Use Data Mining Techniques to Assist Institutions in Achieving Enrollment Goals: A Case Study

Tongshan Chang

The University of California Office of the President

CAIR Conference in Pasadena

11/13/2008

Email: Tongshan.Chang@ucop.edu

Data Mining: Concepts

- **SAS**: "the process of sampling, exploring, modifying, modeling, and assessing (SEMMA) large amounts of data to uncover previously unknown patterns, which can be utilized as a business advantage." (Applying Data Mining, 2005, p. 1-3)

- **Microsoft**: "Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data." (http://technet.microsoft.com/en-us/library/ms174949.aspx)

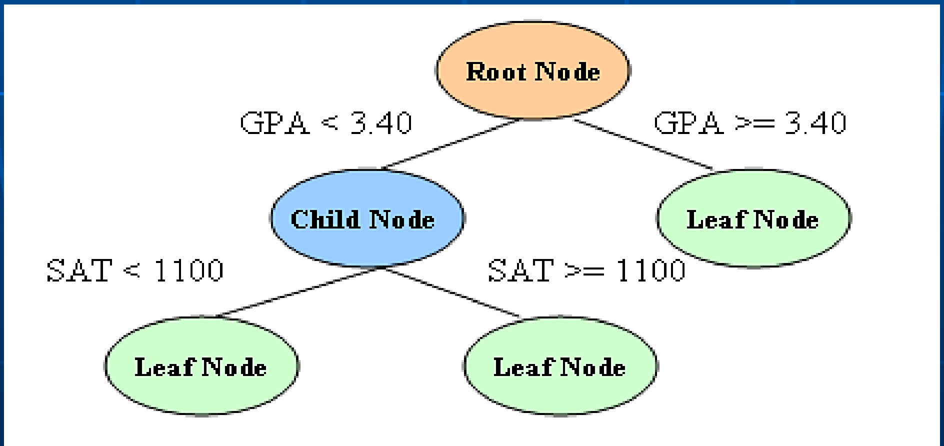
- **Berry and Linoff**: "Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. ...the goal of data mining is to allow a corporation to improve its marketing, sales, and customers." (Data Mining Techniques, p.7).

Data Mining: What Can We Do with It?

- **Classification**: discrete outcomes: yes or no
- **Estimation**: continuous values outcomes
- **Prediction**: the same as classification or estimation, but classifying according to some predicted future behavior or estimated future value
- **Association Rules**: determine which things go together
- **Clustering**: segment a heterogeneous population into a number of more homogeneous subgroups or clusters
- **Description and Profiling**: simply describe what is going on in a complicated database

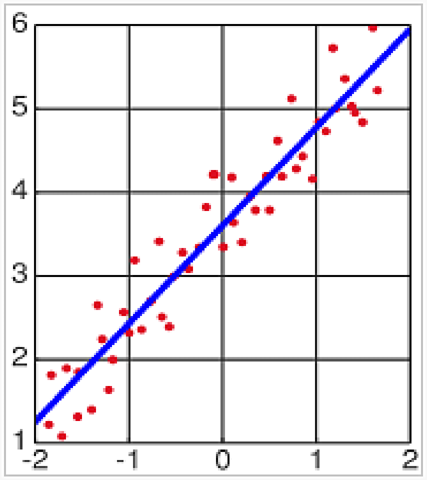
Data Mining: Techniques—Decision Tree

- **Decision Tree**
  - Divide up a large collection of records into smaller sets of records using decision rules
  - Process: Record → Root Node → Child Node → Leaf Node
  - The PATH is an expression of the rules used to classify the records.
    - **3 paths in this tree**
      - **GAP>=3.40**
      - **GPA<2.40 → SAT>=1100**
      - **GPA<3.40 → SAT < 1100**



Data Mining: Techniques—Regression (Logistic Regression)

Regression



Example of linear regression with one independent variable.

Logistic Regression

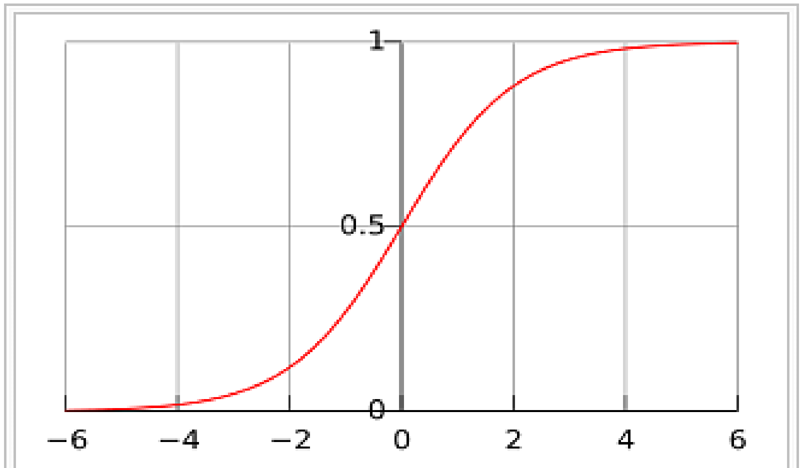
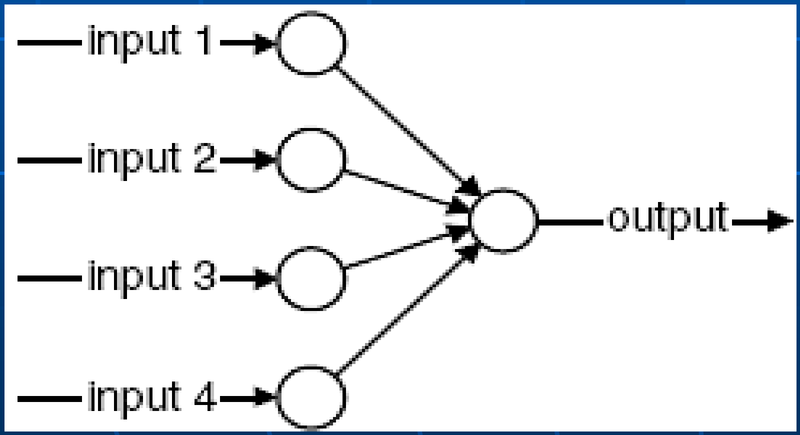


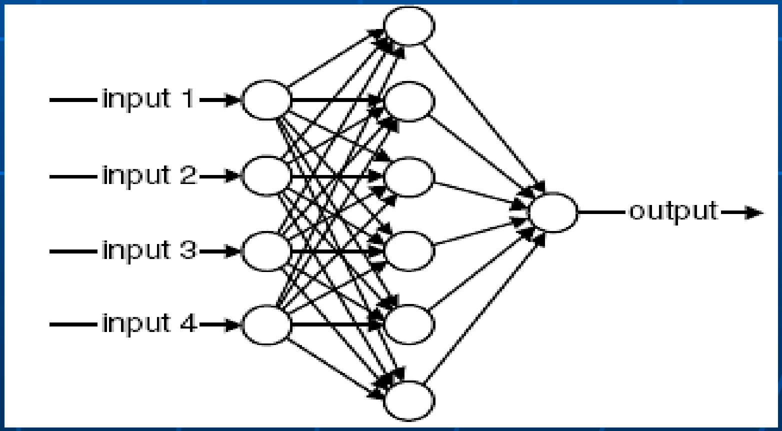
Figure 1. The logistic function, with *z* on the horizontal axis and *f(z)* on the vertical axis.

*Chart Source: Wikipedia, http://en.wikipedia.org/wiki/Logistic_regression

Data Mining: Techniques—Neural Network

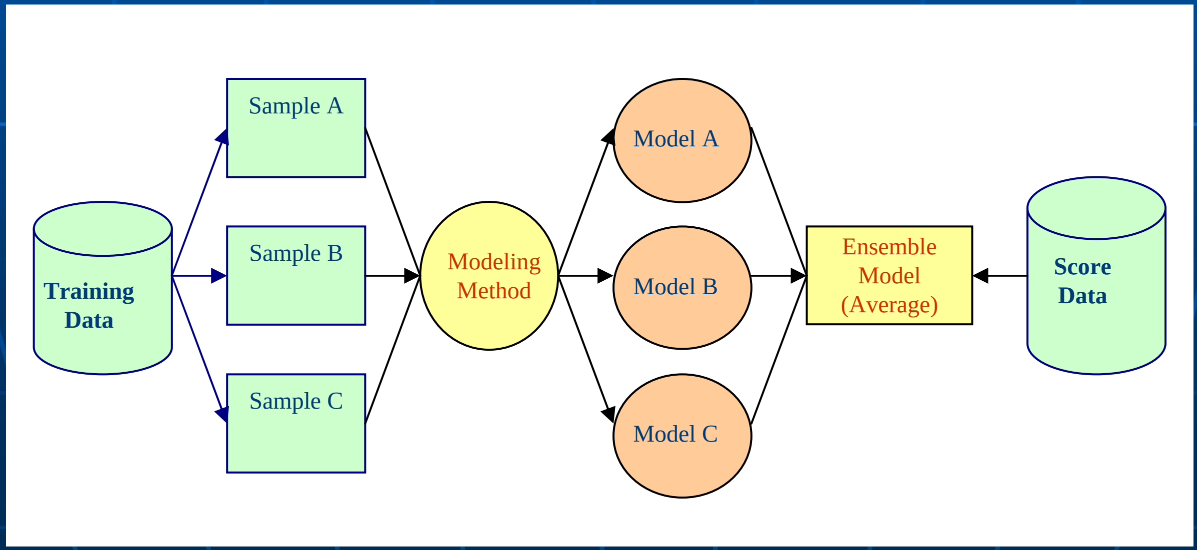
- **Neural Network**
  - Similar property of biological neurons
  - Interconnected artificial neurons
  - Inputs → Hidden Layer → Output(s)
  - Weights
    - Inputs and Hidden Layer
    - Hidden Layer and Output





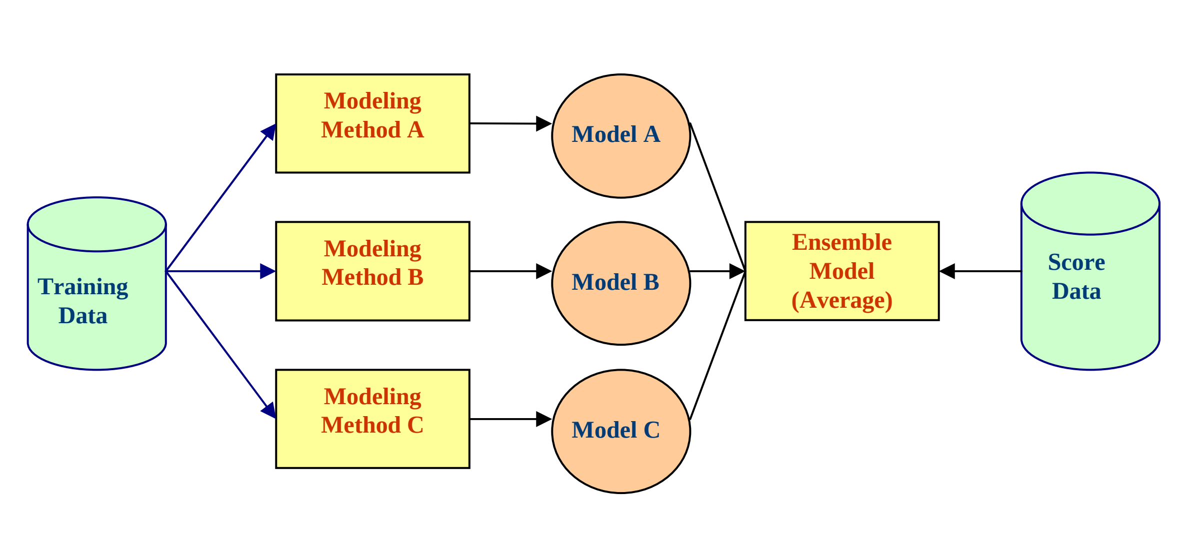
Data Mining: Techniques—Ensemble

- **Ensemble:** Averaging the posterior probabilities for class targets or the predicted values for interval targets from multiple models
- **Methods:**
  - Different models from the same modeling method based on separate samples of training data set



Data Mining: Techniques—Ensemble

- **Ensemble:** Averaging the posterior probabilities for class targets or the predicted values for interval targets from multiple models
- **Methods:**
  - Different models from the same modeling method based on three separate samples of training data set
  - Different models from the different modeling methods based on the same training data set

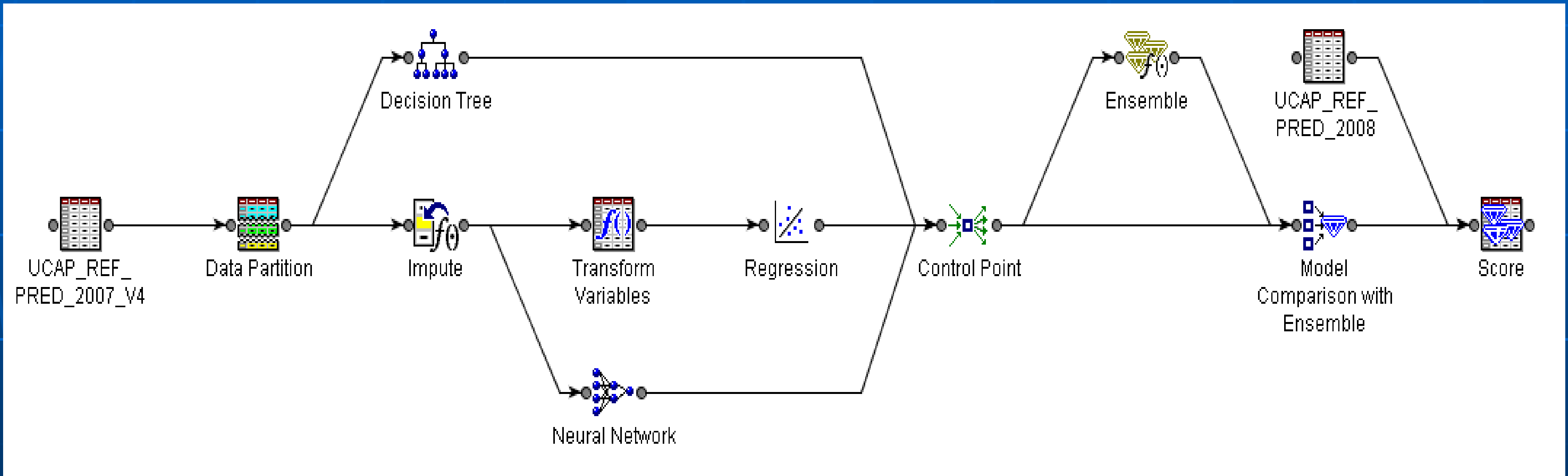


Data Mining: Applications in Institutional Research

- **College admissions yield** (Chang, 2006)
- **Retention** (Herzop, 2006; Sujitparapitaya, 2006)
- **Time to degree** (Eykamp, 2006; Herzop, 2006)
- **Enrollment management** (Aksenova, Zhang, & Lu, 2006)
- **Course offerings** (Luan, 2006; Dai, Yeh, & Lu, 2007)
- **Student performance** (Dede &Clarke, 2007; Heathcote & Dawson, 2005; Minaei-Bidgoli, 2004; Ogor, 2007)
- **Graduation rate** (Baily, 2006)
- **Student experience survey study** (Yu, et. al, 2007)

A Case Study Using SAS Enterprise Miner

Assist Institutions in Achieving Enrollment Goals



A Case Study Using SAS Enterprise Miner—Background

- **Paths to Eligibility for CA Residents at UC**
  - Eligibility in the Statewide Context
  - Eligibility in the Local Context (ELC)
  - Eligibility by Examination Alone
- **Admissions:**
  - UC guarantees to admit all CA eligible applicants, but does not guarantee to admit everyone in terms of the campus or the program he/she applied to.

A Case Study Using SAS Enterprise Miner—Background

- **Referral Pool:**
  - Eligible, not admitted
  - To the referral pool
  - Two UC campuses: Riverside and Merced
  - Don’t know until April, too late, so the yield rate is low
- **Early Referral Pool:**
  - A letter to those who may be in the referral pool
  - Admit those who would like to consider these two campuses
- **Question:** Who do we send a letter to?

A Case Study Using SAS Enterprise Miner—Purpose

- Predict UC applicants who are qualified to UC admissions systemwide, but not admitted to the campus they applied to
- Two campuses use the information to make Early Referral Pool admissions offers and try to enroll more students.

A Case Study Using SAS Enterprise Miner—Data Description

- **UC Freshman Application Data**
  - **Data Sets**:
    - Fall 2007 data, training data
    - Fall 2008 data, target data
  - **Observations (Eligible Applicants)**:
    - Fall 2007: 45,393
    - Fall 2008: 48,356
  - **Elements**
    - Student demographic and academic information
    - Family information
    - Application information (campuses, major, etc.)
- **CDE School Performance Data**
  - Academic Performance Index (API)

A Case Study Using SAS Enterprise Miner—Variables

Variable	Data Type	Description
Referral Pool	Dichotomous	Dependent variable: 1=in referral pool, 0=not in referral pool
Ethnicity	Categorical	7 categories
First Language	Categorical	3 categories: English Only, English and Another Language, and Another Language
Campus(es) Applied to	Categorical	7 variables, one for each campus: e.g. CAMP_BK: 1=applied to UC Berkeley, 0=not applied to UC Berkeley
Parent's Educational Level	Categorical	5 Categories: HS or Less, 2 Year College, 4 Year College and Post Ed. Study, Missing
Family Income	Continuous	
Home Location	Categorical	5 Categories: San Francisco Bay Areas, CA North, LA County, CA South, and Other
Discipline	Categorical	7 Variables, one for each campus: 5 categories for each variable: Engineering, Science, Social Science, Humanities, Others.
Outreach Programs	Dichotomous	Participated at least one or not participated in any one.
API Ranking	Categorical	1 to 10 for public schools, missing for private schools
High School GPA	Continuous	Weighted, Capped GPA
UC Score (SAT or ACT)	Continuous	Highest of converted SAT or ACT score, including 2 highest SAT subject tests

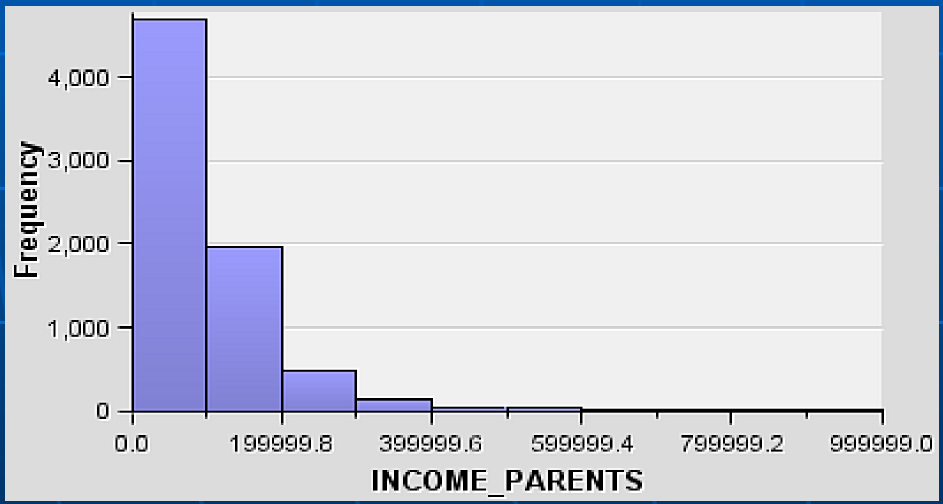
A Case Study Using SAS Enterprise Miner—Missing Value Imputation

- **Categorical Variable**: not necessary, "MISSING" is a category.
- **Continuous Variable**:
  - Discard vs. Impute
    - For data accuracy, simply discard, but reduce data drastically
    - Scoring problem: records with missing values will not be scored
    - Decision tree modeling: not necessary
    - Logistic regression and neural network modeling: ignore all records with missing values
    - Compare models: on the same set of observations
  - SAS Methods: 11— mean, median, mid-range, tree, etc.
  - Method for This Project: median, tree, mean, etc. were used, but the best method is mean

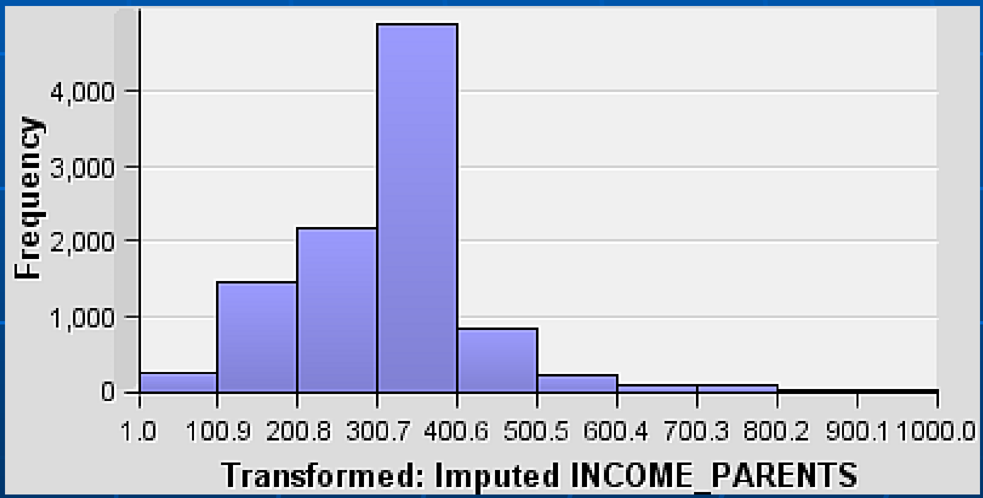
A Case Study Using SAS Enterprise Miner—Data Transformation

- **Transformation**: highly skewed distribution, a great deal of influence
- **Decision tree and neural network modeling**: Flexible
- **Logistic regression modeling**: Transformation may yield a better fitting model

Before Transformation

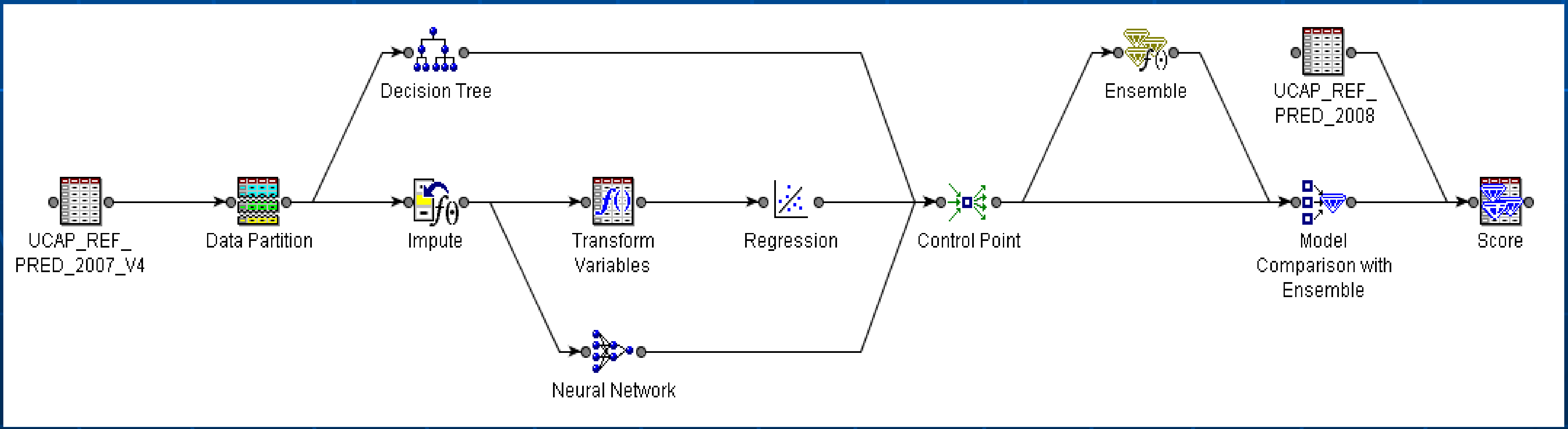


After Transformation



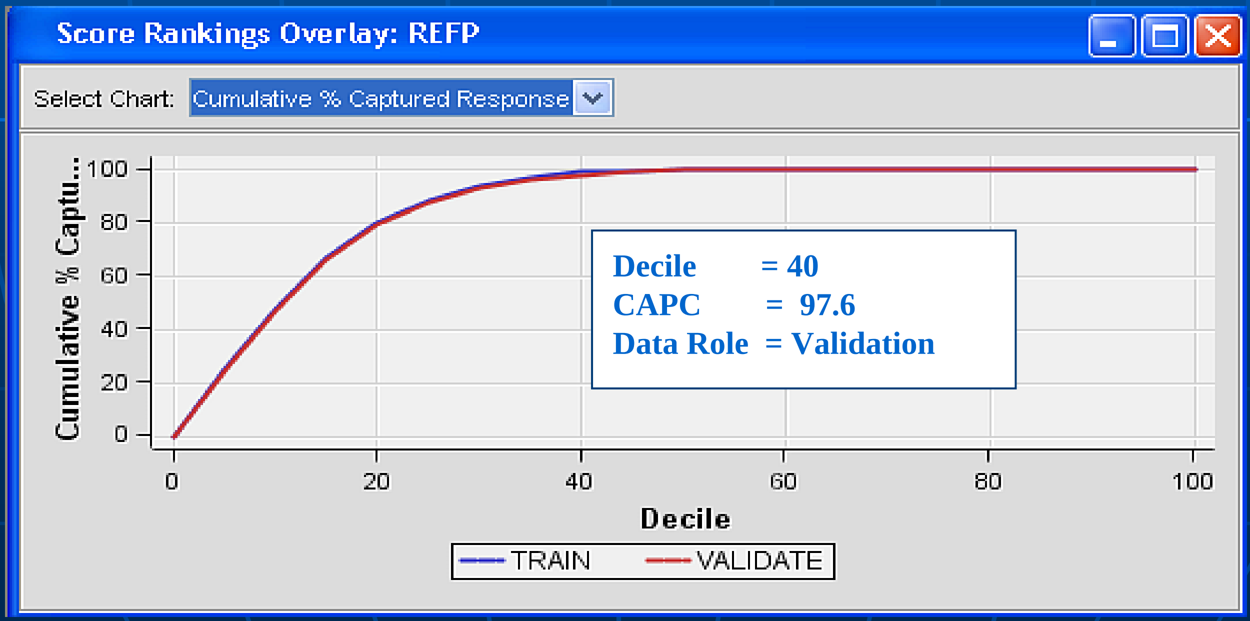
A Case Study Using SAS Enterprise Miner—Modeling Process

- **Data Partition**:
  - **Training Data Set**: Preliminary model fitting
  - **Validation Data Set**: Monitoring and tuning the model to improve its generalization
  - **Test Data Set**: Estimate of Generalization
- **Data Set Percentage**: User decides, but each observation is allowed to use only once, 40%, 30%, and 30%.
- **Four Models**: Decision Tree, Logistics Regression, Neural Network, and Ensemble



A Case Study Using SAS Enterprise Miner—Model Assessment

- **Score Rankings Overlay**
  - **Lift**
  - **Cumulative Lift**
  - **Gain**
  - **% Response**
  - **Cumulative % Response**
  - **% Captured Response**
  - **Cumulative % Captured Responses**



A Case Study Using SAS Enterprise Miner—Model Assessment

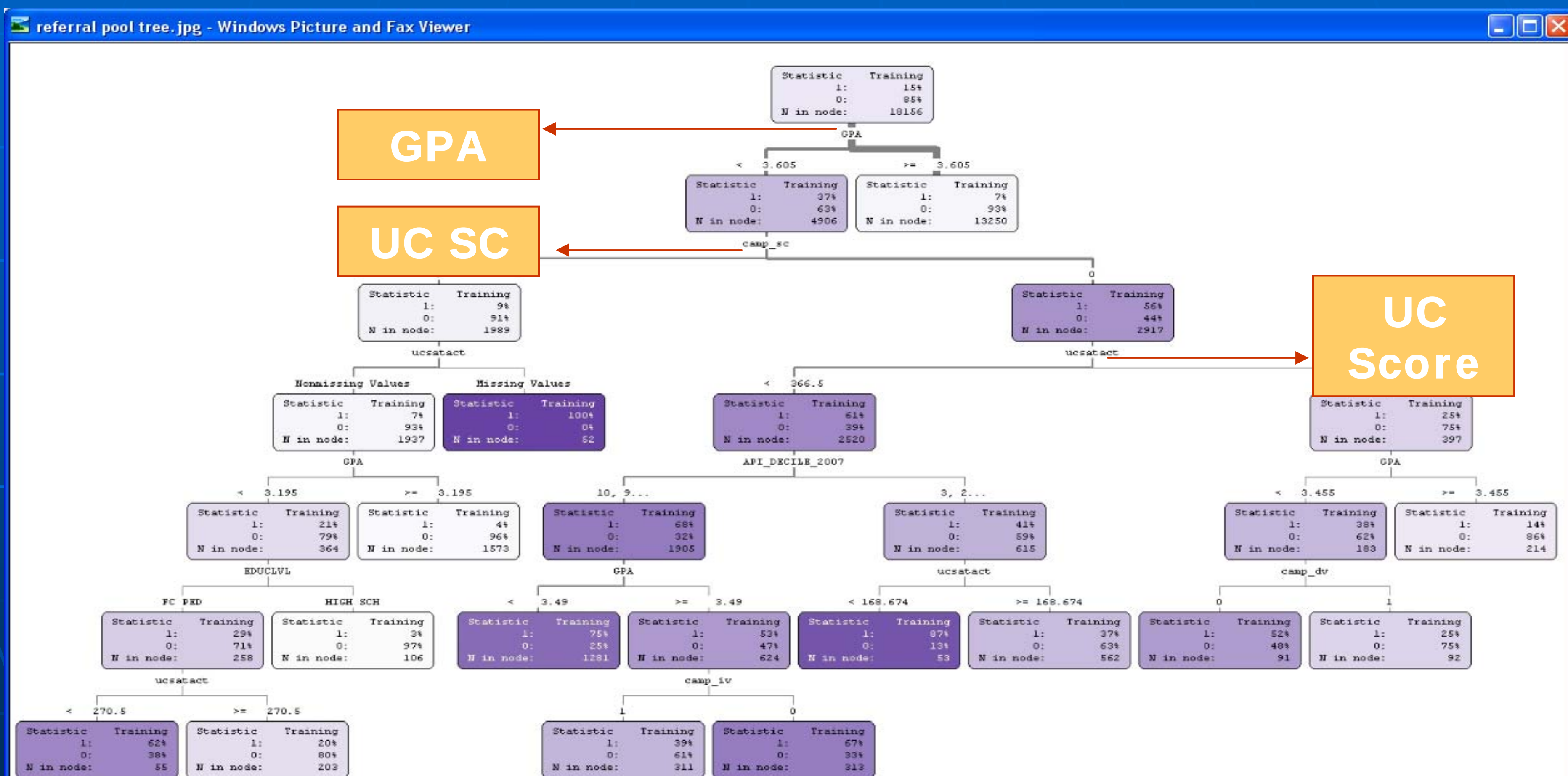
- **Fit Statistics**

Fit Statistics

TARGET	Fit statistics	Statistics Label	Train	Validation	Test
REFP	_AIC_	Akaike's Information Criterion	8388.369	.	.
REFP	_ASE_	Average Squared Error	0.071947	0.073726	0.073328
REFP	_AVERR_	Average Error Function	0.228364	0.235384	0.235675
REFP	_DFE_	Degrees of Freedom for Error	18108	.	.
REFP	_DFM_	Model Degrees of Freedom	48	.	.
REFP	_DFT_	Total Degrees of Freedom	18156	.	.
REFP	_DIV_	Divisor for ASE	36312	27234	27240
REFP	_ERR_	Error Function	8292.369	6410.444	6419.8
REFP	_FPE_	Final Prediction Error	0.072329	.	.
REFP	_MAX_	Maximum Absolute Error	0.990743	0.999043	0.997095
REFP	_MSE_	Mean Square Error	0.072138	0.073726	0.073328
REFP	_NOBS_	Sum of Frequencies	18156	13617	13620
REFP	_NW_	Number of Estimate Weights	48	.	.
REFP	_RASE_	Root Average Sum of Squares	0.26823	0.271526	0.270791
REFP	_RFPE_	Root Final Prediction Error	0.26894	.	.
REFP	_RMSE_	Root Mean Squared Error	0.268585	0.271526	0.270791
REFP	_SBC_	Schwarz's Bayesian Criterion	8763.093	.	.
REFP	_SSE_	Sum of Squared Errors	2612.545	2007.863	1997.446
REFP	_SUMW_	Sum of Case Weights Times F...	36312	27234	27240
REFP	_MISC_	Misclassification Rate	0.103602	0.104502	0.105874

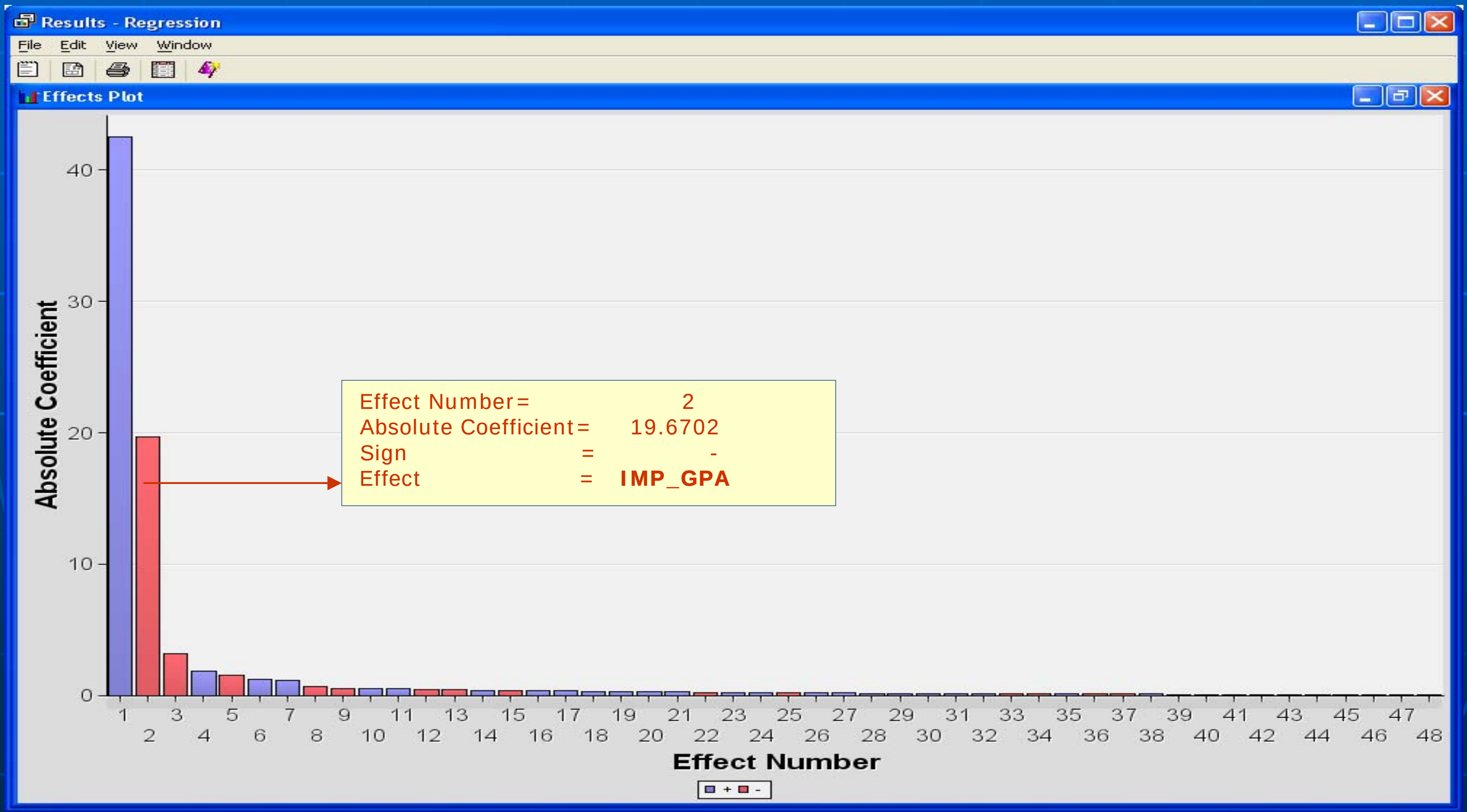
A Case Study Using SAS Enterprise Miner—Model Assessment

- **Importance of a variable in modeling: Tree Map**



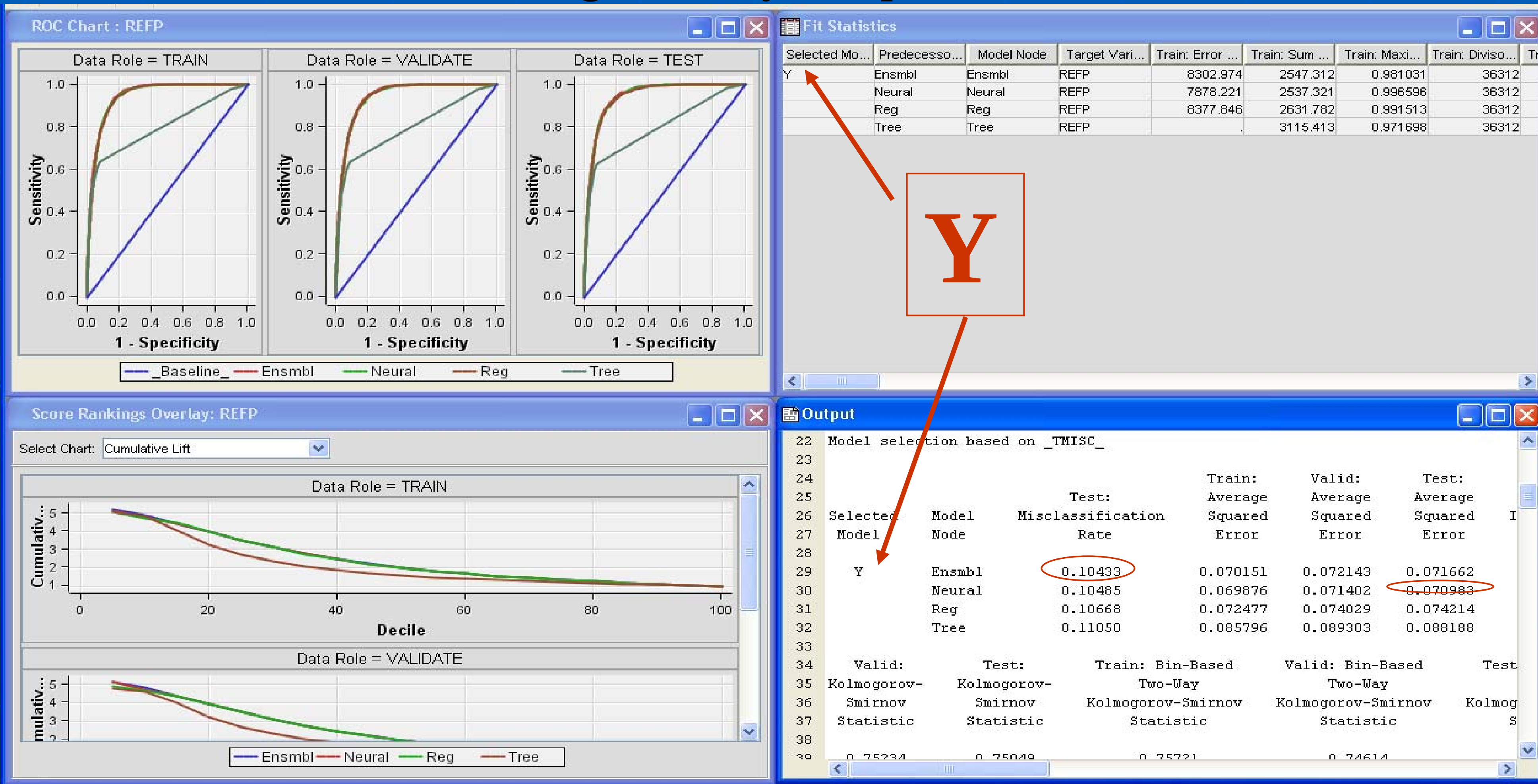
A Case Study Using SAS Enterprise Miner—Model Assessment

- **Tree:** The closer a variable is to the root node, the more prominent in the model.
- **Regression Effects Plot**: Displays a ranked plot of the effect scores most prominent in the model



A Case Study Using SAS Enterprise Miner—Model Comparison

- **Receiver Operating Characteristics (ROC) Chart:** Measure of the predictive accuracy of a model.
- **Fit Statistics, Score Rankings Overlay, Output**



A Case Study Using SAS Enterprise Miner—Scoring and Deployment

- **Scoring: Process to apply the model to new cases**
  - Generate SAS Code
  - Cleaning target data
  - Calculate probability
- **Deployment:**
  - A list of students with a probability equal to or above 40% to two campuses
  - Campuses sent a letter to selected students
  - Campus made offers to those students who responded, allowed campuses to review their applications (Early Referral Pool)

- **Results: Comparison with the actual referral pool**
  - **Accuracy**
    - In terms of the number, accuracy rate: **93%**
    - In terms of individual students,

Predicted Probability	Predicted Referral Pool	Actual Referral Pool	Cumulative Accuracy Rate	Predicted Referral Pool as Cumulative % of Total Population	Actual Referral Pool as Cumulative % of the Entire Referral Pool
90 - 100%	65	52	80.0%	0.1%	0.6%
80 - 89%	353	275	77.9%	0.7%	3.3%
70 - 79%	2,732	2,018	73.9%	5.6%	23.9%
60 - 69%	4,986	3,555	71.3%	10.3%	42.0%
50 - 59%	6,659	4,597	69.0%	13.8%	54.3%
40 - 49%	8,209	5,518	67.2%	17.0%	65.2%

A Case Study Using SAS Enterprise Miner—Results

- **Results: Comparison with the actual referral pool**
  - **Accuracy**
    - **In terms of the number, accuracy rate: 93%**
    - **In terms of individual students,**
  - **Yield**

	2005	2006	2007	2008		
				Total	Early Referral Pool	Traditional Referral Pool[2]
Actual Referral Pool	6,170	6,090	6,923	9,300	1,099	8,201
SIRs[1] from Actual Referral Pool	392	398	465	769	241	528
Referral Pool Yield Rate	6.4%	6.5%	6.7%	**8.3%**	**21.9%**	6.4%
Total SIRs from All Admits	3,691	4,006	4,412	5,770		
Referral Pool SIRs as % of Total SIRs	10.6%	9.9%	10.5%	**13.1%**		

Data Mining Workshop Information

Summer Program for Educators Teaching Data Mining

- **Track 1**: Basic SAS programming; **Track 2**: SAS Enterprise Miner
- **Location**: CSU Long Beach, the SAS Campus in Cary, NC
- **Time**: Early August
- **Registration Fee**: No
- **Text Books**: Free
- **Breakfasts and Lunches**: Every day and free
- Invited people only
- Invitation letter is sent out early February
- Contact the SAS Institute in January
- Contact person: Susan Walsh, susan.walsh@sas.com