

**RAMAIAH INSTITUTE OF TECHNOLOGY**

**Department of Information Science and Engineering**

**(SEP- DEC 2022)**

**Deep learning - IS741**

**(Assignment 2)**

**Predicting the Mileage/Miles Per Gallon (mpg) of a car**

**Submitted by**

**1MS19IS099- Ravi tejas**

**Predicting the Mileage/Miles Per Gallon (mpg) of a car using some features of the car.**

This repository contains the files of my data science and first machine learning project.The files presented above are: two csv data files that contain some technical specs of some cars,a file with the extension '.pkl' which is the final model with the best prediction accuracy,an a jupyter notebook with the '.ipynb' file extension which contains the python code for the whole project, and the images of the charts presented in this project.

**About the dataset**

The dataset is downloaded from UCI Machine Learning Repository.The link to the data is <https://archive.ics.uci.edu/ml/datasets/auto+mpg>

**Content**

* Title: Auto-Mpg Data
* Source: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.The dataset was used in the 1983 American Statistical Association Exposition
* Relevant information: The data concerns city-fuel consumption in miles per gallon,to be predicted in terms of 3 multivariate discrete and 5 continuous attributes.The number of instances are 398 and the number of attributes are 9 including the class attribute.The attribute information are:
  1. mpg - Milieage/Miles Per Galon
  2. cylinders - the power unit of the car where gasoline is turned into power
  3. displacement - engine displacement of the car
  4. horsepower - rate of the engine performance
  5. weight - the weight of a car
  6. acceleration - the acceleration of a car
  7. model - model of the car
  8. origin - the origin of the car
  9. car - the name of the car

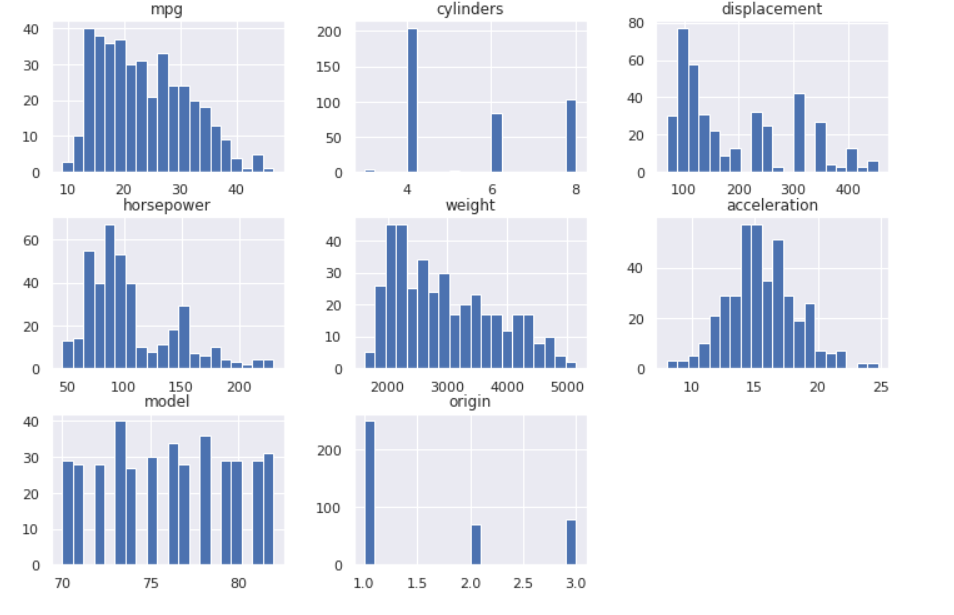
**Insights from the dataset**

Before applying a machine learning algorithm on the data to solve the problem,an Exploratory Data Analysis was made to know more about the data and below are some insights we gained from performing the analysis.

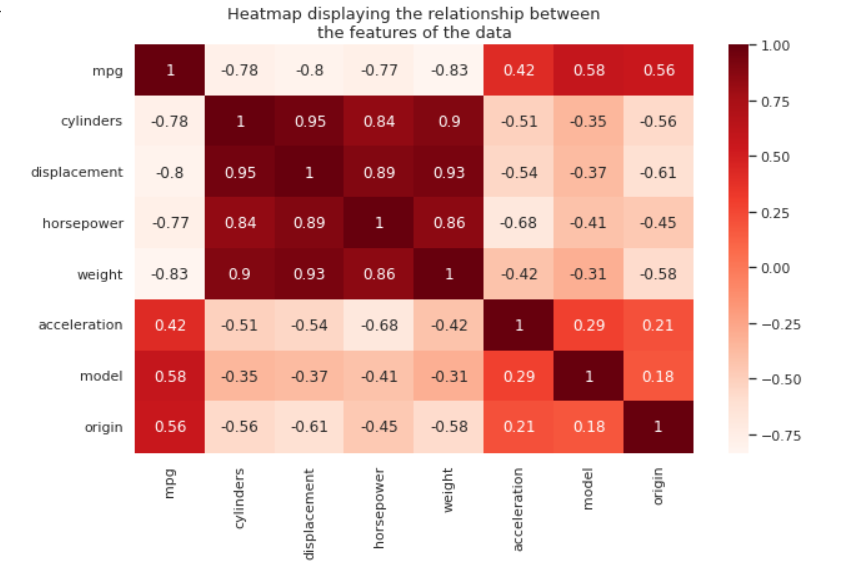
**Distributions of the variables/features.**

* The acceleration of the cars in the data is normally distributed and the most of the cars have an acceleration of 15 meters per second squared.
* Half of the total number of cars (51.3%) in the data have 4 cylinders.
* Our output/dependent variable (mpg) is slightly skewed to the right.

**Relationship between the Miles Per Gallon (mpg) and the other features.**

****

* We can see that there is a relationship between the mpg variable and the other variables and this satisfies the first assumption of Linear regression.



* There is a strong negative correlation between the displacement,horsepower,weight,and cylinders.This implies that,as any one of those variables increases,the mpg decreases.
* The displacement,horsepower,weight,and cylinders have a strong positive correlations between themselves and this voilates the non-multicollinearity assumption of Linear regression.Multicollinearity hinders the performance and accuracy of our regression model.To avoid this, we have to get rid of some of these variables by doing feature selection.
* The other variables.ie.acceleration,model and origin are not highly correlated with each other.

**Training a regression model for the prediction.**

After making a feature selection,the variables I used in the prediction are the acceleration,model,origin,and horsepower.The final regression model that appeared to have the highest accuracy score is the GradientBoostingRegressor model with the following scores:

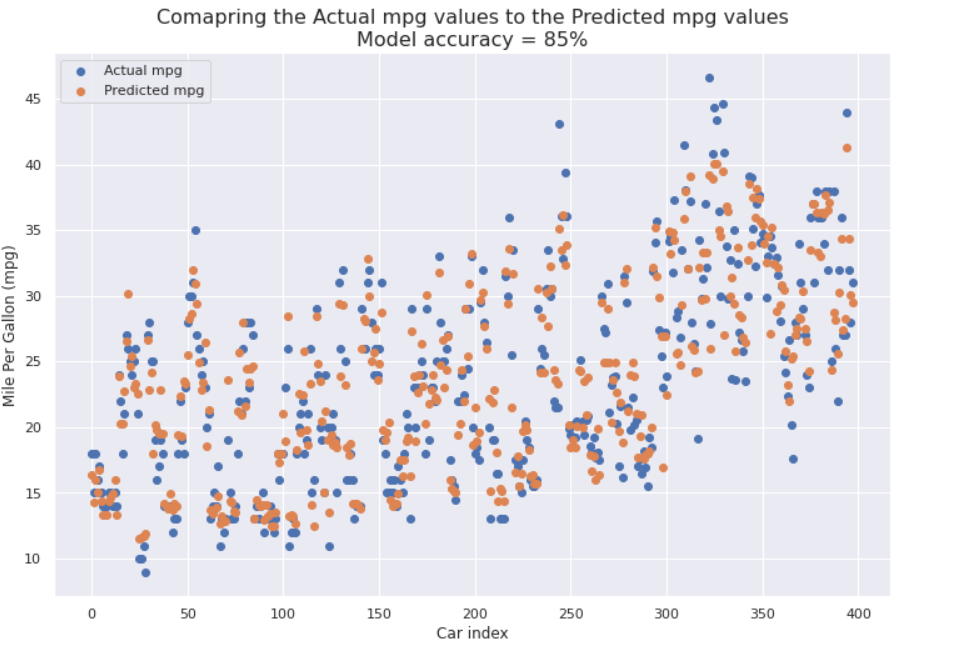
Train score: 0.9040700029438588

Test score: 0.8379811577934175

Overall model accuracy: 0.8379811577934175

Mean Squared Error: 8.888595699098659

I therefore made predictions using the predictor varibles to see how well the model predicts and visualized the actual mpg values recorded and the mpg values predicted by the model to see how close our predictions are to the actual values.



We can see from the above scatter plot that our model made a good predictions as the values of the actual mpg and the predicted mpg are very close to each other.We can confidently say that we have succeeded in training a model that predicts the Mileage Per Gallon (mpg) of a car given the acceleration,model,origin and the horsepower of a car.