

The project was to review purchase data for the Widget Buyers and determine which buyers are most likely to purchase widgets. This data was based on the historical data that supplied from previous purchases. Using SAS Enterprise Miner, there were three models used for comparison – Decision Tree, Neural Network, and Regression Line. The target determined was the “WidgBuy” and level was “Binary”. These factors were initialized upon the import of the data for all three models to utilize in the evaluation.

### The Decision Tree

A decision tree is a power form of multiple variable analysis. Decision trees can be used with several variables for manipulation. Based on the strength of the entropy the variable can be of marginal relevance or very relevant to create rules. If the variable is determined to be relevant a split can be made thusly paring down the data to only the critical relevant data. Each of the rules create a hierarchy of nodes to allow for a prediction with a reduced number of variables.

Event Classification Table			
Data Role=TRAIN Target=WidgBuy Target Label=WidgBuy			
False Negative	True Negative	False Positive	True Positive
0	6	3	11

Figure 1 Decision Tree Confusion Matrix

The confusion matrix for the Decision Tree showed a misclassification rate of 15 % (3/20) possibility of a Type I error by incorrectly identifying the buyer with 3 False Positives. While the Type I error isn't quite as dangerous as the Type II, the object is to be able to make a prediction with as few errors as possible.

### The Neural Network

The Neural Network is able to capture complex relationships while acquiring knowledge through learning. Each network knowledge is stored in the hidden layer and connected through weights. Neural Networks are most useful when the desired output is unknown by looking for linear and non-linear relationships.

The classification table demonstrates no false positives or negatives making it an optimum choice with no errors.

Event Classification Table			
Data Role=TRAIN Target=WidgBuy Target Label=WidgBuy			
False Negative	True Negative	False Positive	True Positive
0	9	0	11

Figure 2 Neural Network Confusion Matrix

## The Regression Model

The Linear Regression model predicts scores based on the predictor variable and criterion variable by trying to find the best-fitting straight line through the points. This confusion matrix had the exact same data result as the neural network model.

Event Classification Table

Data Role=TRAIN Target=WidgBuy Target Label=WidgBuy

False Negative	True Negative	False Positive	True Positive
0	9	0	11

Figure 3 Regression Confusion Matrix

## The Comparison

The cumulative lift shows the neural network and regression have the same lift at 1.8 meaning we should stop training at 55.

The Neural Network as stated in model comparison is the best model to use based on the data. There were three model nodes. Neural network and Regression were nearly identical in the average squared error. In the ROC curve it showed sensitivity to true positives for the .5 cut offs.



Figure 4 Cumulative Lift

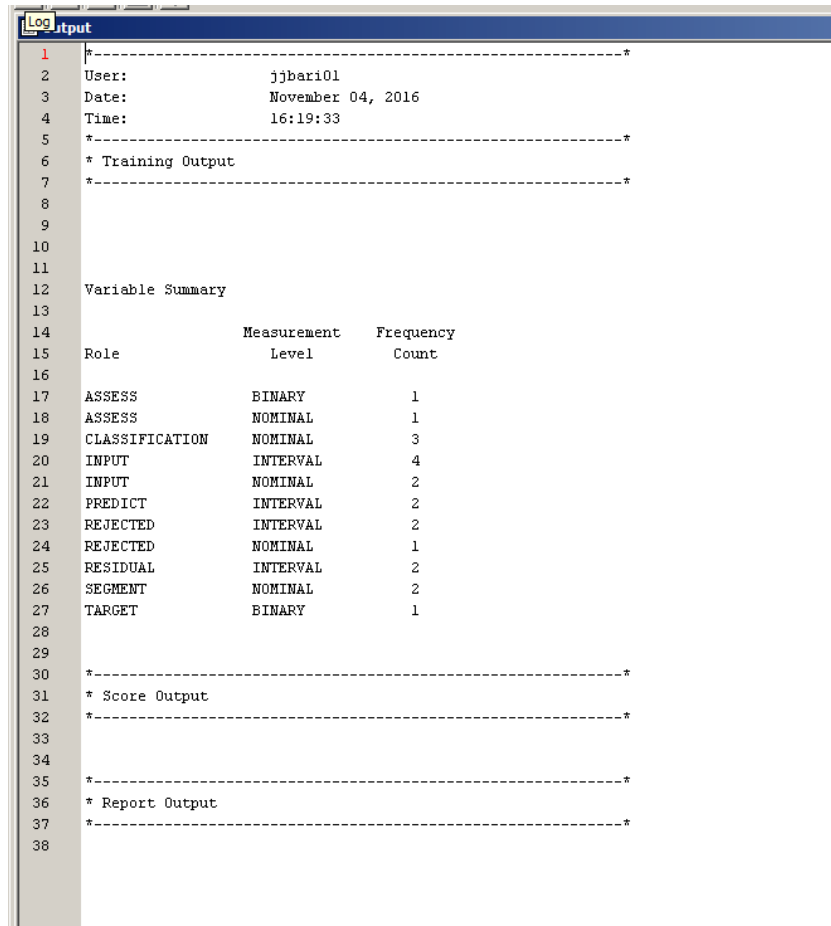
The assumptions made are how people make decision to buy items. The available information shows that people will buy widgets and can be influenced through their economic status, age, location and norms.

The weights in the neural network modules provide a variance between -8.4165 which has a negative impact to the 4.3 with the most influence being the age and income the most likely determinates.

The Node rules showed that income has the strongest relevance to a predicted buy at .89(close to one). Their age being the second determining factor – age over 30 most likely because they have more discretionary funds. The location didn't make a mention as relevant in the purchase of a widget.

## The SAS Code

The SAS code was difficult to determine which parts were relevant. While it's an overview of the frequency, I couldn't determine how I could change the measurements to affect the frequency to manipulate the data.

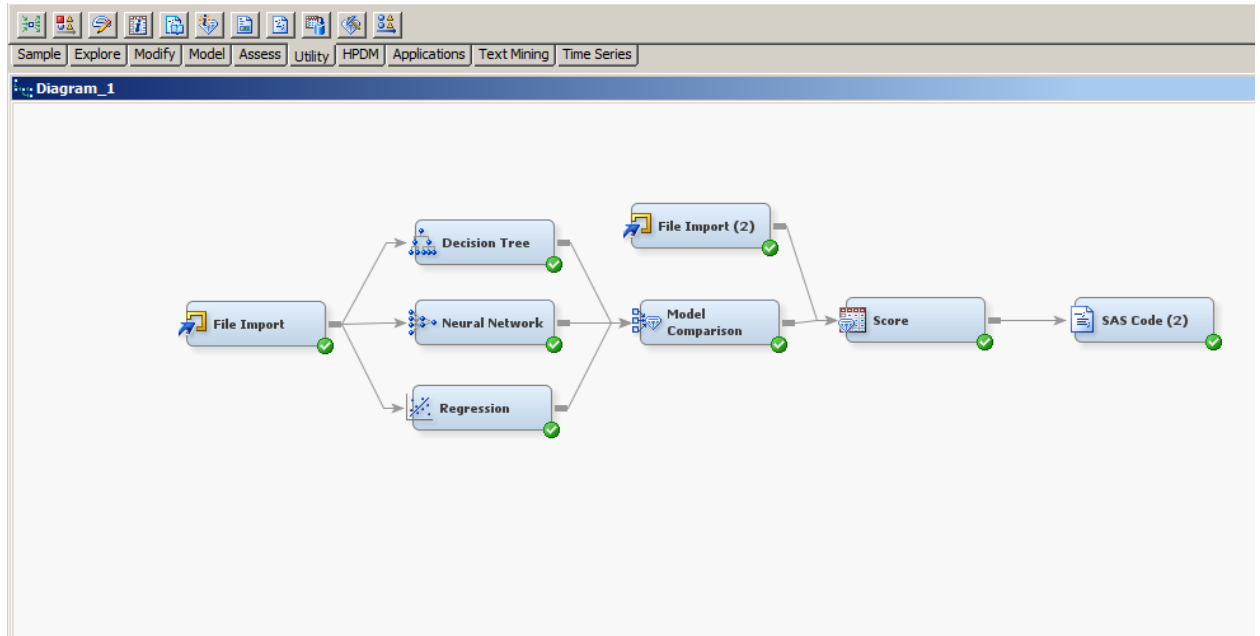


The screenshot shows the SAS Log Output window. It contains a 'Variable Summary' table and several section headers. The 'Variable Summary' table lists variables, their roles, measurement levels, and frequencies. Section headers include 'Training Output', 'Score Output', and 'Report Output'.

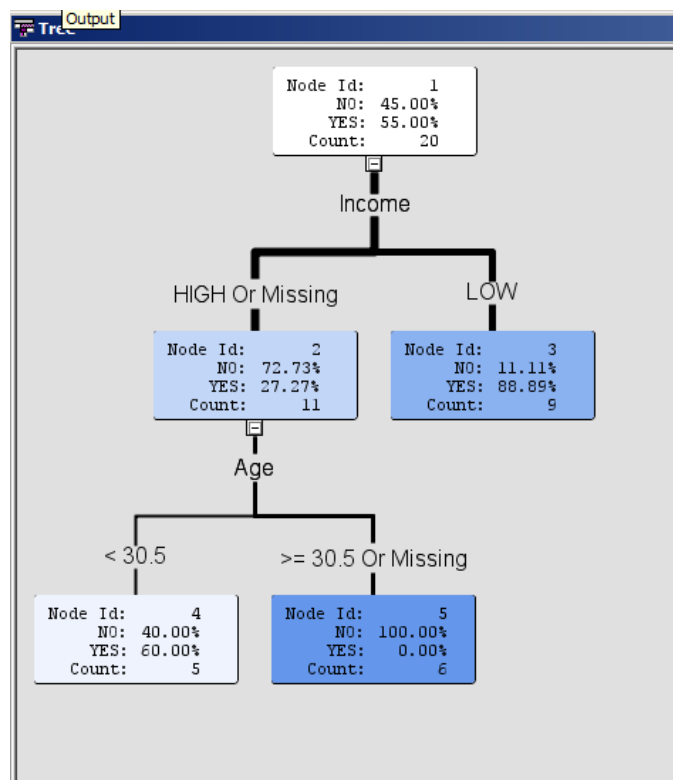
Variable	Role	Measurement Level	Frequency Count
ASSESS		BINARY	1
ASSESS		NOMINAL	1
CLASSIFICATION		NOMINAL	3
INPUT		INTERVAL	4
INPUT		NOMINAL	2
PREDICT		INTERVAL	2
REJECTED		INTERVAL	2
REJECTED		NOMINAL	1
RESIDUAL		INTERVAL	2
SEGMENT		NOMINAL	2
TARGET		BINARY	1

Figure 5 SAS Code

a) Workflow/Diagram



b) The Tree diagram from the decision tree



c) Node Rules

```

Node Rules
1  *-----*
2  Node = 3
3  *-----*
4  if Income IS ONE OF: LOW
5  then
6  Tree Node Identifier = 3
7  Number of Observations = 9
8  Predicted: WidgBuy=Yes = 0.89
9  Predicted: WidgBuy=No = 0.11
10
11 *-----*
12 Node = 4
13 *-----*
14 if Income IS ONE OF: HIGH or MISSING
15 AND Age < 30.5
16 then
17 Tree Node Identifier = 4
18 Number of Observations = 5
19 Predicted: WidgBuy=Yes = 0.60
20 Predicted: WidgBuy=No = 0.40
21
22 *-----*
23 Node = 5
24 *-----*
25 if Income IS ONE OF: HIGH or MISSING
26 AND Age >= 30.5 or MISSING
27 then
28 Tree Node Identifier = 5
29 Number of Observations = 6
30 Predicted: WidgBuy=Yes = 0.00
31 Predicted: WidgBuy=No = 1.00
32
33

```

d) Relative Importance Table

Variable Importance			
Variable Name	Label	Number of Splitting Rules	Importance
Income	Income	1	1.0000
Age	Age	1	0.7228
X5	X5	0	0.0000
X2	X2	0	0.0000
Residence	Residence	0	0.0000
X4	X4	0	0.0000

e) Lift & ROC charts

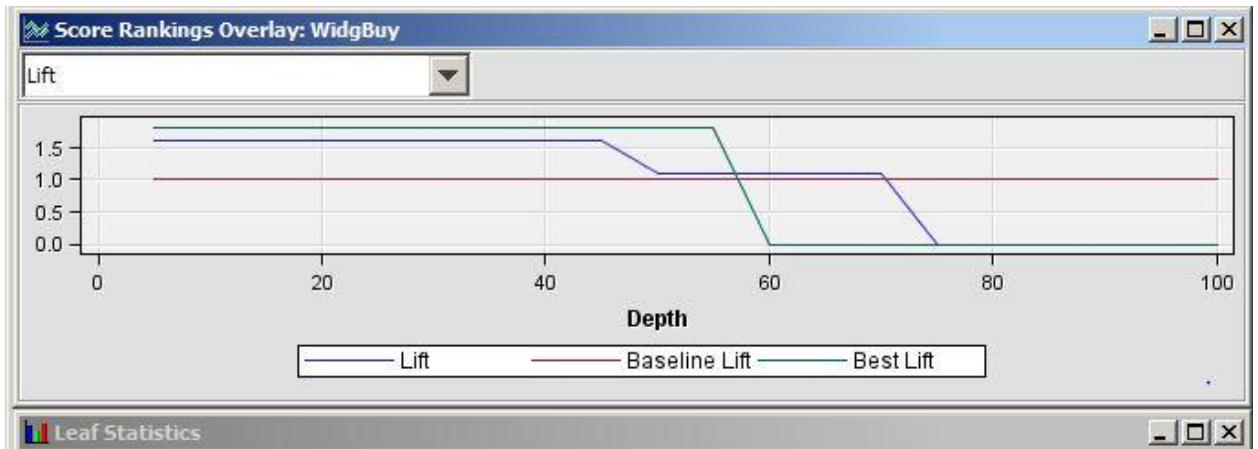


Figure 6 Lift Decision Tree

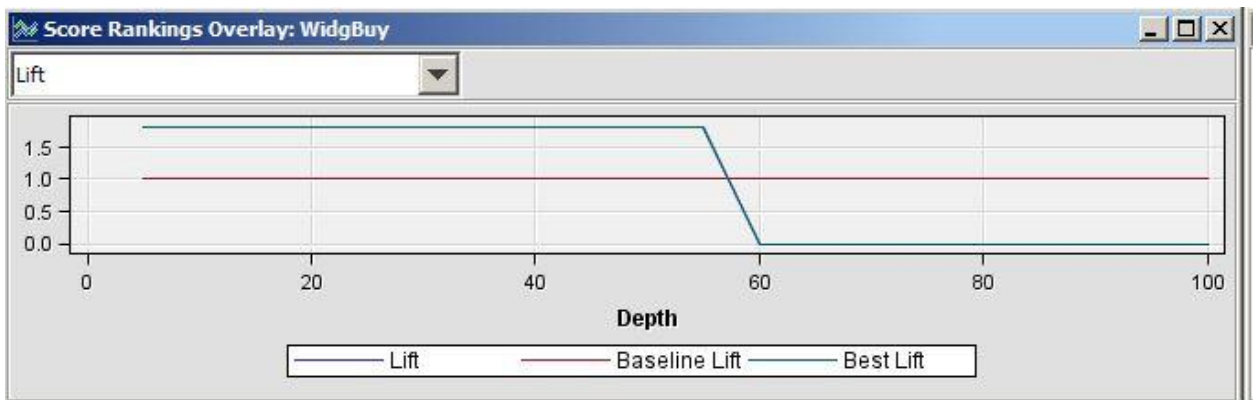


Figure 7 Lift Neural Network

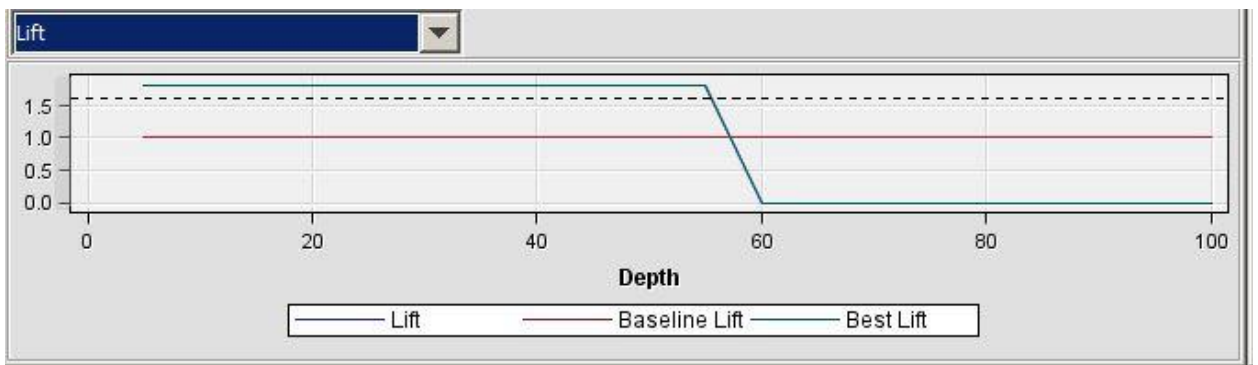


Figure 8 Lift Regression

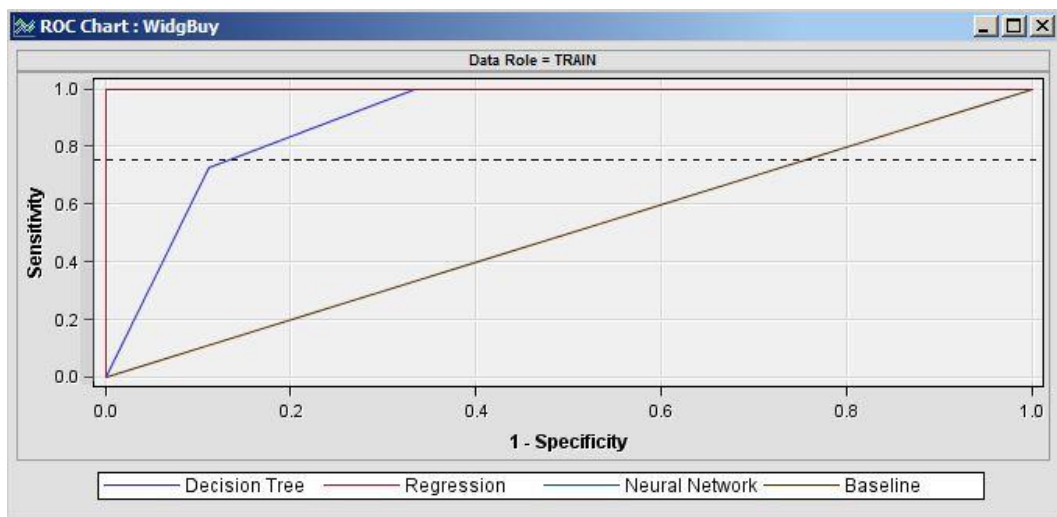
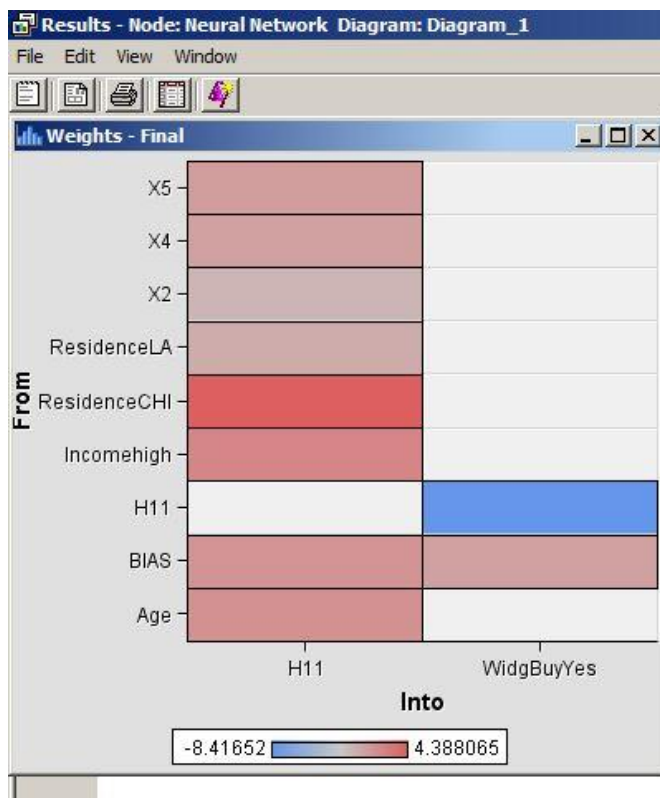


Figure 9 ROC

f) Neural Network Final Weights



g) Effects for the Regression Model

