DATA MINING & MACHINE LEARNING - C.A

1. LINEAR REGRESSION

1.1 BUSINESS UNDERSTANDING

This dataset is referenced from the UCI Machine Learning Repository, named Auto MPG which is about prediction of the Fuel consumption in the miles of an automobile, especially a car. In terms of MPG (Miles Per Gallon)

Link of the dataset: https://archive.ics.uci.edu/ml/datasets/Auto+MPG

1.2. DATA UNDERSTANDING AND PREPARATION

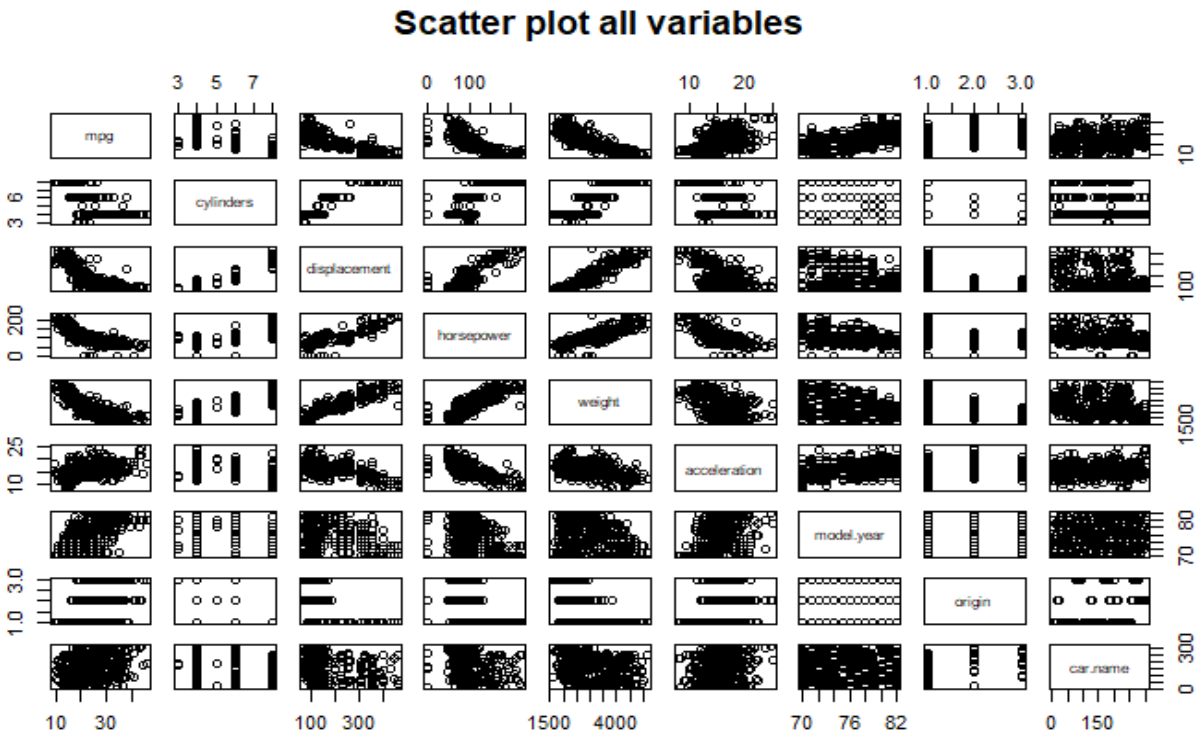
1.2.1 Data Understanding

This dataset has 9 variables and total of 398 instances. Variables such as MPG -Miles per Gallon, Cylinders, Displacement, horsepower, weight, accelerator, model year, origin and car name, In this variables, where mpg, displacement, horsepower, weight and accelerator are continuous integer values and Cylinders, Model year, Origin are Multi valued discrete integers and finally car name is the string with each instances respectively.

1.2.2 Preparation

As part of the dataset preparation, which needs to be pre-processed before we are importing to the R-studio, this dataset was not available in direct csv format instead as. data file and which is to be processed in spreadsheet and data processing work and cleaning the csv file has done.

Splitting the dataset into 80 percentage for the train dataset and 20 percentage for the test data set and removing the variables which contain (strings)



1.3. MODELLING

1.3.1 Fitting Multiple Linear Regression of model 1:

In the first model of Multiple linear regression in train dataset, I have added independent variable such as cylinders, displacement, horsepower, weight, origin, acceleration and model year. With dependent variable MPG. The results show us the **Multiple R-squared** value as **0.8179** and **Adjusted R-squared as 0.8138**, From there starts the backward elimination for higher P-value Variables and less significant variables. And this model has the highest R-squared values which has the high probability of fitting. We shall further get this by evaluating its Mean Square Error (MSE) and will find the best fitted model.

1.3.2 Fitting Multiple Linear Regression of model 2:

After removing the variable Acceleration which has high p- value now again the model is fitted, which calculated **Multiple R-squared** as **0.8168** and the **Adjusted R-squared** as **0.8139**. Here Acceleration which has higher p value is removed for the next model.

1.3.3 Fitting Multiple Linear Regression of model 3:

Here Again in this model after removing the unnecessary variable and we calculated the linear regression and we received the **Multiple R-squared** value as **0.8152**, and **Adjusted R-squared** value as **0.8129**.

1.4. EVALUATION

I have carried two phases of evaluation, one is model evaluation and Prediction evaluation, First, model evaluation is based on the Mean Square error and prediction of test dataset is based upon the Root Mean Square Error (RMSE)

1.4.1 Model Evaluation

Generally, we have evaluated the model performance based upon the Multiple R-Squared value and Adjusted R-Square values, here additionally we shall find based upon the mean square errors of each model. Based upon this lowest value of MSE, which is the **First model with highest R-Squared value and Adjusted R- Square value is best Fitted for Linear Regression**

```
> model1_summ<-summary(model1)
> mean(model1_summ$residuals^2) # 11.41574
[1] 11.41574
> # model 2: Mean Squared Error
> model2_summ<-summary(model2)
> mean(model2_summ$residuals^2) # 11.48437
[1] 11.48437
> # model 3 : Mean Squared Error
> model3_summ<-summary(model3)
> mean(model3_summ$residuals^2) # 11.58157
[1] 11.58157
```

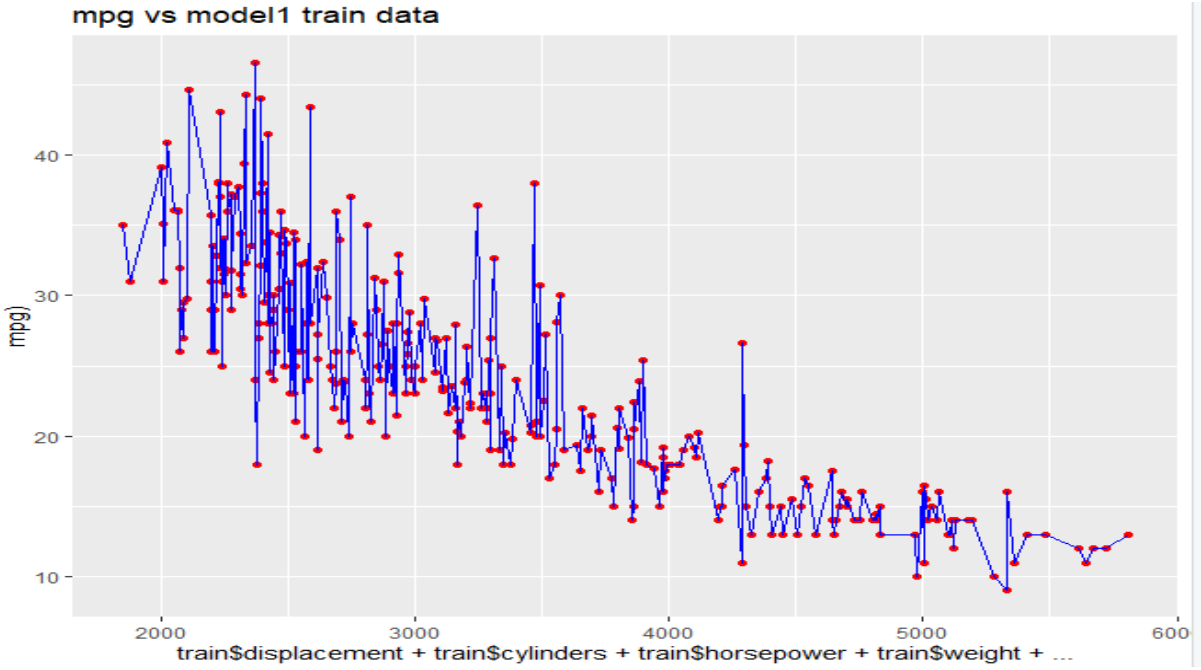
1.4.2 Prediction Evaluation

Prediction evaluation is based upon the test data and best predicted model is found using the RMSE (Actual -Predicted Values) and from this analysis we conclude that the **Model 1 is the best fit with lowest RMSE values**

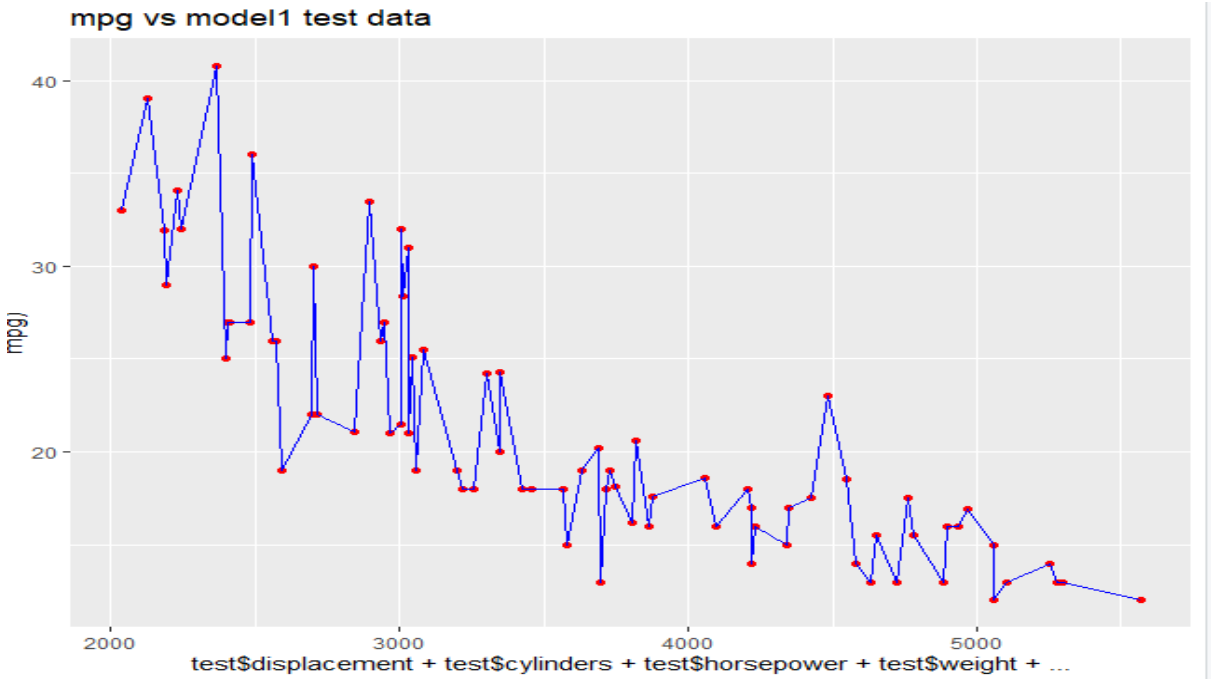
```
> # step 3.2 RSME calculation for(ACTUAL-PREDICITED) models
> # for model 1
> sqrt(mean((test$mpg - prediction1)^2))  # 2.998772
[1] 2.998772
> # for model 2
> sqrt(mean((test$mpg - prediction2)^2)) # 3.022216
[1] 3.022216
> # for model 3
> sqrt(mean((test$mpg - prediction3)^2)) # 3.030346
[1] 3.030346
```

1.4.3 Plots for best fitted Model: 1

Training:



Prediction:



2.POLYNOMIAL REGRESSION

2.1. BUSINESS UNDERSTANDING

This dataset is referenced from the UCI Machine Learning Repository, named Auto MPG which is about prediction of the Fuel consumption in the miles of an automobile, especially a car. In terms of MPG (Miles Per Gallon)

Link of the dataset: https://archive.ics.uci.edu/ml/datasets/Auto+MPG

2.2. DATA UNDERSTANDING AND PREPARATION

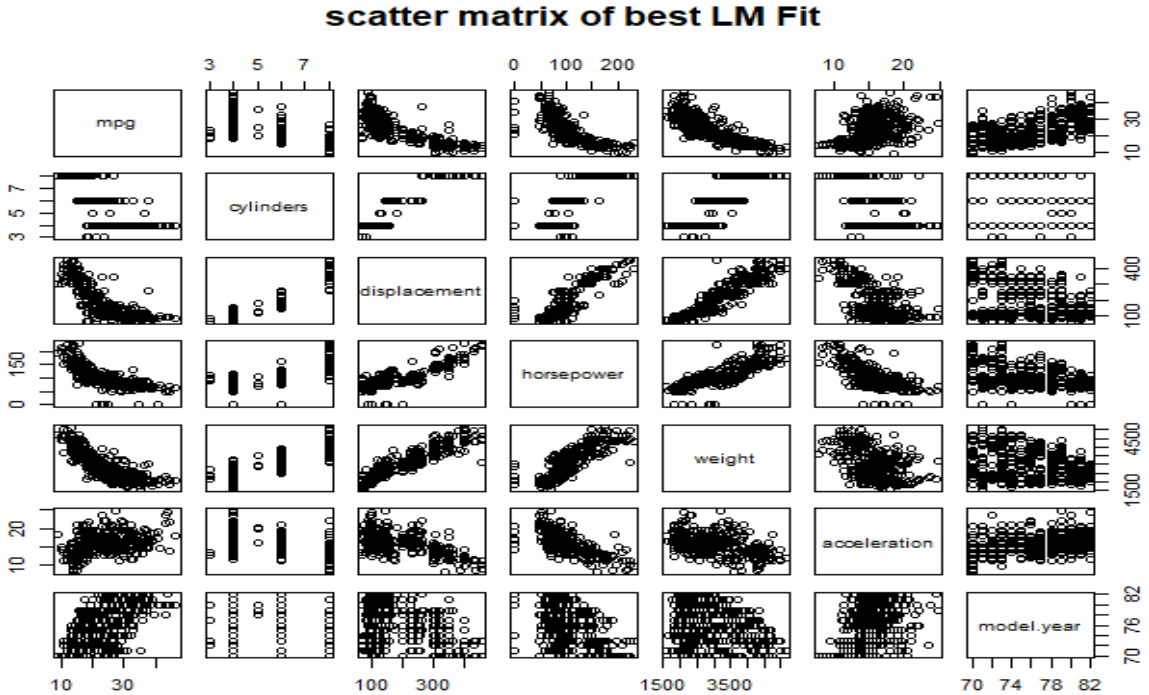
2.2.1 Data Understanding

This dataset has 9 variables and total of 398 instances. Variables such as MPG -Miles per Gallon, Cylinders, Displacement, horsepower, weight, accelerator, model year, origin and car name, In this variables, where mpg, displacement, horsepower, weight and accelerator are continuous integer values and Cylinders, Model year, Origin are Multi valued discrete integers and finally car name is the string with each instances respectively.

2.2.2 Preparation

As part of the dataset preparation, which needs to be pre-processed before we are importing to the R-studio, this dataset was not available in direct csv format instead as. data file and which is to be processed in spreadsheet and data processing work and cleaning the csv file has done.

Splitting the dataset into 80 percentage for the train dataset and 20 percentage for the test data set and removing the variables which contain (strings)



2.3. MODELLING

2.3.1 Fitting Multiple Polynomial Regression of model 1:

In the polynomial regression of the first model-1 I have taken the best fitted model from the linear regression mode and calculated with the degree 2 , where I get the **Multiple R-squared** value is **0.8896** and the **Adjusted R-squared** value is **0.8759**

2.3.2 Fitting Multiple Polynomial Regression of model 2:

In this model 2, I have modified degree to 3 and calculated the Poly function and the results we are showing improvements **Multiple R-squared** value is **0.9324** and the **Adjusted R-squared value is 0.9088.** And I removed the Horsepower which has the high p value in it for the next model.

2.3.3 Fitting Multiple Polynomial Regression of model 3:

In this Model degree was modified to 4 and the results were showing drastic improvements **Multiple R-squared** value is **0.9664** and **the Adjusted R-squared** value is **0.9211**

2.3.4 Fitting Multiple Polynomial Regression of model 4:

Here the degree was modified to 5, but this model started to over fitting, which had the **Multiple R-squared** value is **1** and **Adjusted R-squared** is **Nan** , which is found to be suspicious and overfitted., so best option to avoid this model

2.3.5 Fitting Multiple Polynomial Regression of model 5:

Again, in order to reconfirm the overfitting based upon the degree I modified to 10 and results were same as the model 5 so which was obviously overfitted and which should be avoided

2.4 EVALUATION

In the evaluation phase we have predicted the models from 1 to 5 and calculated the Root Mean Square Error, (Actual-Predicted)

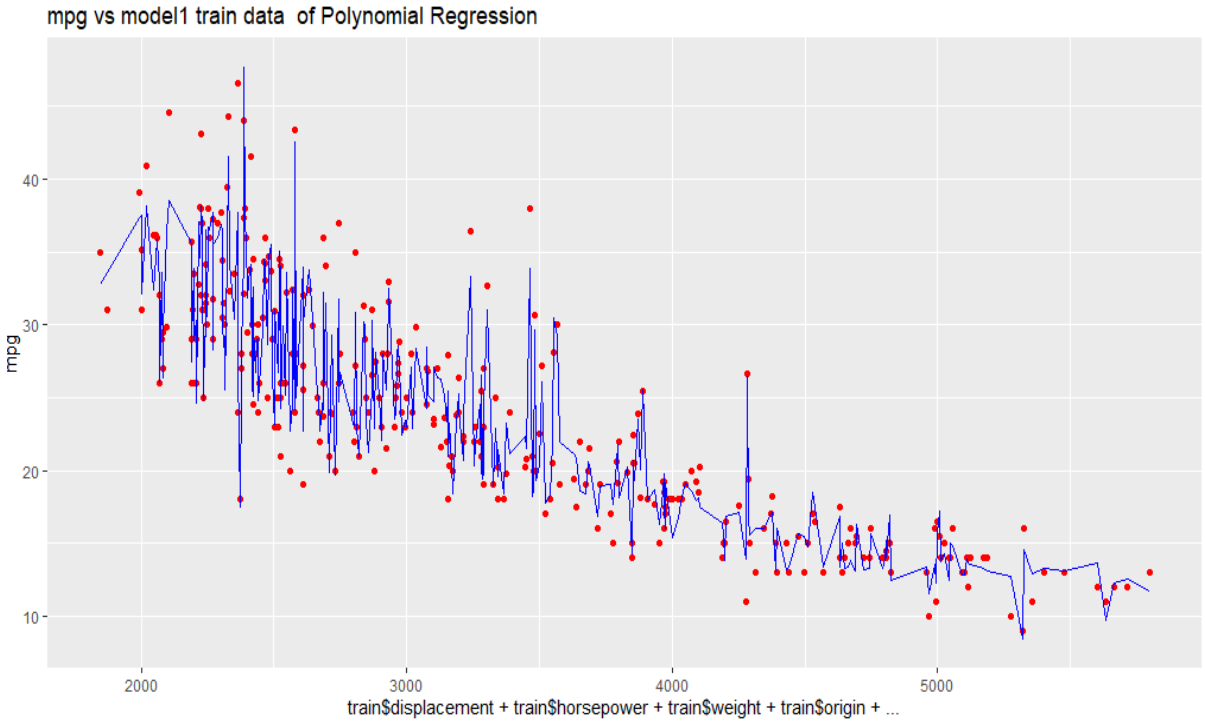
In the prediction stage of test dataset using RSME we found the models which were exact over fitted and found the best lowest value

```
# for model 1
sqrt(mean((test$mpg - prediction1)^2)) # 2.116572
# for model 2
sqrt(mean((test$mpg - prediction2)^2)) # 3.115849
# for model 3
sqrt(mean((test$mpg - prediction3)^2)) # 58.31147
# for model 4
sqrt(mean((test$mpg - prediction4)^2)) # 117849
# for model 5
sqrt(mean((test$mpg - prediction5)^2))  # 704280007
```

Model 3, surprisingly which had good R-Square value, shows Very High RSME levels and model 4,5 from the beginning they were very much suspicious on their R-Square values, with RSME we found they are not fit. Now comes between Model 1 and Model 2 though model 1 is less compared to model 2- in terms of R-Squared values But still, **I conclude that Model 1 has low RSME Values and which is much more exact for prediction and fitted one.**

2.4.1 Plot:

Training of Best Model: Model-1



Testing of Best Model: Model-1