Nga Nguyen

CIS 445

Project 4

11/21/17

In this project, we used several nodes in SAS in order to build and test several different types of models being used in different ways to analyze and interpret the data from the Real Estate files which contained information about the features of low-end residential properties in two different types of neighborhoods of a medium sized mid-west US city. Initially, once the file import node was used to import the data, the variable levels for several variables were changed according to Table 1 so that the data could be properly processed. The role of SalePrice was changed to that of the target variable because it was the output that we were trying to predict based on the various variables that were included in the dataset.

The Stat Explore node was used to examine the distribution of all the variables. It 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. The Variable Selection node was used to define the variables I wanted to use. The interactive binning node allowed us to split and combine bins to get a better distribution of the data. The transform variables node was used so that we can improve the fit of a model to our data and it allows us to create interaction variables.

The filter node allows us to identify and remove any outliers from our data sets. The data partition node gives us the ability to divide our data set into a training set, a validation set and a test set. There was no test set in this case. The data was divided into a training set and a validation set only. The neural network node was used to train specific neural network configurations. The regression node allows us to fit linear and logistic regression models to the

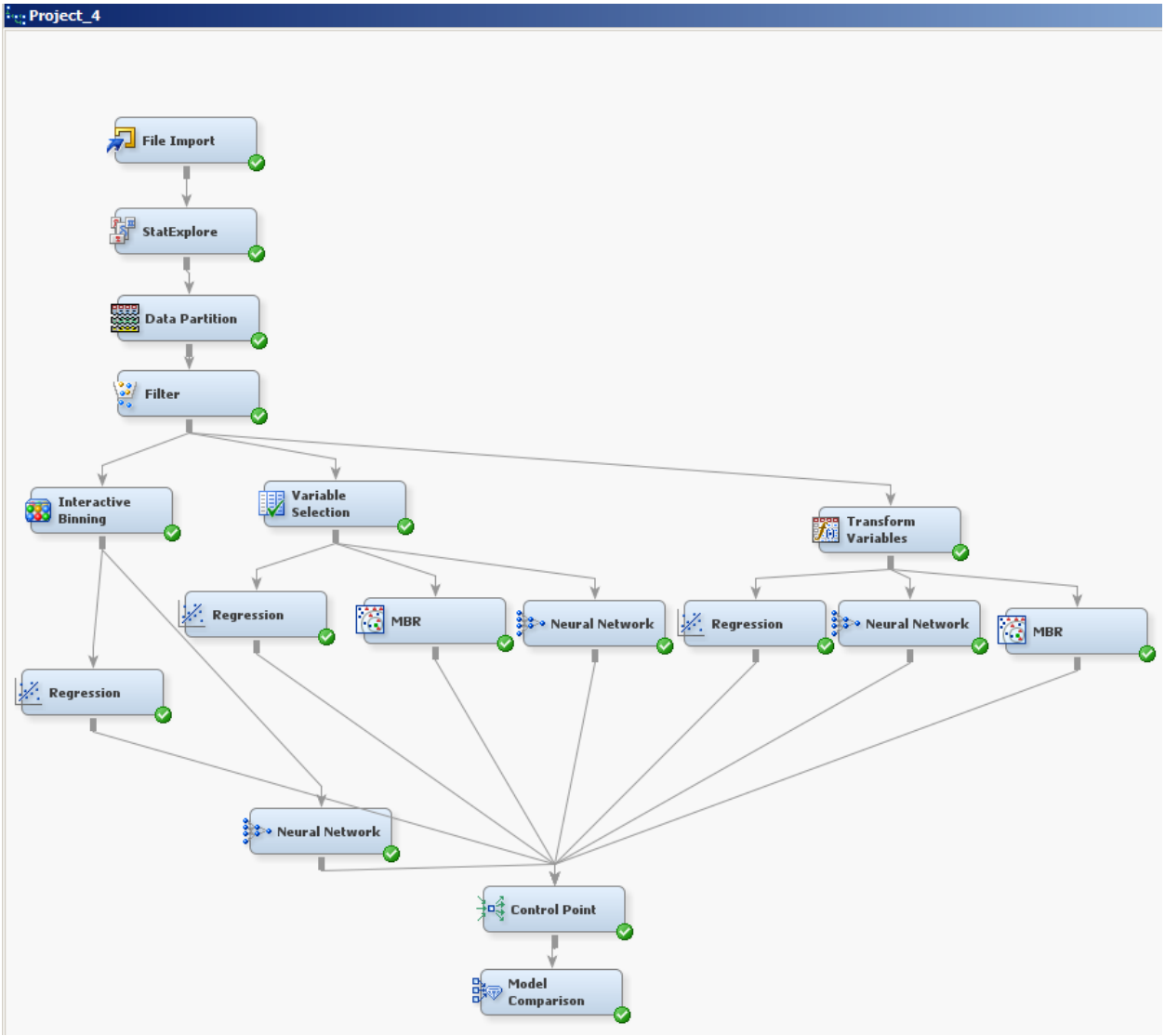
data that we have. The MBR node lets us use the k-nearest neighbor algorithm to categorize/predict any observations that we come across. The model comparison node to compare each of the models and predictions.

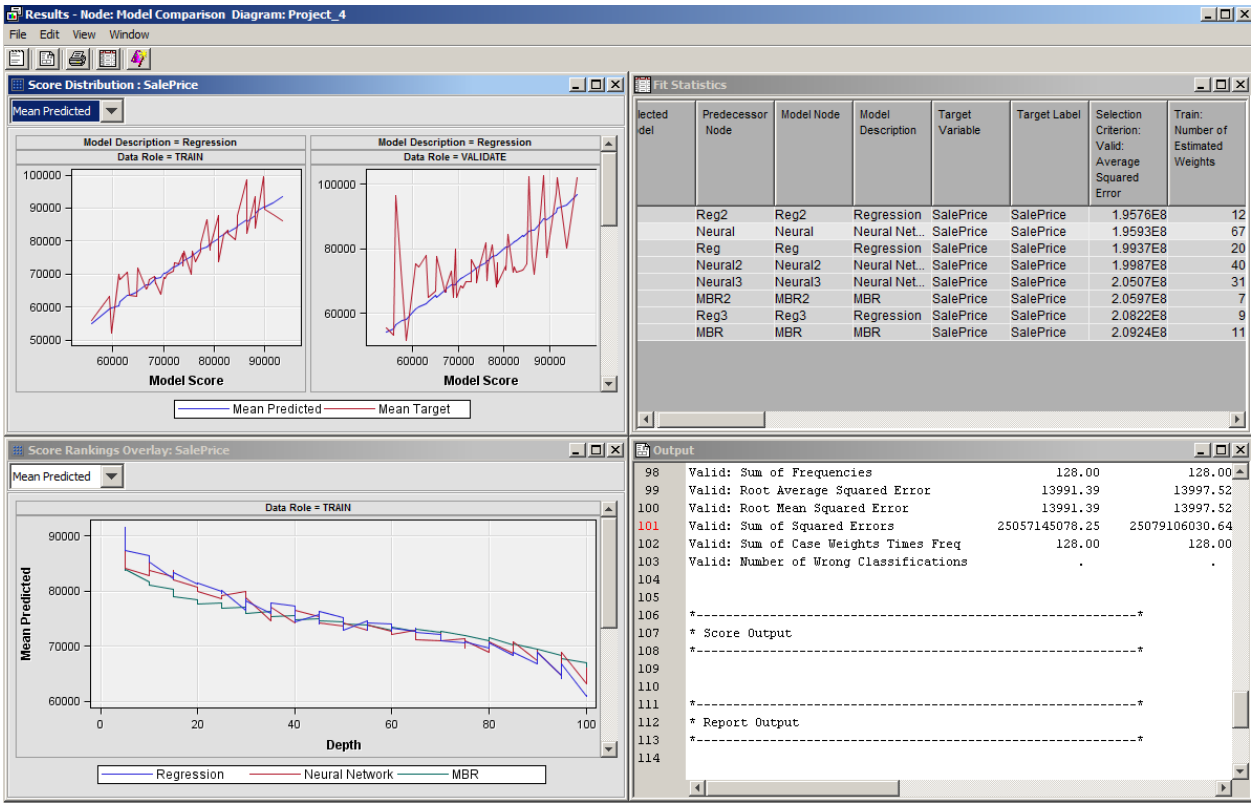
For this project Dillon imported the data and figured out in what order we should arrange all of our models. Dillon was also able to identify what outliers that we must keep and which ones could possibly be thrown out. Dillon was also able to use the log function to show our data in a more presentable format. I was able to categorize the data and set the target output and analyze the data from the results.

According to the output results, Regression 2 (Variable Selection) was able to perform the best out of all the models. The root mean squared error (RMSE) for regression 2 was the smallest measuring at $12,752.96. While MBR performed the worse with a mean squared error of $13731.30.

Looking at the mean predicted graphs seem to me to be very sporadic, which is caused by all the variables and their impacts on the target. The graph represents the error between the predicted and target/outcome values. The error for regression 2 is the best and has the best error on the mean predicted graph.

In the end using our data set of 321 records of housing data that was separated into 193 training entries and 128 validation entries. Since there is not a lot of data to be utilized by our models we can not have a very low RMSE, which shows the error in the base units. This is important because you can compare the data back into the real unit (dollars in this case). After looking at the RMSE we have concluded that regression 2 (the regression model that utilizes variable selection) is the best model for the data we have utilizing the nodes and tools we have learned so far without using decision trees.


Project_4
File Import
StatExplore
Data Partition
Filter
Interactive Binning
Variable Selection
Transform Variables
Regression
MBR
Neural Network
Regression
Neural Network
MBR
Regression
Neural Network
Control Point
Model Comparison



lected del	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Number of Estimated Weights
	Reg2	Reg2	Regression	SalePrice	SalePrice	1.9576E8	12
	Neural	Neural	Neural Net...	SalePrice	SalePrice	1.9593E8	67
	Reg	Reg	Regression	SalePrice	SalePrice	1.9937E8	20
	Neural2	Neural2	Neural Net...	SalePrice	SalePrice	1.9987E8	40
	Neural3	Neural3	Neural Net...	SalePrice	SalePrice	2.0507E8	31
	MBR2	MBR2	MBR	SalePrice	SalePrice	2.0597E8	7
	Reg3	Reg3	Regression	SalePrice	SalePrice	2.0822E8	9
	MBR	MBR	MBR	SalePrice	SalePrice	2.0924E8	11

Selected M	Predecess	Model No	Model Des	Target Var	Target Lab	Selection Criterion: V	Train: Nun	Train: Sum	Train: Sum	Train: Tota	Train: Moc	Train: Deg	Train: Ave	Train: Roo	Train: Divi	Train: Sum	Train: Mea	Train: Root Mean Squared Error
Y	Reg2	Reg2	Regressior	SalePrice	SalePrice	1.96E+08	12	193	193	193	12	181	1.53E+08	12350.14	193	2.94E+10	1.63E+08	12752.96318
	Neural	Neural	Neural Ne	SalePrice	SalePrice	1.96E+08	67	193	193	193	67	126	1.31E+08	11430.81	193	2.52E+10	2.00E+08	14147.19464
	Reg	Reg	Regressior	SalePrice	SalePrice	1.99E+08	20	193	193	193	20	173	1.47E+08	12112.87	193	2.83E+10	1.64E+08	12793.89102
	Neural2	Neural2	Neural Ne	SalePrice	SalePrice	2.00E+08	40	193	193	193	40	153	1.35E+08	11616.93	193	2.60E+10	1.70E+08	13047.40669
	Neural3	Neural3	Neural Ne	SalePrice	SalePrice	2.05E+08	31	193	193	193	31	162	1.61E+08	12669.56	193	3.10E+10	1.91E+08	13828.74472
	MBR2	MBR2	MBR	SalePrice	SalePrice	2.06E+08	7	193	193	193	7	186	1.71E+08	13060.26	193	3.29E+10	1.77E+08	13303.74922
	Reg3	Reg3	Regressior	SalePrice	SalePrice	2.08E+08	9	193	193	193	9	184	1.65E+08	12835.51	193	3.18E+10	1.73E+08	13145.67702
	MBR	MBR	MBR	SalePrice	SalePrice	2.09E+08	11	193	193	193	11	182	1.78E+08	13334.25	193	3.43E+10	1.89E+08	13731.29684