Car Price Prediction Project

Submitted by:

Parth Rupavatiya

ACKNOWLEDGMENT

- I would like to express my special thanks to my SME Rashi Mathur, who gave me the golden opportunity to do this wonderful project on the topic used car price prediction.
- Some of the articles and research papers, I find useful for completion of this project:

REFERENCES:

1) Praful Rane, Deep Pandya, Dhawal Kotak, “USED CAR PRICE PREDICTION”, IRJET-2021.

INTRODUCTION

Business Problem Framing:

- The price of a new car 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 prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase.
- Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features.
- Due to covid 19 impact in the market, some cars are in demand hence making them costly and some are not in demand hence cheaper. Therefore, our client is facing problems with their previous car price valuation machine learning models.
- To overcome this problem we have developed a machine learning model which will be highly effective.

Conceptual Background of the Domain Problem:

- In this project, we scraped data of used cars from different website like olx, cars24, droom etc. After that we will build a machine learning model to predict the price of used cars.
- This model will then be used by our client to understand how exactly the used car prices vary after covid-19. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Motivation for the Problem Undertaken:

As we know, used car industries are very large market and there are various companies working in the domain. We take this project because data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in used car sales and purchases.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem:

- We use some mathematical, statistical and analytics approaches in this project. Which is described below:

1) **Z-score:** Outliers are extreme values that fall a long way outside of the other observations. In this project we used zscore method to remove outliers. In this procedure we calculate the z-score for each observation. Any z-score greater than 3 or less than -3 is considered to be an outlier. This rule of thumb is based on the empirical rule. From this rule we see that almost all of the data (99.7%) should be within three standard deviations from the mean. To calculate the z-score, we subtract the mean from the data point, and then divide by our standard deviation.

2) **One hot encoding:** One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. One hot encoding is useful for data that has no relationship to each other.
3) **Standard scaler:** we use standard scaler method to scale our data. This method normalizes our data and essential for machine learning algorithms that calculate distance between data. For instance, most of the classifiers calculate the distance between two points by the distance. If one of the features has large value, then distance consider that particular feature. This method is necessary, where large and small values present in our data. This method transform our data with mean = 0 and standard deviation = 1.

Data Sources and their formats:

- We scrape used car data from different websites like droom, quikr, cars24 etc. The data contains various features of used cars from different locations and we need to do analysis of that data and make machine learning model.
- This dataset contains 5319 records and 8 features.

Data Pre-processing Done:

For cleaning or pre-processing the data we use some techniques which are described below:

- First, we check null values present in the data and find only one null value in fuel column, so we dropped that row. We also dropped unnamed 0 column because it represents only index of the data and not useful for model.

- Next, we convert all dataset into uppercase and remove extra trailing edges of location and driven kilometres data.
- Furthermore, we remove rows of location whose frequency is less than 3 because those locations do not impact our model and also rename some locations.
- After that, we remove rows of brand feature whose frequency is less than 4 because those Brands do not impact our model and also rename some brand names.
- Then, we remove rows of manufacturing year and fuel features, whose frequency is less than 2 because those data is wrong and not related to that particular column and also rename some data of both columns.
- Then, we remove comma (“,”) from the driven kilometres and car price features and then convert that features into integer and float data type respectively. We also convert manufacturing year feature into integer data type.
- At last, in car price feature some car owners like to ask for price to buyer and instead to give car price they stated ask for price. So, we convert that type of data with mean of car prices.
- After analysis of the data, we remove outliers using z-score method, and encoding categorical data into numerical data using one hot encoding method.
- At last, we scale data with mean = 0 and standard deviation = 1 using standard scaler method.

Data Inputs- Logic- Output Relationships:

- We used regression machine learning models because our target variable is car price. So, we need to find price of used cars.
- There are many regression models but here we used some of them models.
- First we split our training dataset into two segments: training and testing. We take 85% data for training and 15% data for testing. For splitting data we use train test split method. Below is the code for splitting the data:

```
# split train and test data from dataset.
from sklearn.model_selection import train_test_split,cross_val_score
x_train,x_test,y_train,y_test = train_test_split(x_s,y,test_size=0.15,random_state=0)
```

1) 85% of the observation as training set--> x_train
2) The associated target for each observation in x_train --> y_train
3) 15% of the observation as test set--> x_test
4) The target associated with the test set--> y_test.

- After splitting data we passed training data to machine learning models. The fitted model will first be used to generate prediction on the test set (x_test). Next, the predicted class labels are compared to the actual observed class label (y_test) to see the difference between them.

Libraries Used:

- We used many libraries used in this project, which is described below:
  1) Numpy: This library is used for scientific computing. It supports multidimensional arrays and matrices.
  2) Pandas: This library is used for data analysis and modelling convenient in python. Pandas simplify analysis by converting CSV, JSON, and TSV data files or a SQL database into a data frame with rows and columns.
  3) Matplotlib and Seaborn: Both libraries are used for data visualization.
  4) Scikit-learn: The Python library, Scikit-Learn, is built on top of the Matplotlib, Numpy, and SciPy libraries. It has wide range of algorithms.
  5) SciPy: SciPy stands for Scientific Python. It is a scientific computation library that uses Numpy underneath. It also provides more utility functions for optimization, stats and signal processing.

Models Development and Evaluation

Identification of possible problem-solving approaches (methods):

1) Data reading and understanding
2) Data cleaning
3) Data analysis
4) Handling outliers
5) Handling skewness
6) Encoding data
7) Scaling
8) Train test split
9) Machine learning algorithms

Testing of Identified Approaches (Algorithms):

The regression algorithm that we used is:

1) Linear regression
2) K-Neighbors regressor
3) Decision tree regressor
4) Extra tree regressor
5) Lasso regression
6) Ridge regression
7) Random forest regressor
8) Adaboost regressor
9) Gradient boosting regressor
10) Xgboost regressor
11) Catboost regressor

Building machine learning models:

- We use many algorithms to find best model, but here we describe only best model.
- We find catboost regressor as a best model. Catboost builds upon the theory of decision trees and gradient boosting. The main idea of boosting is to sequentially combine many weak models and thus through greedy search create a strong competitive predictive model.
- Below is the code of our model with catboost regressor with hyperparameter tuning:

```
cat = CatBoostRegressor(max_depth=7,iterations=1000,learning_rate=0.2)
cat.fit(x_train,y_train)
train=cat.score(x_train,y_train)
pred_cat=cat.predict(x_test)
print("Model training accuracy:",train)
print("r2_score:",r2_score(y_test,pred_cat))
print("mean absolute error:",mean_absolute_error(y_test,pred_cat))
print("root mean squared error", np.sqrt(mean_squared_error(y_test,pred_cat)))
```

Output:

```
Model training accuracy: 0.9534057499186895
r2_score: 0.7997998450112115
mean absolute error: 133134.49767846425
root mean squared error 248397.30594318893
```

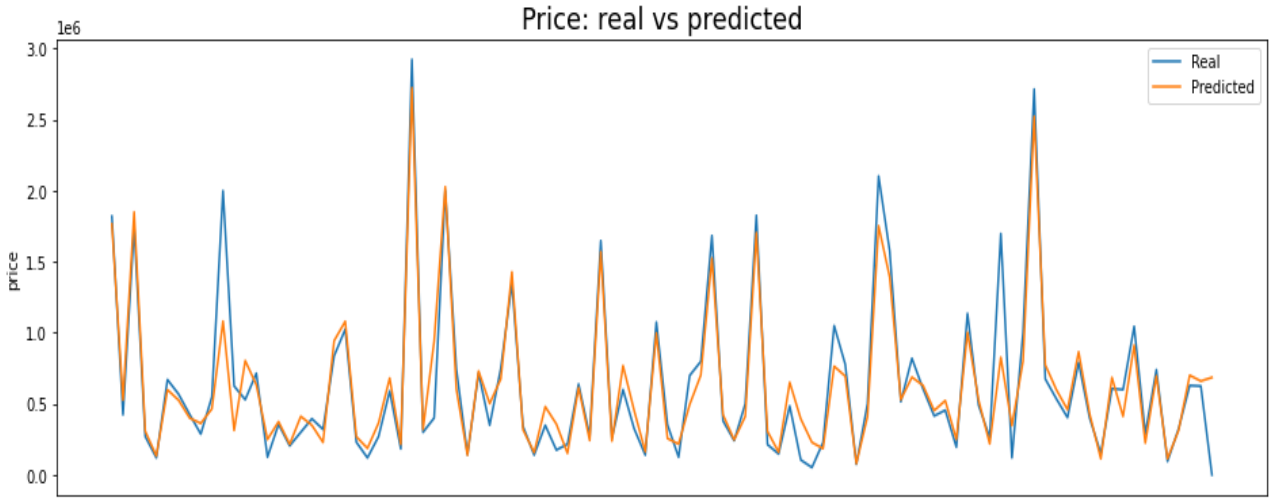
- We get 95% training model accuracy and 79% test data r2_score accuracy; also we get good mean absolute error (133134.49) among all other algorithms. Now, we have to confirm that model is not going through underfitting or overfitting.
- So, we check catboost regressor training model accuracy using cross validation to confirm that our model is not going through underfitting or overfitting. Below is the code for cross validation:

```
# check accuracy of catboost model with cross-validation
accuracy = cross_val_score(cat,x_s,y,cv=4,scoring="r2")
print(accuracy)
print("Accuracy of Model with Cross Validation is:",accuracy.mean() * 100)
```

- Here, we use CV = 4, that means our training set is divided into 4 parts and provide mean accuracy of those 4 parts.
  Output:

```
[0.60289665 0.66505246 0.67902011 0.69479964]
Accuracy of Model with Cross Validation is: 66.04422155078996
```

- We get 66% r2_score accuracy using cross validation that means our model is not underfitted or overfitted.
- Now, let's see the how the predicted values are comparing to real values for catboost regressor in graphical format:



- Now, check accuracy of all used algorithms:

Algorithms	Model training accuracy	R2 score	Mean absolute error	Root mean squared error
Linear regression	0.81	-3.26	4.36	1.00
K-Neighbors regressor	0.46	0.14	270426.10	513030.76
Decision tree regressor	0.99	0.63	149755.71	333633.15
Extra tree regressor	0.99	0.71	139489.22	294895.80

Lasso regression(HP)	0.97	0.81	122474.07	237659.30
Ridge regression(HP)	0.97	0.74	134788.41	278163.67
Random forest regressor(HP)	0.70	0.60	206434.44	350512.49
Adaboost regressor	0.43	0.42	310006.05	420023.63
Gradient boosting regressor(HP)	0.83	0.75	165784.24	246410.63
Xgboost regressor(HP)	0.87	0.78	153156.12	258801.35
Catboost regressor(HP)	0.95	0.79	133134.49	248397.30

Note: In above table HP means with Hyperparameter Tuning.

Key Metrics for success in solving problem under Consideration:

- We use three types of metrics for solving problem. Which is described below:
    1) R2_score: R2 score is the percentage of variation explained by the relationship between two variables. Range of the r2 score is varies from 0 to 1. Mathematical formula of the r2 score is as below:

$$R^2 = 1 - SS_{res} / SS_{tot}$$

Where,

SS_{res} is the sum of squares of the residual errors.

SS_{tot} is the total sum of the errors.

2) Mean absolute error (MAE): The Mean Absolute Error, also known as MAE, is one of the many metrics for summarizing and assessing the quality of a machine learning model.
   - Mean absolute error subtract the predicted value from actual value as below:

     Prediction Error → Actual Value - Predicted Value
   - This prediction error is taking for each record after which we convert all error to positive. This is achieved by taking Absolute value for each error as below:

     Absolute Error → |Prediction Error|
   - Finally we calculate the mean for all recorded absolute errors (Average sum of all absolute errors). Below is the formula of MAE:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

3) Root mean squared error (RMSE): Root Mean Square Error is the measure of how well a regression line fits the data points. RMSE can also be construed as Standard Deviation in the residuals.

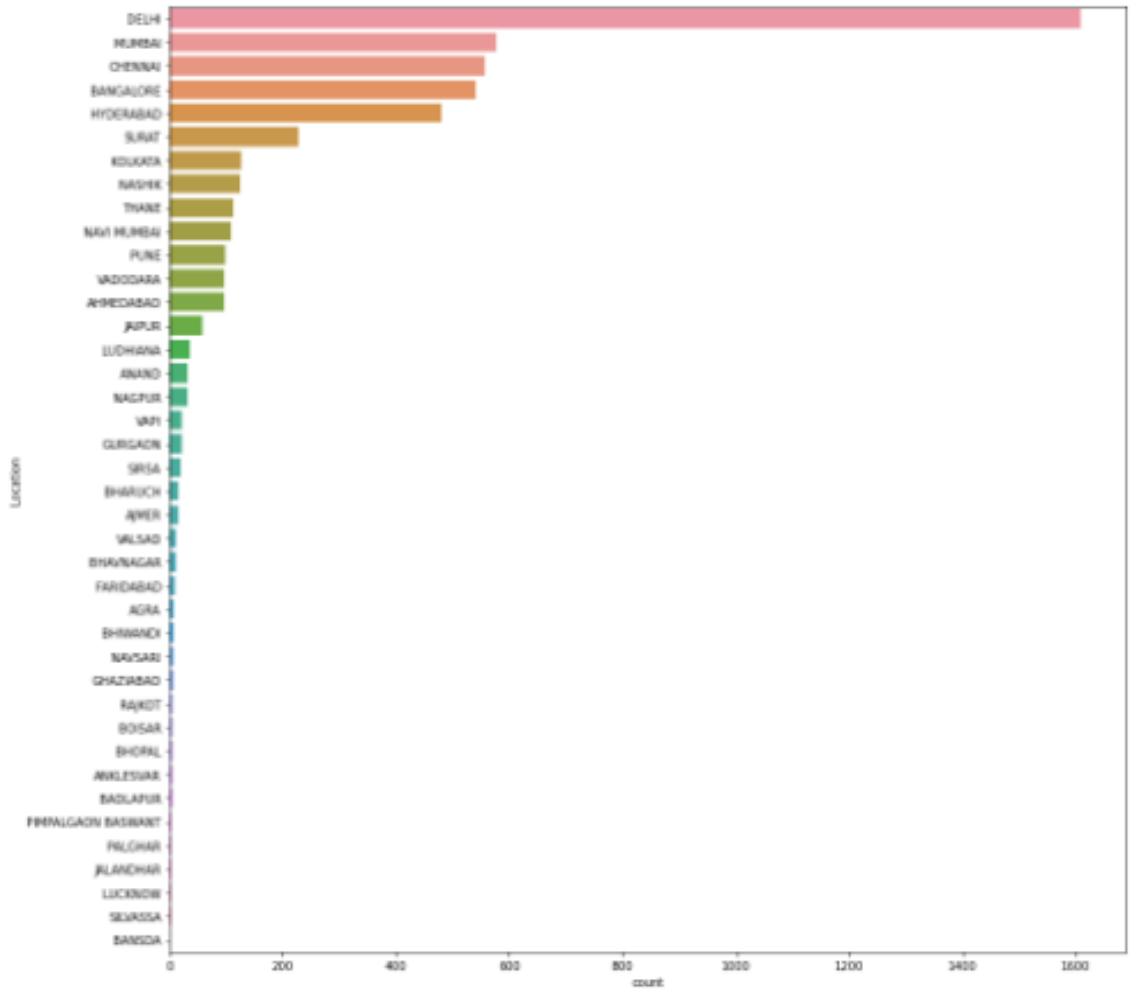
   - It is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

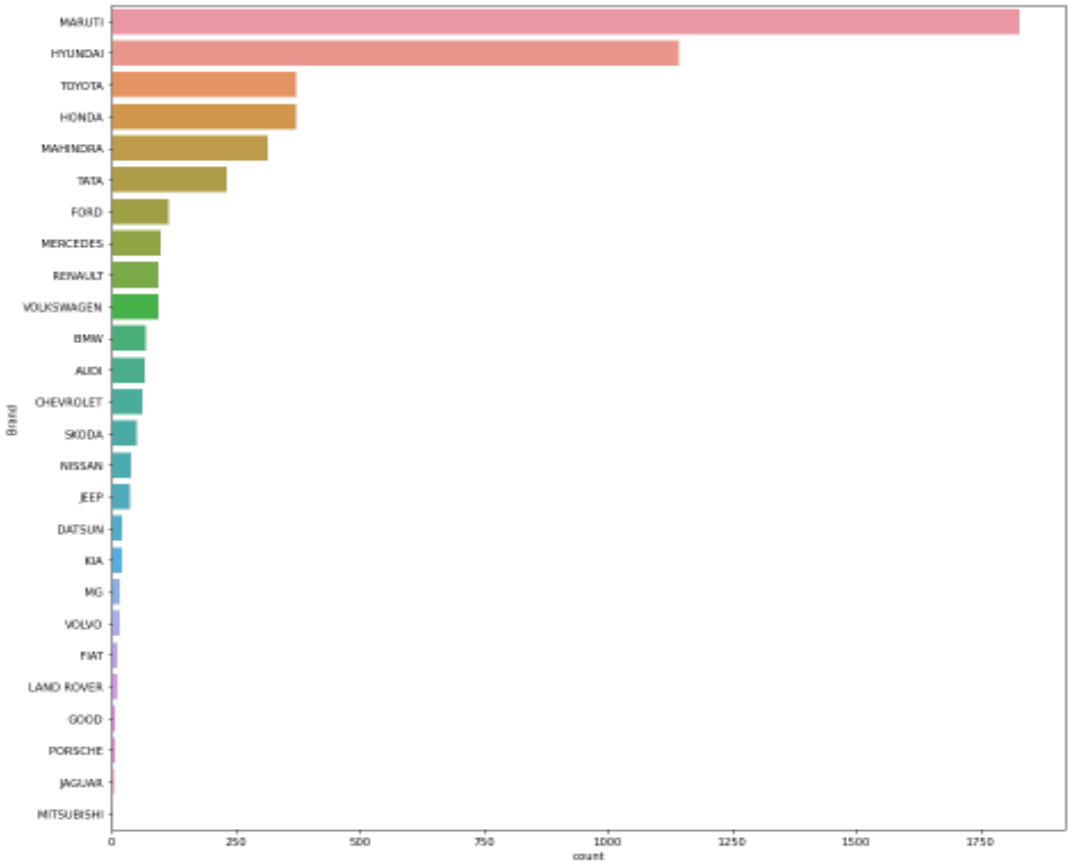
   - ŷ1, ŷ2,… ŷn are predicted values
   - y1,y2,…yn are observed values
   - n is the number of observations

Visualizations:

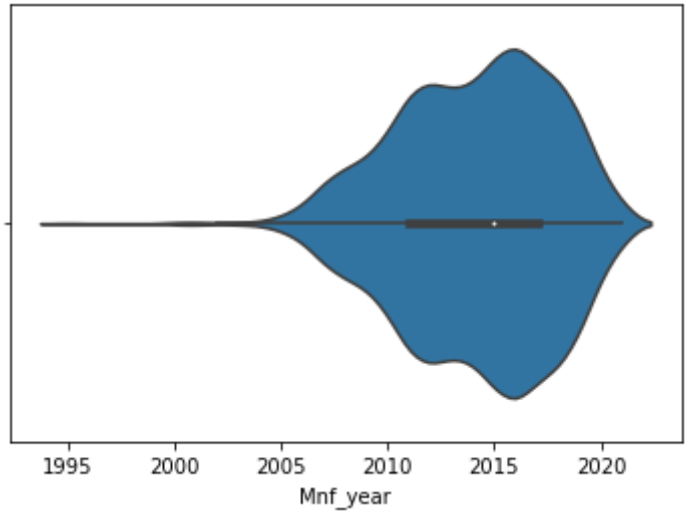
- Now, let's look at the analysis of the data.



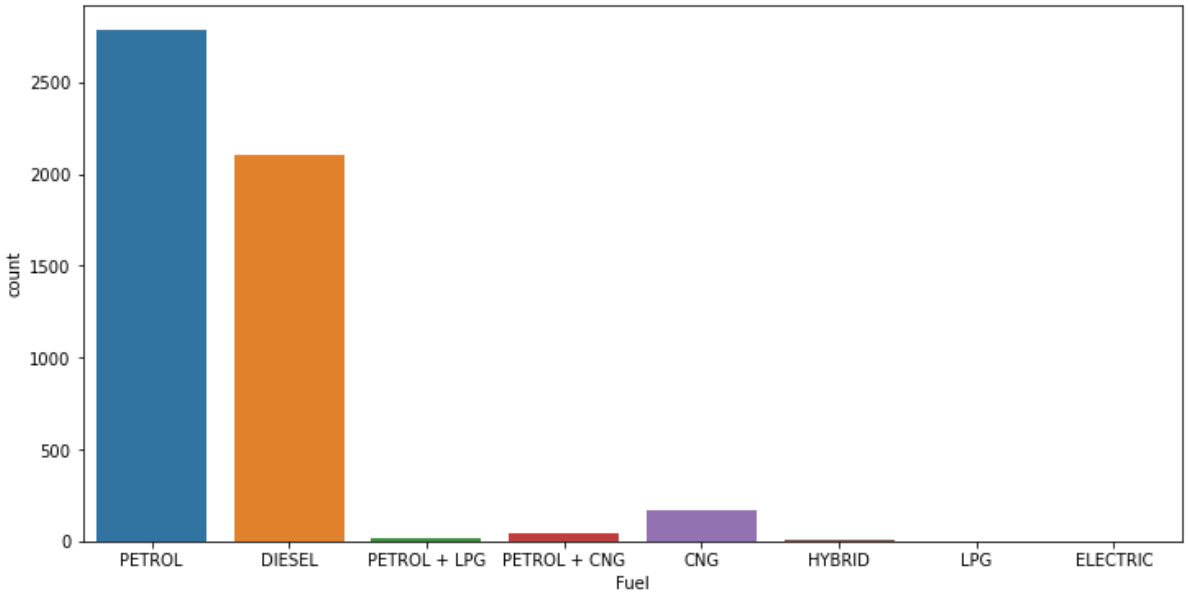
From the above graph we can see that, most of the used cars are from Delhi, Mumbai, Chennai and Bangalore in our dataset.



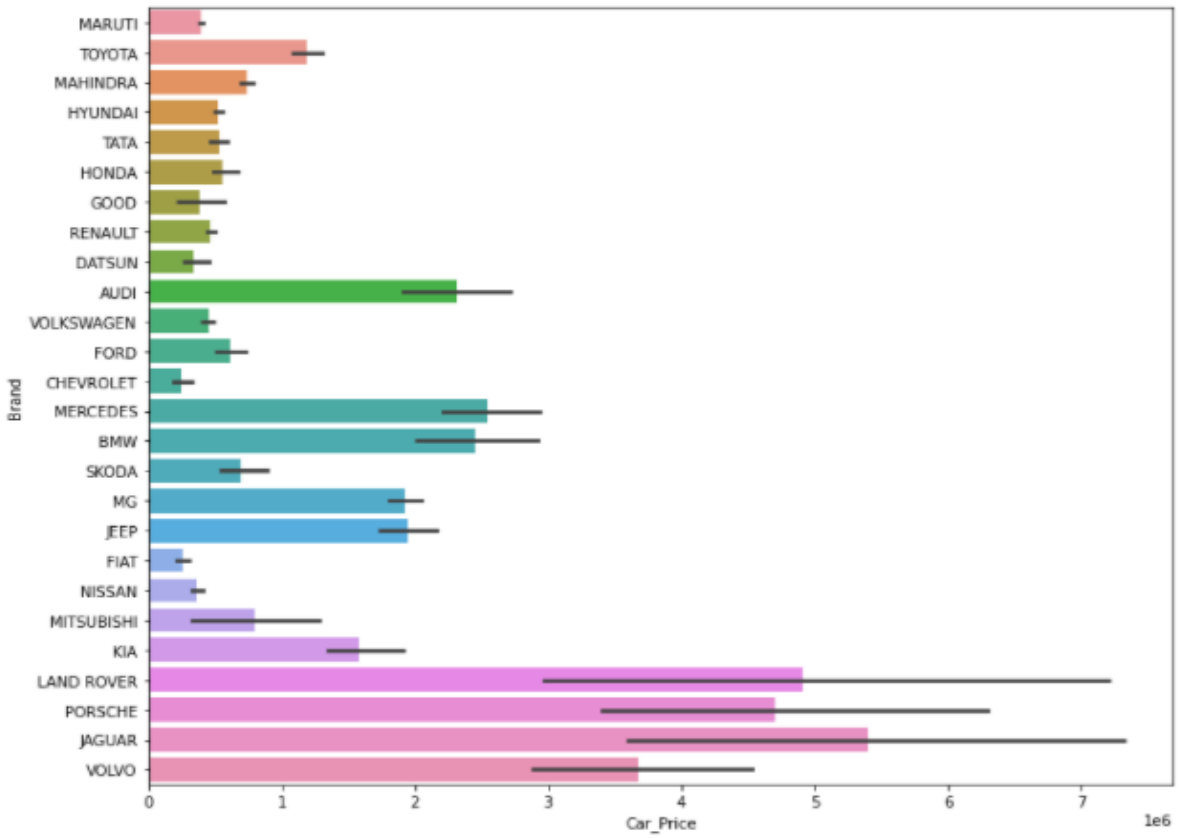
Majority of the car is from Maruti Company, while lest number of cars is from Mitsubishi Company in our dataset.



From the above violin plot we can see that, most of the cars manufacturing year is between 2011 to 2017.

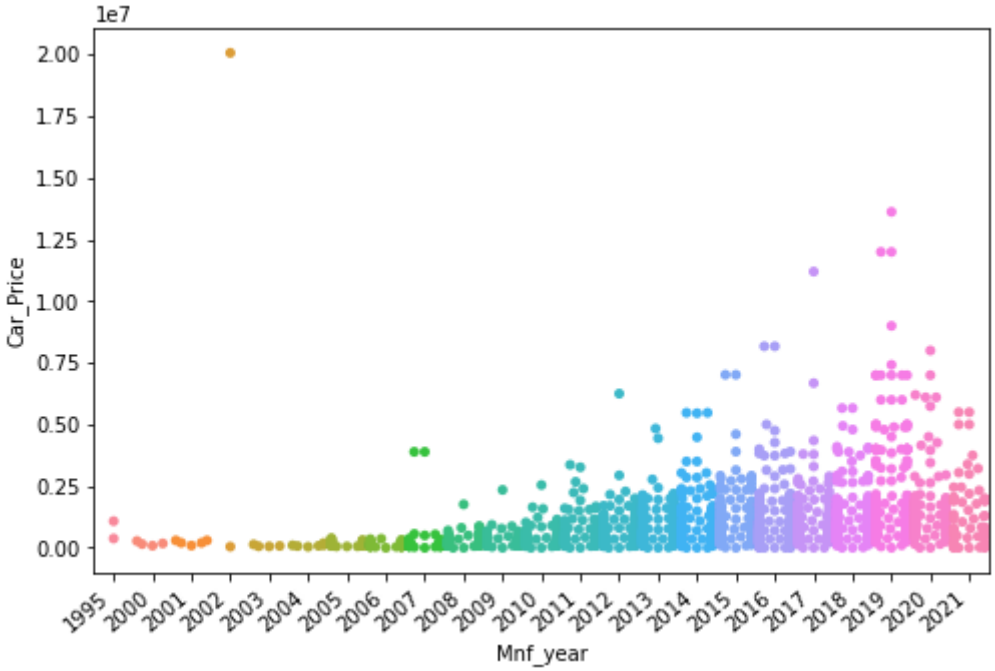


From the above bar plot we can see that, majority number of cars is run by petrol and diesel, while least number of cars is run by electric power.

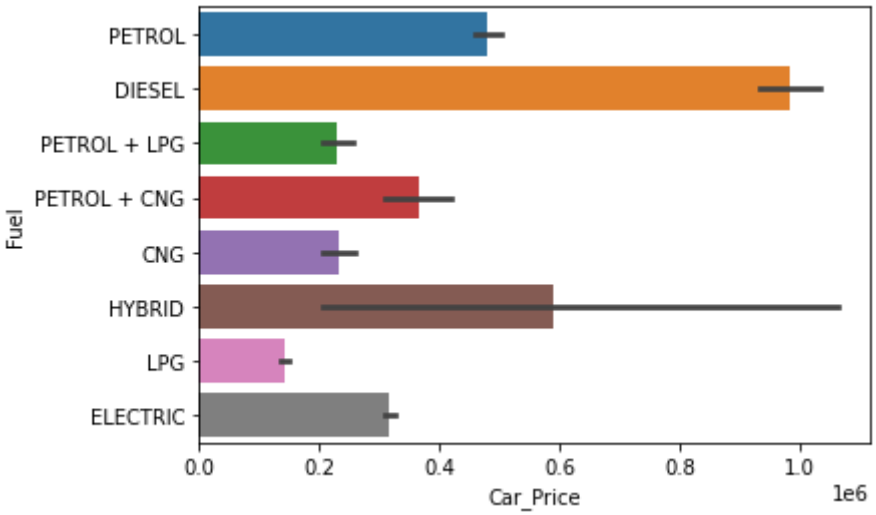


From the above bar plot we can see that, jaguar companies used car have maximum price compare to other companies car. While

Chevrolet and fiat companies used car have low price among other entities.

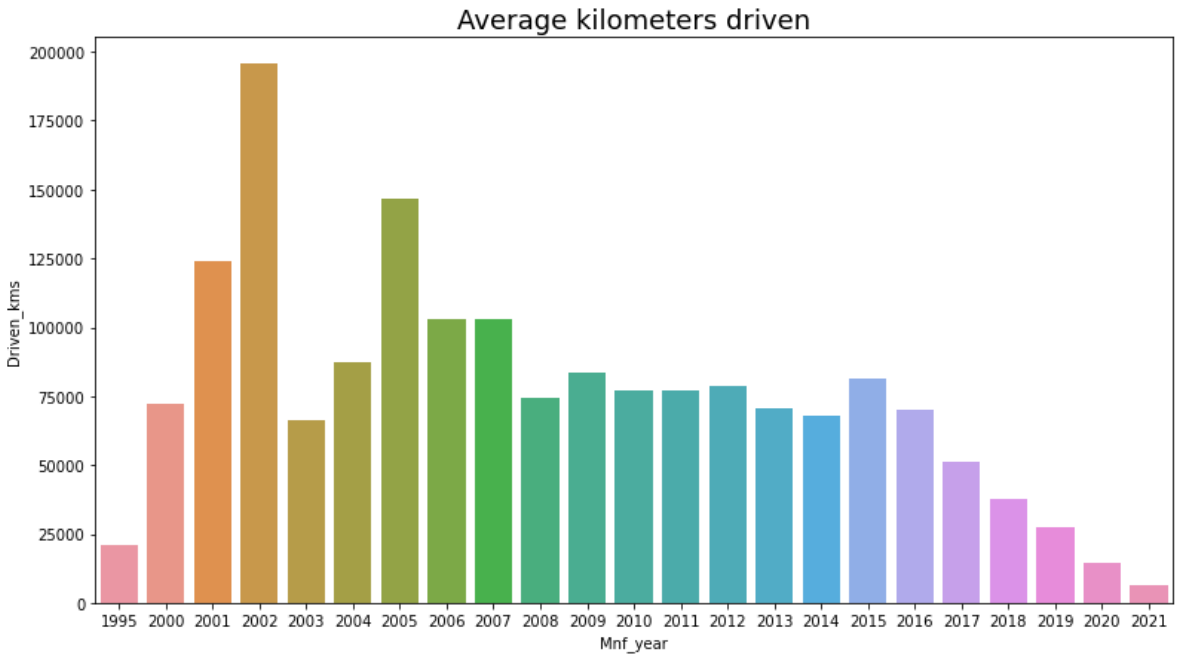


From the above swarm plot we can see that, car price is gradually increases over the years.

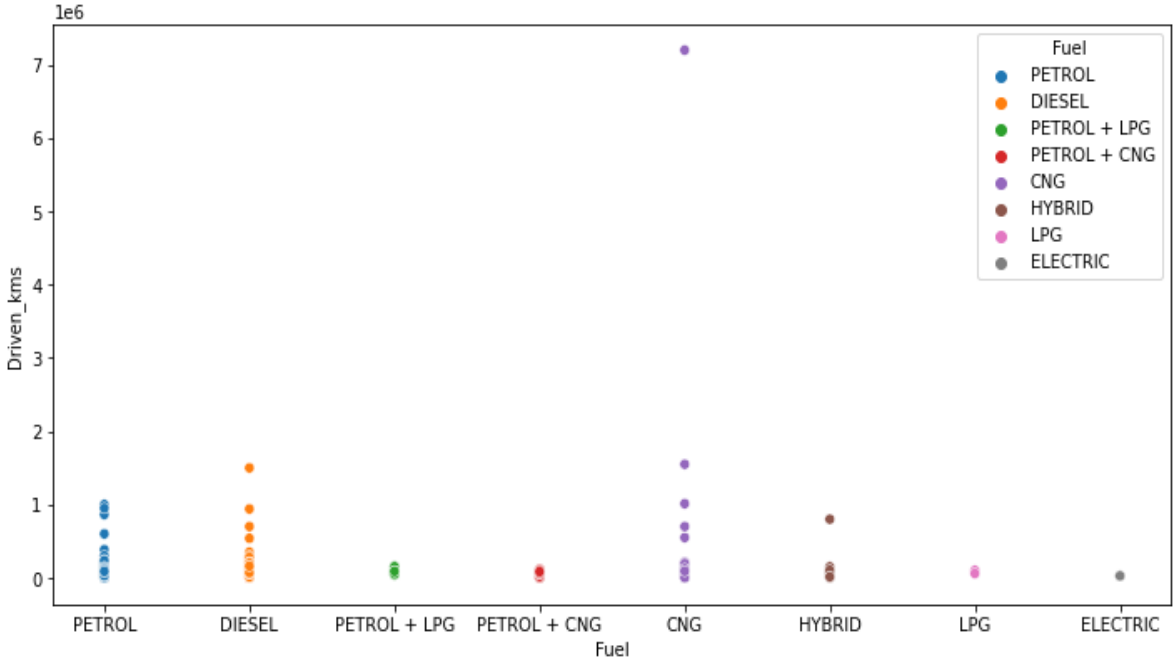


- From the above bar plot we can see that, diesel cars price is maximum, while LPG cars price is lowest compare to other categories.
- Petrol+LPG and CNG cars price is almost similar.

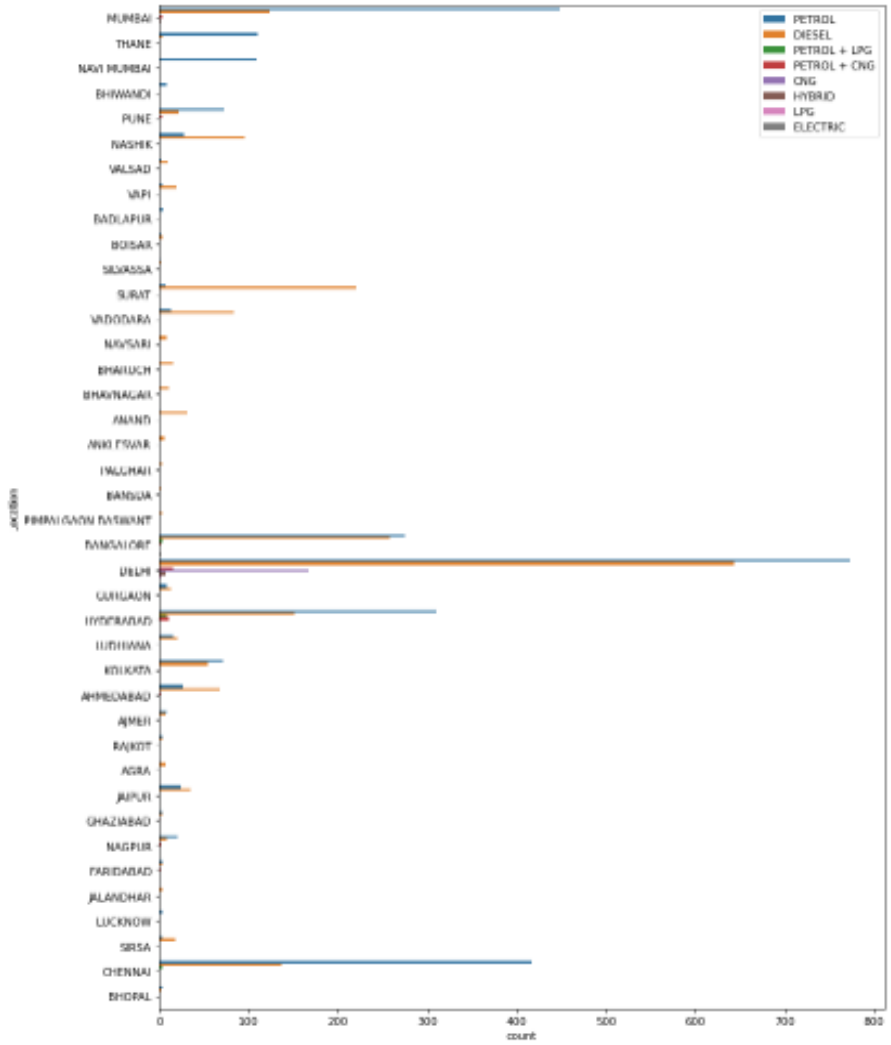
- Hybrid cars price is second highest among all other categories.



- From the above bar plot we can see that, average kilometres driven rises up from year 1995 until 2002 and after 2015 it is linearly goes down until 2021.
- There is not much difference in average kilometres driven between years 2008 to 2015.



The above scatter plot shows that more number of petrol, diesel and CNG cars seems to give more kilometres driven than other fuel type cars.



From the above graph we can see that, in majority of the locations petrol and diesel are most used fuel type in used cars.

CONCLUSION

Key Findings and Conclusions of the Study:

- The purpose of this article was twofold: to understand the pattern of used cars market and make predictive model, which is able to effectively predict the price of used cars.
- We use many algorithms to find best model and best result were observed of the catboost regressor with 79% r2 score accuracy with good mean absolute error.
- There are many variables important to predict the price of houses. Like driven kilometres, car model, fuel of car etc.
- By using machine learning model our client can decide whether to increase or decrease the price of used cars.

Scope of future work:

- In future we may add large historical data of car price which can help to improve accuracy of the machine learning model.
- For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.