

Submitted by:

Ankit Prasad

**ACKNOWLEDGMENT**

I want to express my great appreciation to Sajid Choudhary for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. I would also like to thank the staff of the Data Trained for helping me with the problems I faced during the research work. Articles from the "Medium" platform were beneficial during the whole process. It helped me clear my concepts.

**INTRODUCTION**

With the covid 19 impact in the market, we have seen many changes in the car market. Some cars are in demand, making them costly, and some are not in order, hence cheaper. One of our clients works with small traders, who sell used cars. With the change in the 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 a car price valuation model.

**Objective**

This project contains two-phase**-**

**Data Collection Phase**

We scraped at least 5000 used cars data. In this section, You need to scrape the data of used cars from websites. We used web scraping for this. We fetched data for different locations. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometres, fuel, number of owners, location and at last target variable Price of the car.

**Model Building Phase**

After collecting the data, we built a machine learning model. Before model building, we did data pre-processing steps.

Followed the complete life cycle of data science. Include all the steps like.

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

**Analytical Problem Framing**

In the whole research process various mathematical, statistical and analytics modelling has been done. There has been reduction of the columns because few of them was not necessary for the problem solving like Id. And few of them was removed due to very less correlation with dependent variable. Since the dataset contains a lot of features hence feature selection has been also done. To fix the outliers we used z score method. After this also there was a lot of skewness in dataset so power transform has been used. To check the accuracy r2 score was used also for cross validation cross\_val \_score is used.

**DATA/ DATA PREPROCESSING:**

* The dataset contains 6384 rows and 12 columns
* Price is our dependent variable.
* We created new features from old ones.
* All columns were object data types we converted Kilometre, year and Price in numeric.
* There are no null values in the dataset.
* There were few empty cells in the dataset
* Removed empty cells
* Trimmed few columns

Hardware and Software Requirements and Tools Used

* Acer Aspire 5- i5 8th generation, 8gb ram, NVidia mx130 integrated graphic,
* JuypterNotebook/Google chrome
* Libraries and packeges used:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings("ignore")

from sklearn.preprocessing import LabelEncoder

from sklearn.feature\_selection import VarianceThreshold

from sklearn.feature\_selection import mutual\_info\_regression

from sklearn.feature\_selection import SelectPercentile

from sklearn.preprocessing import StandardScaler

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

from sklearn.preprocessing import power\_transform

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error, mean\_absolute\_error, r2\_score

from sklearn.linear\_model import LinearRegression

from sklearn.tree import DecisionTreeRegressor

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import ExtraTreesRegressor

from sklearn.svm import SVR

from sklearn.model\_selection import cross\_val\_score

from sklearn.ensemble import BaggingRegressor

from sklearn.ensemble import AdaBoostRegressor

from sklearn.ensemble import GradientBoostingRegressor

from sklearn.model\_selection import GridSearchCV

the library used here is sklearn, numpy, matplotlib, pandas and seaborn. The matpotplotlib and seaborn library has been used to make charts to visualize and understand the problem, correlation, outliers and many other things, the pandas and NumPy library is used to handle dataset and perform various tasks. The seaborn library is used for model building and cross validation of the models.

**Model/s Development and Evaluation**

The approach to solve this problem was to get the domain knowledge to understand the data better. Which values can be the part of the data and which is not? After exploring the data, it is found that though the data has no missing value. It has extreme outliers and unrealistic value. We used Z-Score method to remove outliers. There was some skewness in the data, power transform method has been used so it dealt skewness. To check the accuracy, mean square error, mean absolute error, r2 score was used also for cross validation cross\_val \_score is used

**Algoritham used for Traning and testing:**

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error, mean\_absolute\_error, r2\_score

from sklearn.linear\_model import LinearRegression

from sklearn.tree import DecisionTreeRegressor

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.ensemble import ExtraTreesRegressor

from sklearn.svm import SVR

from sklearn.model\_selection import cross\_val\_score

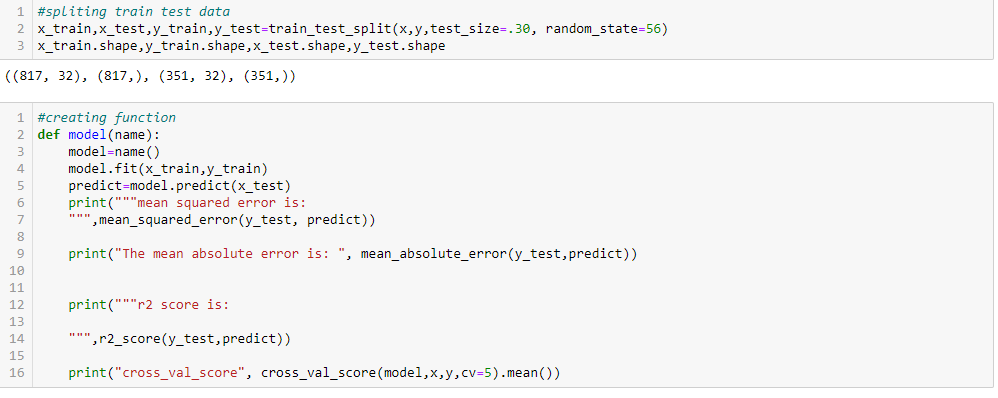
from sklearn.ensemble import BaggingRegressor

from sklearn.ensemble import AdaBoostRegressor

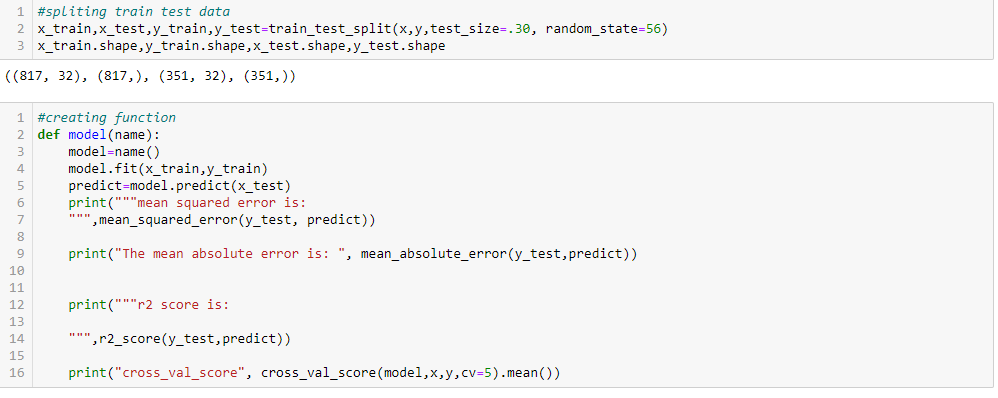
from sklearn.ensemble import GradientBoostingRegressor

from sklearn.model\_selection import GridSearchCV

**Performance of the model:**

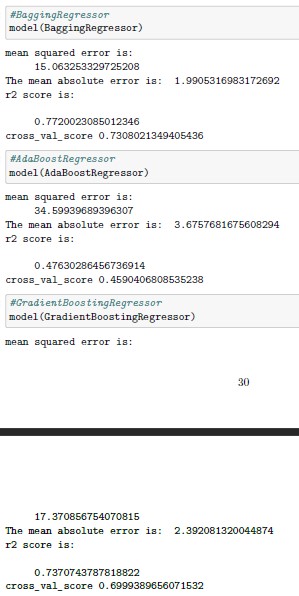


By the train test split method, 70 percent of the data has been taken for the model building while 30 percent of the data has been reserved for checking the model's performance.



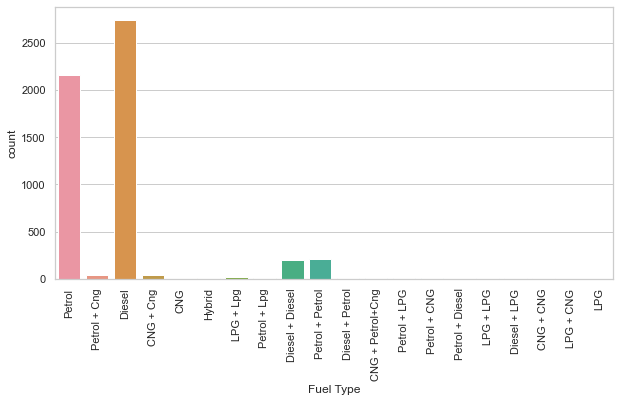
The code above has been used to speed up the model training and its evaluation process. Here the function name model ha been created which take the name of model as argument.

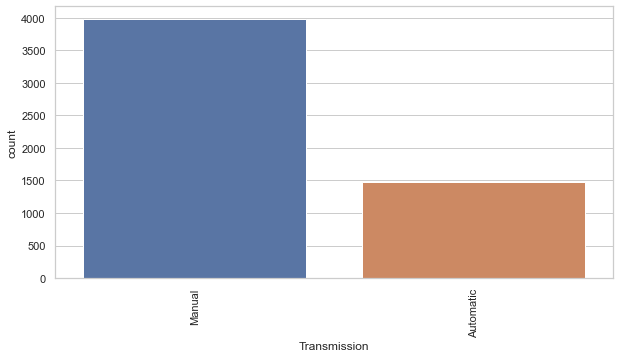


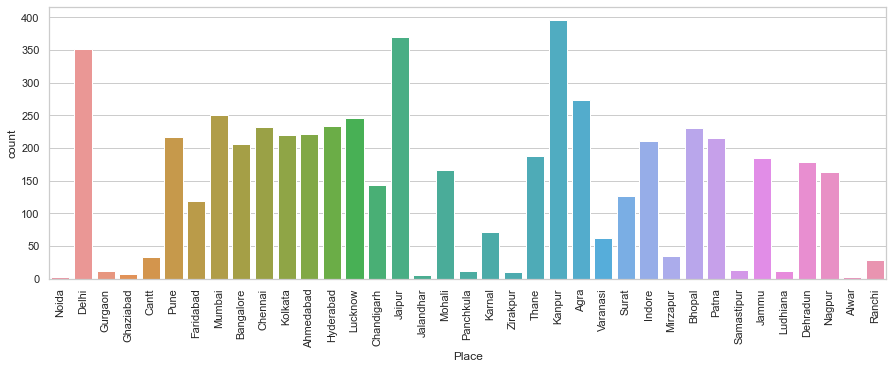


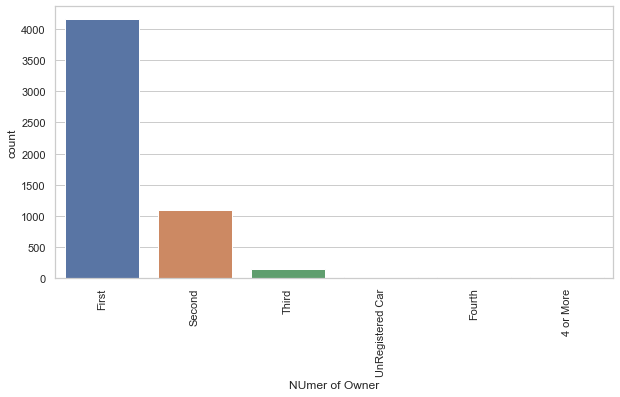


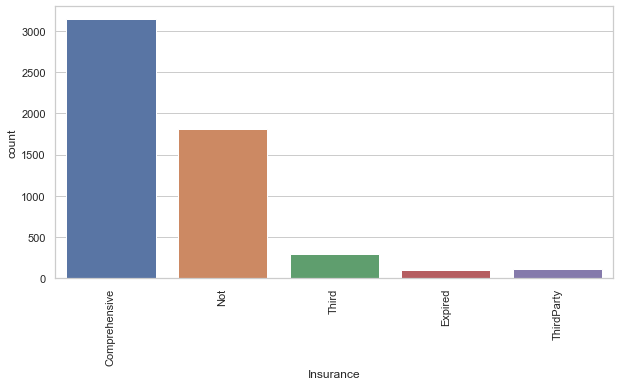
**Visualization:**

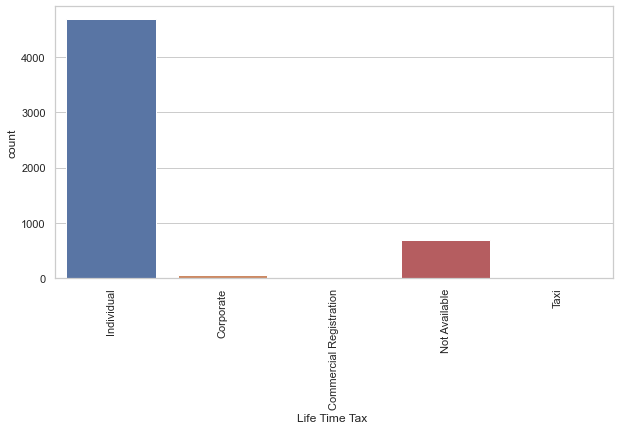
****

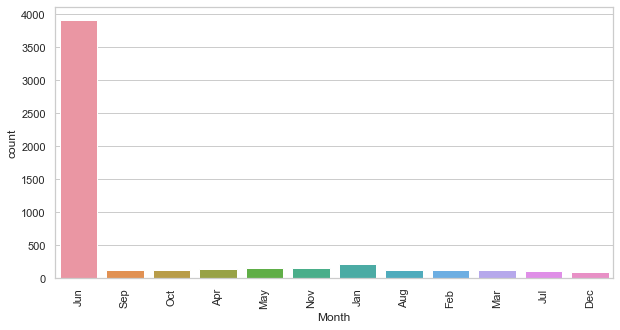
****

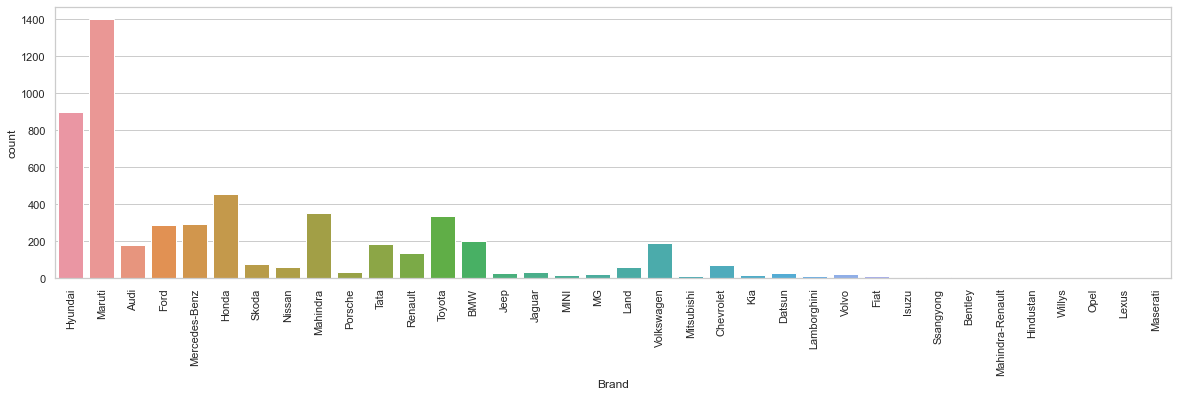
****

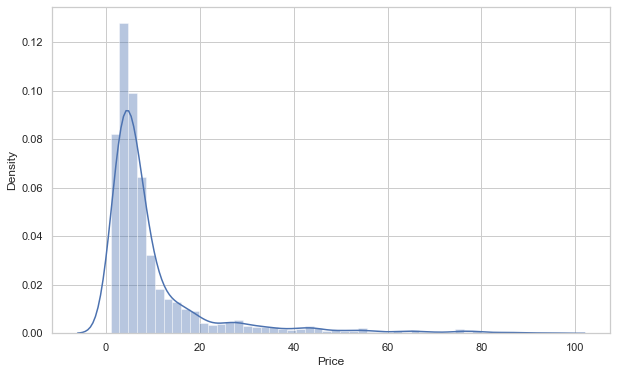
****

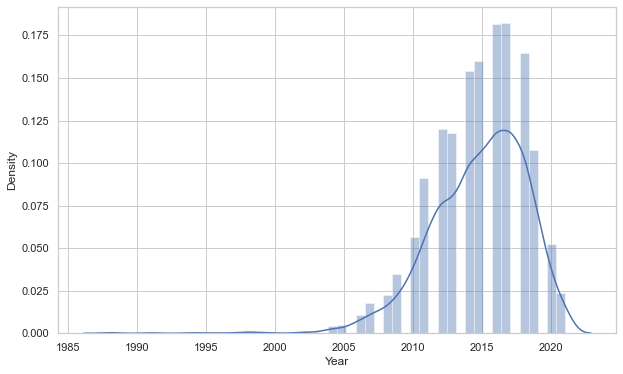
****

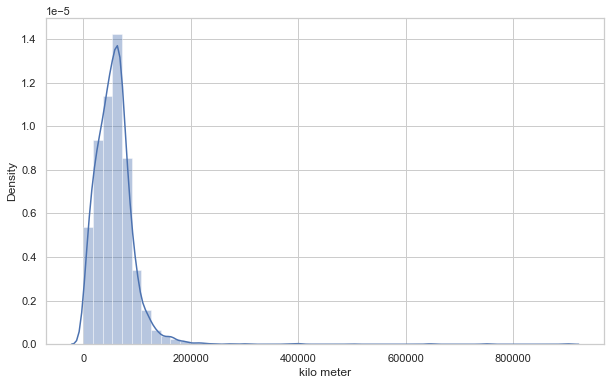
****

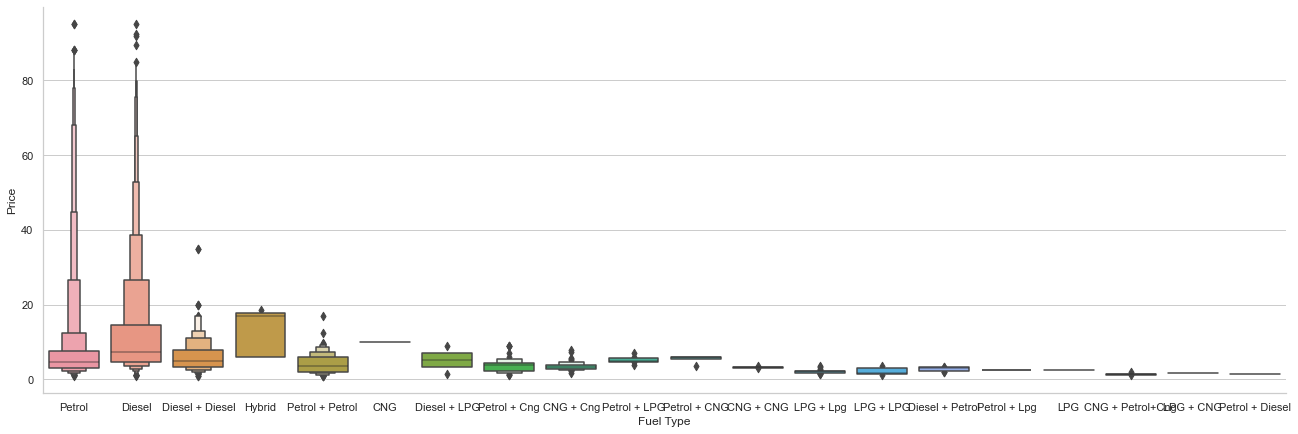
****

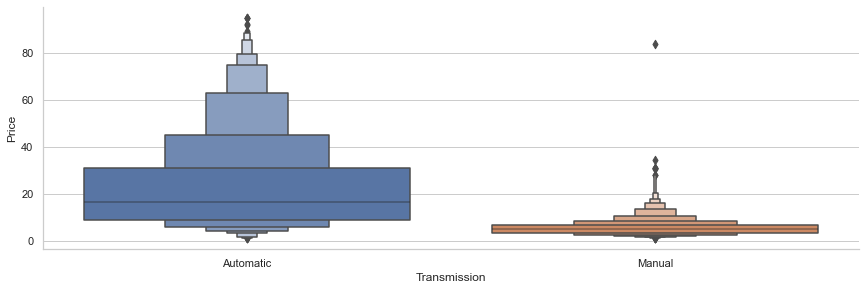
****

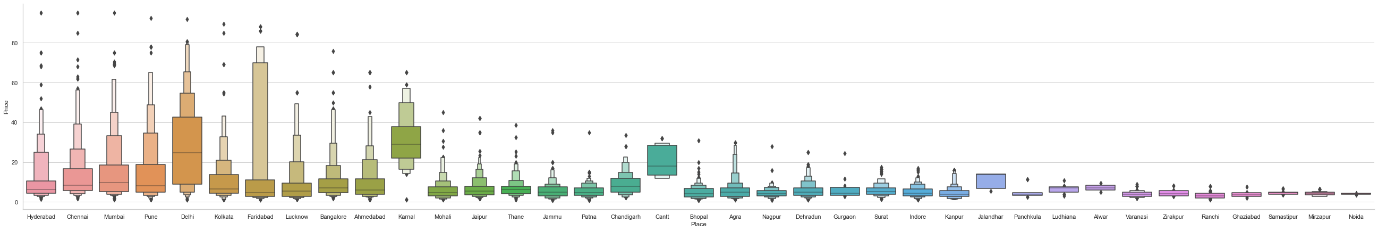
****

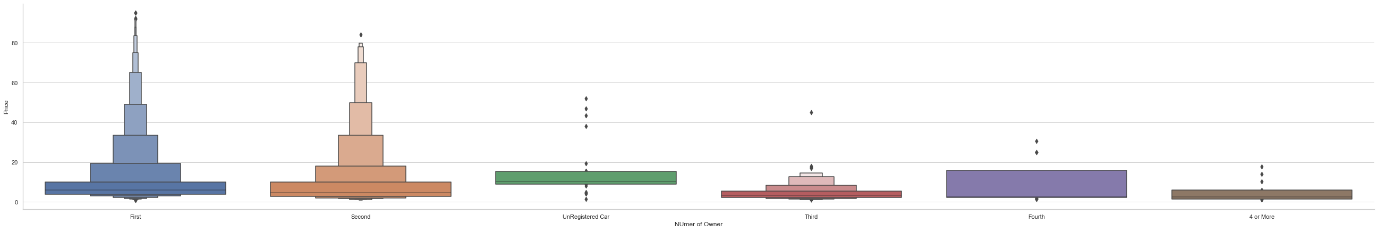
****

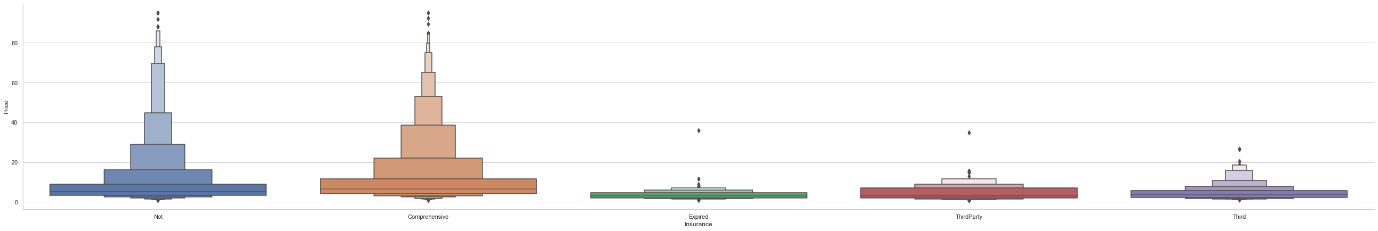
****

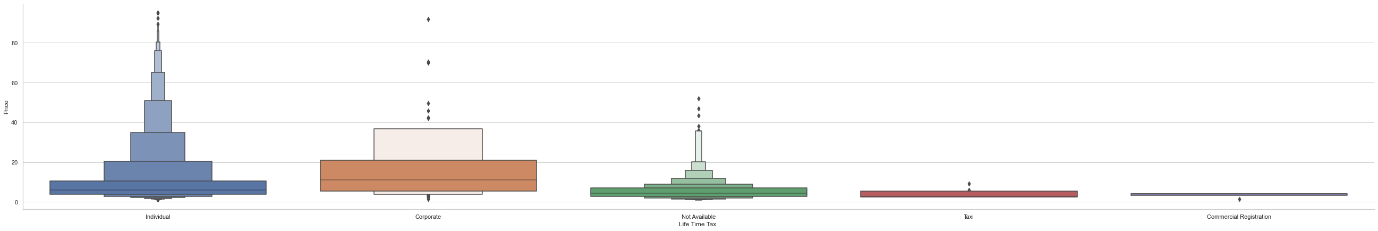
****

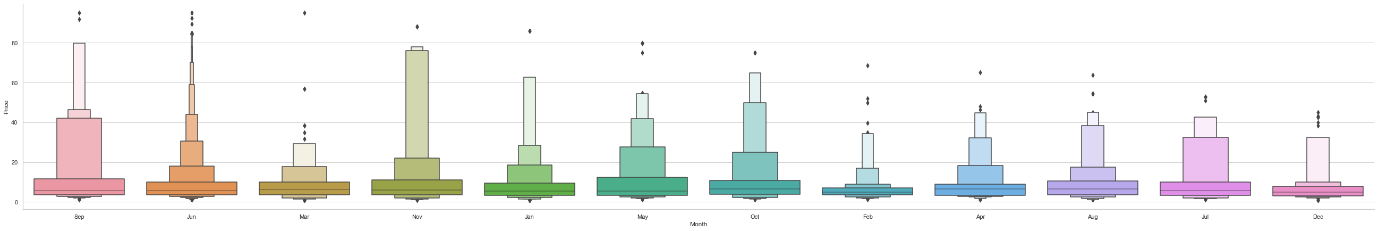
****

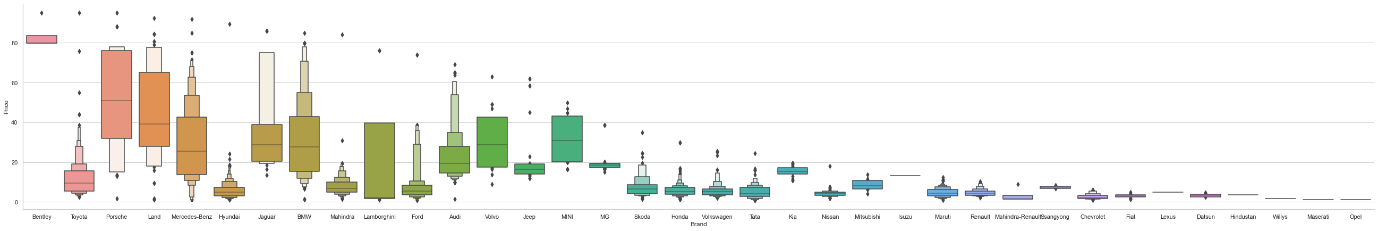
****

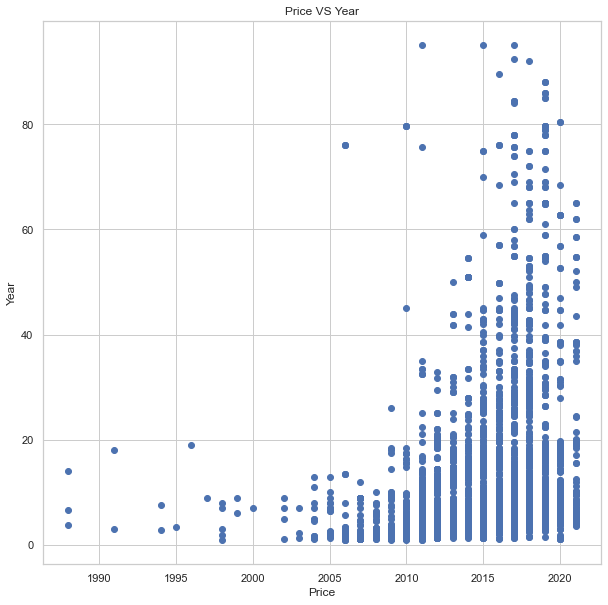
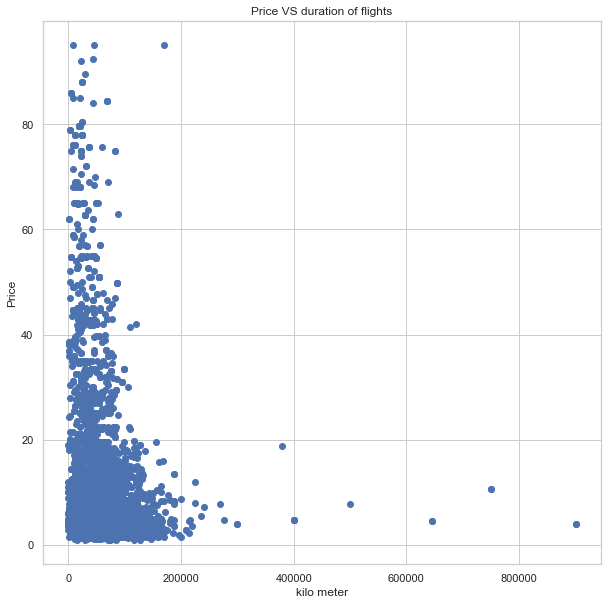
****

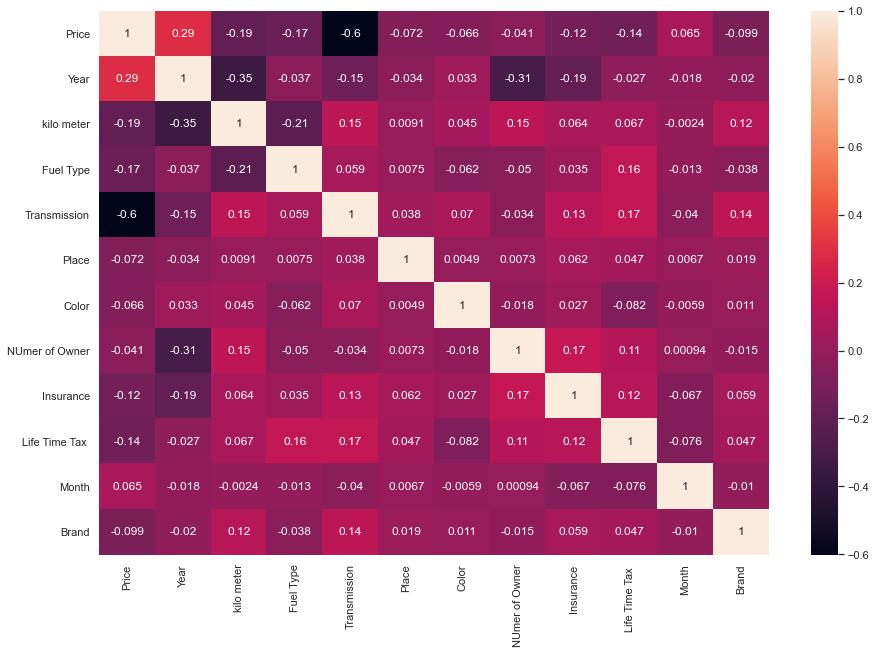
****

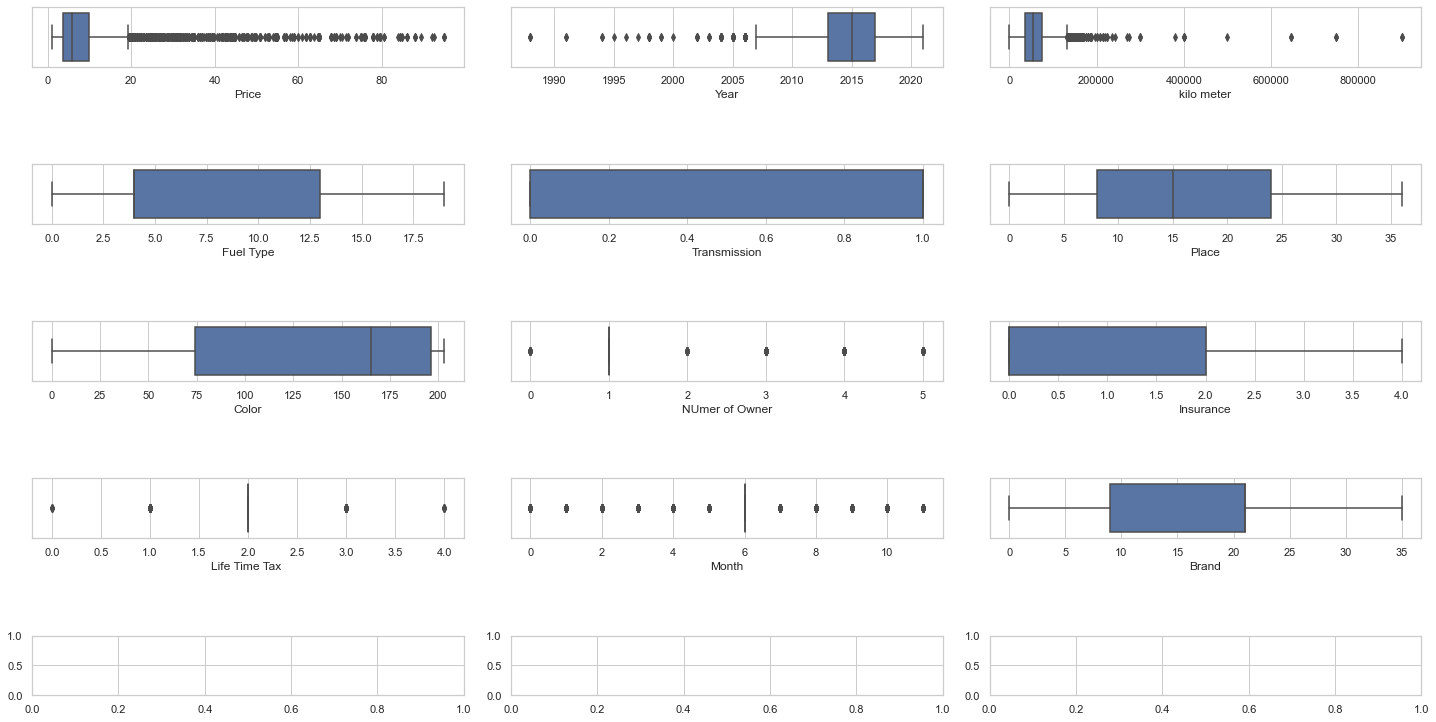
****

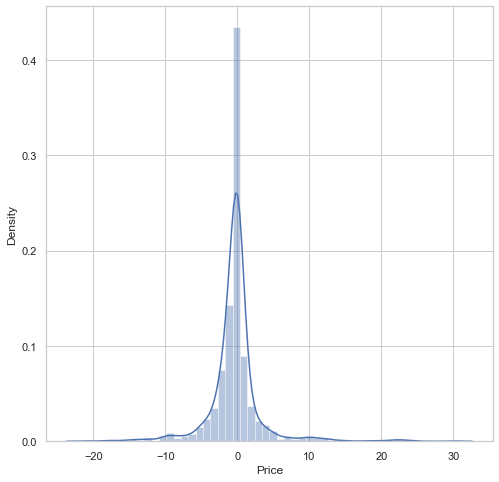
****

****

**  
**

****

****

****

**Observations:**

# The number of cars with diesel option in more followed by Petrol.

# Manual Transmission is more available.

# Noida, Jaipur and Kanpur have highest number of cars listed for sale.

# Most of the cars are available in white followed by Silver.

# Most cars are fist hand

# Most Cars have Comprehensive insurance while a lot of cars doesn't come with insurance

# Majority of cars are listed in June.

# Maruti cars are most listed cars followed by Hyundai

# Majority of cars are listed below 20 lakhs.

# Majority of cars model are between 2010-2017

# Very Few Cars HAs been used above 20k Kilometre

# Diesel and Petrol fuel type has higher price in comparison to others.

# Automatic cars are pricier than Manual.

#The average price of Delhi is higher than others

# There are cars in Hyderabad and Chennai whose price is heights.

# The price starts decreasing as the number of owners starts increasing.

# Third Party and expired insurance has very less value.

# Individual life time tax has more value.

# The price in September and Jan is high.

# Bentley has highest Price but it is few in numbers

# Toyota, Porsche and land rover are high price brands

# Latest the model Highest the price

# The less driven car has more value

# year and transmission are most correlated with price

**CONCLUSION**

This paper showed the model training process for the prediction of the Sale Price. One of the objectives of the paper was to check the important variable for the prediction of the price and how these variables describe the price. Through model training and evaluating its performance. ExtraTreesRegressor proved to be as best model. As the difference between the r2score and cross validation score was minimum. This project has increased my understanding of the concept. During the research I came across various challenges and while solving them I learned a lot of new things. For example. How to plot different charts. For example, I learned how to plot subplot. I learned new libraries and how to use them. I explored various methods for feature selection. Also, I came to understand how can multicollinearity can cause problem during the model training. The limitation of the solution provided is that the data carried a lot of unrealistic values. Apart from that my laptop took to much time while running certain command where I lost a lot of precious time.