Modeling Freshmen Outcomes using
SAS Enterprise Miner

Nora Galambos, PhD
Office of Institutional Research, Planning & Effectiveness

NEAIR Annual Conference
Newport, RI 2013

Stony Brook University

Data Mining

- Knowledge discovery by extracting information from large amounts of data

- Uses analytic tools for data-driven decision making

- Uses modeling techniques to apply results to future data

- Incorporates statistics, pattern recognition, and mathematics

Enterprise Miner Interface

Project Panel →

Properties Panel →

SEMMA Tools Palette ←

Diagram Workspace

Sample
Explore
Modify
Model
Assess

Project Flow Workspace

Data Nodes

Merge Node   **SAS Code Node**

Stony Brook University

Variable Selection View

Merge Node View

Merge Variable Specification

SAS Code Node

SAS Code Output

Results - Node: SAS Code Diagram: NEAIR

File Edit View Window

Output

The FREQ Procedure

twoplus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2264	13.50	2264	13.50
1	14501	86.50	16765	100.00

twoplus2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	14501	86.50	14501	86.50
1	2264	13.50	16765	100.00

```
*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Stony Brook University

Filter and Data Partition Nodes

Data Node Variable Selection/Configuration

Stony Brook University

Filter Node Properties Panel

- Filter rare values
- Choose whether to keep missing values
- Create cutoffs

.. Property	Value
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
⊟ Class Variables	
Class Variables	[...]
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percent	0.01
Maximum Number of Levels	25
⊟ Interval Variables	
Interval Variables	[...]
Default Filtering Method	Standard Deviations from th
Keep Missing Values	Yes
Tuning Parameters	[...]
Score	
Create score code	Yes
Status	

Stony Brook University

Filter Node Variable Selection

Interactive Categorical Filter

Filtering Class Categories
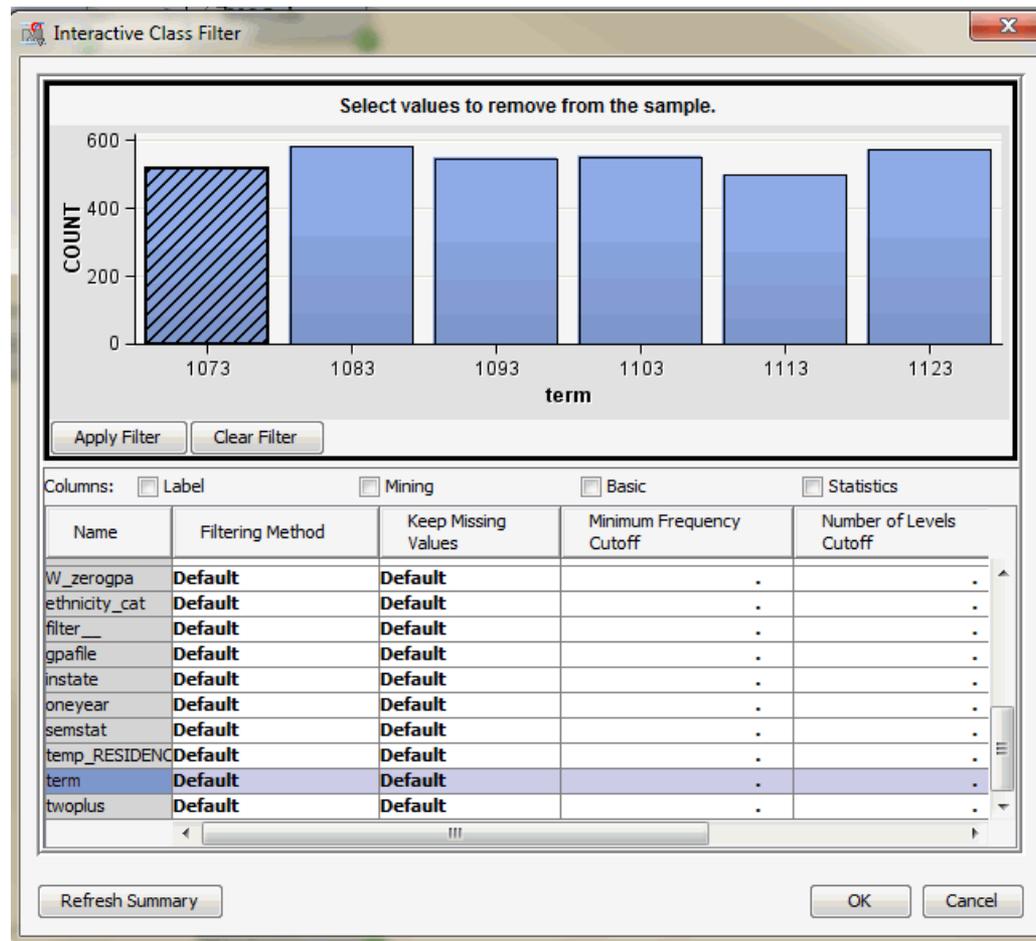
Interactive Interval Filter

Training, Validation, and Test Partitions

Find the correct level of model complexity. A model that is not complex enough may lack the flexibility to represent the data, underfitting. When the model is too complex it can be influenced by random noise, overfitting.

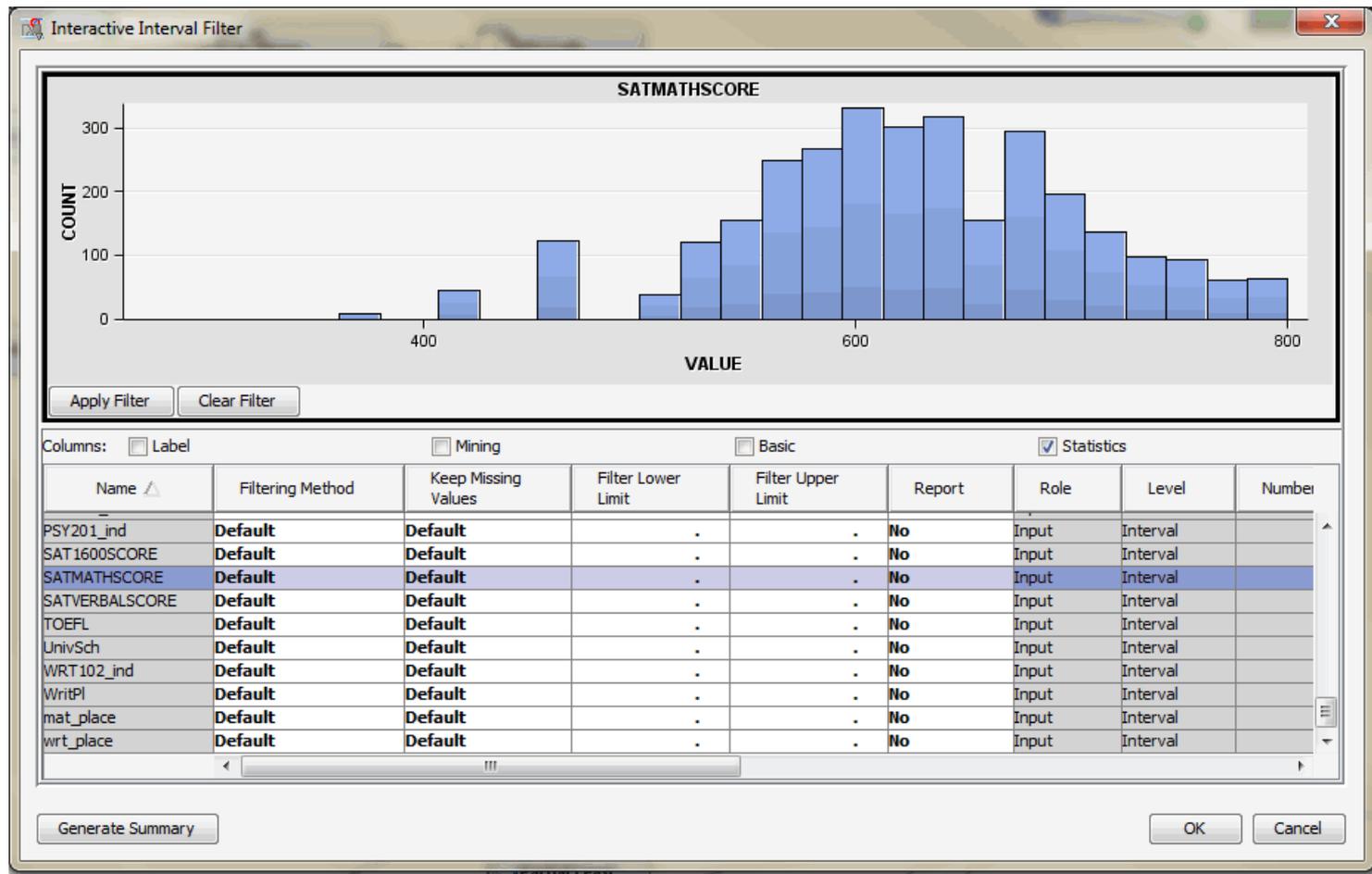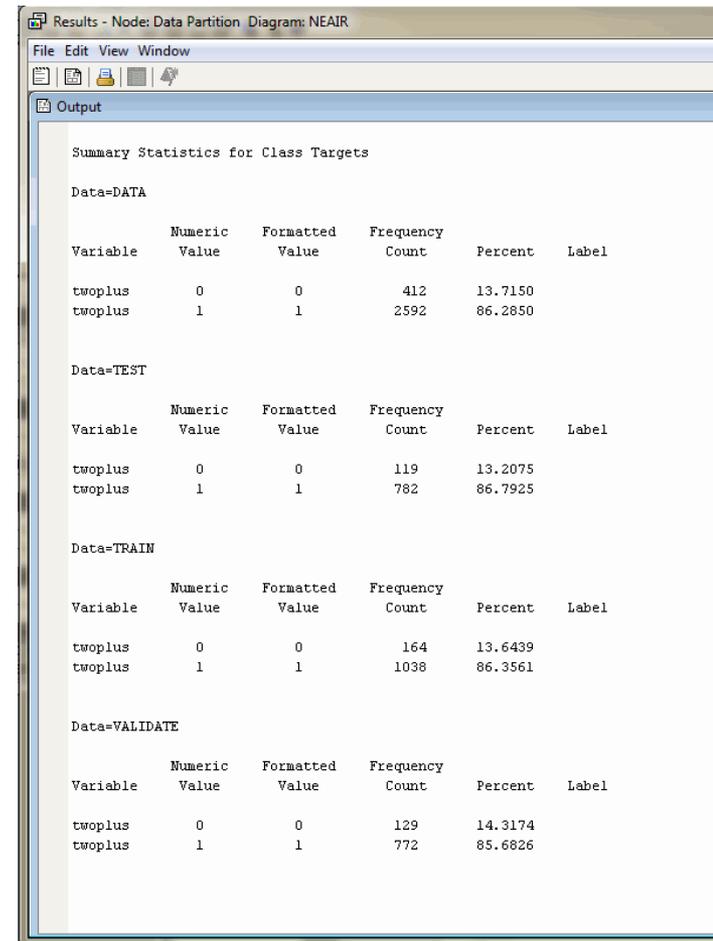Partitioning is used to avoid over- or underfitting. The training partition is used to build the model. The validation partition is set aside and is used to test the accuracy and fine tune the model. The test partition is used for evaluating how the model will work on new data.

Results - Node: Data Partition Diagram: NEAIR
File Edit View Window
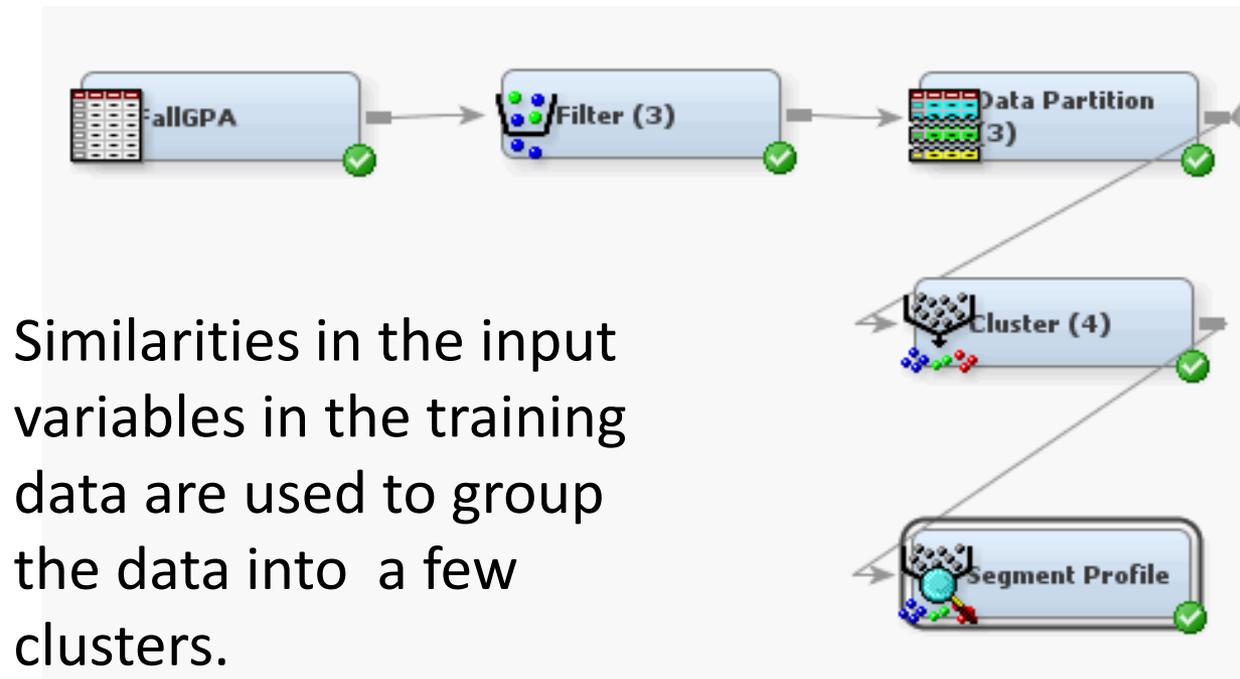
Output

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
twoplus	0	0	412	13.7150	
twoplus	1	1	2592	86.2850	

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
twoplus	0	0	119	13.2075	
twoplus	1	1	782	86.7925	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
twoplus	0	0	164	13.6439	
twoplus	1	1	1038	86.3561	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
twoplus	0	0	129	14.3174	
twoplus	1	1	772	85.6826	

Stony Brook University

Cluster Analysis and Segment Profile Nodes

Similarities in the input variables in the training data are used to group the data into a few clusters.

Stony Brook University

Cluster Analysis Results
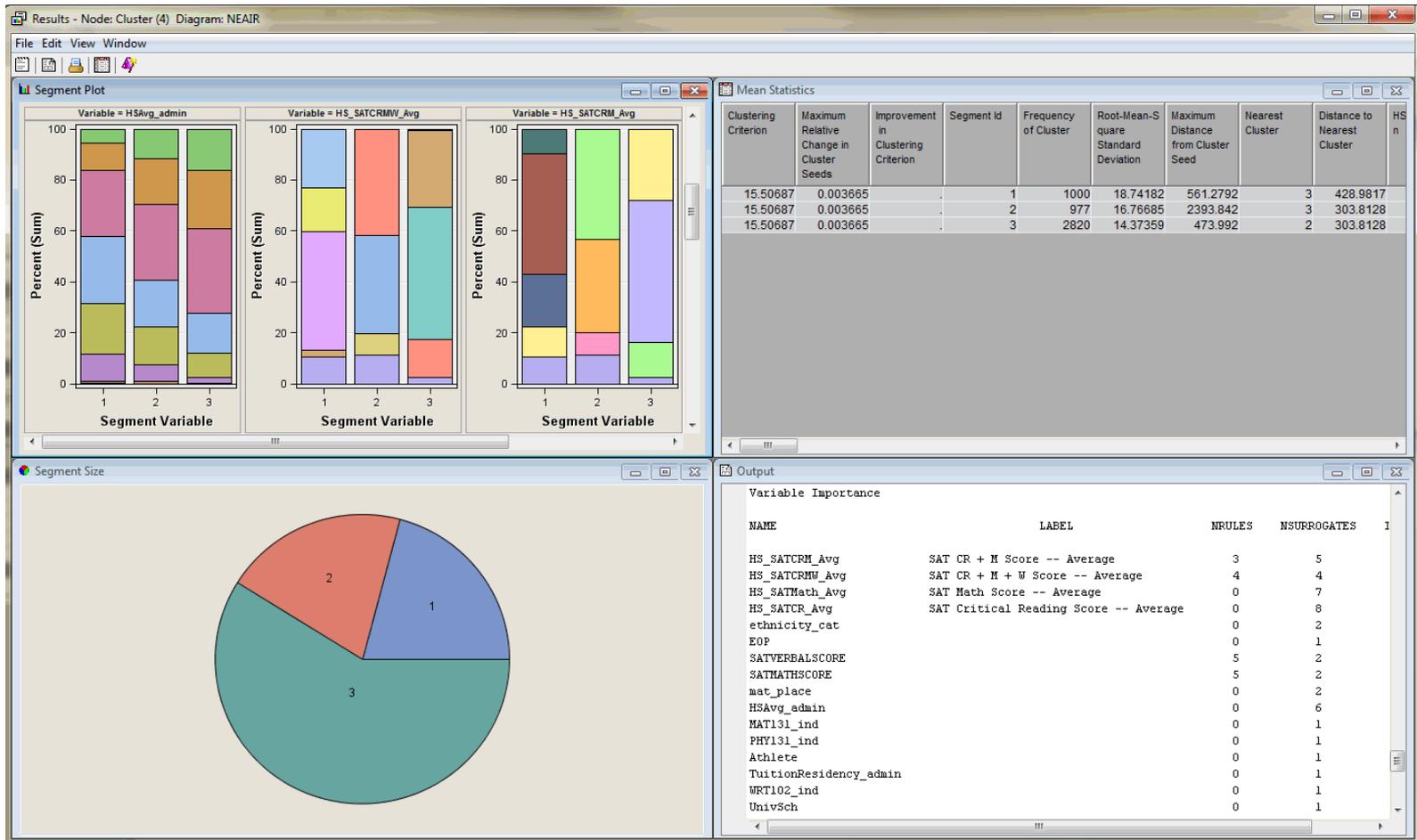
Segment Profile

Segment Profile Detail

Stony Brook University

Segment Profile Graphic Comparisons
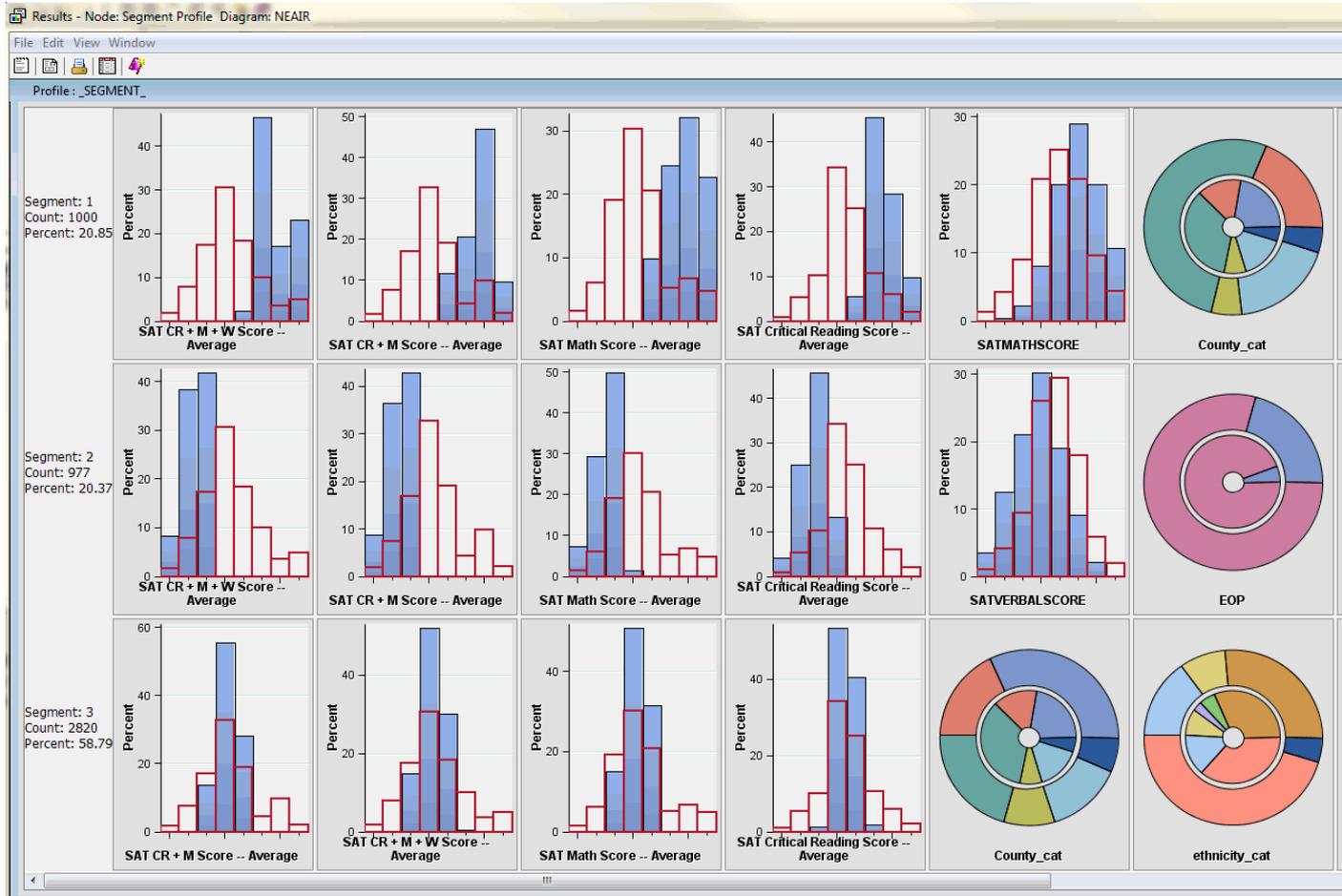
Full Enterprise Miner Model
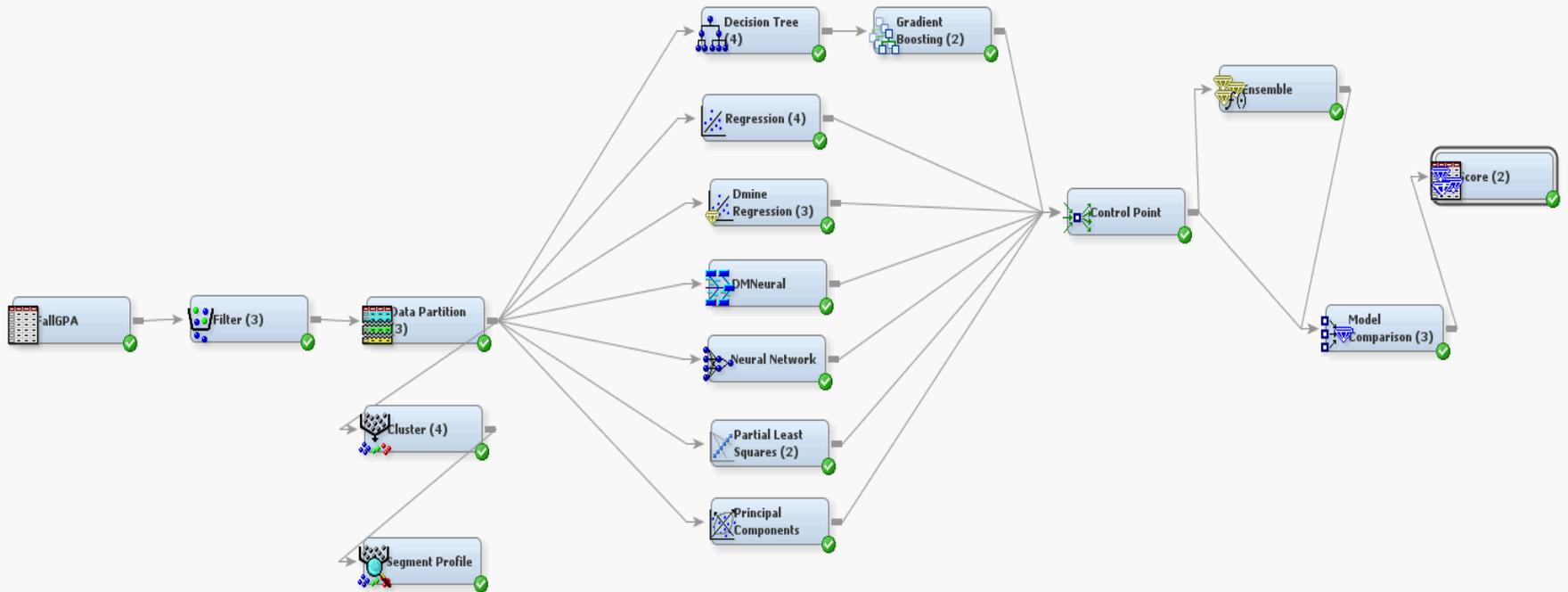
Decision Tree Configuration

Splitting Rule	
Interval Criterion	ProbF
Nominal Criterion	ProbChisq
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	5
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Kass Adjustment	Before
Inputs	No
Number of Inputs	1
Split Adjustment	Yes
Output Variables	
Leaf Variable	Yes

Stony Brook University

Interactive Decision Tree Building:
Categorical Outcome
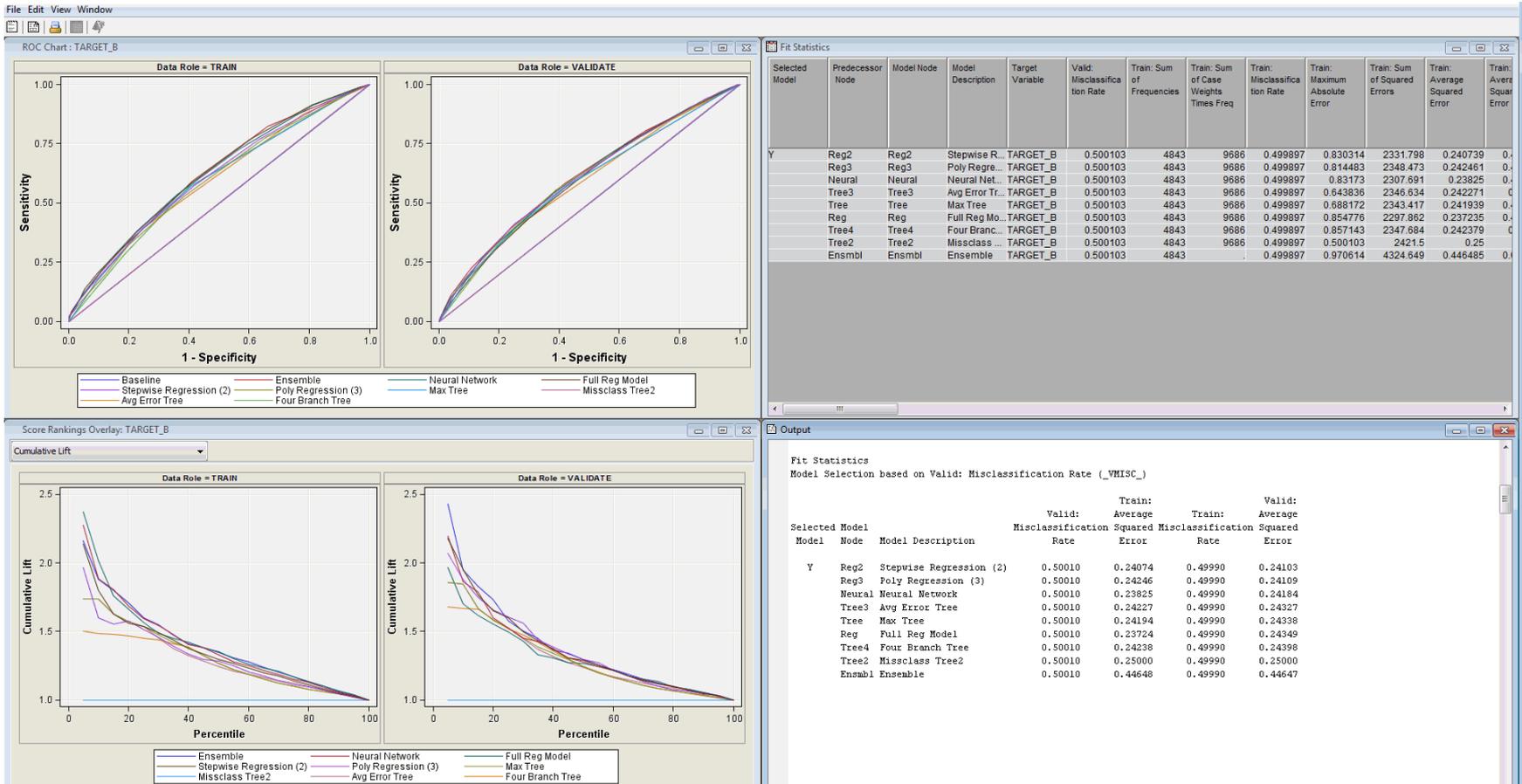## First Semester Freshmen GPA above/below 2.00
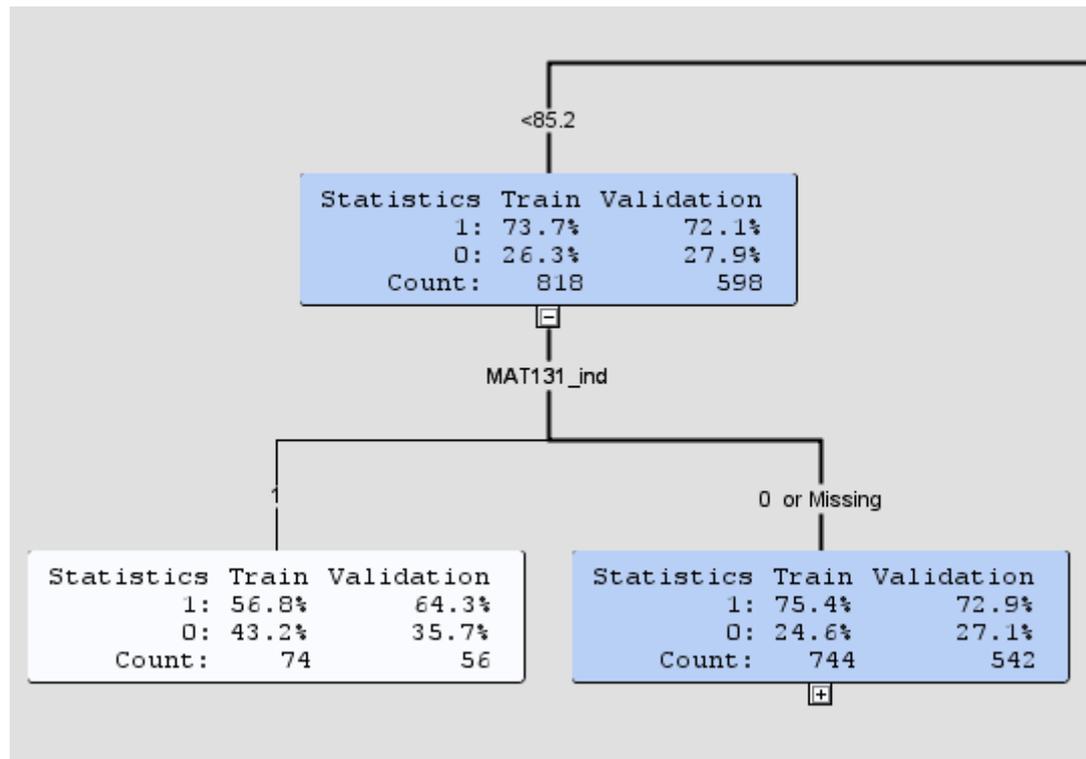
Adding Tree Branches and Leaves
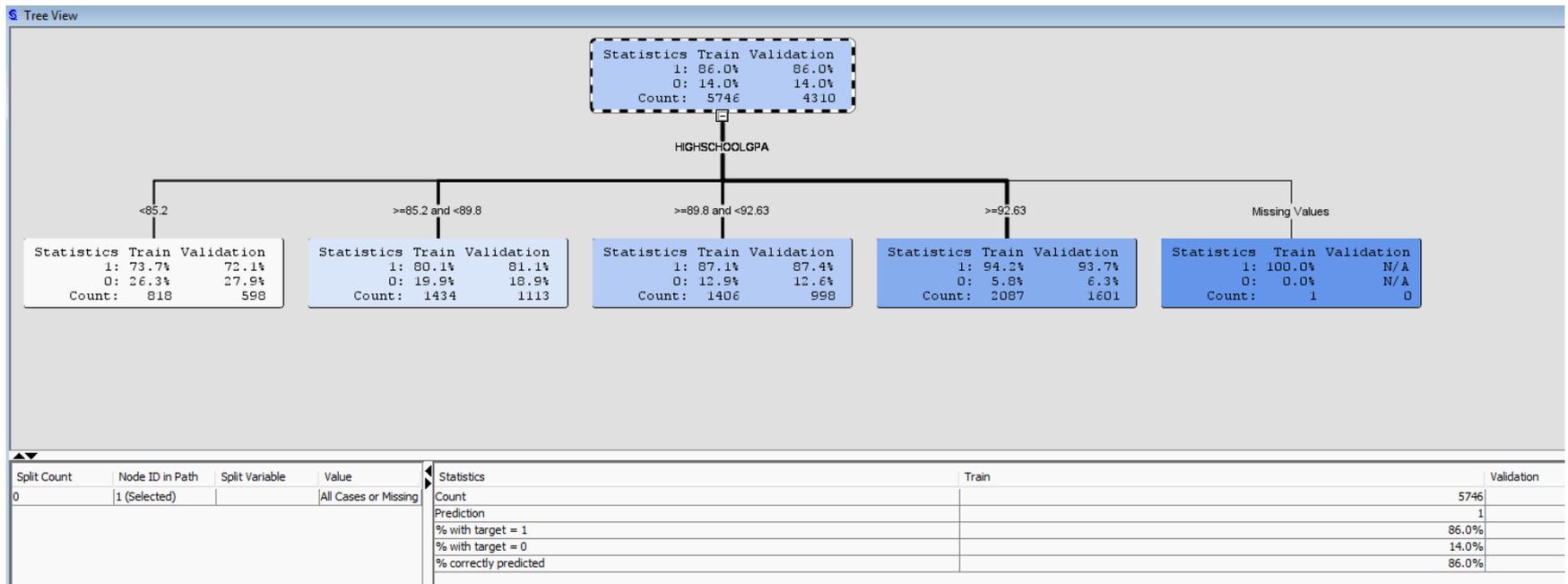
Evaluating a Decision Tree
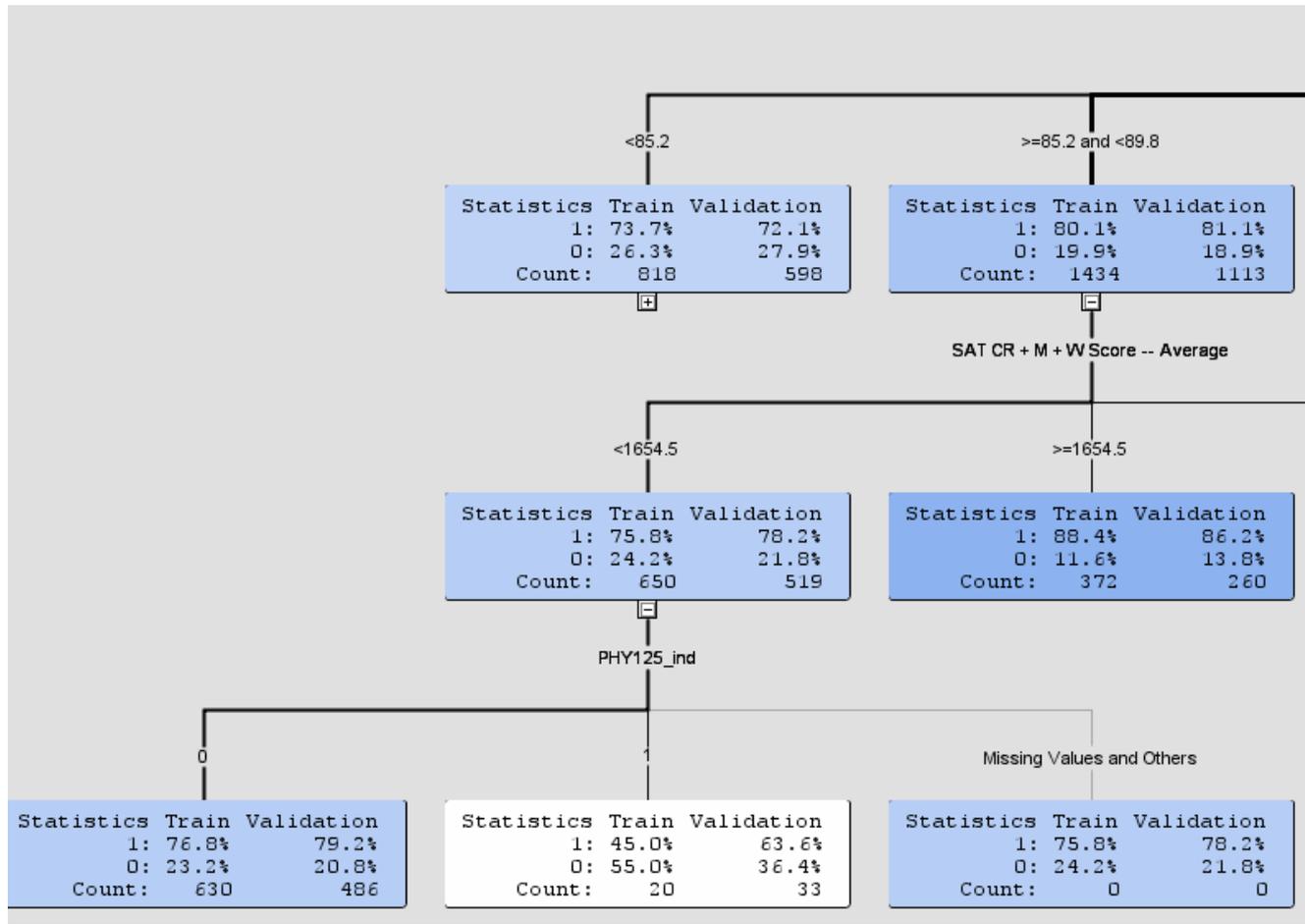# with a Categorical Outcome
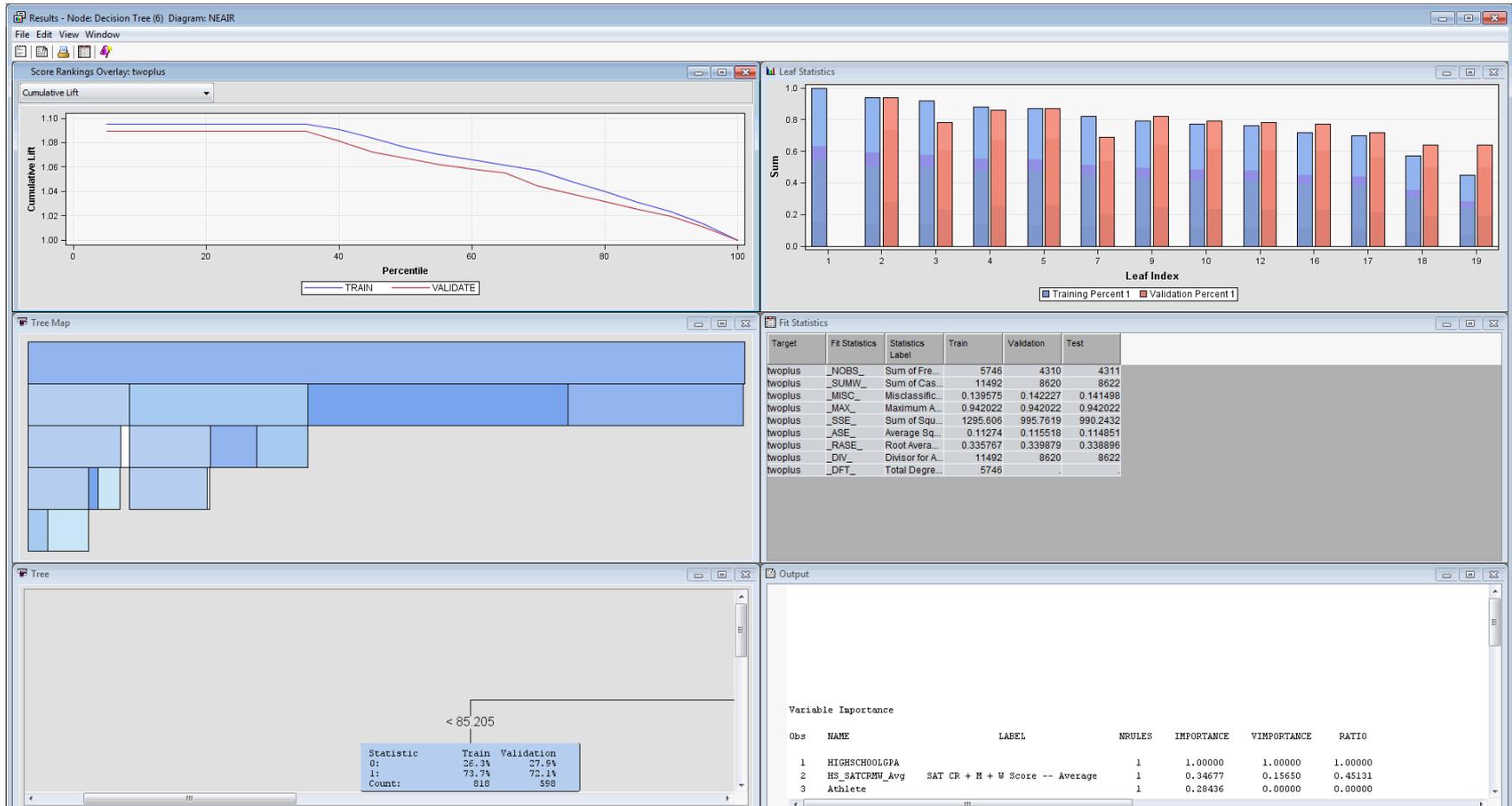
Receiver Operator Curves and Cumulative Lift

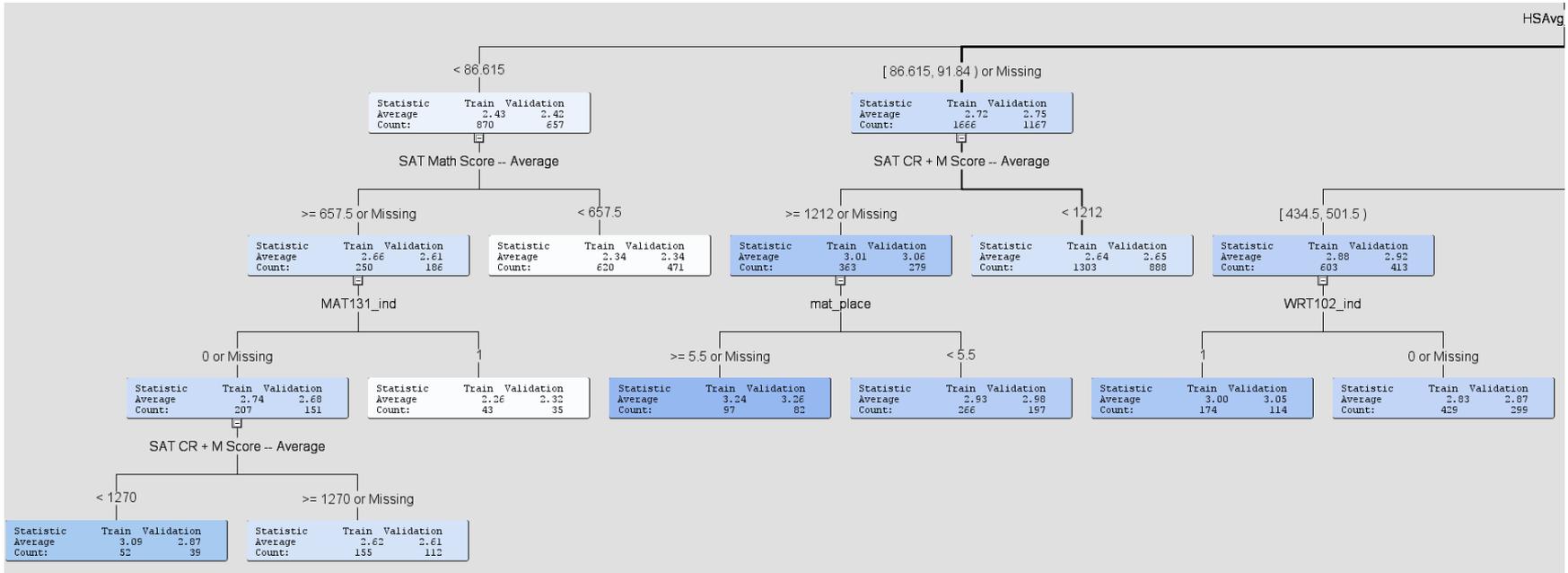# Decision Tree with Interval Outcome

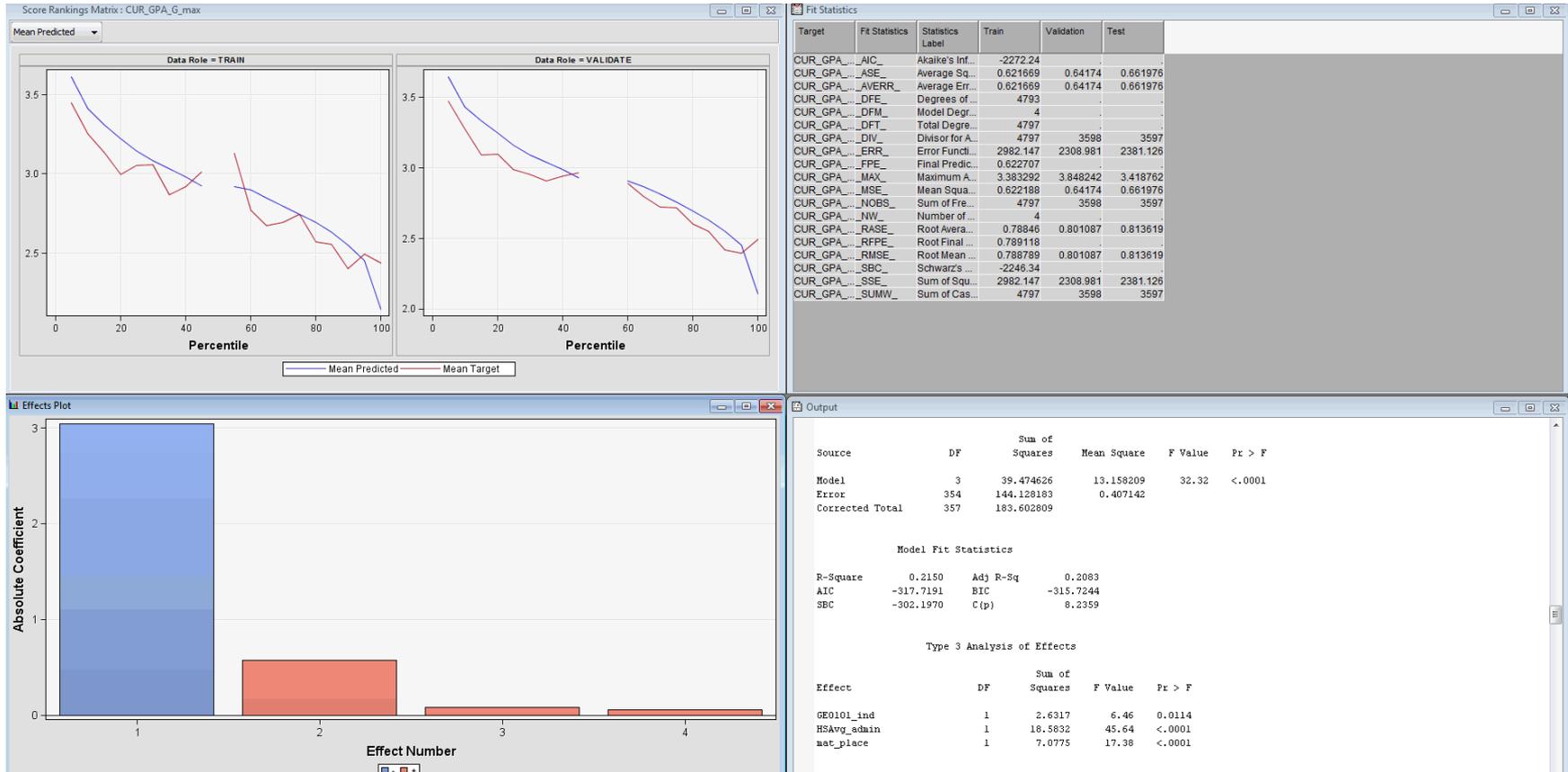# Using Decision Tree to Predict
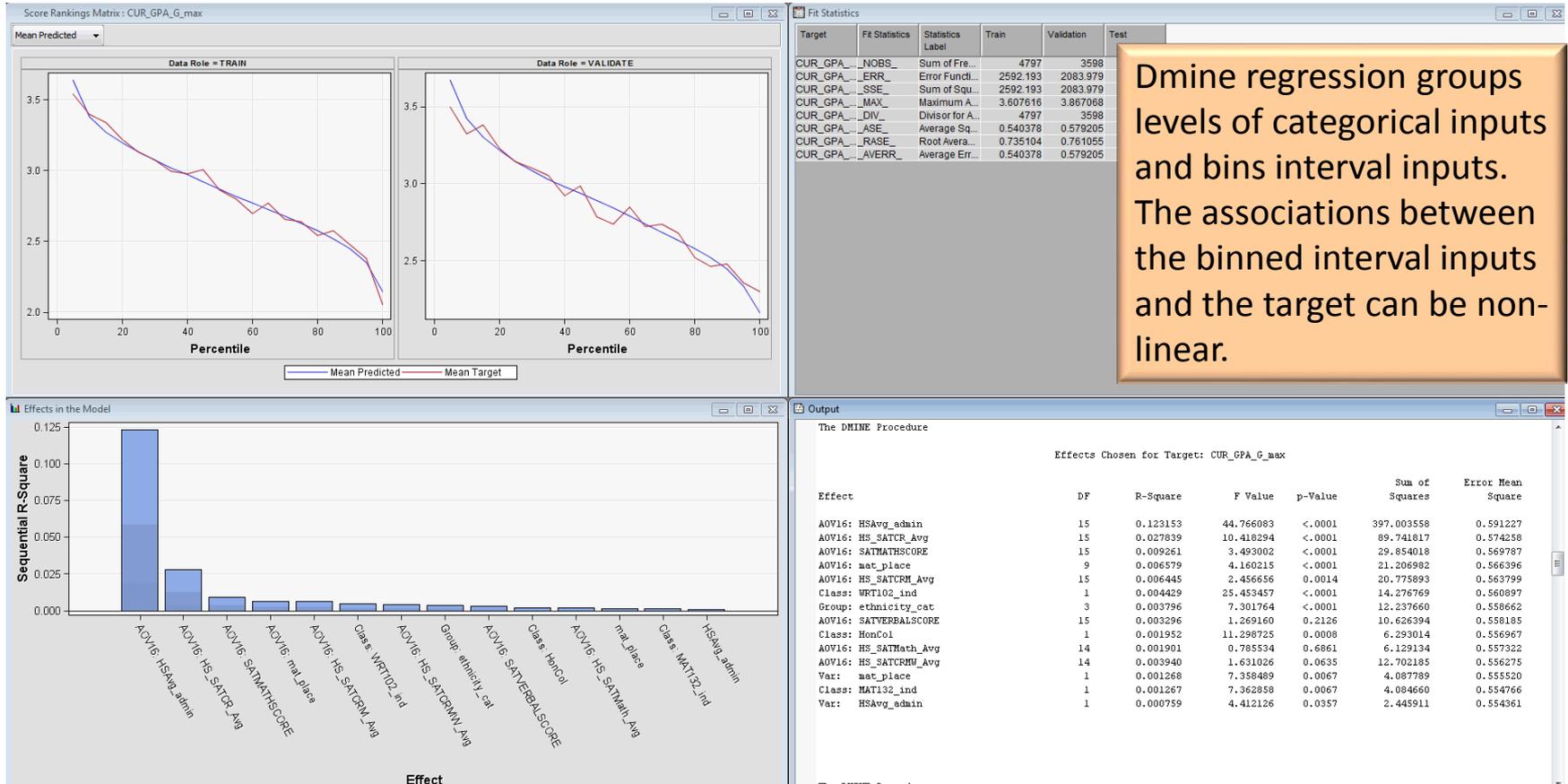# First Semester Freshmen GPA

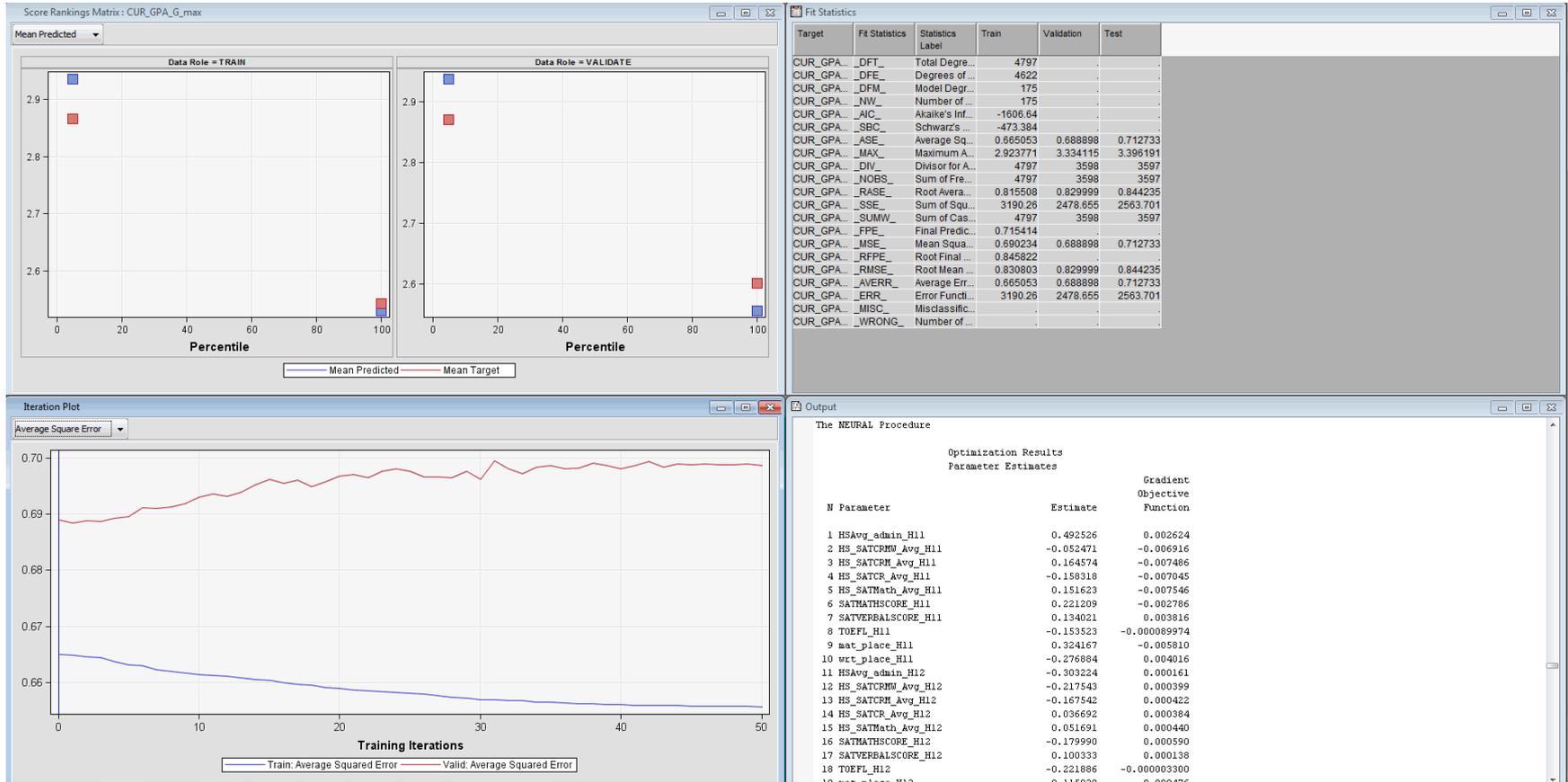# Decision Tree View

Linear Regression Model
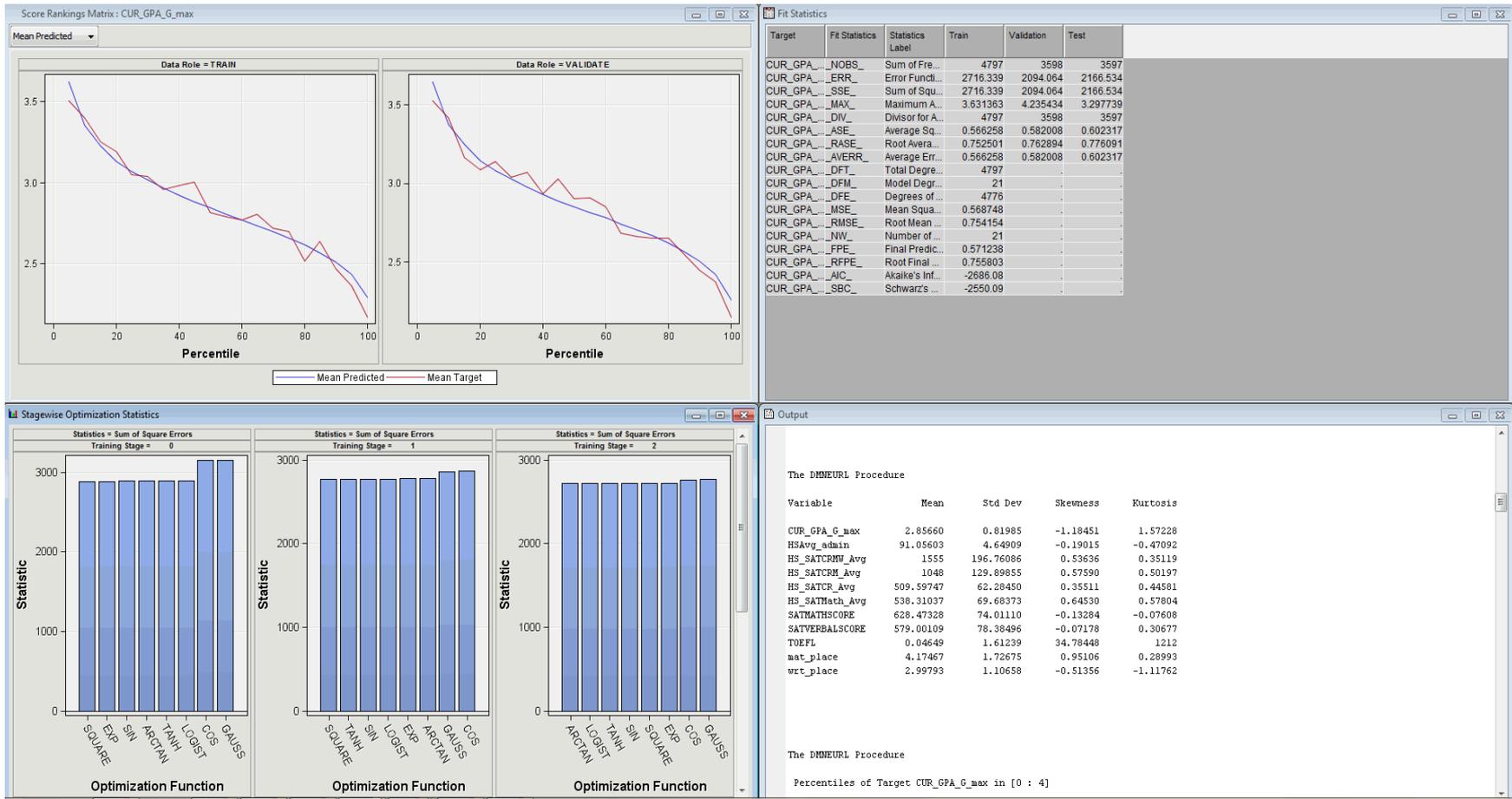
Dmine Regression Model

Dmine regression groups levels of categorical inputs and bins interval inputs. The associations between the binned interval inputs and the target can be non-linear.
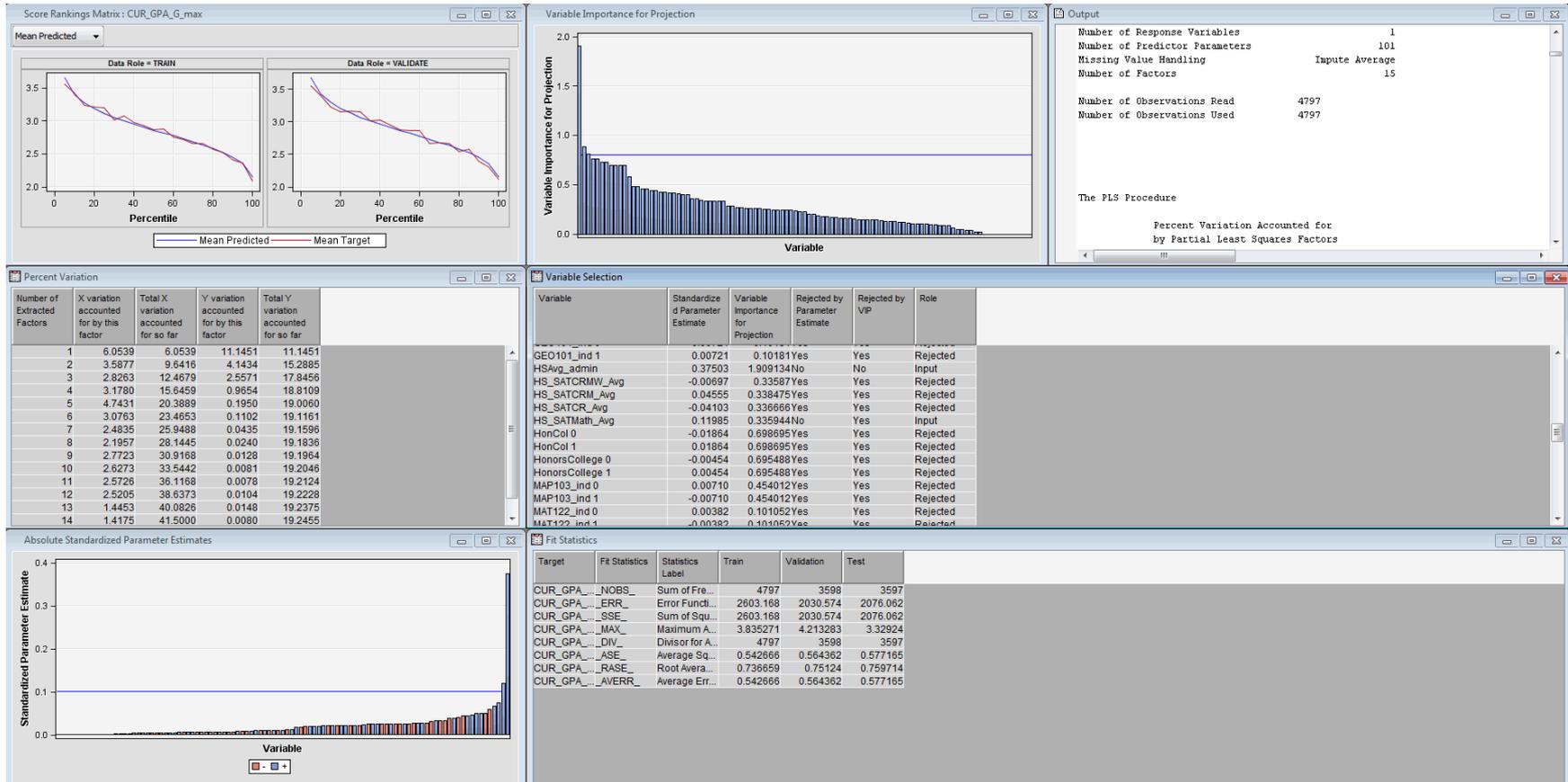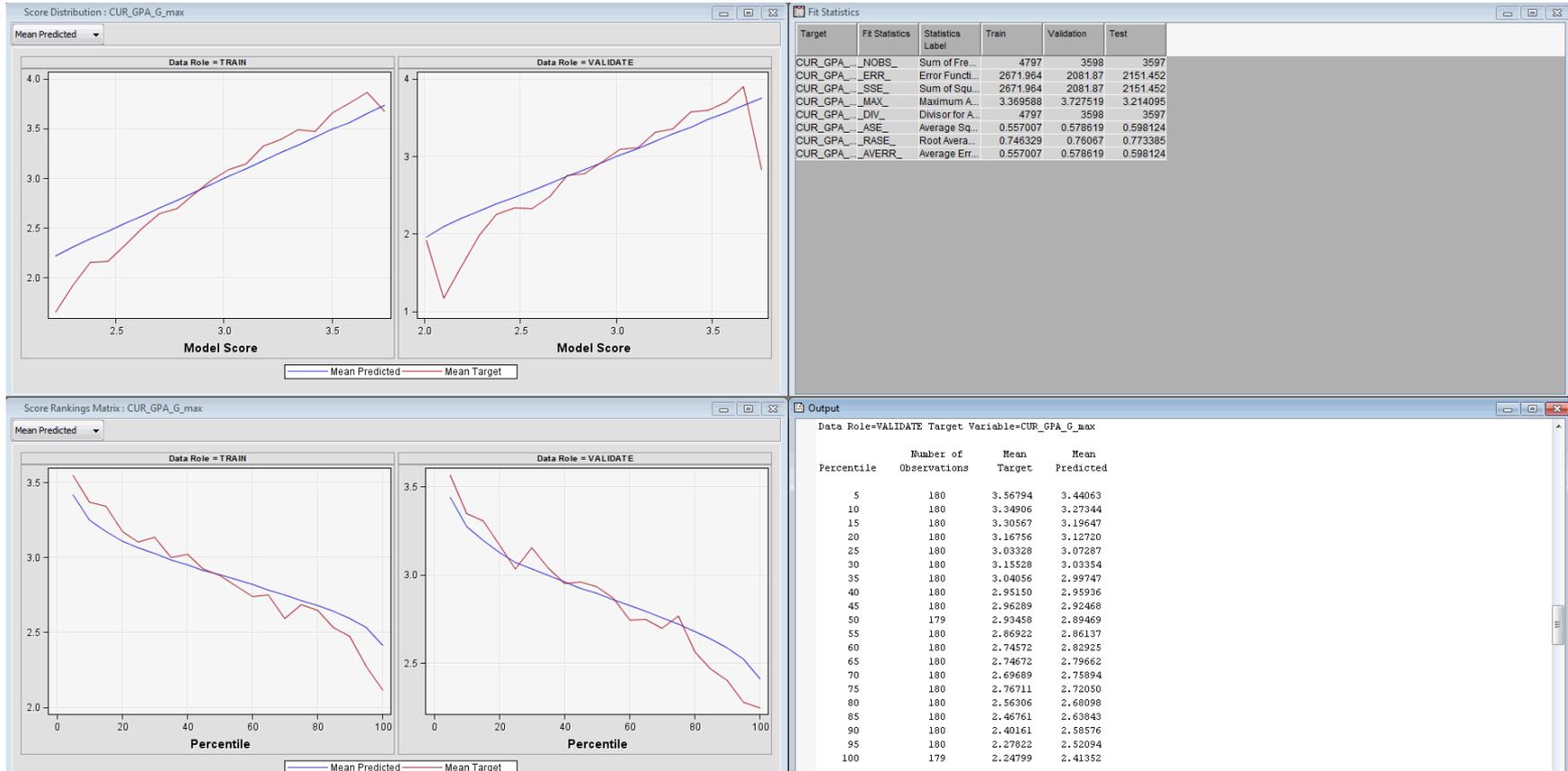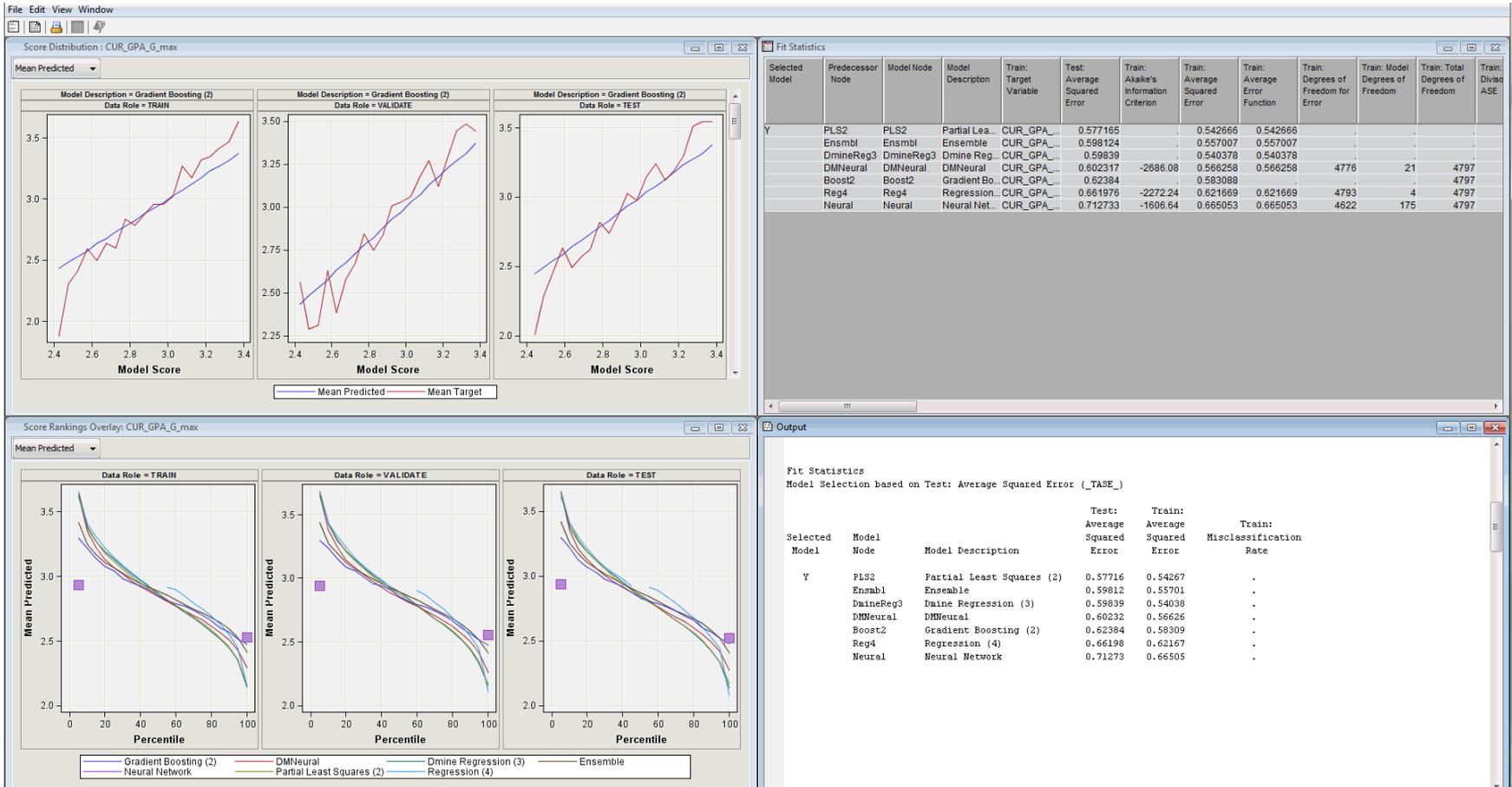
Neural Network Model
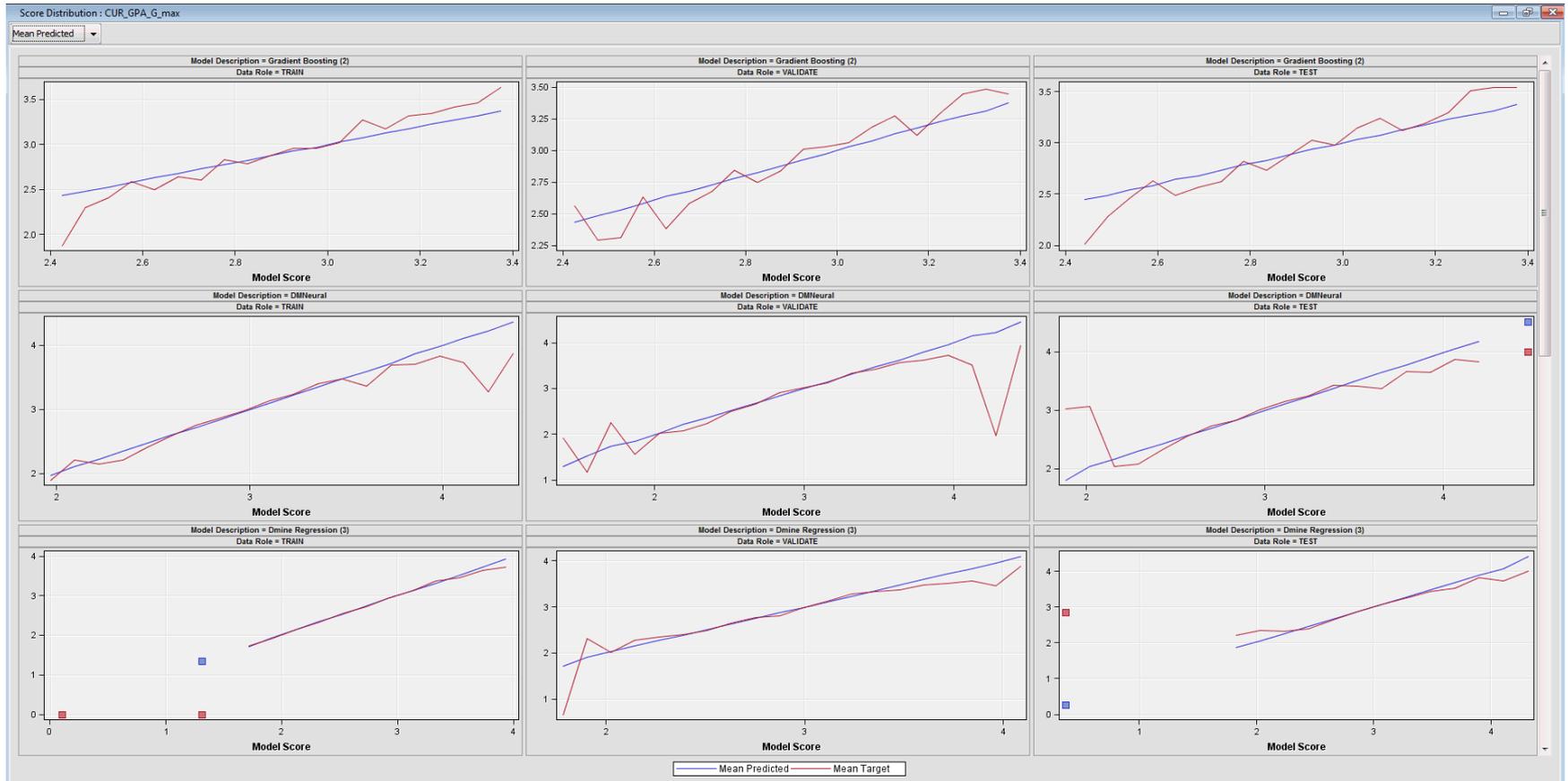
Dmneural Model

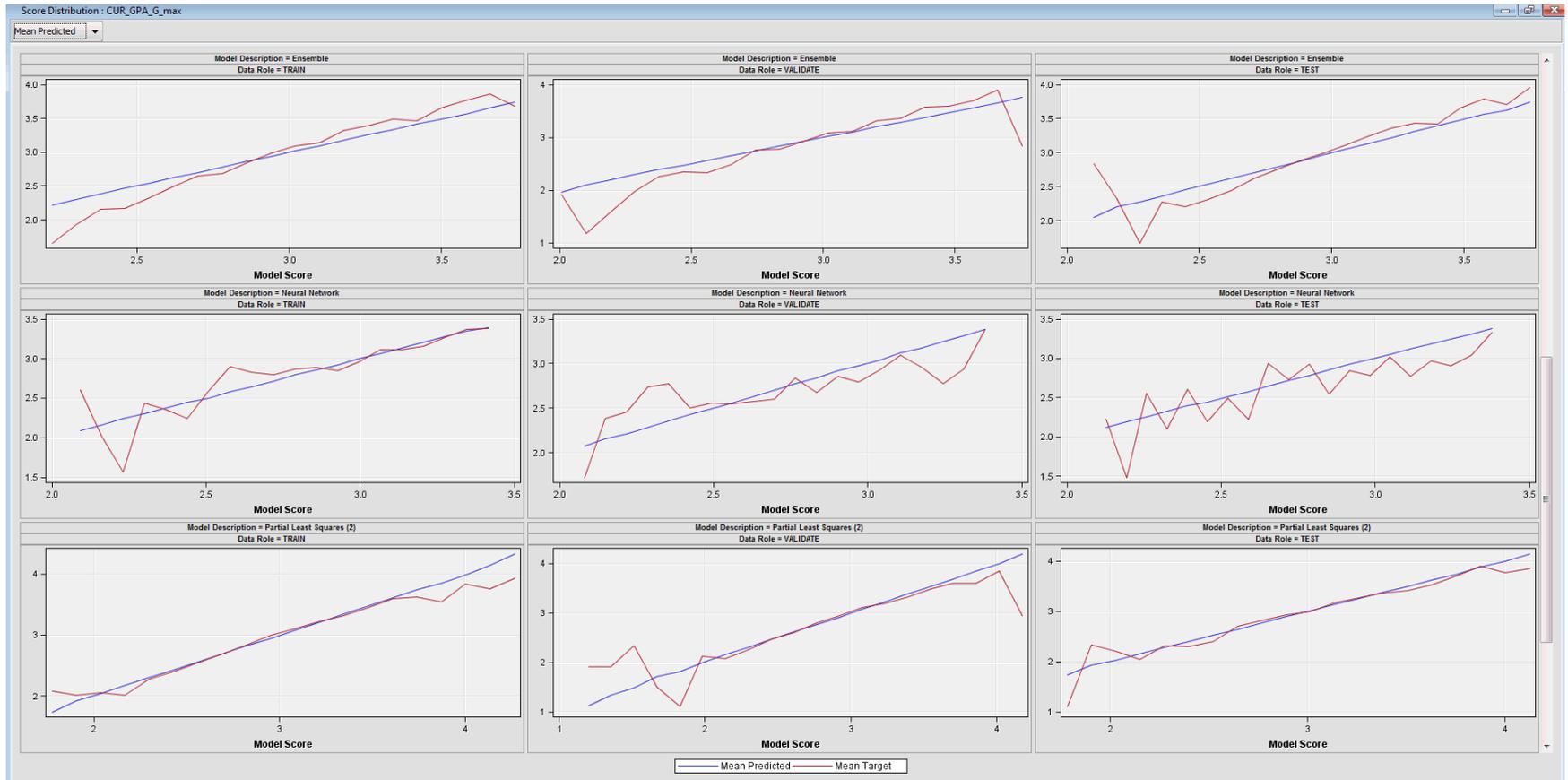# Partial Least Squares Regression Model

Ensemble Node Model

Model Comparison

Model Comparison Graphs 1

Model Comparison Graphs 2

Model Comparison Graphs 3

Score Node Output
for Partial Least Squares Model

```
Interval Variable Summary Statistics

Variable Name=P_CUR_GPA_G_max

Statistics     Label                     TRAIN      VALIDATE       TEST

  MEAN         Mean                       2.86          2.87       2.85
  STD          Standard Deviation         0.36          0.37       0.37
  N                                    4797.00       3598.00    3597.00
  MIN          Minimum                    1.70          1.12       1.72
  P25          25th Percentile            2.61          2.61       2.59
  MEDIAN       Median                     2.84          2.85       2.82
  P75          75th Percentile            3.08          3.11       3.08
  MAX          Maximum                    4.33          4.25       4.19
```

SAS Code
to Run Partial Least Squares Model on New Data

Model Package