

Megan Balcom

CIS 445

11/4/2016

Project 2

Confusion Matrix: Regression

Classification Table

Data Role=TRAIN Target Variable=WidgBuy Target Label=WidgBuy

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	100	100	9	45
YES	YES	100	100	11	55

The classification accuracy of identifying Widget Buyers versus Non-Widget Buyers was 100%, which is excellent. However, the sample size is too small for this model to be used in a real world setting.

Confusion Matrix: Neural Network

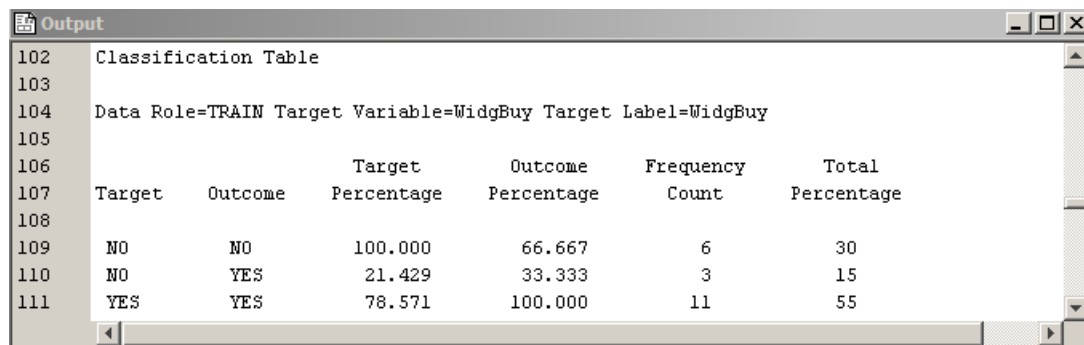
Classification Table

Data Role=TRAIN Target Variable=WidgBuy Target Label=WidgBuy

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	100	100	9	45
YES	YES	100	100	11	55

The classification accuracy of identifying Widget Buyers versus Non-Widget Buyers was 100%, which is excellent. However, the sample size is too small for this model to be used in a real world setting.

Confusion Matrix: Decision Tree



102	Classification Table					
103						
104	Data Role=TRAIN Target Variable=WidgBuy Target Label=WidgBuy					
105						
106			Target	Outcome	Frequency	Total
107	Target	Outcome	Percentage	Percentage	Count	Percentage
108						
109	NO	NO	100.000	66.667	6	30
110	NO	YES	21.429	33.333	3	15
111	YES	YES	78.571	100.000	11	55

The classification accuracy of identifying Widget Buyers versus Non-Widget Buyers was not as good as the Neural Network and Regression models. Also, the sample size is too small for this model to be used in a real world setting.

Roc and Lift Charts

ROC charts depict the global performance of the models for cutoffs within the range [0,1]. A ROC curve can be interpreted by examining a variety of characteristics about the chart. For example, the closer the curve to the left-side and the top of the ROC space, this means the test is more accurate. Conversely, the farther away the curve is from the left and top, the less accurate the test is. In the case of this ROC chart, it seems to suggest that the Decision Tree Model is not very accurate. However, the Neural Network and Regression models performed very well and demonstrated a high level of accuracy.

Lift is a measure of the effectiveness of a predictive model. This is calculated as the ratio between results gathered with and without a given predictive model. The Regression and Neural Network models did produce a lift of 0.2 more than the Decision Tree Model. While this cumulative lift chart does not show the expected number of positive results, it is visually clear that the Regression and Neural Network models produced a greater number of positive results compared to the Decision Tree.

Rules Generated By Decision Tree

The rules generated by the decision tree are explicit and find understandable relationship among independent and dependent variables. In this case, the first split in the tree is based off of Income – either it is high or missing, or it is low. After that, on the high income or missing income branch, there is another rule that looks at the age of the buyer.

- (1) If Income = Low, then the person is a widget buyer
- (2) If Income = High or Missing, and Age <30.5, then the person is a widget buyer
- (3) If Income = High or Missing, and Age >= 30.5, then the person is not a widget buyer

Importance of Variables

Even though this is a small data set, it is important to have an ability to find a way to clearly identify which variables have the most predictive power. For this project, SAS EM is able to generate the relative importance of independent variables. As illustrated in the previous answer, SAS EM determined that Income and Age were the most important variables out of the seven initial variables. In the Decision Tree model, SAS determined that the most entropy that could be recognized in that model occurred only if the branching occurred in the order of branching off of Income first, then by Age.

Effects Generated for Logistic Regression Coefficients

The Income and Age variables have the most predictive power – 1.0 and 0.7 respectively. Income is the most predictive followed by Age.

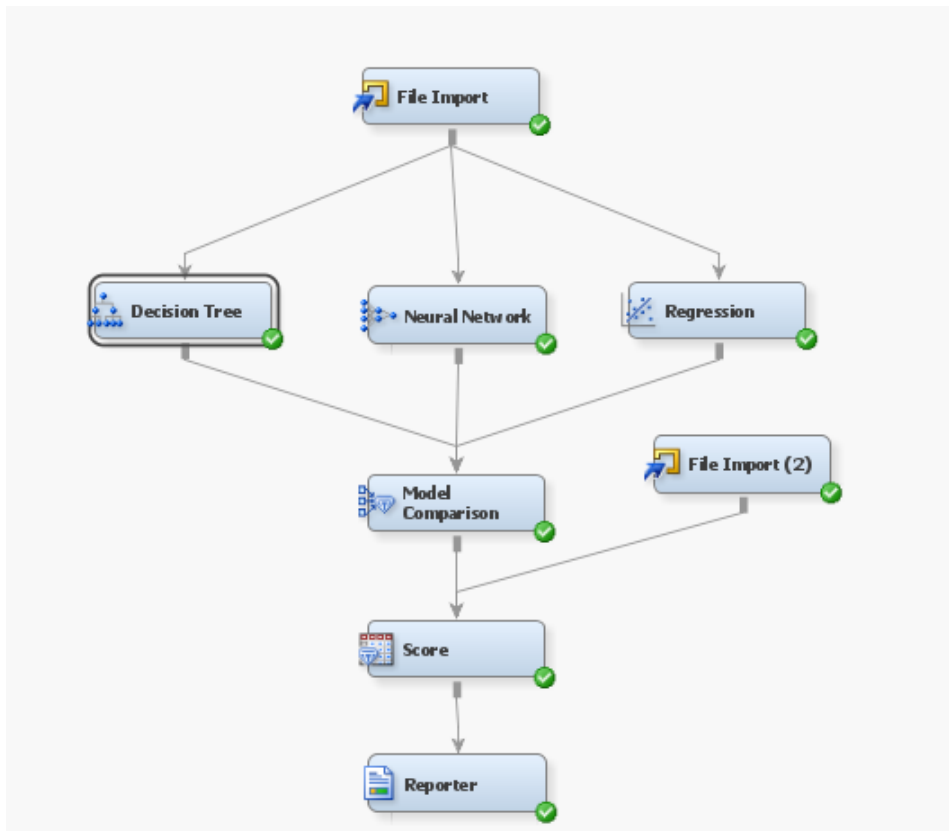
The logistic regression model identified the same variables as the decision tree in terms of their predictive power.

Weights of the Neural Network

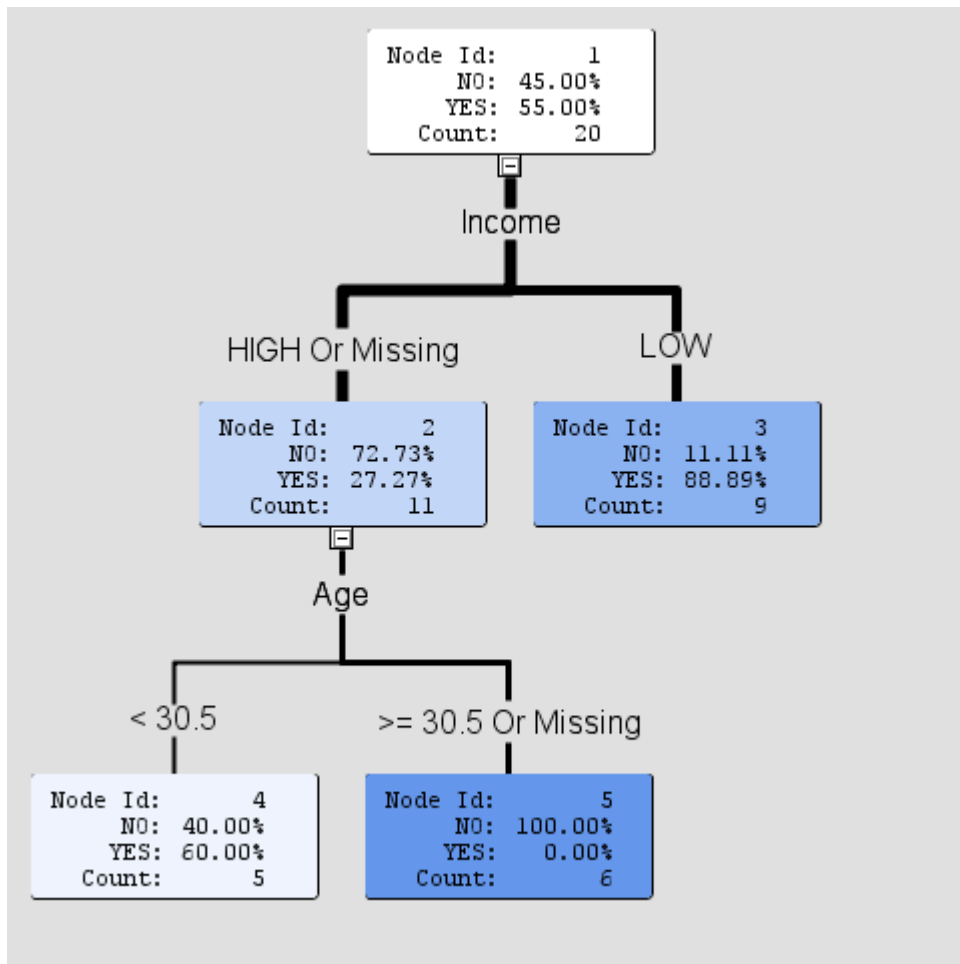
The weight table is a little hard to make sense of, however there are some pieces of discernable information after examining the table of final weights. There are two variables that seem to stick out as possessing both the lowest value weight and the highest value weight: H13 and Chicago residence. There is not an extremely clear picture that emerges from the table of weights, though.

Classification of Widget Buyers vs. Non-Widget Buyers

14 people were non-widget buyers and six were non-widget buyers.



1.a



1.b

```

*-----*
Node = 3
*-----*
if Income IS ONE OF: LOW
then
Tree Node Identifier   = 3
Number of Observations = 9
Predicted: WidgBuy=Yes = 0.89
Predicted: WidgBuy=No  = 0.11

*-----*
Node = 4
*-----*
if Income IS ONE OF: HIGH or MISSING
AND Age < 30.5
then
Tree Node Identifier   = 4
Number of Observations = 5
Predicted: WidgBuy=Yes = 0.60
Predicted: WidgBuy=No  = 0.40

*-----*
Node = 5
*-----*
if Income IS ONE OF: HIGH or MISSING
AND Age >= 30.5 or MISSING
then
Tree Node Identifier   = 5
Number of Observations = 6
Predicted: WidgBuy=Yes = 0.00
Predicted: WidgBuy=No  = 1.00

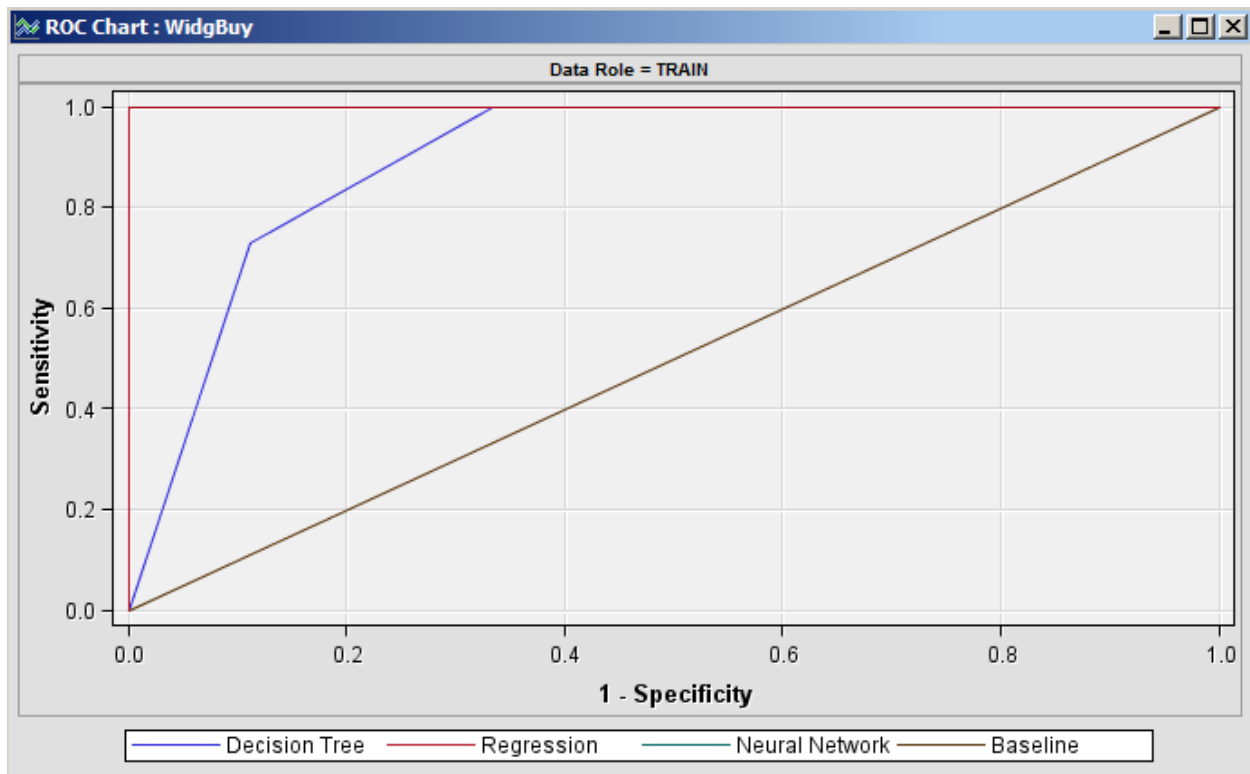
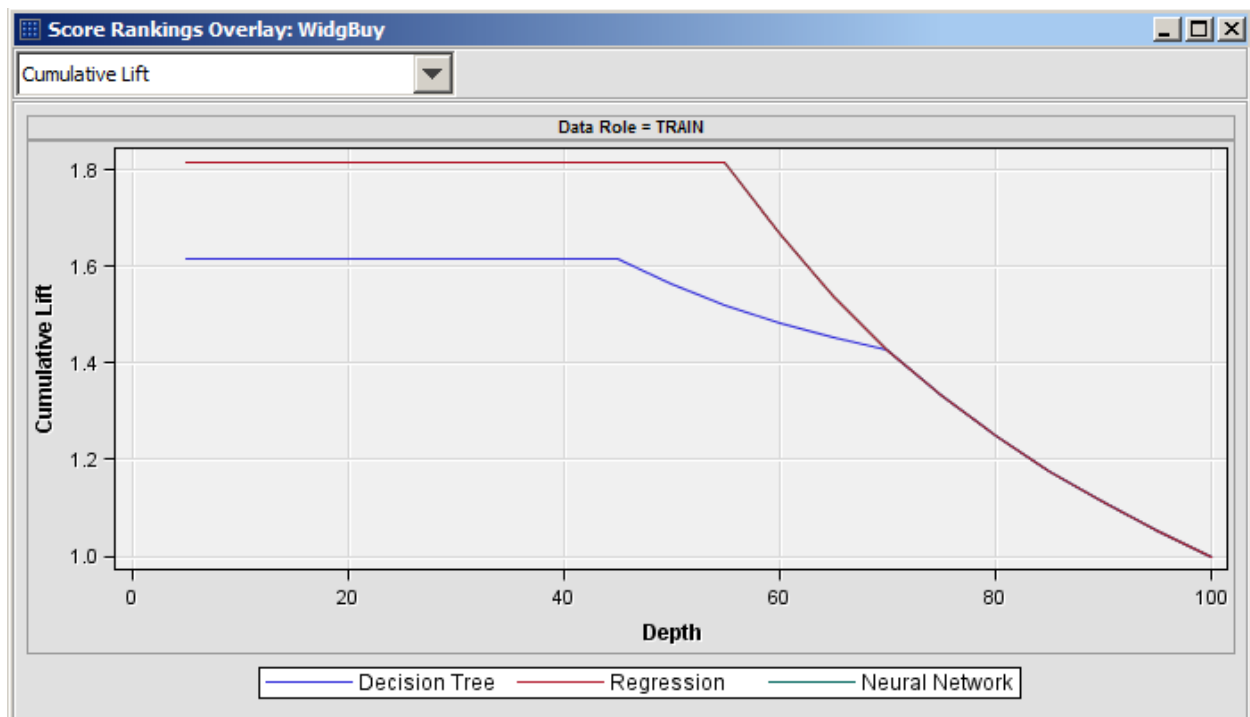
```

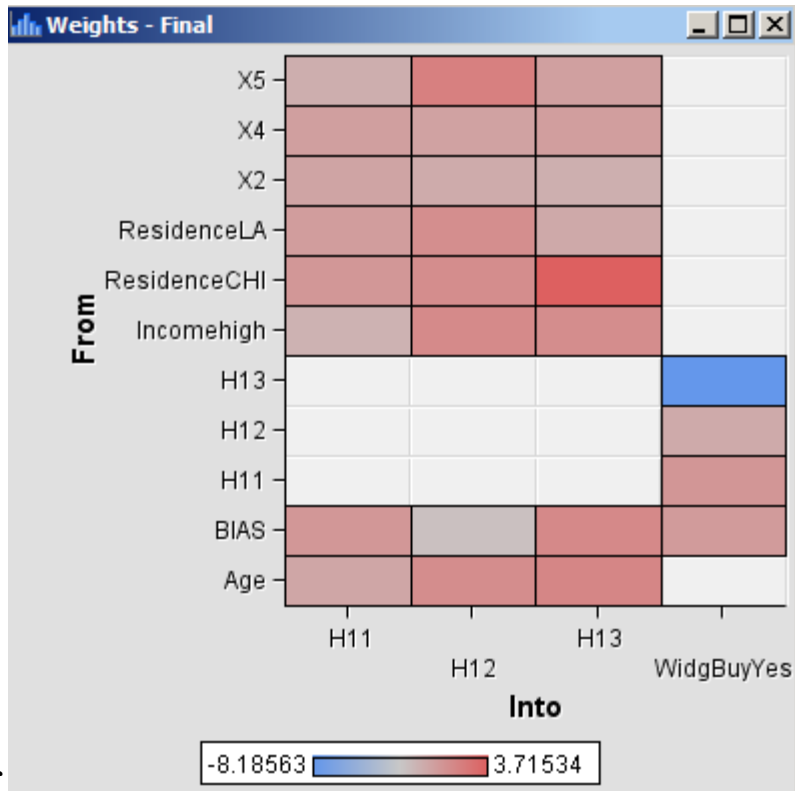
1.c

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

1.d

1.e





1.f

1.g

