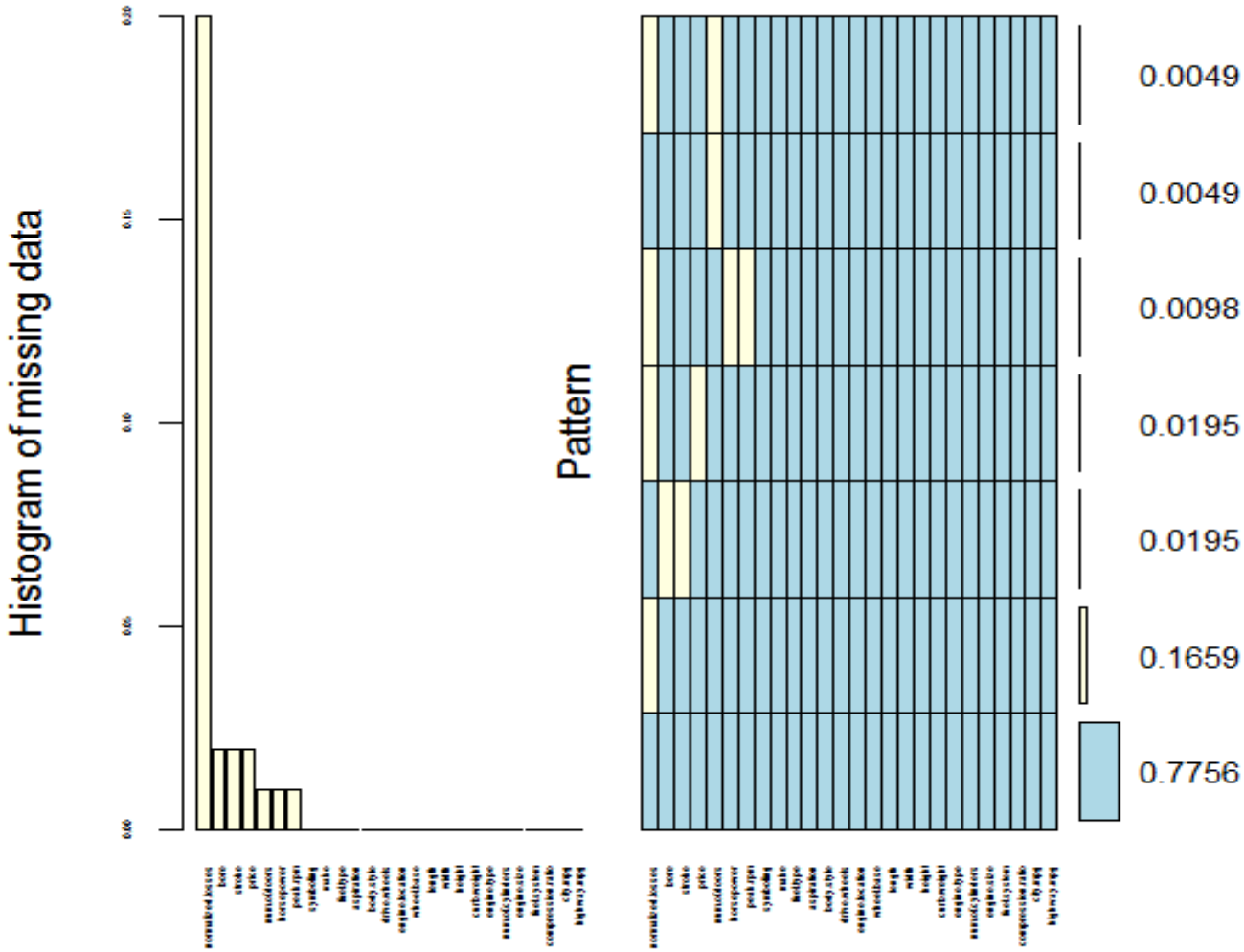
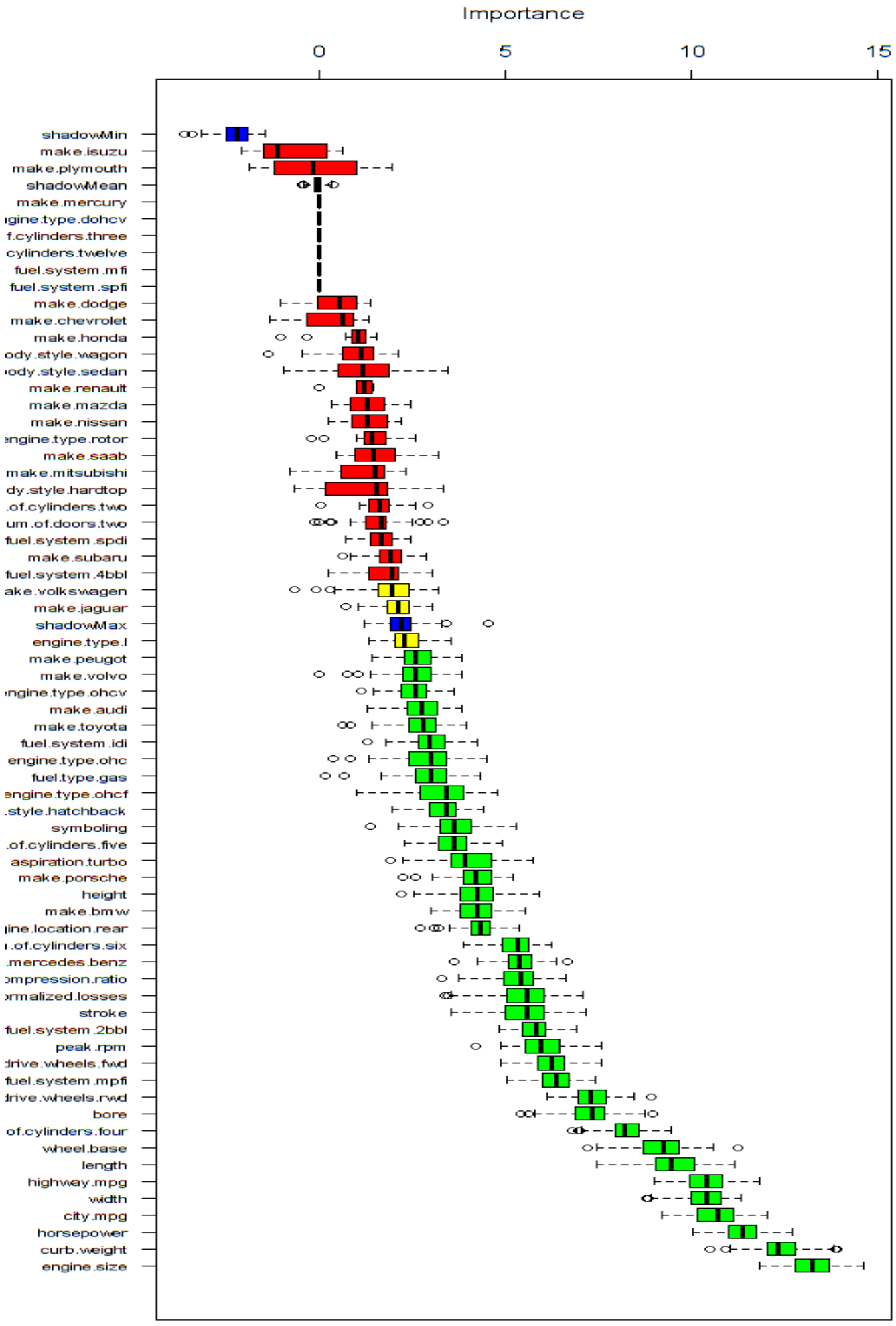
Output Summary:-

The dataset consists of the automobile details and the task is to build a model for price prediction. **Caret** package has been mainly used, the implementation of **Boruta** to find important features has also been shown.

- **Data Loading and exploration –** The Data is loaded in a way to treat unwanted strings as NAs. The structure, summary and missing values are checked. The following plot is observed for the missing values:



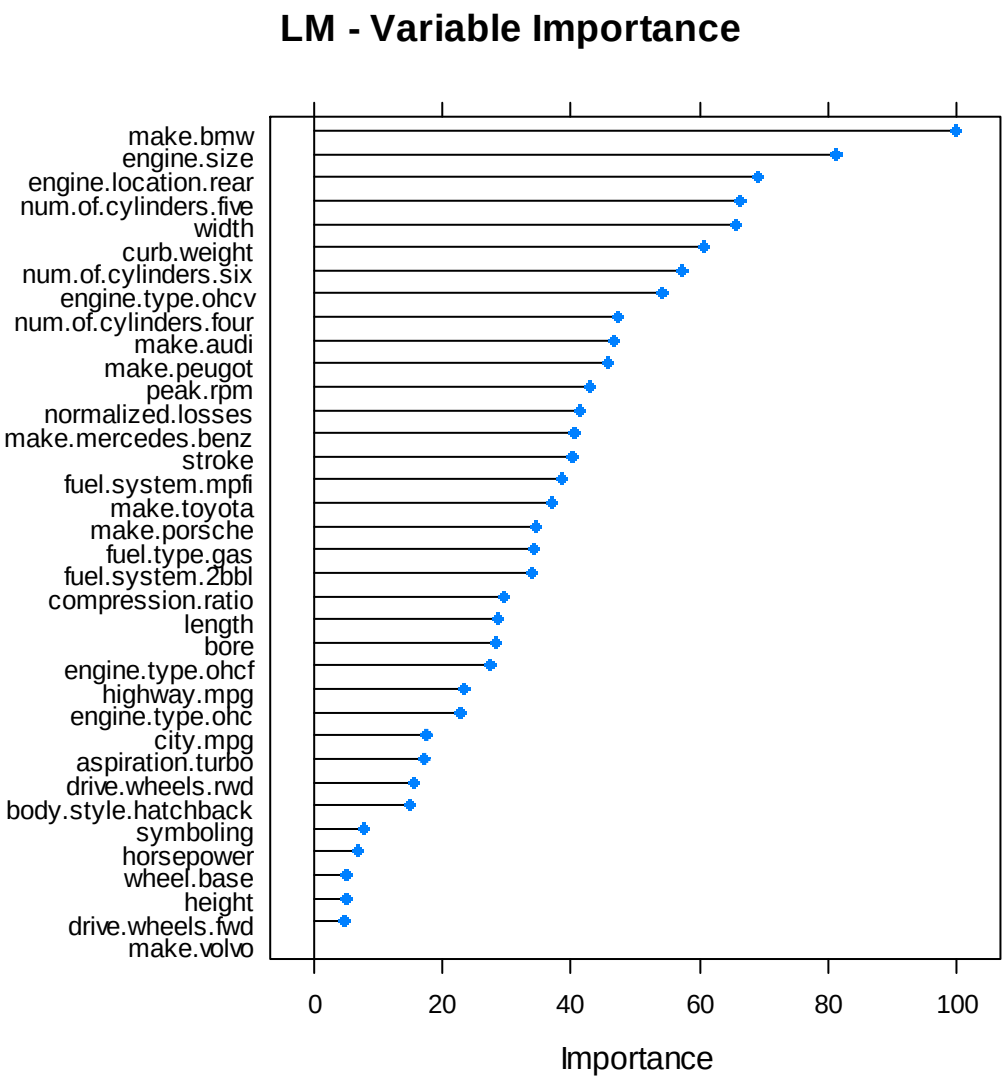
- **Preprocessing –** The data is scaled, centered and the missing values are imputed using the KNN algorithm with the help of caret. Categorical variables are then converted to numerical using the one-hot encoding
- **Feature selection using Boruta –** Boruta package is used to find the important features out of several features present in the dataset. The following plot is obtained:


Importance
0
5
10
15
shadowMin
make.isuzu
make.plymouth
shadowMean
make.mercury
igine.type.dohcv
f.cylinders.three
cylinders.twelve
fuel.system.mfi
fuel.system.spfi
make.dodge
make.chevrolet
make.honda
ody.style.wagon
ody.style.sedan
make.renault
make.mazda
make.nissan
ngine.type.rotor
make.saab
make.mitsubishi
dy.style.hardtop
.of.cylinders.two
um.of.doors.two
fuel.system.spdi
make.subaru
fuel.system.4bbl
ake.volkswagen
make.jaguar
shadowMax
engine.type.l
make.peugot
make.volvo
ngine.type.ohcv
make.audi
make.toyota
fuel.system.idi
engine.type.ohc
fuel.type.gas
engine.type.ohcf
.style.hatchback
symboling
.of.cylinders.five
aspiration.turbo
make.porsche
height
make.bmw
ine.location.rear
.of.cylinders.six
.mercedes.benz
ompression.ratio
rmalized.losses
stroke
fuel.system.2bbl
peak.rpm
drive.wheels.fwd
fuel.system.mpfi
drive.wheels.rwd
bore
of.cylinders.four
wheel.base
length
highway.mpg
width
city.mpg
horsepower
curb.weight
engine.size

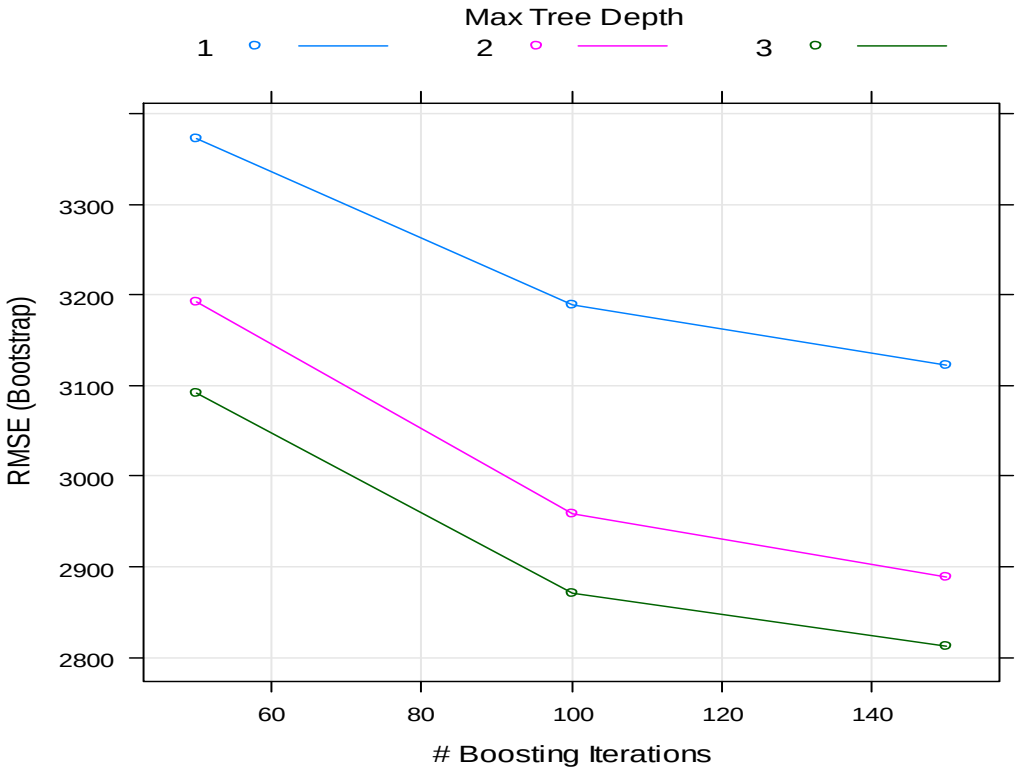
Using the getSelectedAttributes function, the important features are selected for model bulding.

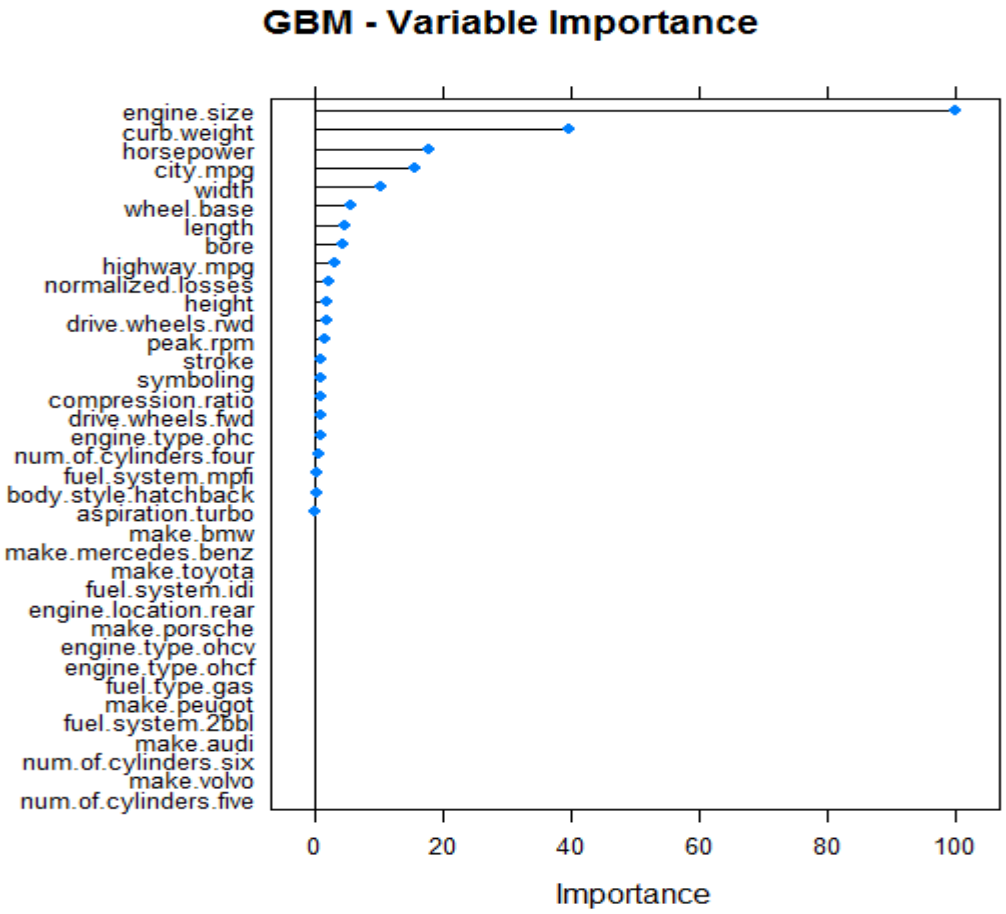
- **Splitting the data into train and test set –** The dataset is split into training and testing in the ratio of 8:2 using createDataPartition function.
- **Model Training –** Model training is performed on the training dataset using the train function from the caret package. Three algorithms – Linear model, GLM and NNET has been used to perform predictions. The following plots have been obtained for the three algorithms :

Output Plot for LM :

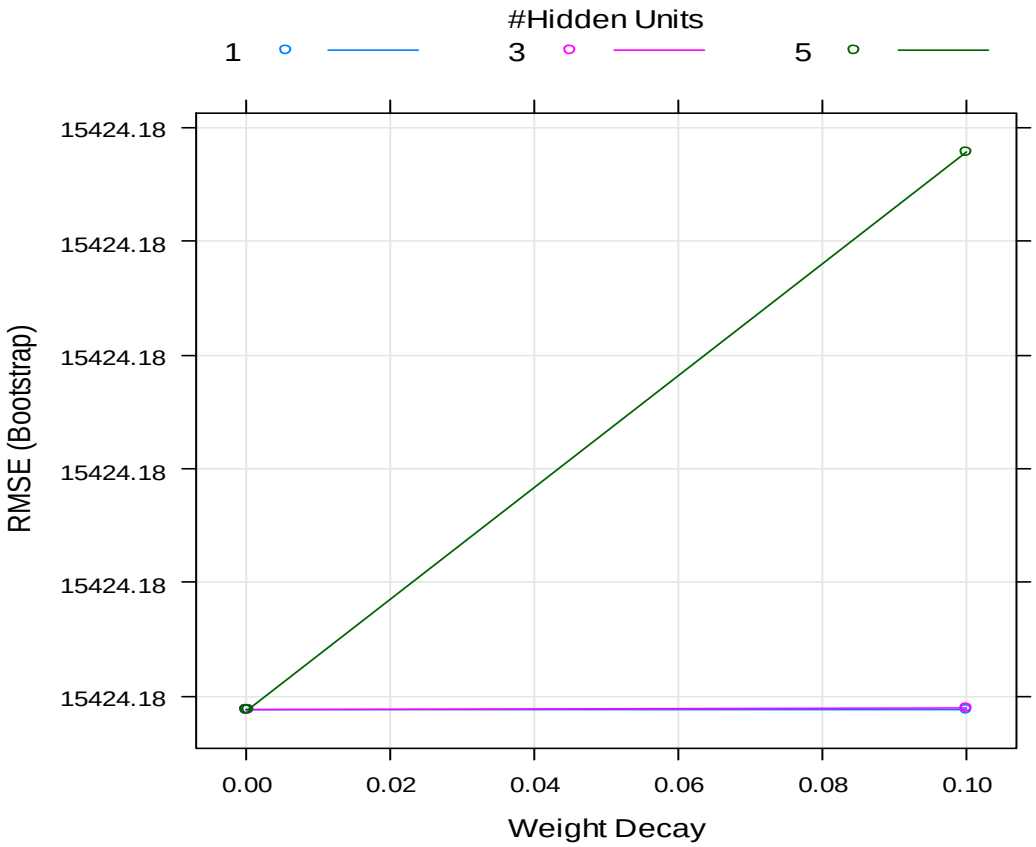


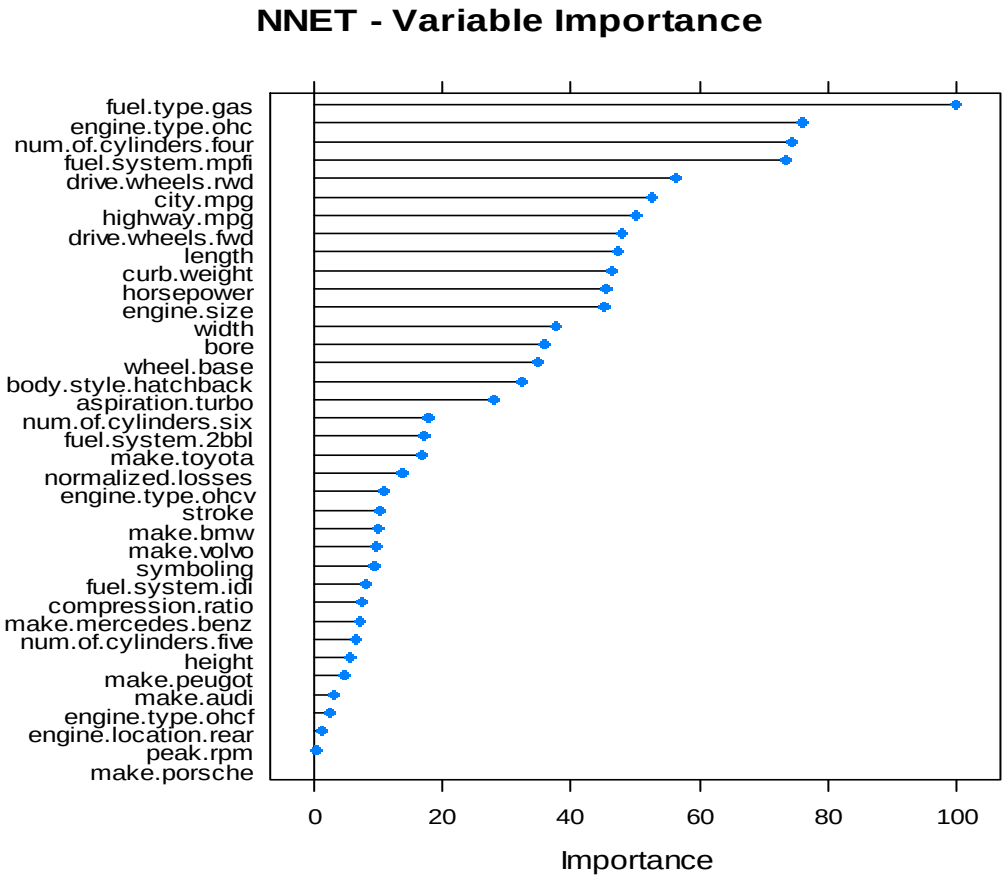
GBM Plots :


Max Tree Depth
1
2
3
RMSE (Bootstrap)
3300
3200
3100
3000
2900
2800
60
80
100
120
140
Boosting Iterations


GBM - Variable Importance
engine.size
curb.weight
horsepower
city.mpg
width
wheel.base
length
bore
highway.mpg
normalized.losses
height
drive.wheels.rwd
peak.rpm
stroke
symboling
compression.ratio
drive.wheels.fwd
engine.type.ohc
num.of.cylinders.four
fuel.system.mpfi
body.style.hatchback
aspiration.turbo
make.bmw
make.mercedes.benz
make.toyota
fuel.system.idi
engine.location.rear
make.porsche
engine.type.ohcv
engine.type.ohcf
fuel.type.gas
make.peugot
fuel.system.2bbl
make.audi
num.of.cylinders.six
make.volvo
num.of.cylinders.five
0
20
40
60
80
100
Importance

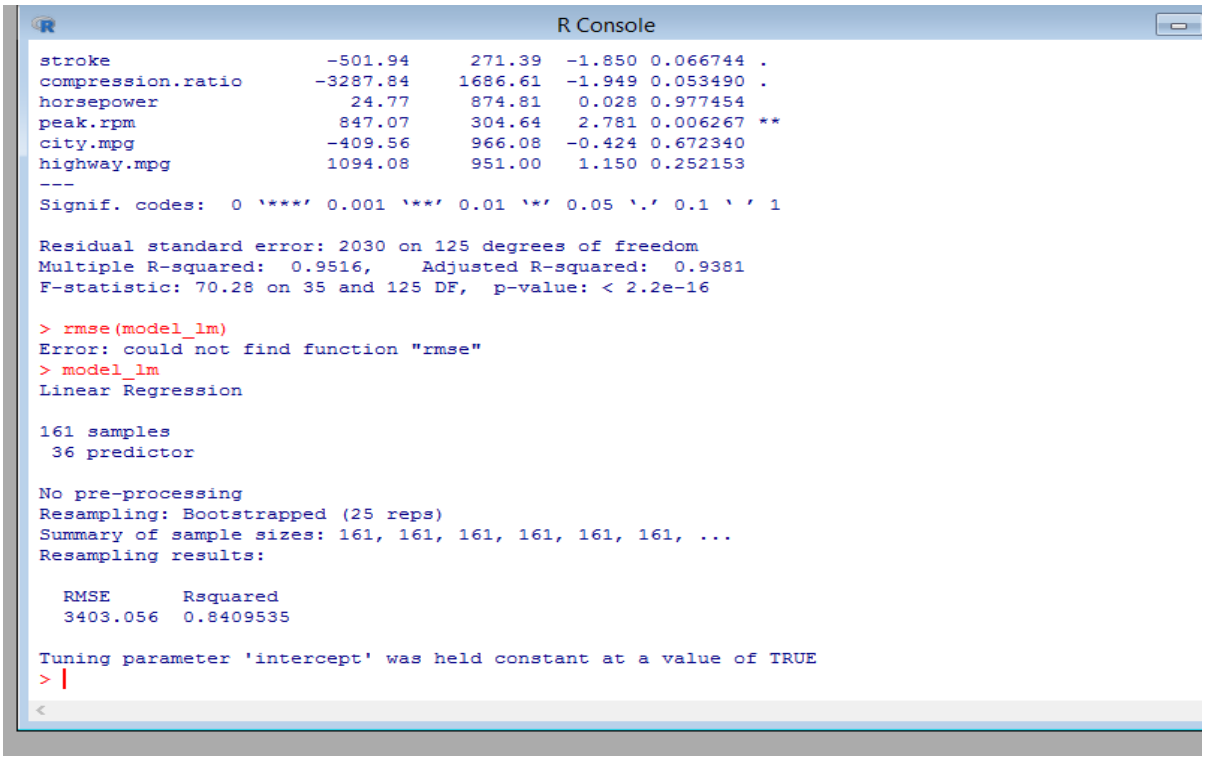
NNET Output Plots :





- **Price Prediction & Model Evaluation –** The prices can be predicted using any of the above used algorithms. Linear Regression model has been used to compute the required errors:

Mean Absolute Error = mean(abs(predictions - testSet$price))
MAE = 2178.836
RMSE = 3403.056



Relative Absolute Error = (mean(abs(predictions- testSet$price))) / (mean(abs(mean(testSet$price)-testSet$price)))
RAE = 0.3782024

Relative Squared Error = (mean((predictions-testSet$price)^2))/(mean((mean(testSet$price)-testSet$price)^2))
RSE = 0.1866436

Coefficient of Determination(R^2) = 0.8409535