Francis Analytics

Actuarial Data Mining Services

Capitalizing on Decision Trees to Advance Predictive Capabilities

Louise.francis@data-mines.com

www.data-mines.com

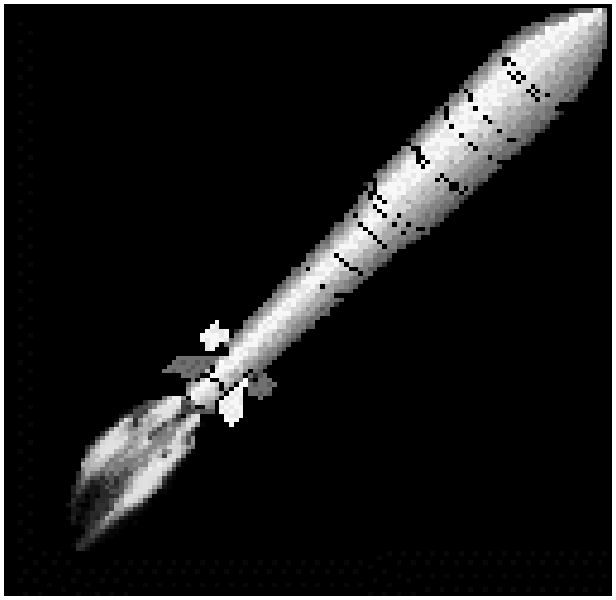
Outline of Workshop

- History of data mining and using trees for data mining
- Background in conventional approaches to modeling – the tree ancestors
- The data
- Trees and Tree applications
- Software
- Where to find data
- References

Trees within the context of data mining

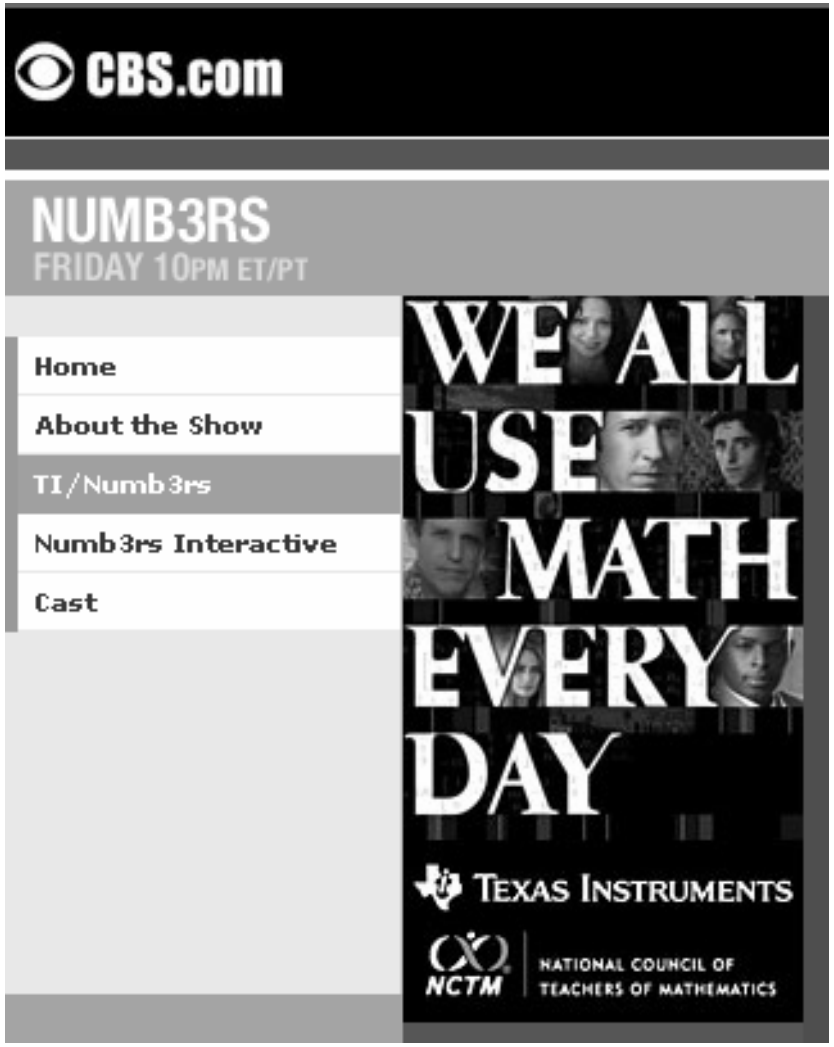
Why Predictive Modeling?

- Better use of data than traditional methods
- Advanced methods for dealing with messy data now available
- Decision Trees a popular form of data mining



Core Part of a Business Strategy

Data Mining Goes Prime Time



Real Life Insurance Application – “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

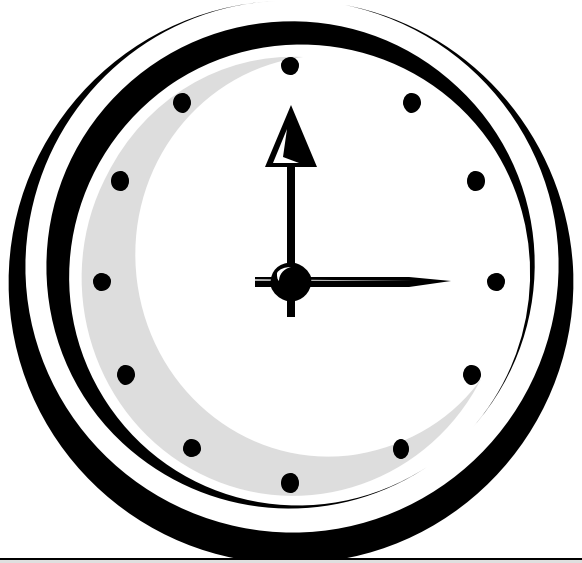
By PATRICK HEALY

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

Timeline of Casualty Actuarial Evolution



Week of 1/1/1900				
1/1/1900	1/2/1900	1/3/1900	1/4/1900	1/5/1900
Stone Age - 1914	**Pre-Industrial Age: 1970s**	**Industrial Age: 1970s**	**Computer Age: 1990s**	**Current Era**

CHAID: The First DM Paper: 1990

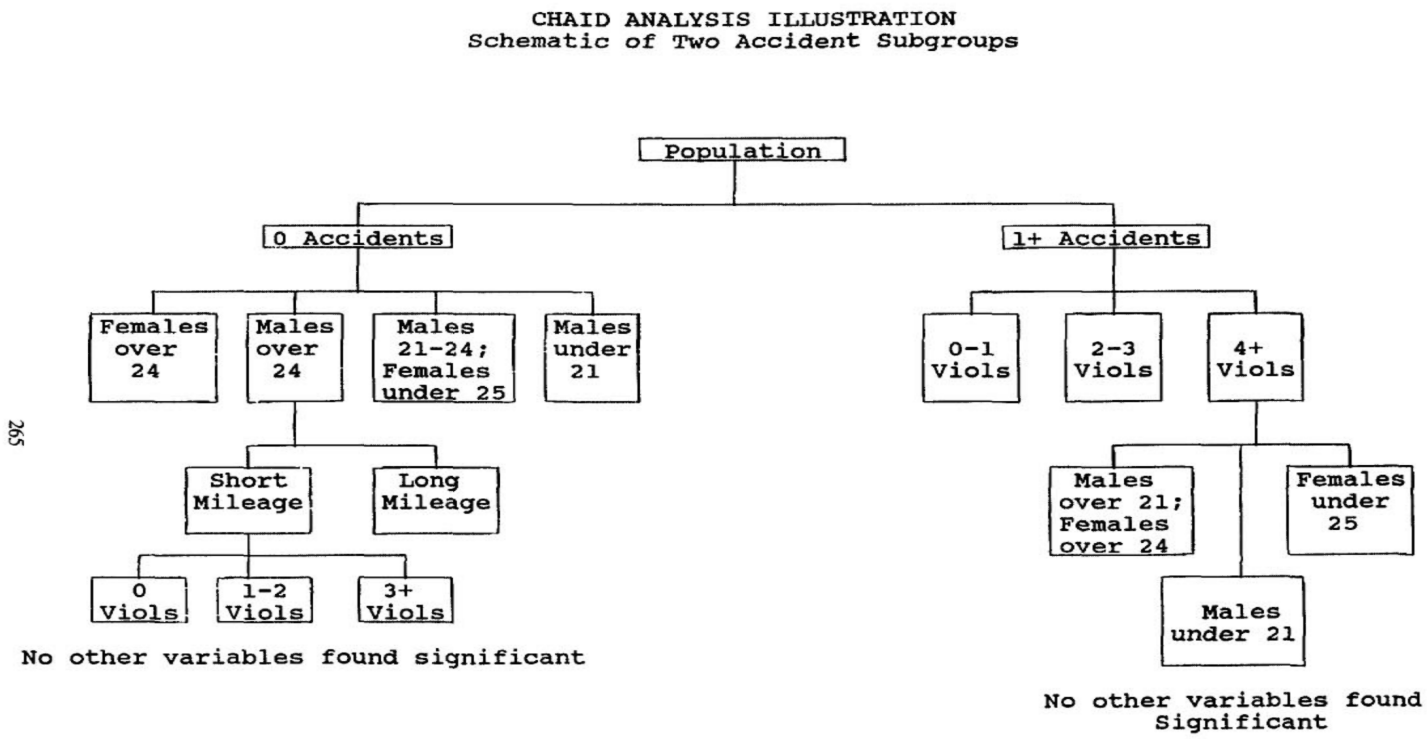


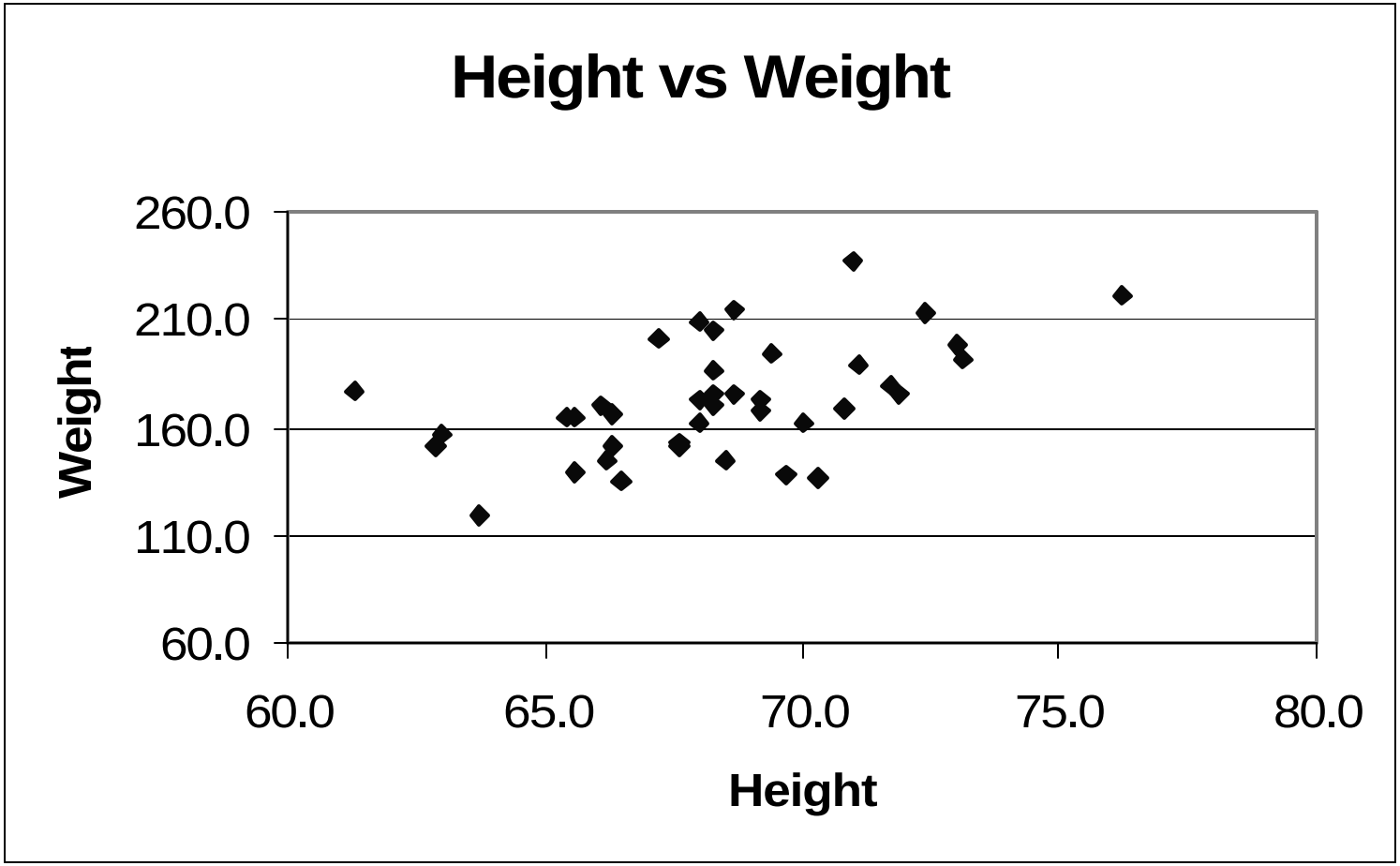
Exhibit A

Major Kinds of Data Mining

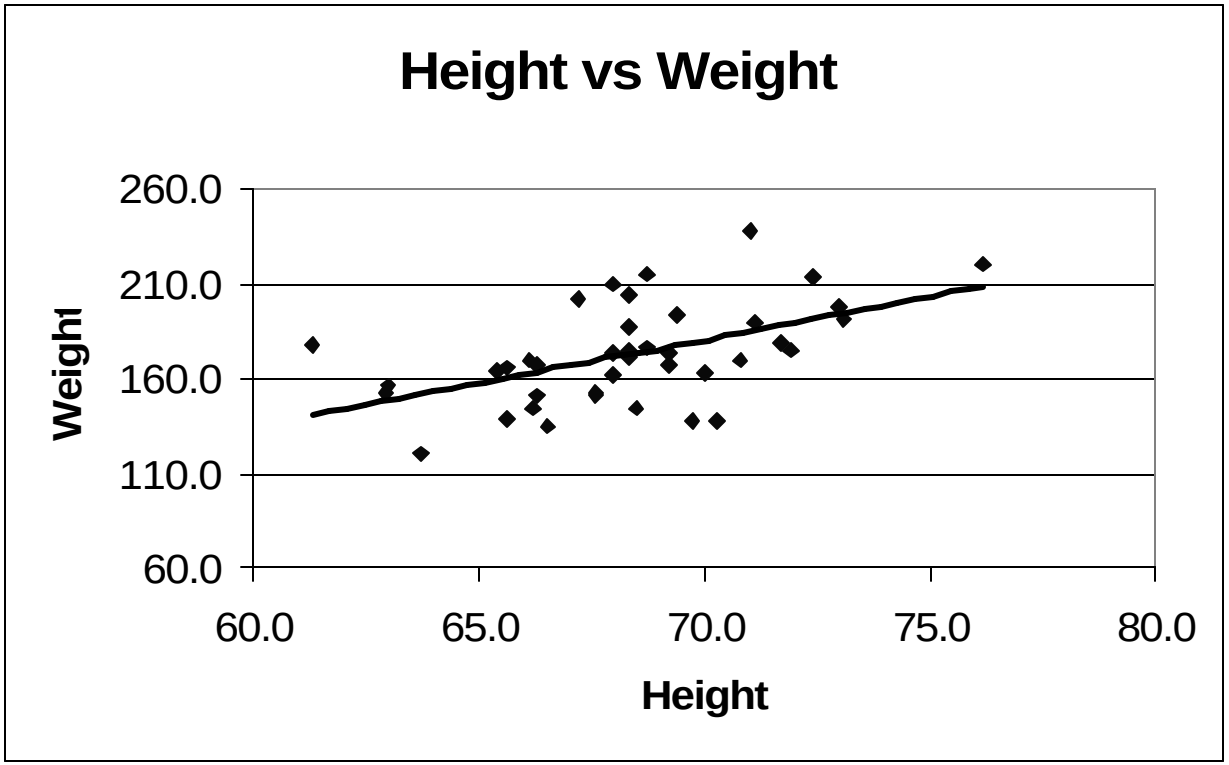
- Supervised learning
  - Most common situation
  - A dependent variable
    - Severity
    - Loss ratio
    - Fraud/no fraud
  - Some methods
    - Regression
    - CART
    - CHAID
    - Some neural networks
- Unsupervised learning
  - No dependent variable
  - Group like records together
    - A group 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

Background – Conventional Statistical Models

Linear Models



Model : Fit a Line Through Data

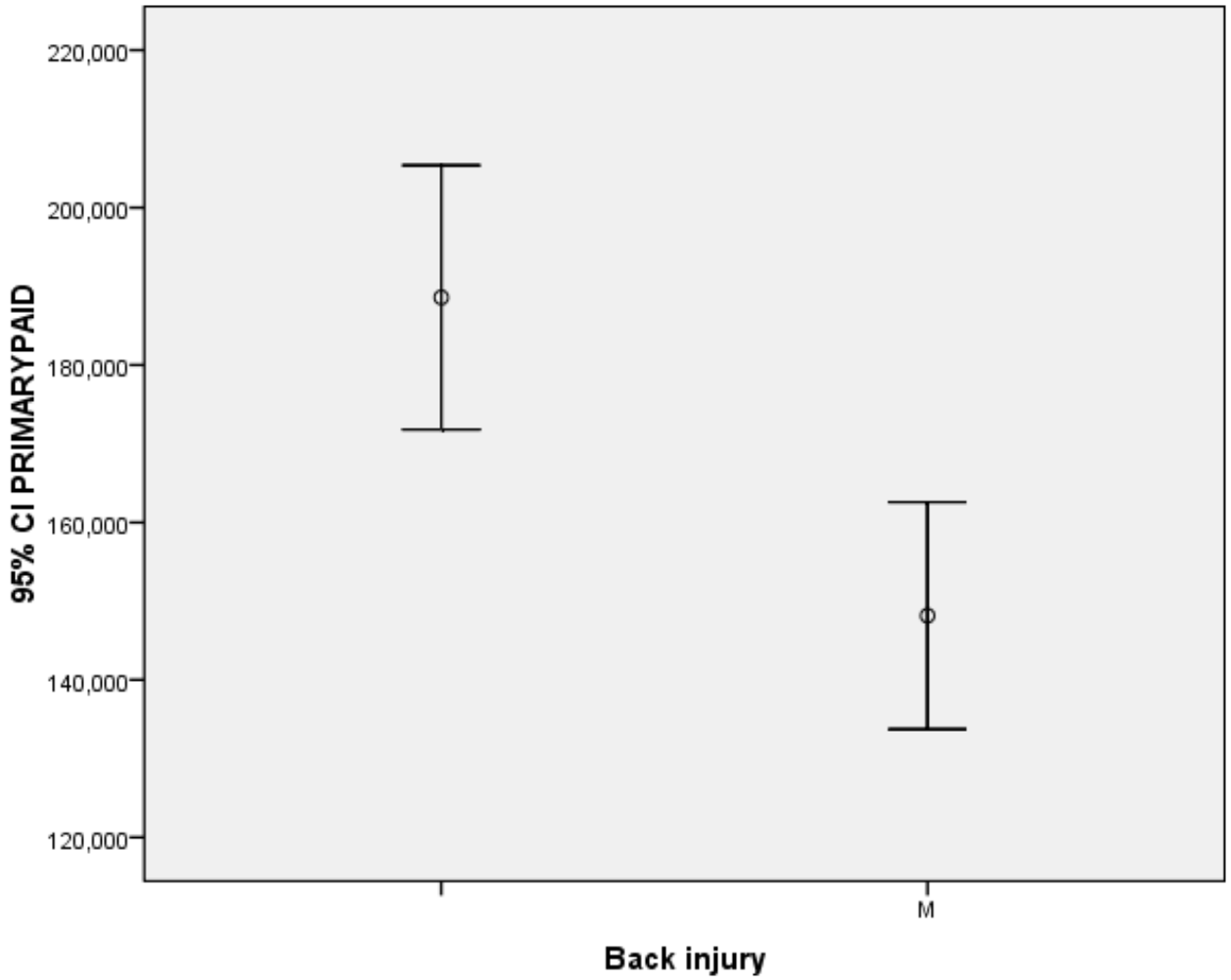


Linear Model

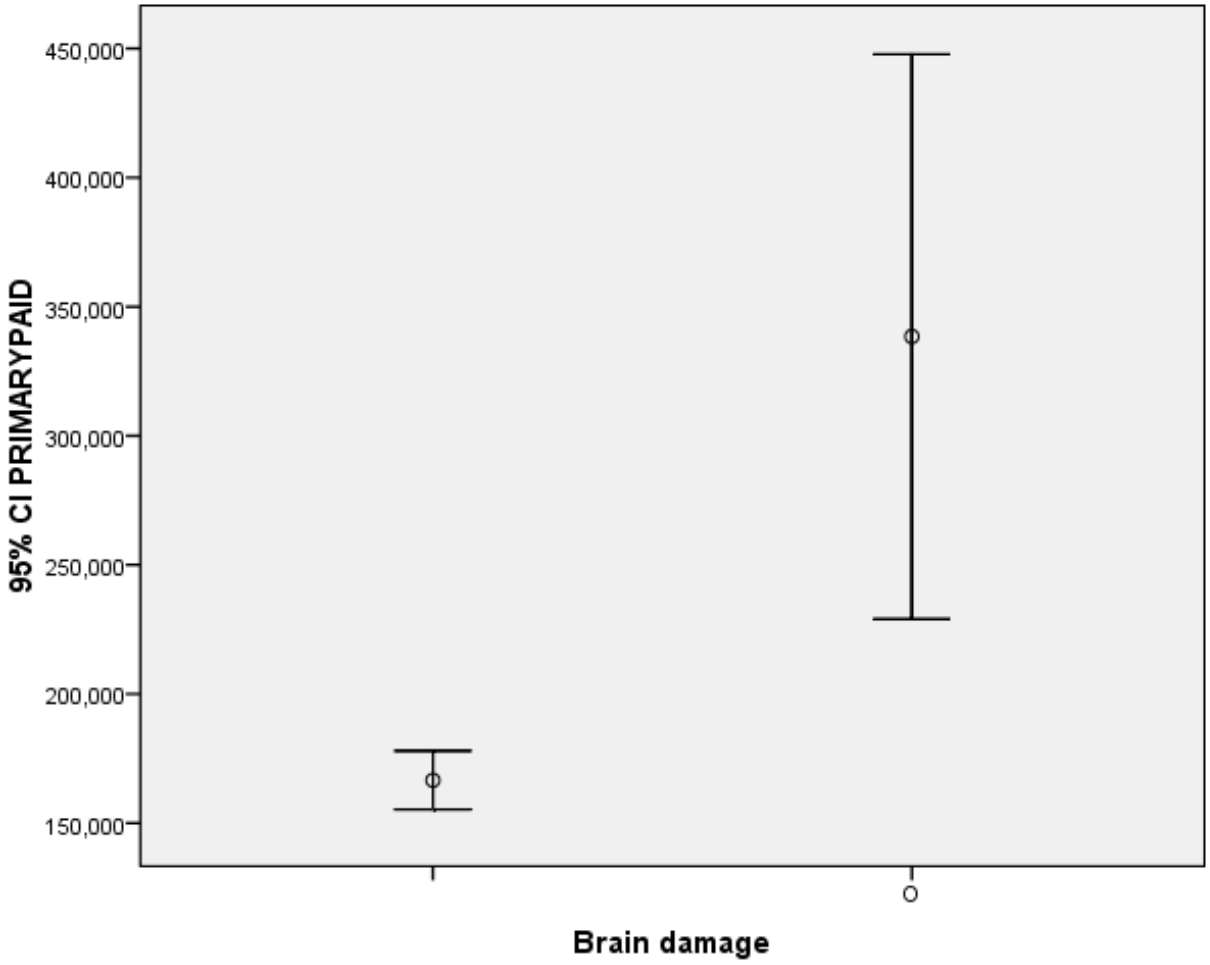
- $Y = a + b * x + e$

- Y is dependent variable (weight)
- X is independent variable (height)
- A is intercept or constant
- e is random error
  - e = Observed Y – Model Y

Another Linear Model: Analysis of Variance (ANOVA)



Another Linear Model: Analysis of Variance (ANOVA)



ANOVA for Categorical Independents

- Y = a + b1 * category1 + b2 * category2 + e
- Y is continuous dependent variable
- Category variables are indicator variables
- A is base category

Two Categories

- Model $Y = a_i$, where i is a category of the independent variable
- In traditional statistics we compare a_1 to a_2

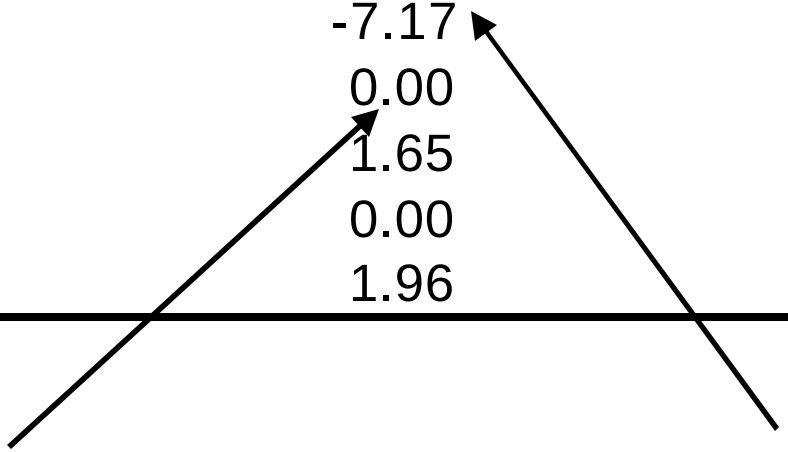
Fitting ANOVA With Two Categories Using A Regression

- Create A Dummy Variable for Attorney Involvement
- Variable is 1 If Attorney Involved, and 0 Otherwise

Attorneyinvolvement-insurer	Attorney	TotalSettlement
Y	1	25000
Y	1	1300000
Y	1	30000
N	0	42500
Y	1	25000
N	0	30000
Y	1	36963
Y	1	145000
N	0	875000

If Only Two Categories: T-Test for test of Significance of Independent Variable

	Variable 1	*Variable 2*
Mean	124,002	440,758
Variance	2.35142E+11	1.86746E+12
Observations	354	1448
Hypothesized Mea	0	
df	1591	
t Stat	-7.17	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.00	
t Critical two-tail	1.96	



Use T-Test from Excel Data Analysis Toolpak

More Than 2 Categories

- If there are K Categories-
- Create k-1 Dummy Variables
  - $Dummy_i$ = 1 if claim is in category i, and is 0 otherwise
- The k^{th} Variable is 0 for all the Dummies
- Its value is the intercept of the regression

Design Matrix

Severity	Injury	Dummy 1	Dummy 2	Dummy 3	Dummy 4	Dummy 5	Dummy 6	Dummy 7	Dummy 8
-	BRUISE	0	1	0	0	0	0	0	0
271.53	OTHER	0	0	0	0	0	0	0	0
751.71	STRAIN	0	0	1	0	0	0	0	0
762.08	FRACTURE	0	0	0	0	1	0	0	0
796.75	CUT/PUNCT	1	0	0	0	0	0	0	0
382.20	BRUISE	0	1	0	0	0	0	0	0
171.35	EYE	0	0	0	0	0	0	1	0

Injury Code	Injury_Backinjury	Injury_Multipleinjuries	Injury_Nervouscondition	Injury_Other
1	0	0	0	0
1	0	0	0	0
12	1	0	0	0
11	0	1	0	0
17	0	0	0	1

Top table Dummy variables were hand coded, Bottom table dummy variables created by XLMiner.

Regression Output for Categorical Independent

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.16
R Square	0.03
Adjusted R Square	0.02
Standard Error	19,621.92
Observations	4,112.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	4.36E+10	5.45E+09	14	0
Residual	4103	1.58E+12	3.85E+08		
Total	4111	1.62E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6,410.86	954.05	6.72	0.00	4,540.40	8,281.32
Dummy 1	(5,130.72)	1,130.93	(4.54)	0.00	(7,347.96)	(2,913.48)
Dummy 2	(2,153.48)	1,147.89	(1.88)	0.06	(4,403.96)	97.00
Dummy 3	1,140.73	1,148.45	0.99	0.32	(1,110.86)	3,392.31
Dummy 4	(2,332.76)	1,683.84	(1.39)	0.17	(5,634.00)	968.48
Dummy 5	8,148.78	1,716.79	4.75	0.00	4,782.94	11,514.61
Dummy 6	(4,205.91)	1,656.39	(2.54)	0.01	(7,453.34)	(958.48)
Dummy 7	(5,871.33)	2,299.01	(2.55)	0.01	(10,378.63)	(1,364.03)
Dummy 8	(5,532.85)	2,516.55	(2.20)	0.03	(10,466.65)	(599.04)

Assumptions of Regression

- Errors independent of value of X
- Errors independent of value of Y
- Errors independent of prior errors
- Errors are from normal distribution
- Linearity

Generalized Linear Models (GLMs)

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

Generalized Linear Models
Common Links for GLMs

The identity link: h(Y) = Y

The log link: h(Y) = *ln*(Y)

The inverse link: h(Y) = $\frac{1}{Y}$

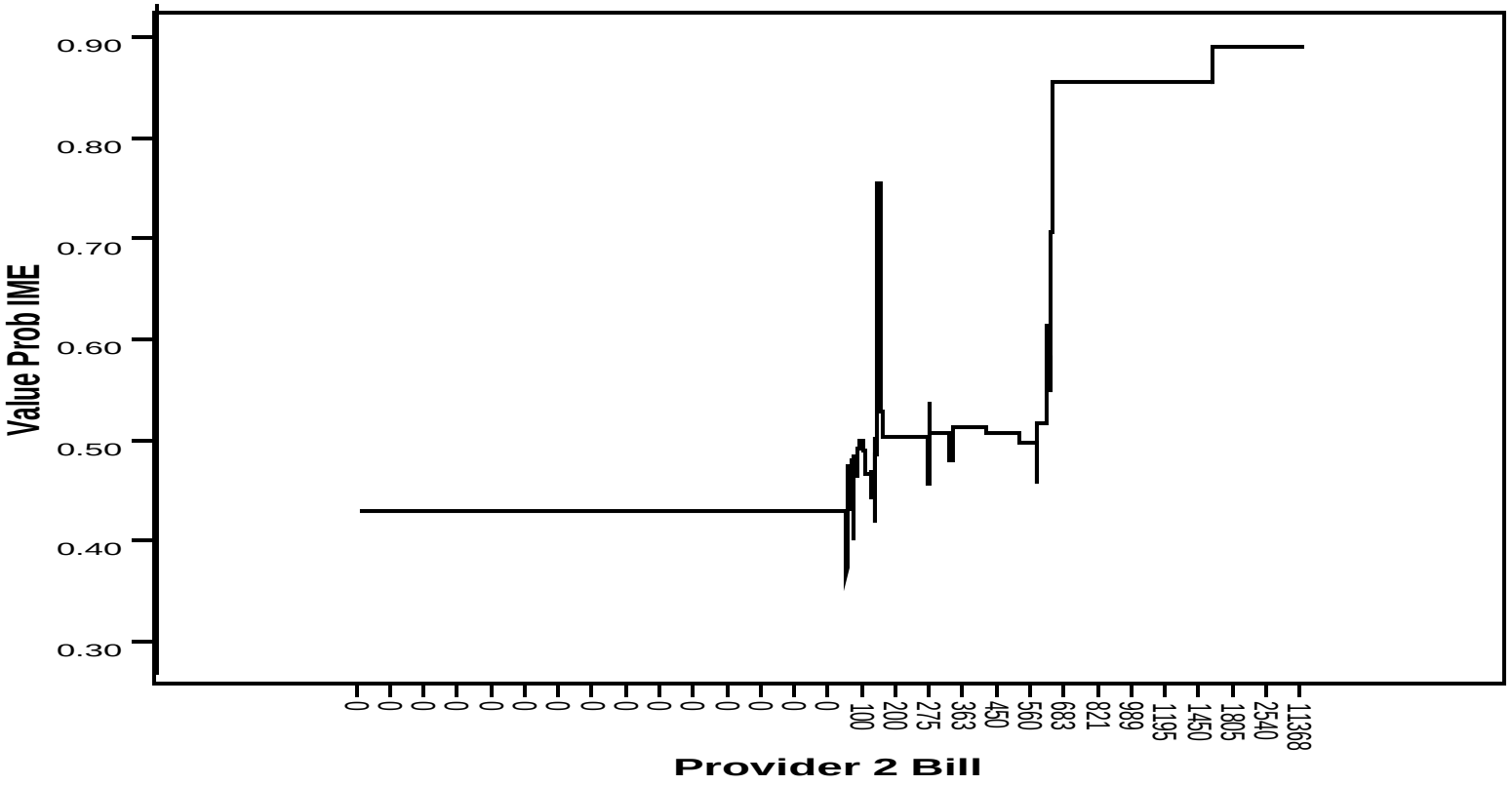
The logit link: h(Y) = $\ln(\frac{Y}{1-Y})$

The probit link: h(Y) = $\Phi(Y)$, Φ denotes the normal CDF

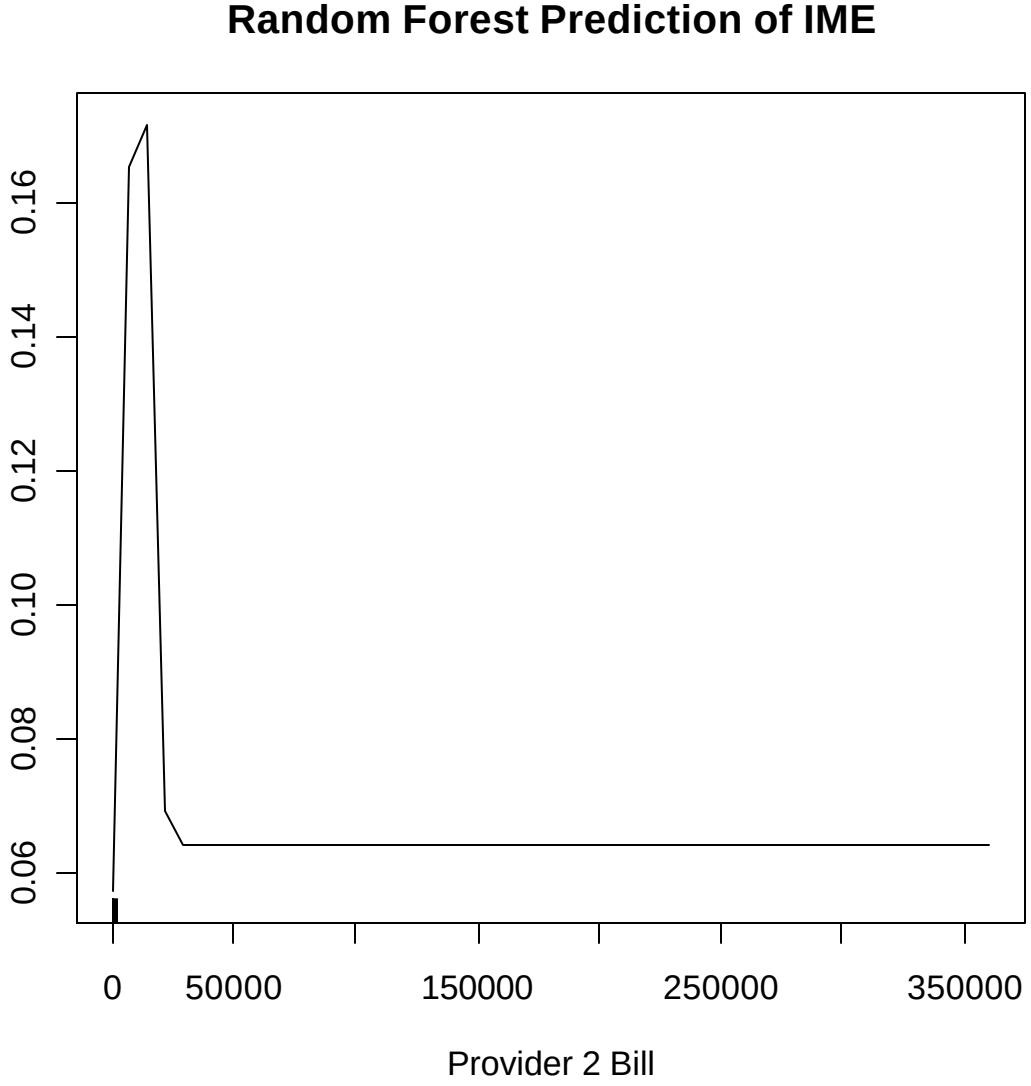
Summary Classical Models

- Classical models are linear
- Regression and ANOVA are among the most commonly used techniques in statistical modeling
- Classical models do not handle non-linearity, interactions and other real-life complications as efficiently as data mining methods such as trees
- The CART tree models are closely related to ANOVA
- The CHAID tree models are also related to ANOVA (and to categorical analysis using the Chi-Squared), but with significant differences

An Insurance Nonlinear Function: Provider Bill vs. Probability of Independent Medical Exam

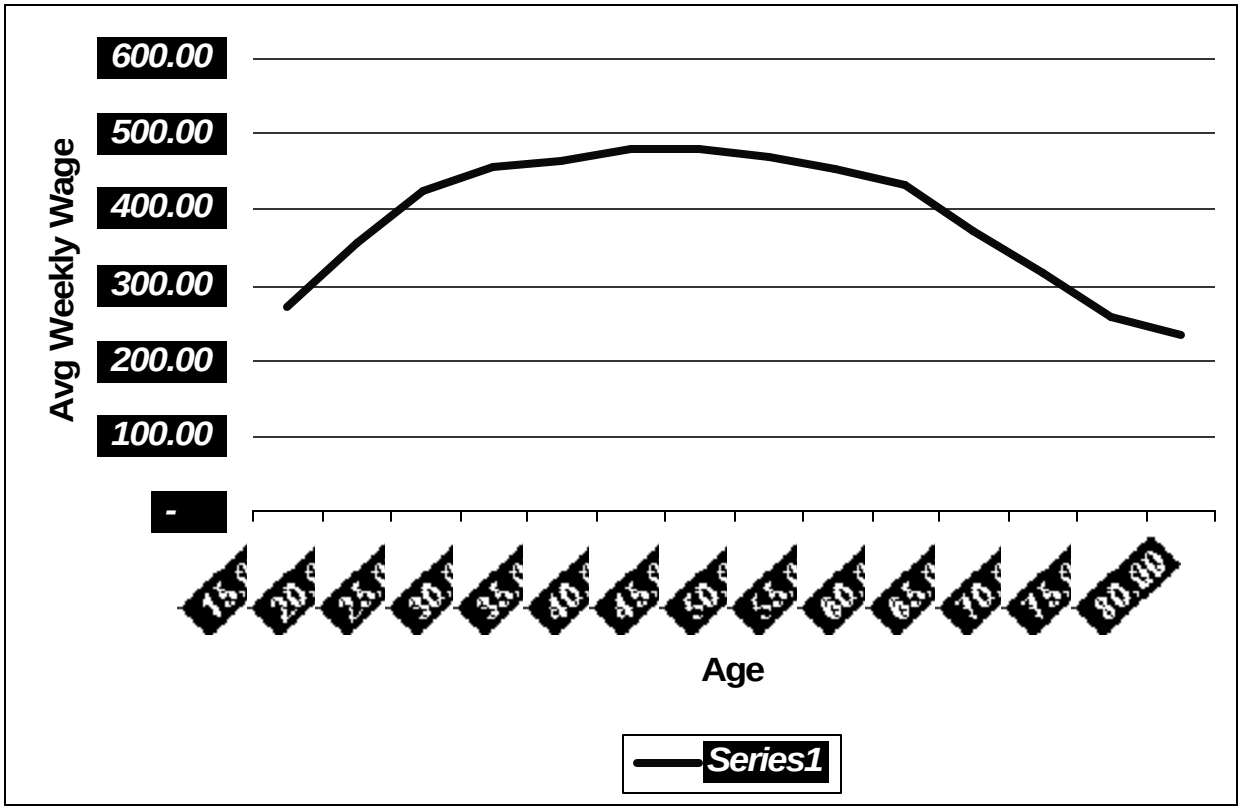


Data Complexities: Nonlinearities



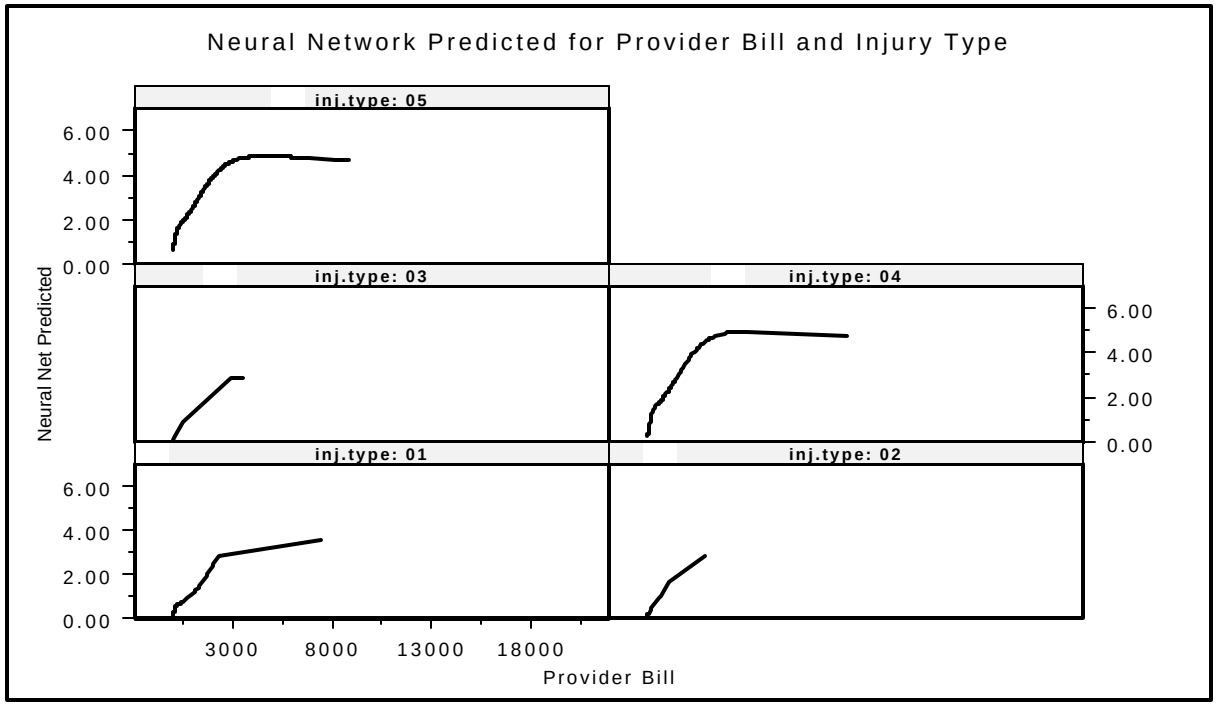
Data Complexities – Nonlinearities

Avg Wage vs Age



Interactions

- Effect of a predictor variable on dependent variable depends on the values of another variable(s)



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

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
- Handles missing data

The Data used for Illustration

1. Described in “Distinguishing the Forest From the Trees”, Derrig and Francis, 2005 CAS Winter Forum
2. Texas Dept of Insurance “WC” Closed Claims – will be posted at www.data-mines.com

The Fraud Database

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)

Predictor Variables

- Claim file variables
  - Provider bill, Provider type
  - Injury
- Derived from claim file variables
  - Attorneys per zip code
  - Docs per zip code
- Using external data
  - Average household income
  - Households per zip

TYPES OF FRAUD

- WORKERS' COMPENSATION
- Employee Fraud
- -Working While Collecting
- -Staged Accidents
- -Prior or Non-Work Injuries
- Employer Fraud
- -Misclassification of Employees
- -Understating Payroll
- -Employee Leasing
- -Re-Incorporation to Avoid Mod

Insurance Fraud- The Problem

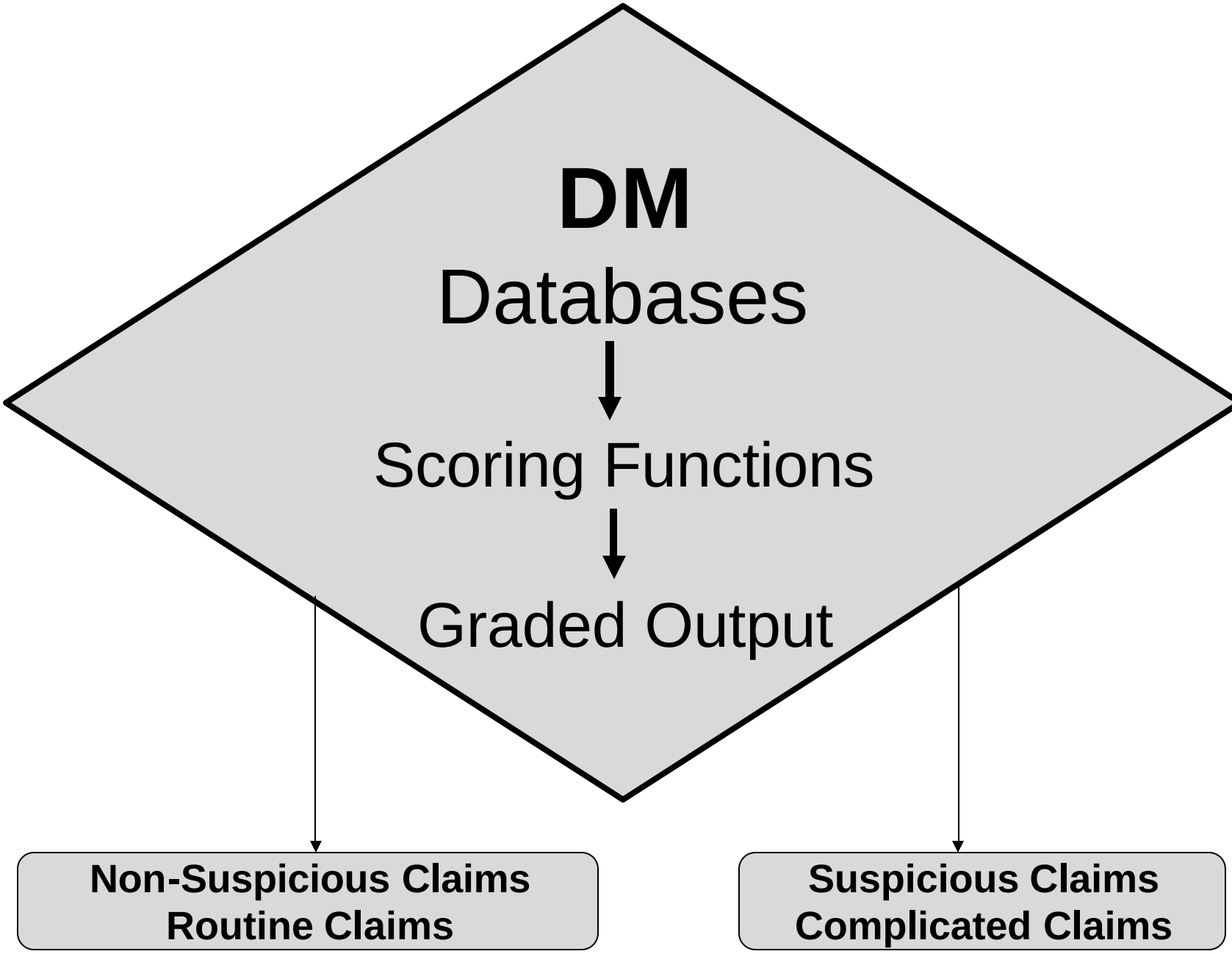
- ISO/IRC 2001 Study: Auto and Workers Compensation Fraud a Big Problem by 27% of Insurers.
- CAIF: Estimation (too large)
- Mass IFB: 1,500 referrals annually for Auto, WC, and (10%) Other P-L.

FRAUD IDENTIFICATION

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

REAL PROBLEM-CLAIM FRAUD

- 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


DM
Databases
Scoring Functions
Graded Output
Non-Suspicious Claims
Routine Claims
Suspicious Claims
Complicated Claims

Underwriting Red Flags

- **Prior Claims History (Mod)**
- **High Mod versus Low Premium**
- **Increases/Decreases in Payroll**
- **Changes of Operation**
- **Loss Prevention Visits**
- **Preliminary Physical Audits**
- **Check Yellow Pages**
- **Check Websites**

Texas Data

The Texas database

- Closed claims published on Texas department of insurance web site
- Selected “work related” claims
- Only claims over a threshold (i.e., $25,000)

Some Work Related Liability Data

Closed Claims from Tx Dept of Insurance

- Total Payment and primary payment
- Initial Indemnity reserve
- Policy Limit
- Attorney Involvement
- Lags
  - Closing
  - Report
- Injury
  - Sprain, back injury, death, etc
- Data, along with some of analysis will be posted on internet

WC Ratemaking Application

- The most common application
- Check CAS web site
- www.casact.org
- Go to continuing education part of site
- Get overheads from Predictive Modeling Seminar

WC Claims Related Applications

- Reserving
- Identify complex claims and assign to experienced adjustor
- Identify simple claims and pay
- Identify claims for subrogation
- Identify fraud and abuse

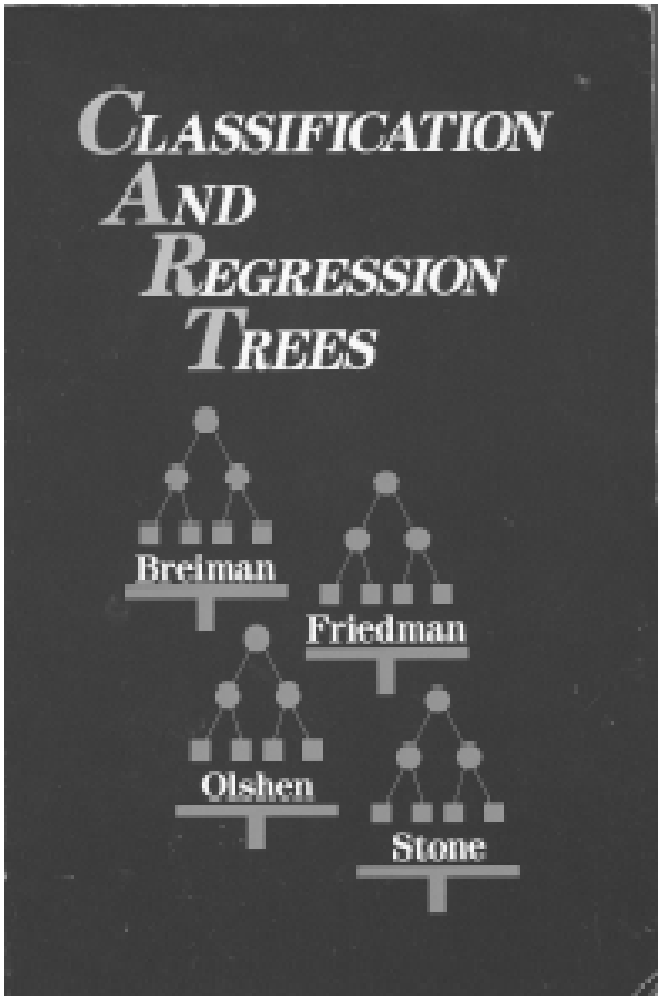
Decision Trees

Two Useful Data Mining Methods

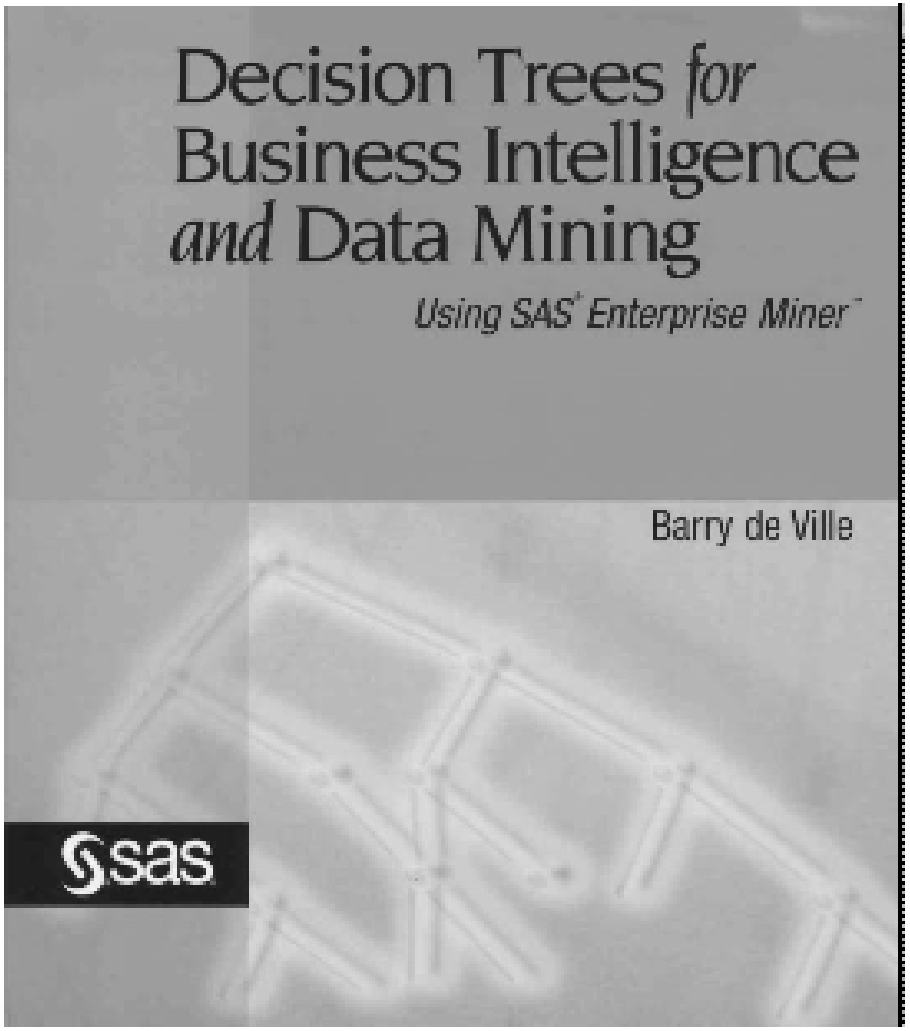
- Decision Trees
  - Classification and Regression Trees (CART)
  - Chi Square Automatic Interaction Detection (CHAID)

The Classic Reference on Trees

Brieman, Friedman Olshen and Stone, 1993



Practical Introductory Reference



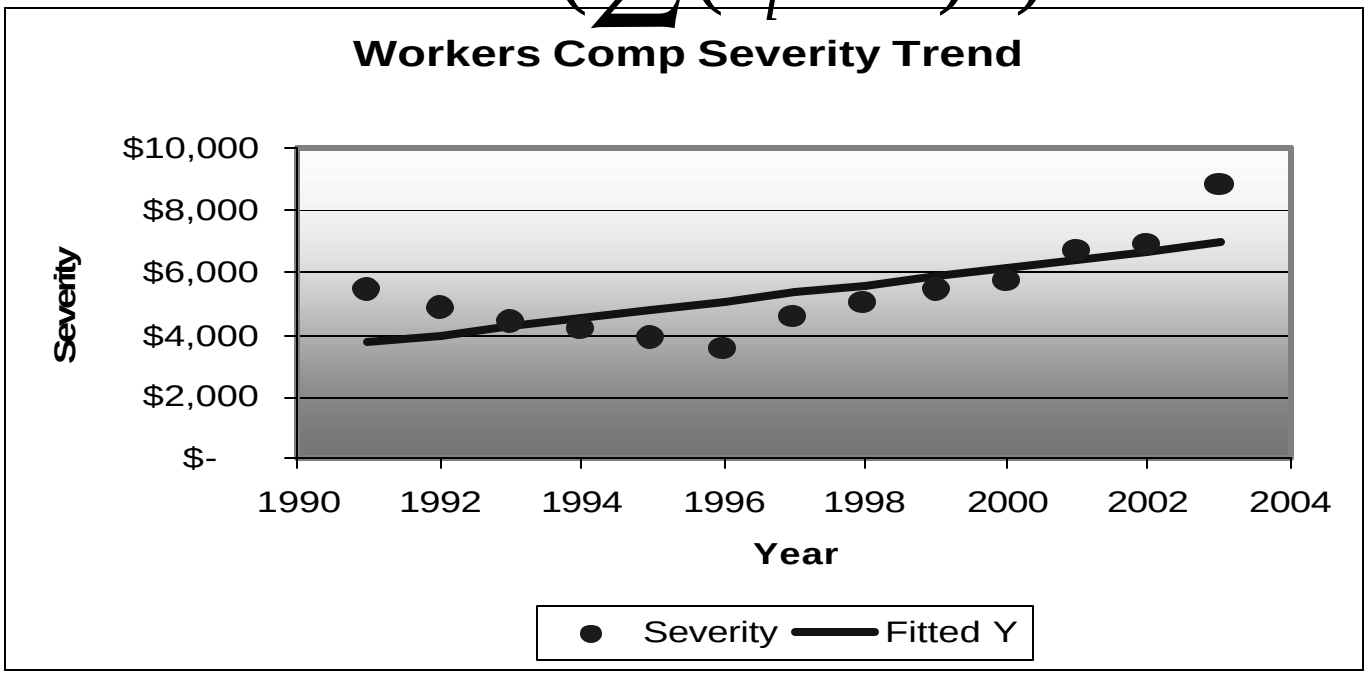
Decision Trees: Recursive Partitioning

- To partition on numeric (or ordered variables)
  - Select one of the variable
  - Test all possible binary splits of the variable
    - If (x<t) data to left, otherwise it goes to right
  - Compute mean (or proportion for categorical dependent) variable for each possible two-way split of the data
  - Select the split that optimizes a goodness of fit statistic
  - Do this for all variables

Classical Statistics: Regression

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

$$\min\left(\sum (Y_i - \widehat{Y})^2\right)$$



Regression Trees

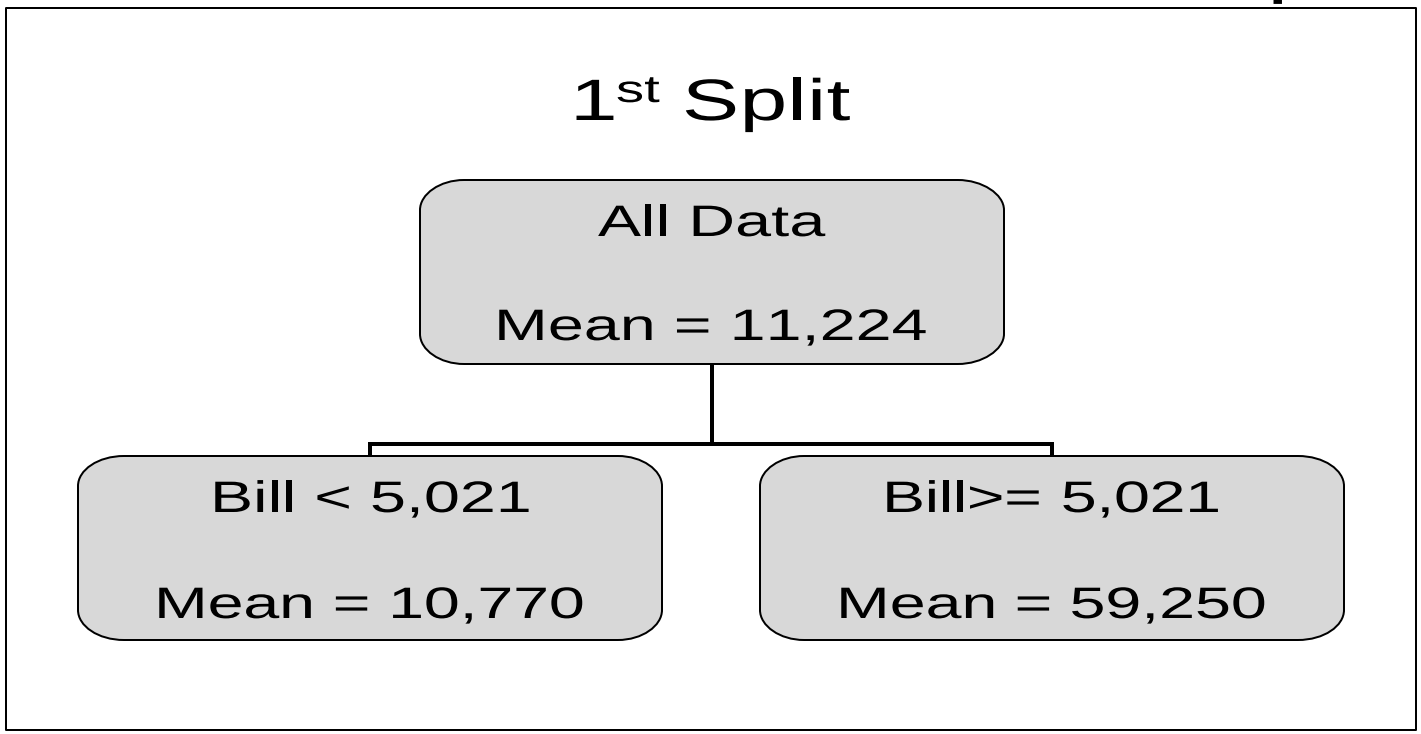
- Tree-based modeling for ***continuous*** **target variable**
  - most intuitively appropriate method for loss ratio analysis
- Find split that produces greatest separation in

$$\mathbf{?}\,[\mathbf{y} - \mathbf{E(y)}]^{\mathbf{2}}$$

- i.e.: find nodes with minimal *within variance*
  - and therefore greatest *between variance*
  - like credibility theory i.e.: find nodes with minimal *within variance*

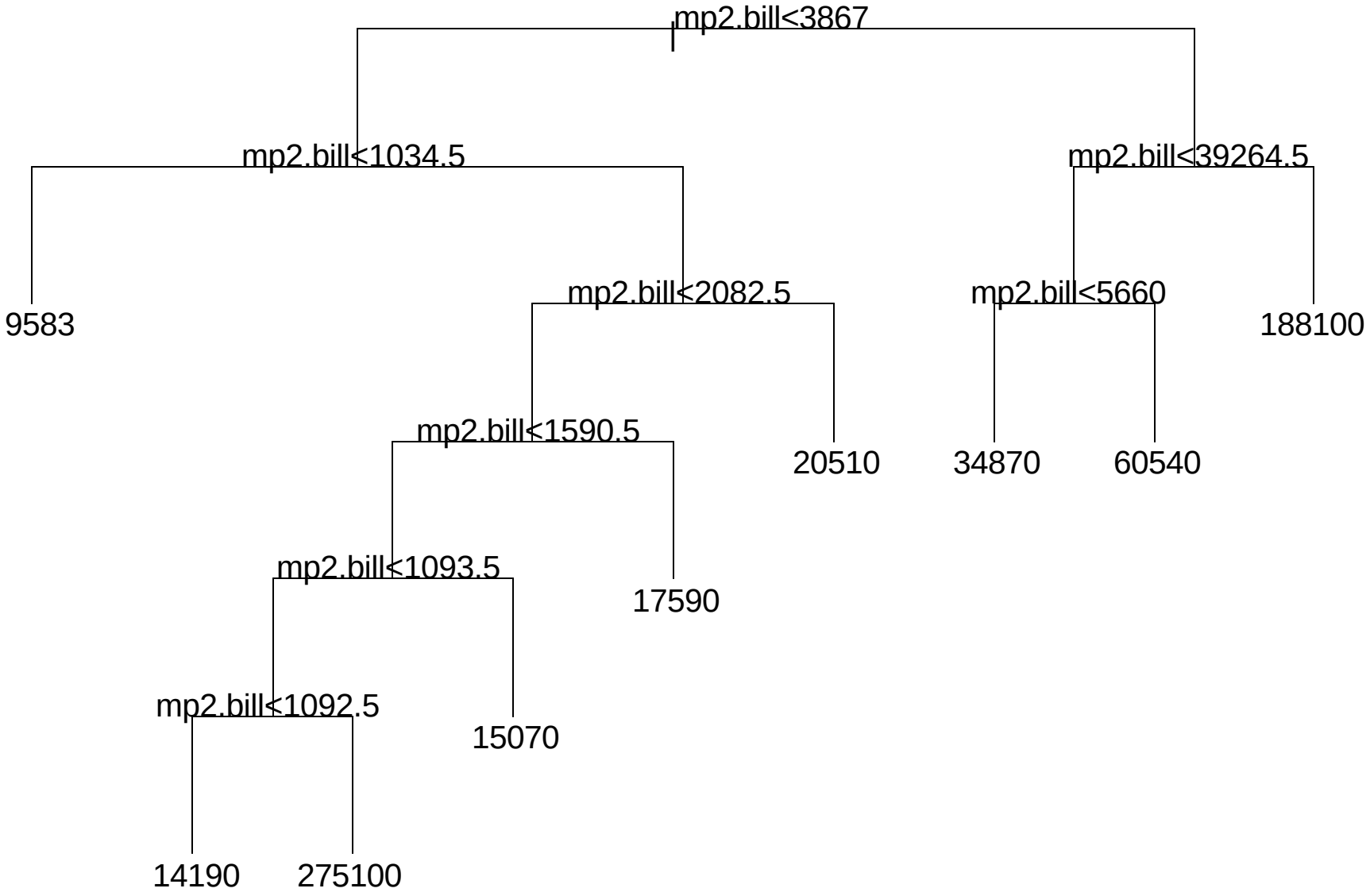
- Every record in a node is assigned the same expectation➔ model is a *step function*

CART – Example of 1st split on Provider 2 Bill, With Paid as Dependent



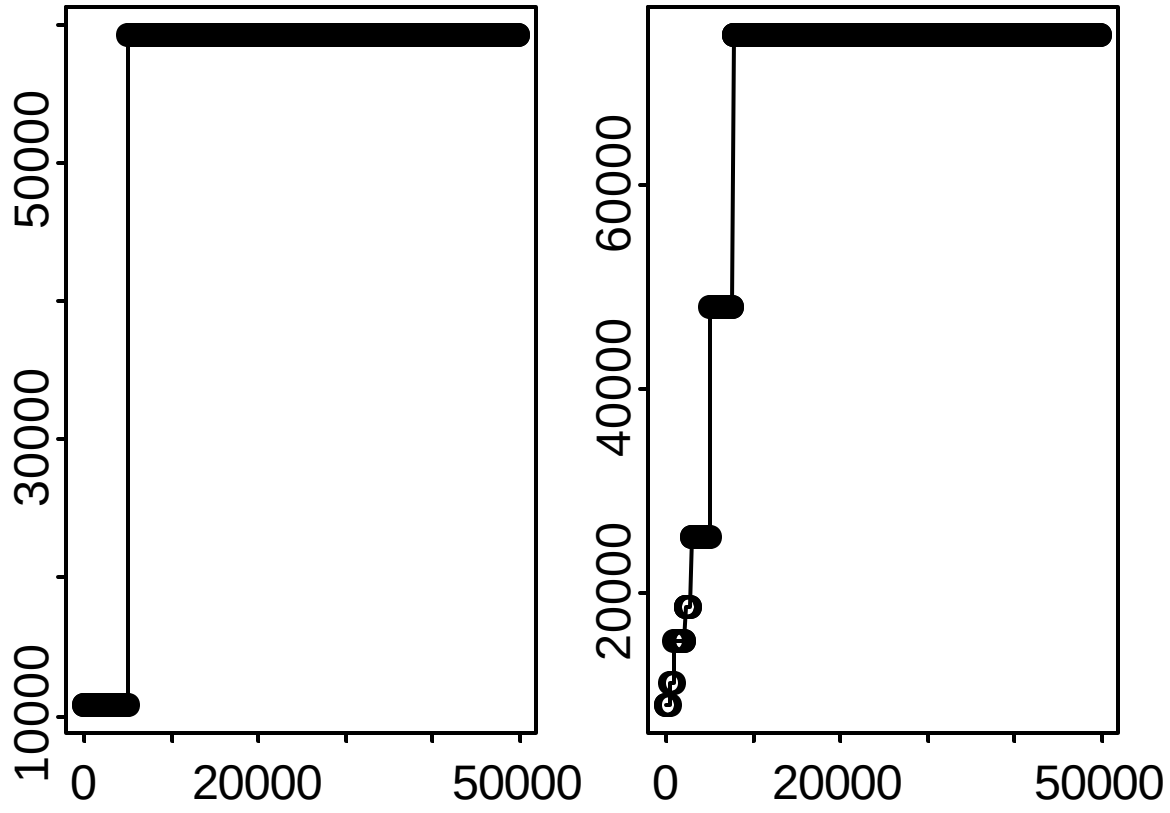
- For the entire database, total squared deviation of paid losses around the predicted value (i.e., the mean) is 4.95x1013. The SSE declines to $4.66x10^{13}$ after the data are partitioned using $5,021 as the cutpoint.
- Any other partition of the provider bill produces a larger SSE than $4.66x10^{13}$. For instance, if a cutpoint of $10,000 is selected, the SSE is $4.76*10^{13}$.

Continue Splitting to get more homogenous groups at terminal nodes



CART Step Function Predictions with One Numeric Predictor

Total Paid as a Function of Provider 2 Bill



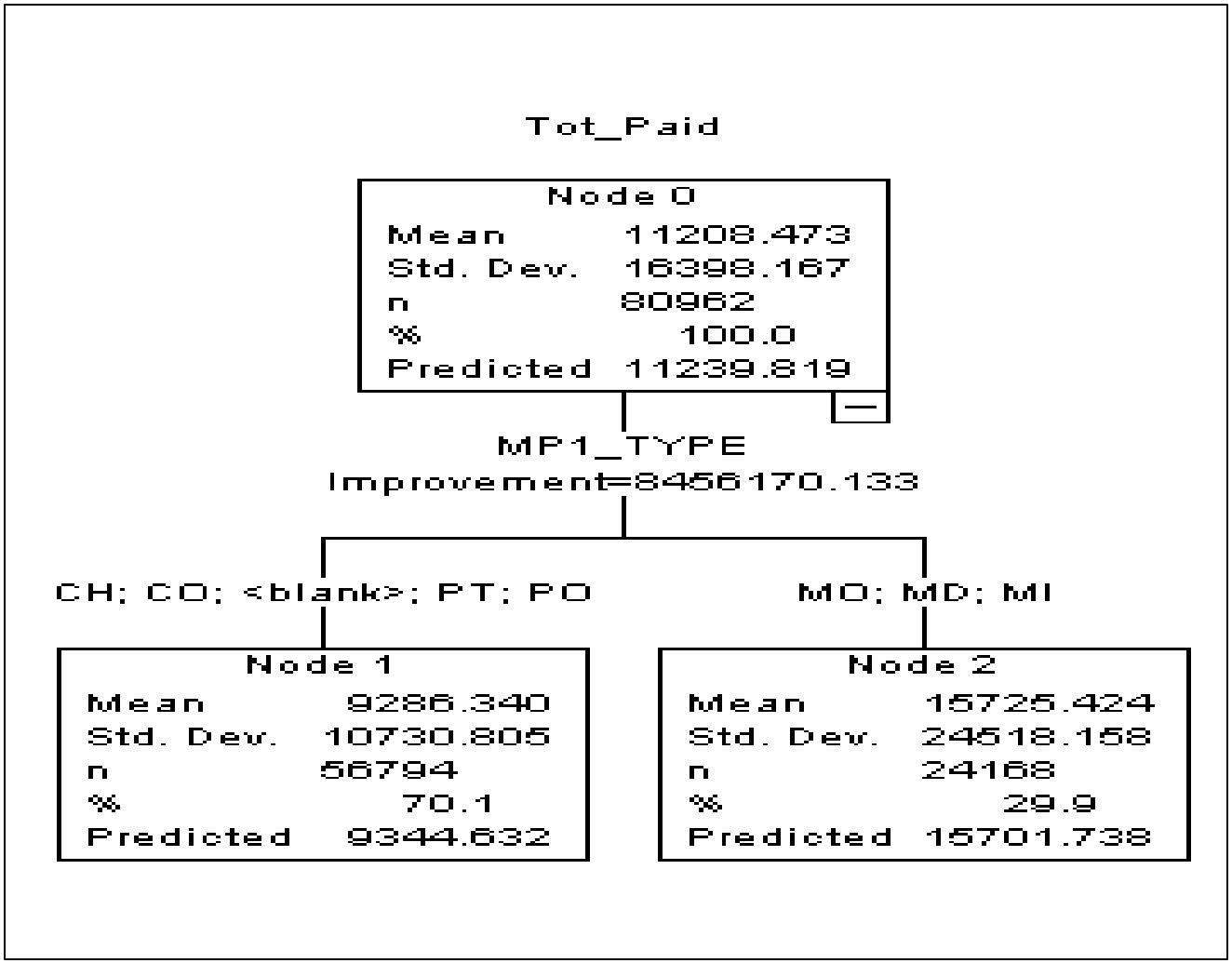
Classification Trees: Categorical Dependent

- Find the split that maximizes the difference in the probability of being in the target class
- Find split that minimizes *impurity*, or number of records not in the dominant class for the node
- Common goodness of fit measures are GINI index and entropy (deviance)

Decision Trees: Partitioning Cont

- For categorical variables
  - Select one of the variable
  - Test all possible two way groupings of the variable
    - Because order is irrelevant there may be many possibilities
  - Compute mean (or proportion of categorical dependent) variable for each possible two-way split
  - Select the split that optimizes goodness of fit
  - Do this for all variables
- Select the variable whether categorical or numeric with the best value of GOF measure

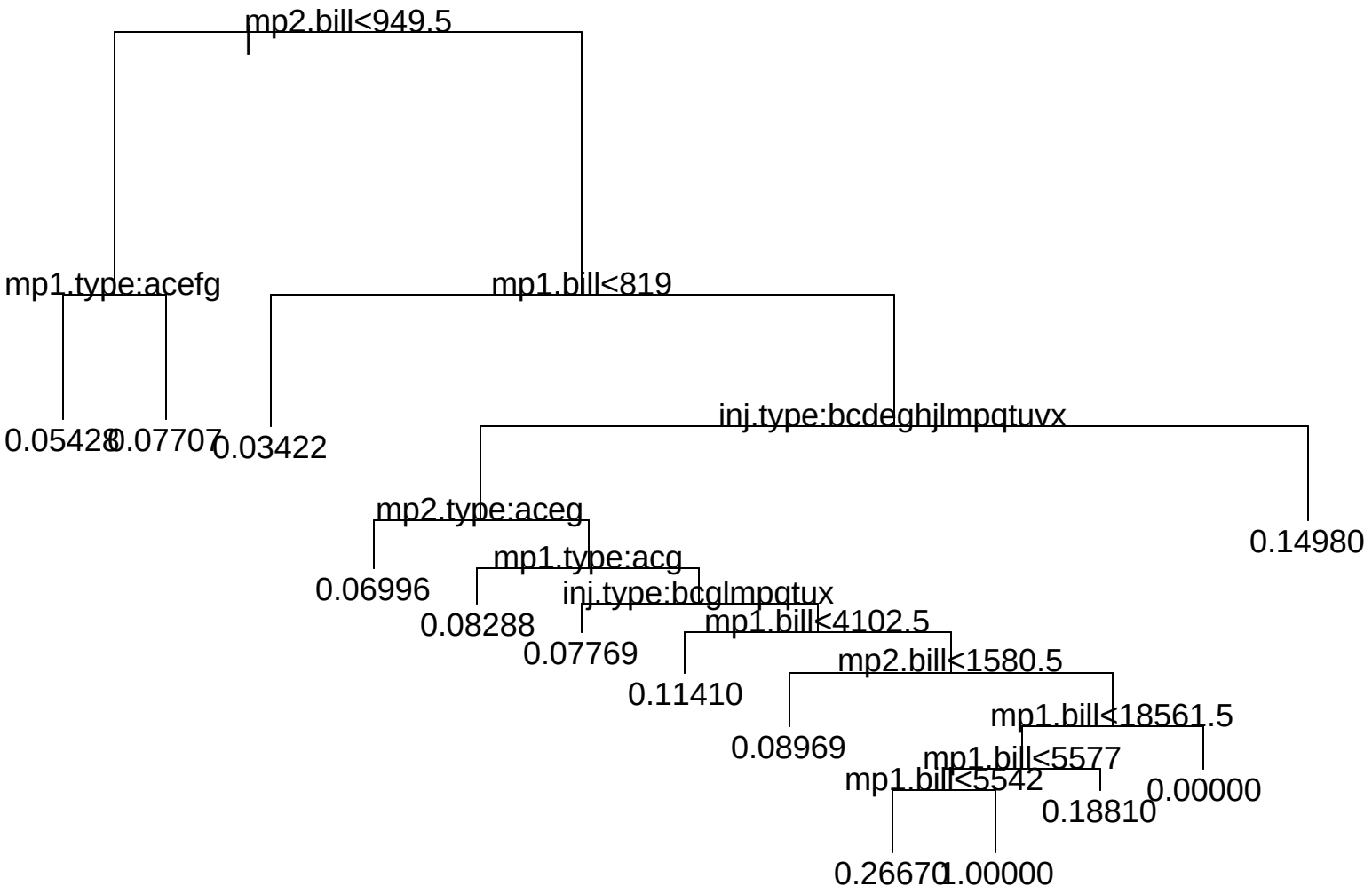
Recursive Partitioning: Categorical Variables



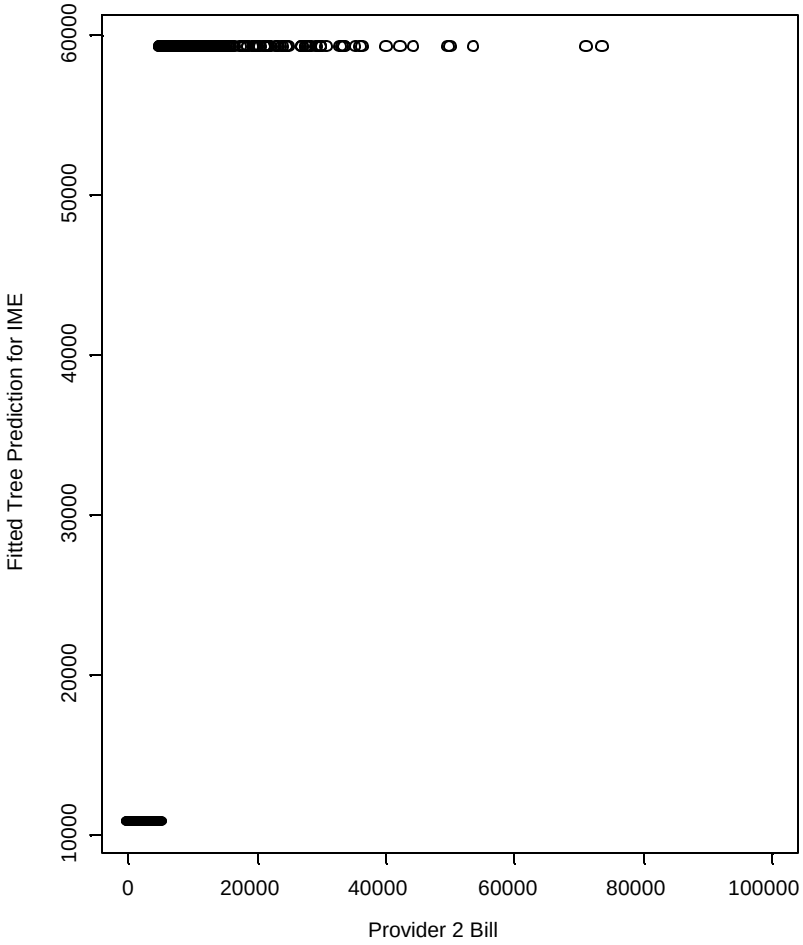
Decision Trees Cont.

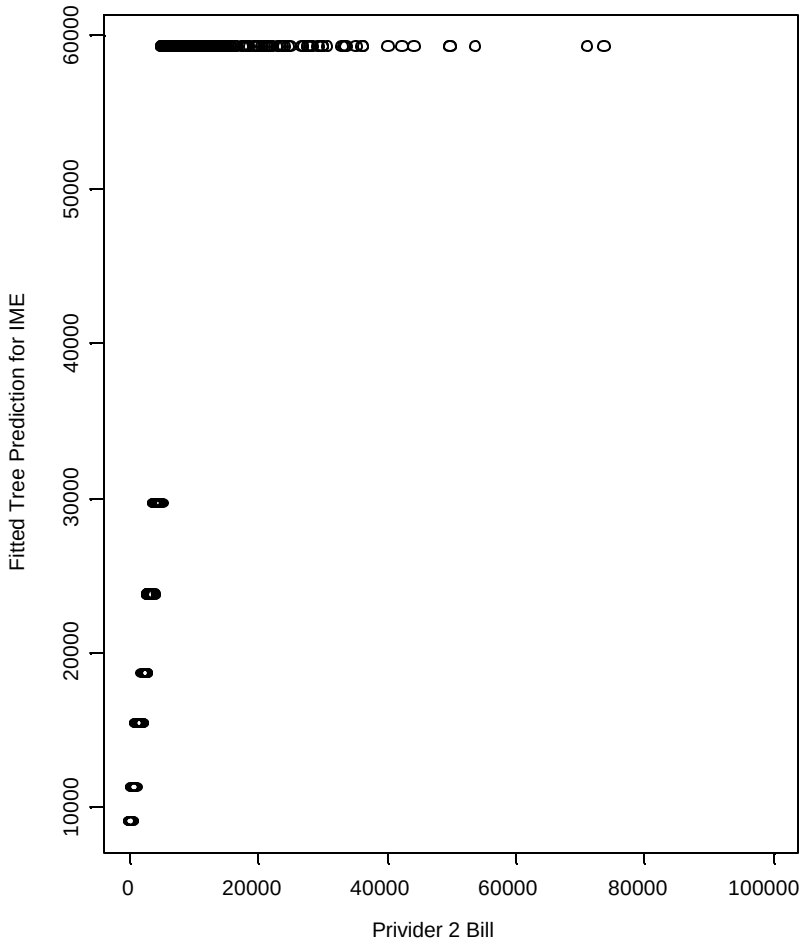
- After splitting data on first node, then
  - Go to each child node
  - Perform same process at each node, i.e.
  - Examine variables one at a time for best split
  - Select best variable to split on
  - Can split on different variables at the different child nodes

Tree Example



CART


Fitted Tree Prediction for IME
10000
20000
30000
40000
50000
60000
0
20000
40000
60000
80000
100000
Provider 2 Bill


Fitted Tree Prediction for IME
10000
20000
30000
40000
50000
60000
0
20000
40000
60000
80000
100000
Privider 2 Bill

Measures for Splitting

- Gini index for a node **$p(1-p)$**
  - where p = relative frequency of defectors
- Entropy for a node **$-Sp\log p$**
  - $-[p*\log(p) + (1-p)*\log(1-p)]$
  - Max entropy/Gini when p=.5
  - Min entropy/Gini when p=0 or 1
  - Similar to deviance under binomial assumption
- Gini might produce *small* but pure nodes
- The "twoing" rule strikes a balance between *purity* and creating roughly *equal-sized nodes*

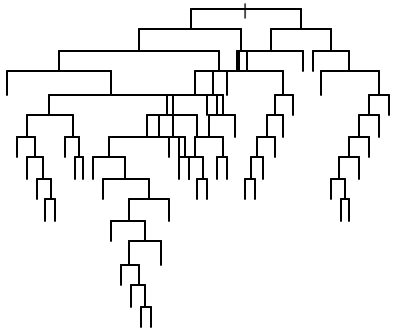
$$\frac{p_L p_R}{4} \sum_j (p(j \mid t_L) - p(j \mid t_R))^2$$

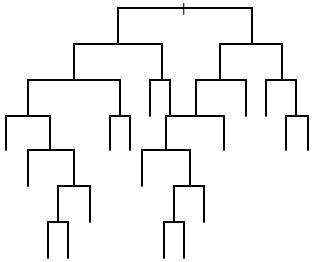
How CART Selects the Best Tree

- Use **cross-validation** (CV) to select the optimal decision tree.
- Use **validation sample** (say one third of data)
- Built into the algorithm.
  - Essential to the method; not an add-on
- Basic idea: "grow the tree" out as far as you can…. Then "prune back".
- CV: tells you when to stop pruning.

Growing & Pruning

- One approach: stop growing the tree early.
  - But how do you know when to stop?
- CART: just grow the tree all the way out; then prune back.
  - Sequentially collapse nodes that result in the smallest change in purity.
  - “weakest link” pruning.





CART: Missing Data

- Find a surrogate variable: Find the variable with the second best goodness of fit for the split at a given node.
- If not missing for the record, use it or
- Find variable with the third best goodness of fit for the split, etc.
- Common alternative: Code missing as a valid value

Weakest-Link Pruning

- Sequentially collapse nodes that result in the smallest change in goodness of fit.
- This gives a nested sequence of trees that are all sub-trees of T_0.

$$T_0 \gg T_1 \gg T_2 \gg T_3 \gg \ldots \gg T_k \gg \ldots$$

- Theorem: the sub-tree T_a of T_0 that minimizes a risk measure is in this sequence
  - Gives us a simple strategy for finding best tree
  - Often simple risk measure like 1 standard deviation

CHAID

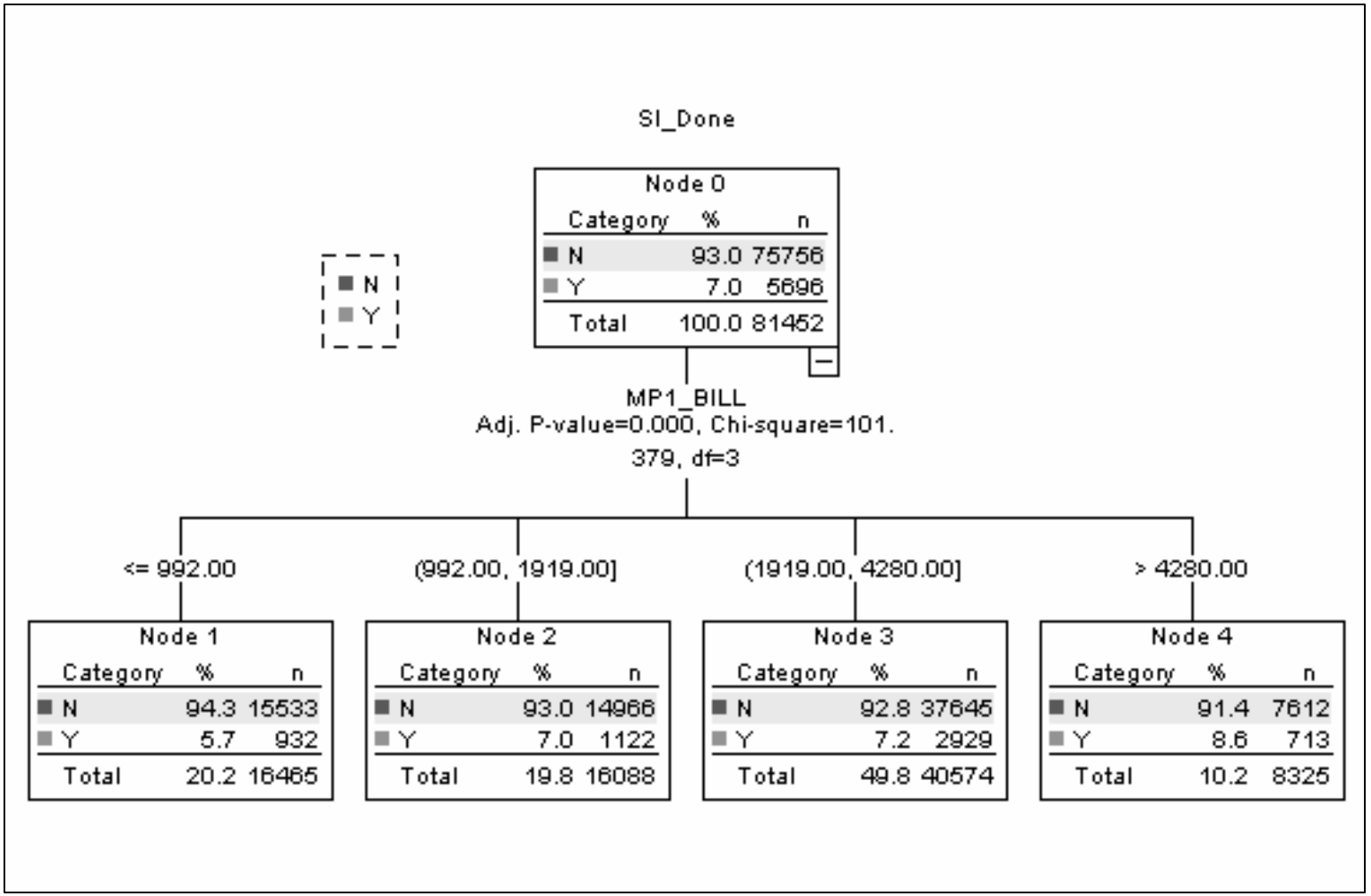
CHAID

- The goodness of fit measure is the Chi Square (F-statistic when the dependent variable is continuous)

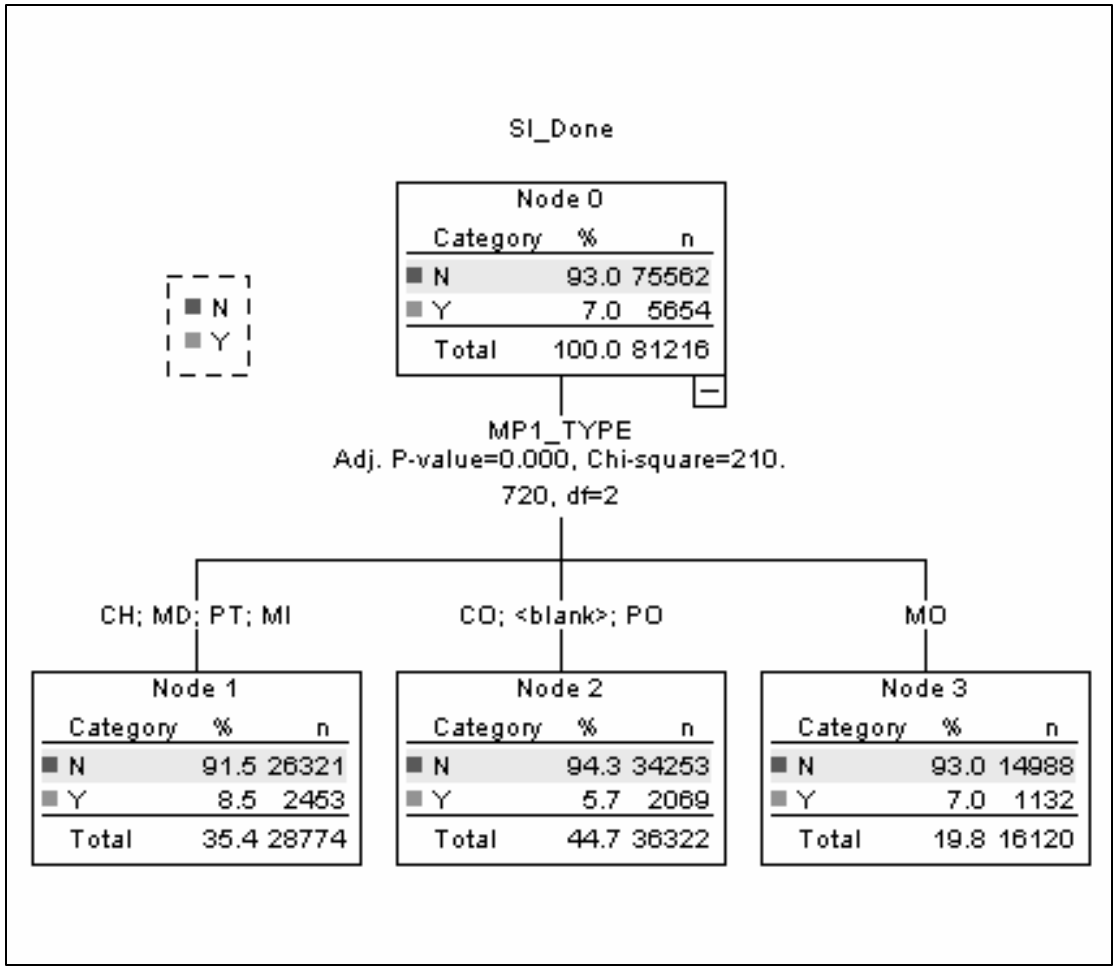
- Splits do not need to be binary

$$\mathrm{X}^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}$$, j=category, O is Observed, E is Expected

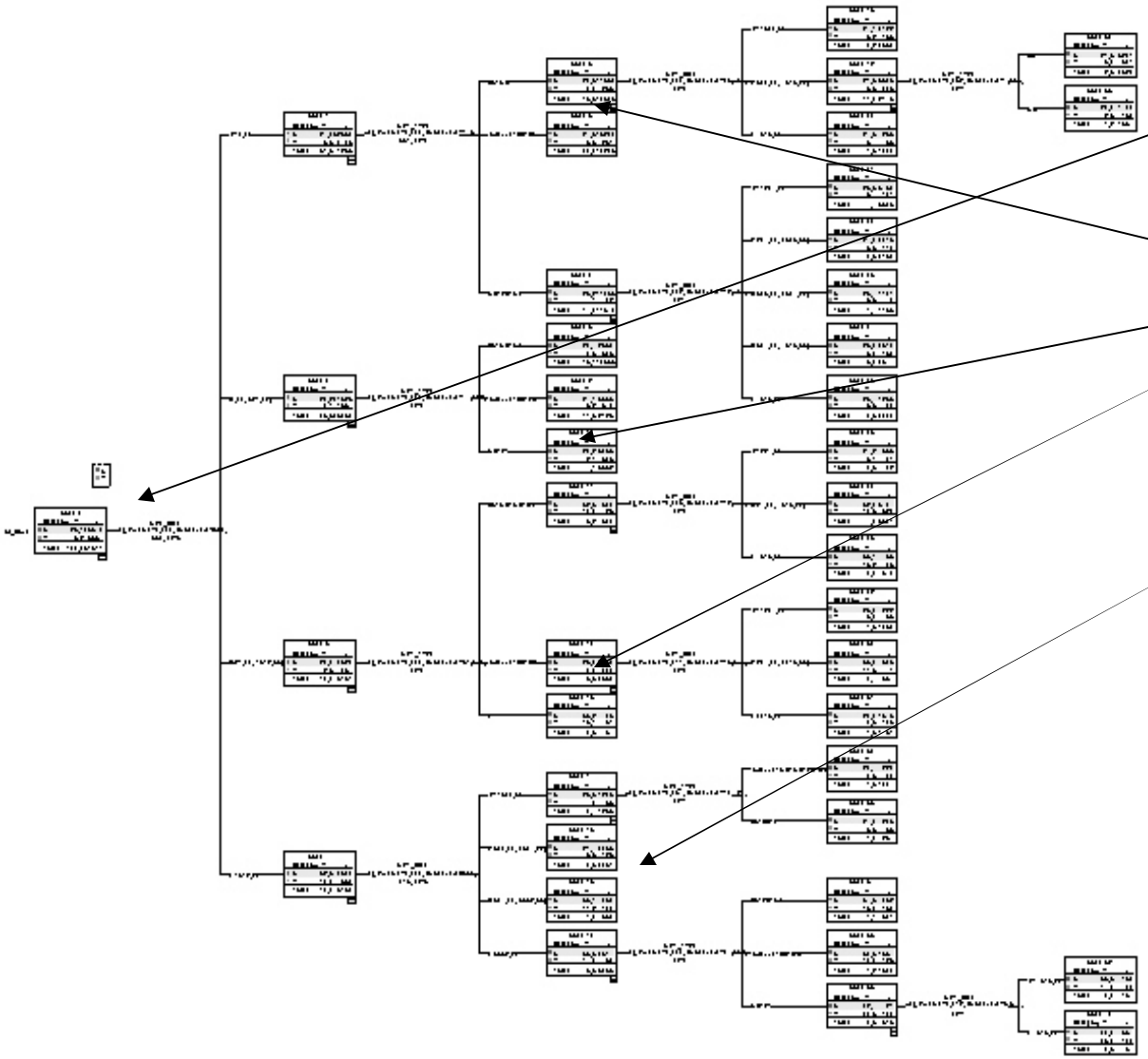
CHAID Tree: Numeric Predictor



CHAID Tree: Categorical Predictor



Full CHAID Model



- Prov 2 Bill
- Prov 1 Type
- Prov 1 Bill

CHAID categories

- CHAID uses either the Chi-Squared (categorical dependant) or F-Test (numeric dependant) to combine categories that are not significantly different from each other

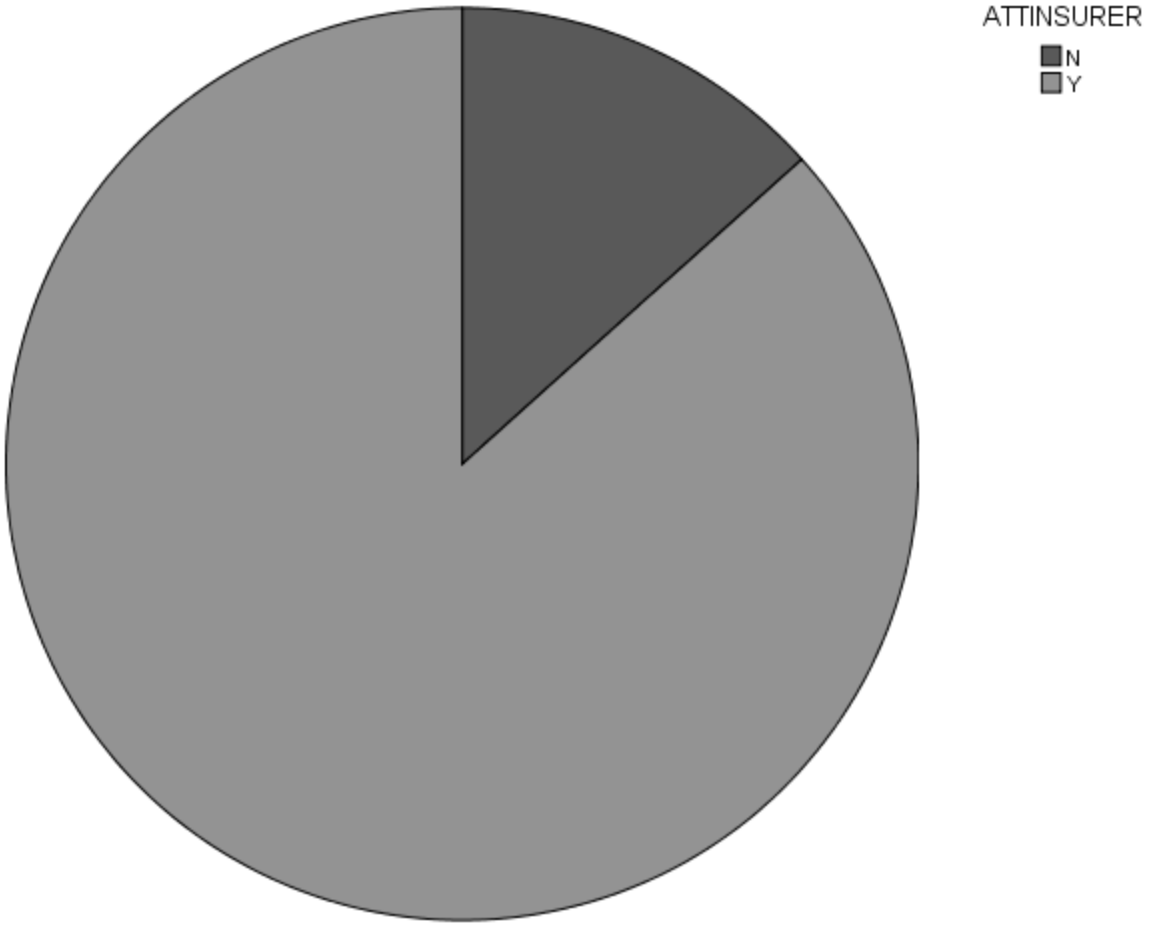
CHAID Example from Texas WC data

- Dependant variable
  - Primary paid losses
  - Whether claim is 1998 or 2005 data base
- Predictors
  - Initial reserve
  - Report lag
  - Injury type
  - Attorney involvement

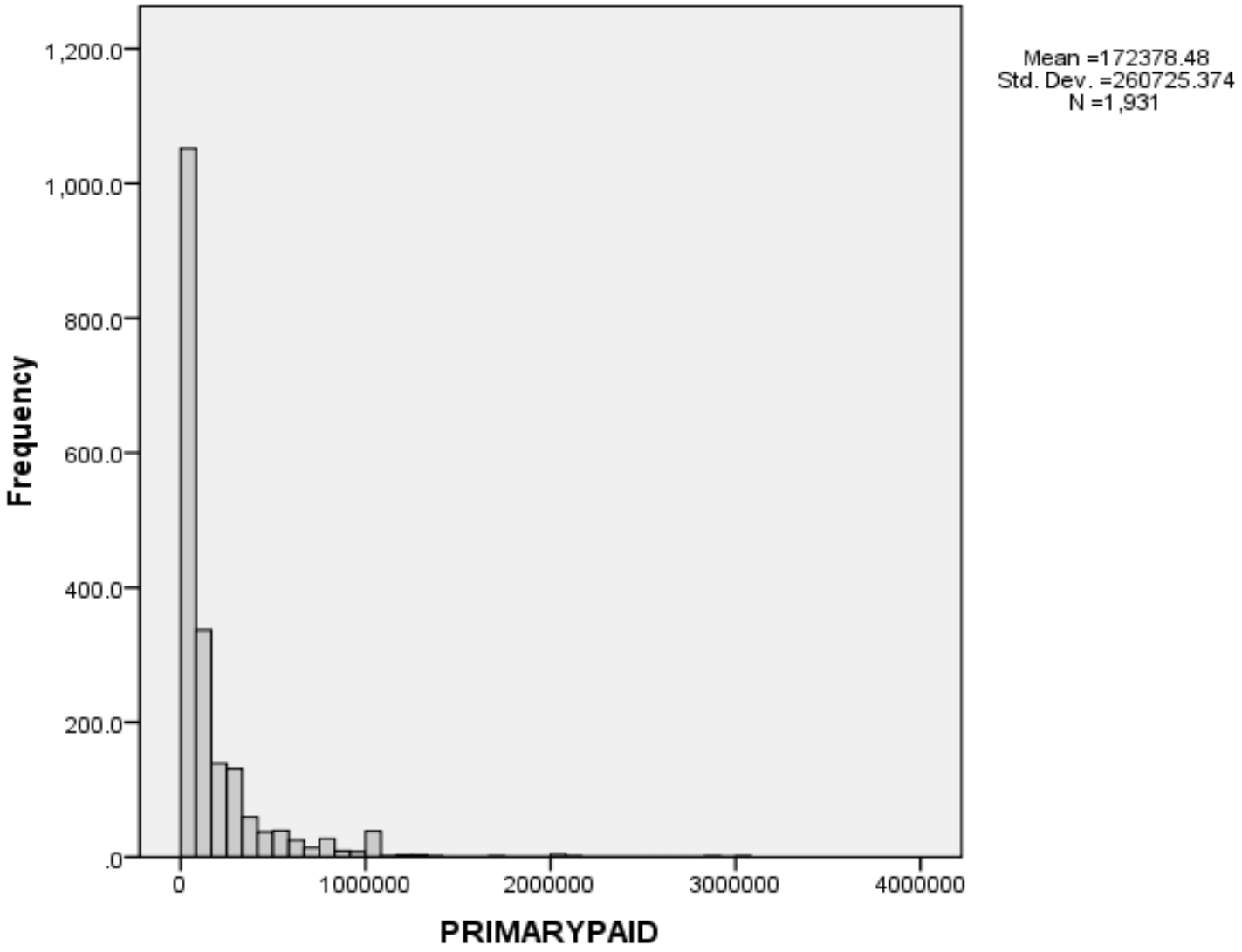
Two Types of Dependent Variables

- Categorical
  - Fraud/No Fraud
  - Lawyer/No Lawyer
  - Simple/Complex Claim
- Numeric/Continuous
  - Paid Losses
  - Economic Damages
  - Claim Severity

Categorical - Lawyer



Continuous – Primary Paid



Distribution of Primary Paid

Statistics

PRIMARYPAID

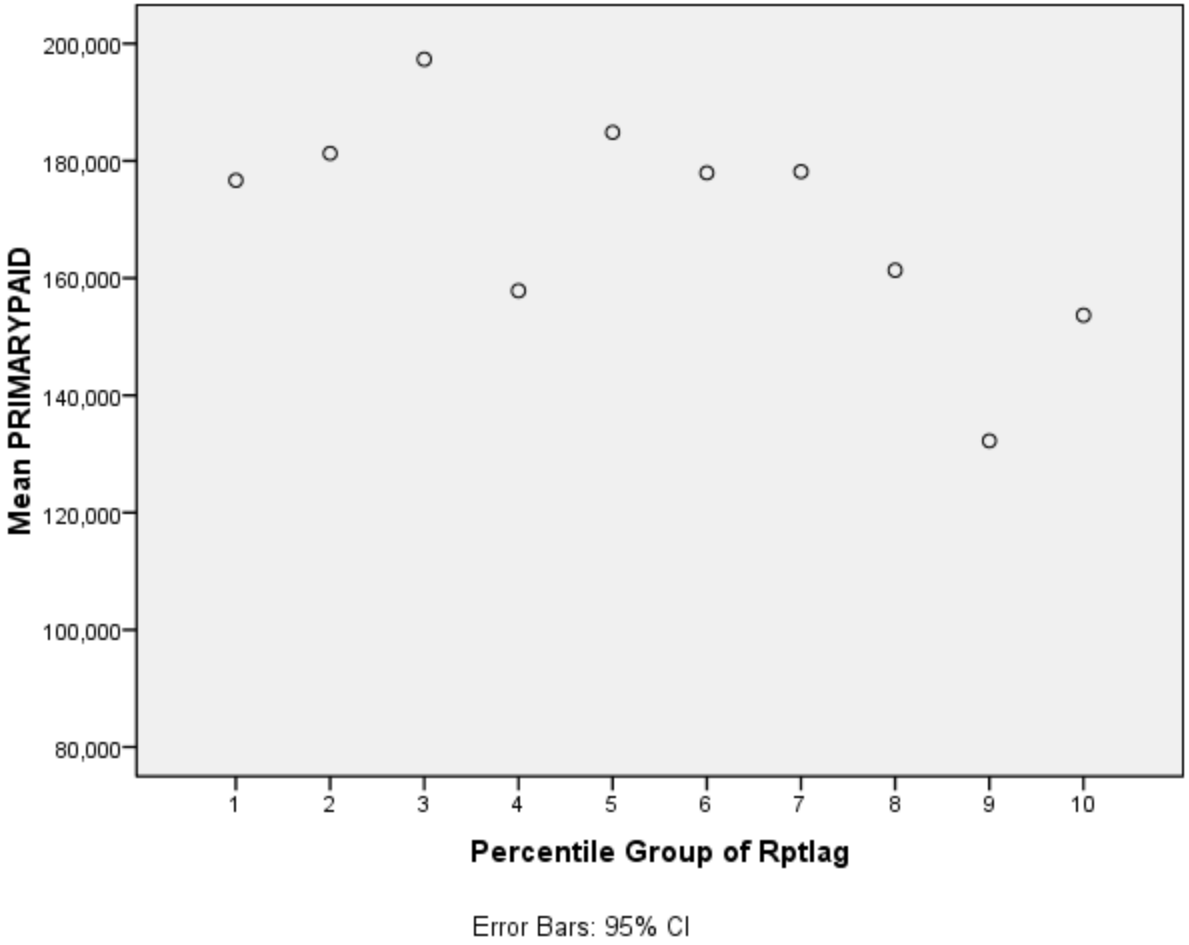
N	Valid	1,931
	Missing	0
Mean		172,378
Minimum		0
Maximu		3,000,000
Percentiles	10	25,000
	20	30,000
	30	40,000
	40	50,000
	50	75,000
	60	100,000
	70	150,000
	80	250,000
	90	467,840

Distribution of Predictors

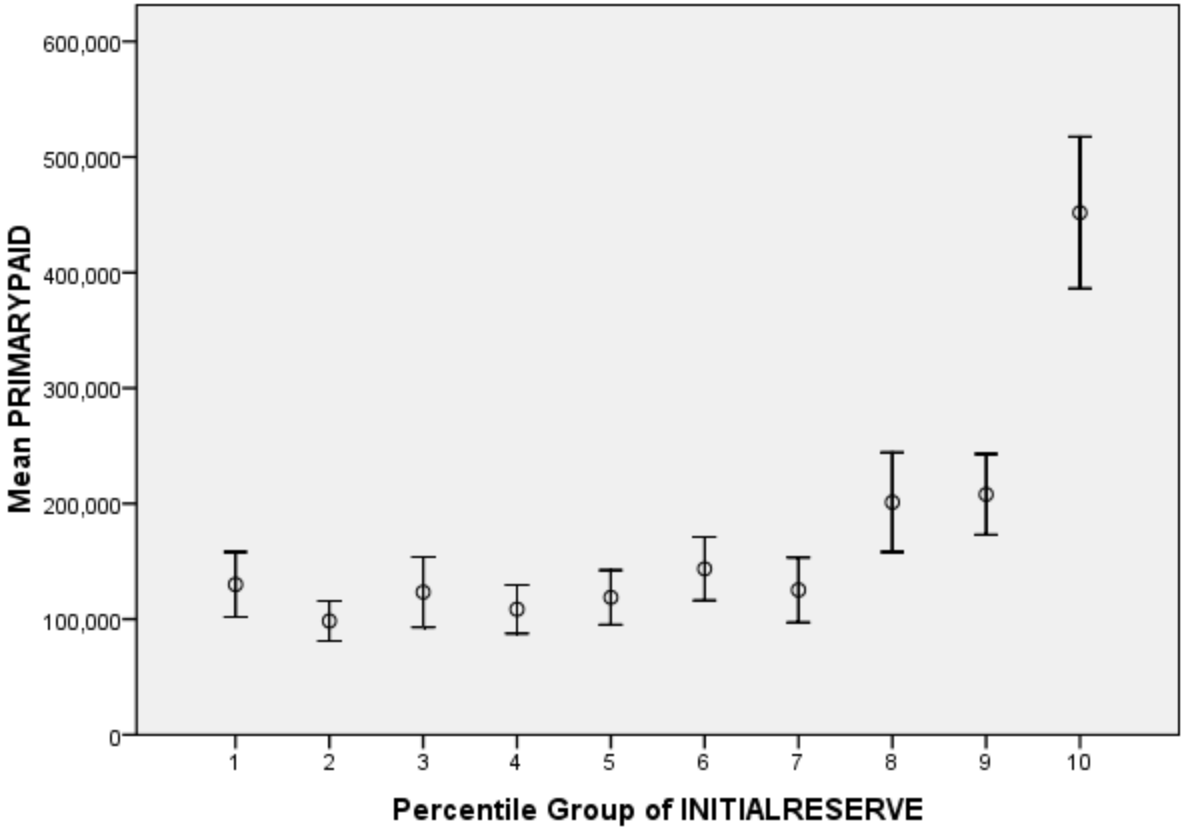
Statistics

		INITIALRESERVE	ReportLag
N	Valid	1931	1931
	Missing	0	0
Mean		74,125.97	315.6189
Minimu		0	.00
Maximu		2,000,000	18781.00
Percentiles	10	3,500.00	1.0000
	20	6,500.00	2.0000
	30	10,000.00	6.0000
	40	15,000.00	21.0000
	50	22,000.00	66.0000
	60	32,260.00	174.0000
	70	50,000.00	330.4000
	80	80,000.00	553.2000
	90	169,850.00	744.0000

Grouping Variables



Are Initial Reserves and Payment Correlated?



Error Bars: 95% CI

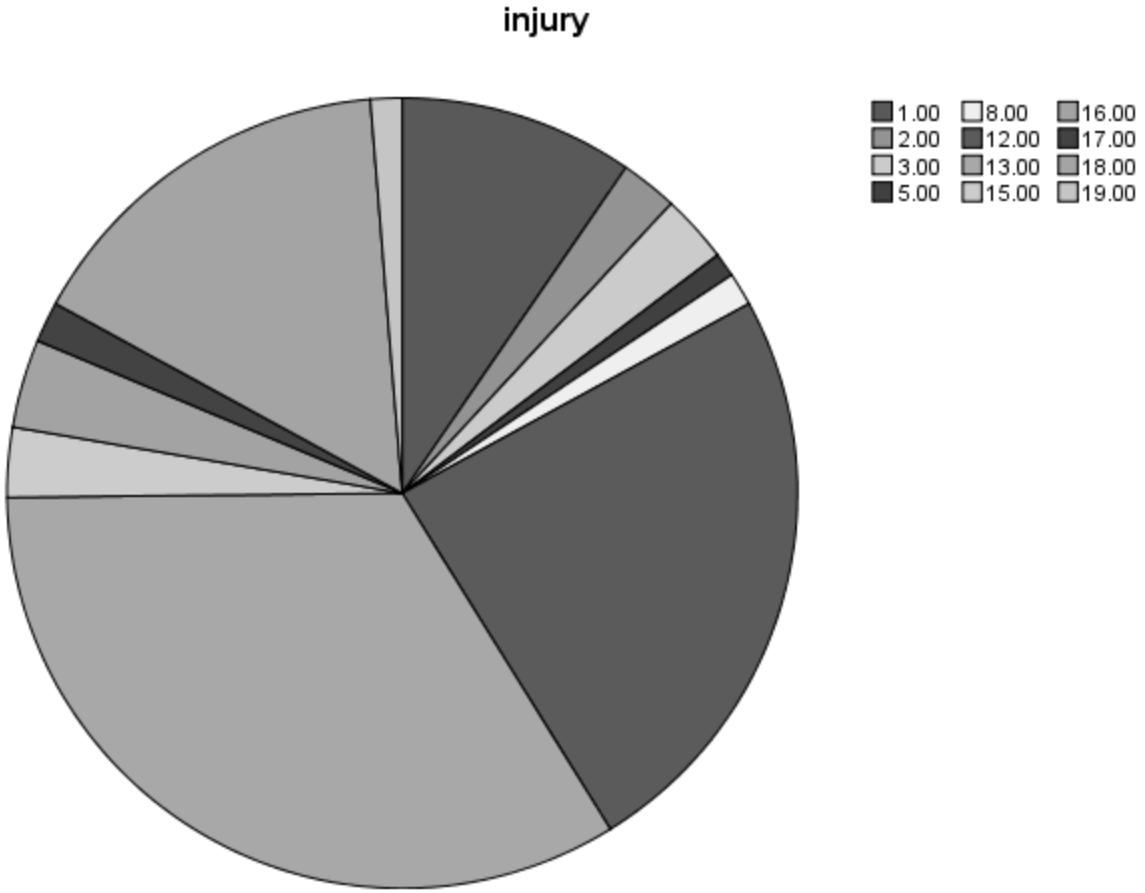
T-Test for Significant Differences

$$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{s_{diff}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_1}{n_2}}}$$

$$df = smaller(n_1 - 1, n_2 - 1)$$

Injury Type

- Largest: Multiple,
- Back and Spinal



Use T-Statistic to Combine Categories

Inj	Injury	Average of PAID	Count	StdDev	Diff	SD(Dfff)	df	T	Significance
8	Respiratorycon	$90,208	26	125,151	8,690	35,733	18	0.24	0.4053
5	POISON1	$98,898	19	113,199	21,499	27,974	18	0.77	0.2261
18	Other	$120,397	304	181,313	9,745	12,901	303	0.76	0.2253
12	Multipleinjuries	$130,141	466	164,813	1,815	21,884	44	0.08	0.4671
2	Amputation	$131,956	45	137,579	8,576	21,920	44	0.39	0.3487
13	Backinjury	$140,532	650	197,256	42,430	26,873	69	1.58	0.0595
16	Scarring	$182,962	70	215,313	11,855	47,806	53	0.25	0.4025
3	Burn	$194,818	54	296,052	36,404	73,181	24	0.50	0.3117
19	Small	$231,222	25	305,463	64,147	73,440	24	0.87	0.1955
15	Braindamage	$295,369	55	302,267	112,664	53,758	54	2.10	0.0204
1	Death	$408,033	185	476,777	6,610	102,901	31	0.06	0.4746
17	Spinalcordinj	$414,643	32	547,283					
Grand Total		$172,378	1931	260,725					

Category Stats After Combination

Inj	Injury	Average of PAID	Count	StdDev	Diff	SD(Dfff)	df	T	Significance
8	Respirator	$90,208	26	125,151	8,690	35,733	18	0.2432	0.4053
5	POISON1	$98,898	19	113,199	21,499	27,974	18	0.7685	0.2261
18	Other	$120,397	304	181,313	9,745	12,901	303	0.7554	0.2253
12	Multipleinj	$130,141	466	164,813	1,815	21,884	44	0.0829	0.4671
2	Amputatio	$131,956	45	137,579	8,576	21,920	44	0.3913	0.3487
13	Backinjury	$140,532	650	197,256	42,430	26,873	69	1.5789	0.0595
16	Scarring	$182,962	70	215,313	11,855	47,806	53	0.248	0.4025
3	Burn	$194,818	54	296,052	36,404	73,181	24	0.4975	0.3117
19	Small	$231,222	25	305,463	64,147	73,440	24	0.8735	0.1955
15	Braindama	$295,369	55	302,267	86,845	49,785	54	1.7444	0.0434
1	**Death/Spi**	**$382,214**	**240**	**442,901**					
Grand Total		$172,378	1931	260,725					

More Than Two Categories

- Use F-Test instead of T-Test
- With More than 2 categories, we refer to it as an Analysis of Variance (ANOVA)

Example with Categorical Dependent

- Two data sets were combined to create database
- 1998 closed claims and 2005 closed claims
- Are there significant differences between data from the different dates?
- Use trees to answer this question
- Dependent variable is the database and is binary

Data with Close Year as Dependent Variable

code	Injury	1998	2005	Grand Total	Cell %		Row %
1	Death	85	100	185	45.9%	54.1%	9.6%
12	Multipleinjuries	237	229	466	50.9%	49.1%	24.1%
2	Amputation	23	22	45	51.1%	48.9%	2.3%
19	Small	13	12	25	52.0%	48.0%	1.3%
13	Backinjury	362	288	650	55.7%	44.3%	33.7%
18	Other	180	124	304	59.2%	40.8%	15.7%
17	Spinalcordinjuri	19	13	32	59.4%	40.6%	1.7%
16	Scarring	43	27	70	61.4%	38.6%	3.6%
15	Braindamage	35	20	55	63.6%	36.4%	2.8%
3	BURNHEAT	35	19	54	64.8%	35.2%	2.8%
8	Respiratorycon	19	7	26	73.1%	26.9%	1.3%
5	POISON1	15	4	19	78.9%	21.1%	1.0%
Grand Total		1066	865	1931			
Col %		55.2%	44.8%				

Compute Expected Counts

Expected

code	Injury	1998	2005	Grand Total	*Row %*	
1	Death	102.1	82.9	185.0	55.2%	44.8%
12	Multipleinjuries	257.3	208.7	466.0	55.2%	44.8%
2	Amputation	24.8	20.2	45.0	55.2%	44.8%
19	Small	13.8	11.2	25.0	55.2%	44.8%
13	Backinjury	358.8	291.2	650.0	55.2%	44.8%
18	Other	167.8	136.2	304.0	55.2%	44.8%
17	Spinalcordinjuri	17.7	14.3	32.0	55.2%	44.8%
16	Scarring	38.6	31.4	70.0	55.2%	44.8%
15	Braindamage	30.4	24.6	55.0	55.2%	44.8%
3	BURNHEAT	29.8	24.2	54.0	55.2%	44.8%
8	Respiratorycon	14.4	11.6	26.0	55.2%	44.8%
5	POISON1	10.5	8.5	19.0	55.2%	44.8%
Grand Total		1066.0	865.0	1931.0		
Col %		55.2%	44.8%			

Chi Square Indicates Database Years Have Different Injury Mix

CHI Squared Statistic

Code	Injury	1998	2005	Grand Total
1	Death	2.9	3.5	6.4
12	Multipleinjuries	1.6	2.0	3.6
2	Amputation	0.1	0.2	0.3
19	Small	0.0	0.1	0.1
13	Backinjury	0.0	0.0	0.1
18	Other	0.9	1.1	2.0
17	Spinalcordinjuri	0.1	0.1	0.2
16	Scarring	0.5	0.6	1.1
15	Braindamage	0.7	0.9	1.6
3	BURNHEAT	0.9	1.1	2.0
8	Respiratorycon	1.5	1.9	3.4
5	POISON1	1.9	2.4	4.3
Grand Total		11.2	13.8	25.0
df	11.0			
Significance				**0.009038066**

USE CHAID to Combine Categories:

Actual

code	Injury	1998	2005	Grand Total	Cell Percent		Row %
18	Other	180	124	304	59.2%	40.8%	90.5%
17	Spinalcordinjuri	19	13	32	59.4%	40.6%	9.5%
Total		199	137	336			
Col %		0.5922619	0.4077				

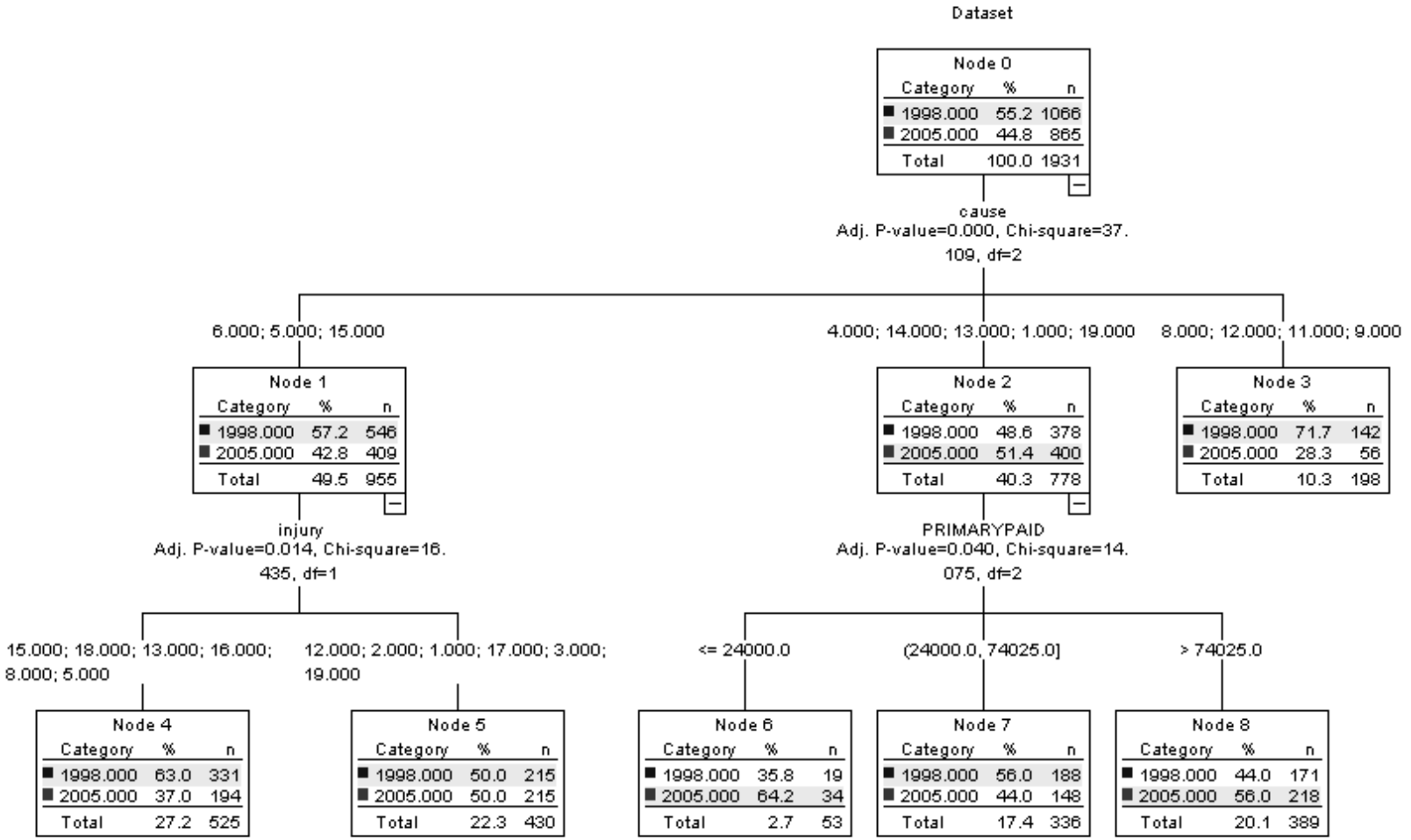
Expected

code	Injury	1998	2005	Grand Total
18	Other	180.05	123.95	304
17	Spinalcordinjuri	18.95	13.05	32

Chi Squared

code	Injury	1998	2005	Grand Total
18	Other	0.0	0.0	304
17	Spinalcordinjuri	0.0	0.0	32
Total		1.259E-05	0	**1.25943E-05**
df	1			
Not Significant				0.997168437

Put Other Predictors in and Get Full Tree



Exhaustive CHAID

- Technique uses statistical rules to stop splitting. No pruning.
- Improves on technique for splitting
- Seems to give more accurate predictions than classical CHAID

CHAID Reference

- Biggs, De Ville, and Suen
- “A Method of Choosing Multiway Partitions For Classification and Decision Trees”
- Journal of Applied Statistics. Vol 18, No 1, 1991

Continuous Dependent Variable

Primary Paid

Possible Numeric Predictors

- Initial reserves
  - Hypothesis: Higher reserves → Higher payment
- Report Lag
  - Hyp: Longer report lag -> Higher payment

Is There A Relationship Between Initial Reserves and Paid?

Report

Mean

Percentile Group of …	PRIMARYPAID
1	130,010
2	98,425
3	123,442
4	108,579
5	118,833
6	143,590
7	125,250
8	201,185
9	208,002
10	451,879
Total	170,306

Is Difference Significant?

ANOVA

PRIMARYPAID

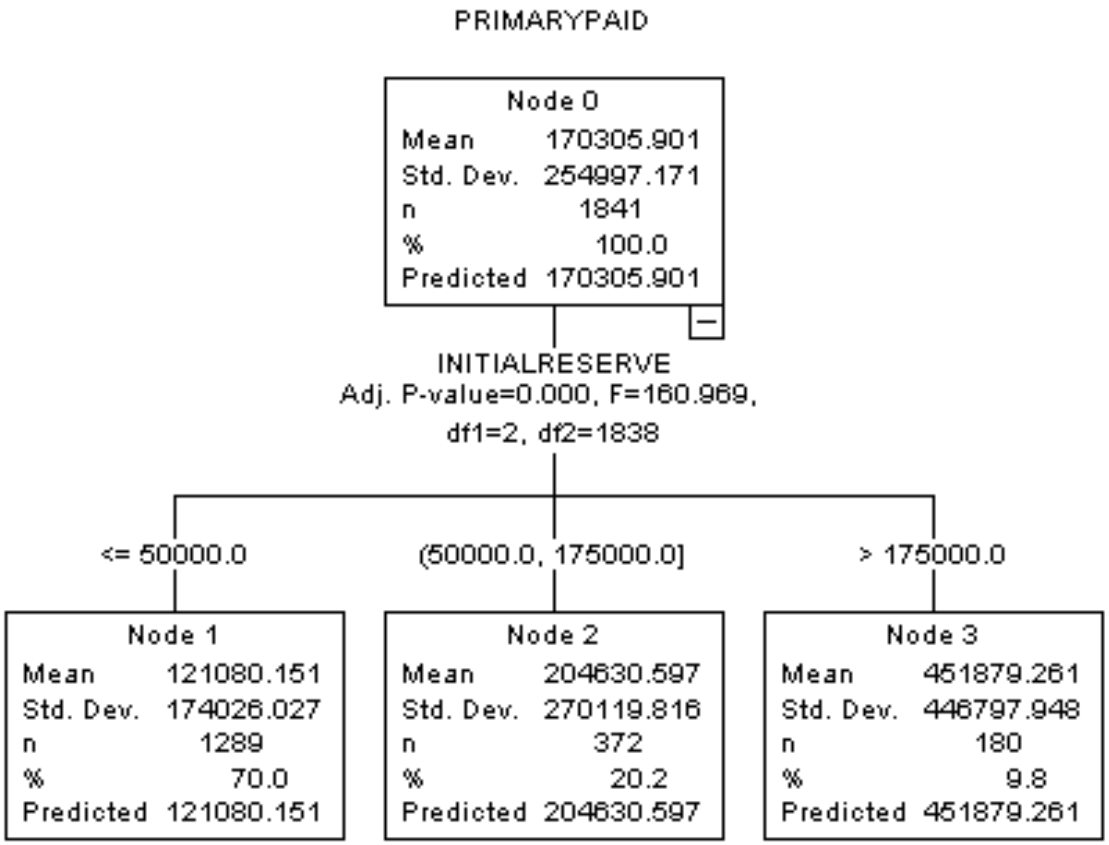
	Su Sq	d	Mean Sq	F	Sig.
Between Grou	5.817E11	9	6.464E10	.994	.443
Within Grou	1.191E14	1831	6.503E10		
Total	1.196E14	1840			

Is Difference Significant?

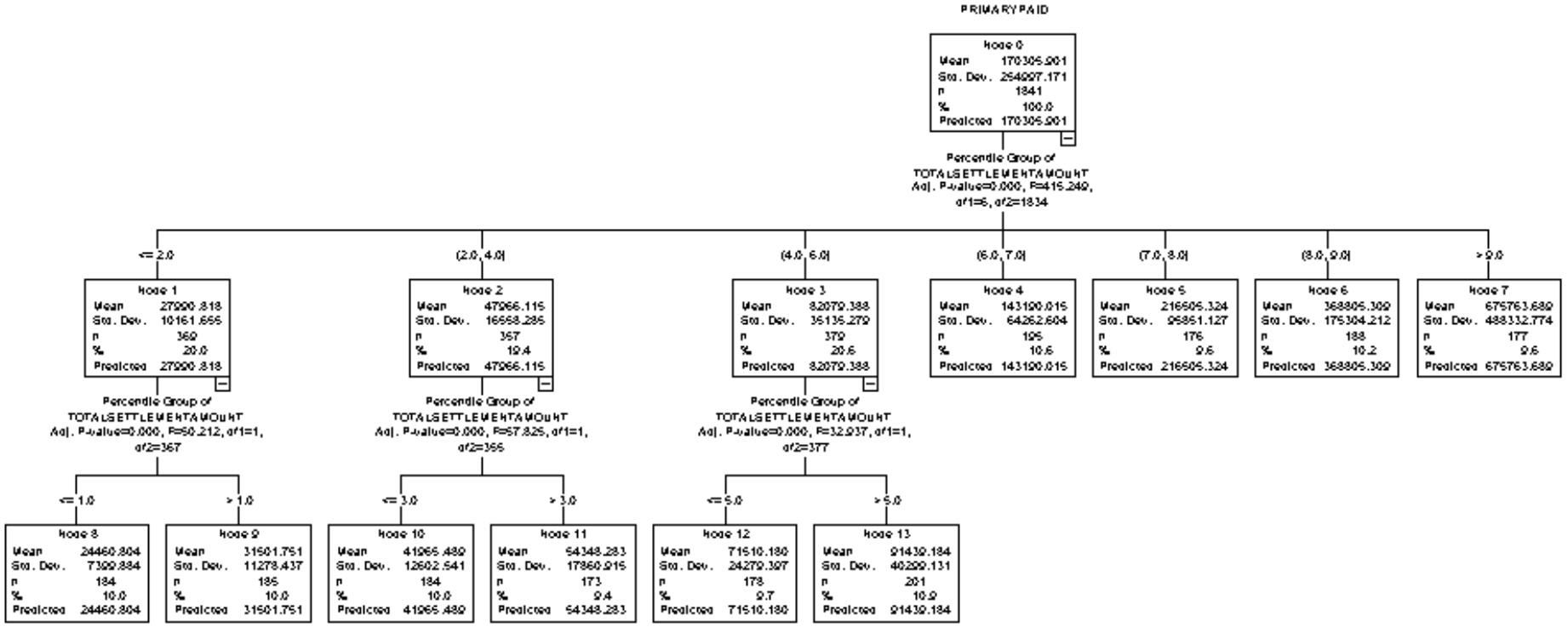
ANOVA Table

			Sum of Sq	df	Mean Sq	F	Sig.
PRIMARYPAID * Percentile Group of INITIALRESERVE	Between Grou	(Combined	1.810E13	9	2.011E12	36.269	.000
		Linearity	9.448E12	1	9.448E12	170.371	.000
		Deviation from Linearity	8.654E12	8	1.082E12	19.506	.000
		Within Grou	1.015E14	1831	5.546E10		
		Total	1.196E14	1840			

Us Trees to Get Best Grouping



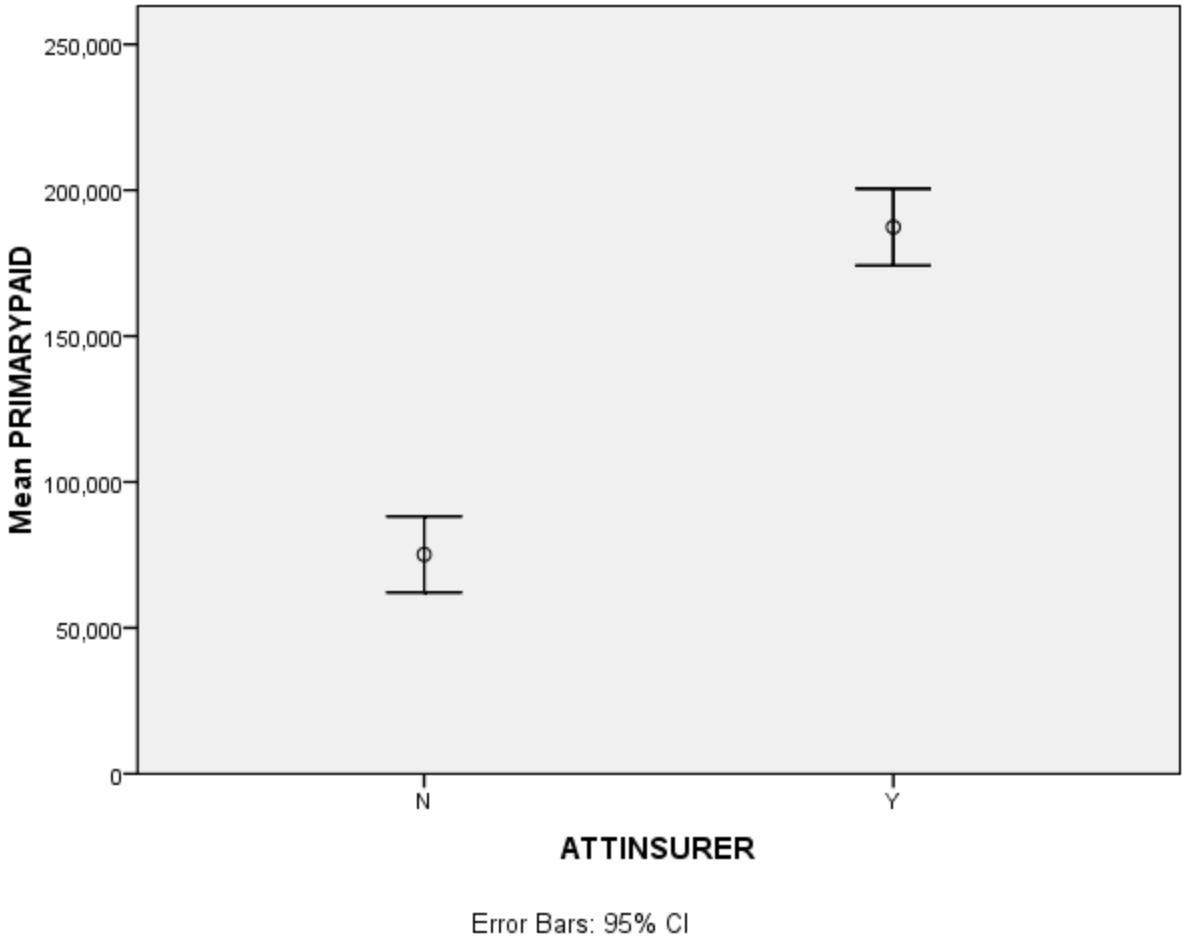
Use Trees to Group



Categorical Predictor

- Injury
- Cause of Loss
- Attorney Involvement
- County of Injury

Attorney Insurer



Primary Paid by Attorney Involvement

Report

PRIMARYPAID

ATTINSURER	Mean	N
N	75,141.16	258
Y	187,373.84	1673
Total	172,378.48	1931

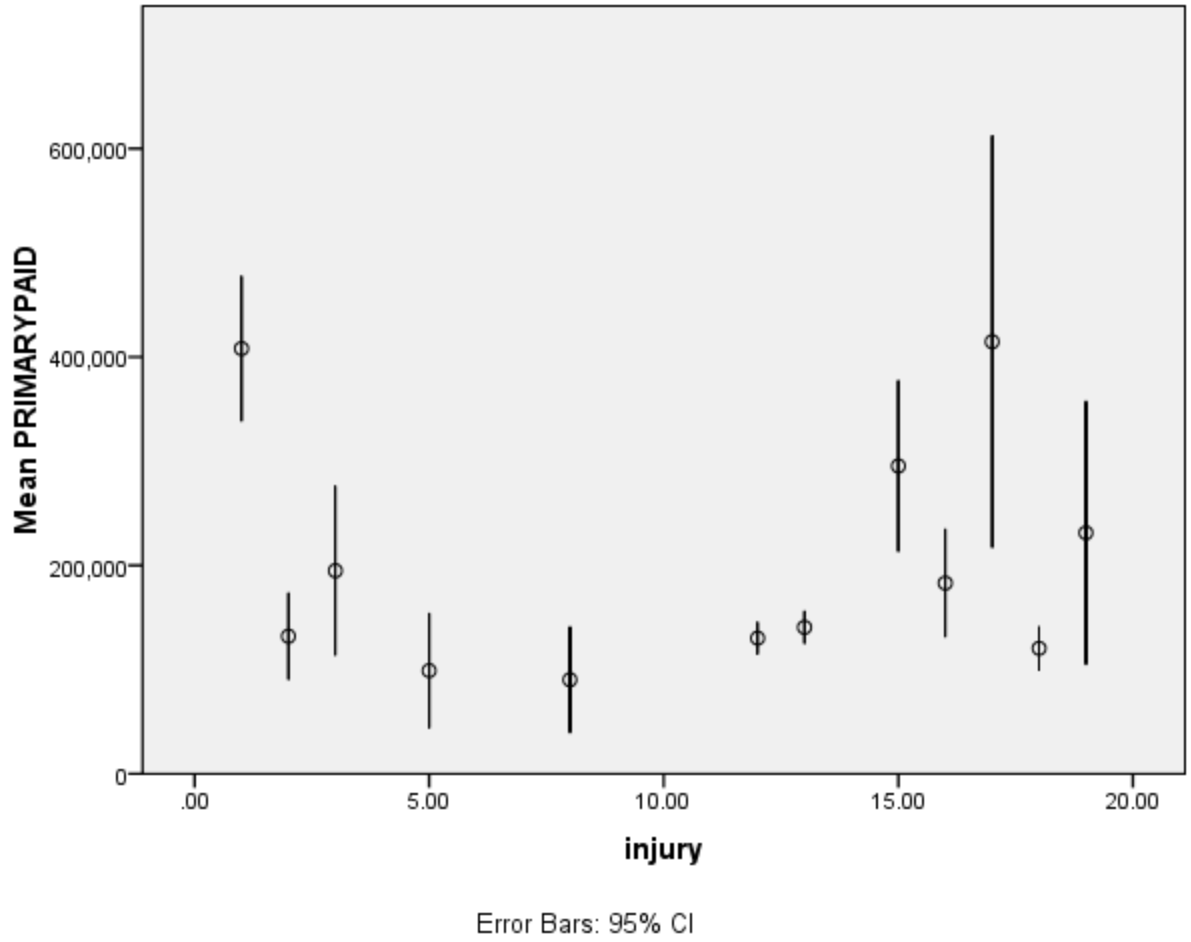
Significance Test

ANOVA Table[a]

		Su Sq	d	Mean Sq	F	Sig.
PRIMARYPAID * ATTINSURER	Between Grou	2.816E12	1	2.816E12	42.306	.000
		1.284E14	1929	6.655E10		
		1.312E14	1930			

a. The grou

Avg Paid by Injury Type



Using ANOVA to Estimate Relationship

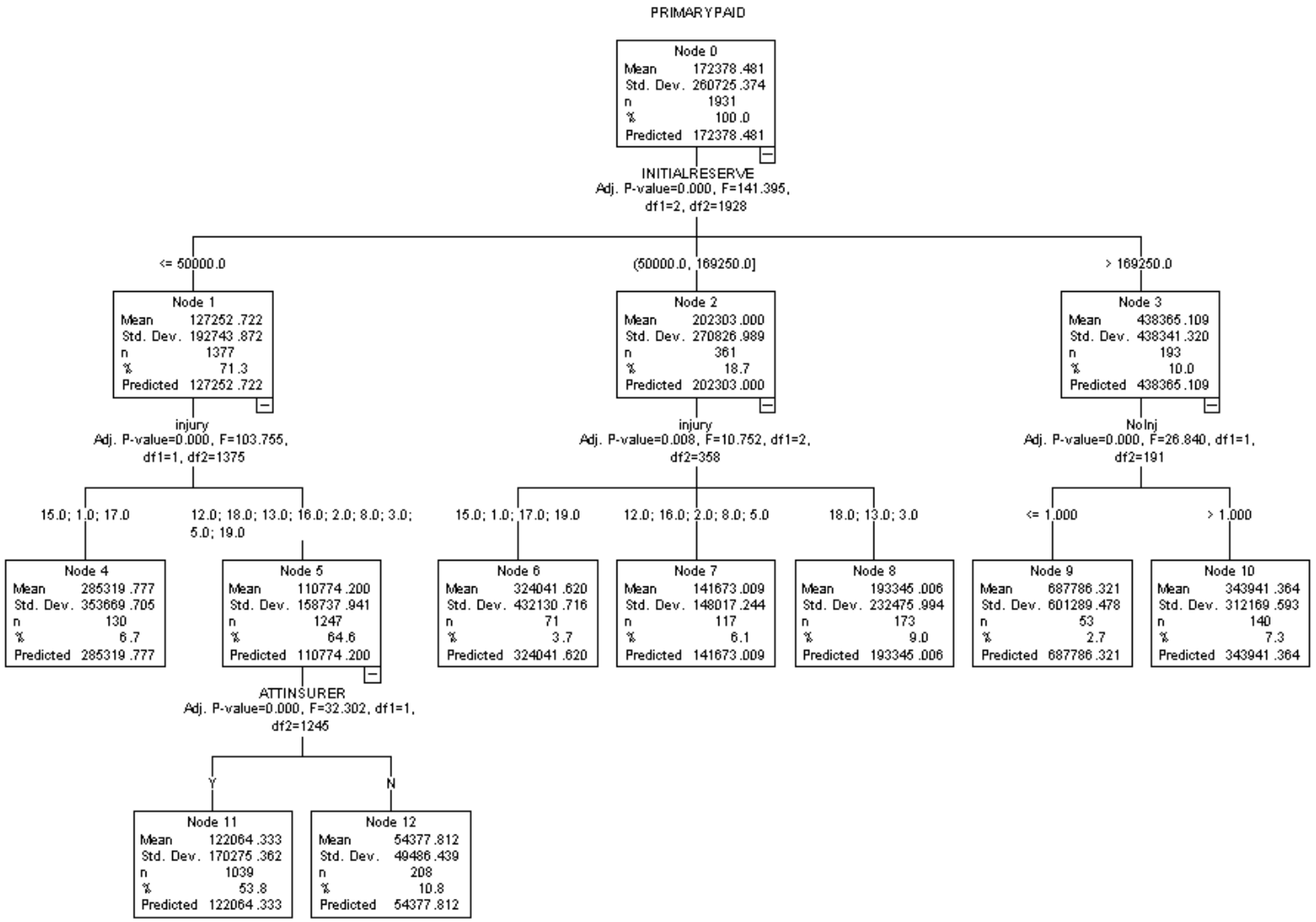
Parameter Estimates

Depend

Parameter	B	Std	t	Sig.	95% Confid		Partial Eta Sq
					Lower Bou	Upper Bou	
Intercept	231,221.800	49051.036	4.714	.000	135022.861	327420.739	.011
[inju	176,810.892	52260.309	3.383	.001	74317.925	279303.859	.006
[inju	-99,265.867	61177.391	-1.623	.105	-219247.023	20715.290	.001
[inju	-36,404.041	59328.703	-.614	.540	-152759.549	79951.467	.000
[inju	-132,323.853	74644.485	-1.773	.076	-278716.687	14068.982	.002
[inju	-141,014.146	68698.396	-2.053	.040	-275745.507	-6282.786	.002
[inju	-101,080.566	50349.594	-2.008	.045	-199826.238	-2334.894	.002
[inju	-90,689.417	49985.426	-1.814	.070	-188720.881	7342.047	.002
[inju	64,146.709	59157.775	1.084	.278	-51873.577	180166.995	.001
[inju	-48,259.386	57142.728	-.845	.398	-160327.758	63808.987	.000
[inju	183,420.919	65465.219	2.802	.005	55030.469	311811.369	.004
[inju	-110,825.254	51028.094	-2.172	.030	-210901.600	-10748.908	.002
[inju	0.000[a]	.	.	.	.	.	.

a. This parameter is set to zero becau

Full Tree for Primary Paid



Ensemble Trees: Fit More Than One Tree

- Fit a series of trees
- Each tree added improves the fit of the model
- Average or Sum the results of the fits
- There are many methods to fit the trees and prevent overfitting
    - Boosting: Iminer Ensemble and Treenet
    - Bagging: Random Forest

Free/ Cheap Software for Trees

- R
  - Go to www.r-project.org
  - Download R
  - Download pdf files with manuals
  - Buy book
- WEKA
  - Developed by computer scientists
- XlMiner – Excel add-in

The Data

- Texas closed claim data used for some illustrations can be found at www.data-mines.com

Introductory Modeling Library Recommendations

- *Data Analysis and Graphics Using R: An Example Based Approach,* Maindonald and Braun, Cambridge University Press
- *Data Mining for Business Intelligence, Concepts, Applications and Techniques in Microsoft Office Excel with XLMiner,*Shmueli, Patel and Bruce, Wiley 2007
- *Iterative Approach to Classification Analysis*, Fish, Gallagher, Howard, 1990 CAS Discussion Paper Program, www.casact.org