Megan Balcom

CIS 445

11/15/16

Project 3

Work Flow

The premise of this project is to examine data from a banking institution to determine which model best fits the historical data of people who default on their loans. This workflow begins with inputting data from the data source SAMPSIO.HMEQ, which contains 13 variables and nearly 6000 observations. Using this data set, this workflow builds six neural network (NN) models to compare their classification accuracy. Based on the results, the best model will be selected and can be used to predict whether a new loan applicant to the bank will default on a loan in the future or not.

After the data source is imported, the Stat Explore node provides data and initial breakdowns about that information that helps visualize the data before it will be manipulated by various other logic and rules. Following Stat Explore is Data Partition node which in this project is divided up by 50%, 50%, and 0% for the training, validation, and test sets, respectively. Impute is used next, which takes into account missing values and eliminated them from the analysis. Neural networks are special in this regard because they do not handle missing values well, compared to other modeling techniques.

The flow breaks off into two main branches: a branch with transformed variables, and a branch without transformed variables. Each of these branches will then feed to three different neural networks – those three neural networks on each branch have three values of the number of networks they have; for each branch there are network values of one, three, and five. From there, the results are fed to a model comparison node, where results of all models are displayed.

Confusion Matrix

NN5S performed best compared to the other NN models when examining the confusion matrix results. It classified the most number of true positives and the most number of true negatives and the fewest false positives and false negatives.

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 NN5S (Neural Network with five neurons in the Variable Selection branch), l is the most accurate of the six, but it is still not a great performing model. However, NN1T (Neural Network with one neuron in the Transformed Variables branch), is the least accurate.

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. Similarly to the ROC chart, NN5S performed slightly better than the other neural networks, while NN1T performed the worst. While this cumulative lift chart does not show the expected number of positive results, it is visually clear that the NN with more neurons that had not been transformed initially performed much better than the NN with fewer neurons and had been transformed.

NN Model Selection

Based on the analysis so far, the best NN model to use for future customers to determine loan granting would be NN5S – the neural network with five neurons in its hidden layer and that did not have its variables transformed. It is still far from being a perfect model, but out of the six models, it performed the best.

Transformed vs Non-Transformed

Based on my results, there does not seem to be a definitive benefit to transforming variables. Only two of the NN models performed much better or much worse than the rest of the models, and they were very different from each other. The rest of the transformed and non-transformed variable results seem to still return roughly the same values and it would be difficult to differentiate them from each other by a quick glance at the ROC or lift charts.

Variable Selection

Similarly to variable transformation, in my results a clear benefit of using variable selection did not emerge. If you look at the ROC and lift charts, aside from the best performing model and the worst performing one, it would be difficult to determine which models used variable selection and which did not.

NN Weight Analysis

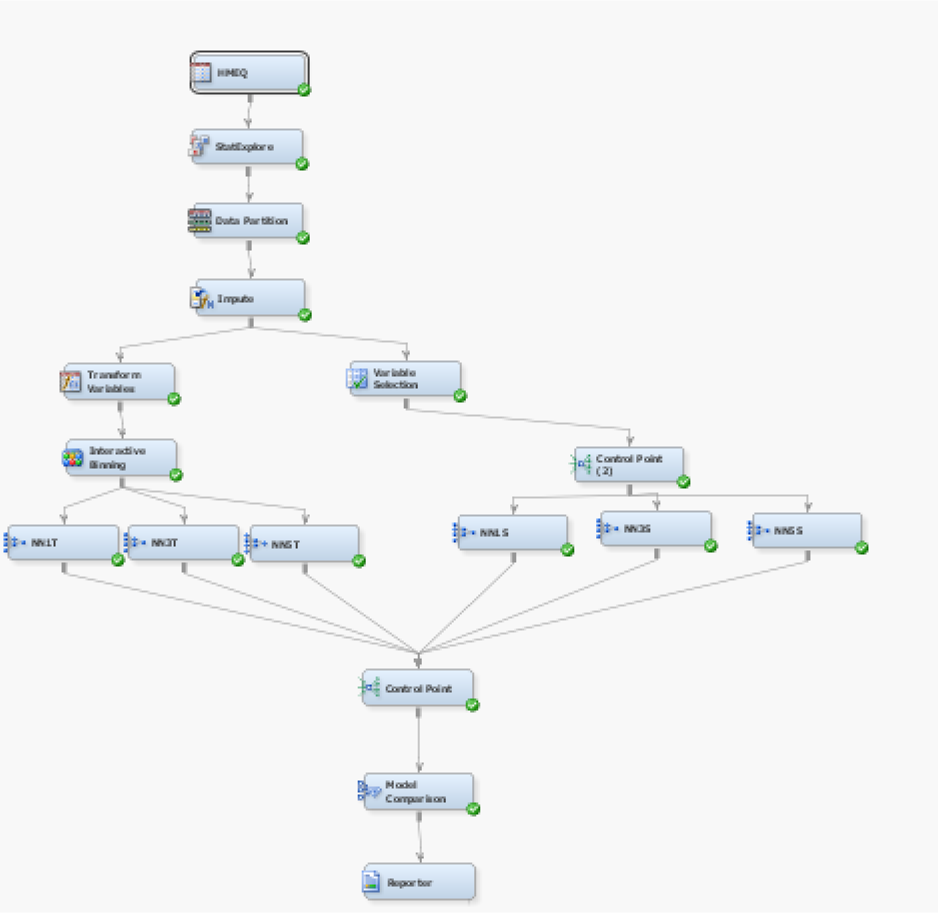
It is difficult to understand what the weights mean and how they contribute to the overall results of the models. For a few models, there are certain criteria that happen to have extremely positive or negative effects, however it is not understandable how that value is utilized and influences the outcome of the model.

Best NN Analysis

My best performing NN has a fairly low false negative classification rate, but a slightly higher false positive classification rate compared to the other neural network models.

Best NN Classification

My best performing NN classifies more false positives (129/321) than false negatives (274/2257).



3.A

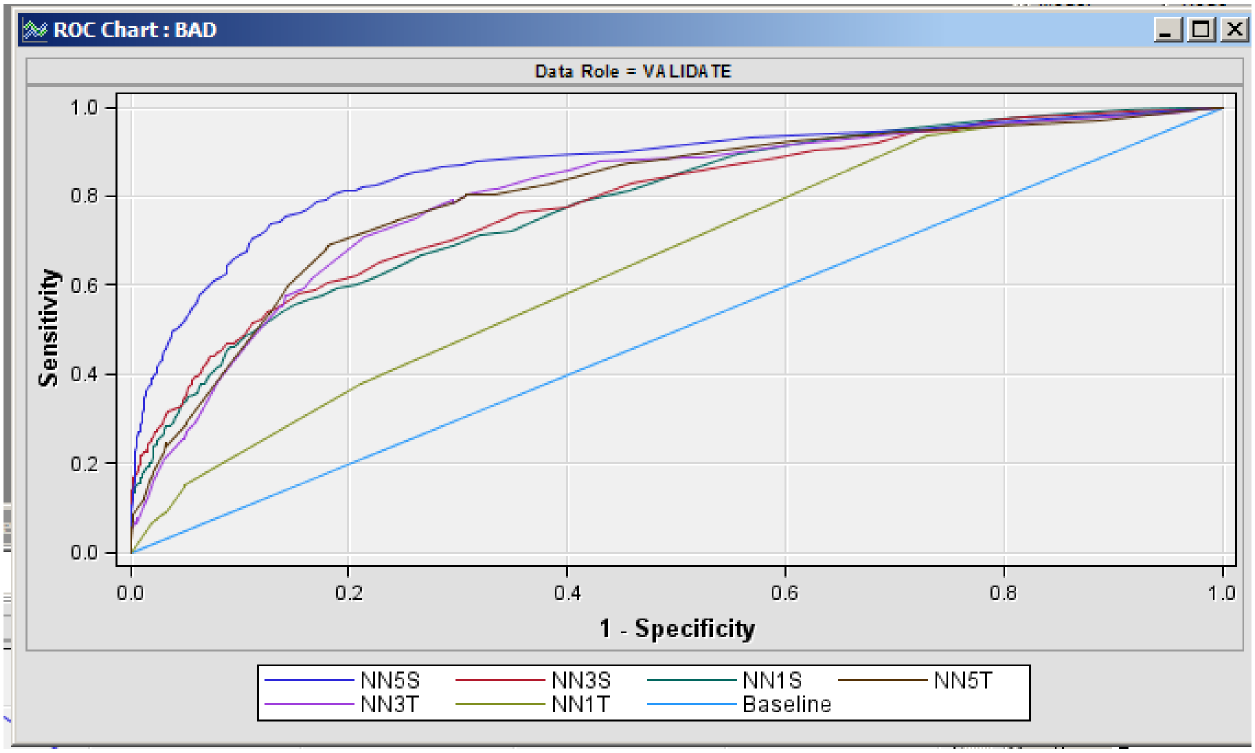
3.B

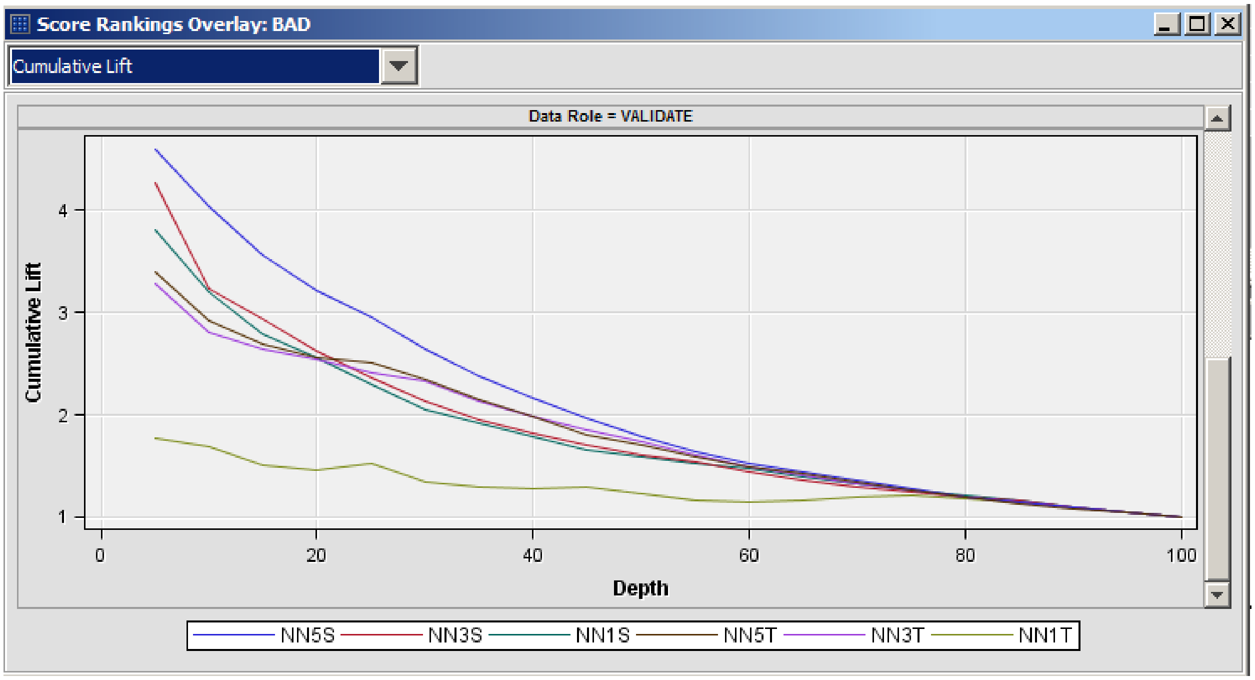
```
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

 Model      Model       Data                Target   False     True      False     True
 Node    Description    Role       Target   Label   Negative  Negative  Positive  Positive

Neural      NN1T       TRAIN        BAD               561      2348        37        33
Neural      NN1T       VALIDATE     BAD               556      2339        47        39
Neural2     NN3T       TRAIN        BAD               417      2286        99       177
Neural2     NN3T       VALIDATE     BAD               442      2266       120       153
Neural3     NN5T       TRAIN        BAD               410      2284       101       184
Neural3     NN5T       VALIDATE     BAD               423      2265       121       172
Neural4     NN1S       TRAIN        BAD               386      2322        63       208
Neural4     NN1S       VALIDATE     BAD               406      2278       108       189
Neural5     NN3S       TRAIN        BAD               353      2310        75       241
Neural5     NN3S       VALIDATE     BAD               385      2262       124       210
Neural6     NN5S       TRAIN        BAD               276      2314        71       318
Neural6     NN5S       VALIDATE     BAD               274      2257       129       321
```

3.C


ROC Chart : BAD
Data Role = VALIDATE
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
NN5S
NN3S
NN1S
NN5T
NN3T
NN1T
Baseline


Score Rankings Overlay: BAD
Cumulative Lift
Data Role = VALIDATE
Cumulative Lift
4
3
2
1
0
20
40
60
80
100
Depth
NN5S
NN3S
NN1S
NN5T
NN3T
NN1T