

# BAN 210 – FINAL ASSESSMENT

## ANALYSIS ON THE AUTO MPG DATASET

BY-

STUDENT NAME: Rohan Sharma

STUDENT ID: 135812212

## INTRODUCTION:

In order to estimate the class of the target variable in the Auto MPG data for the final project of BAN 210, I utilised predictive modelling in the research to forecast the values of Target variable and also used linear regression and neural network models in the evaluation below. I also conducted a study to determine which model will produce the most accurate forecast.

## OBJECTIVE OF THE ANALYSIS:

My analysis's purpose is to answer the following questions:

- Predict the value of our target variable.
- Determine which model performs 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

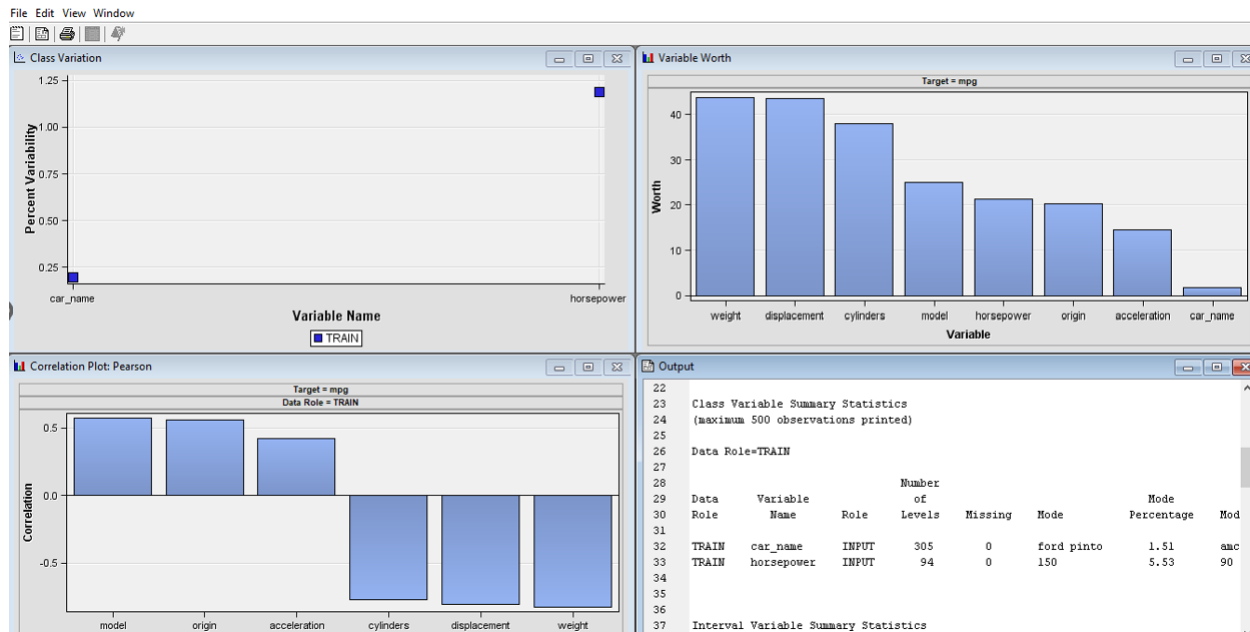
In the first step, the dataset is imported and read on the user system using the "File Import" node. The file may be read by providing the path link on the "Import File" option in the node's Properties. Set Role as the "Target" for the Class variable in the Properties section by clicking the "Variable" button. The other attributes are known as "Input" variables since they are independent variables.

The screenshot displays the SAS Miner interface. On the left, the 'Final Project-Roshan Nair' project is open, showing a tree view with 'Data Sources', 'Diagrams', and 'Model Packages'. The 'Variables - FIMPORT' dialog box is open, showing the 'Import File' path as 'C:\Users\AutoLogon\Desktop'. The 'Variables' table lists attributes with their roles and levels. The 'mpg' variable is highlighted as the 'Target' variable.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
acceleration	Input	Interval	No		No	.	.
car_name	Input	Nominal	No		No	.	.
cylinders	Input	Interval	No		No	.	.
displacement	Input	Interval	No		No	.	.
horsepower	Input	Nominal	No		No	.	.
model	Input	Interval	No		No	.	.
mpg	Target	Interval	No		No	.	.
origin	Input	Interval	No		No	.	.
weight	Input	Interval	No		No	.	.

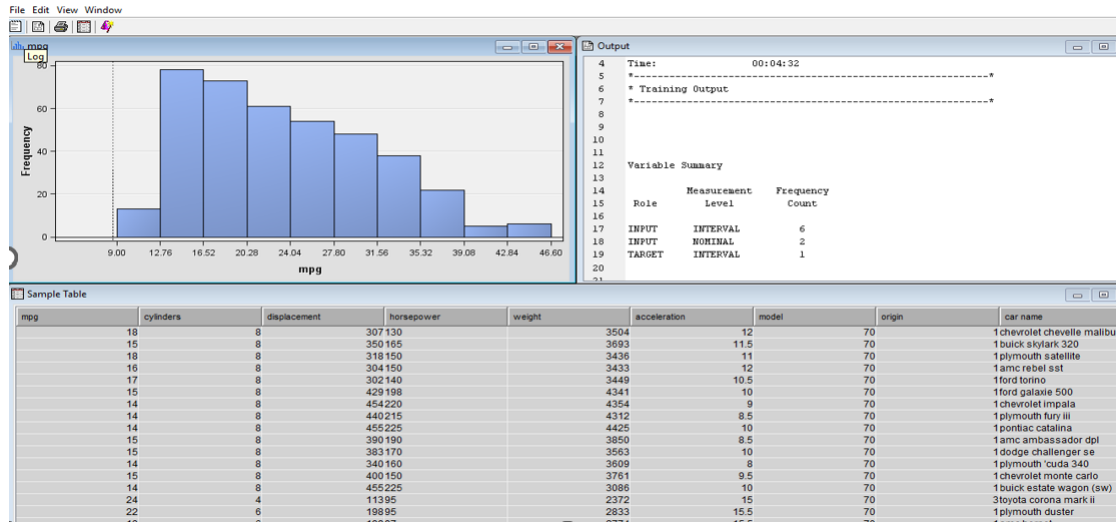
## Step 2: Stat Explore

The StatExplore node is a versatile tool for examining variable distributions and statistics in our data sets. The outcome of the StatExplore node is shown in the screenshot below.



### Step 3: Graph Explore

The Graph Explore node is a powerful visualization tool that allows us to graphically explore enormous quantities of data to identify patterns and trends and to highlight extreme values in the database, and the output that we obtained is seen below.



### Step 4: Data Partition

To avoid overfitting and underfitting, divide the data into 20% Validation and 80% Train datasets.

The population distribution is depicted in the snapshot of the results window below.

Property	Value
<b>General</b>	
Node ID	Part
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	80.0
Validation	20.0
Test	0.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	6/15/22 9:10 PM
Run ID	1b16a686-5b72-4848-b8
Last Error	
Last Status	Complete
Last Run Time	8/11/22 12:04 AM
Run Duration	0 Hr. 0 Min. 5.76 Sec.
Grid Host	
User-Added Node	No

```

43
44
45
46 Summary Statistics for Interval Targets
47
48 Data=DATA
49
50
51 Variable      Maximum      Mean      Minimum      Number of      Missing      Standard
52                                                             Observations      Deviation      Label
53   mpg          46.6      23.514572864          9          398          0      7.8159843126
54
55
56 Data=TRAIN
57
58
59 Variable      Maximum      Mean      Minimum      Number of      Missing      Standard
60                                              Observations      Deviation      Label
61   mpg          46.6      23.523899371         10          318          0      7.7904370653
62
63
64 Data=VALIDATE
65
66
67 Variable      Maximum      Mean      Minimum      Number of      Missing      Standard
68                                              Observations      Deviation      Label
69   mpg          44.6      23.4775          9          80          0      7.966241669
70

```

## Step 5: Data Transformation

I utilised transformation nodes in this stage since they are important when we want to increase the fit of a model to the data. The output is as follows:

Property	Value
<b>General</b>	
Node ID	Trans
Imported Data	...
Exported Data	...
Notes	
<b>Train</b>	<b>Exported Data</b>
Variables	Set of tables exported by this node.
Formulas	...
Interactions	...
SAS Code	...
Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.1
Group Missing	No
Number of Bins	Variables

File Edit View Window

Log

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	cylinders			318	0	3	8	5.462264	1.721132	0.516372	-1.40819	
Input	Original	displacement			318	0	68	455	193.4984	105.4378	0.709186	-0.78992	
Input	Original	weight			318	0	1613	5140	2955.83	844.5248	0.493584	-0.86103	
Output	Computed	PWR_displacement	(max(displacement-68, 0.0)/387)**0.25		318	0	0	1	0.69397	0.17981	-0.20964	-0.5122	Transformed displacement
Output	Computed	SQRT_cylinders	sqrt(max(cylinders-3, 0.0)/5)		318	0	0	1	0.655952	0.249751	0.194671	-1.16663	Transformed cylinders
Output	Computed	SQRT_weight	sqrt(max(weight-1613, 0.0)/3527)		318	0	0	1	0.585582	0.201954	-0.0221	-0.89208	Transformed weight

Output

```

22
23 Computed Transformations
24 (maximum 500 observations printed)
25
26
27 Input Name      Role      Input Level      Name      Level      Formula
28
29 cylinders      INPUT    INTERVAL    SQRT_cylinders  INTERVAL    sqrt(max(cylinders-3, 0.0)/5)
30 displacement  INPUT    INTERVAL    PWR_displacement  INTERVAL    (max(displacement-68, 0.0)/387)**0.25
31 weight        INPUT    INTERVAL    SQRT_weight      INTERVAL    sqrt(max(weight-1613, 0.0)/3527)
32
33
34 *-----*
35 * Score Output

```

## Step 6: Imputation

In this phase, I utilised the Impute node, which is used to replace missing values in data sets used for data mining. The output is as follows:

File Edit View Window

Output

```

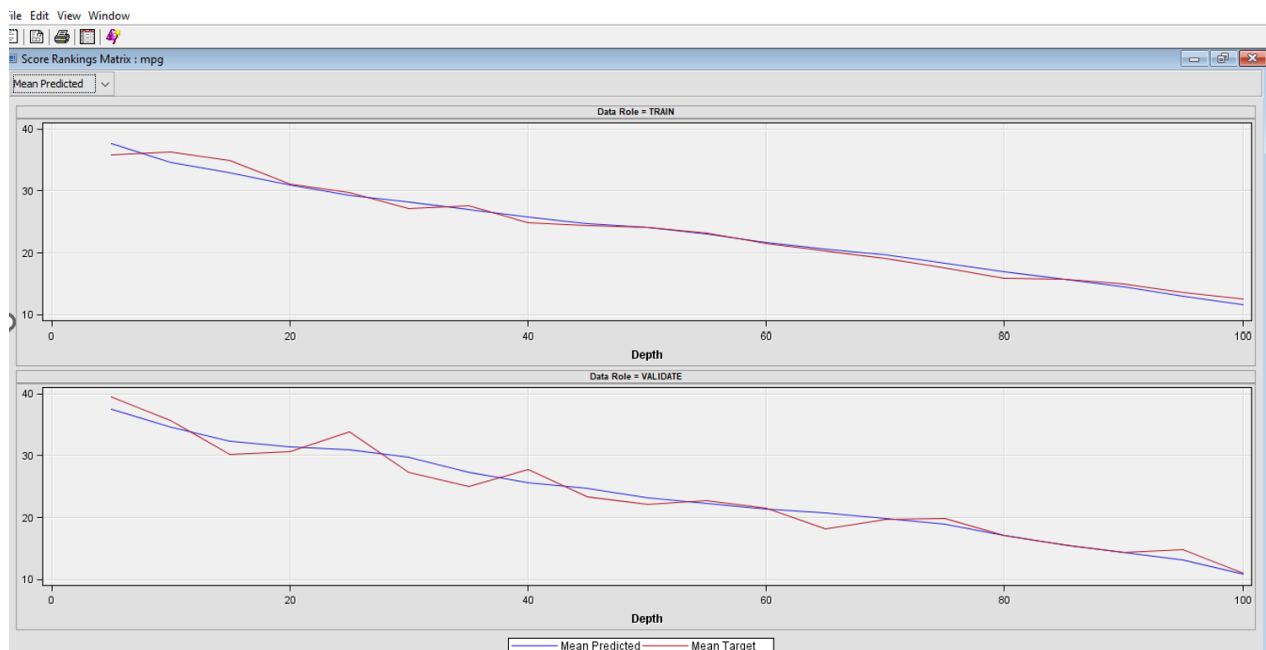
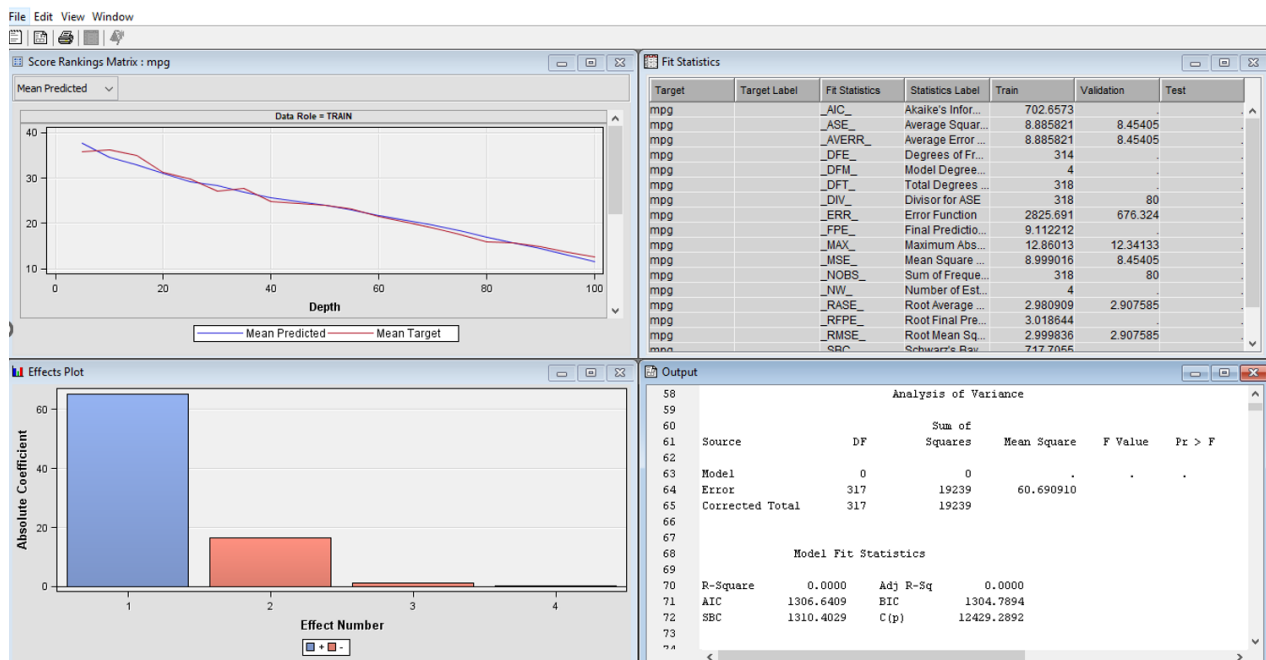
1 *-----*
2 User:          AutoLogon
3 Date:          August 11, 2022
4 Time:          16:54:32
5 *-----*
6 * Training Output
7 *-----*
8
9
10
11
12 Variable Summary
13
14      Measurement      Frequency
15      Role      Level      Count
16
17 INPUT    INTERVAL    6
18 INPUT    NOMINAL     2
19 TARGET   INTERVAL    1
20
21
22 *-----*
23 * Score Output
24 *-----*
25
26
27 *-----*
28 * Report Output
29 *-----*
30

```

## Step 7: Regression Model

Because we are predicting on a categorization variable, we have utilised the Linear Regression model. Logit has been selected under Properties after connecting the Regression node to the data Transformation node.

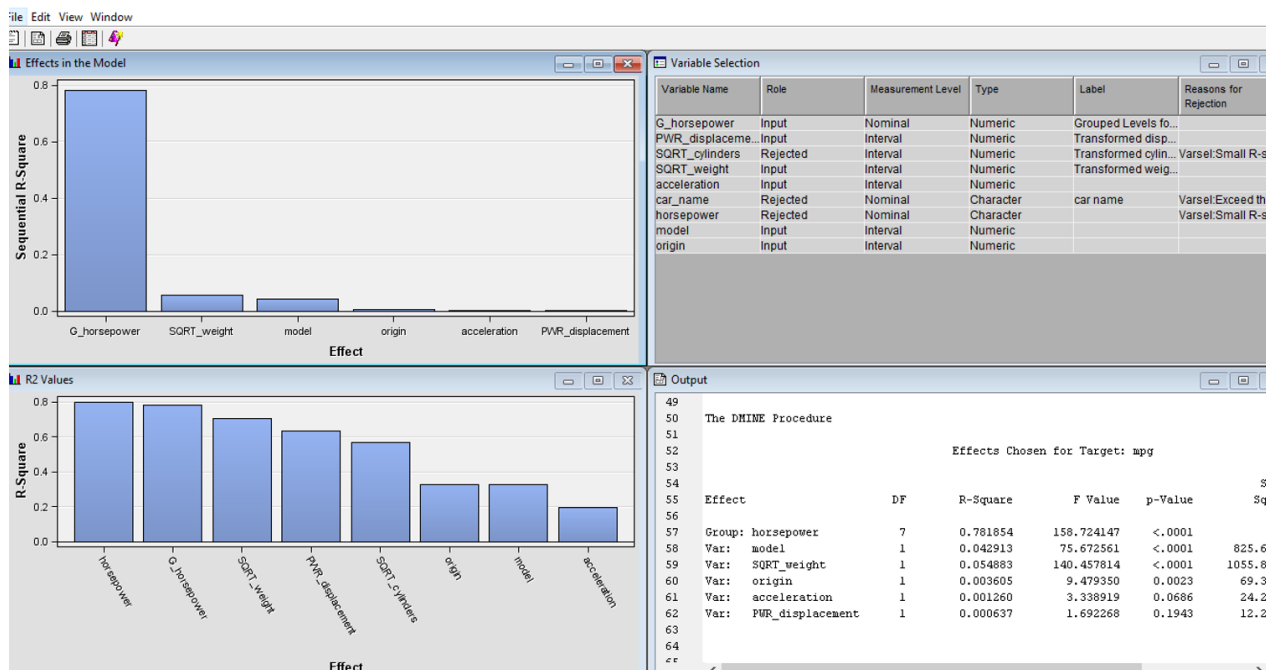
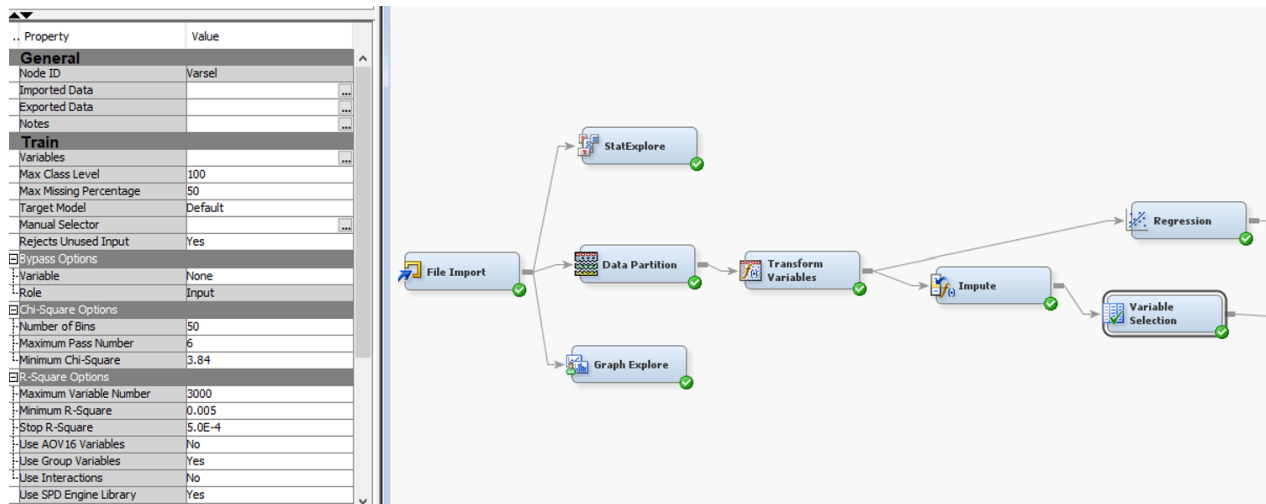
The Regression output yielded the following results:



## Step 8: Variable Selection

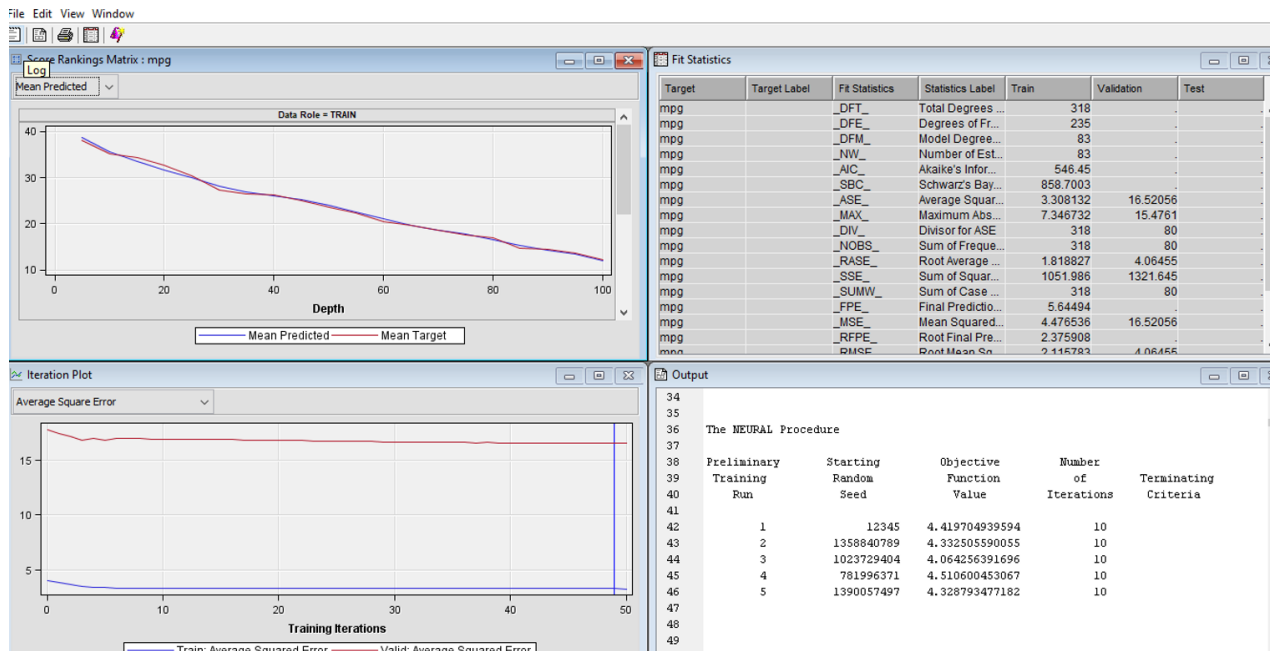
The Variable Selection node helps us reduce the amount of inputs by marking input variables that are unrelated to the objective as rejected. The output is as follows:





## Step 9: Neural Network

In addition to training, the Neural Network node allows us to score the training, validation, test, and score data sets. The following are the outcomes:



File Edit View Window

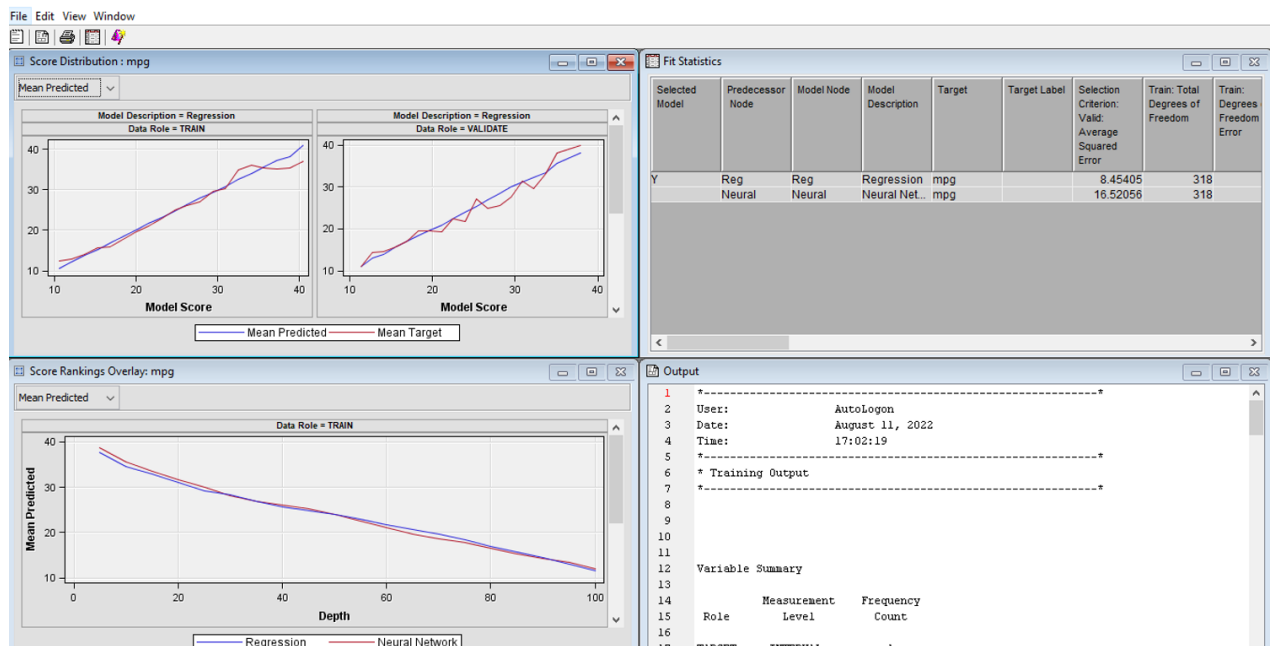
Log

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
mpg		_DFT_	Total Degrees of Freedom	318		
mpg		_DFE_	Degrees of Freedom for Error	235		
mpg		_DFM_	Model Degrees of Freedom	83		
mpg		_NW_	Number of Estimated Weights	83		
mpg		_AIC_	Akaike's Information Criterion	546.45		
mpg		_SBC_	Schwarz's Bayesian Criterion	858.7003		
mpg		_ASE_	Average Squared Error	3.308132	16.52056	
mpg		_MAX_	Maximum Absolute Error	7.346732	15.4761	
mpg		_DIV_	Divisor for ASE	318	80	
mpg		_NOBS_	Sum of Frequencies	318	80	
mpg		_RASE_	Root Average Squared Error	1.818827	4.06455	
mpg		_SSE_	Sum of Squared Errors	1051.986	1321.645	
mpg		_SUMW_	Sum of Case Weights Times Frequency	318	80	
mpg		_FPE_	Final Prediction Error	5.64494		
mpg		_MSE_	Mean Squared Error	4.476536	16.52056	
mpg		_RFPE_	Root Final Prediction Error	2.375908		
mpg		_RMSE_	Root Mean Squared Error	2.115783	4.06455	
mpg		_AVER_	Average Error Function	3.308132	16.52056	
mpg		_ERR_	Error Function	1051.986	1321.645	
mpg		_MISC_	Misclassification Rate			
mpg		_WRONG_	Number of Wrong Classifications			

## Step 10: Model Comparison

We may compare the performance of competing models using various benchmarking criteria by utilising the Model Comparison node. The following are the outcomes:



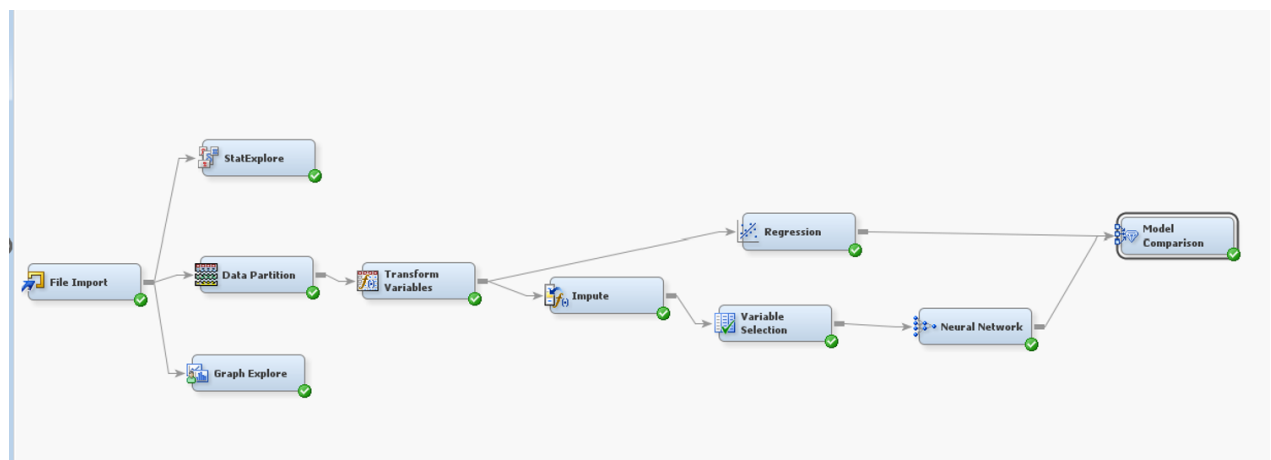
File Edit View Window

Log Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Divisor for ASE	Train: Sum of Frequencies	Train: Root Average Squared Error
Y	Reg	Reg	Regression	mpg		8.45405	318	314	4	4	702.6573	717.7055	8.885921	12.86013	318	318	2.980
	Neural	Neural	Neural Net...	mpg		16.52056	318	235	83	83	546.45	858.7003	3.308132	7.346732	318	318	1.816

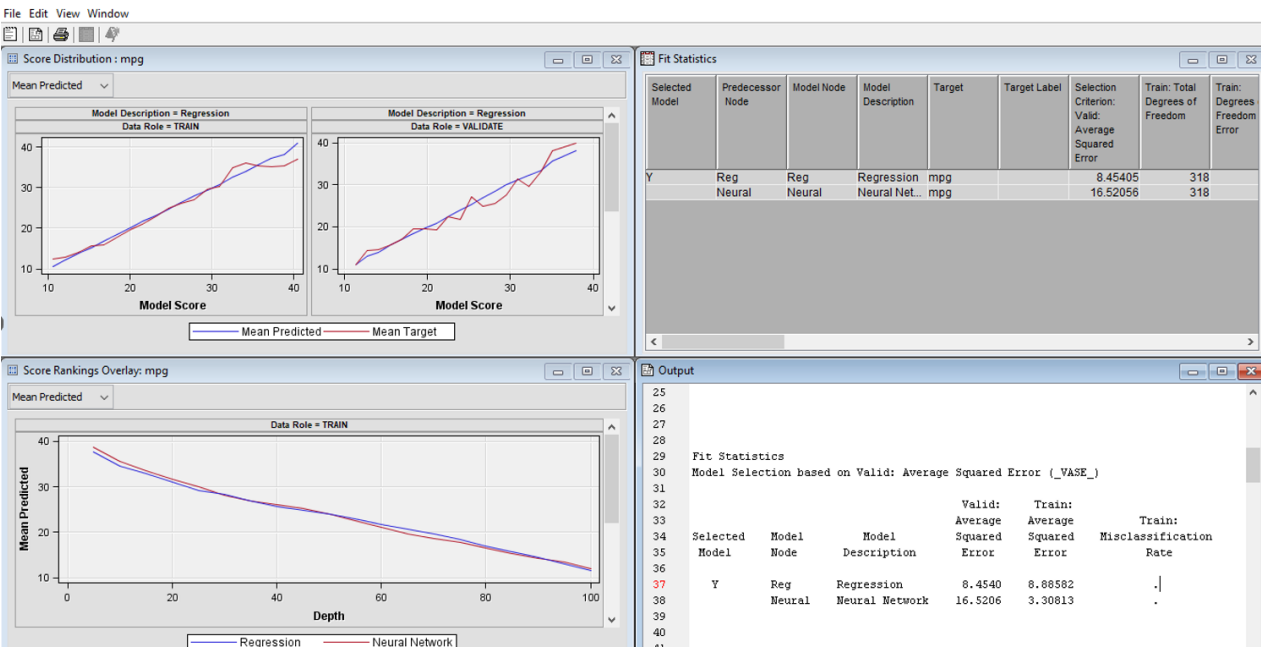
## Step 11: Final Diagram

I received the following final diagram:



CONCLUSION:

I compared the two models by calculating the Mean Squared Error (MSE). Because the MSE Score of the Linear Regression model is lower than that of the Neural Network model, it outperforms the Neural Network model.



GITHUB LINK:

Please find below the GitHub:

DECLARATION:

I, **Rohan Sharma**, 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.