Data Mining

CAS 2004 Ratemaking Seminar
Philadelphia, Pa.

Louise Francis, FCAS, MAAA
Louise_francis@msn.com

Objectives

- Answer the question: Why use data mining?
- Introduce the main data mining methods
  - Decision Trees
  - Neural Networks
  - MARS
  - Clustering

The Data

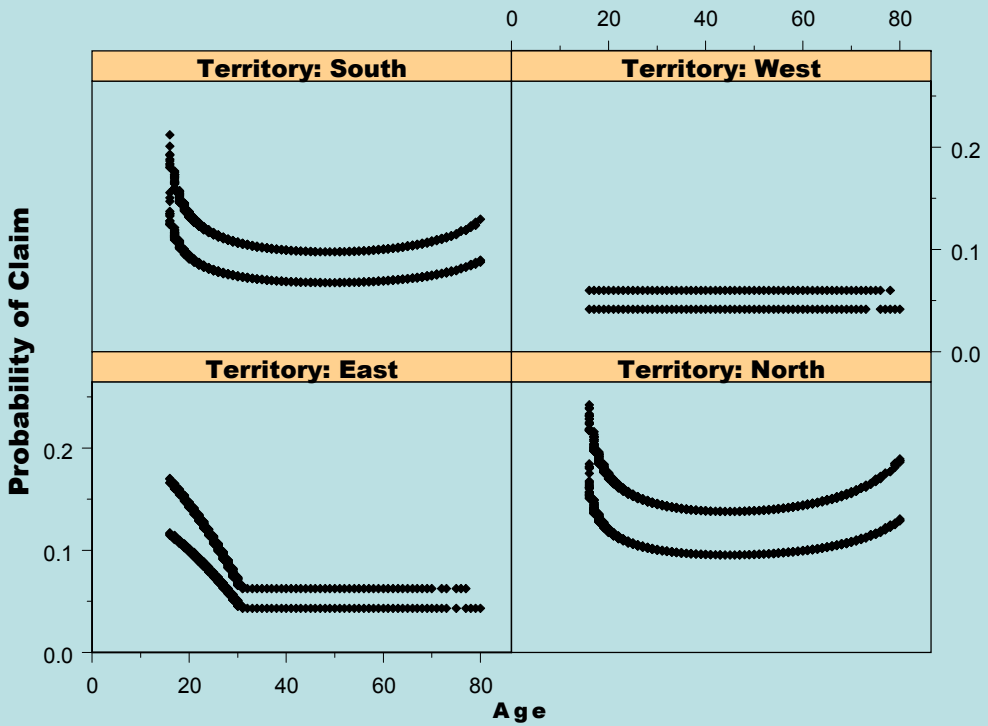
- Simulated Data for Automobile Claim Frequency
- Three Factors
  - Territory
    - Four Territories
  - Age
    - Continuous Function
  - Mileage
    - High, Low

Data Challenges

- Nonlinearities
  - Relation between dependent variable and independent variables is not linear or cannot be transformed to linear
- Interactions
  - Relation between independent and dependent variable varies by one or more other variables
- Correlations
  - Predictor variables are correlated with each other

Simulated Example: Probability of Claim vs. Age

by Territory and Mileage Group



Francis Analytics and Actuarial
Data Mining, Inc.

Claim Frequency Data

Claim Count

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	13579	90.5	90.5	90.5
	1	1349	9.0	9.0	99.5
	2	70	.5	.5	100.0
	3	2	.0	.0	100.0
	Total	15000	100.0	100.0	

Independent Probabilities for Each Variable

Claim Count * Territory

Mean

		Claim Count
Territory	East	.06
	North	.13
	South	.10
	West	.05
	Total	.10

Claim Count * Mileage Group

Mean

		Claim Count
Mileage Group	High	.12
	Low	.08
	Total	.10

Mean

		Claim Count
Age Group	18.5	.13
	25.0	.11
	35.0	.09
	45.0	.09
	55.0	.09
	65.0	.10
	75.0	.13
	85.0	.18
	Total	.10

Decision Trees

- Recursively partitions the data
  - Often sequentially bifurcates the data – but can split into more groups
- Applies goodness of fit to select best partition at each step
- Selects the partition which results in largest improvement to goodness of fit statistic

Goodness of Fit Statistics

- Chi Square ⇒ CHAID (Fish, Gallagher, Monroe-Discussion Paper Program, 1990)

$$C = \sum_{i,k} \frac{(\text{Observed-Expected})^2}{\text{Expected}}$$

- Deviance ⇒ CART

$$D_i = -2\sum_{k} n_{ik} \log(p_{ik}) \text{ (categorical)}$$

$$\text{D} = \sum_{\text{cases j}} (\text{y}_\text{j} - \mu_j)^2 \text{ (or RSS for continuous variables)}$$

Goodness of Fit Statistics

- Gini Measure $\Rightarrow$ CART

$$i = 1 - \sum_k p_k{}^2$$

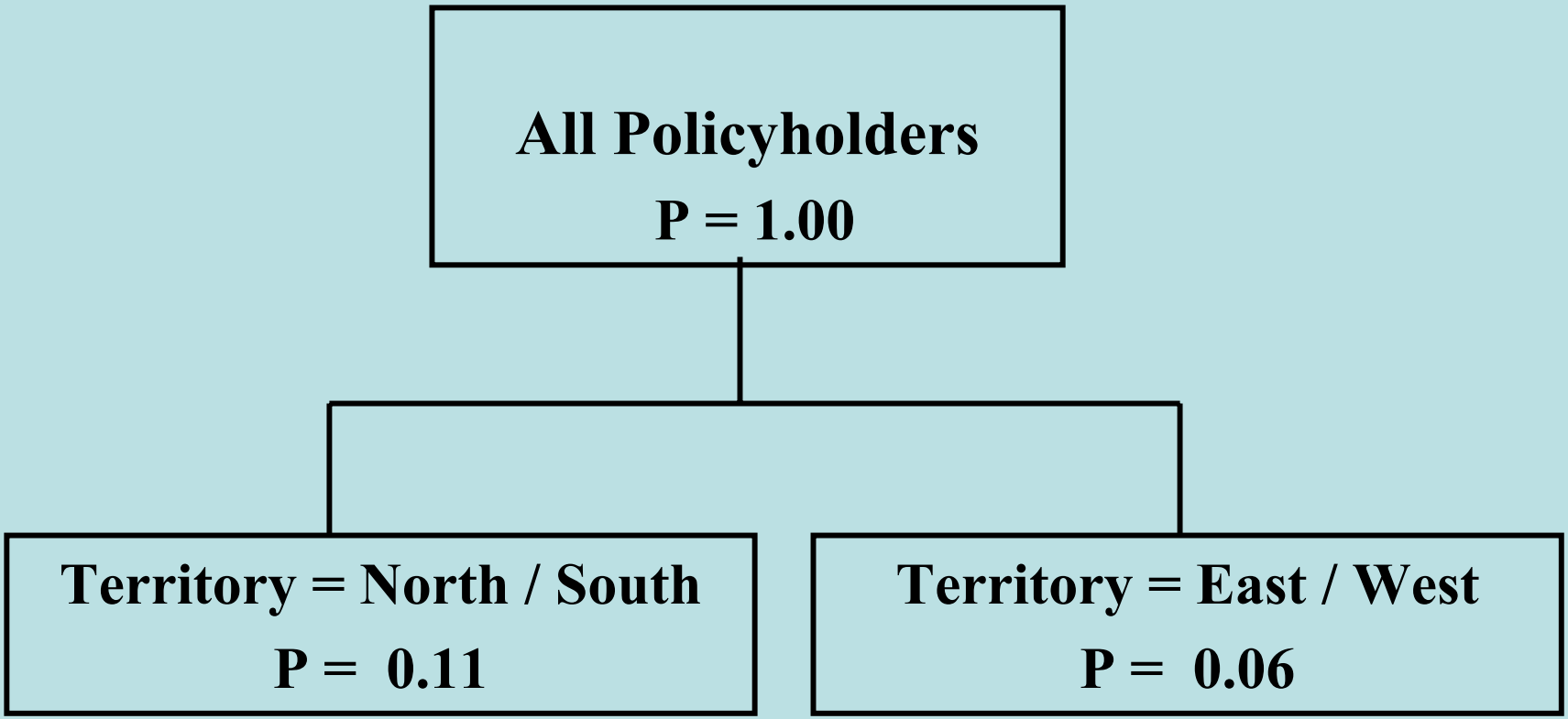
Goodness of Fit Statistics

- Entropy $\Rightarrow$ C4.5

$$I(E) = -\log_2\left(\frac{E}{N}\right) = -\log_2(p_E)$$

$$H = -\sum_k p_k \log_2(p_k)$$

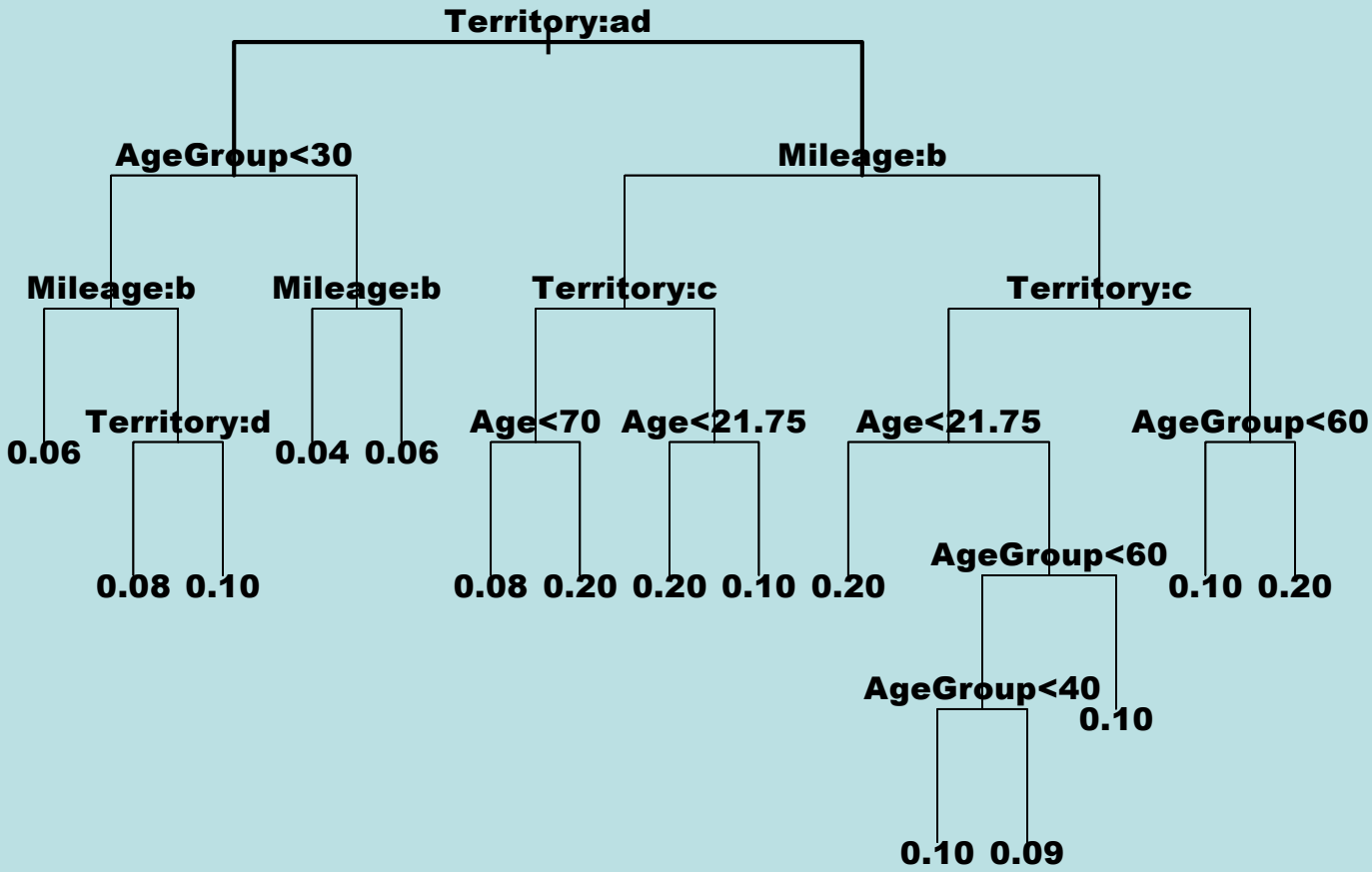
First Split


All Policyholders
P = 1.00
Territory = North / South
P = 0.11
Territory = East / West
P = 0.06

Example of Goodness of Fit Calculation

Example of Deviance Calculation					
	No Claims	Claims	Claims	No Claims	
	N	N	p	p	Deviance
Root Node	13,579	1,421	0.905	0.095	4,082.64
"North/South"	9,854	1,198	0.892	0.108	3,294.12
"East/West"	3,725	223	0.944	0.056	744.76
Total First Split					4,038.88
Change in Deviance					43.76

Example of Fitted Tree

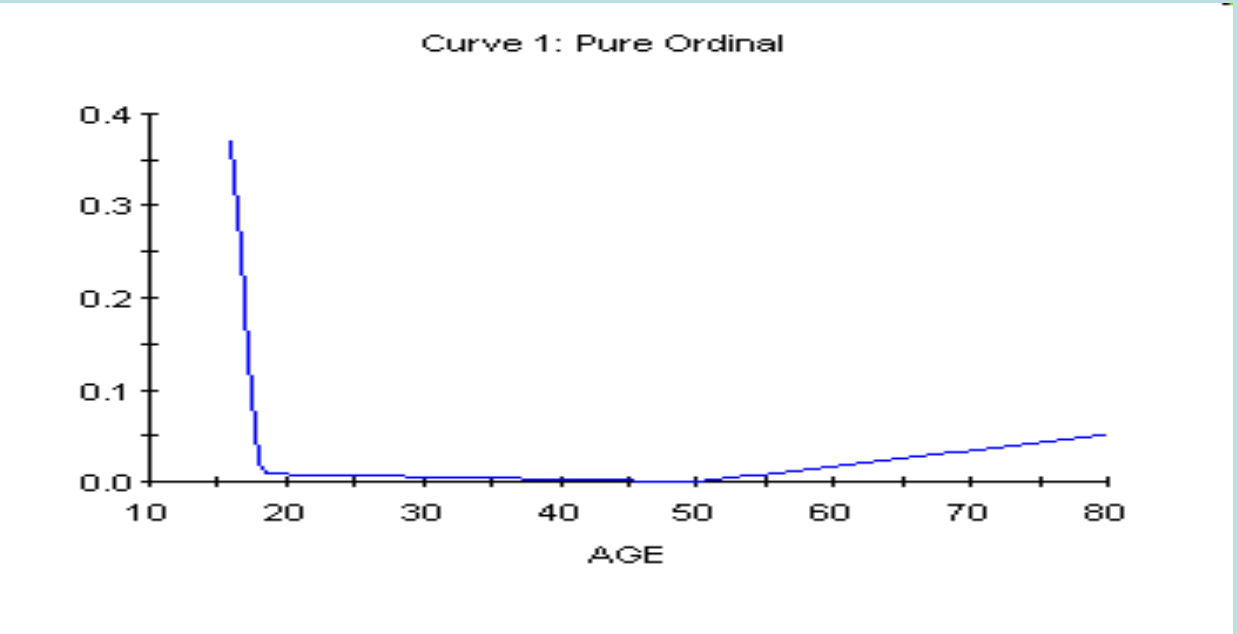


MARS

- Multivariate Adaptive Regression Splines
- An extension of regression which
  - Uses automated search procedures
  - Models nonlinearities
  - Models interactions
  - Produces a regression-like formula

Nonlinear Relationships

- Fits piecewise regression to continuous variables



Interactions

- Fits basis functions (which are like dummy variables) to model interactions
  - An interaction between Territory=East and Mileage can be modeled by a dummy variable which is 1 if the Territory=East and mileage =High and 0 otherwise.

Goodness of Fit Statistics

- Generalized Cross-Validation

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - k/N}\right]^2$$

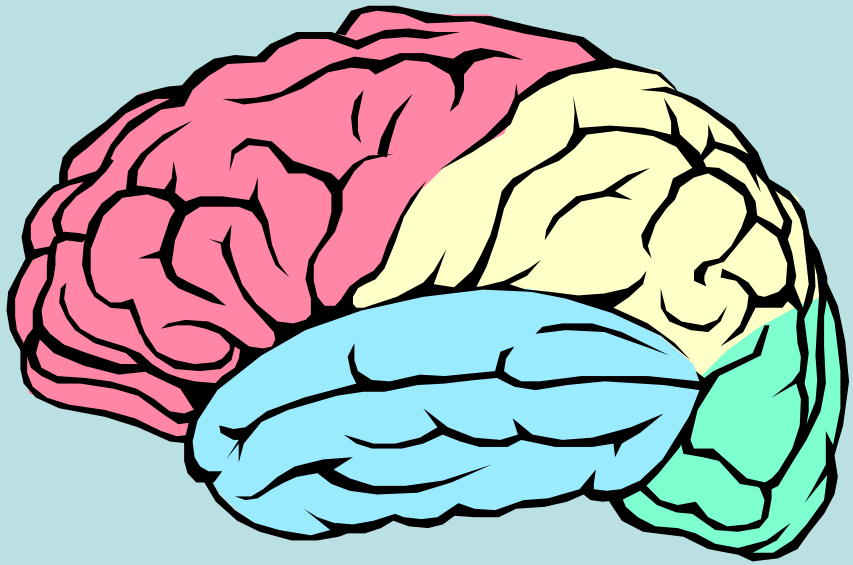
where N is the number of observations
y is the dependent variable
x is the independent variable(s)
k is the effective number of parameters or degrees of freedom in the model.

Fitted MARS Model

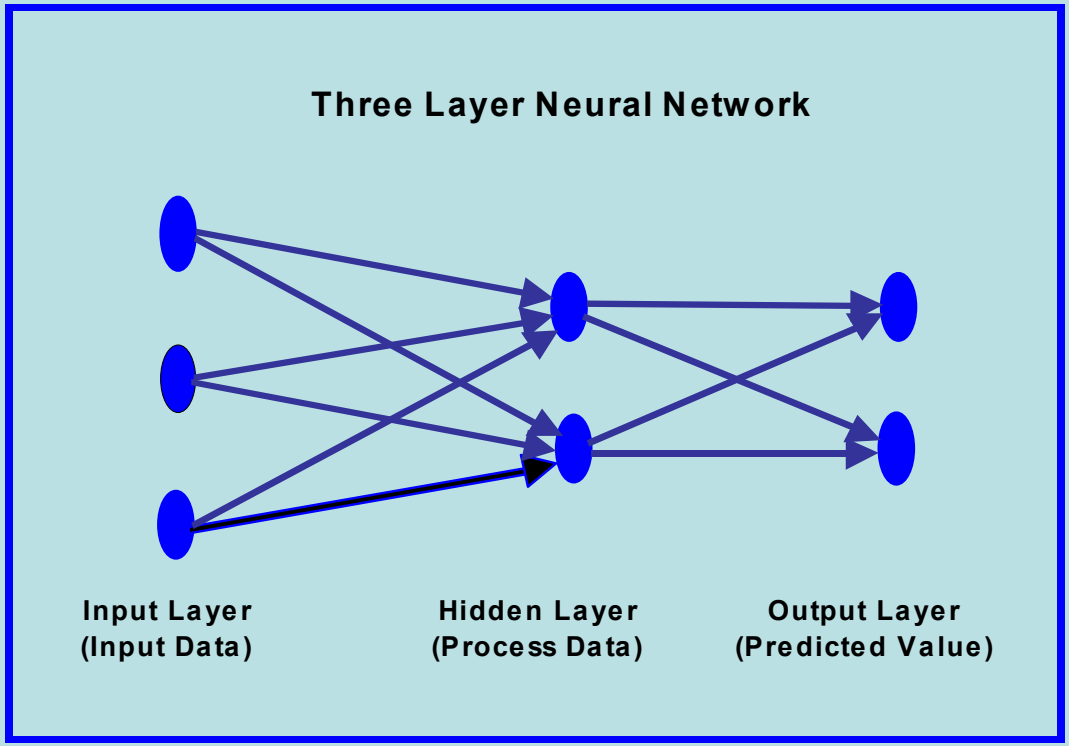
```
Basis Functions:
 BF1 = ( TERRITORY = 2 OR TERRITORY = 3);
 BF3 = (MILEAGE = HIGH);
 BF5 = ( TERRITORY = 1 OR TERRITORY = 2);
 BF7 = max(0, AGE - 50.000);
 BF8 = max(0, 50.000 - AGE );
 BF9 = max(0, AGE - 18.000);
 BF10 = max(0, 18.000 - AGE );
 BF11 = ( TERRITORY = 2 OR TERRITORY = 4) * BF10;
 BF13 = ( TERRITORY = 1) * BF9;
 BF17 = max(0, AGE - 19.000) * BF3;
 BF18 = max(0, 19.000 - AGE ) * BF3;
 BF19 = max(0, AGE - 22.000) * BF3;

Model
 Y = -3.887 + 0.044 * BF1 + 0.032 * BF5 - 0.121 * BF7 +
0.124 * BF8+ 0.123 * BF9 - 0.071 * BF11 - .979823E-03 *
BF13 - 0.011 * BF17 - 0.049 * BF18 + 0.011 * BF19;
```

Neural Networks

- Developed by artificial intelligence experts – but now used by statisticians also
- Based on how neurons function in brain



Neural Network Structure



Neural Networks

- Fit by minimizing squared deviation between fitted and actual values
- Can be viewed as a non-parametric, non-linear regression
- Often thought of as a “black box”
  - Due to complexity of fitted model it is difficult to understand relationship between dependent and predictor variables

Understanding the Model:
Variable Importance

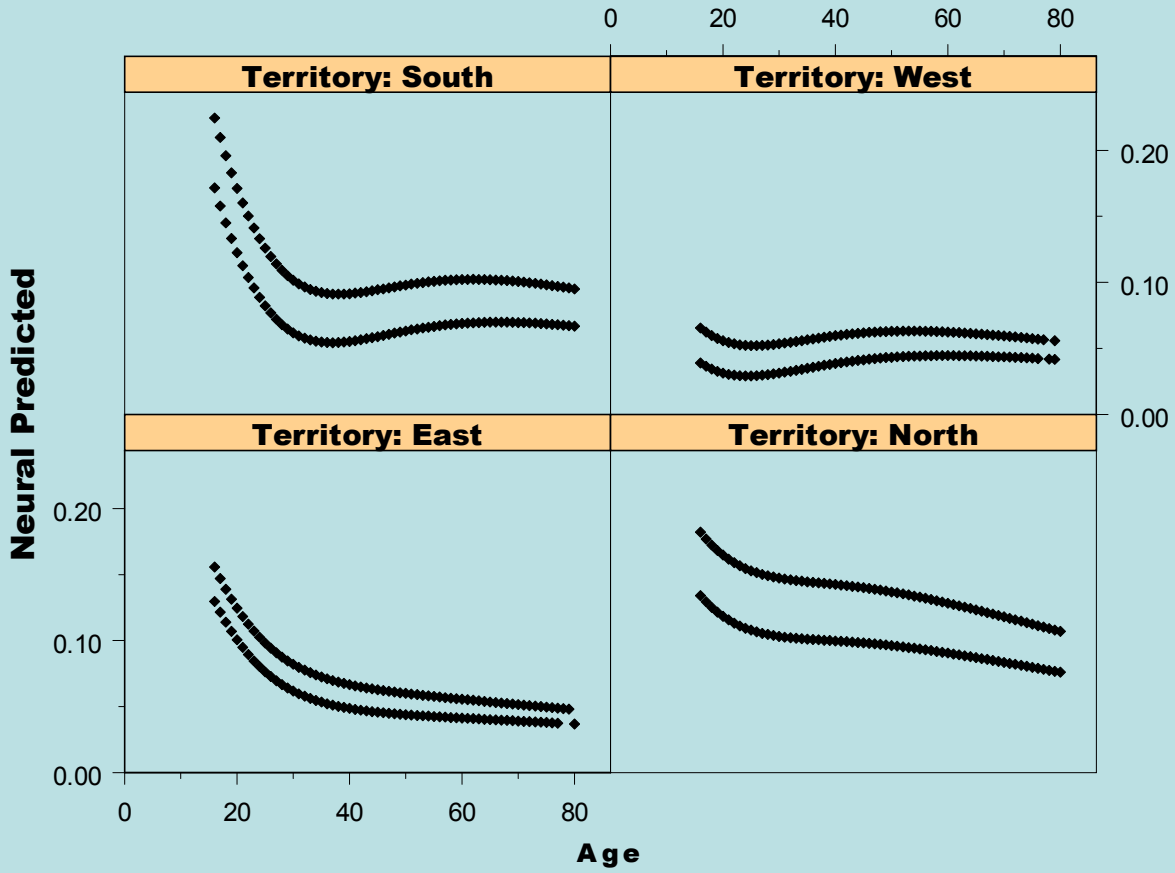
- Look at weights to hidden layer
- Compute sensitivities:
  - a measure of how much the predicted value's error increases when the variables are excluded from the model one at a time

Importance Ranking

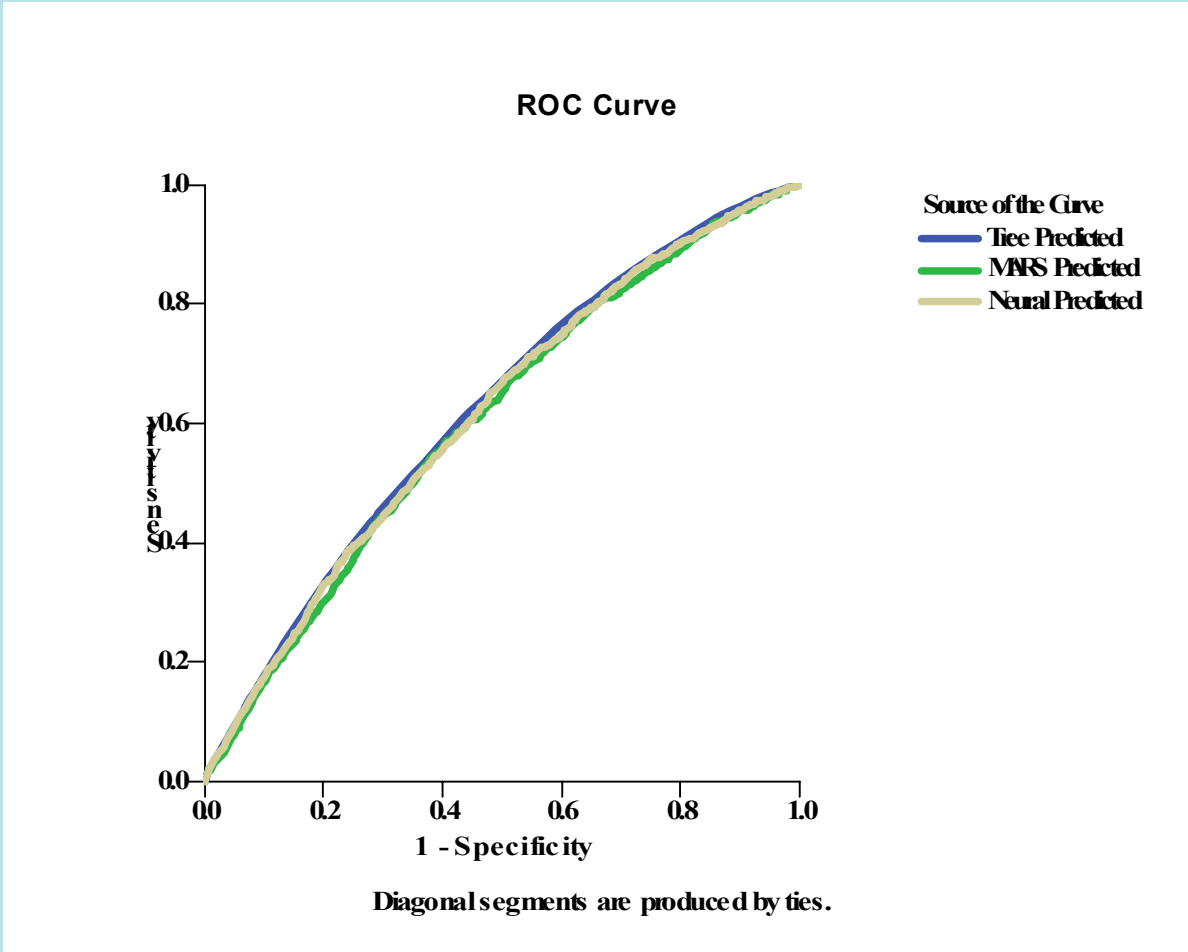
- Neural Network and Mars ranked variables in same order

Variable	Neural Net Rank	MARS Rank
Territory	1	1
Age	2	2
Mileage	3	3

Visualizing Fitted Neural Network



ROC Curves for the Data Mining Methods



Correlation

- Variable gender added
- Its only impact on probability of a claim: correlation with mileage variable – males had higher mileage
  - MARS did not use the variable in model
  - CART used it in two places to split tree
  - Neural Network ranked gender as least important variable

How the Methods Did

Correlation with "True" Claim Frequency

Correlations

		True frequency	Tree Predicted	MARS Predicted	Neural Predicted
True frequency	Pearson Correlation	1	.895**	.923**	.954**
	Sig. (2-tailed)	.	.000	.000	.000
	N	15000	15000	15000	15000
Tree Predicted	Pearson Correlation	.895**	1	.840**	.924**
	Sig. (2-tailed)	.000	.	.000	.000
	N	15000	15000	15000	15000
MARS Predicted	Pearson Correlation	.923**	.840**	1	.892**
	Sig. (2-tailed)	.000	.000	.	.000
	N	15000	15000	15000	15000
Neural Predicted	Pearson Correlation	.954**	.924**	.892**	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	15000	15000	15000	15000

**. Correlation is significant at the 0.01 level (2-tailed).

Unsupervised Learning

- Common Method: Clustering
- No dependent variable – records are grouped into classes with similar values on the variable
- Start with a measure of similarity or dissimilarity
- Maximize dissimilarity between members of different clusters

Dissimilarity (Distance) Measure

- Euclidian Distance

$$d_{ij} = \left(\sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \right)^{1/2} \quad \text{i, j = records k=variable}$$

- Manhattan Distance

$$d_{ij} = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right| \right)$$

Binary Variables

		Row Variable			
Column Variable			1	0	
	0	a	b	a+b	
	1	c	d	c+d	
		a+c	b+d		

Binary Variables

- Sample Matching

$$d = \frac{b + c}{a + b + c + d}$$

- Rogers and Tanimoto

$$d = \frac{2(b + c)}{(a + d) + 2(b + c)}$$

Example: Fraud Data

- Data from 1993 closed claim study conducted by Automobile Insurers Bureau of Massachusetts
- Claim files often have variables which may be useful in assessing suspicion of fraud, but a dependent variable is often not available
- Variables used for clustering:
  - Injury type
  - Provider type
  - Legal representation
  - Prior Claim
  - SIU Investigation

Results for 2 Clusters

Cluster	Lawyer	Back Claim Or Sprain	Chiro or PT	Prior Claim
1	77%	73%	56%	26%
2	3%	29%	14%	1%

Cluster	Suspicious Claim	Average Suspicion Score
1	56%	2.99
2	3%	0.21

Beginners Library

- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997
- Kaufman, Leonard and Rousseeuw, Peter, *Finding Groups in Data,* John Wiley and Sons, 1990
- Smith, Murry, *Neural Networks for Statistical Modeling*, International Thompson Computer Press, 1996

Data Mining

CAS 2004 Ratemaking Seminar
Philadelphia, Pa.

Louise Francis, FCAS, MAAA
Louise_francis@msn.com