CAR PRICE PREDICTION PROJECT

Submitted by:

MIENGANDHA SINHA

ACKNOWLEDGEMENT

I would like to express my gratitude to the Company to give this project to me. In making of this project I hereby used to take help from the references which is given by the company as sample documentation and details related to project and professionals and SME guided me a lot in the project and the other previous projects helped me and guided me in completion of the project.

INTRODUCTION

- Car Price Prediction Problem

This project aims to predict the price of an used Car by taking it's Company name, it's variant, price details, how it work whether it work manual or by automatic and many other parameters it can be predicted. So, here in this it required to the making of model of the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

- Conceptual Background of the Domain Problem

A primary objective of this project is to estimate used car prices by using some attributes that are highly correlated with a Price. As it is said in the statement that with the impact of covid19, some of the people are facing problems with their previous car price valuation through machine learning models. So they are looking for new machine learning models from new data to help in making car price valuation model. By doing some research on this project, are able to trained the models and predicting things make the previous background to work efficiently.

- ## Review of Project

As it is said, some of the people who works with small traders, who sell used cars. With the change in market due to the impact placed, are facing such problems with their previous car price valuation machine models. This is all because of some cars are in demand hence making them costly and some are not in demand hence they are cheaper and it creates a lot of changes in the car market. From the collecting data phase to the model building phase gives important variables in the used car model. There all types of cars in the datasets gives various information. And then the model building do all the data visualizations, data pre-processing steps, evaluating the model, data cleaning and selecting the best model for the project.

- Motivation for the Problem Undertaken

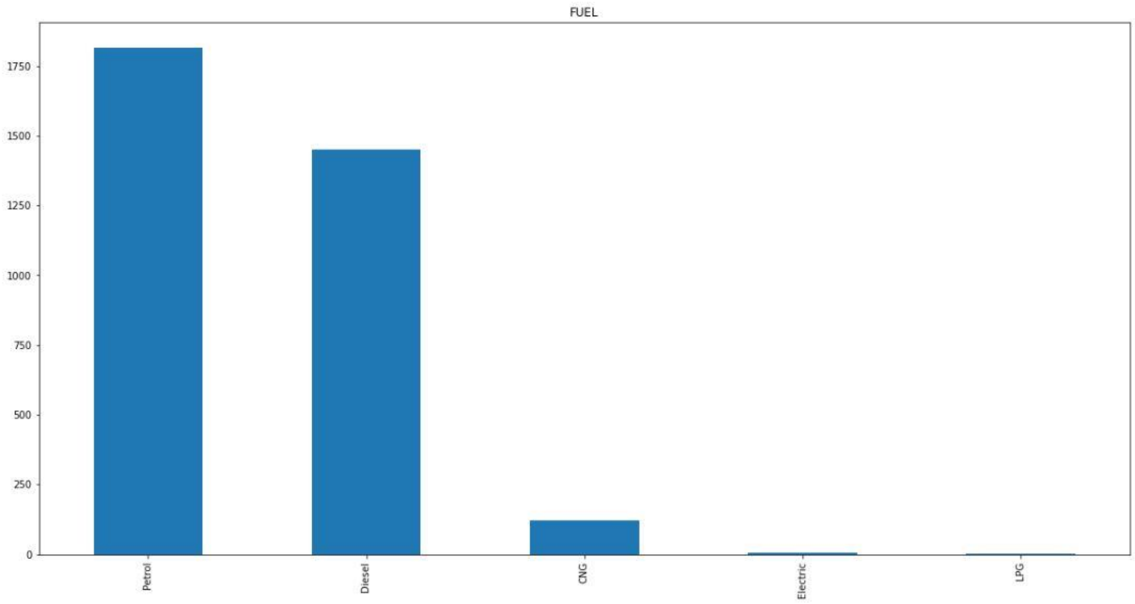
Here, the datasets have the total of 3394 entries with 8 rows, no null values, EDA has to be performed to see whether it gain or loss in the variable and its compare to the price among every aspects, to build machine learning models, to determine the optimal values of Hyper parameters and the selection of the best model, by predicting of the value can help to the clients and for the further change in the market from the new data.

ANALYTICAL PROBLEM FRAMING

- Mathematical/Analytical Modelling of the Problem

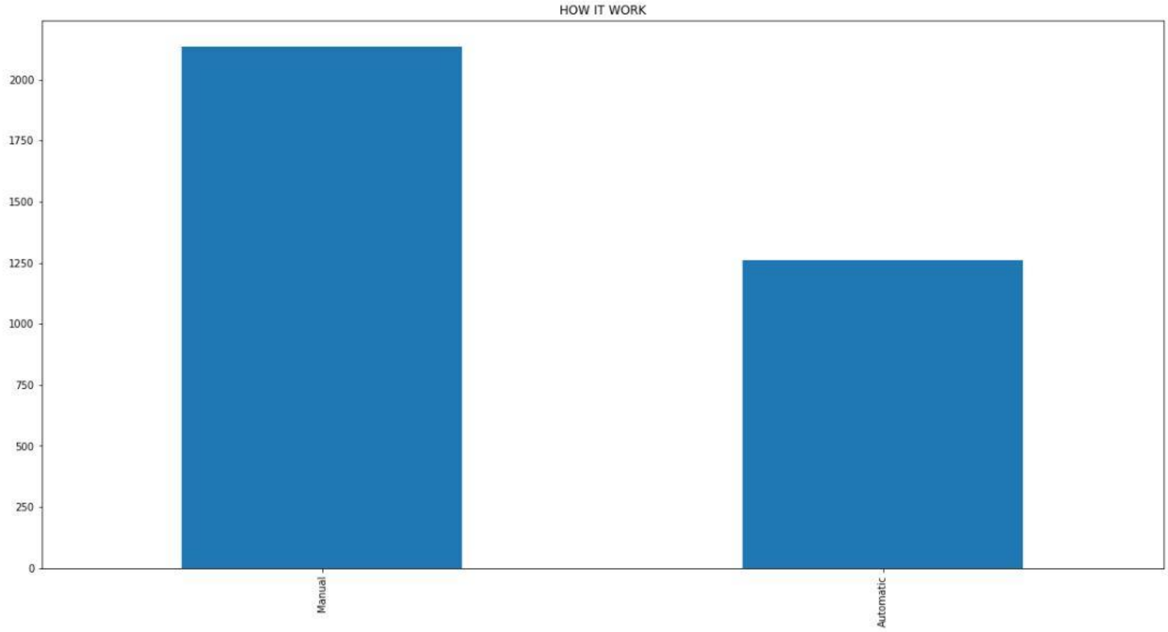
In this project, mathematical/analytical modelling are used. Checking the null values found that having no null values in the datasets, the type of data frame is in pandas, data frame info tells that int64(1 variable), object(7 variables), by using the data visualization they are:

1. It is about the fuel used by the used car:
   Petrol has the highest accuracy among other fuel



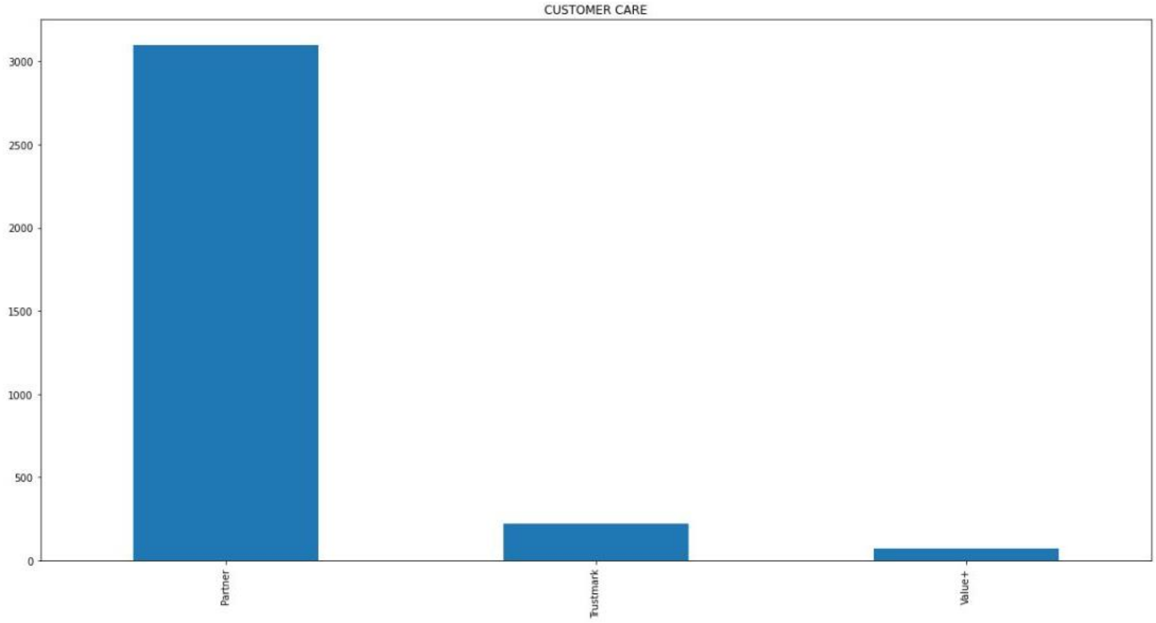
2. It is about how the used car work whether it work manually or automatically:

Manual work is on the high as compared to automatic.

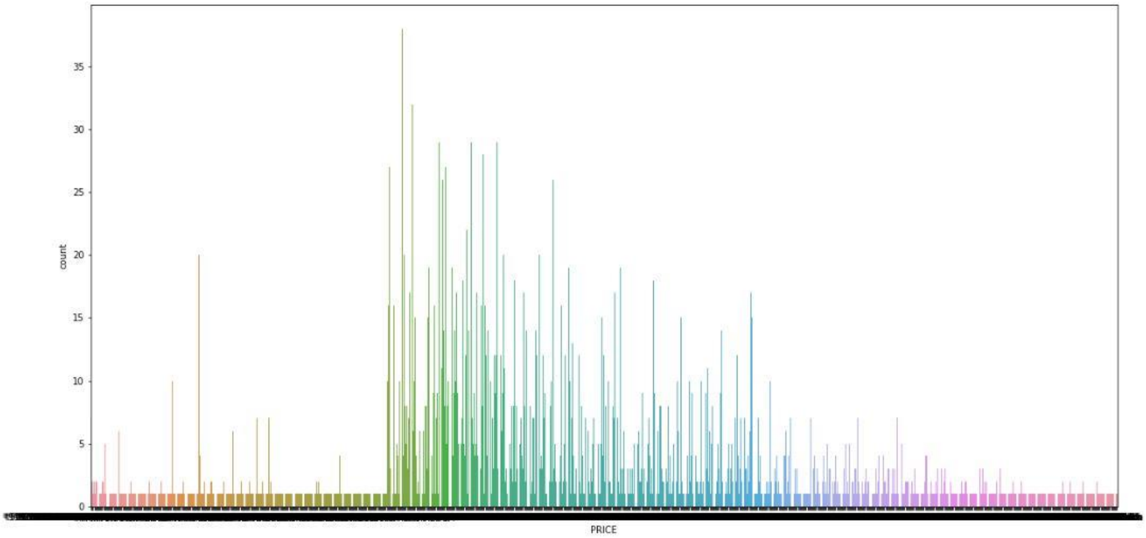


3. It is about the assurance of the used car means it is for customer care purpose:

Partner has the highest accuracy

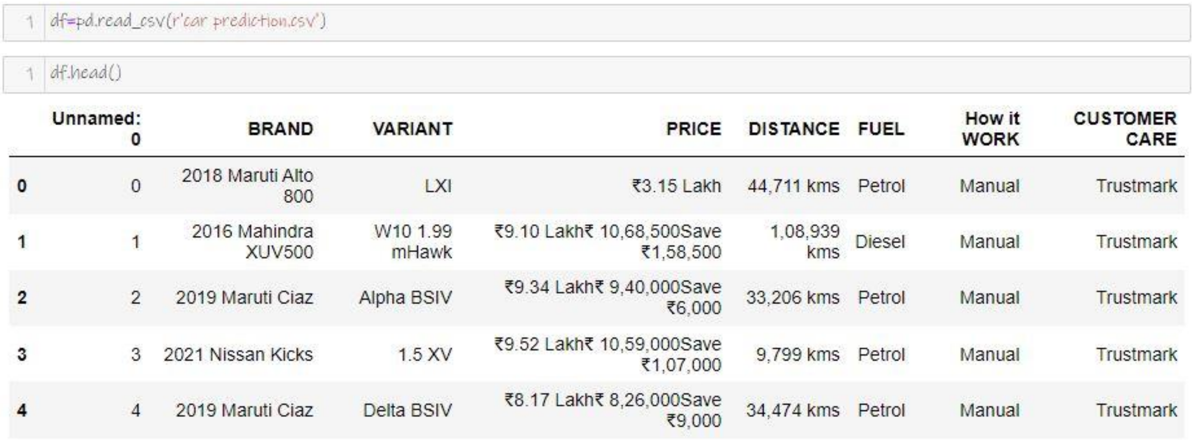


4. The count plot of the price of used car:



- ## Data Sources and their formats

The data sources and their formats are from .csv file.



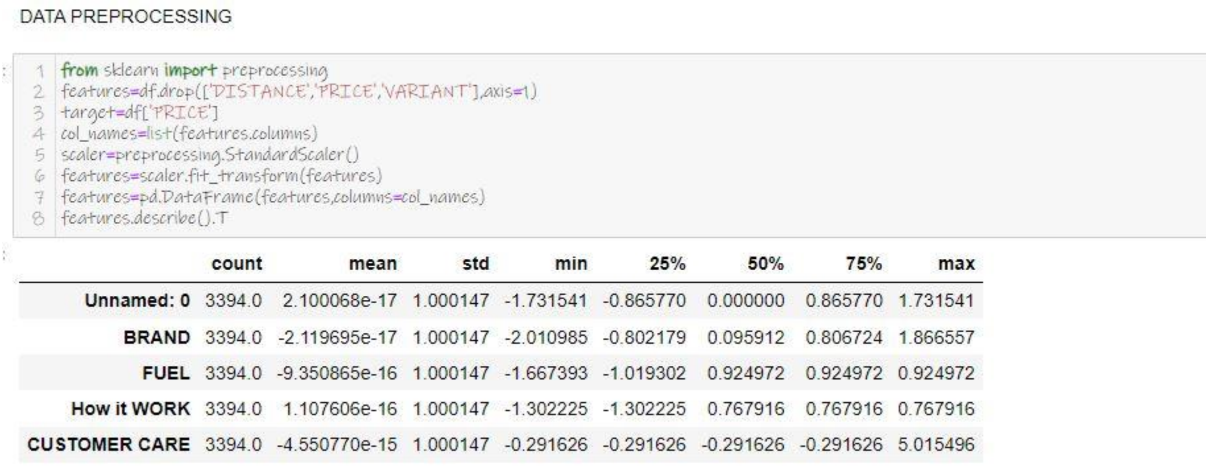
```
df=pd.read_csv(r'car prediction.csv')
```

```
df.head()
```

	Unnamed: 0	BRAND	VARIANT	PRICE	DISTANCE	FUEL	How it WORK	CUSTOMER CARE
0	0	2018 Maruti Alto 800	LXI	₹3.15 Lakh	44,711 kms	Petrol	Manual	Trustmark
1	1	2016 Mahindra XUV500	W10 1.99 mHawk	₹9.10 Lakh₹ 10,68,500Save ₹1,58,500	1,08,939 kms	Diesel	Manual	Trustmark
2	2	2019 Maruti Ciaz	Alpha BSIV	₹9.34 Lakh₹ 9,40,000Save ₹6,000	33,206 kms	Petrol	Manual	Trustmark
3	3	2021 Nissan Kicks	1.5 XV	₹9.52 Lakh₹ 10,59,000Save ₹1,07,000	9,799 kms	Petrol	Manual	Trustmark
4	4	2019 Maruti Ciaz	Delta BSIV	₹8.17 Lakh₹ 8,26,000Save ₹9,000	34,474 kms	Petrol	Manual	Trustmark

- ## Data Preprocessing Done

The steps followed for the cleaning of the data is Label Encoder after then importing preprocessing there transform the target columns into features then lastly it has to set/fir for the data frame.



DATA PREPROCESSING

```
from sklearn import preprocessing
features=df.drop(['DISTANCE','PRICE','VARIANT'],axis=1)
target=df['PRICE']
col_names=list(features.columns)
scaler=preprocessing.StandardScaler()
features=scaler.fit_transform(features)
features=pd.DataFrame(features,columns=col_names)
features.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	3394.0	2.100068e-17	1.000147	-1.731541	-0.865770	0.000000	0.865770	1.731541
BRAND	3394.0	-2.119695e-17	1.000147	-2.010985	-0.802179	0.095912	0.806724	1.866557
FUEL	3394.0	-9.350865e-16	1.000147	-1.667393	-1.019302	0.924972	0.924972	0.924972
How it WORK	3394.0	1.107606e-16	1.000147	-1.302225	-1.302225	0.767916	0.767916	0.767916
CUSTOMER CARE	3394.0	-4.550770e-15	1.000147	-0.291626	-0.291626	-0.291626	-0.291626	5.015496

- Data Inputs- Logic- Output Relationships

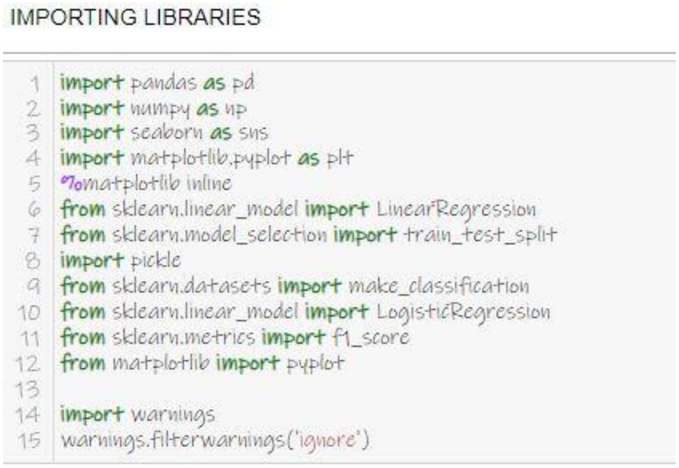
  The relationships between inputs and outputs can be studied extracting weights of the trained model. Regression is that relationships between them can be blocky or highly structured based on the training data. It requires the data scientist to train the algorithm with both labeled inputs and desired outputs.

- State the set of assumptions (if any) related to the problem under consideration

  Presumptions are by using regression label encoding, classifier, selection of the best models, confusion matrix that it means the relationship between the dependent and independent variables look fairly linear. Thus, our linearity assumption is satisfied.

- Hardware and Software Requirements and Tools Used

By importing many libraries are


IMPORTING LIBRARIES

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pickle
from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score
from matplotlib import pyplot

import warnings
warnings.filterwarnings('ignore')
```

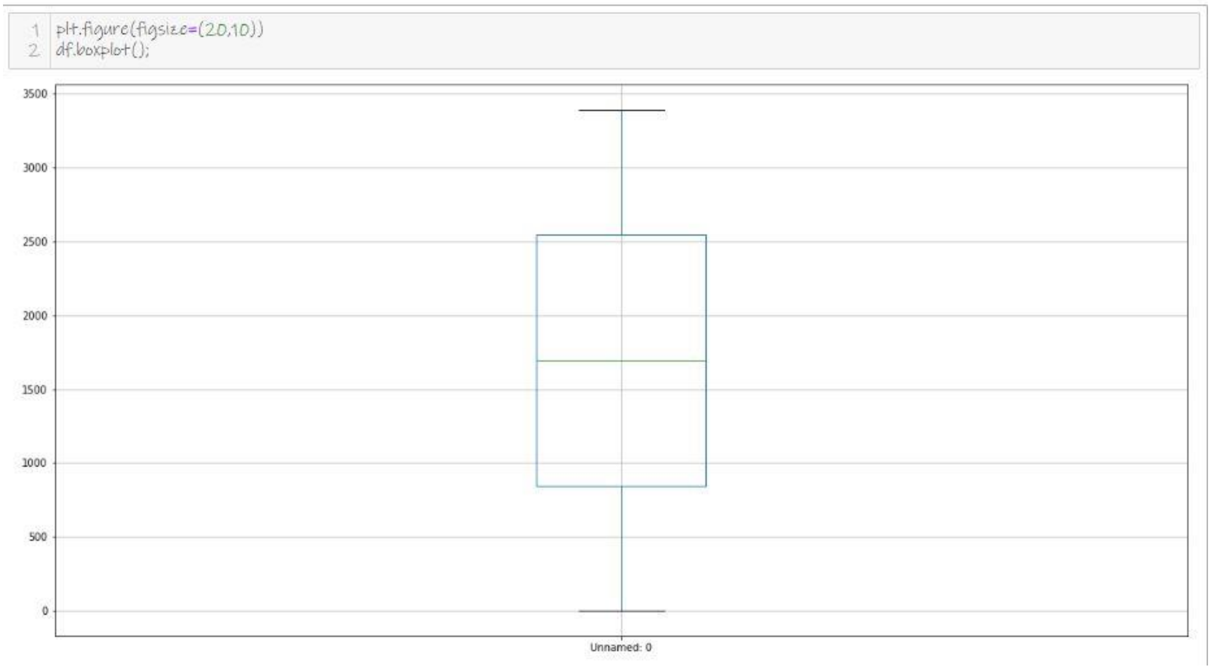
MODEL/s DEVELOPMENT AND EVALUATION

- Identification of possible problem-solving approaches(methods)

  The collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modelling or designing surveys and studies. The approaches/methods of identification are descriptive and inferential statistics which are describes as the properties of sample and population data, and inferential statistics which uses those properties to test hypotheses and draw efficient conclusions in terms of outputs.

- Testing of Identified Approaches(Algorithms)

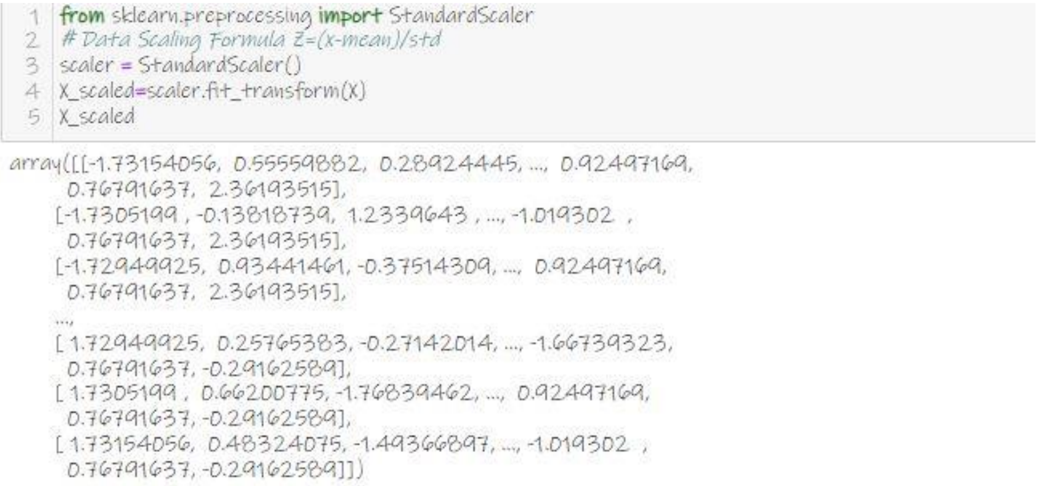
There is no outliers.



```
plt.figure(figsize=(20,10))
df.boxplot();
```

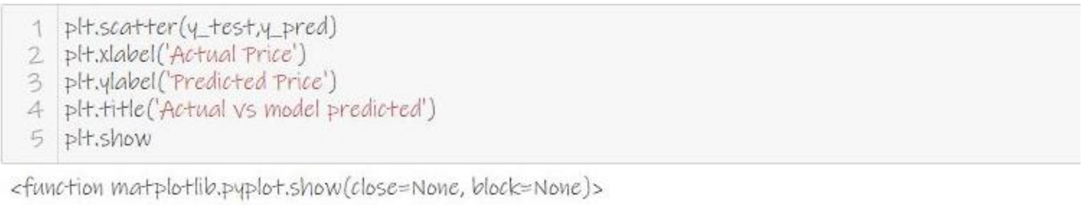
- Run and Evaluate selected models

DATA SCALING



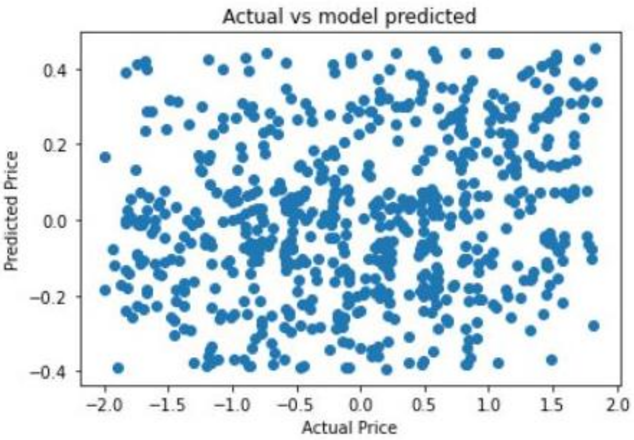
```
from sklearn.preprocessing import StandardScaler
# Data Scaling Formula Z=(x-mean)/std
scaler = StandardScaler()
X_scaled=scaler.fit_transform(X)
X_scaled
```

```
array([[-1.73154056,  0.55559882,  0.28924445, ...,  0.92497169,
         0.76791637,  2.36193515],
       [-1.7305199 , -0.13818739,  1.2339643 , ..., -1.019302  ,
         0.76791637,  2.36193515],
       [-1.72949925,  0.93441461, -0.37514309, ...,  0.92497169,
         0.76791637,  2.36193515],
       ...,
       [ 1.72949925,  0.25765383, -0.27142014, ..., -1.66739323,
         0.76791637, -0.29162589],
       [ 1.7305199 ,  0.66200775, -1.76839462, ...,  0.92497169,
         0.76791637, -0.29162589],
       [ 1.73154056,  0.48324075, -1.49366897, ..., -1.019302  ,
         0.76791637, -0.29162589]])
```

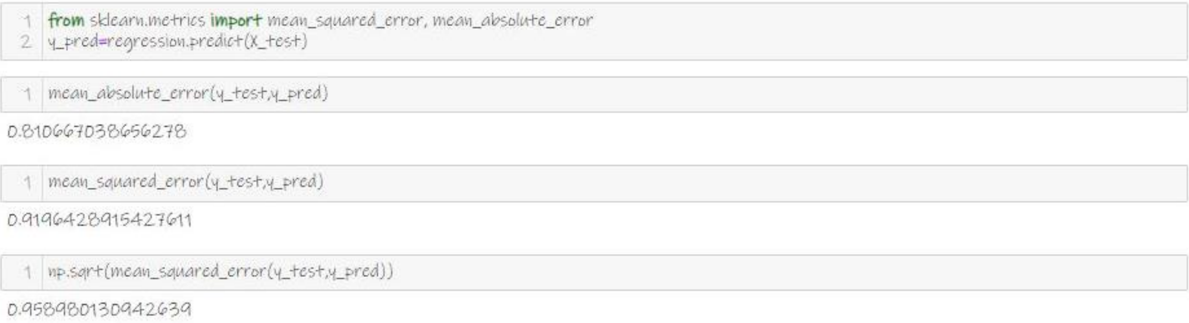


```
plt.scatter(y_test,y_pred)
plt.xlabel('Actual Price')
plt.ylabel('Predicted Price')
plt.title('Actual vs model predicted')
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



MODEL EVALUATION



```
from sklearn.metrics import mean_squared_error, mean_absolute_error
y_pred=regression.predict(X_test)
```

```
mean_absolute_error(y_test,y_pred)
```

```
0.8106670386562278
```

```
mean_squared_error(y_test,y_pred)
```

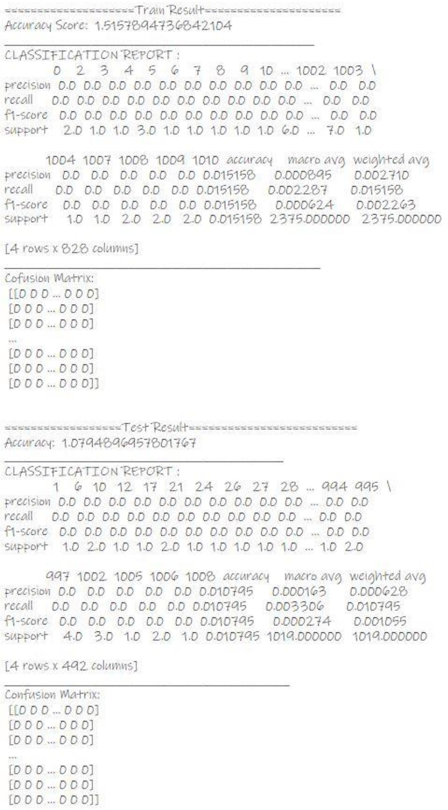
```
0.9196428915427611
```

```
np.sqrt(mean_squared_error(y_test,y_pred))
```

```
0.958980130942639
```

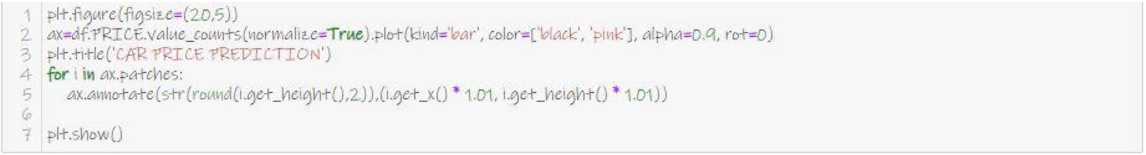
GRADIENT BOOSTING CLASSIFIER:

Accuracy score: 1%

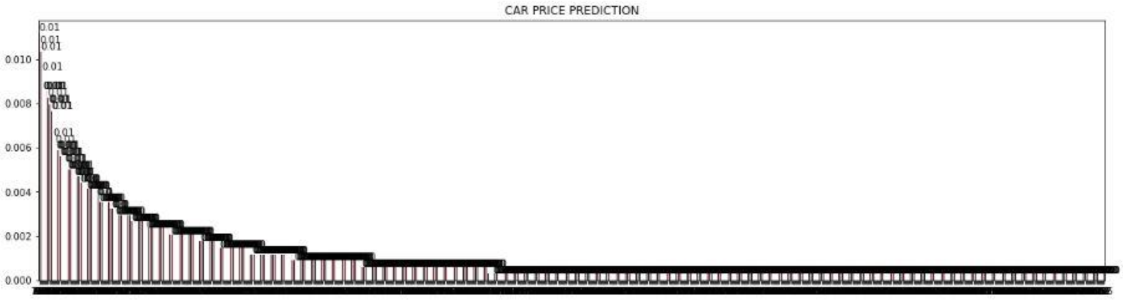


```
===================Train Result====================
Accuracy Score: 1.5157894736842104
_______________________________________________
CLASSIFICATION REPORT :
            0    2    3    4    5    6    7    8    9   10  ...  1002  1003  \
precision  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...   0.0   0.0
recall     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...   0.0   0.0
f1-score   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...   0.0   0.0
support    2.0  1.0  1.0  3.0  1.0  1.0  1.0  1.0  1.0  6.0  ...   7.0   1.0

           1004  1007  1008  1009  1010  accuracy    macro avg  weighted avg
precision   0.0   0.0   0.0   0.0   0.0  0.015158     0.000895      0.002710
recall      0.0   0.0   0.0   0.0   0.0  0.015158     0.002287      0.015158
f1-score    0.0   0.0   0.0   0.0   0.0  0.015158     0.000624      0.002263
support     1.0   1.0   2.0   2.0   2.0  0.015158  2375.000000   2375.000000

[4 rows x 828 columns]
_______________________________________________
Cofusion Matrix:
 [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]


=================Test Result========================
Accuracy: 1.0794896957801767
_______________________________________________
CLASSIFICATION REPORT :
            1    6   10   12   17   21   24   26   27   28  ...  994  995  \
precision  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0
recall     0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0
f1-score   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...  0.0  0.0
support    1.0  2.0  1.0  1.0  2.0  1.0  1.0  1.0  1.0  1.0  ...  1.0  2.0

           997  1002  1005  1006  1008  accuracy    macro avg  weighted avg
precision  0.0   0.0   0.0   0.0   0.0  0.010795     0.000163      0.000628
recall     0.0   0.0   0.0   0.0   0.0  0.010795     0.003306      0.010795
f1-score   0.0   0.0   0.0   0.0   0.0  0.010795     0.000274      0.001055
support    4.0   3.0   1.0   2.0   1.0  0.010795  1019.000000   1019.000000

[4 rows x 492 columns]
_______________________________________________
Confusion Matrix:
 [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

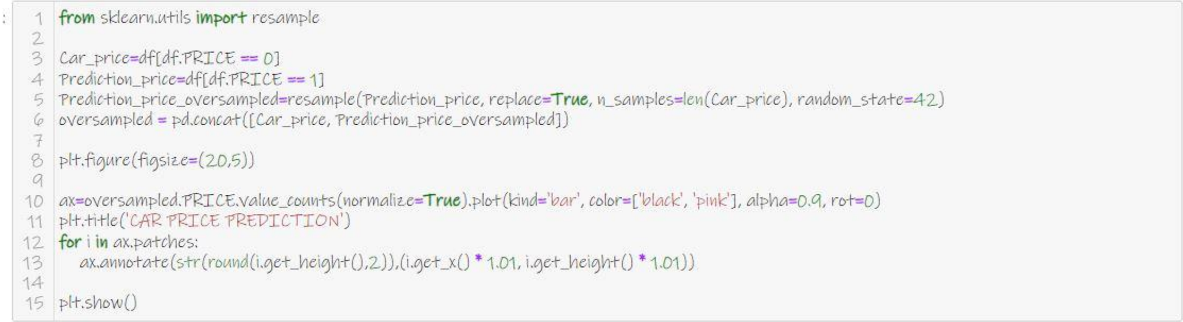
- Visualizations



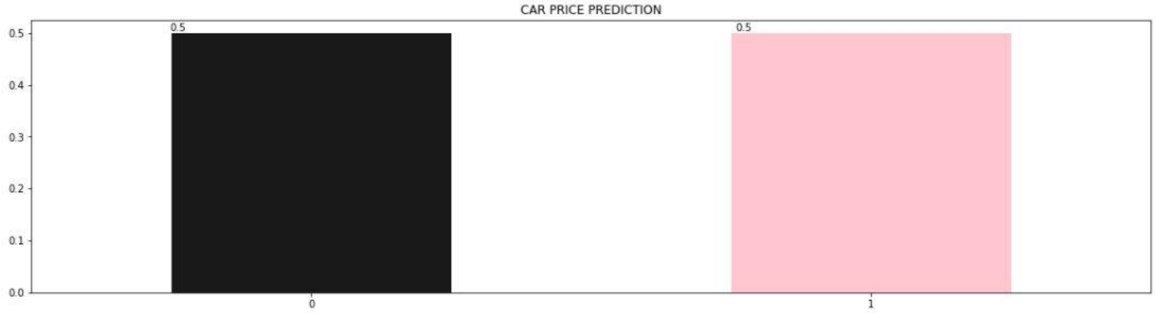
```python
plt.figure(figsize=(20,5))
ax=df.PRICE.value_counts(normalize=True).plot(kind='bar', color=['black', 'pink'], alpha=0.9, rot=0)
plt.title('CAR PRICE PREDICTION')
for i in ax.patches:
    ax.annotate(str(round(i.get_height(),2)),(i.get_x() * 1.01, i.get_height() * 1.01))

plt.show()
```



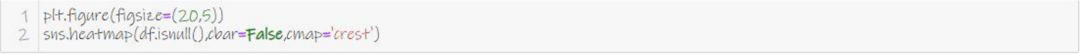
RESAMPLING THE LABEL



```
from sklearn.utils import resample

Car_price=df[df.PRICE == 0]
Prediction_price=df[df.PRICE == 1]
Prediction_price_oversampled=resample(Prediction_price, replace=True, n_samples=len(Car_price), random_state=42)
oversampled = pd.concat([Car_price, Prediction_price_oversampled])

plt.figure(figsize=(20,5))

ax=oversampled.PRICE.value_counts(normalize=True).plot(kind='bar', color=['black', 'pink'], alpha=0.9, rot=0)
plt.title('CAR PRICE PREDICTION')
for i in ax.patches:
    ax.annotate(str(round(i.get_height(),2)),(i.get_x() * 1.01, i.get_height() * 1.01))

plt.show()
```

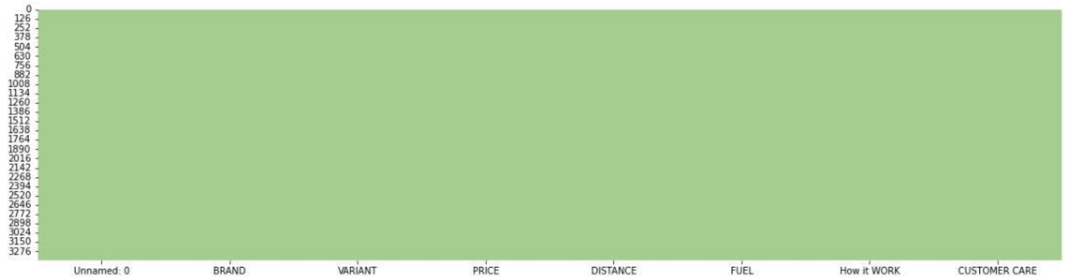


USING THE HEATMAP FOR THE DATASETS

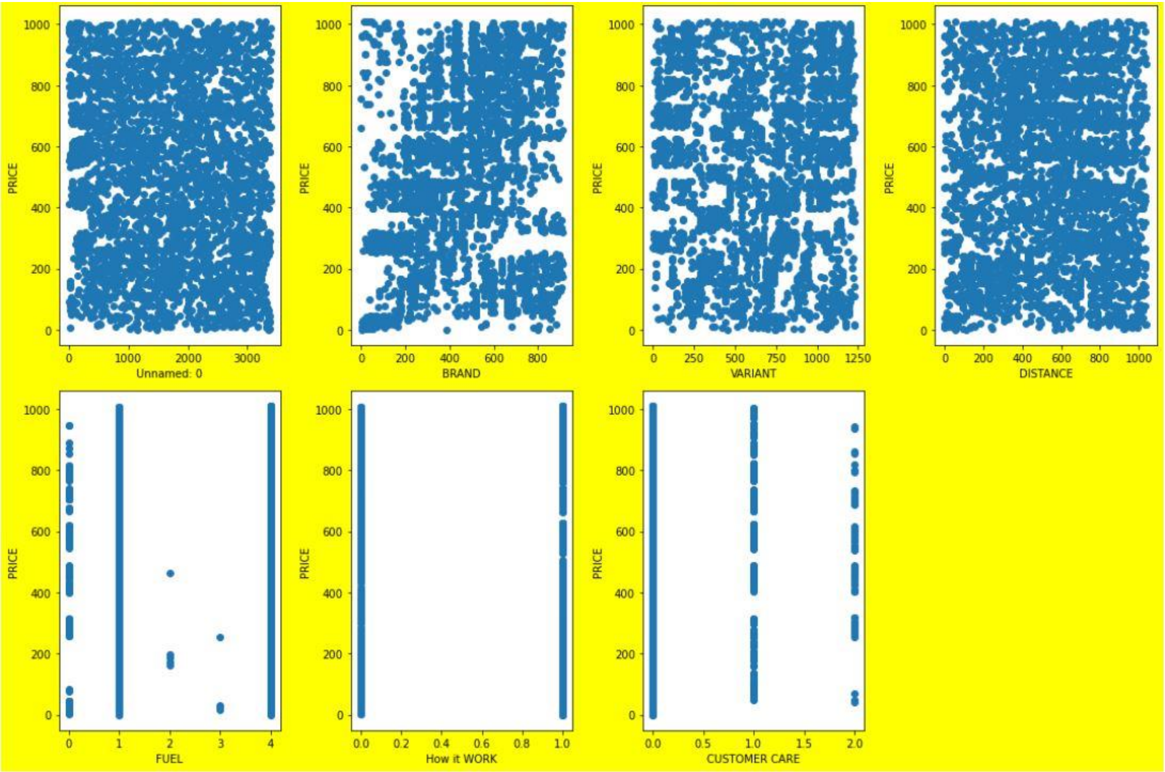


```
plt.figure(figsize=(20,5))
sns.heatmap(df.isnull(),cbar=False,cmap='crest')
```

```
<AxesSubplot:>
```



VISUALIZING THE RELATIONSHIPS



The relationship between the dependent and independent variables look good in linear. Thus, our linearity assumption is satisfied.

- Interpretation of the Results

  The results were interpreted from the visualizations, preprocessing and modelling:

  1. The purpose of this case is to understand and evaluate used car prices and to develop a strategy that utilizes data mining techniques to predict used cars prices, to help guide the individuals looking to buy or sell of used

cars to give them a better insight into the automotive sector.

2. The model of the independent variables and dependent variables are exactly vary with the variables.

3. It can accordingly manipulating the strategy of the areas that will yield high returns as it make easier for the clients.

4. By visualizations there are many things to be noted when it will according to work each other it means that from which aspect it is going to work when between the brand of the car and the price of cars get compare and between the price of the car and which fuel is used in cars compare so, it will be predicted.

5. By preprocessing the data it means that from the help of label encoder helps the dataset column to transform to fit another column in to it.

CONCLUSION

With the impact of the covid19, there are many changes took placed in the surrounding as well as big impact in the market. As this project is about the car price prediction case so, here am talking only about changes in the car market. As it is said in the statement that in the car market, some of the cars are in demand hence making them costly and some are not in demand hence cheaper, facing problems with their previous car price valuation machine learning models.
I would like to conclude here that doing research in the topic found that in the market, some of the cars are in demand and the clients are ready to buy or sell their used cars in the market through some websites. By using some parameters to predict the data according to the statement. There are many things to be noted when it will according to work each other like preprocessing the data, cleaning, visualizing, scaling and label encoder helps that dataset column to transform to fit another column into it. The model of the project are ready to analyse the independent and dependent variable.

THANK YOU