**BAN 210 – FINAL PROJECT**

**ANALYSIS ON THE AUTO MPG DATASET**

**Roshan Nair**

**134526219**

**Predictive Analytics (BAN210ZDD.07046.2224)**

**PROFESSOR Uzair Ahmad**

**DATE: 12th August 2022**

**ANALYSIS ON THE AUTO MPG DATASET**

**INTRODUCTION:**

For the final project of BAN 210, in order to estimate the class of the target variable in the Auto MPG data, I used predictive modeling in the study to predict the values of Target variable and also used linear regression and neural network models in the assessment below. In order to determine which model will make the prediction the most accurate, I have also performed an analysis.

**OBJECTIVE OF THE ANALYSIS:**

The goal of my analysis is to answer the following questions:

• To predict value of our target variable.

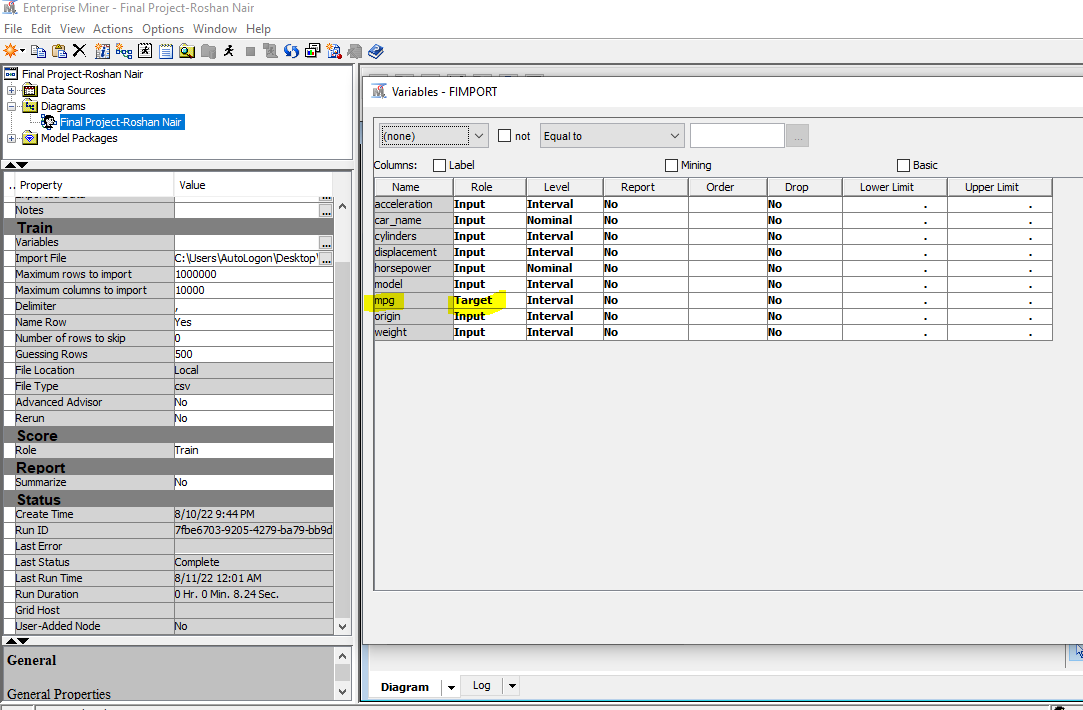
• To check which model is performing better and by how much accuracy.

**METHODOLOGY AND INFERENCES:**

Below are the steps I followed using SAS Miner to analyze the dataset:

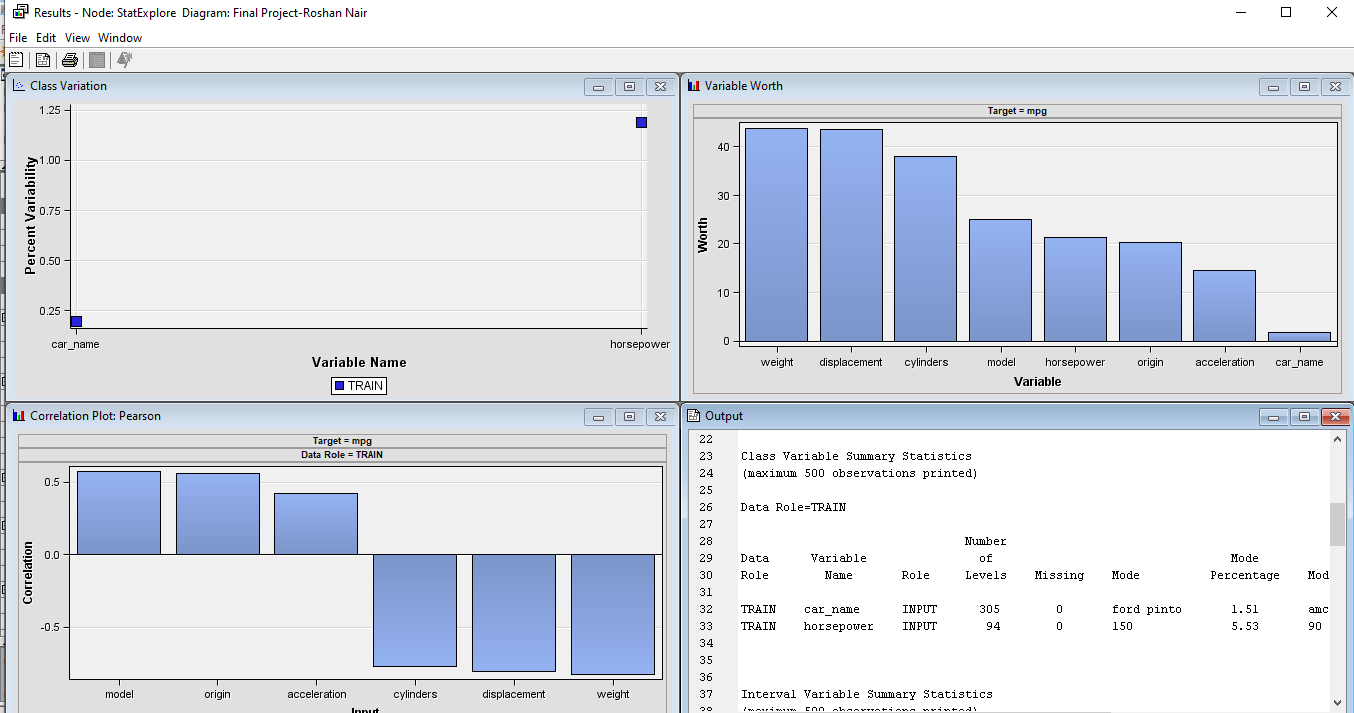
**STEP 1: FILE IMPORT**

The dataset is imported and read on the user system in the first phase using the "File Import" node. By including the link to the path on the "Import File" option in the node's Properties, the file can be read. Set Role as the "Target" for the Class variable using the "Variable" under the Properties section. The remaining characteristics are the independent variables and are referred to as "Input" variables.

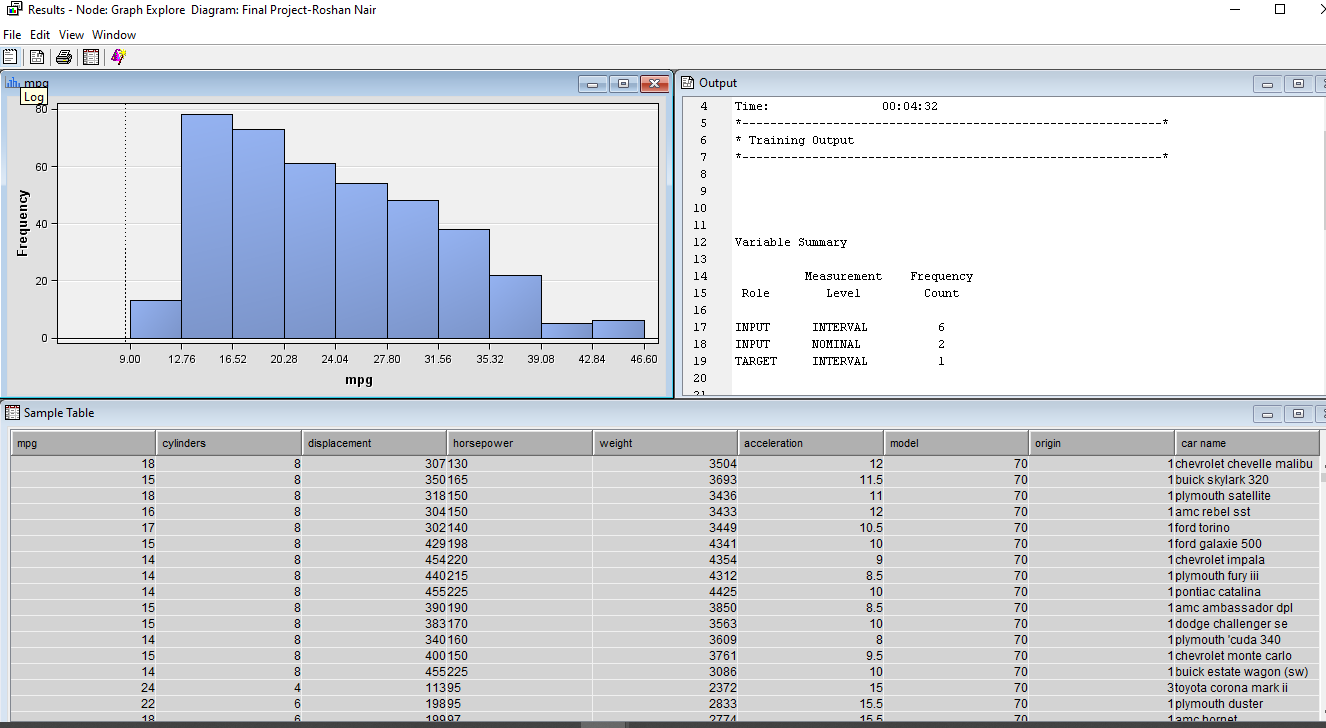


**STEP 2: STAT EXPLORE**

The StatExplore node is a multipurpose tool that we can use to examine variable distributions and statistics in our data sets. The Stat Explore node's result is seen in the screenshot that follows:

**STEP 3: GRAPH EXPLORE**

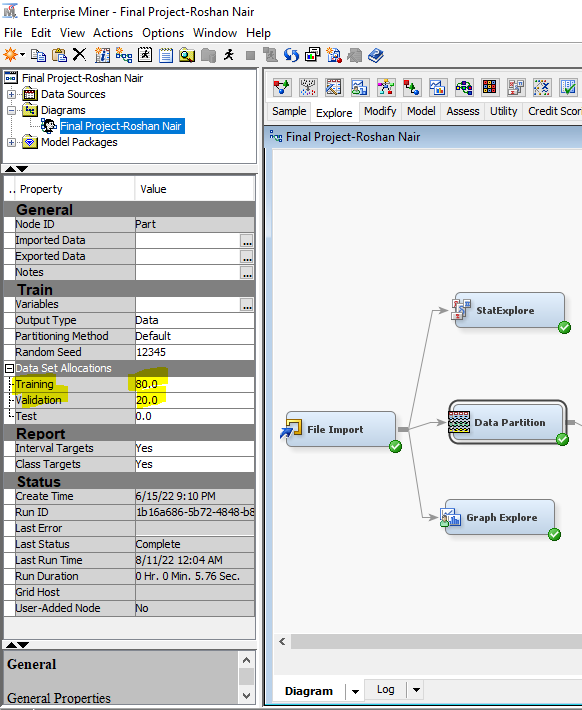
The Graph Explore node is an advanced visualization tool that enables us to explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database and following is the output that we received.



**STEP 4: DATA PARTITION**

To avoid any overfitting and under fitting, by partitioning the data as 20% Validation dataset and 80% Train dataset.

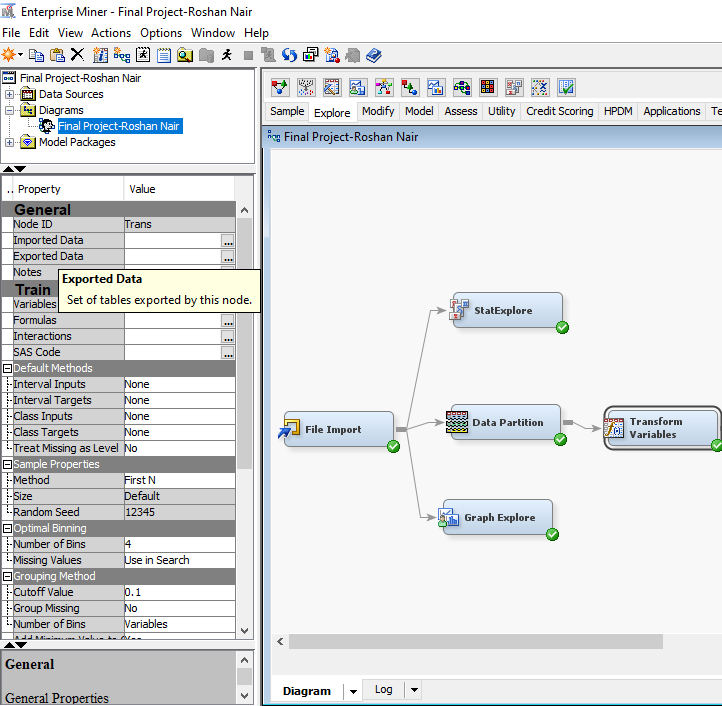
The below screenshot of the results window shows the distribution of the population.

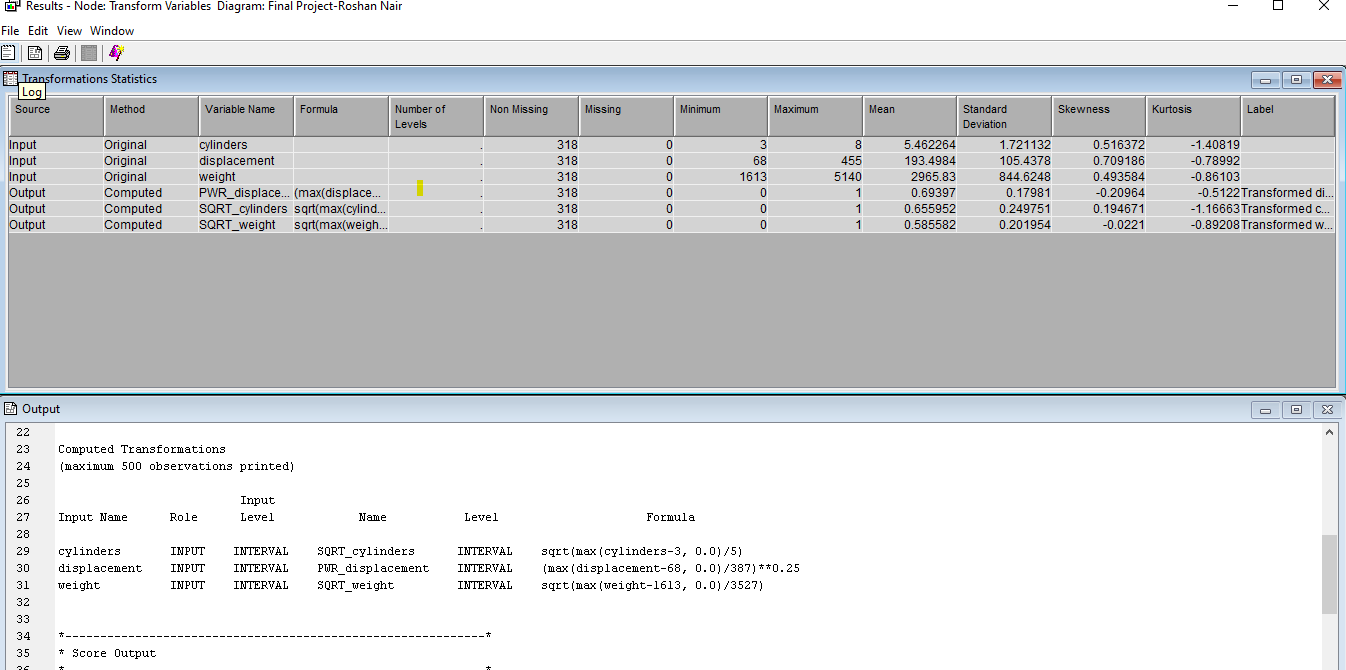




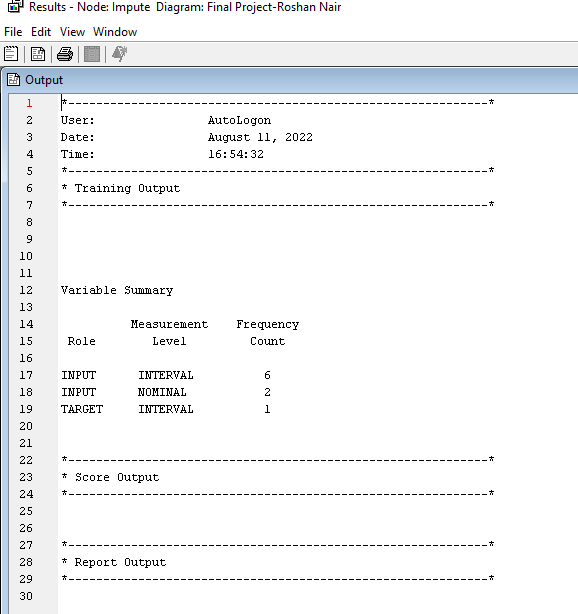
**STEP 5: DATA TRANSFORMATION**

In this step I have used transformation node because they are useful when we want to improve the fit of a model to the data. Following is the output:





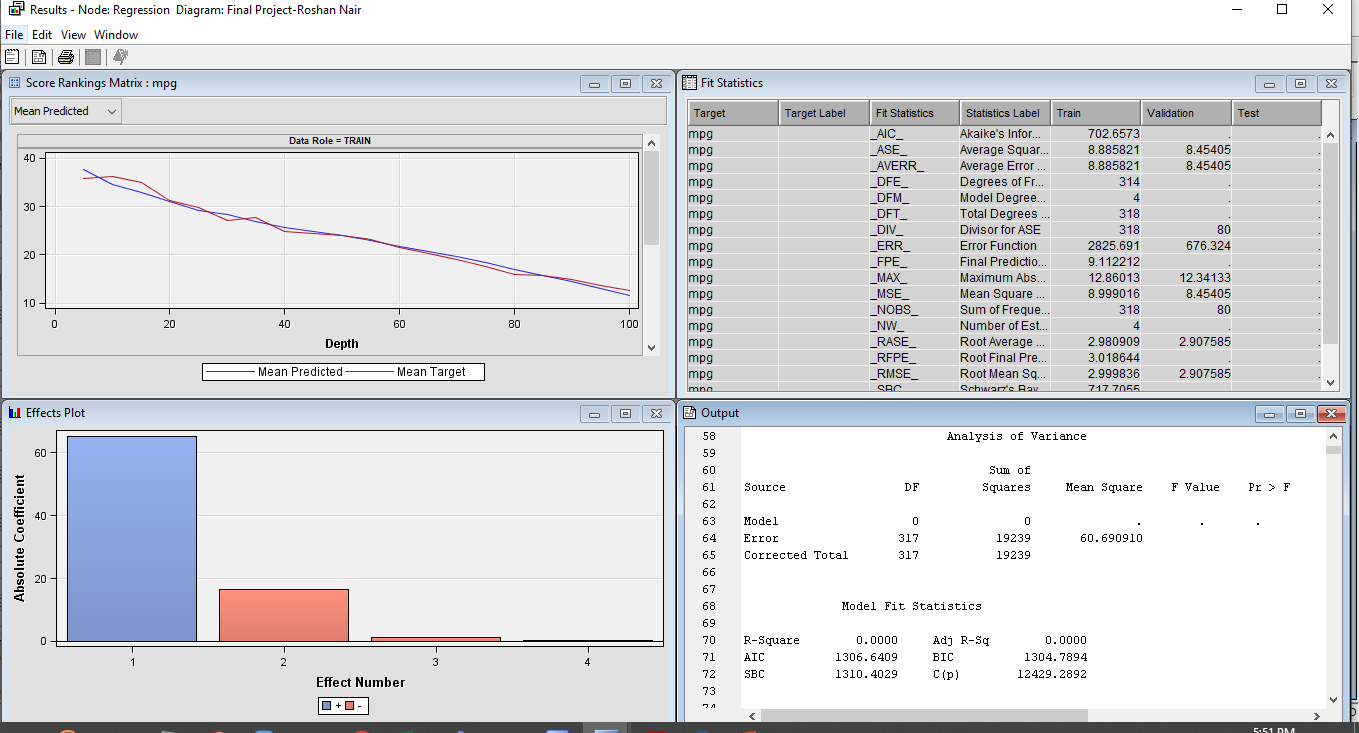
**STEP 6: IMPUTATION**

In this step I have used Impute node because Impute node to replace missing values in data sets that are used for data mining. Following is the output:

**STEP 7: REGRESSION MODEL**

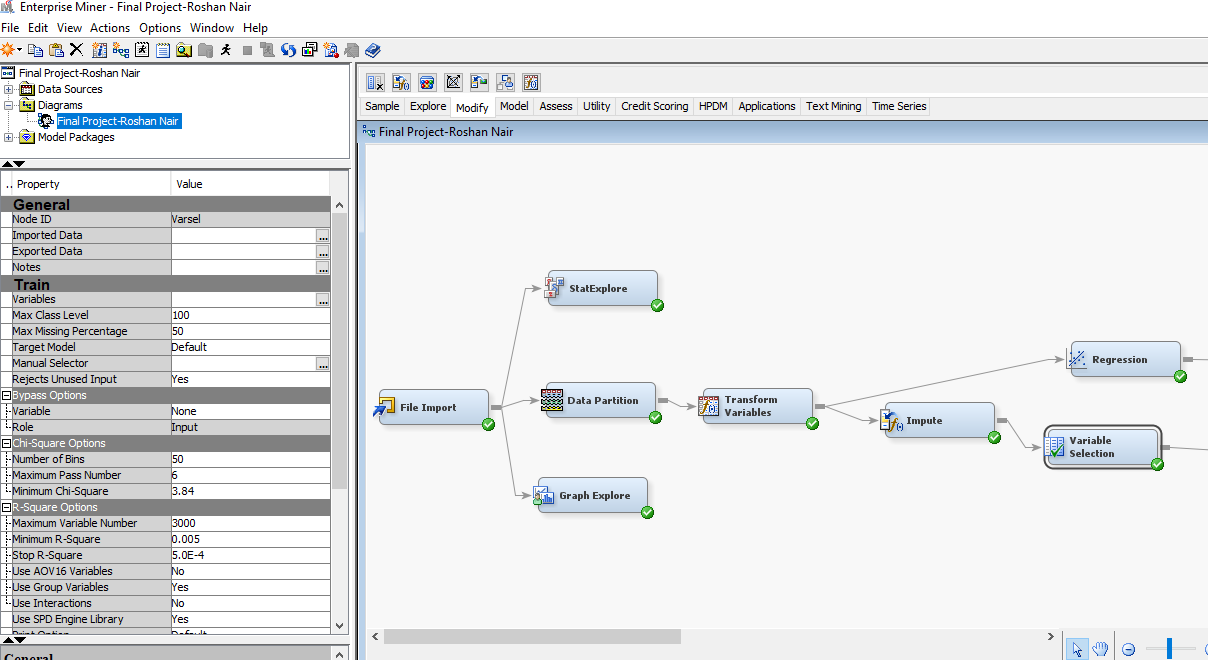
Because we are running prediction on classification variable, Linear Regression model has been used. The Regression node has been connected to the data Transformation node and Logit has been selected under Properties.

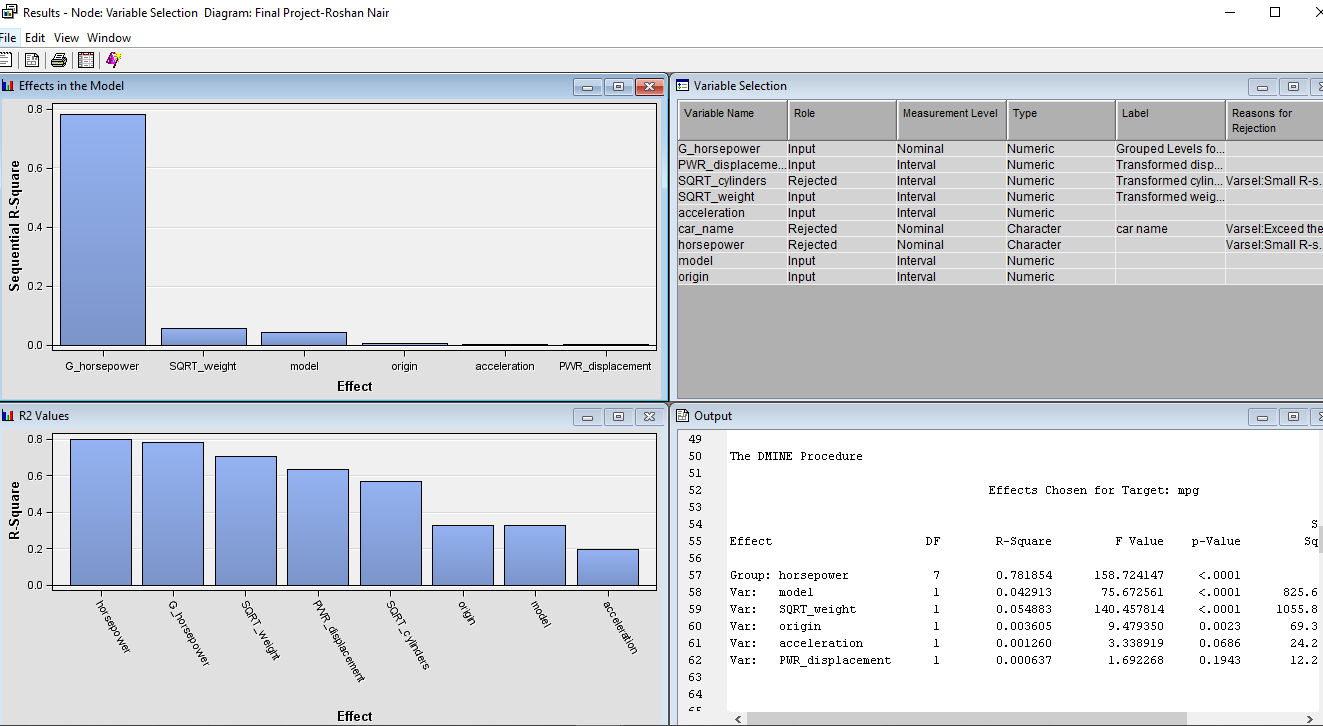
The following has been obtained from the Regression output:



**STEP 8: VARIABLE SELECTION**

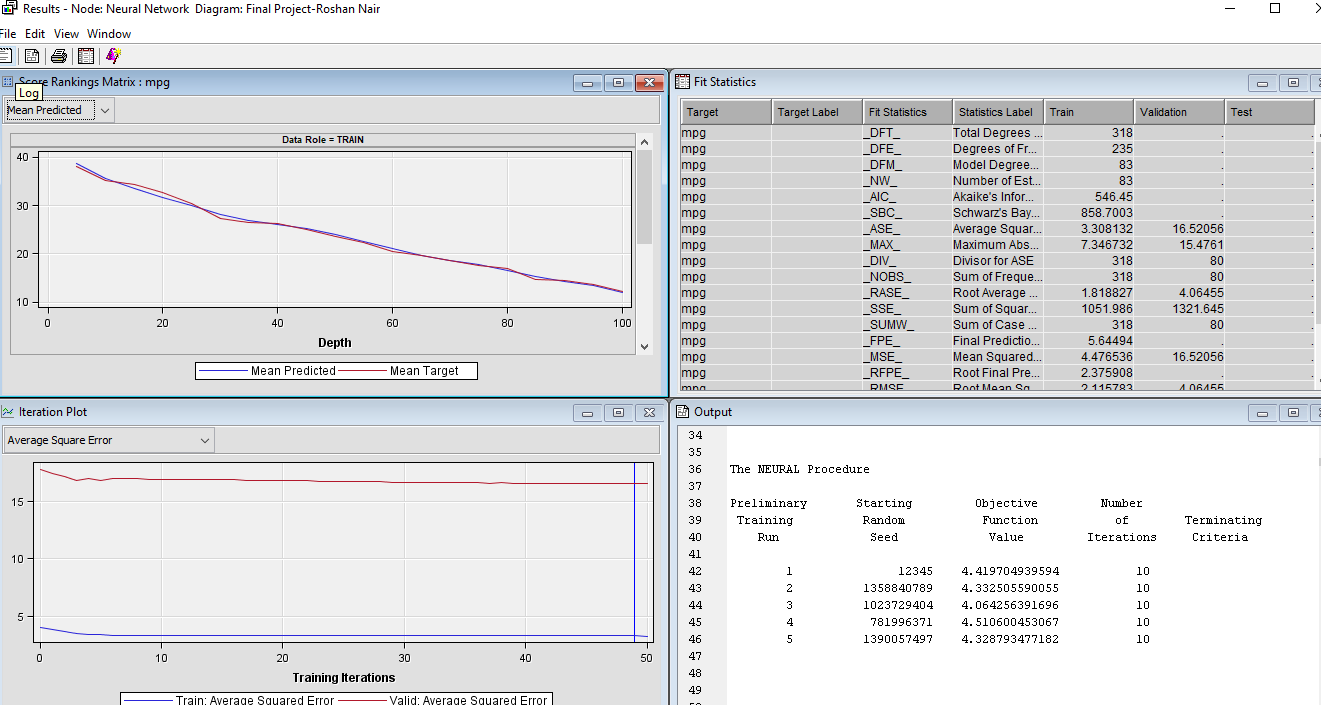
The Variable Selection node assists us in reducing the number of inputs by setting the status of the input variables that are not related to the target as rejected. Following is the output:

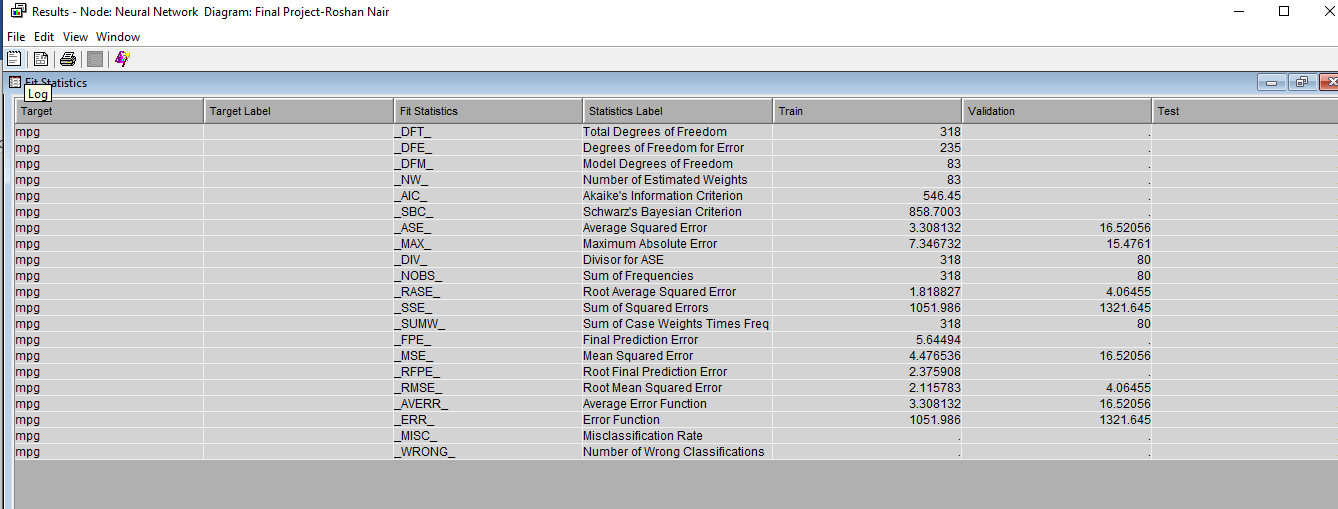




**STEP 9: NEURAL NETWORK**

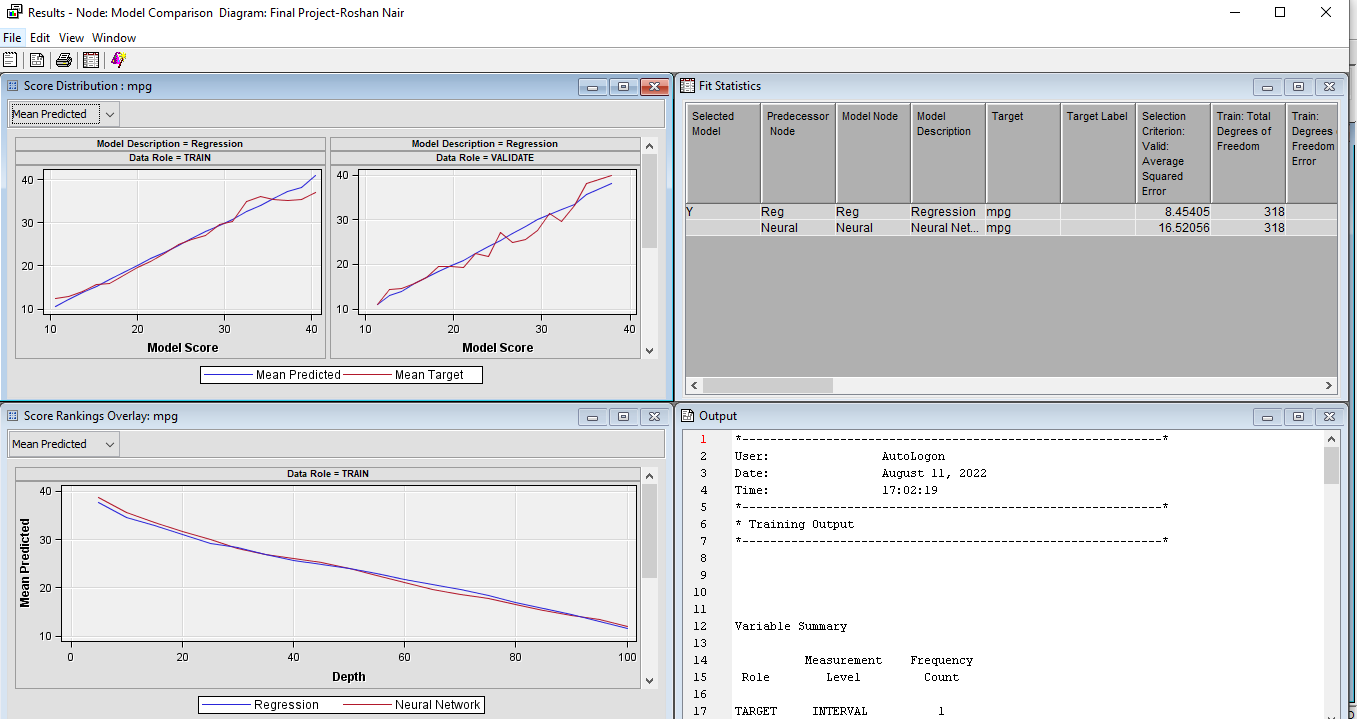
The Neural Network node allows us to score the training, validation, test, and score data sets in conjunction with training. Below are the results obtained:

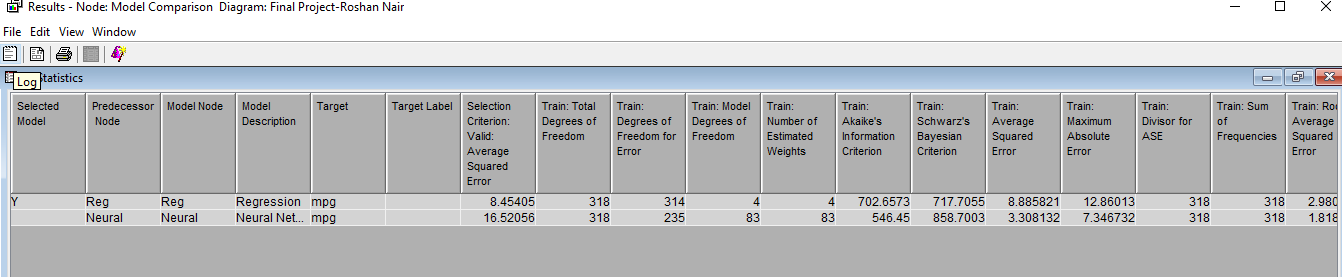




**STEP 10: MODEL COMPARISON**

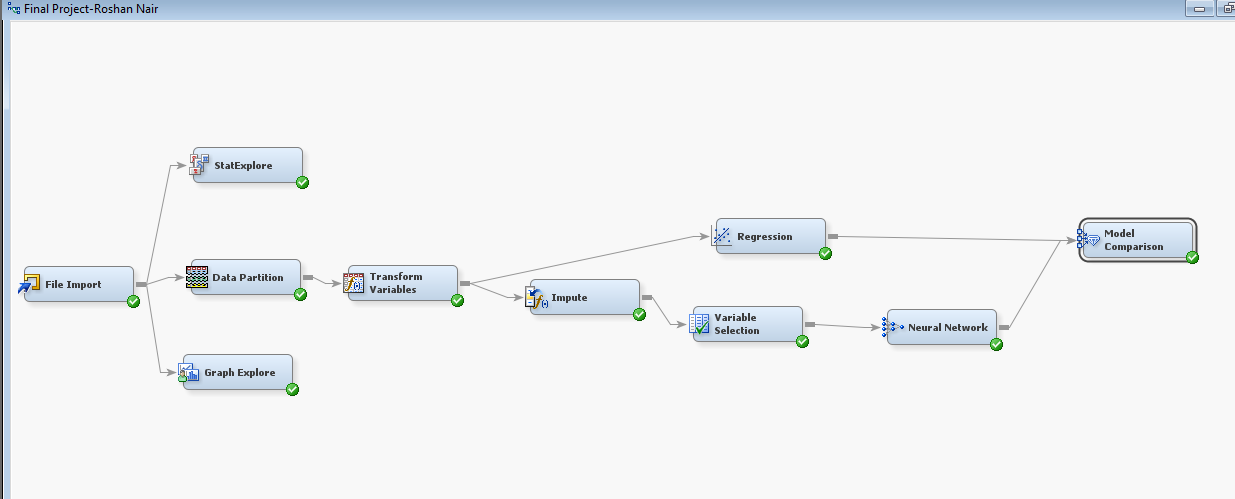
The Model Comparison node enables us to compare the performance of competing models using various benchmarking criteria. Below are the results obtained:





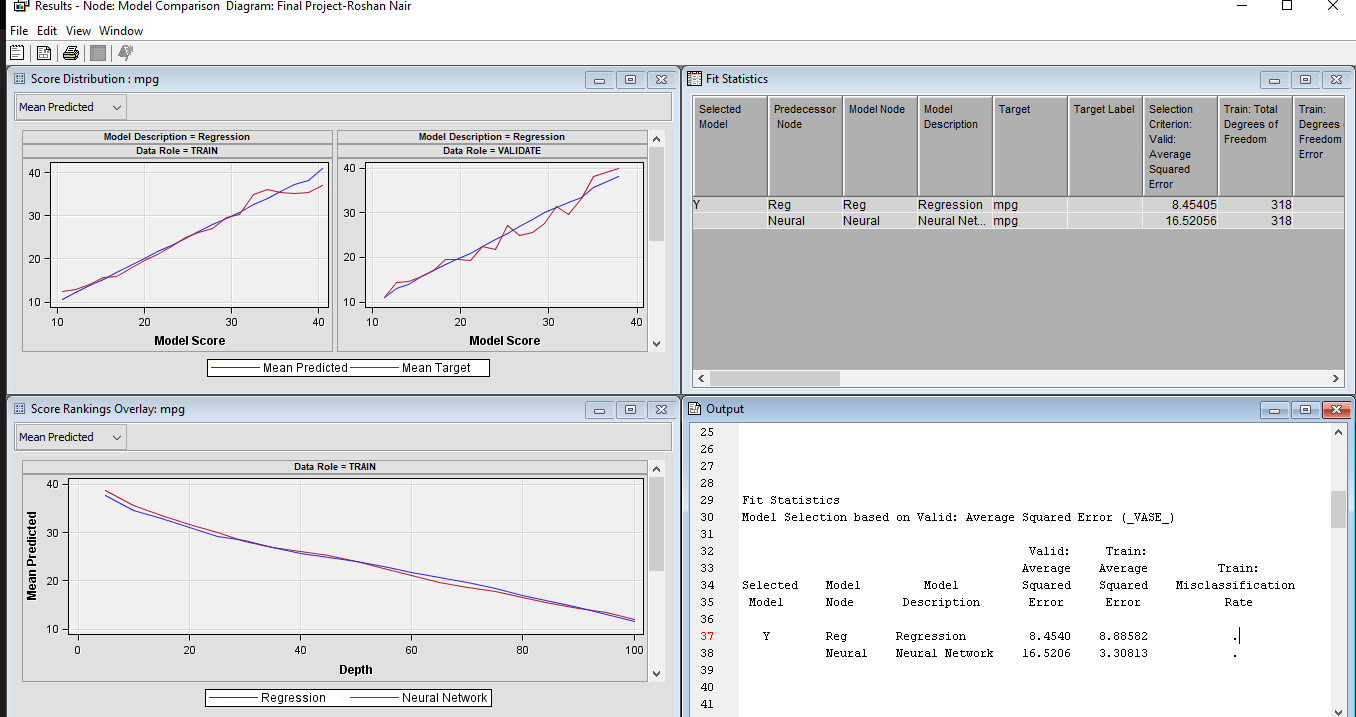
**STEP 11: FINAL DIAGRAM**

I have obtained the final diagram as below:



**CONCLUSION:**

I compared the two models inferring Mean Squared Error (MSE). Linear Regression model (MSE-8.45) is better than the Neural Network model, as the MSE Score is lower as compared to the Neural Network model (MSE-16.52) and hence it gives us the best prediction value of our target variable.



**GITHUB LINK:**

Please find below the GitHub link for the final assessment:

<https://github.com/Roshaa97/BAN-210-Final-Project--Roshan-Nair>

**DECLARATION:**

I, **Roshan Nair**, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.