Project Four

Ryan English

The diagram I built for determining the best way to predict the value of homes in poorer areas, has two distinct paths: One with transforming the variables and filtering and one without. Other than those nodes the paths match with three different models: Linear Regression, Neural Network, and MBR. I then compare the models and determine the best one.

Starting from the beginning I input the data. The given excel spreadsheet contains 363 records of house sale transaction in poorer areas and the features of the home. By mining the features, we hope to gain knowledge on how to determine price based on these parameters.

After I input the data, I just go ahead and look at the Statistical Data, with the Stats Explorer. I did this in an effort to get a better understanding of each of the parameters we are given for the homes. I wanted to see if there where there are abnormalities in the statistics. As we can see in the graph Age has some negative values.

After having a look at the Stats Explorer, I added in the Data Partition node. This node takes the inputted data and separates it into 70% training data and 30% validation data. This way we can train to better fit the models. We don't separate any score data, because we don't have sufficient amount of data.

Next, I filter out all the outliers. Based on the stats earlier, the only outliers that I could determine (also based on the instructions) was the Age that was negative. I added a custom filter to get rid of these variables, since it seemed like bad data.

Here is where we split off into two paths. The first I am going to discuss is going directly into the models. Here we take the Linear Regression, Neural Network and MBR and compare the results. My Model Comparison gave the Neural Network the win. All of the models, however, had massive mean square errors. In a real-world scenario, I don't think this Neural Network would be much help.

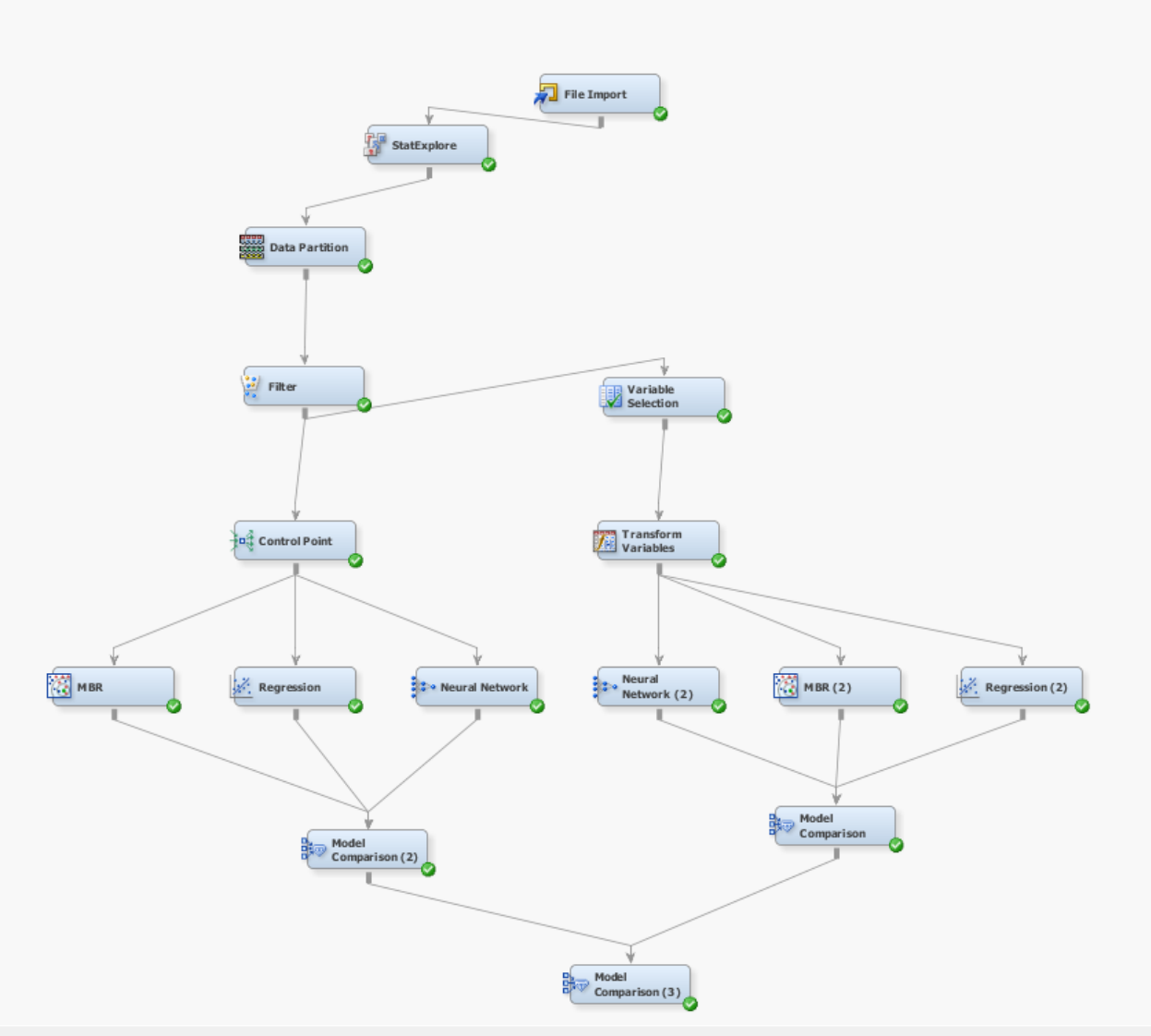
The other path on the other hand, before getting to the models selected the most meaningful variables by using the Variable Select node. The selected variables based on R^2 where Total Area and Age. The variables removed where Neighborhood, Baths, and Second Floor. After we selected some

variables we moved to transform the variables. I left the Transform Variable node at default to let it do what it thought best. I looked into the different variables and they all had 4 bins; and I didn't see any reason to make a custom rule.

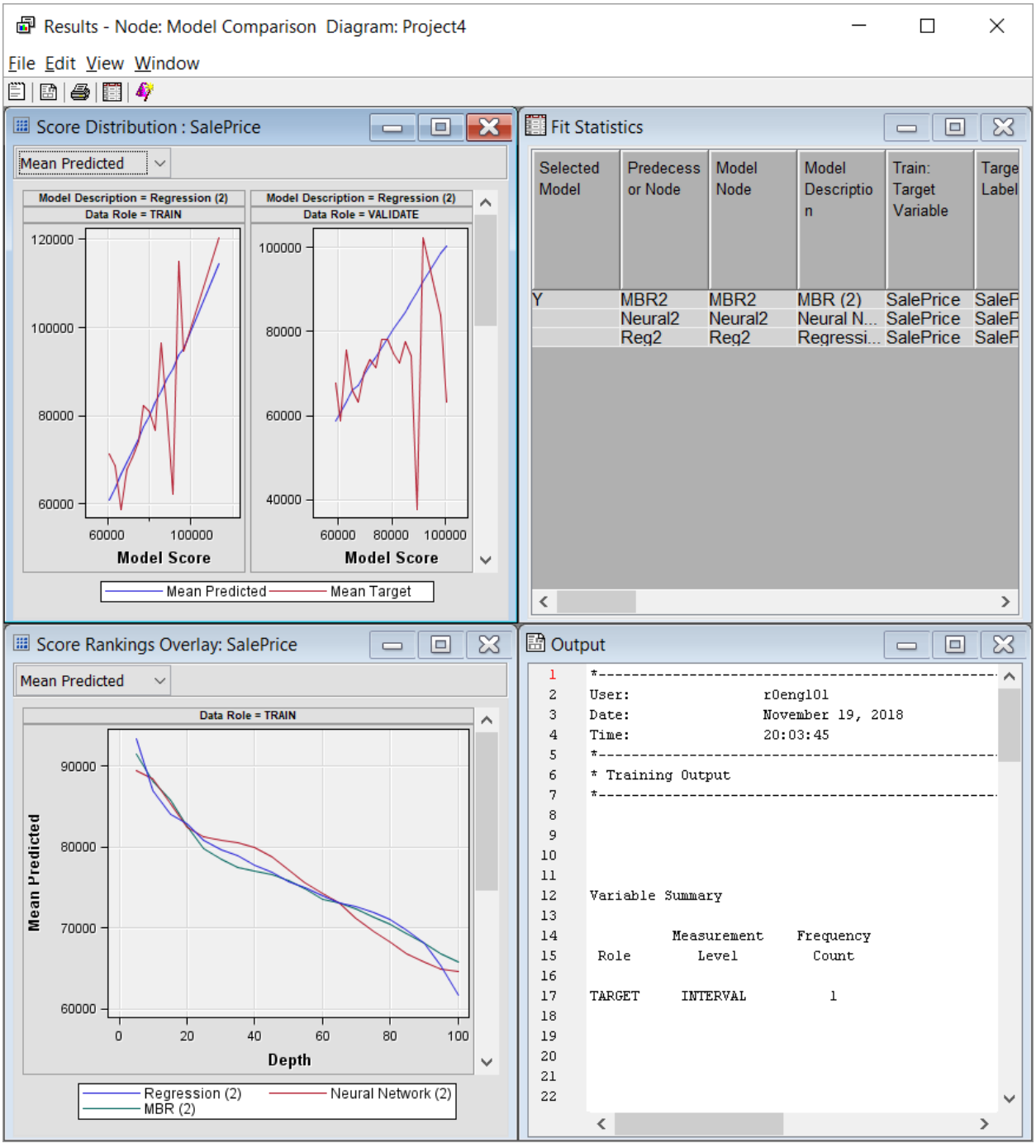
Just like before, I compare these models and determine that the MBR is the best model for this path. Unfortunately, the Average Square Error is still massive (but I assume this is in $ Thousands). Interestingly enough, the only value that was really different between the three models was the Average Square Error; and the MBR had almost have the amount as the other three.

The last node I use is to compare the best of each path with one more model comparison node. The MBR won, but not by a huge margin. It would seem that removing all the variables, selecting the best, and transforming the variables did not have a massive outcome with the ~400 records; but I can image with about ~4000 records the scale might be a bit bigger.

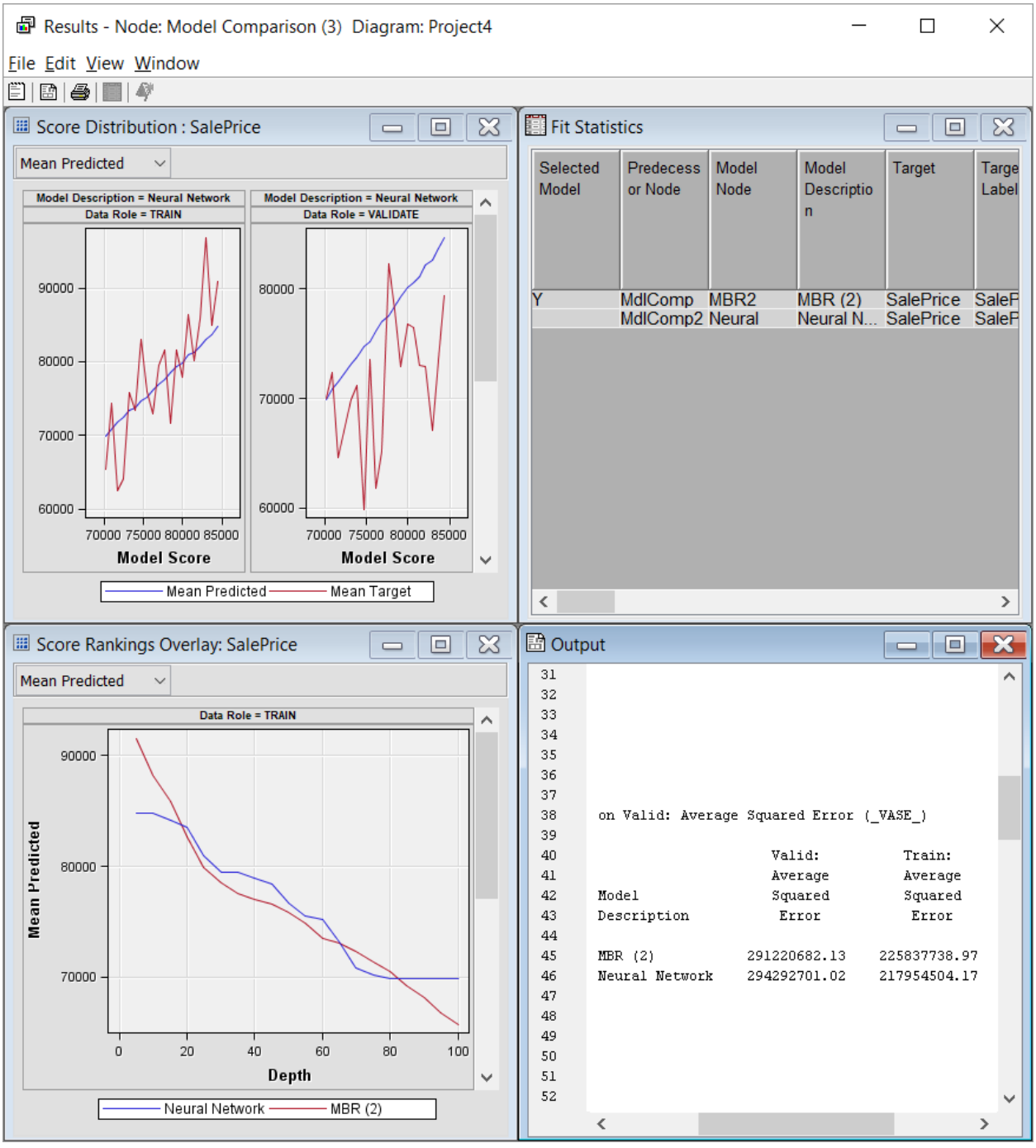
This assignment was much harder than the previous; made me really understand how massive and difficult it is to use SAS enterprise miner. I struggled for a long time with determining the order of which the nodes need to be in and what properties to change. I think, by the end of it I finally got a moderately good diagram. For some reason, I never was able to get my ROC charts to show. I had it set to Yes but I never could see them.



Final Diagram



Model Comparison: Path with Transform and Variable Selection



Model Comparison: Default