PROJECT NAME:-

[CAR PRICE PREDICTION]

SUBMITTED BY:

SHIVAM SHARMA

INTRODUCTION

Problem Statement:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model. This project contains two phase..

Data Collection Phase:-

We have to scrape at least 5000 used cars data. We can scrape more data as well, more the data better the model In this section We need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) We need web scraping for this. number of columns for data doesn’t have limit, and our creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give us a hint about important variables in used car model. We can make changes to it, you can add or we can remove some columns, it completely depends on the website from which we are fetching the data. Try to include all types of cars in our data for example- SUV, Sedans, Coupe, minivan, Hatchback. Note – The data which we are collecting is important to us.

Model Building Phase:-

After collecting the data, we need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Review of Literature

THE TOPIC IS ABOUT THE PREDICTING THE SALE PRICE OF THE CAR WHICH IS FETCHED FROM THE CARS24.COM AS WE CAN MAKE A MODEL SO THAT THEY CAN EASILY PREDICT

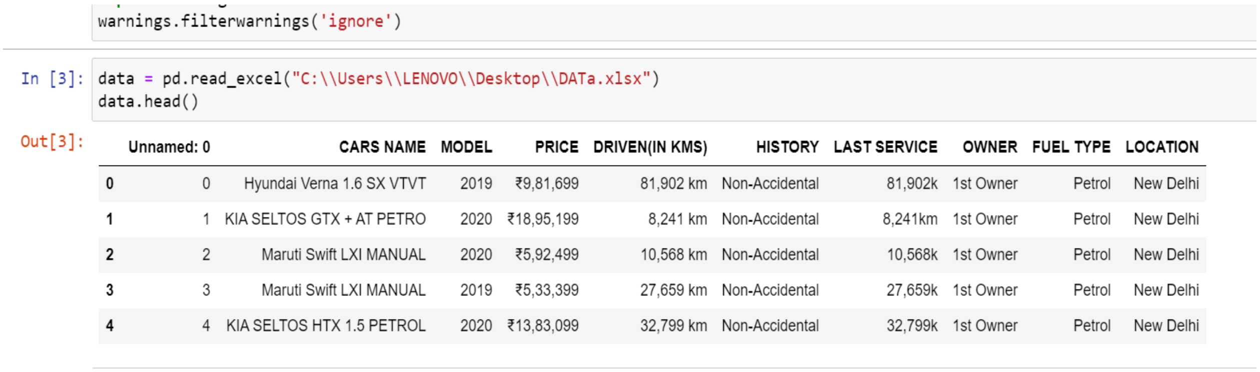
THE PRICE OF A CAR AS THEY CAN TAKE ADVANTAGE AND MAKE PROFIT FROM SELLING A CAR..

Motivation for the Problem Undertaken

Due to impactof covid19 in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data.

Data Sources and their format

- Cars24.com
- Data fetched from excel depository



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

```
In [3]: data = pd.read_excel("C:\\Users\\LENOVO\\Desktop\\DATa.xlsx")
        data.head()
```

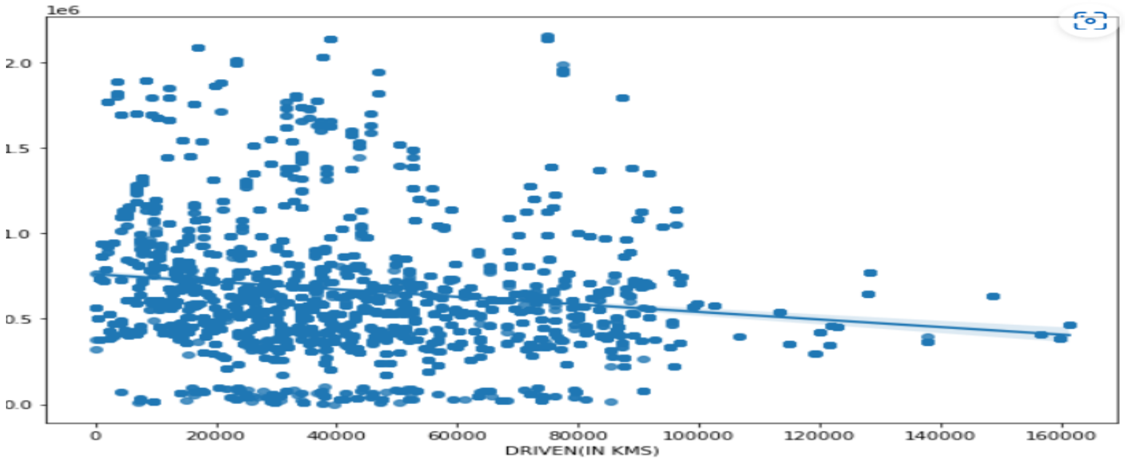
Out[3]:

	Unnamed: 0	CARS NAME	MODEL	PRICE	DRIVEN(IN KMS)	HISTORY	LAST SERVICE	OWNER	FUEL TYPE	LOCATION
0	0	Hyundai Verna 1.6 SX VTVT	2019	₹9,81,699	81,902 km	Non-Accidental	81,902k	1st Owner	Petrol	New Delhi
1	1	KIA SELTOS GTX + AT PETRO	2020	₹18,95,199	8,241 km	Non-Accidental	8,241km	1st Owner	Petrol	New Delhi
2	2	Maruti Swift LXI MANUAL	2020	₹5,92,499	10,568 km	Non-Accidental	10,568k	1st Owner	Petrol	New Delhi
3	3	Maruti Swift LXI MANUAL	2019	₹5,33,399	27,659 km	Non-Accidental	27,659k	1st Owner	Petrol	New Delhi
4	4	KIA SELTOS HTX 1.5 PETROL	2020	₹13,83,099	32,799 km	Non-Accidental	32,799k	1st Owner	Petrol	New Delhi

Data Preprocessing and EDA

The steps followed for the cleaning and EDA of the data:-

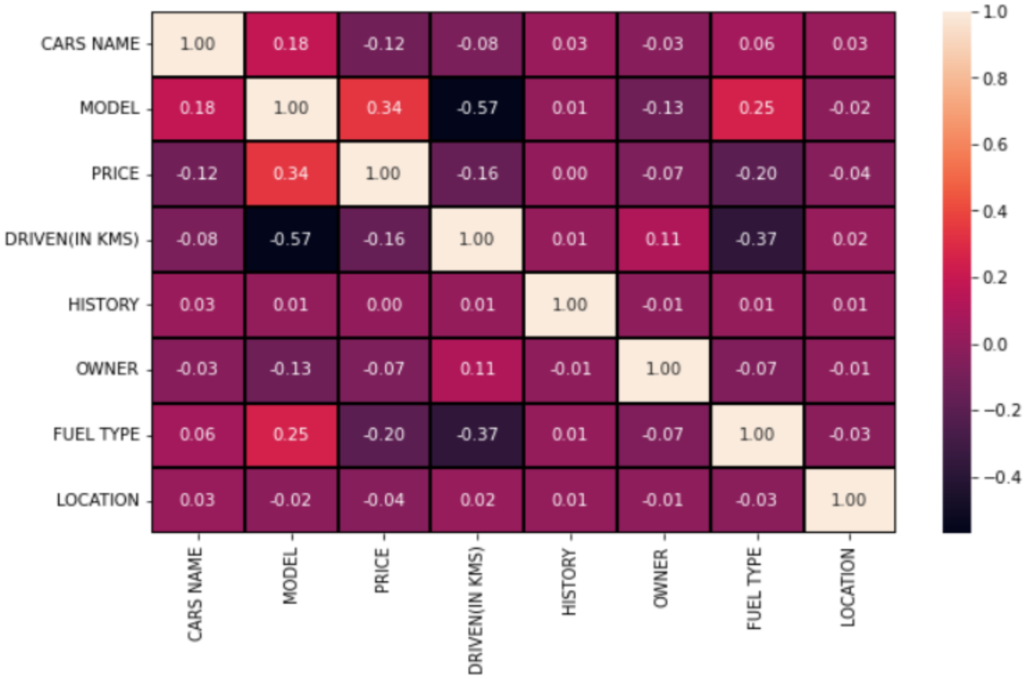
1) First we analyse the data by simply DATA.INFO() method..
2) Then we used Drop function for delete some columns from the dataset which having some Index (ex- Unnamed:0)
3) Then we analyse the data with countplot visualization techniques with column on x axis by which it is easy to interpret the data.
4) Then we replace some columns value which has object datatype so that they can be easily understandable by machine i.e- int datatypes..
5) After that we plot a graph of regression plot with the help of regplot to analyse the Price trends with many columns..



6) Now we will check the correlation between the independent variables and between the independent variables and dependent variables.

7) Then we Plot a Heatmap to check clearly the correlations between all the variables.

```
<AxesSubplot:>
```



8) Since the data Is continuous data so we are not checking the skewness and outliers.

9) After these methods we seprate the lables and features in different dataframe so that we can easily apply some methods on the labels and features differently..

10) After this we applied scaling on the features (x) by the help of Standard Scaler…

11) Now we are ready for the model building……

Algorithm used in this project:-

For building machine learning models there are several models present inside the Sklearn module.

Sklearn provides two types of models i.e. regression and classification. Our dataset's target variable is to predict the sale price of the car. So for this kind of problem we use regression models.

But before the model fitting we have to seprate the predictor and target variable, then we pass this variable to the train_test_split method to create the training set and testing set for the model training and prediction.

We can build as many models as we want to compare the accuracy given by these models and to select the best model among them.

I have selected 5 models:

1. Linear Regression

2. Lasso
3. Random forest Regressor
4. Adaboost Regressor
5. Gradient boosting Regressor

BEST Algo. From these And why?

Gradient boosting Regressor is the best algo from all of these algo which is used in this data to predict because the difference between the cross val score and the accuracy score is minimum for this algo and it also give better **ACCURACY(approx. 97%)** after Hypertuningwith GRIDSEARCHCV that's why we USED THIS **Algo.**

- **Save the model for later predictions by the help of pickle..**

# Now my model is ready to predict

CONCLUSIONS

THE DATAFRAME CONTAINS THE DATA WHICH RELATED TO THE SALE PRICE PREDICTION OF THE CAR..

We got our best model i.e **GRADIENT BOOSTING REGRESSOR with the accuracy score of 96.7%.** HERE our model predicts ROOT MEAN SQUARED ERROR OF 183574.60 THAT IS VERY LOW THAN OTHERS..

KEYFINDING:-

WE DON'T CHECK THE OUTLIERS AND SKEWNESS IN THIS DATASET BECAUSE THE DATA IS CONTINOUS ..