**Car Price Prediction Project**

Submitted by:

Tushar Saini

**Acknowledgement**

I would like to express my sincere gratitude to fliprobo technologies for supporting me in this project. I wish to express my sincere gratitude to my SME, Ms. Sapna Verma, for his enthusiasm, patience, insightful comments, helpful information, practical advise and unceasing ideas that helped me tremendously at all times in my project time. Her immense knowledge, profound experience and professional expertise in Data Science has enabled me to complete this research successfully. Without her support and guidance, this project would not have been possible.

I also wish to express my sincere thanks to the DataTrained for helping me to get trained in the field of Data Science. Without them I don’t know anything about Data Science. They help me to learn the Data Science from scratch. The faculty their helped me a lot in every aspect.

I also like to thank websites like geeksforgeeks.com, towardsdatascience.com, etc for providing the information regarding many topics of data science.

1. **Introduction:-**

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 scrape at least 5000 used cars data. I am scrapping the data from 'Carwale.com'. We have to fetch data for different locations. The number of columns for data doesn’t have limit, it’s up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, and number of owners, location and at last target variable Price of the car.

This data is to give you a hint about important variables in used car model. We can make changes to it, we can add or we can remove some columns, it completely depends on the website from which we are fetching the data. I tried to include all types of cars in my data for example- SUV, Sedans, Coupe, minivan, Hatchback.

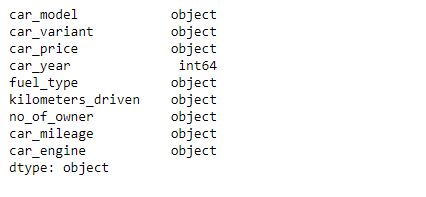
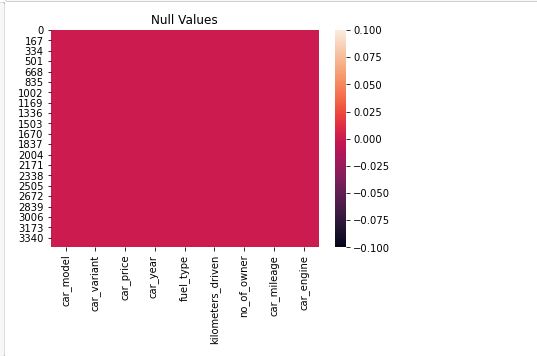
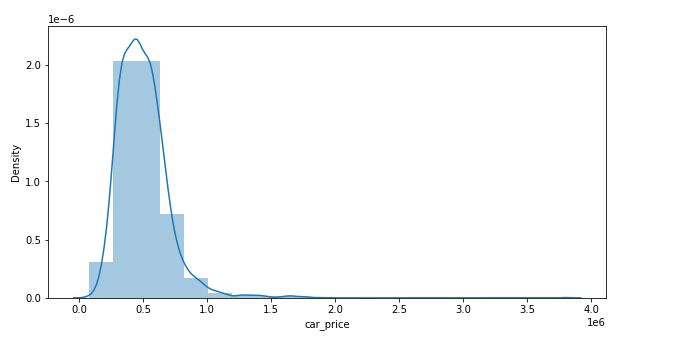
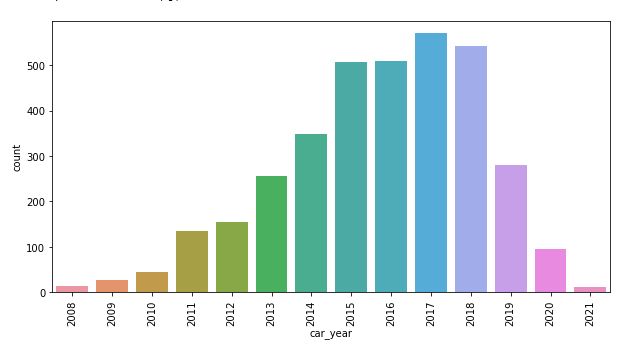
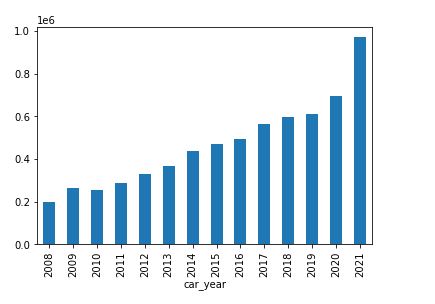
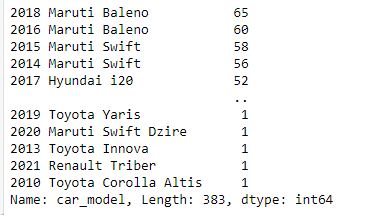
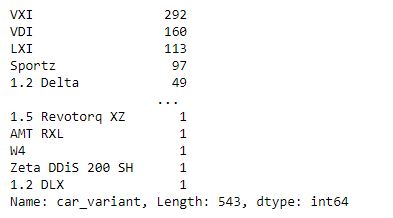
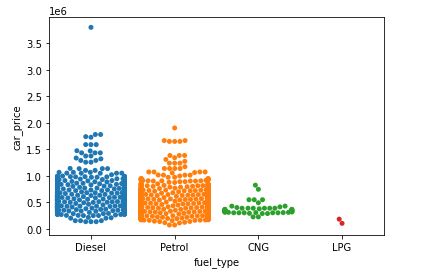
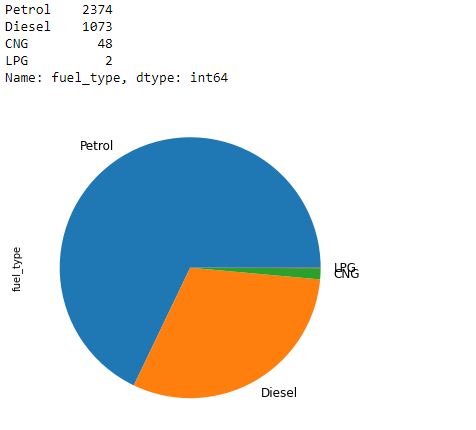
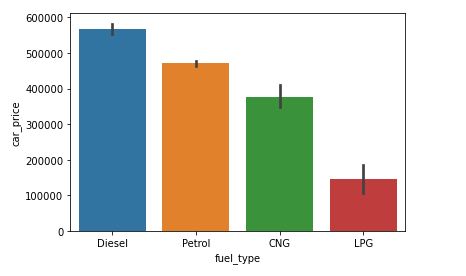
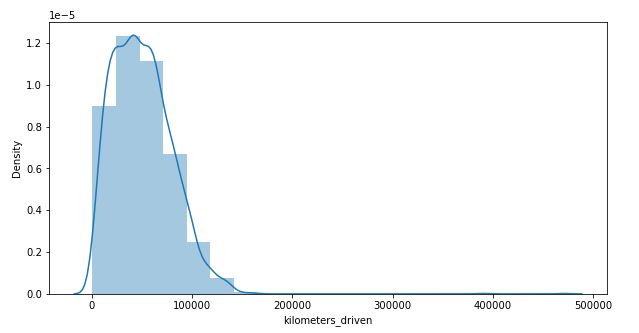
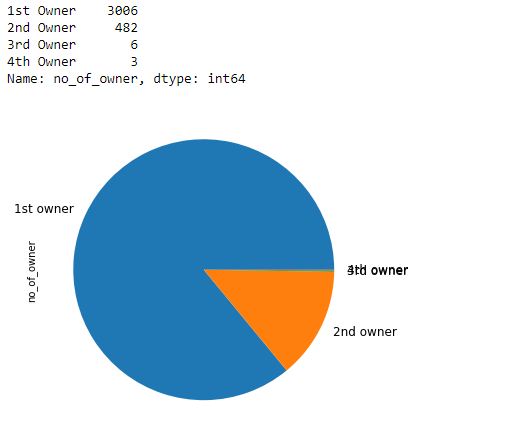
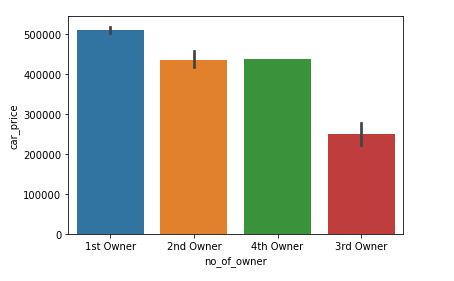
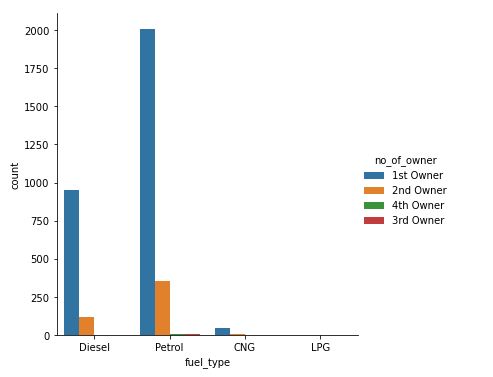
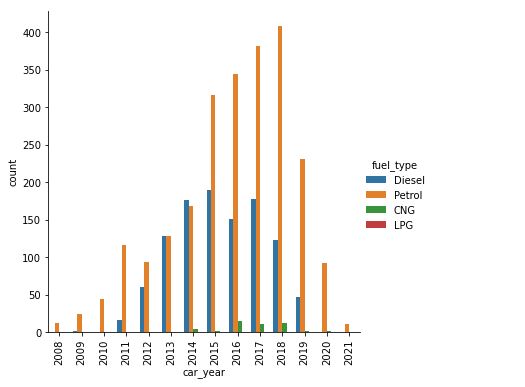
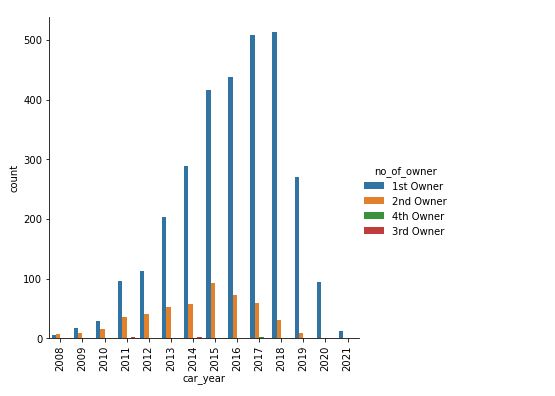
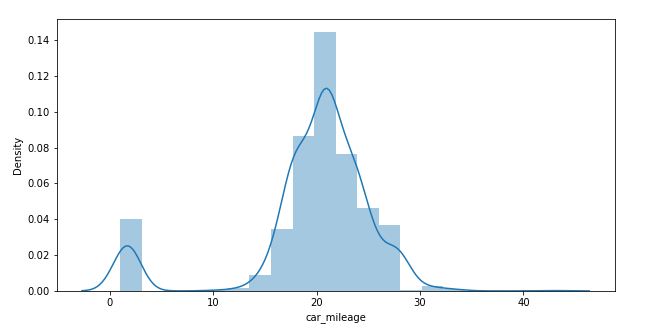
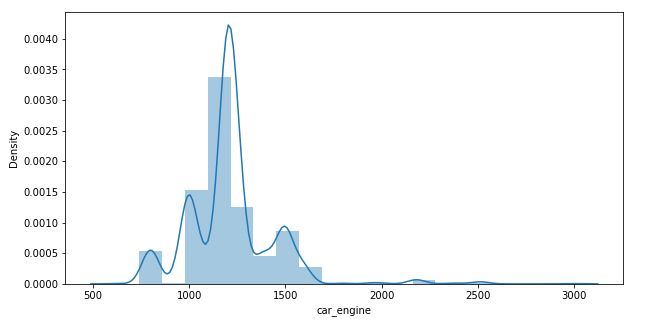
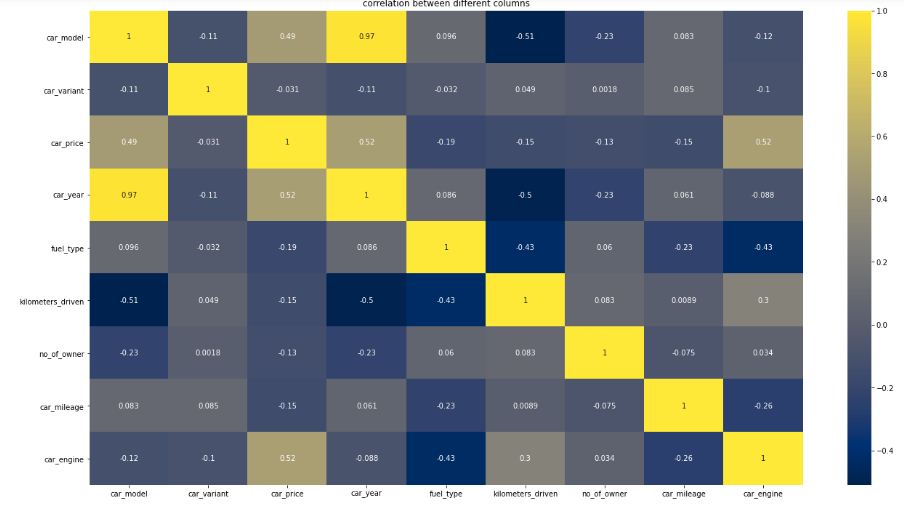
I fetched the data from cardekho.com, as it has all the data which I need for my model building like model of the car, variant of the car, owner of the car, manufacturing year, mileage of the car, engine capacity and lastly the price of the car. I fetched the data from many different locations like Delhi, Gurugram, Noida, Gujarat, Rajasthan, Mumbai, Pune, Banglore, Hydrabad, Uttar Pradesh.

After collecting the data, you 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.

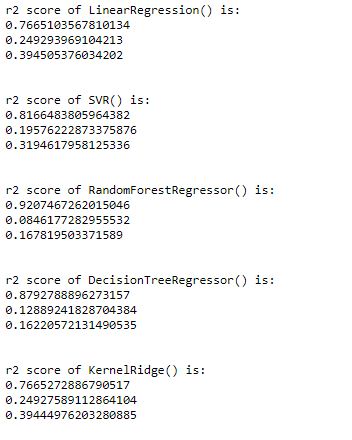
1. **Analytical Problem Framing:-**

* **Data Visualization :-**

By analyzing the data, I have found some important features of that dataset. The analyzed information that I found are given below:-

* There are 3497 rows and 9 columns in the dataset and it doesn’t have any null values.
* I also observe that the datatypes of columns in the dataset are a mixer of ‘int’, ‘object’.
* There are no null values in the dataset.
* There are 383 unique car models in the dataset.
* That there are 543 unique cars variant in the dataset.
* The average prices of the used cars are around 500000 rupees.
* We can observe that majority of the cars are from year 2015 to 2018.
* The price of the car are gradually decreasing as the car becomes old.
* We can observe that '2018 Maruti Baleno' has the highest values in the dataset that means people are selling this car more.
* We can observe that 'vxi' variant of the car has the highest value that means people are selling this variant more.
* The car prices of diesel and petrol car are almost same. The prices of CNG or LPG cars are very less as compared to diesel or petrol cars.
* Petrol cars are in more numbers on sale as compared to diesel, CNG, LPG.
* Diesel cars have high price as compared to any other fuel type.
* Most of the cars are driven less than 100000 kms.
* Majority of the cars are owned by their 1st owner.
* Car price are higher for 1st owner car as compared to others.
* We can observe from the above graph that in every fuel type, the majority of the car owner is 1st.
* We can observe that in almost every year majority of the car are petrol type but in year 2013, it is same for diesel and petrol and in 2014 diesel cars more as compared to petrol.
* We can observe that in every year majority of the cars are owned by the 1st owner.
* We can observe that majority of the cars have a mileage of around 20kmpl.
* We can observe that majority of the engines are from 1000CC to 1500CC.
* We can observe that 'car\_engine' and 'car\_year' are the most positively correlated columns with the 'car\_price' and 'fuel\_type' is the most negatively correlated column with the 'car\_price'.
* **Data Pre-Processing:-**
* We can also drop column 'unnamed: 0' as it is for no use.
* We have converted the data type of price column from object to int as price should be an int column.
* We have converted the datatype of kilometers\_driven column from object to float as kilometers should be a float column.
* We have converted mileage from 'km/kg' into 'kmpl' then converting it into float as mileage should an float column.
* We have converted the car\_engine column in float from object as engine capacity should be a float column.
* We have encoded the dataset using OrdinalEncoder.
* There are some outliers present in the car\_price but the price can be high depend on the car model and variant. So we are not removing outliers from it.
* We observe that there are some skewness present in the ‘car\_price’, ‘kilometers\_driven’, ‘no\_of\_owner’, ‘car\_mileage’, ‘car\_engine’. We have removed the skewness from these columns using ‘Power\_Transform’ except ‘no\_of\_owner’ as it is a classification column.
* We have scaled the data using StandardScalar.
* **Hardware and Software Requirements and Tools Used:-**
* We used anaconda jupyter notebook to do research on this project.
* I use python to do my coding for the dataset.
* We use different kind of libraries present in the jupyter notebook like pandas, numpy, sklearn, seaborn, matplotlib, joblib, etc.
* I have used dell inspiron 1440 which has 3GB Ram and windows 7.
* I use pandas to import that data make it in a dataframe.
* Sklearn is used to import all the model and methods that are used for pre-processing and model building of the dataset.
* Matplotlib and Seaborn are used to plot various graphs and figures to visualize the data.

1. **Model/s Development and Evaluation:-**

* For building a machine learning model first I split the dataset into independent and target variable.
* Independent variable that is ‘x’ contains all the other columns except ‘car\_price’ because ‘car\_price’ is our target value which we have to predict.
* Target variable which is ‘y’ contains the ‘car\_price’ column.
* Now I find the best random state for one model and I use that random state for other models also.
* First I use ‘Linear Regression’ as it is a Regression problem. I make a loop and take 0-200 random state. That loop will give me the accuracy score for every random state from 0-200.
* At random state 99 the testing accuracy and training accuracy is highest.
* Now I test this random state with other models also.
* I used a total of 5 models which are ‘Linear Regression’, ‘SVR’, ‘RandomForest Regression’, ‘DecisionTree Regression’, ‘KernelRidge’.
* ‘Random Forest Regression’ gives me the best accuracy score that is 92.07%.
* Now I check whether my models are underfitted or overfitted. For that I use cross validation score and taking the k-fold value of 3.
* The Minimum difference in accuracy score and cross validation score is for RandomForest Regression (3.89) so it is our best model.
* After that I have done hyperparameter tuning of the model to increase the model accuracy score.
* After doing the hyperparameter tuning our model score has increased from 88.11% to 92.31%.

1. **Conclusion:-**
   * From the above project I find that the majority of the people are selling there car at a price around 500000 rupees.
   * Selling the old car helps people to buy new one and the person who want to buy a good car in less price.
   * I made a machine learning model to check what will be the car price by taking different parameters of the car. My model is 92.31% correct.
   * I have learnt that Random Forest Regression algorithm is good for price prediction problems.
   * I have learned that through visualization we can understand the core of the data. Visualization is very important for a data science project as it will give us information about the dataset.
   * I have used 5 algorithims. What I have learned that it might be possible some algorithms can be used for a specific problem. So we have to check what is the problem through which we can choose our algorithm.