CAR PRICE PREDICTION

Project report submitted to the Amrita Vishwa Vidyapeetham in partial fulfilment of the requirement for the Degree of

B.Tech. Computer Science and Engineering
Artificial Intelligence

Submitted by:

Abhinav Pandey (AM.EN.U4AIE21088)
Abhishek. A. (AM.EN.U4AIE21003)
Anfas Hassan V (AM.EN.U4AIE21014)
Aravind MJ (AM.EN.U4AIE21017)
Rithesh R (AM.EN.U4AIE21054)
D.S.S. Sandeep Chandra (AM.EN.U4AIE21079)

July 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING AMRITA VISHWA VIDYAPEETHAM
(Estd. U/S 3 of the UGC Act 1956)
Amritapuri Campus Kollam -690525

BONAFIDE CERTIFICATE

Your Guides

Project Guide

Coordinator name

Project Coordinator

Reviewer
Chairperson
Dept. of Computer Science & Engineering
Place: Amritapuri
Date: 11 July 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AMRITA VISHWA VIDYAPEETHAM

(Estd. U/S 3 of the UGC Act 1956)

Amritapuri Campus Kollam -690525

DECLARATION

We,**Abhinav Pandey, Abishek A., Anfas Hassan, Aravind MJ, Rithesh R, D.S.S. Sandeep Chandra** hereby declare that this project entitled **CAR PRICE PREDICTION** is a record of the original work done by us under the guidance of **DR GEORG GUTJAHR** and **DR GOPAKUMAR G** , Dept. of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, that this work has not formed the basis for any degree/diploma/associations/fellowship or similar awards to any candidate in any university to the best of our knowledge.

Place: Amritapuri Date: 12 July 2022

Signature of the student

Signature of the Project Guide

ACKNOWLEDGEMENTS

Humble pranams at the lotus feet of Amma, Sri Mata Amritanandamayi devi

We thank Dr Georg Gutjahr, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, for her timely advice and constructive suggestions. We thank the faculty of the CSE Department for giving us an opportunity to explore our capabilities and widen our knowledge. We also thank our team members for their cooperation and hard work put into this project. We also thank our batch mates for their support and our parents for their support and encouragement.

INTRODUCTION

An automobile company is entering the market and it is planning to set up a manufacturing unit here. It also plans to produce cars locally to give competition to existing companies. They have planned to do a contract with an automobile consulting company to understand the factors on which the pricing of the cars depends. Specifically, they want to understand the main features about a car that has more effect on its pricing.

The company wants to know:

- Which variables are significant in predicting the price of a car
- How well those variables describe the price of a car

In our project we have planned to solve the same problem. We want to model the price of cars with the available independent variables. We have split every model into a training and testing set. We have used independent variables in the training set to predict the price and correspondingly check the accuracy of the models with the actual price in the testing set. We have made 12 different models and fit single as well as multi variables in that model according to analysis. We have used many python libraries such as sklearn, matplot, seaborn, etc.

This provides a solution for the company to manage and understand the variation of prices with the variables. They can accordingly manipulate the design of the cars, the business strategy etc to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

DATA SET

Our dataset consists of 205 rows and 26 columns. The rows include the no of data per car. The column includes independent variables or features of the car. The variables are:

- Car id:
  The id of the car given
  Data type: int
  Examples: 205, 103

- Symboling:
  The degree to which the auto is more risky than its price indicates.
  Data type: int
  Examples: 3,2,1

- Car Name :
  The name of the car
  Data type: object
  Examples: alfa romeo stelvio, audi 100 ls

- Fuel type :
  The type of fuel used in the engine
  Data type: object
  Examples: gas, diesel

- Car body :
  Its the vehicle frame of the car
  Data type: object
  Examples: convertible,sedan

- Drive wheel :
  Wheel and tire assembly that pushes or pulls a vehicle down the road.
  Data type: object
  Examples: fwd,rwd,awd

- Engine location :
  The location of engine
  Data type: object
  Examples: front,rear

- Wheel base :
  The horizontal distance between the centers of the front and rear wheels.
  Data type: float
  Examples: 172.2, 176.6

- Car length , width and height:

The dimensions of the car
Data type: float
Examples: 168.8, 64.1, 48.4

- Curb weight :
The weight of the vehicle including a full tank of fuel and all standard equipment
Data type: int
Examples: 2548, 2823

- Engine type :
The type of the engine
Data type: int
Examples: dohc,ohcv

- Cylinder number :
The number of cylinders in the car
Data type: object
Examples: four,eight

- Engine size :
The size of the engine
Data type: int
Examples: 90,98

- Fuel system :
The system used in the car
Data type: object
Examples: mpfi,2bbl

- Bore ratio :
The ratio of bore to stroke
Data type: float
Examples: 3.47,2.68

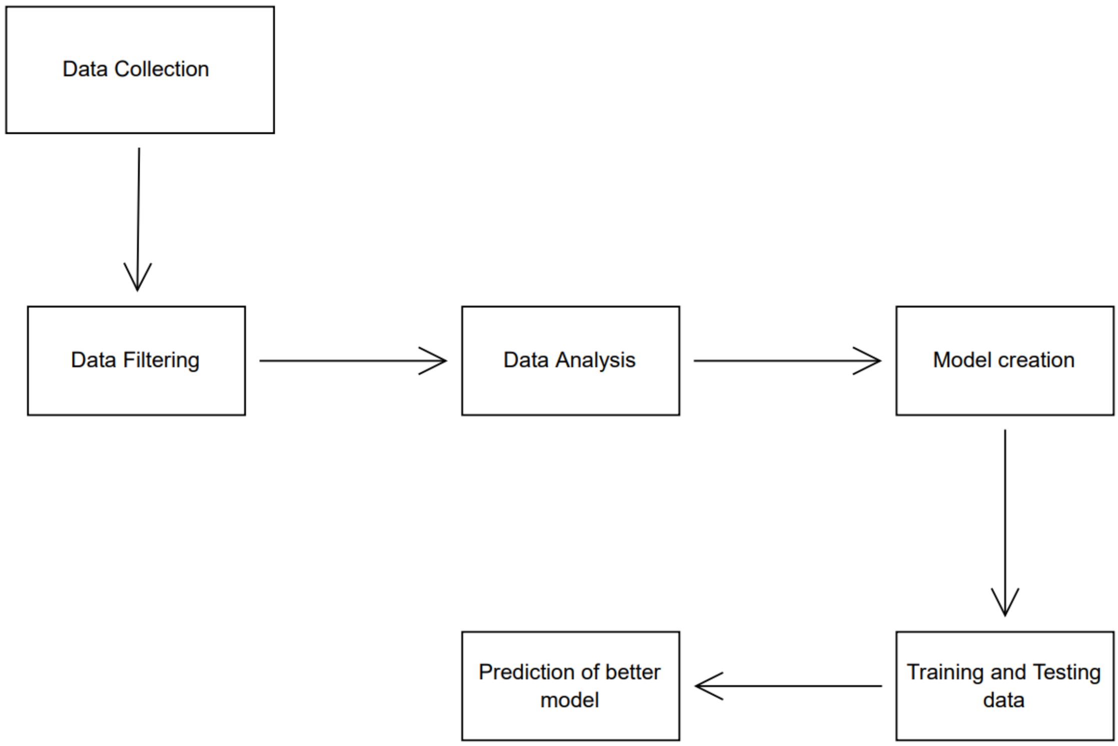
- Stroke :
A phase of the engine's cycle , during which the piston travels from top to bottom or vice versa.
Data type: float
Examples: 3.47,2.68

- Horsepower :
Power produced in a car
Data type: int
Examples: 154,115

- Price :
The price of the specific car
Data type: float Examples: 13950 , 17450

METHODOLOGY

Approach for car price prediction proposed in this paper is composed of several steps as shown



Data is collected from a dataset from Kaggle. It consists of 205 rows each of the car and its information. It includes columns such as fueltype, horsepower, cylinder number etc. But this data was not ready to be used. We have to filter and process the data before we put regression on it. We first understood the structure and features of the data.

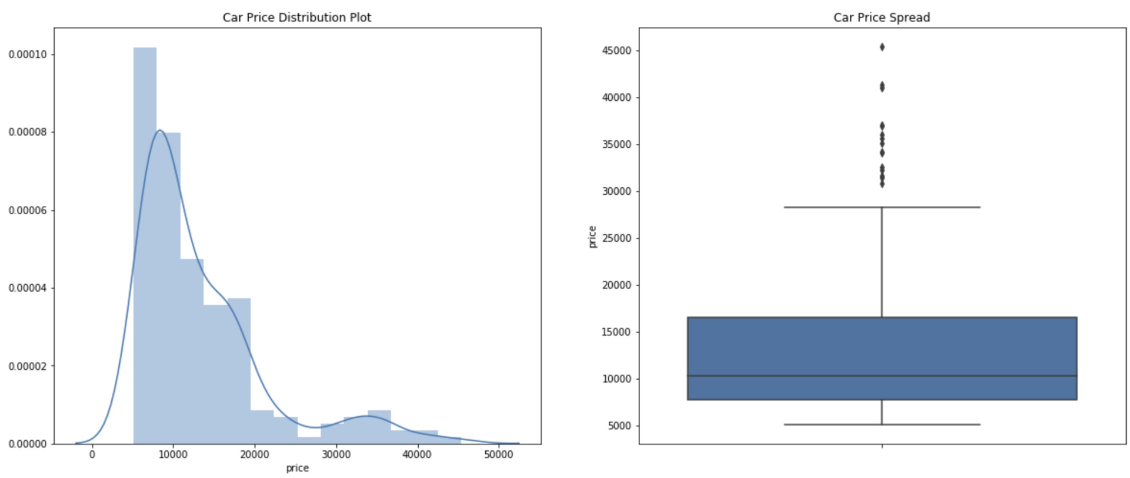
Firstly, the company names of the cars were very vast. So, we have filtered the names of cars with the actual names of the company. For instance, alfa-romeo giulia and alfa-romeo stelvio are cars from same company but different name. We have only considered the company name like alfa romeo.

Second task was to remove any errors. The company names were misspelt which made them different data. So, we corrected some names as show:

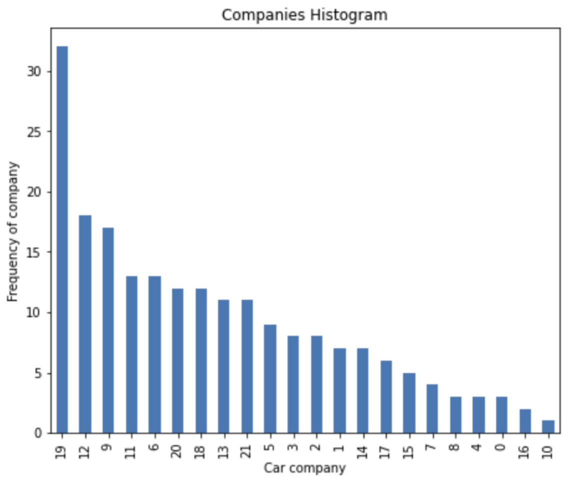
- maxda = mazda
- Nissan = nissan
- porsche = porcshce
- toyota = toyouta
- vokswagen = volkswagen = vw

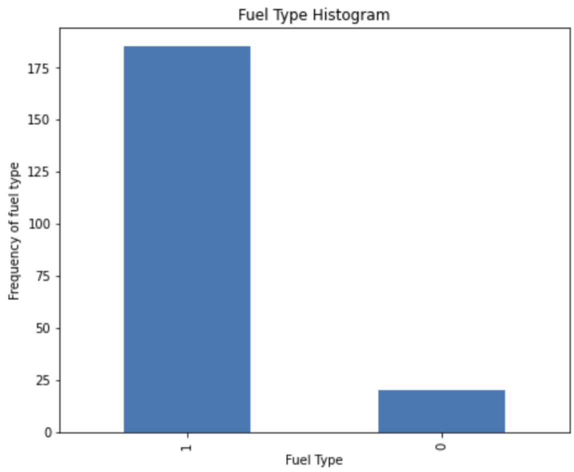
Third task was to remove the duplicates from the dataset if there were any. Fourth and main task was to encode our data. We have a lot of strings in our data, but our regression model cannot take strings as an input. So, we used labelEncoder to transform string into int value according to the set. Now our data is filtered and ready to be used.

We move on to data analysis and visualisation. We crossed checked every independent variable and how it varies in the data set with the price. We have visualised the spread of the variable with the price and accordingly selected the variable for the model.

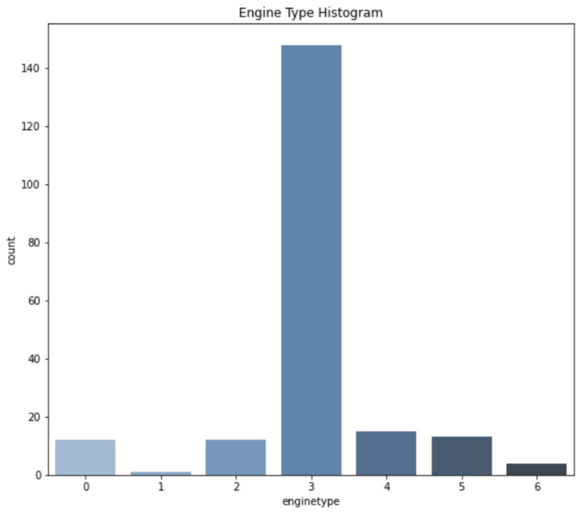


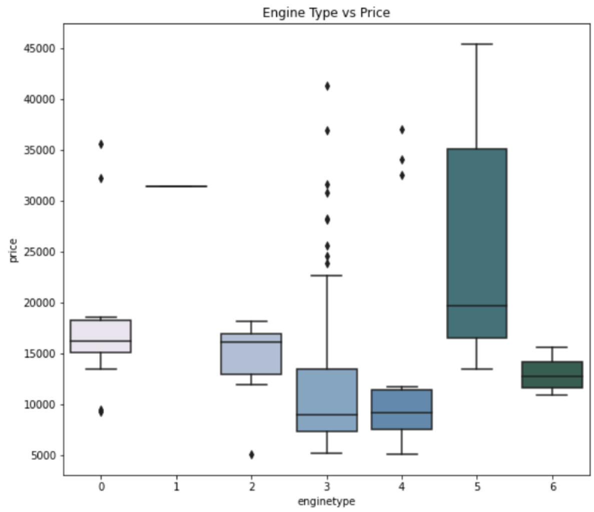
- The plot seemed to be right-skewed, meaning that the most prices in the dataset are low(Below 15,000).
- There is a significant difference between the mean and the median of the price distribution.
- The data points are far spread out from the mean, which indicates a high variance in the car prices.(85% of the prices are below 18,500, whereas the remaining 15% are between 18,500 and 45,400.



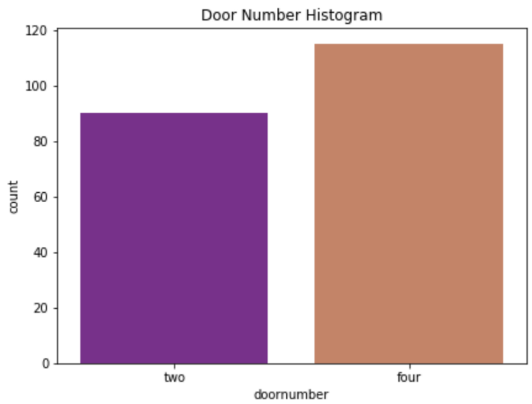


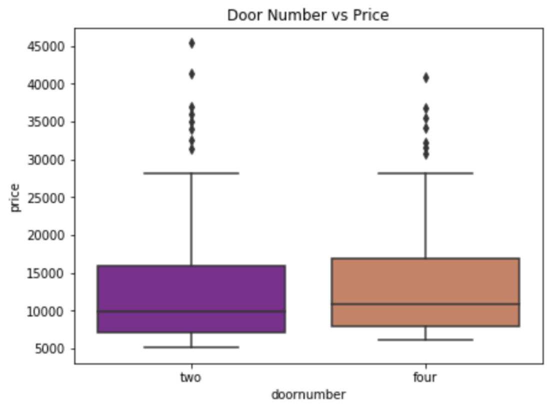
In the histogram bar graph, we can see that there is much variation in the company name but the fuel type histogram is not useful as there are only two types petrol and diesel.

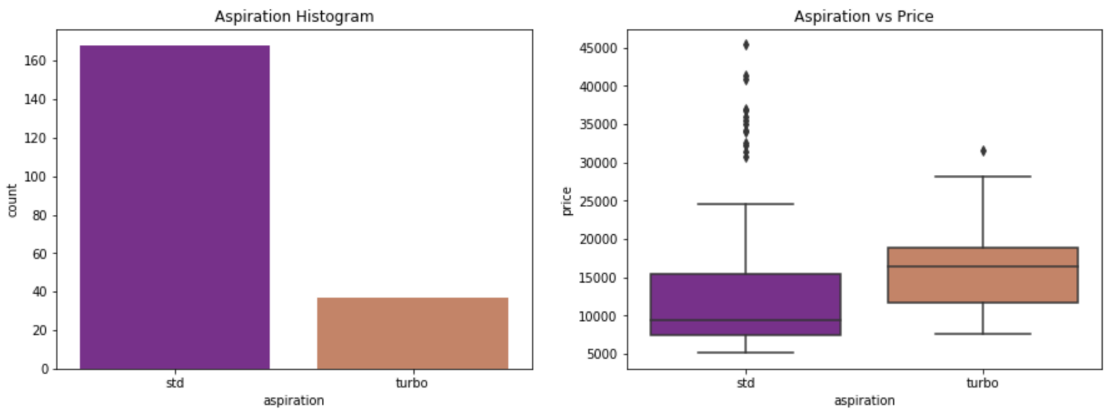




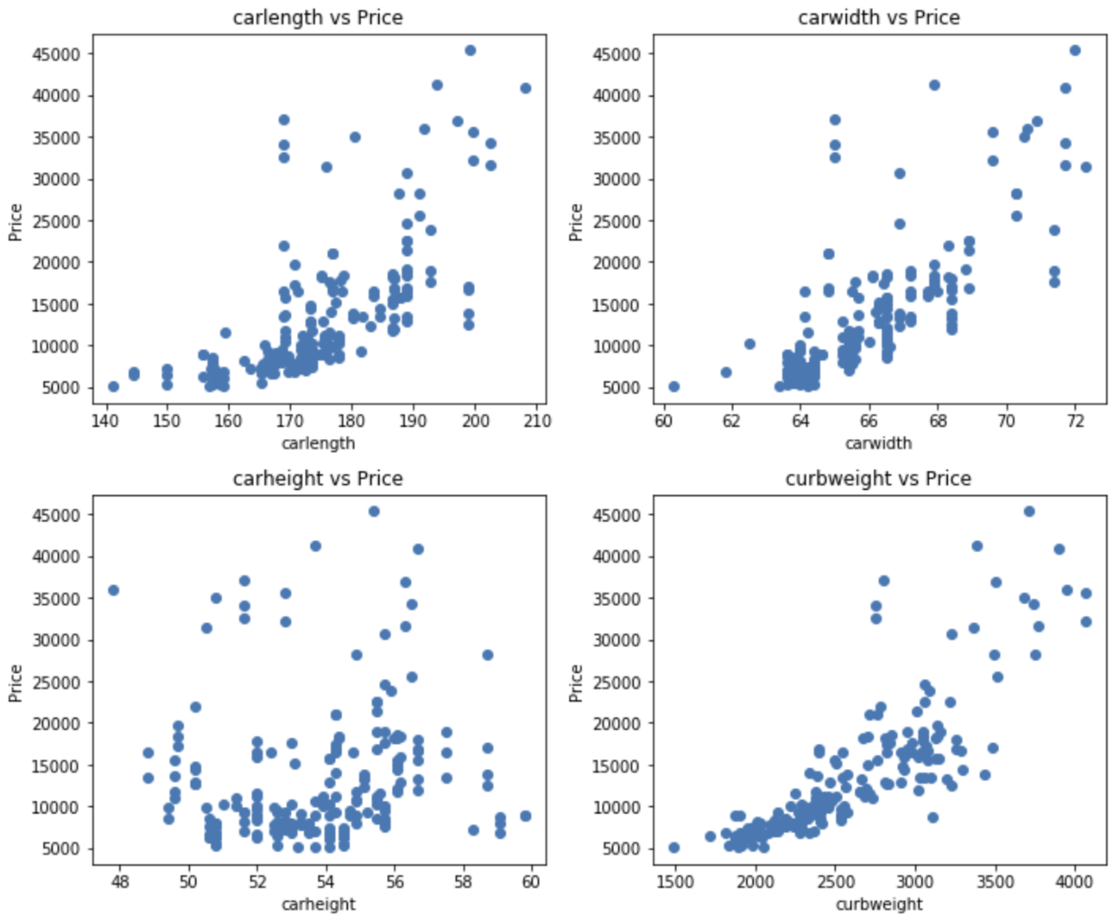
Here we can analyse how much the different types of engines are varying the prices. Third engine type is the most used of all the data. But the fifth is the one which has highest price range and others have low price range.







Door-number variable is not affecting the price much. There is no significant difference between the categories in it. It seems aspiration with turbo has a higher price range than the std(though it has some high values outside the whiskers)



carwidth, carlength and curbweight seems to have a positive correlation with price. carheight doesn't show any significant trend with price.

Now that we've analysed the data, we have to fit the data in the models. We've used 12 models. 6 of them are of single variables like highwaympg, citympg, horsepower, enginesize, wheelbase and boreratio. Some showed less accuracy, so we are going to further use the ones which show more accuracy and less MSE. Then there are 5 models which are made to test combined variables.

We have used the sklearn python library to split the testing and training set. There is a separate table for actual price for testing of the models. We first define the model, put our variable or combined variables in it and then determine the size of the testing set. We have also imported LinearRegression from sklearn and used that to predict the prices from the model. Then we test it in the testing set. We calculate the mean squared error of the predicted and actual prices of cars. We also determine the accuracy of the model.

Example:

Model(Prediction using enginesize):

```
The predicted values by this model is:
[11992.72061451 43483.43893767  5292.56777979 12327.72825624
 10317.68240583 10150.17858496  6967.60598847  8140.13273455
 26900.56067175 11992.72061451  8140.13273455]
MSE: 6144.821980004657
Accuracy: 73.9279943295746
```

RESULTS

- Prediction using highway mpg

```
The predicted values by this model is:
[10638.16701235 24530.06133791 19127.65798908 12181.7108263
 13725.25464025 12953.48273328  7551.07938445  7551.07938445
 19899.42989605 18355.8860821   8322.85129143]
MSE: 8670.001912036572
Accuracy: 48.09679008820164
```

- Prediction using citympg

```
The predicted values by this model is:
[11734.92618908 22175.03540292 20568.86475464 13341.09683736
 14947.26748565 13341.09683736  8522.58489252  6916.41424423
 20568.86475464 18159.60878221  8522.58489252]
MSE: 9208.007813681877
Accuracy: 41.45535335870263
```

- Prediction using horsepower

```
The predicted values by this model is:
[12014.45095302 25636.4747883  17964.30044429 10605.27607351
 14989.37569866 10605.27607351  7473.77634126  7473.77634126
 25323.32481508 11701.30097979  7630.35132787]
MSE: 7443.064297724783
Accuracy: 61.74753978061042
```

- Prediction using enginesize

```
The predicted values by this model is:
[11992.72061451 43483.43893767  5292.56777979 12327.72825624
 10317.68240583 10150.17858496  6967.60598847  8140.13273455
 26900.56067175 11992.72061451  8140.13273455]
MSE: 6144.821980004657
Accuracy: 73.9279943295746
```

- Prediction using wheelbase

```
The predicted values by this model is:
[12056.54837176 29342.8676904  10670.72530402 11400.10586599
 11400.10586599 14390.56617006  9503.71640487 14390.56617006
 16651.64591216 24456.01792522 10087.22085445]
MSE: 8913.165800009598
Accuracy: 45.14454758787443
```

- Prediction using boreratio

```
The predicted values by this model is:
[13076.97444037 20255.34452539 13076.97444037 13382.43699718
 10633.2739859  10938.7365427   7578.6484178   8189.57353142
 17506.18151411 15062.48105963 10327.81142909]
MSE: 9998.047760401667
Accuracy: 30.97820573982223
```

- Prediction using enginesize and boreratio

```
The predicted values by this model is:
[12037.54811929 43048.14336905  5621.89204293 12394.7995057
 10141.88543948 10017.96262016  6569.3715754   7765.04855394
 26841.17733675 12274.59390612  8020.32863207]
MSE: 6116.264033693828
Accuracy: 74.16976952480022
```

- Prediction using enginesize and horsepower

```
The predicted values by this model is:
[10940.48707128 28014.31014816 17512.48324494  9878.30617788
 14065.47008713 10545.36720449  6396.76221687  7516.19750877
 24847.63091274 12849.60780198  7205.69742733]
MSE: 6926.390547057232
Accuracy: 66.87394014272734
```

- Prediction using stroke, horsepower, enginesize

```
The predicted values by this model is:
[11694.04652353 40922.54114056  9810.09573196 11112.81598546
 11744.04526778 10730.92000348  6114.95242019  8261.82488638
 27288.74299418 15269.64348334  7040.6393813 ]
MSE: 5153.279421675414
Accuracy: 81.66321412175945
```

- Prediction using boreratio, horsepower and enginesize

```
The predicted values by this model is:
[11475.42816292 40576.54498161  8570.05866654 11306.6182886
 11138.33850558  9726.93198887  6579.68509621  7163.33214137
 27483.40328807 13857.24462064  7426.49898697]
MSE: 5432.870687067083
Accuracy: 79.61951255305274
```

- Prediction using boreratio, horsepower, enginesize and stroke

```
The predicted values by this model is:
[11664.23992307 41050.9784228   9810.89784196 11043.75993545
 11851.6540931  10812.01573246  6241.52262586  8429.88837722
 27317.63201988 15311.14079607  7050.03057583]
MSE: 5122.63795465136
Accuracy: 81.88062736199548
```

CONCLUSION

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In this research, PHP scripts were built to normalise, standardise and clean data to avoid unnecessary noise for linear regression.Then after analysing the fitting the data in the models, we used it in the testing and training sets and analysed which performed better.

REFERENCES

Dataset: https://www.kaggle.com/code/goyalshalini93/car-price-prediction-linear-regression-rfe

Python Libraries:

https://scikit-learn.org/stable/

https://numpy.org

https://numpy.org

https://seaborn.pydata.org

Colab link:
https://colab.research.google.com/drive/1INp_tTRYZTZeJJY1HKyBumMJnmi2AVl9?usp=sharing