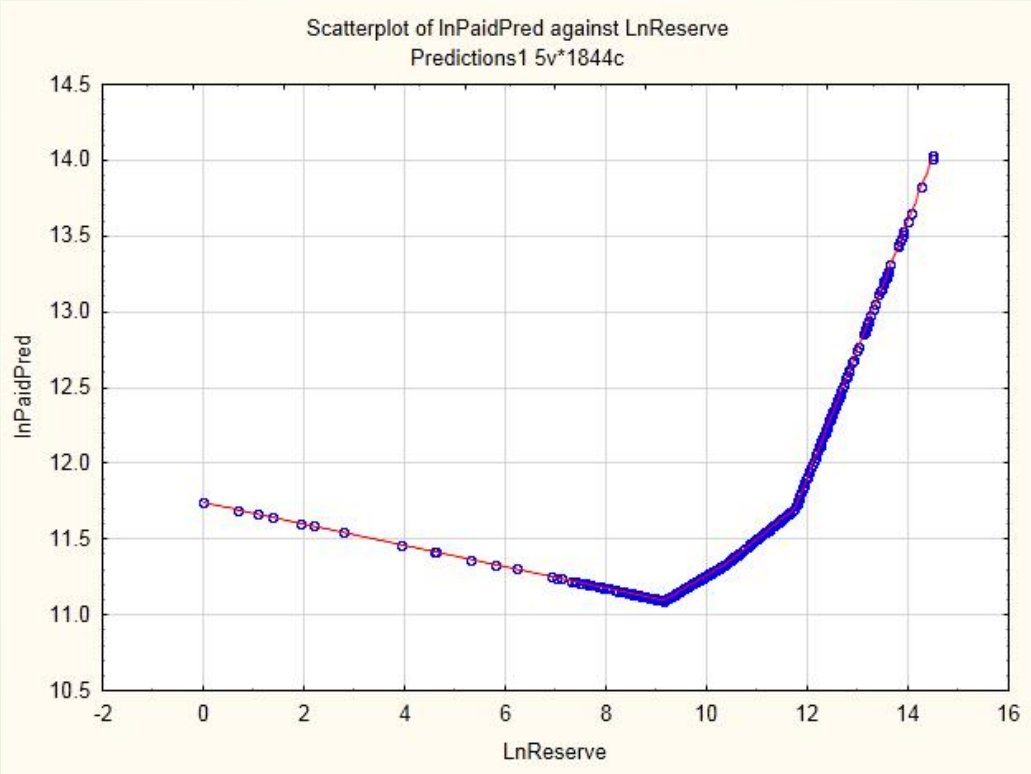
- Classification
  - Target variable is categorical
- Prediction
  - Target variable is numeric

A Casualty Actuary's Perspective on Data Modeling

- The Stone Age: 1914 – …
  - Simple deterministic methods
    - Use of blunt instruments: the analytical analog of bows and arrows
  - Often ad-hoc
  - Slice and dice data
  - Based on empirical data – little use of parametric models
- The Pre – Industrial age: 1970 - …
  - Fit probability distribution to model tails
  - Simulation models and numerical methods for variability and uncertainty analysis
  - Focus is on underwriting, not claims
- The Industrial Age – 1985 …
  - Begin to use computer catastrophe models
- The 20th Century – 1990…
  - European actuaries begin to use GLMs
- The Computer Age 1996…
  - Begin to discuss data mining at conferences
  - At end of 20st century, large consulting firms starts to build a data mining practice
- The Current era – A mixture of above
  - In personal lines, modeling the rule rather than the exception
    - Often GLM based, though GLMs evolving to GAMs
  - Commercial lines beginning to embrace modeling

Data Complexities: Nonlinearities

MARS Prediction of Primary Paid Severity



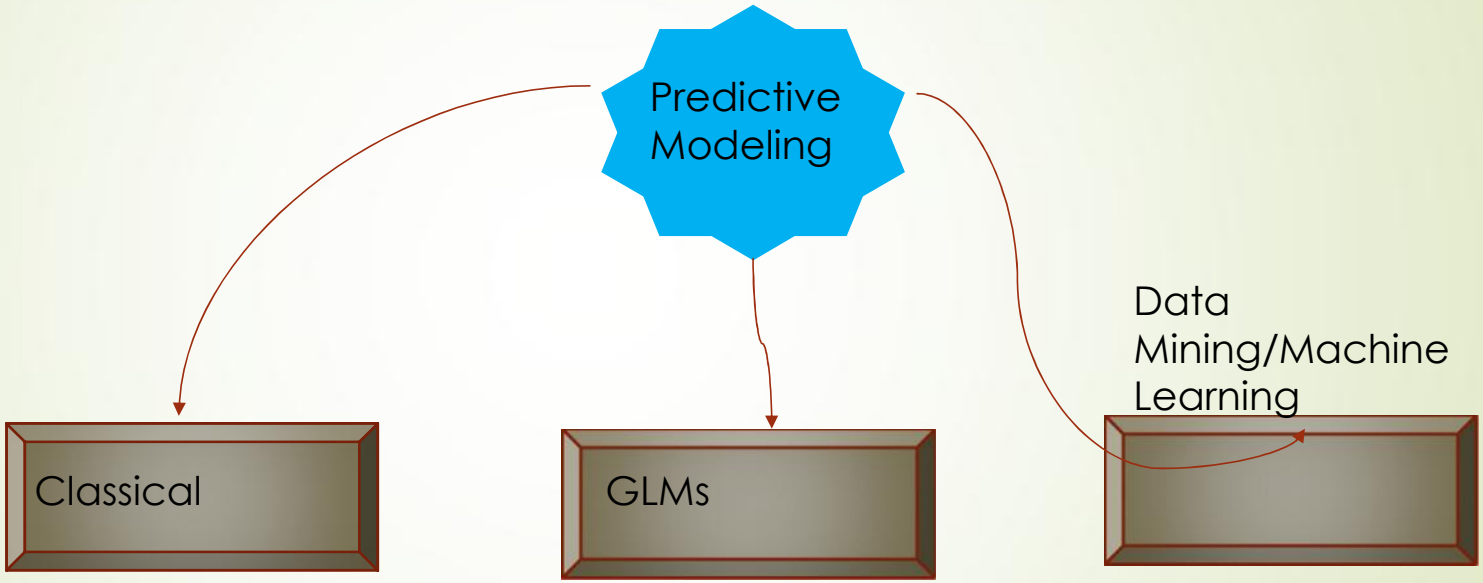
Major Kinds of Data Mining

- **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
- Newer Methods
  - Ensemble
  - SVM
  - Deep learning
  - Text Mining

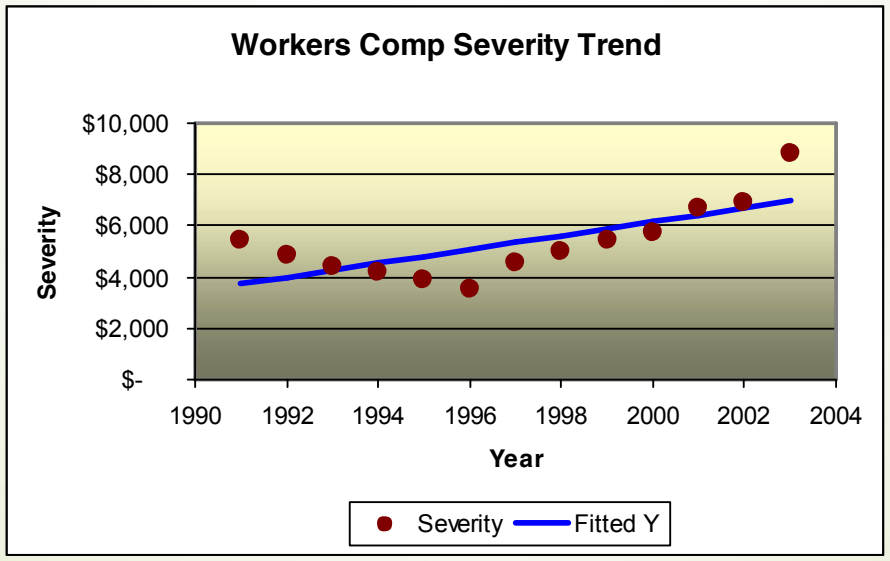
Predictive Modeling



Classical Statistics: Regression

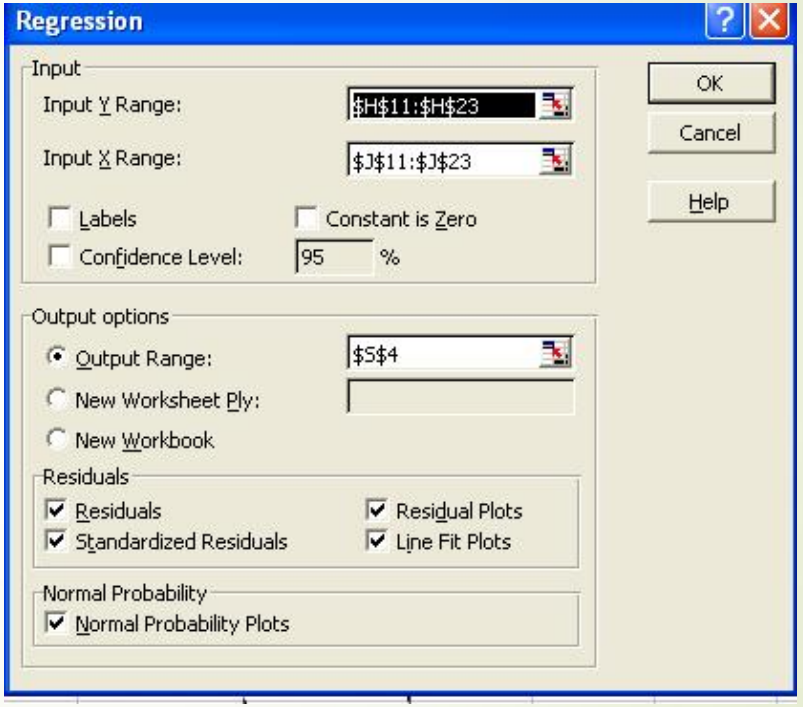
- Estimation of parameters: Fit line that minimizes deviation between actual and fitted values

$$\min(\sum (Y_i - \hat{Y})^2)$$



Linear Modeling Tools Widely Available: Excel Analysis Toolpak

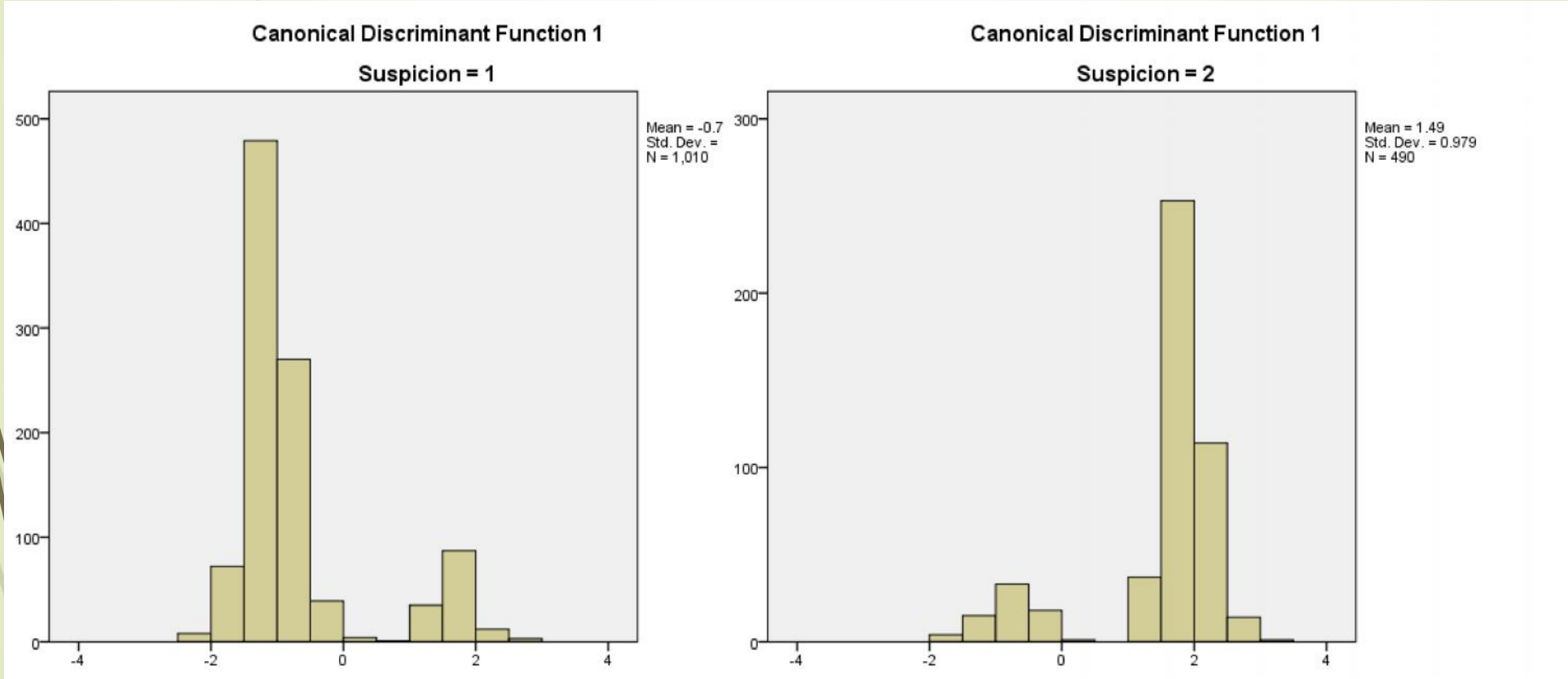
- Install Data Analysis Tool Pak (Add In) that comes with Excel
- Click Tools, Data Analysis, Regression



Goodness of Fit

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.303				
R Square	0.092				
Adjusted R Square	0.091				
Standard Error	1.206				
Observations	1818				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	266.29	266.29	183.07	0.00
Residual	1816	2641.50	1.45		
Total	1817	2907.79			

Classical Model: Discriminant Analysis



Generalized Linear Models (GLMs)

- Relax normality assumption
  - Exponential family of distributions
- Models some kinds of nonlinearity

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

Linear Model vs GLM

- Regression:

$$Y_i = \mu_i + \varepsilon$$
$$\mu_i = X'\mathrm{B}$$
$$\varepsilon \sim N(0, \sigma^2)$$

- GLM:

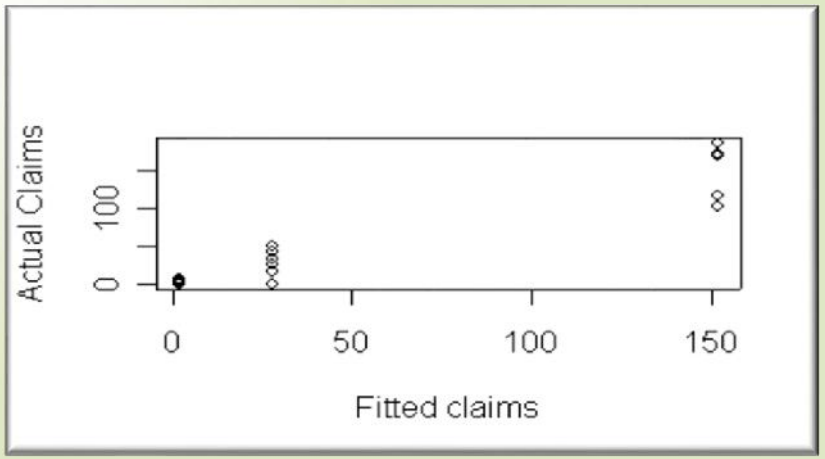
$$Y = h(\mu_i) + \varepsilon$$
$$h(\mu) = X'\mathrm{B}$$
$$\varepsilon \sim \text{exponential family}$$
$$\text{h is a link function}$$

Estimating Parameters

- As with nonlinear regression, there usually is not a closed form solution for GLMs
- A numerical method used to solve for parameters
- For some models this could be programmed in Excel – but statistical software is the usual choice
- If you can't spend money on the software, download R for free

GLM fit for Poisson Regression

```
>devage<-as.factor((AGE)
>claims.glm<-glm(Claims~devage, family=poisson)
>summary(claims.glm)
Call:
glm(formula = Claims ~ devage, family = poisson)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-10.250  -1.732  -0.500   0.507  10.626
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.73540    0.02825 167.622  < 2e-16 ***
devage2     -0.89595    0.05430 -16.500  < 2e-16 ***
devage3     -4.32994    0.29004 -14.929  < 2e-16 ***
devage4     -6.81484    1.00020  -6.813 9.53e-12 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 2838.65  on 36  degrees of freedom
Residual deviance:  708.72  on 33  degrees of freedom
AIC: 851.38
```

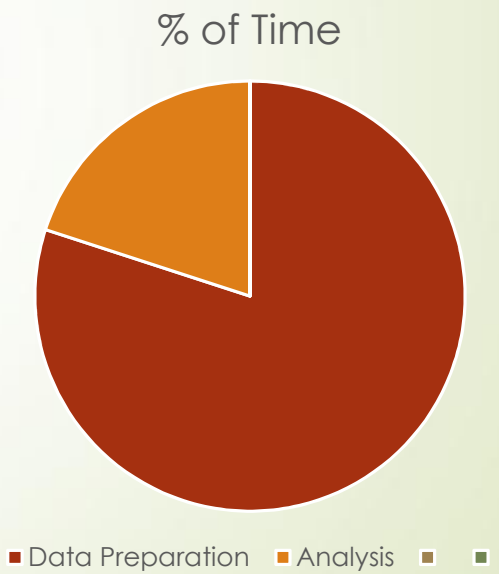


Data Complexities: Missing Data

- 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

Data

- Data Management
- Data quality
  - Francis, "Dancing With Dirty Data", CAS forum, www.casact.org
  - CAS Working Party, "Actuarial IQ", www.casact.org
- Big Data



Examples of Applications

- Claim Frequency, Claim Severity
  - Use features of data to predict
  - Chapter in Predictive Modeling book
  - www.casact.org, "Intro to GLMs"
- Liklihood a claim will occur (life insurance)
- 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"

The Questionable Claims Study Data

- 1993 AIB closed PIP claims
- Simulated data based on research performed on original data
- Dependent Variables
  - Suspicion Score
  - Expert assessment of likelihood of fraud or abuse
- Predictor Variables
  - Red flag indicators
  - Claim file variables

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 Fraud Problem

from: www.agentinsure.com

<< RETURN TO ARTICLES LIST

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.

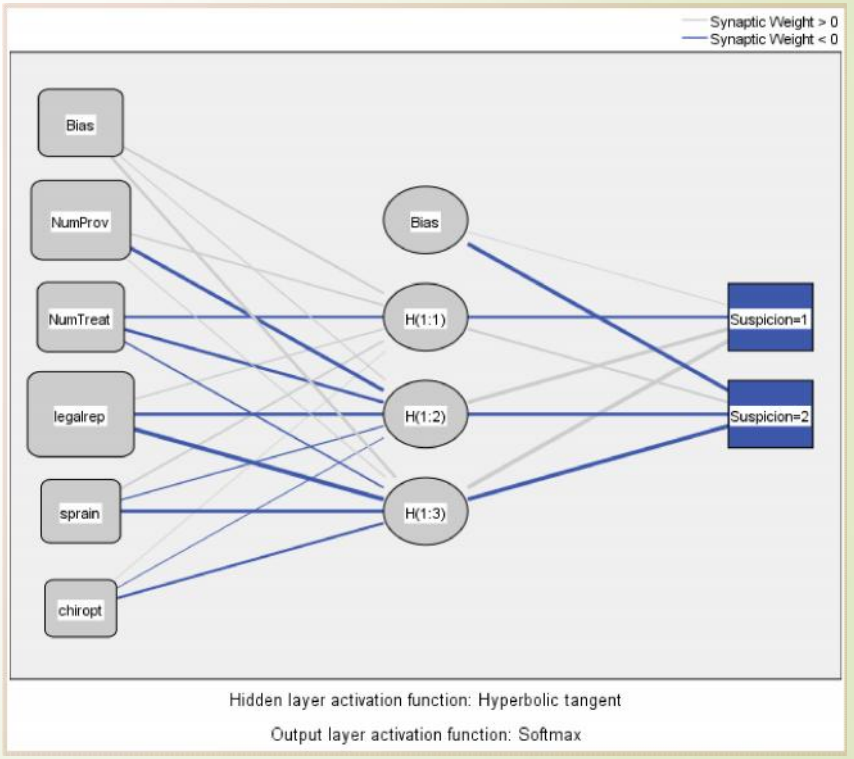
Fraud and Abuse

- Planned fraud
  - Staged accidents
- Abuse
  - Opportunistic
  - Exaggerate claim
- Both are referred to as "questionable claims"

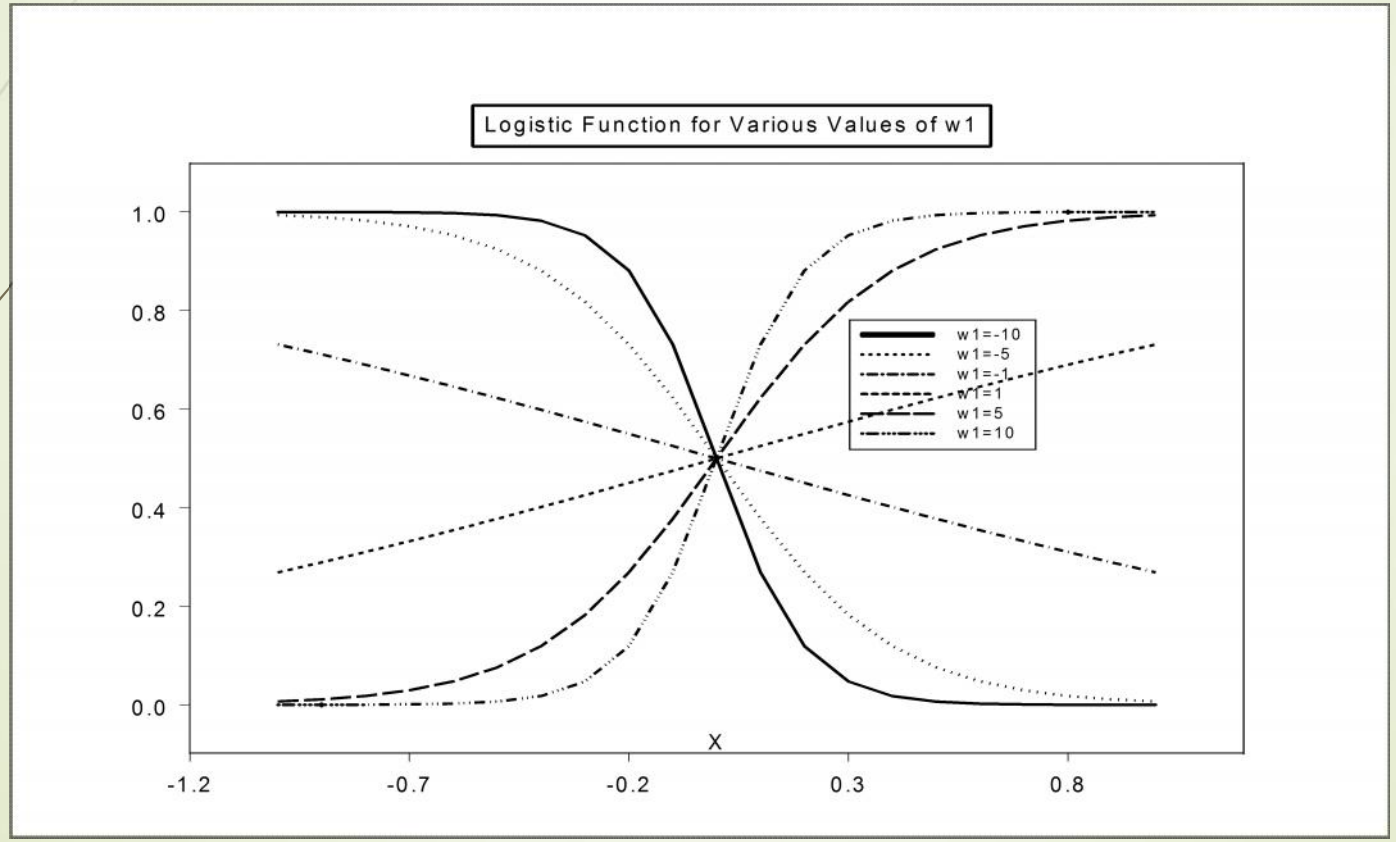
Neural Networks

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

- See Francis, "Neural Networks Demystified", CAS Forum, 2001, www.casact.org



Hidden Layer of Neural Network (Input Transfer Function)



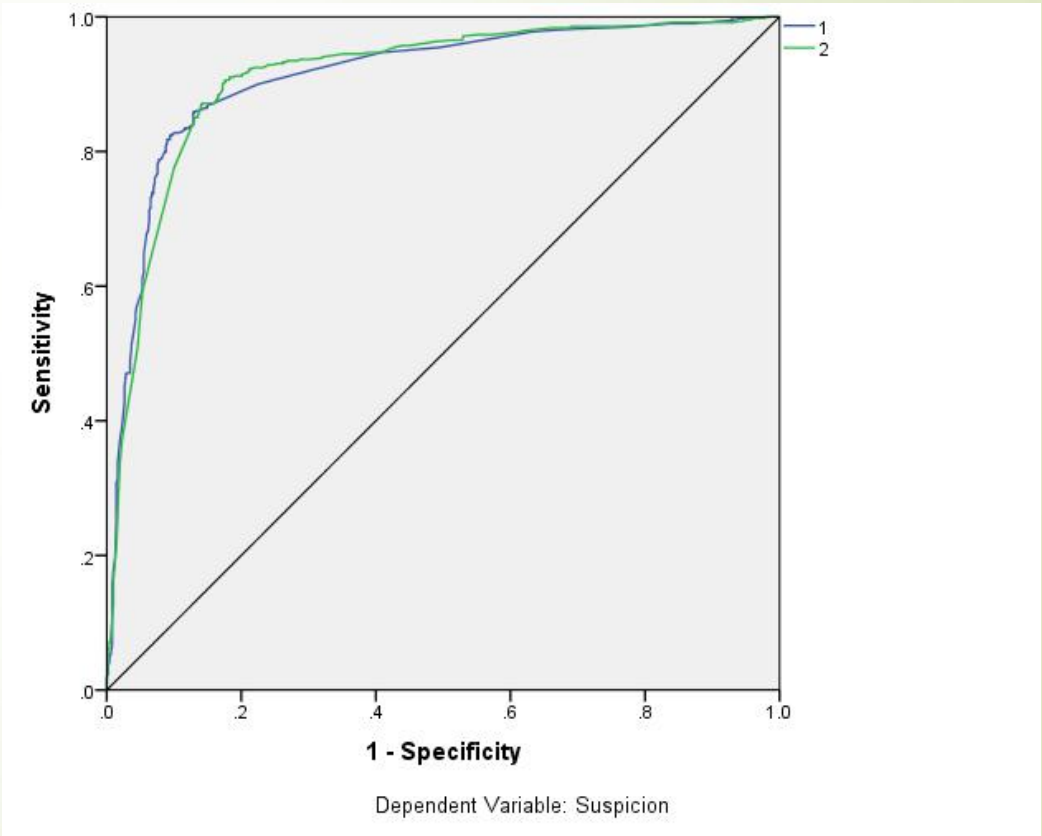
Assessing Results

- Confusion Matrix
- ROC Curve

Classification

Sample	Observed	Predicted 1	Predicted 2	Percent Correct
Training	1	608	95	86.5%
	2	52	278	84.2%
	Overall Percent	63.9%	36.1%	85.8%
Testing	1	271	36	88.3%
	2	24	136	85.0%
	Overall Percent	63.2%	36.8%	87.2%

Dependent Variable: Suspicion



Regression Trees

- Tree-based modeling for ***continuous* target variable**
  - most intuitively appropriate method for loss ratio analysis
- Find split that produces greatest separation in

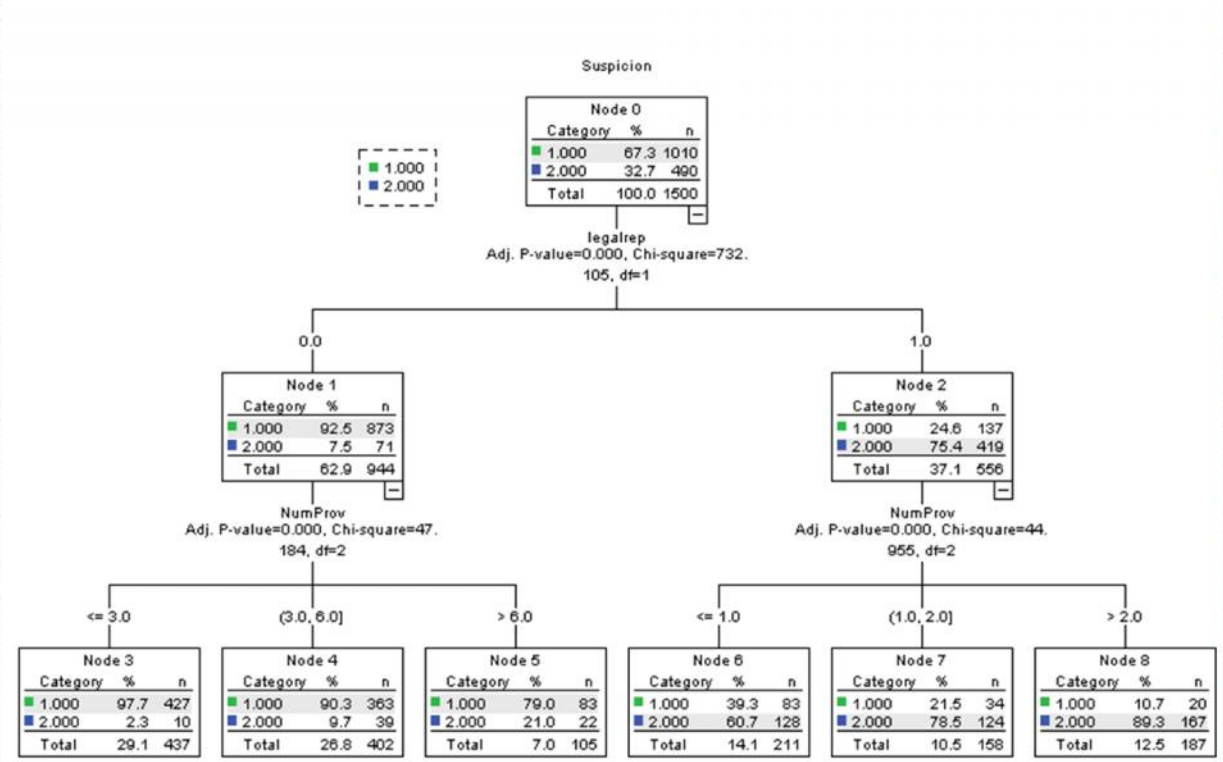
$$\sum y - E(y)^2$$

- 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*

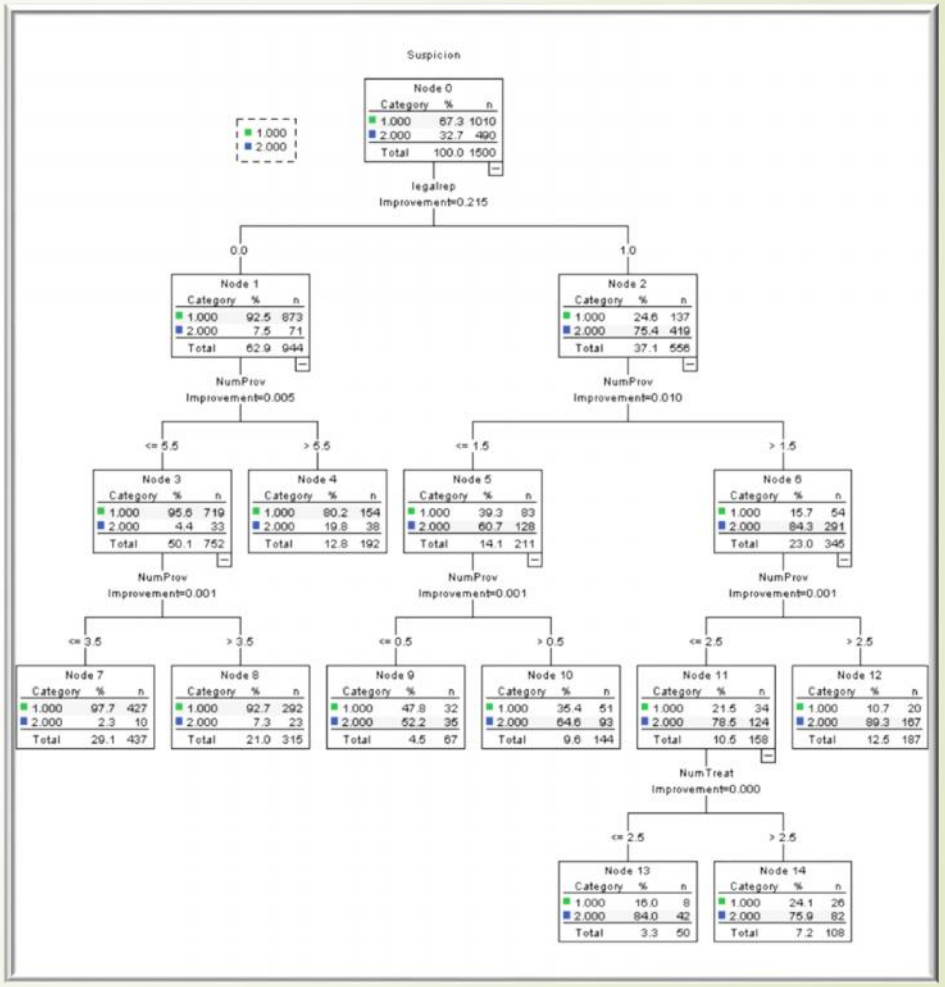
CHAID

- Minimize Chi-Square Statistic



C&RT

- Binary splits
- Gini Index, MSE



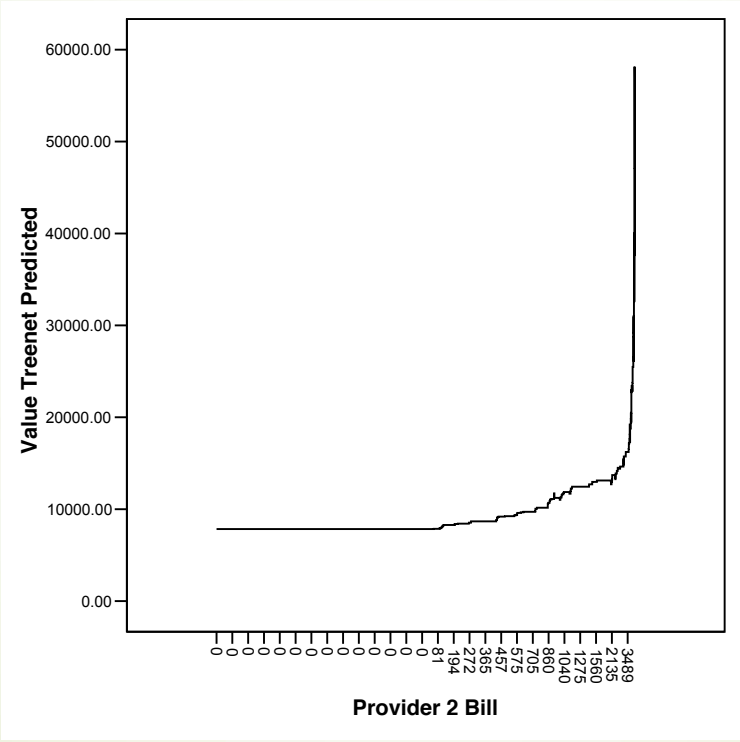
Different Kinds of Decision Trees

- 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

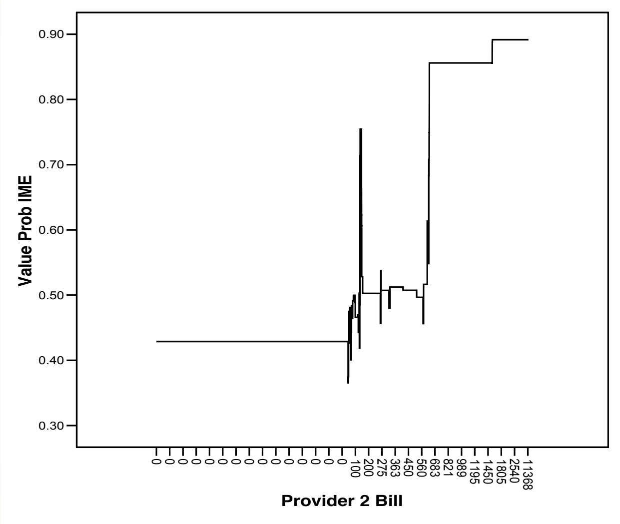
The Methods and Software Evaluated

1) TREENET
2) Iminer Tree
3) SPLUS Tree
4) CART
5) Iminer Ensemble
6) Random Forest
7) Naïve Bayes (Baseline)
8) Logistic (Baseline)

Ensemble Prediction of Total Paid



Ensemble Prediction of IME Requested



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)

Results for IME Requested

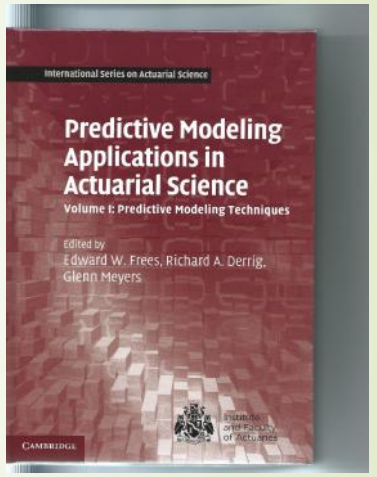
Area Under the ROC Curve – IME Decision				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.669	0.688	0.629	0.701
Lower Bound	0.661	0.680	0.620	0.693
Upper Bound	0.678	0.696	0.637	0.708
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.649	703	0.676	0.677
Lower Bound	0.641	695	0.669	0.669
Upper Bound	0.657	711	0.684	0.685

Results for SIU Referral

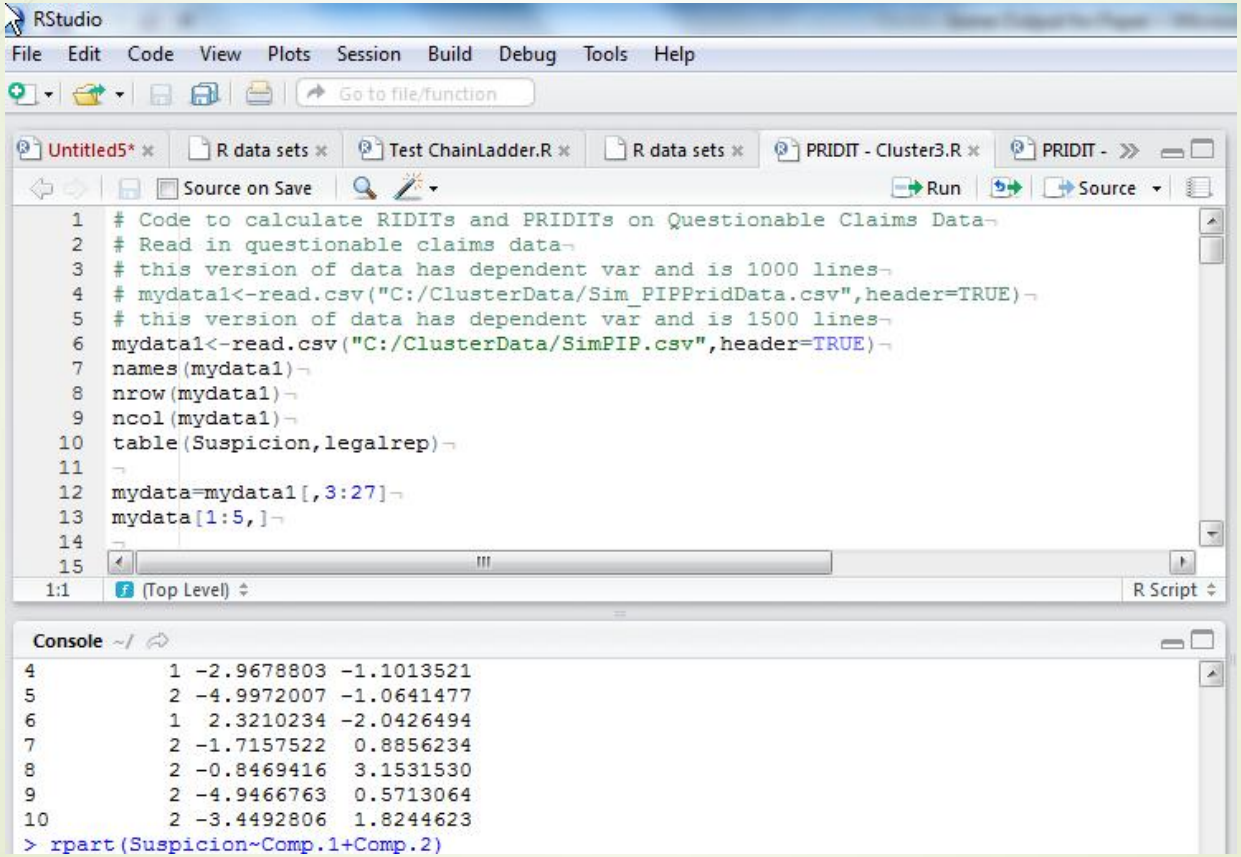
Area Under the ROC Curve – SIU Decision				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.607	0.616	0.565	0.643
Lower Bound	0.598	0.607	0.555	0.634
Upper Bound	0.617	0.626	0.575	0.652
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.539	0.677	0.615	0.612
Lower Bound	0.530	0.668	0.605	0.603
Upper Bound	0.548	0.686	0.625	0.621

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 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



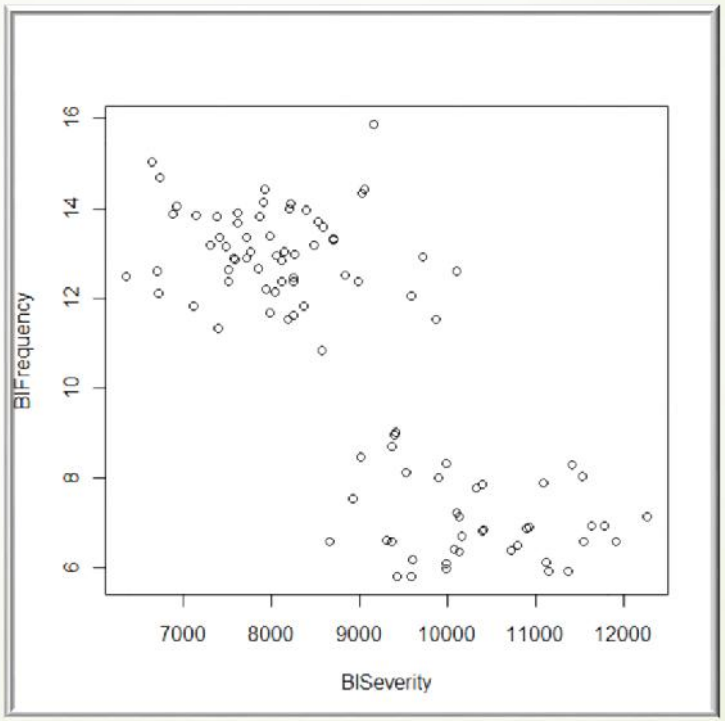
R Libraries

- 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

Dependent Variable Problem: Unsupervised Learning

- 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

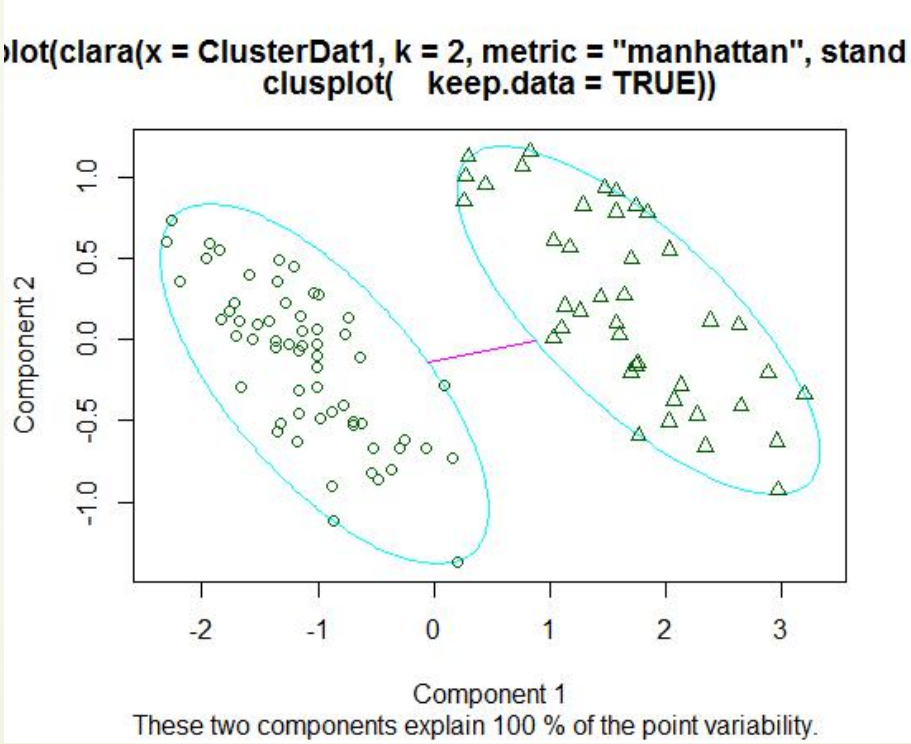
Grouping Records



Clustering

- Hierarchical clustering
- K-Means clustering
- Most frequent is k-means

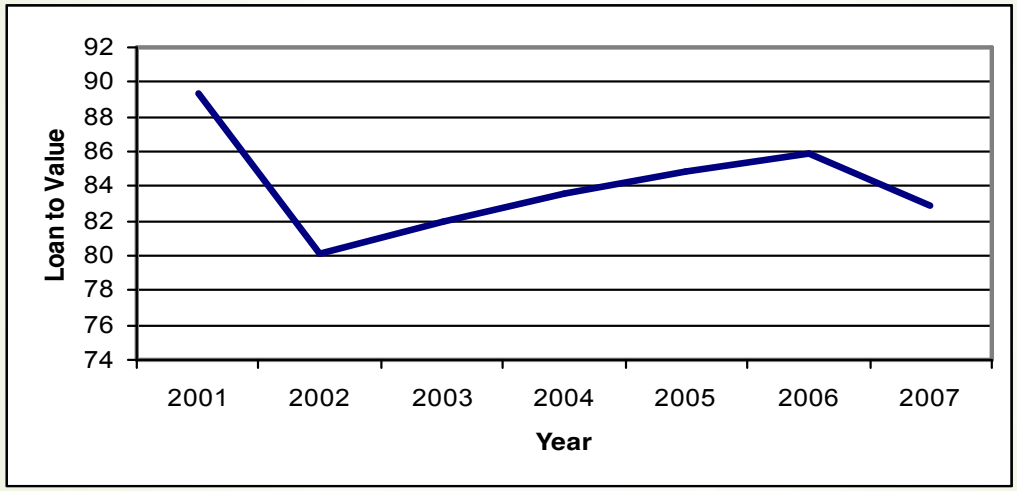
Cluster Plot



The Mortgage Crisis

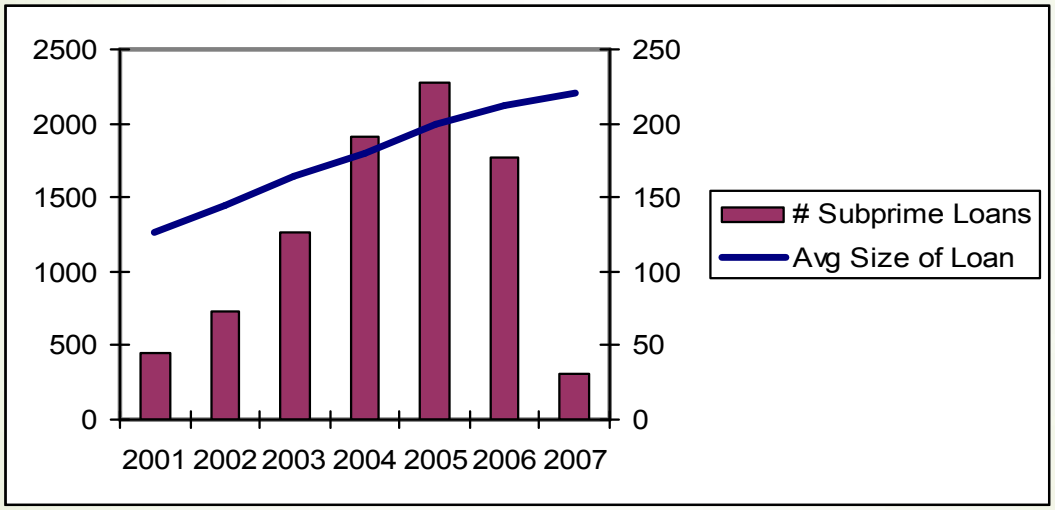
Could simple descriptive statistics have predicted the meltdown?

Time Series of Loan-to-Value



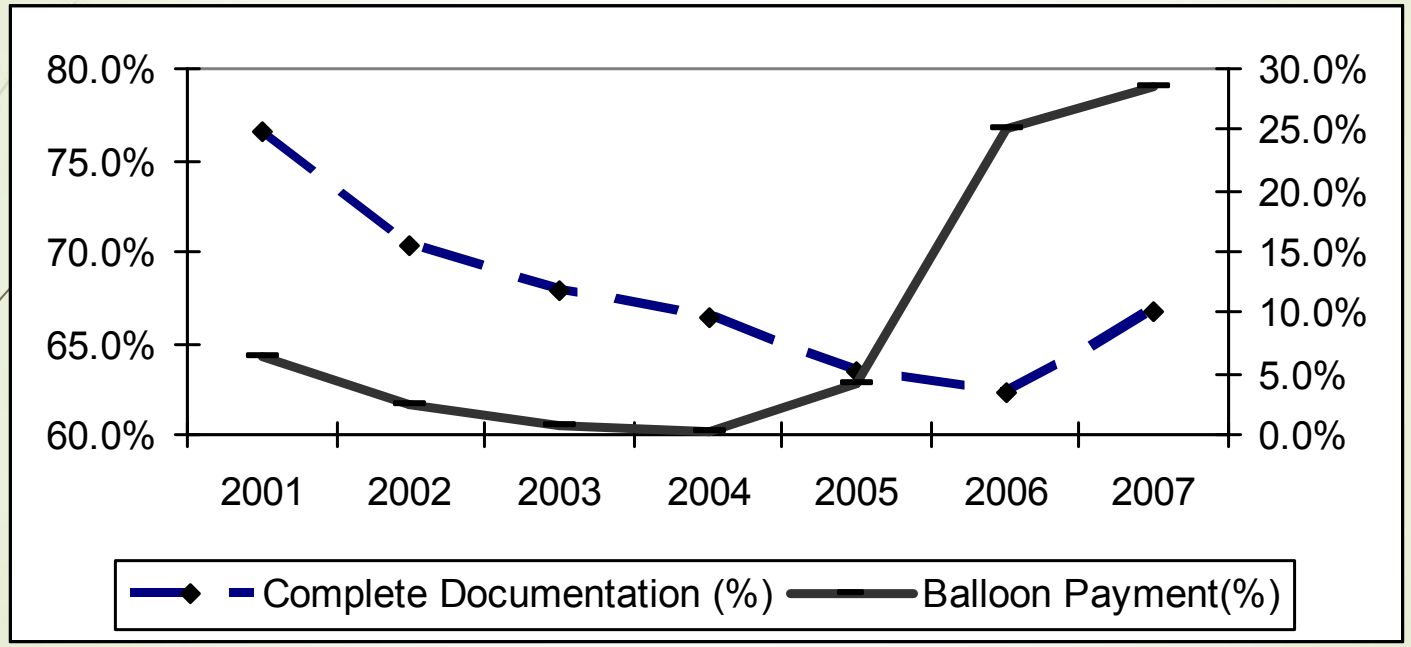
Data from Demyanyk and Van Hemert, “Understanding the Subprime Mortgage Crisis”, 2008

Subprime Loan Volume and Size



Data from Demyanyk and Van Hemert, 2008

Balloon Payments and Completed Documentation



Data from Demyanyk and Hemert, 2008

Observations from HMDA

- HMDA indicates lower income applicants tend to have a higher loan to income ratio
- HMDA cross-state comparison indicates states with a foreclosure problem have consistently higher loan to income ratios compared to states not experiencing a foreclosure problem

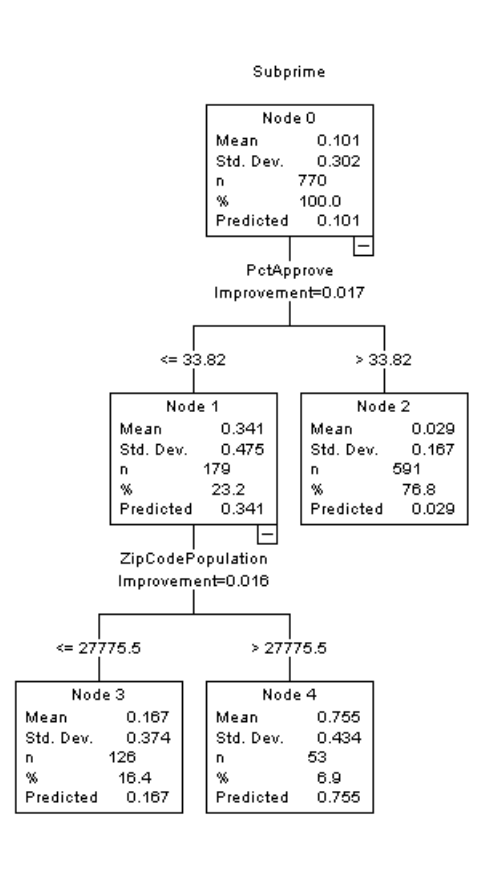
The Data

- HMDA Data
- LISC ZIP Foreclosure Needs Score
  - Subprime component
  - Foreclosure component
  - Disclosure component

  http://www.housingpolicy.org/foreclosure-response.html

- Zip Code Demographic Data

CART Subprime Tree



CART Foreclosure Variable Ranking

Independent Variable	Importance	Normalized Importance
Denial Percent	.027	100.0%
Mean Denial Score	.027	99.9%
PctApprove	.024	88.5%
ZipCodePopulation	.020	72.6%
PctPropNot1-4Fam	.019	69.5%
Median Rate Spread	.017	61.6%
PInCom	.016	60.5%
HouseholdsPerZipcode	.015	56.1%
Mean LTV Ratio	.014	52.7%

Results of Applying Clustering to HMDA Data

- K-means clustering applied to loan characteristics but not result data (i.e., approval)

Table III.5 – Means On Variables[1]

	Cluster		
	1	2	3
Avg Loan Amount	297.23	566.96	163.80
Average Income	165.71	356.66	87.26
Mean LTV[2] Ratio	2.53	2.38	2.48
Rate Spread - mean	4.84	4.54	5.05
Median LTV Ratio	2.29	2.09	2.31
Median Rate Spread	4.40	3.95	4.67
Percent Applicants High LTV	4.4	3.8	4.5
Pct Applicants High Rate Spread	4.7	4.5	5.6
Percent Manufactured, Multi Family Houses	1.9	.4	6.1
Pct Home Improvement	57.8	56.5	65.6
Percent Refinance	52.4	52.5	57.3
Pct Owner Occupied	18.1	28.4	13.5

Library for Getting Started

- Dahr, V, Seven *Methods for Transforming Corporate into Business Intelligence*, Prentice Hall, 1997
- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997, 2003
- 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
- Francis, L.A., Neural Networks Demystified, *Casualty Actuarial Society Forum,* Winter, pp. 254-319, 2001. Found at www.casact.org
- Francis, L.A., "Taming Text: An Introduction to Text Mining", CAS Winter Forum, March 2006, www.casact.org
- Francis, L.A., Martian Chronicles: Is MARS better than Neural Networks? *Casualty Actuarial Society Forum,* Winter, pp. 253-320, 2003.
- 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