# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 3**

Dt : 23-02-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

Astha Patel - AU1940312 - Btech Computer Science and Engineering.

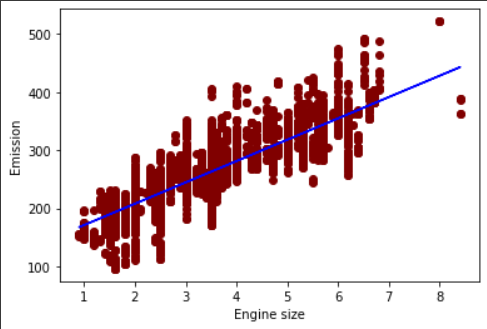
Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

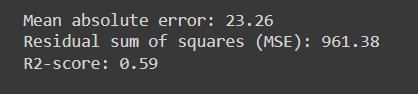
**Task Performed this week**

**Trials on Linear Regression**

**Single Variable Linear Regression**

Considering the dataset , we have different variables to be considered as parameters for the thorough analysis , but here we consider Engine Size (as by correlation we found earlier that engine size and CO2 emission values are positively correlated) as one of the independent variables and we build upon that so we get best fitting lines among the data points for the better determination of CO2 emissions.





**Multiple Variable Linear Regression**

For the determination of CO2 emissions we consider various other variables from the clean dataset which may influence the CO2 emissions in a linear manner which are Number of cylinders , Engine Size and Fuel Consumption etcetera.

The linear regression for the CO2 prediction is done using necessary libraries from sklearn

**from sklearn.metrics import r2\_score, mean\_squared\_error,mean\_absolute\_error**

To evaluate the predicted values using R2 score and Mean Square error

**from sklearn.model\_selection import train\_test\_split**

To split the dataset into train and test data set

**from sklearn.preprocessing import MinMaxScaler**

To scale the values of the variables between 0 to 1 : done as a part of data preprocessing. Due to scaling MSE, R2, coefficient and intercept values are also in range 0 to 1 which is easier to analyze.

After scaling the data, the data is splitted into train set and test set using train\_test\_split() function, where test size is approximately 1/3rd of the total dataset.

**from sklearn.linear\_model import LinearRegression**

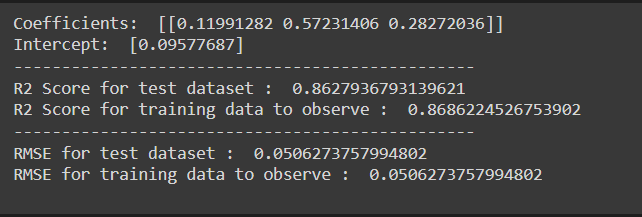
To find a best fit linear regression linear for the training data using LinearRegression().fit(X\_train, Y\_train)

Predict for test set using y\_pred = linearregression.predict(X\_test)

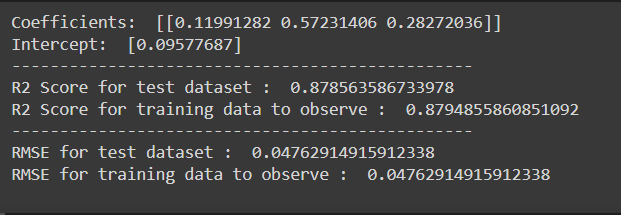
**from sklearn.metrics import r2\_score, mean\_squared\_error**

Measure the error between predicted value and true value using r2\_score and mean\_squared\_error.

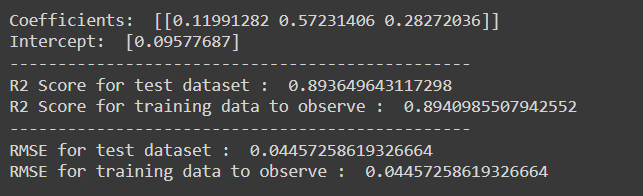
Linear regression for features : Engine Size(L) , Fuel Consumption Hwy (L/100 km), Cylinders to the output CO2 Emissions(g/km)



Linear regression for features : Engine Size(L) , Fuel Consumption Comb (L/100 km), Cylinders to the output CO2 Emissions(g/km)



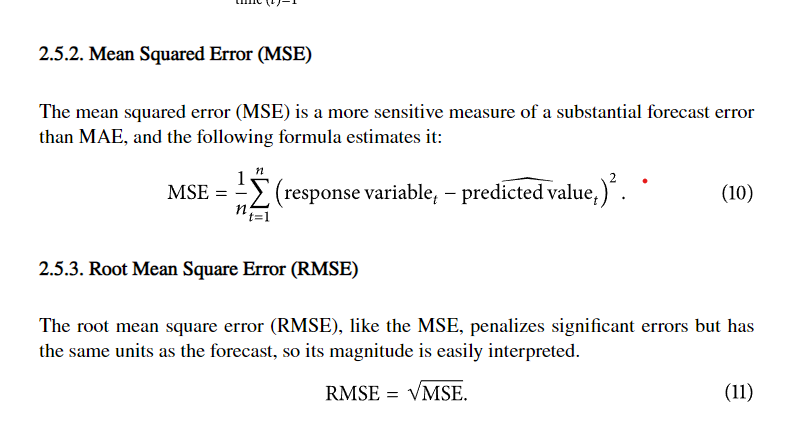
Linear regression for features : Engine Size(L) , Fuel Consumption Comb (mpg), Cylinders to the output CO2 Emissions(g/km).



**Error Measures :**

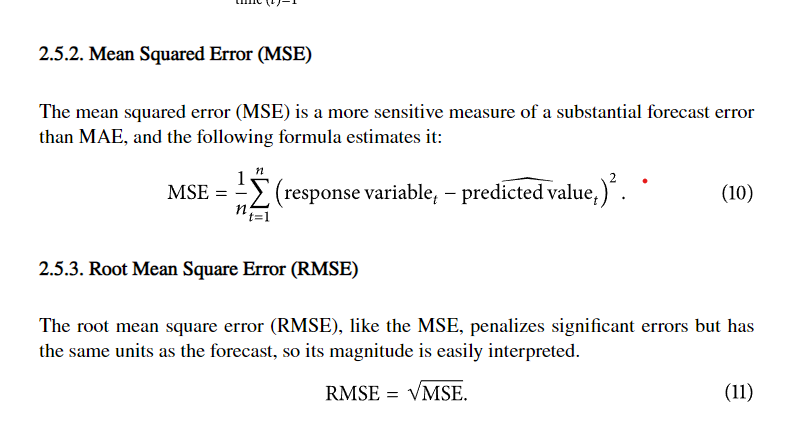
**Mean Squared Error (MSE)**

The Mean Squared Error plays a substantial role for prediction in errors , most prominently it focuses on large errors as the formula states that there could be exponential growth to the errors, especially large errors due to the same.

****

**Root Mean Squared Error (RMSE)**

This term is the square root of Mean Square Error.

****

**R Squared Error**

It is one of the standards to keep a check on the accuracy of the model. The R squared error would justify how data is to be fitted as a regression line onto the data model.

The possible scores for the same can be positive or negative as well. Best score that can be attained by any data model would be 1.0

If the R-squared values are higher than this means the model fits well our dataset.

**Tasks to be Performed Next Week**

1. Check for data outliers using box plot, scatter plot and Interquartile Range.
2. To evaluate and understand linear regression and its efficiency for the data set.
3. If not efficient enough, try to increase efficiency by data preprocessing or other required steps.