Nga Nguyen
CIS 445
Project 3

A. Brief description of nodes in the diagram and each path of the workflow

The nodes in the diagram: Input data node is done first in order to retrieve the data that we want to test and manipulate. This gets the data from HMEQ and allows us to use it. The StatExplore node is next this gives us the ability to examine the statistical properties of our data set and lets us calculate correlation statistics for our input and target variables. After the StatExplore node the Data Partition node gives us the ability to divide our data set into a training set, a validation set and a test set. For this tutorial half of the data was put into a training set, while the other half was put into our validation set. After the Data Partition node there are 3 separate splits to either the Transform Variable node, the Impute node or the Variable Selection node.

The Transform Variable node was used so that we can improve the fit of a model to our data and it allows us to create interaction variables such as INDELINQ and INDEROG. The Interactive Binning variable node follow the transform variable node. This node allowed us to split and combine bins to get a better distribution of the data. After this, I put in another variable selection node to define the variables I wanted to use and then I put in my neural network nodes for 3 neurons in one hidden layer and 5 neurons on another hidden layer, which is used to compare all the paths in my workflow. The Impute node allows us to eliminate any missing values by rejecting any observation that may have an incorrect or missing value. I did a model comparison node to compare each of the models and predictions.

B. Discussion of confusion matrices

The confusion matrices for all 6 neural networks is demonstrated below. When looking at the results, the best neural networks that produced the most True Positives and the most True Negatives were Neural Network 5B and Neural Network 3B. In the training set there were 2979 cases, of that for Neural Network 5B, there were 340 True Positives and 2261 True Negatives for the training set. There were 2981 cases in the validation set, of that number there were 338 True Positives and 2233 True Negatives. The worse performing neural networks were Neural Network 5C and Neural Network 3C. For Neural Network 5C there were 203 True Positives and 2326 True Negatives in the training set. While there were 203 True Positives and 2307 True Negatives in the validation set.

C. Discussion of ROC chart and Lift chart

The ROC charts and Lift charts demonstrate how good a specific model is in terms of its classification accuracy. The ROC chart shows us the global performance of the models for cutoffs of 0.5. The best model will have a curve in the ROCC chart that is closest to the upper left-hand corner of the graph. In the validation set, Neural Network 3B and Neural Network 5B were deemed the best with nearly the same results. The Lift chart measures ow effective the

predictive model is at calculating the results we get from with and with the predictive model that was created.

D. Best Neural Network model

According to the output of the fit statistics the models that were transformed and binned performed the best. The best Neural Network models to choose to classify future customers would be Neural Network 3B which has 3 neurons in its hidden layer and Neural Network 5B which has 5 neurons in its hidden layer.

E. Is it better to transform?

Yes it is better to transform your variables because this will enable you to use the Interactive Binning node which allows you to better classify your results on specific variables. This enables you to make more accurate and predictive models of the data.

F. Does variable selection help with classification accuracy rates?

Variable selection does help with the classification accuracy rates depending on specific circumstances of how you manipulate your data. If you transform your data and do interactive binning before you do variable selection it can improve your accuracy rates. If you do variable selection without doing these two things first it will produce worse results.

G. Is it easy to interpret the weights of the Neural Network model?

No, it is not very easy to interpret the weights of the neural network model the results are put into a sort of "black box" which are very difficult for humans to analyze and understand. The weights of a neural network are where the model's knowledge is encoded, and the explicit equation used by it is unknown.

H. Analysis of best neural network model

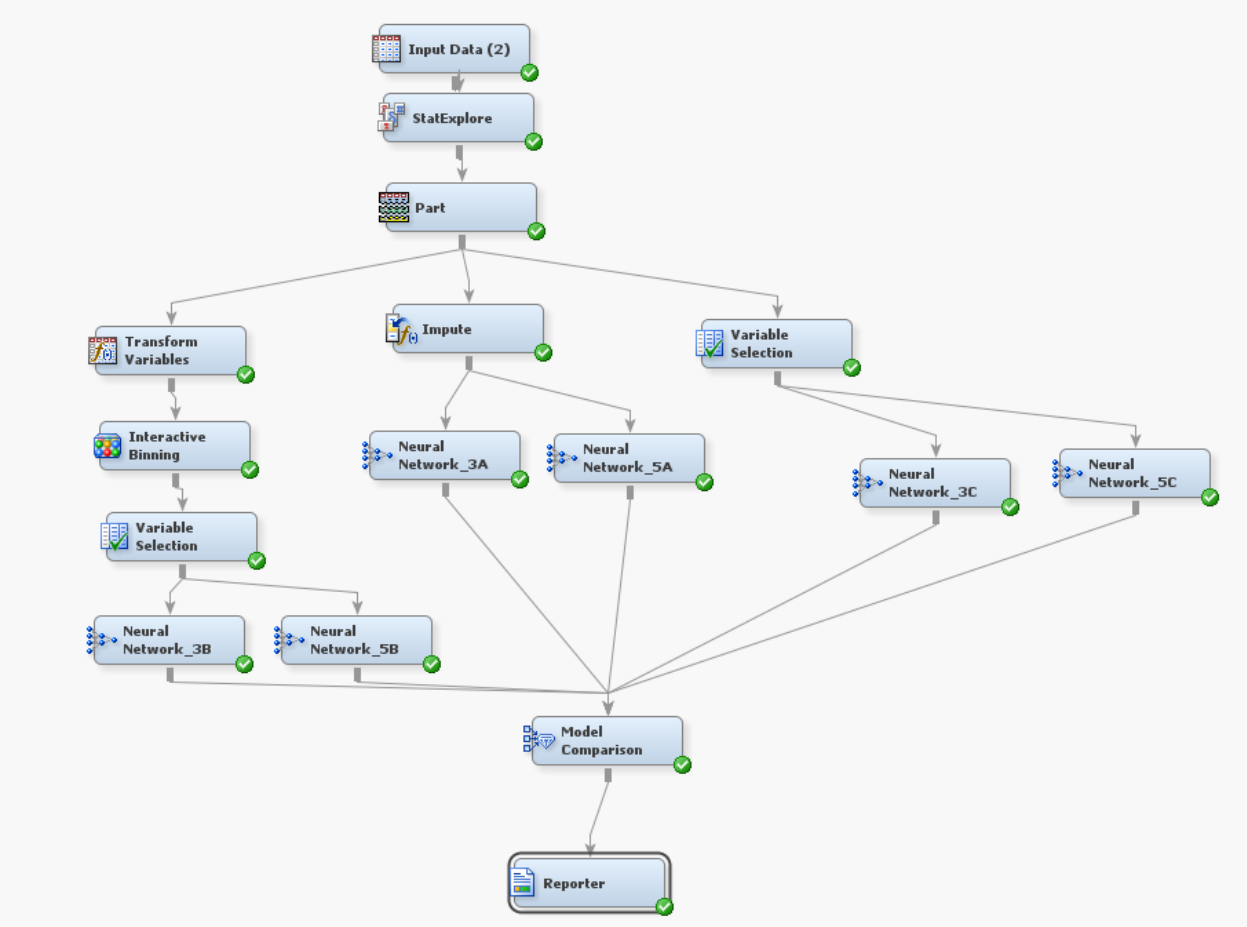
Neural Network_5B was the best neural network model. The overall classification accuracy for the validation set was (338+2233)/(595+2386)=0.862=86.2%. The classification accuracy of "good" transactions is 338/595==0.560=56.8%. The classification accuracy of "bad" transactions is 2233/2386=0.936=93.6%.

I. Does the best Neural Network model classify more good or bad loans?

The best Neural Network model classifies more bad loans than good loans according to the confusion matrices. The classification accuracy of bad loans is 93.6% according to the results of the validation set. It classifies good transactions correctly on 56.8% of the time.

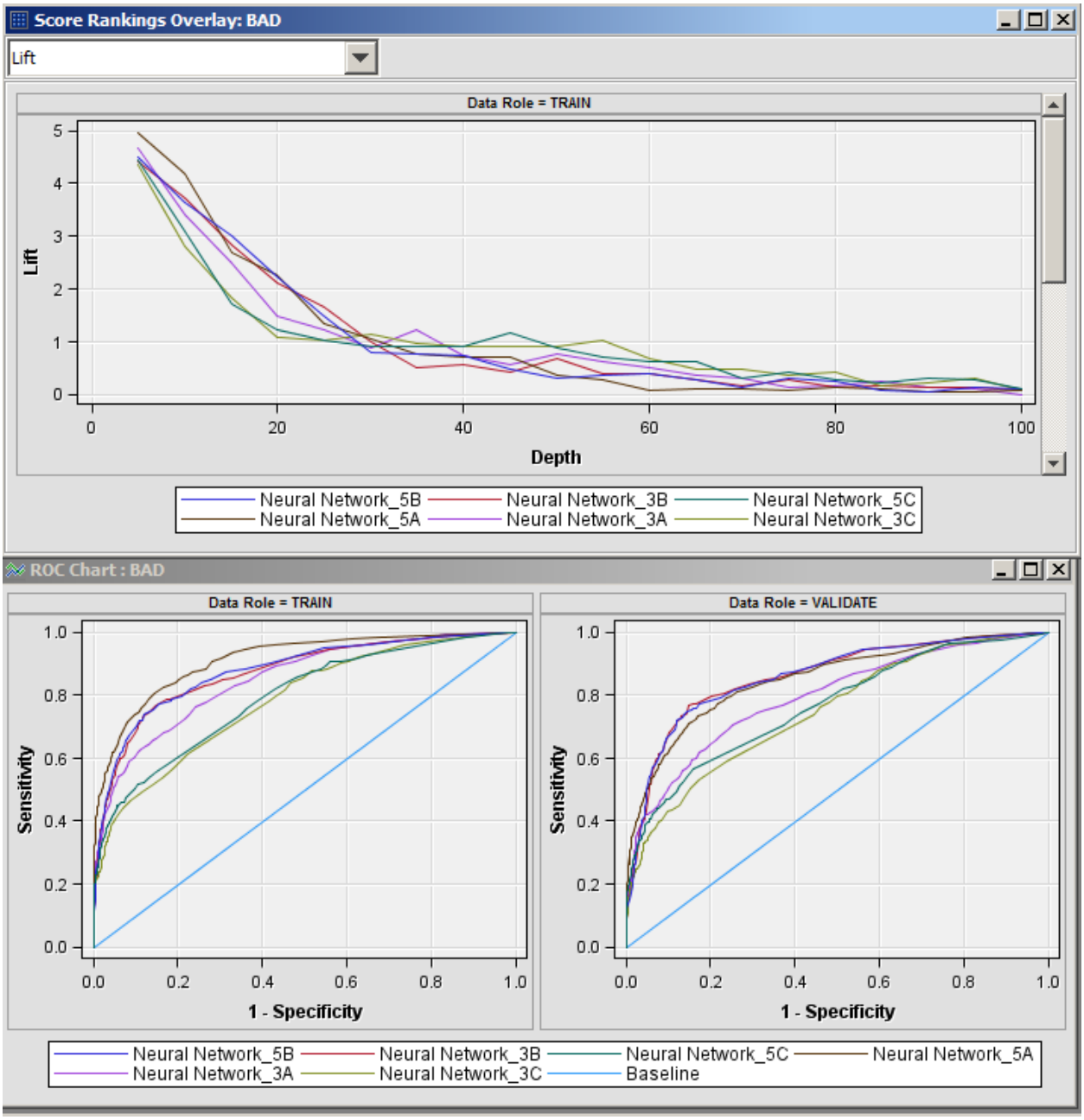
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Neural6	Neural6	Neural Network_5B	BAD		0.137538
	Neural5	Neural5	Neural Network_3B	BAD		0.140557
	Neural3	Neural3	Neural Network_5A	BAD		0.140892
	Neural2	Neural2	Neural Network_3A	BAD		0.150956
	Neural4	Neural4	Neural Network_5C	BAD		0.158001
	Neural	Neural	Neural Network_3C	BAD		0.17209



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

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Neural6	Neural Network_5B	TRAIN	BAD		254	2261	124	340
Neural6	Neural Network_5B	VALIDATE	BAD		257	2233	153	338
Neural5	Neural Network_3B	TRAIN	BAD		264	2257	128	330
Neural5	Neural Network_3B	VALIDATE	BAD		257	2224	162	338
Neural2	Neural Network_3A	TRAIN	BAD		358	2329	56	236
Neural2	Neural Network_3A	VALIDATE	BAD		374	2310	76	221
Neural3	Neural Network_5A	TRAIN	BAD		256	2303	82	338
Neural3	Neural Network_5A	VALIDATE	BAD		295	2261	125	300
Neural	Neural Network_3C	TRAIN	BAD		428	2329	56	166
Neural	Neural Network_3C	VALIDATE	BAD		435	2308	78	160
Neural4	Neural Network_5C	TRAIN	BAD		391	2326	59	203
Neural4	Neural Network_5C	VALIDATE	BAD		392	2307	79	203


Score Rankings Overlay: BAD
Lift
Data Role = TRAIN
Lift
Depth
0
1
2
3
4
5
0
20
40
60
80
100
Neural Network_5B
Neural Network_3B
Neural Network_5C
Neural Network_5A
Neural Network_3A
Neural Network_3C
ROC Chart : BAD
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Neural Network_5B
Neural Network_3B
Neural Network_5C
Neural Network_5A
Neural Network_3A
Neural Network_3C
Baseline