PROJECT REPORT ON:

**“Car Price Prediction Project”**



**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analysing skills. And I want to express my huge gratitude to Shristi Maan (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

Who has inspired me in so many aspects and also encouraged me a lot with his valuable words and with his unconditional support I have ended up with a beautiful Project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today.. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

**Submitted by: - Ankit soran**

**Contents**

1. Introduction
   1. Business Problem Framing:
   2. Conceptual Background of the Domain Problem
   3. Review of Literature
   4. Motivation for the Problem Undertaken
2. Analytical Problem Framing
   1. Mathematical/ Analytical Modelling of the Problem
   2. Data Sources and their formats
   3. Data Pre-processing Done
   4. Hardware and Software Requirements and Tools Used
3. Data Analysis and Visualization
   1. Identification of possible problem-solving approaches (methods)
   2. Testing of Identified Approaches (Algorithms)
   3. Key Metrics for success in solving problem under consideration
   4. Visualization
   5. Run and Evaluate selected models
   6. Interpretation of the Results
4. Conclusion

**INTRODUCTION:**

1.1 Business Problem Framing:

Car price prediction is somehow interesting and popular problem. As per information that was gotten from Cars24 website total vehicles data we were able get were almost 2500 based on the data.

Accurate car price prediction is bit tricky involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, manufacturing year and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent changes in the price of a fuel. Different features like exterior colour, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this report, we applied different methods and techniques in order to achieve higher precision of the used car price prediction.

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.

* 1. Conceptual Background of the Domain Problem

The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling there are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it’s a big corner to pay too much or sell less than its market value.

There are one of the biggest target group that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what are the important features for a used car, then they may consider this knowledge and offer a better service.

* 1. Review of Literature

The second-hand car market has continued to expand even as the reduction in the market of new cars. According to the recent report on India’s pre-owned car market by Indian Blue Book, nearly 4 million used cars were purchased and sold in 2018-19. The second-hand car market has created the business for both buyers and sellers. Most of the people prefer to buy the used cars because of the affordable price and they can resell that again after some years of usage which may get some profit. The price of used cars depends on many factors like fuel type, colour, model, mileage, transmission, engine, and number of seats etc., the used cars price in the market will keep on changing. Thus the evaluation model to predict the price of the used cars is required.

* 1. Motivation for the Problem Undertaken

There are websites that offers an estimate value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car’s market value

**2. Analytical Problem Framing**

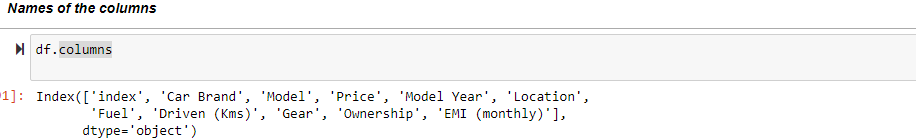
* 1. Mathematical/ Analytical Modelling of the Problem

As a first step I have scrapped the required data from Cars24 website. I have fetched data for different locations and saved it to excel format. In this particular problem I have car price as my target column and it was a continuous column. So clearly it is a regression problem and I have to use all regression algorithms while building the model. There was null values in the dataset. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 5% null values so I decided to drop those columns and some unwanted columns. If I keep those columns as it is, it will create high skewness in the model. Since we have scrapped the data from Cars24 website the raw data was not in the format, so we have use feature engineering to extract the required feature format. To get better insight on the features I have used plotting like pie plot, bar plot, count plot, strip plot and count plot. With these plotting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using z-score method. I have used all the regression algorithms while building model then tuned the best model and saved the best model. At last I have predicted the car-price using saved model.

2.2 Data Sources and their formats

The data was collected from Cars24.com website in excel format. The data was scrapped using selenium and Beautiful Soup. After scrapping required features the dataset is saved as excel file.

Also, my dataset was having 2817 rows and 10 columns including target. In this particular datasets I have object type of data which has been changed as per our analysis about the dataset. The information about features is as follows.



2.3 Data Pre-processing Done

* As a first step I have scrapped the required data using selenium and Beautiful soup

From Cars24 website.

* And I have imported required libraries and I have imported the dataset which was in excel format.
* Then I did all the statistical analysis like checking shape, nunique, value counts, info etc.
* While checking for null values I found null values in the dataset and removed them as they were very less.
* I have also dropped Unnamed: 0.
* Next as a part of feature extraction I converted the data types of all the columns and I have extracted useful information from the raw dataset. Thinking that this data will help us more than raw data.
  1. Hardware and Software Requirements and Tools Used
* While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

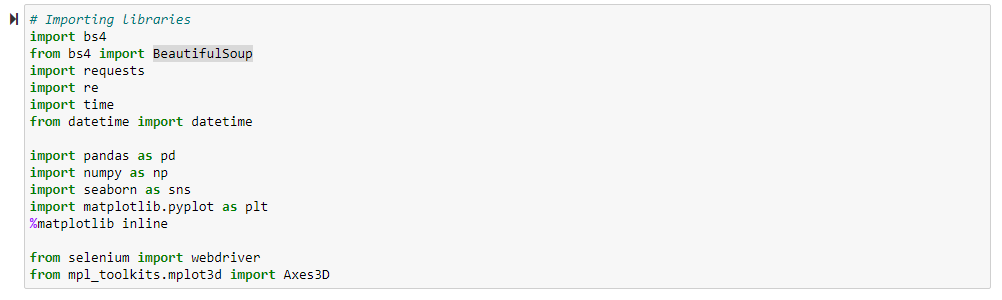
**Hardware required**: -

* 1. Processor — core i5 and above
* 2. RAM — 8 GB or above
* 3. SSD — 250GB or above

**Software/s required**: -

* 1.Anaconda

**Libraries** **required:-**

****

**3. Data Analysis and Visualization**

3.1 Identification of possible problem-solving approaches (methods)

Since the data collected was not in the format we have to clean it and bring it to the proper format for our analysis. To remove outliers I have used z-score method. We have dropped all the unnecessary columns in the dataset according to our understanding. We have checked for correlation between dependent and independent features. Also I have used Standardisation to scale the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms.

* 1. Testing of Identified Approaches (Algorithms)

Since car price was my target and it was a continuous column with improper format which has to be changed to continuous float data type column, so this particular problem was Regression problem. And I have used all Regression algorithms to build my model. By looking into the difference of r2 score and cross validation score I found DecisionTreeRegressor as a best model with least difference. Also to get the best model we have to run through multiple models and to avoid the confusion of over fitting we have go through cross validation. Below are the list of Regression algorithms I have used in my project.

* Lasso
* Ridge
* KNeighborsRegressor
* Decision Tree Regressor
* Gradient boosting Regressor
  1. Key Metrics for success in solving problem under consideration

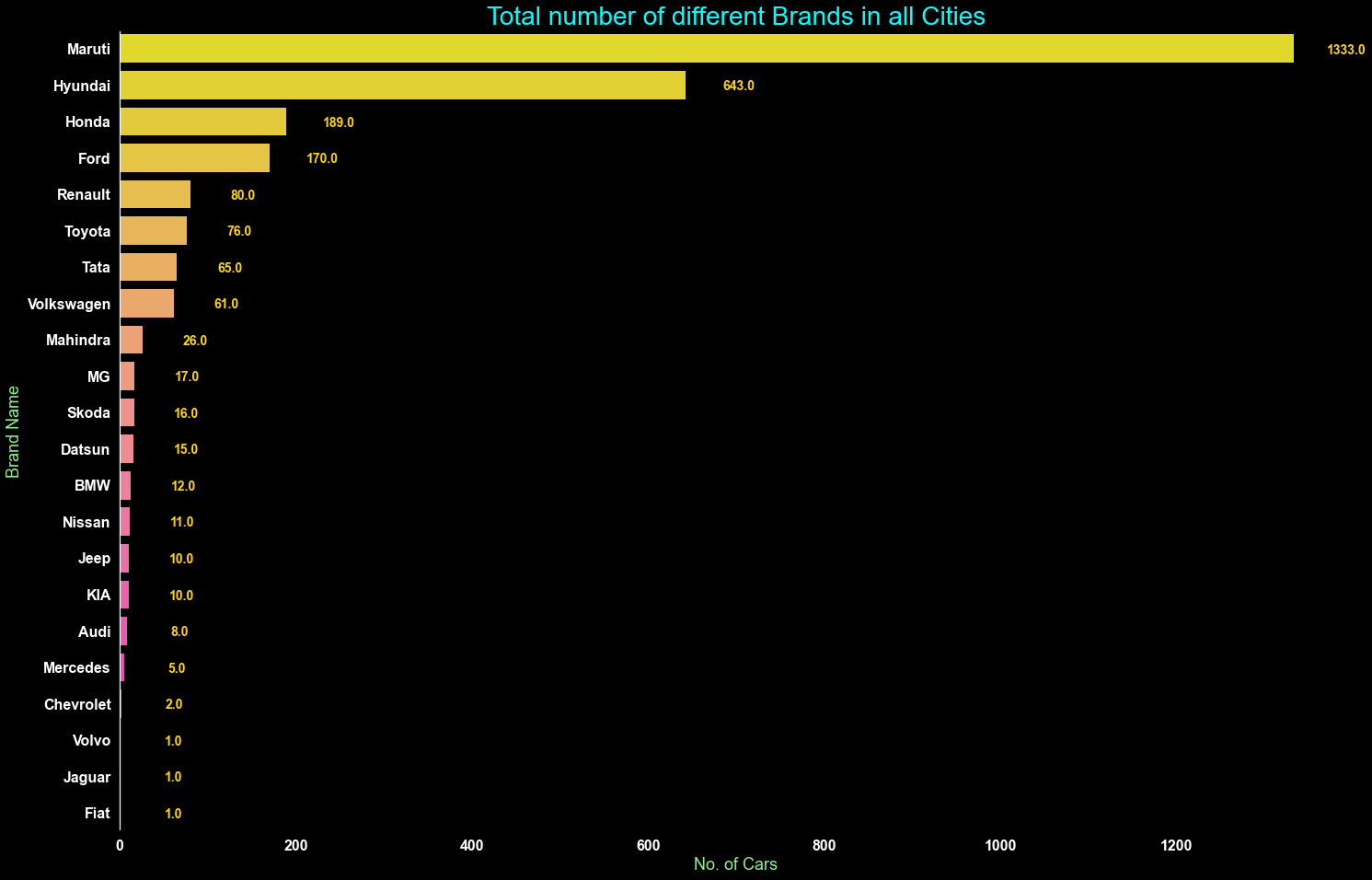
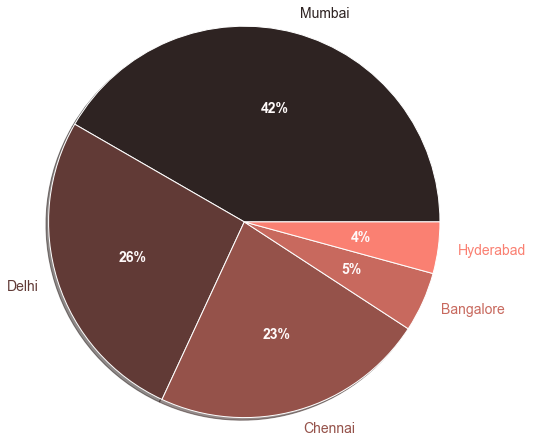
I have used the following metrics for evaluation:

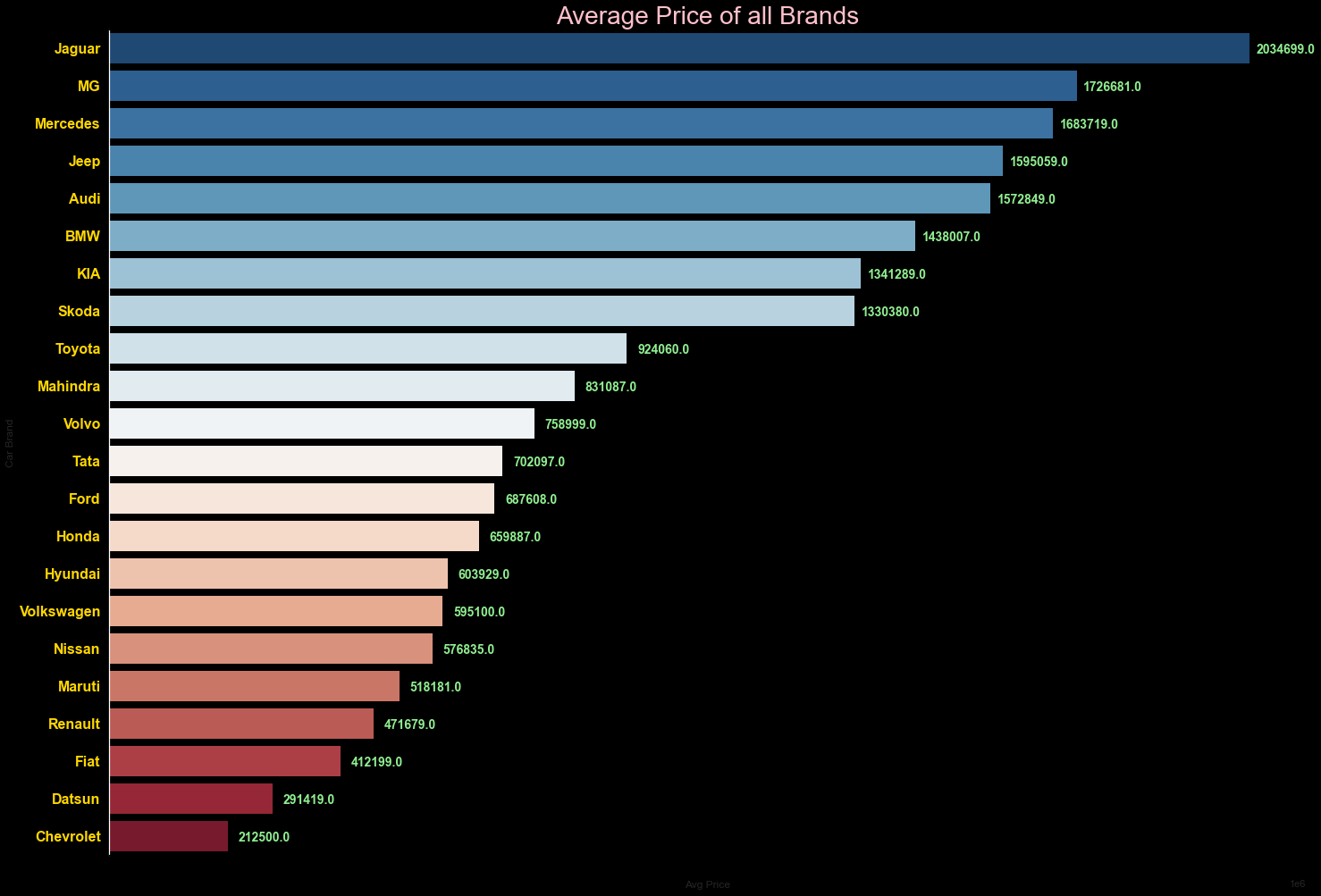
* I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
* I have used root mean square is one of the most commonly used measures for evaluating the quality of predictions.
* I have used r2 score which tells us how accurate our model is.

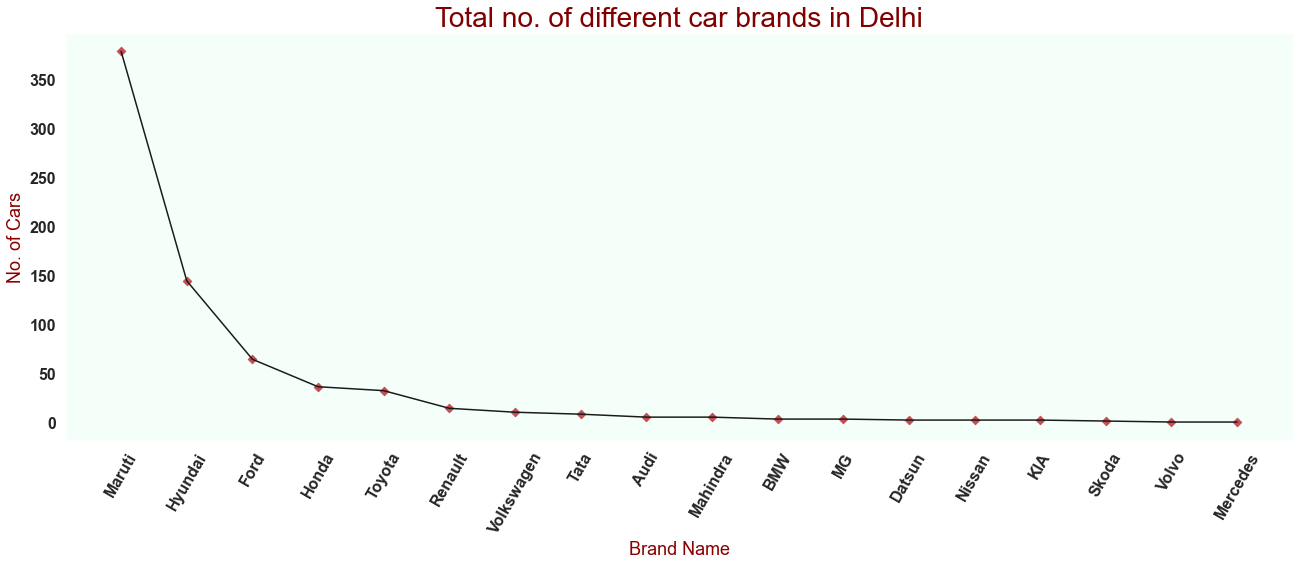
3.4 Visualizations

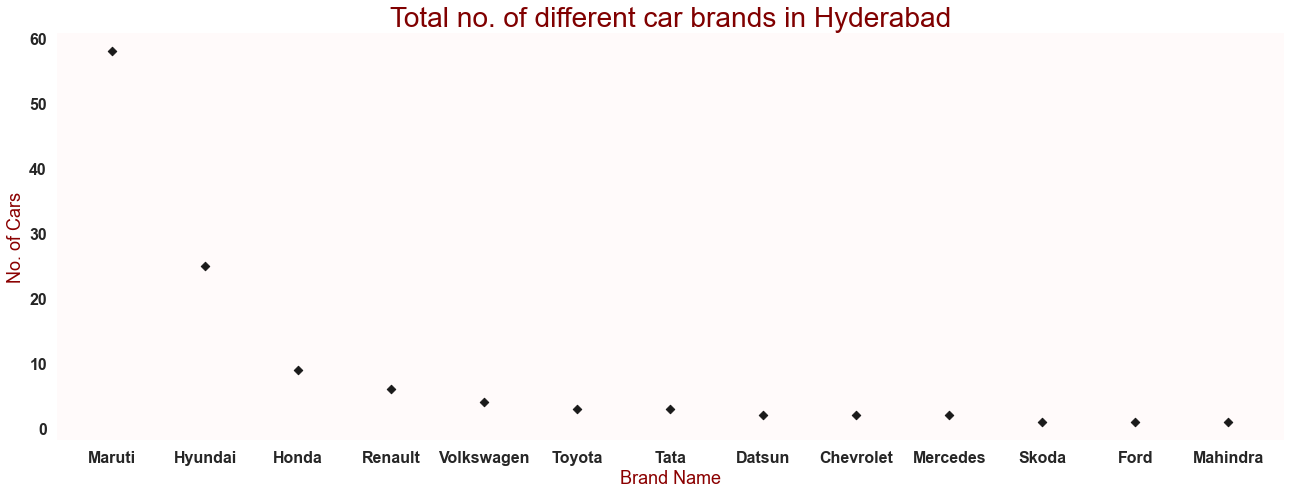
I have used bar plots to see the relation of categorical feature with target.

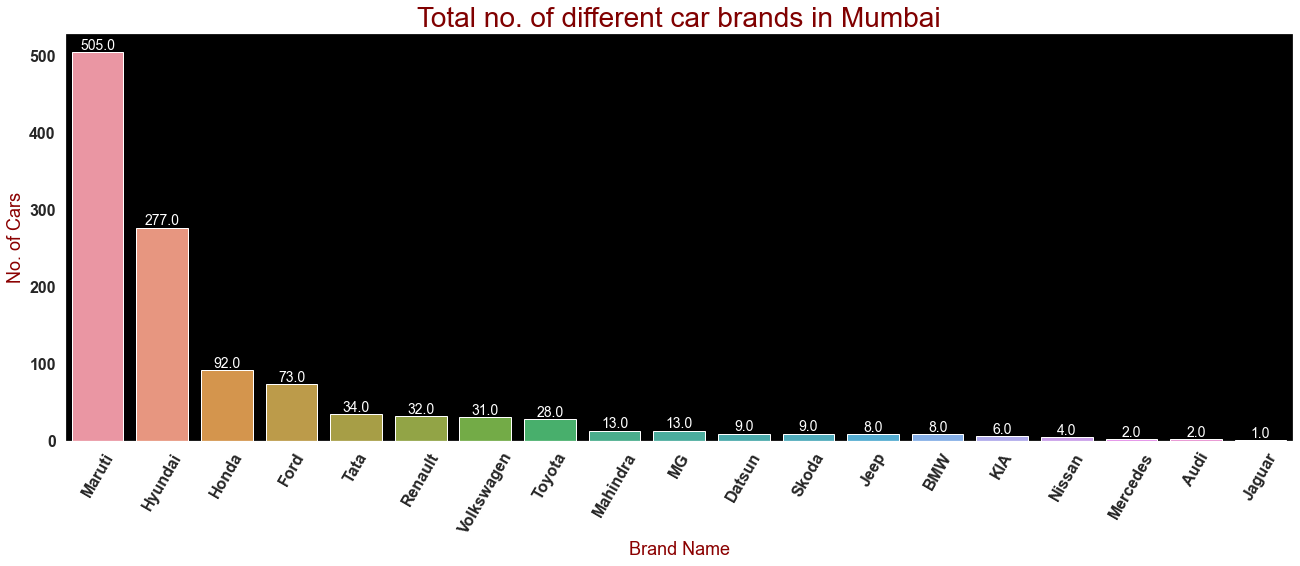
1. **Univariate Analysis for Categorical columns:**

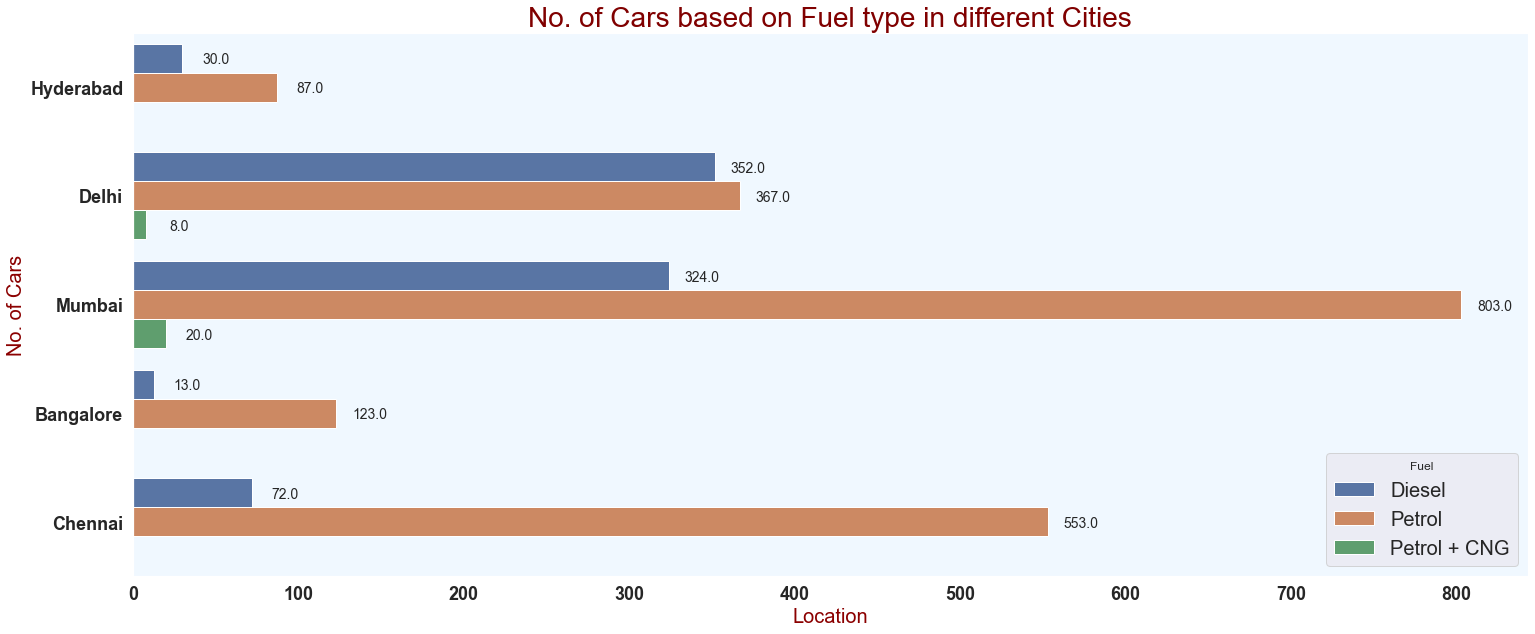
****

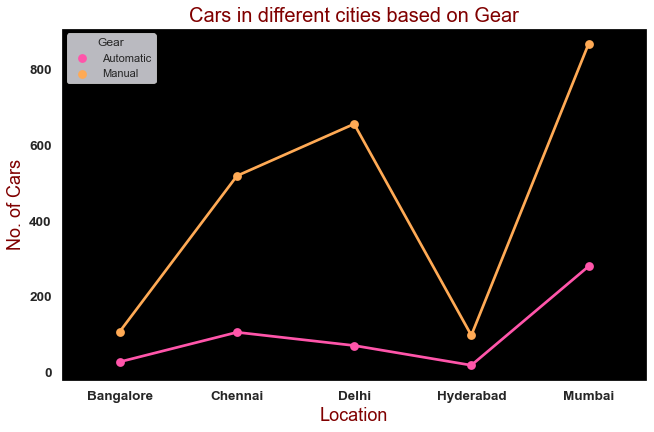
****

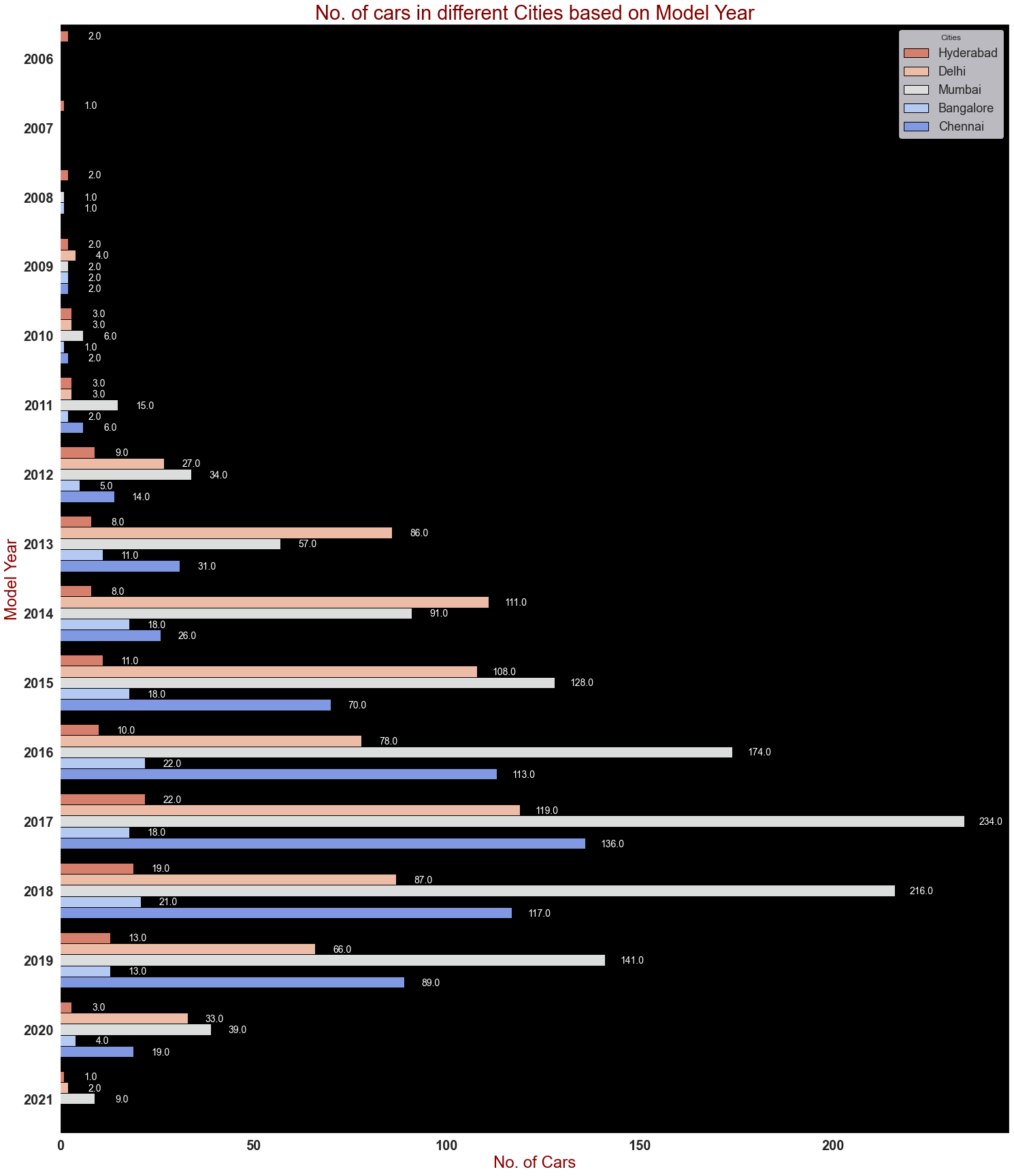








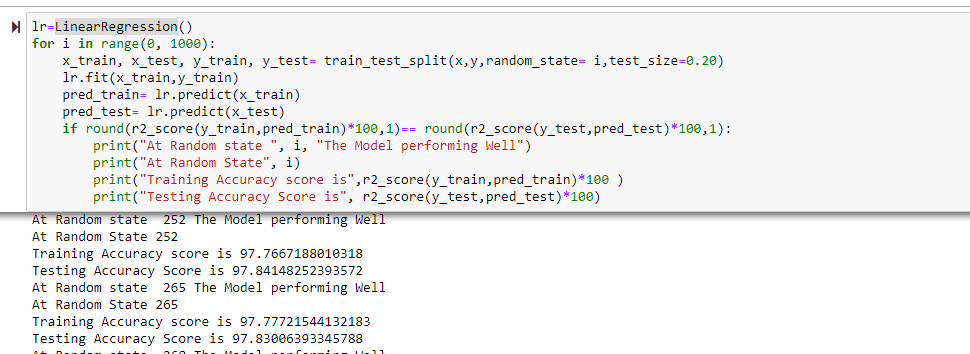


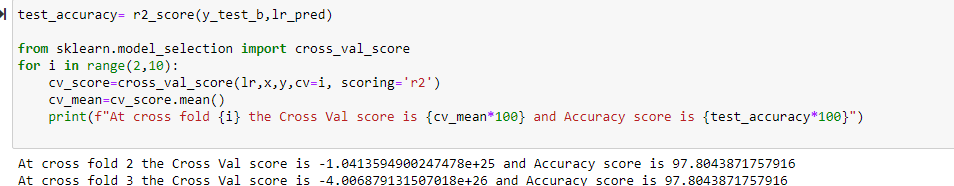


**Observations:**

* Maximum cars are petrol driven and also diesel driven.
* Maximum cars are with Manual gear transmission.
* Maximum cars under sale are Maruti followed by Hyundai.
* Most cars are available from year '2014' to '2019'.
* In Bangalore, Delhi-NCR, Mumbai and New-Delhi we can find maximum cars for sale. Since these are most populated places.

**Model Building:**

****

****

**Lasso**

****

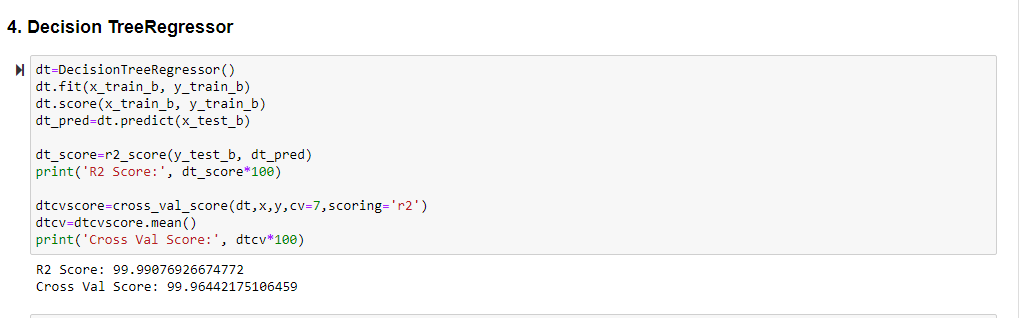
**Ridge**

****

**KNeighbors Regressor**

****

**Decision Tree Regressor**

****

**Gradient boosting Regressor**

****

**I have chosen all parameters of DecisionTreeRegressor, after tuning the model with best parameters I have my model accuracy 99.79%.**

* **Saving the model and Predictions:**
* I have saved my best model using .pkl as follows**.**
* Now loading my saved model and predicting the price values.



**Conclusion**

In this project report, we have used machine learning algorithms to predict the used car prices. We have mentioned the step by step procedure to analyse the dataset and finding the correlation between the features. Thus we can select the features which are correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a hyper parameter tuning was done to the best model and the accuracy has been improved. Hence we calculated the performance of each model using different performance metrics and compared them based on those metrics. Then we have also saved the best model and predicted the used car price. It was good the predicted and actual values were almost same.

THANKYOU