Distinguishing the Forest from the Trees
2008 CAS Fall Meeting

Richard Derrig, PhD,
Opal Consulting
www.opalconsulting.com
Louise Francis, FCAS, MAAA
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com

Data Mining

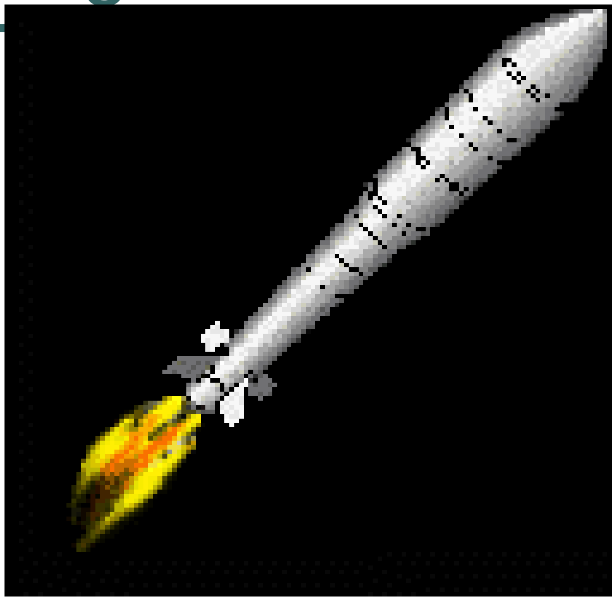
- **Data Mining**, also known as **Knowledge-Discovery in Databases (KDD)**, is the process of automatically searching large volumes of data for patterns. In order to achieve this, data mining uses computational techniques from statistics, machine learning and pattern recognition.
  - **www.wikipedia.org**

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

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



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

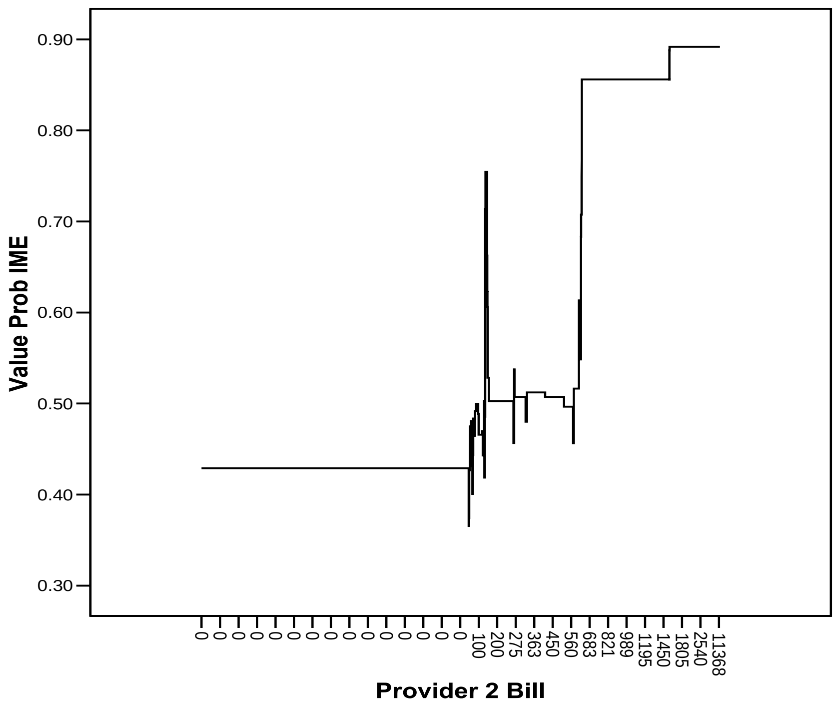
Desirable Features of a Data Mining Method:

- 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

Nonlinear Example Data

Provider 2 Bill (Binned)	Avg Provider 2 Bill	Avg Total Paid	Percent IME
Zero	-	9,063	6%
1 – 250	154	8,761	8%
251 – 500	375	9,726	9%
501 – 1,000	731	11,469	10%
1,001 – 1,500	1,243	14,998	13%
1,501 – 2,500	1,915	17,289	14%
2,501 – 5,000	3,300	23,994	15%
5,001 – 10,000	6,720	47,728	15%
10,001 +	21,350	83,261	15%
All Claims	545	11,224	8%

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



The Fraud Surrogates used as Dependent Variables

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

Predictor Variables

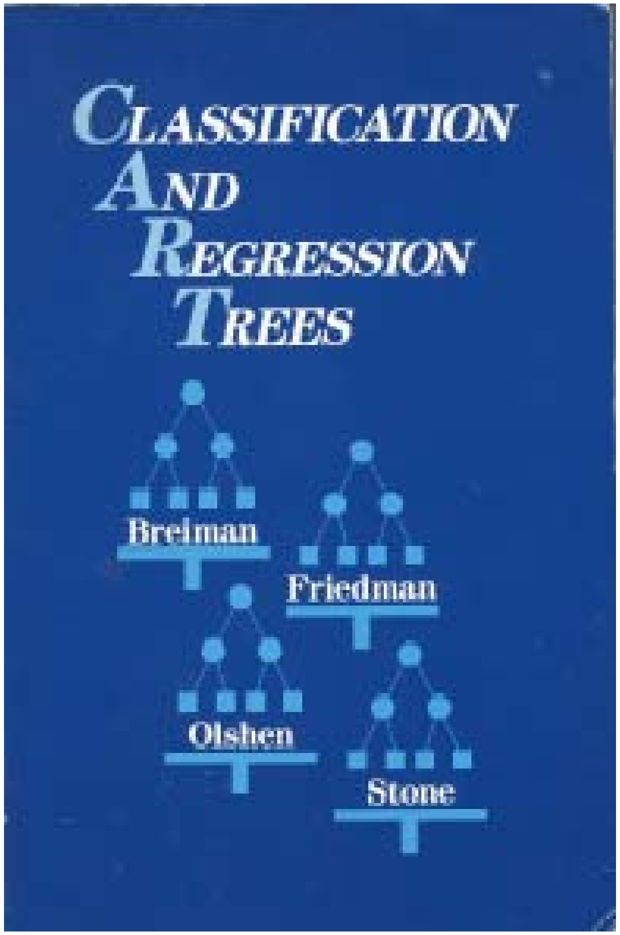
- Claim file variables
  - Provider bill, Provider type
  - Injury
- Derived from claim file variables
  - Attorneys per zip code
  - Docs per zip code
- Using external data
  - Average household income
  - Households per zip

Decision Trees

- In decision theory (for example risk management), a **decision tree** is a graph of decisions and their possible consequences, (including resource costs and risks) used to create a plan to reach a goal. Decision trees are constructed in order to help with making decisions. A decision tree is a special form of tree structure.
  - **www.wikipedia.org**

The Classic Reference on Trees

Brieman, Friedman Olshen and Stone, 1993



Regression Trees

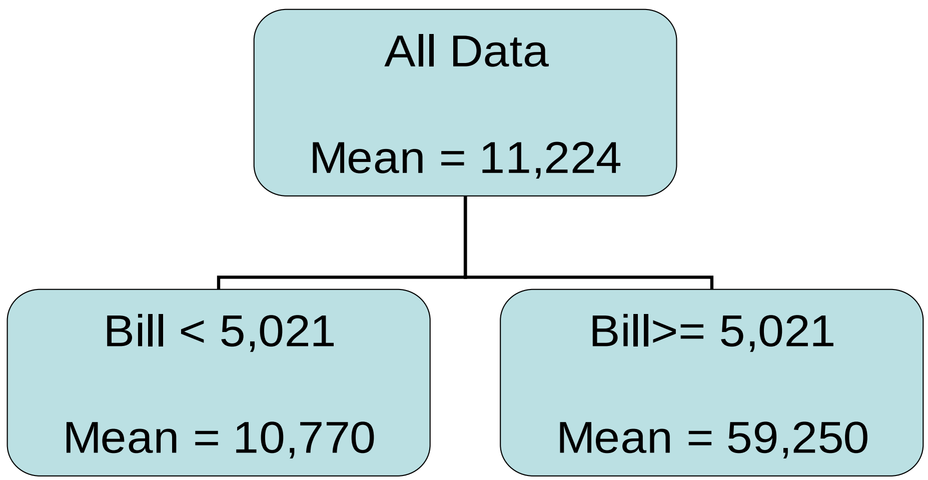
- Tree-based modeling for ***continuous* target variable**
  - most intuitively appropriate method for loss ratio analysis
- Find split that produces greatest separation in

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

- 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*

CART Example of Parent and Children Nodes
Total Paid as a Function of Provider 2 Bill

1st Split



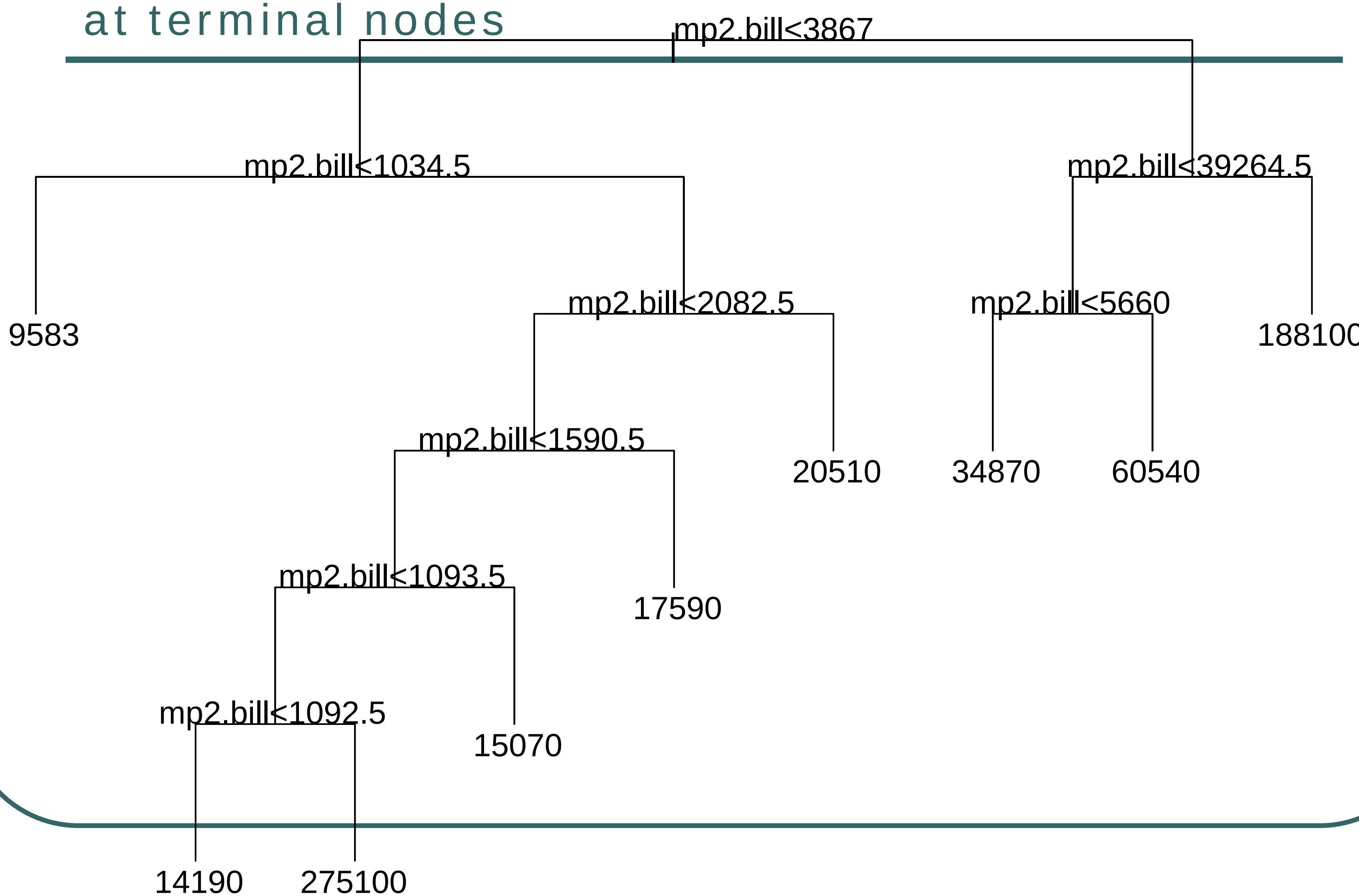
Decision Trees Cont.

- 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

Classification Trees: Categorical Dependent

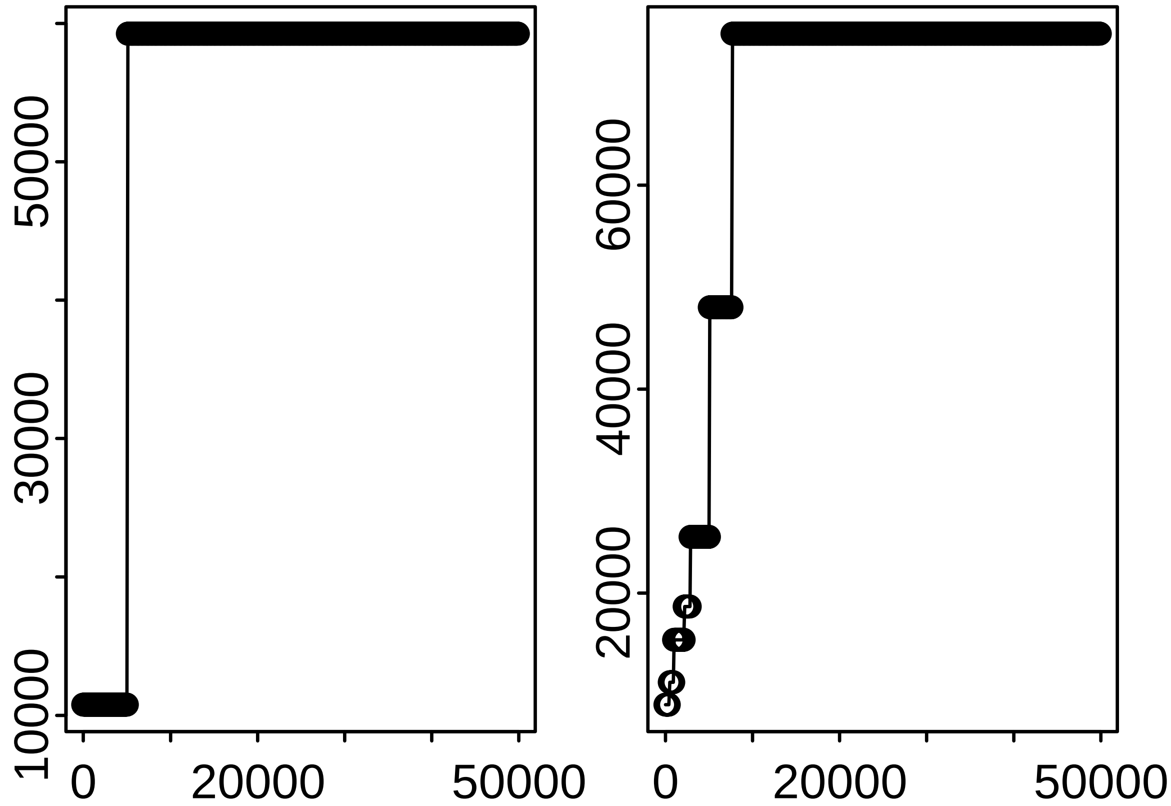
- Find the split that maximizes the difference in the probability of being in the target class
- Find split that minimizes *impurity*, or number of records not in the dominant class for the node
- Common goodness of fit measures are GINI index and entropy (deviance)

Continue Splitting to get more homogenous groups at terminal nodes

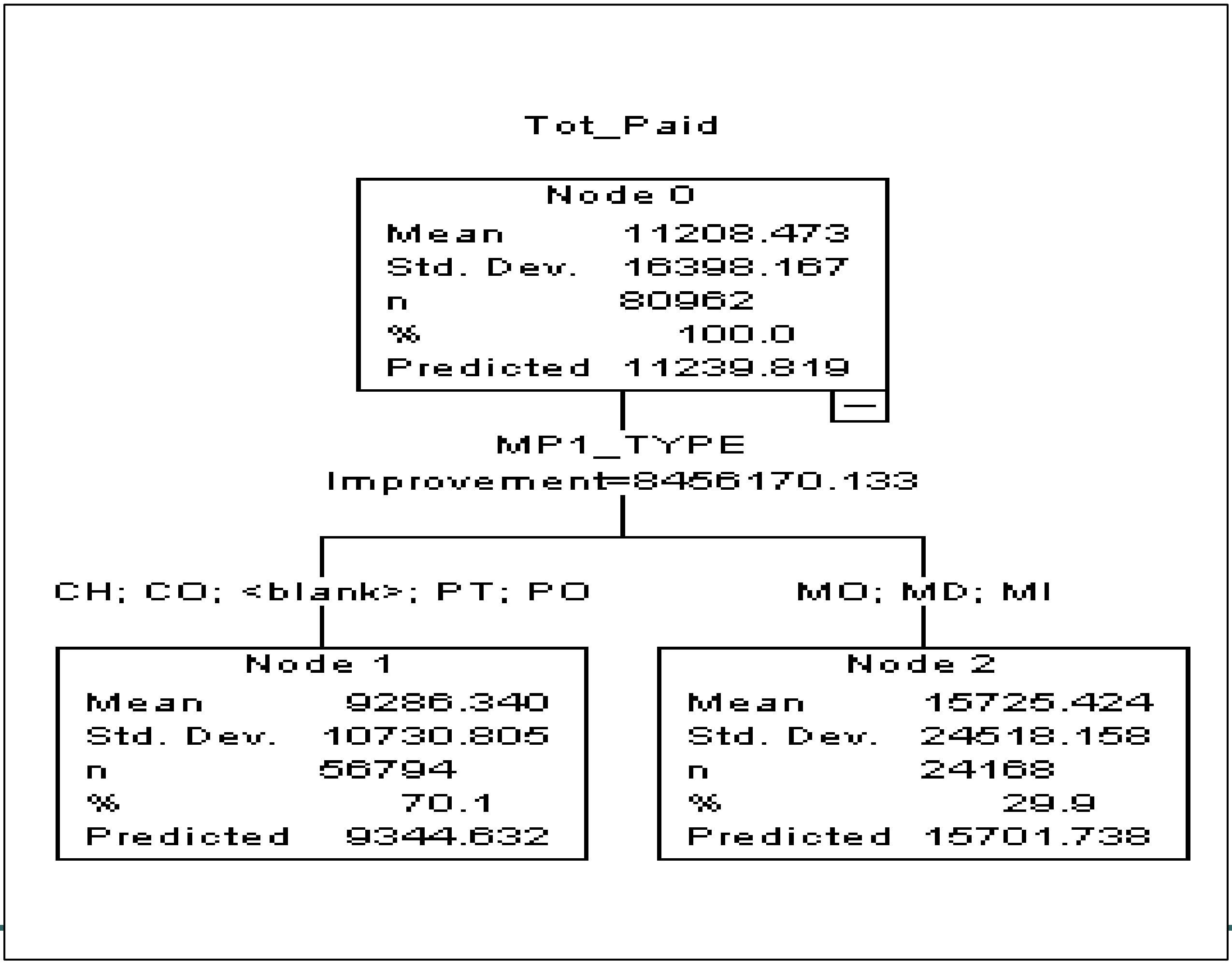


CART Step Function Predictions with One Numeric Predictor

Total Paid as a Function of Provider 2 Bill



Recursive Partitioning: Categorical Variables



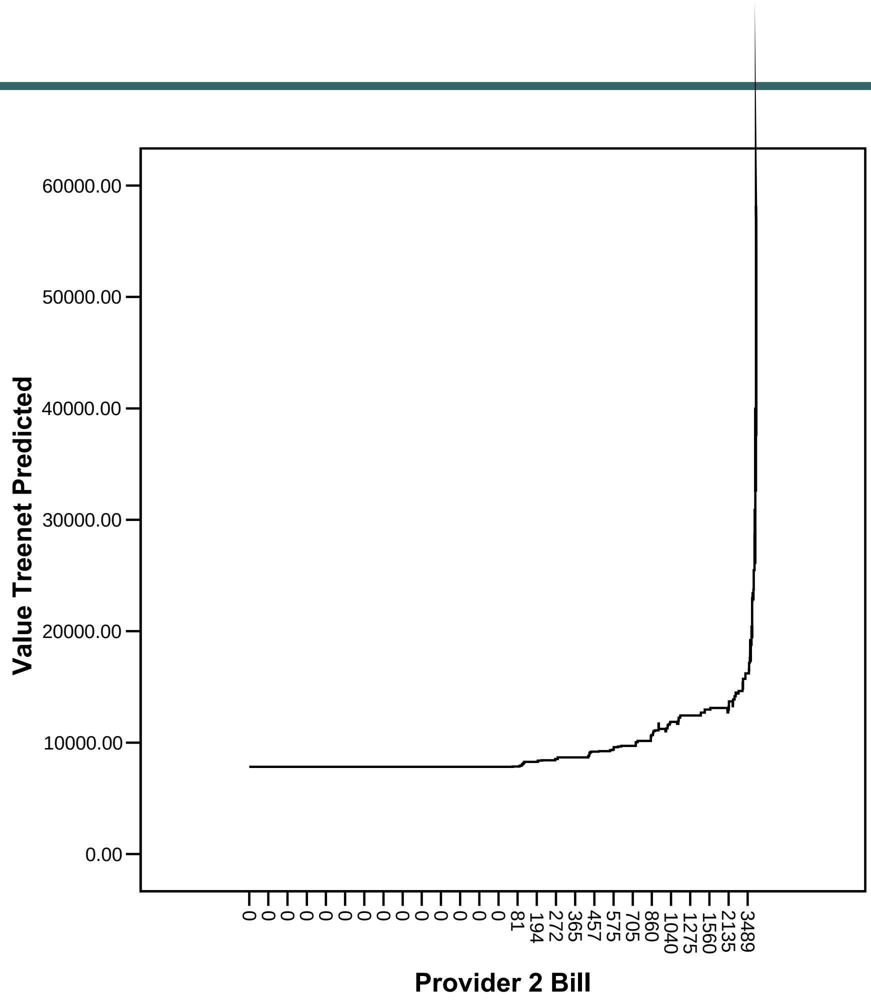
Different Kinds of Decision Trees

- 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

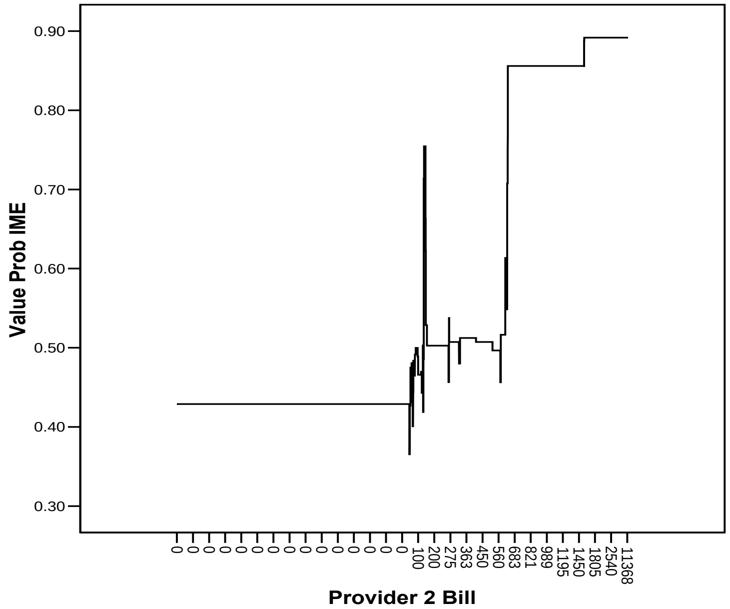
The Methods and Software Evaluated

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

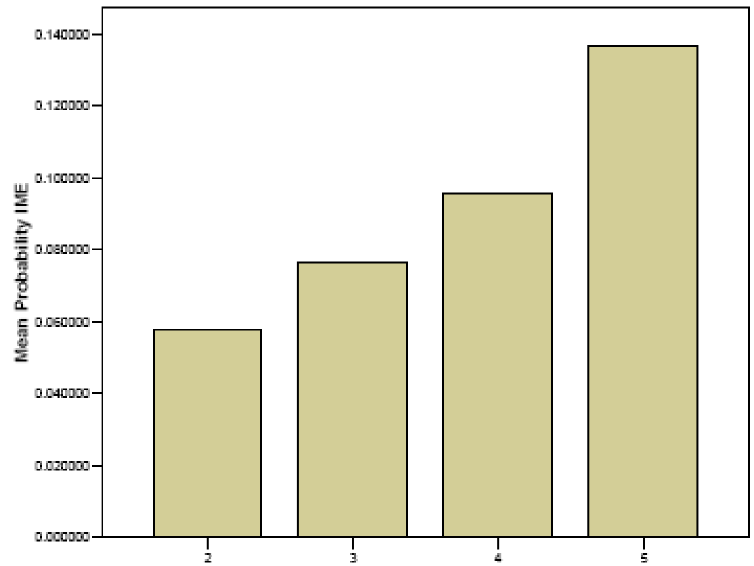
Ensemble Prediction of Total Paid



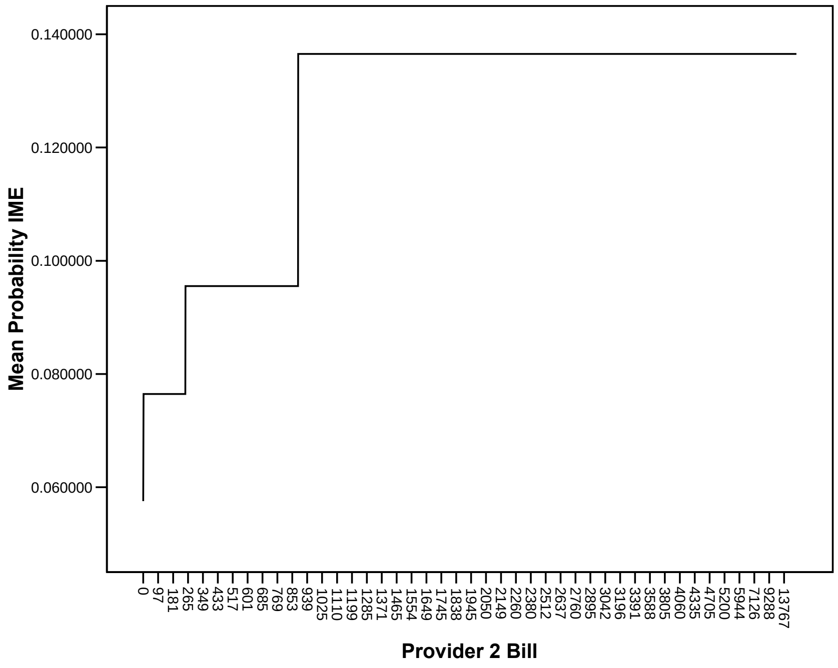
Ensemble Prediction of IME Requested



Bayes Predicted Probability IME Requested vs. Quintile of Provider 2 Bill



Naïve Bayes Predicted IME vs. Provider 2 Bill



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 IME Favorable

Area Under the ROC Curve – IME Favorable				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.651	0.664	0.591	0.683
Lower Bound	0.641	0.653	0.578	0.673
Upper Bound	0.662	0.675	0.603	0.693
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.654	0.692	0.670	0.677
Lower Bound	0.643	0.681	0.660	0.667
Upper Bound	0.665	0.702	0.681	0.687

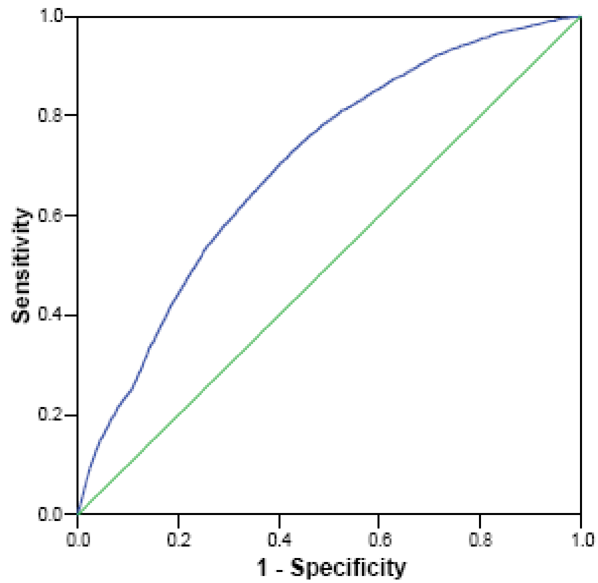
Results for SIU Referral

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

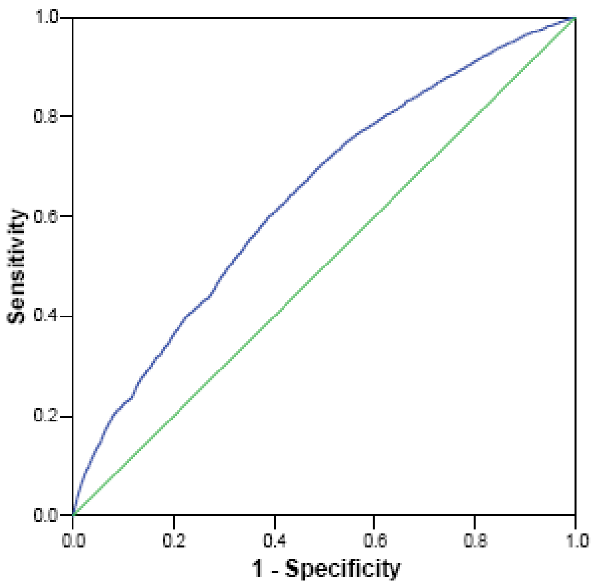
Results for SIU Favorable

Area Under the ROC Curve – SIU Favorable				
	CART Tree	**S-PLUS Tree**	**Iminer Tree**	**TREENET**
AUROC	0.598	0.616	0.547	0.678
Lower Bound	0.584	0.607	0.555	0.667
Upper Bound	0.612	0.626	0.575	0.689
	Iminer Ensemble	**Random Forest**	**Iminer Naïve Bayes**	**Logistic**
AUROC	0.575	0.645	0.607	0.610
Lower Bound	0.530	0.631	0.593	0.596
Upper Bound	0.548	0.658	0.625	0.623

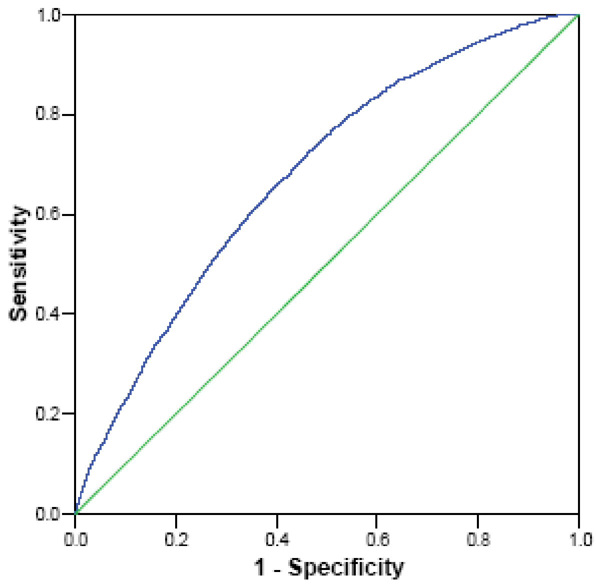
TREENET ROC Curve – IME
AUROC = 0.701



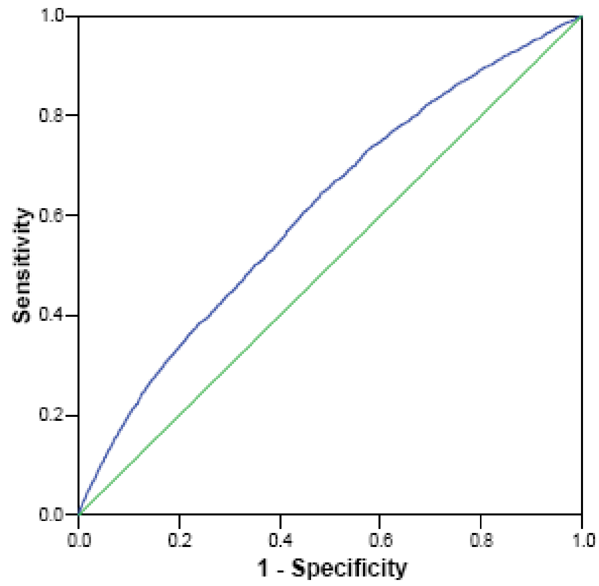
TREENET ROC Curve – SIU
AUROC = 0.677



Logistic ROC Curve – IME
AUROC = 0.643



Logistic ROC Curve – SIU
AUROC = 0.612



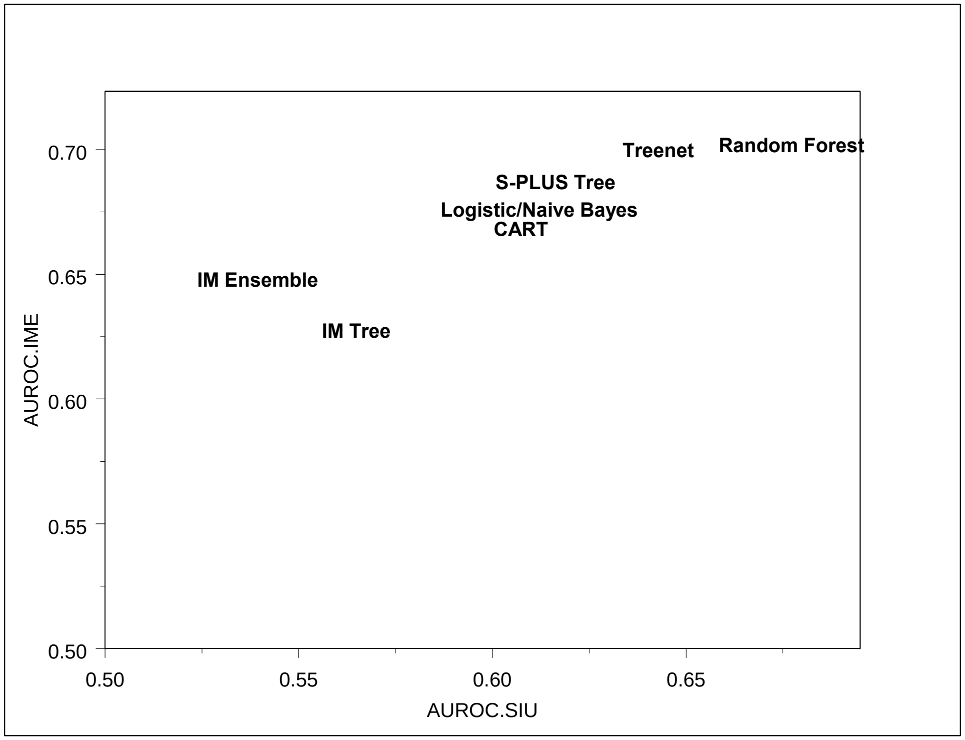
Ranking of Methods/Software – 1st Two Surrogates

Ranking of Methods By AUROC - Decision				
Method	**SIU AUROC**	**SIU Rank**	**IME Rank**	**IME AUROC**
Random Forest	0.645	1	1	0.703
TREENET	0.643	2	2	0.701
S-PLUS Tree	0.616	3	3	0.688
Iminer Naïve Bayes	0.615	4	5	0.676
Logistic	0.612	5	4	0.677
CART Tree	0.607	6	6	0.669
Iminer Tree	0.565	7	8	0.629
Iminer Ensemble	0.539	8	7	0.649

Ranking of Methods/Software – Last Two Surrogates

Ranking of Methods By AUROC - Favorable				
Method	**SIU AUROC**	**SIU Rank**	**IME Rank**	**IME AUROC**
TREENET	0.678	1	2	0.683
Random Forest	0.645	2	1	0.692
S-PLUS Tree	0.616	3	5	0.664
Logistic	0.610	4	3	0.677
Iminer Naïve Bayes	0.607	5	4	0.670
CART Tree	0.598	6	7	0.651
Iminer Ensemble	0.575	7	6	0.654
Iminer Tree	0.547	8	8	0.591

Plot of AUROC for SIU vs. IME Decision



Plot of AUROC for SIU vs. IME Favorable