FLIP ROBO

House Price Prediction

Submitted by:

Prateek AP

Acknowledgment

Data Credit's: Carwalle.com

Would like to thank the FlipRobo team, especially Sapna for all the extensive assistance through the projects.

INTRODUCTION

Business Problem:

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.

Conceptual Background of the Domain Problem

India being a price sensitive economy, getting maximum value for the money is the primary consideration for any buyer and sellers need to have the right balance between price quoted and the profits to be competitive as well as sustainable in the highly populated used car space.

The prices for used cars can differ for a variety of factors, and the customer experience is quite different than buying a new vehicle.

Motivation for the Problem Undertaken:

Since the automobile services are providing a way to the transportation, the problem being looked at is to predict a price value of the used cars. Since the covid season may impact the business in either way profitable or loss for the automotive Industry. Hence our client is expecting us to build a model based on the current situation to predict the car price nearer to the exact value. The main objective of our client is to maximize the profit by knowing the actual price of the already used cars. The main motivation behind this initiation is to help our client get some profit out of selling price.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

On analysis of the scrapped data, Dataset has majority of the features which are categorical and the target variable as continuous values. Which makes it a regression problem.

We have very few features which have predominantly right skewness. We will first interpret the data for our Analysis by getting some statistical parameters like mean, median & quantiles which decides the factors that will help predict the target variable. We will apply the formula to identify & delete the outliers using Zscore method for our better model building.

- ## Data Sources and their formats:

Data set is scrapped by from website carwale.com.

The car details are scrapped is according to different cities Bangalore, Kolkata, Chennai, Mumbai, Ahmedabad, Hyderabad, Pune, Delhi. The data was scrapped based on the cities and had 8 data Frames corresponding to each city. Which was then compiled into 1 single dataset in csv format

Data Preprocessing Done:

1. Loading Dataset
2. Dropping duplicate values
3. Feature Engineering, extracting years old from Year
4. Shape of our dataset having rows 18,340 and 10 columns
5. Looking at statistical parameters
6. No Null values in our dataset.
7. Categorical values are labelled using LabelEncoder
8. Removing Outliers
9. Removing skewness
10. Showing correlation with each other features and dropping City.

Data Inputs- Logic- Output Relationships:

The input data provided, helps to understand the car specifications, independent features which is responsible to find the car price which is the target dependent variables. Any change in the individual features may change the target variables.

State the set of assumptions (if any) related to the problem under consideration:

We have scrapped the data for 8 cities, but while doing so we have some cities which are not a part of the 8 cities we intended but from same State. We have merged the unintended cities with the 8 cities based on location.

Hardware and Software Requirements and Tools Used

Minimum Hardware Requirements:

- 8Gb+ of RAM
- 128 GB or more (256 GB recommended)
- 10Gb+ of free hard drive space
- Working keyboard
- Trackpad/Mouse
- Display
- A power Adapter.

Software's and Tool's Used:

- Jupyter Notebook
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Scikit Learn

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods):

The data set contains 16204 rows of data with no null values related to the cars. Since the dependent feature is continuous variable, we understand that this problem is a Regression Problem. On analysis of the dataset, I found outliers in 2 features because all other features are categorical variables which is not suitable for either removing outliers or skewness. The outliers were corrected by zscore method. The skewness was also reduced using power transform. There were certain columns which had least importance with our target variable, hence those were dropped. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

Testing of Identified Approaches (Algorithms):

The Algorithm's used are as follows:

- Linear Regression:

  -
    ```
    Model                                                Report:
    RMSE                                       913636.8880079273
    MAE                                         598672.448471305
    r2_score                    :              47.25432682467042
    cv_score                    :              45.73828222628279
    Difference between r2_score and cv is 1.51604459838763
    ```
- Ridge:

  -
    ```
    Model                                              Report:
    RMSE                                    913637.6731891622
    MAE                                     598659.9907344144
    r2_score:                              47.254236165172905
    cv_score                   :           45.739093776995105
    Difference between r2_score and cv is 1.5151423881778001
    ```
- Lasso:

  -
    ```
    Model                                                Report:
    RMSE                                       913636.9671607269
    MAE                                        598671.9806678155
    r2_score:                                  47.254317685442494
    cv_score                    :              45.738288895961475
    Difference between r2_score and cv is 1.5160287894810196
    ```
- Decision Tree:

  -
    ```
    Model                                                Report:
    RMSE                                        450156.669273796
    MAE                                       205168.00397641573
    r2_score:                                  87.19537201567415
    cv_score                    :              83.53188978890567
    Difference between r2_score and cv is 3.6634822267684797
    ```
- RandomForestRegressor:

  -
    ```
    Model                                                Report:
    RMSE                                       345135.0408932915
    MAE                                       166965.78865722514
    r2_score:                                  92.47307312278703
    cv_score                    :              90.67620534809902
    Difference between r2_score and cv is  1.7968677746880104
    ```
- ExtraTreesRegressor:

  -
    ```
    Model                                                Report:
    RMSE                                       351282.6958146908
    MAE                                       175989.19784245166
    r2_score:                                  92.20254096720792
    cv_score                    :              90.27574459558106
    Difference between r2_score and cv is 1.926796371626864
    ```

- GradientBoostingRegressor:

  -
    ```
    Model                                                      Report:
    RMSE                                             550309.4965664748
    MAE                                             328483.44503800396
    r2_score:                                        80.86389247771388
    cv_score                      :                  79.18832629101288
    Difference between r2_score and cv is 1.6755661867009906
    ```

Run and evaluate selected models:

The algorithms used for hyper parameter tuning and fitting train & test dataset are

a. GradientBoostingRegressor
b. RandomForestRegressor
c. ExtraTreesRegressor

Final Model:

GradientBoostingRegressor with HyperParameter Tuning.

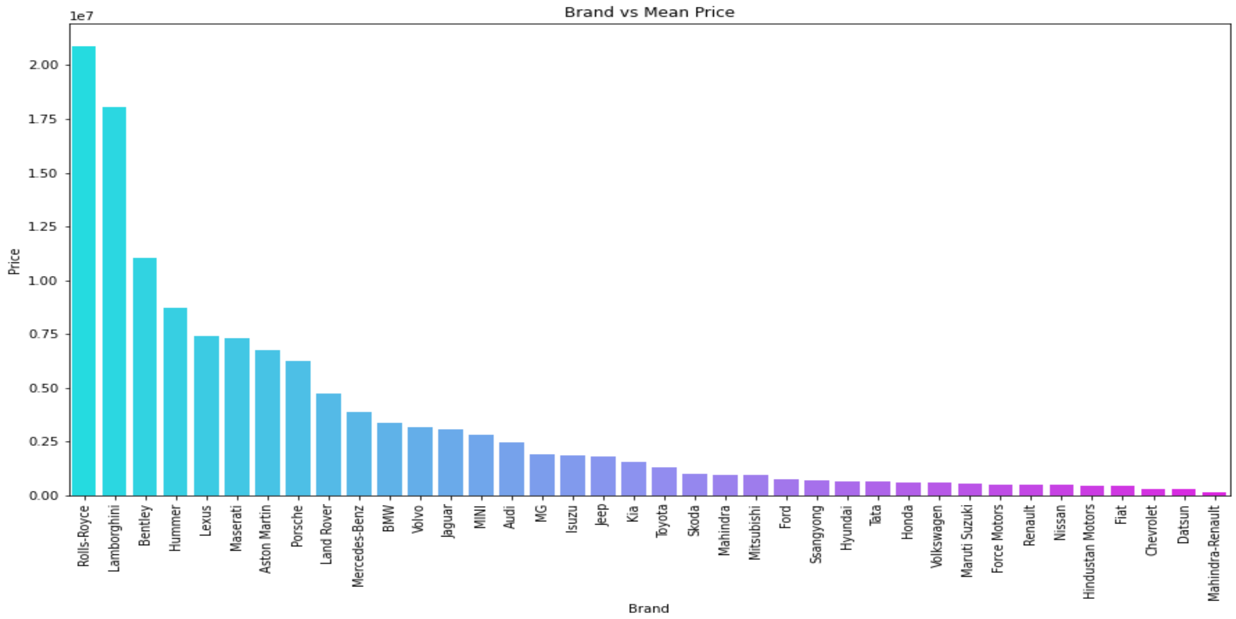
```
Model Report:  GradientBoostingRegressor
RMSE 350958.7091668782
MAE 177899.88335234672
r2_score: 92.21691746601036
cv_score : 91.77573869454856
Difference between r2_score and cv is 0.44117877146179296
```

Key Metrics for success in solving problem under consideration

The Key Metrics used in solving the problem are:

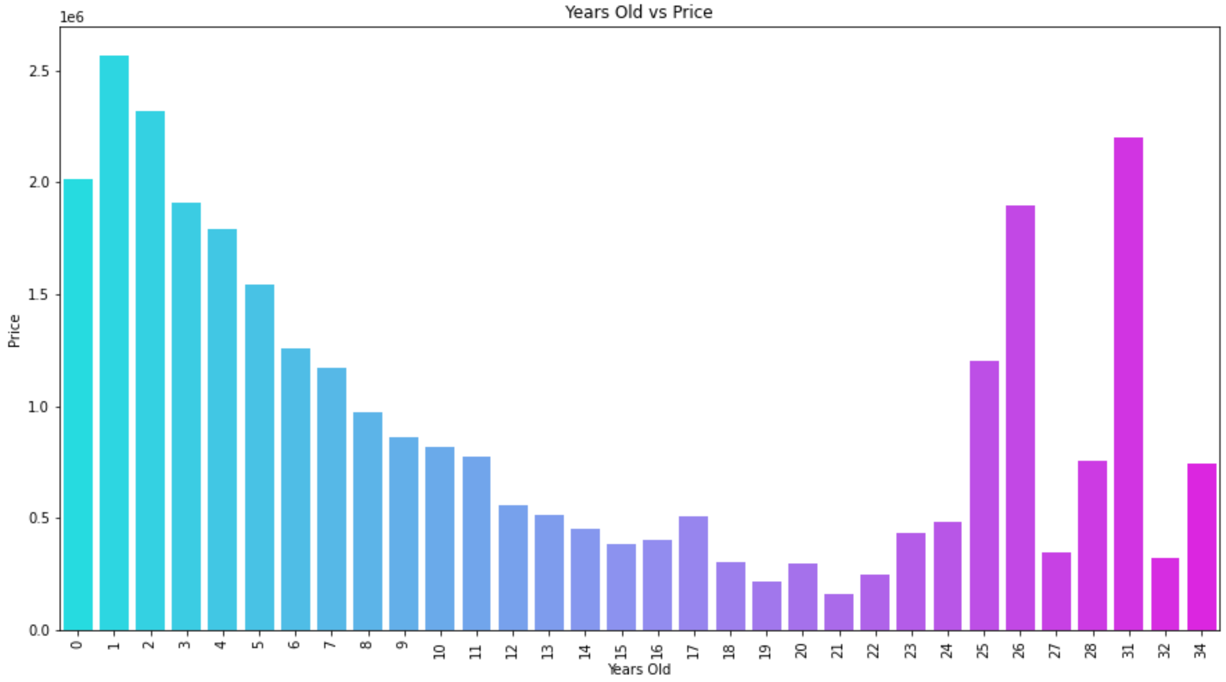
- R2 Score
- Cross-Validation Score
- MSE
- RMSE

Visualizations:

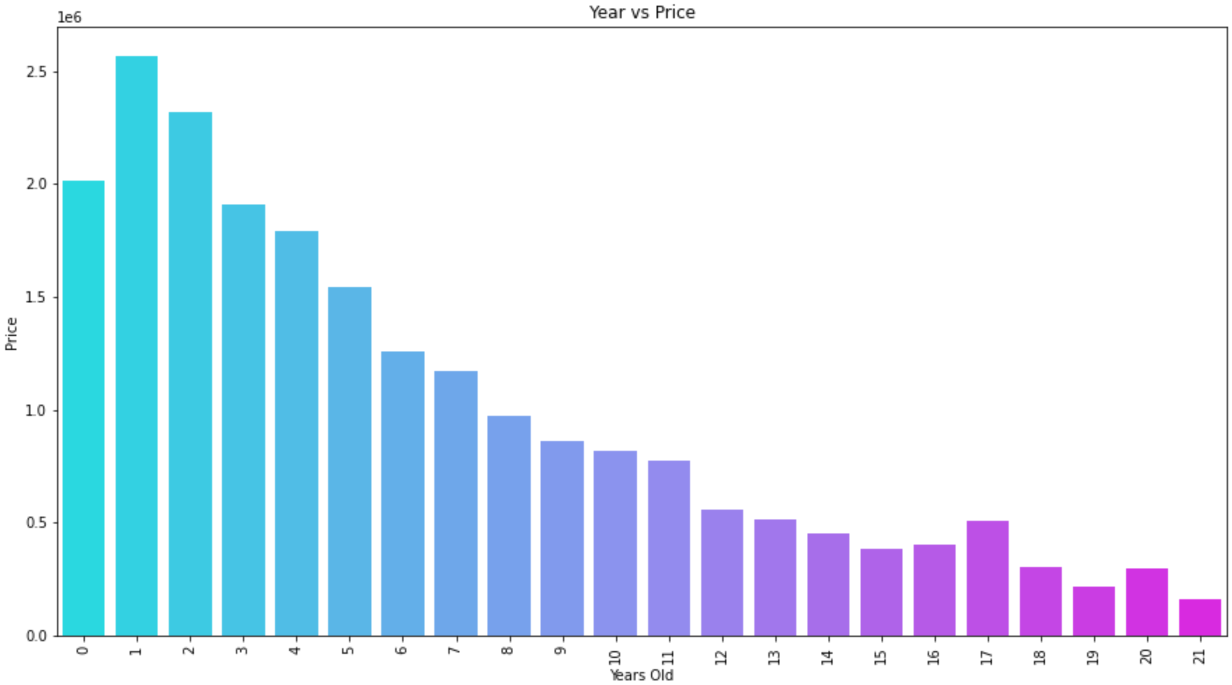


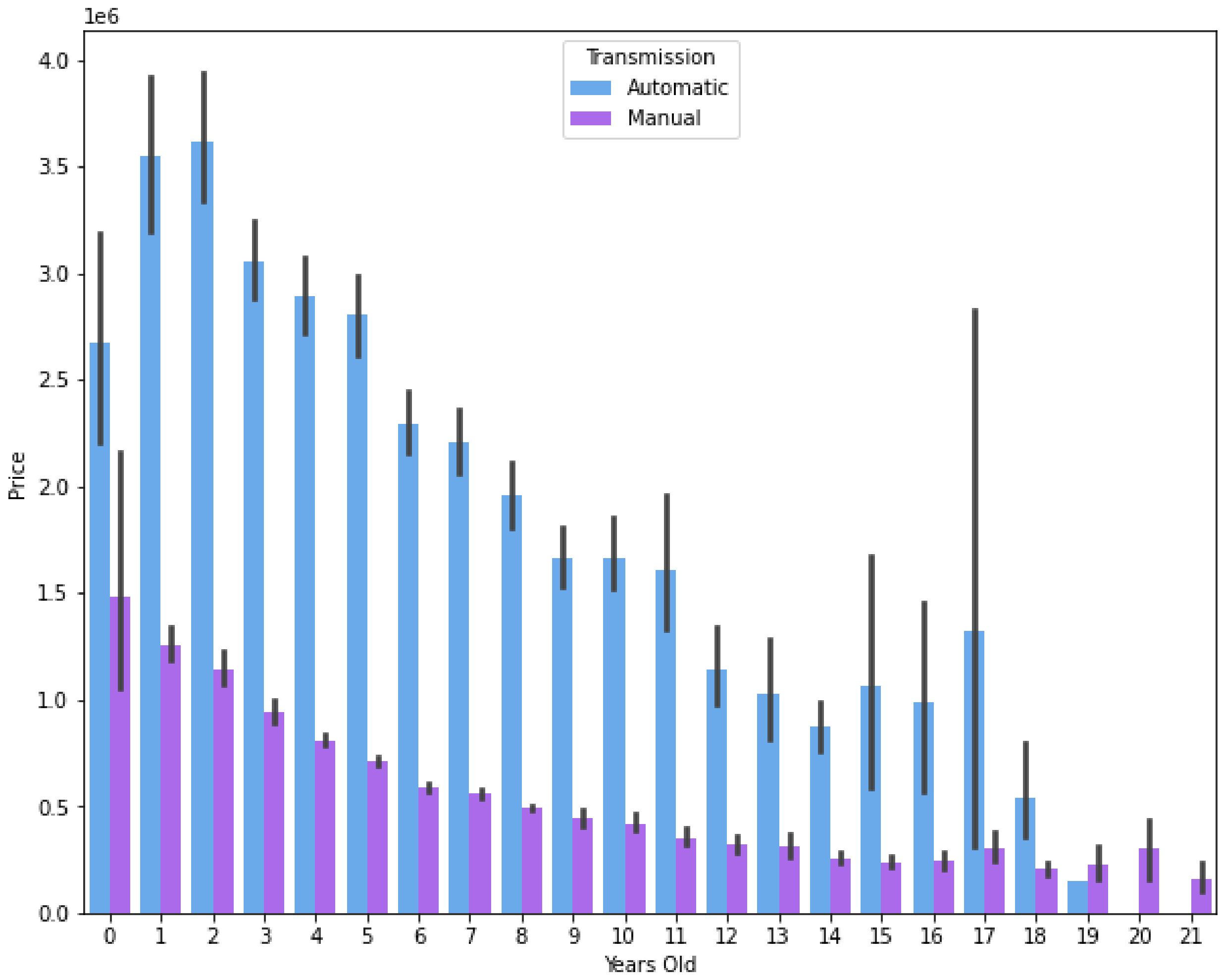
We see the mean price wise distribution of the brands.

- Rolls-Royce are the most expensive.
- Mahindra-Renault is the cheapest.

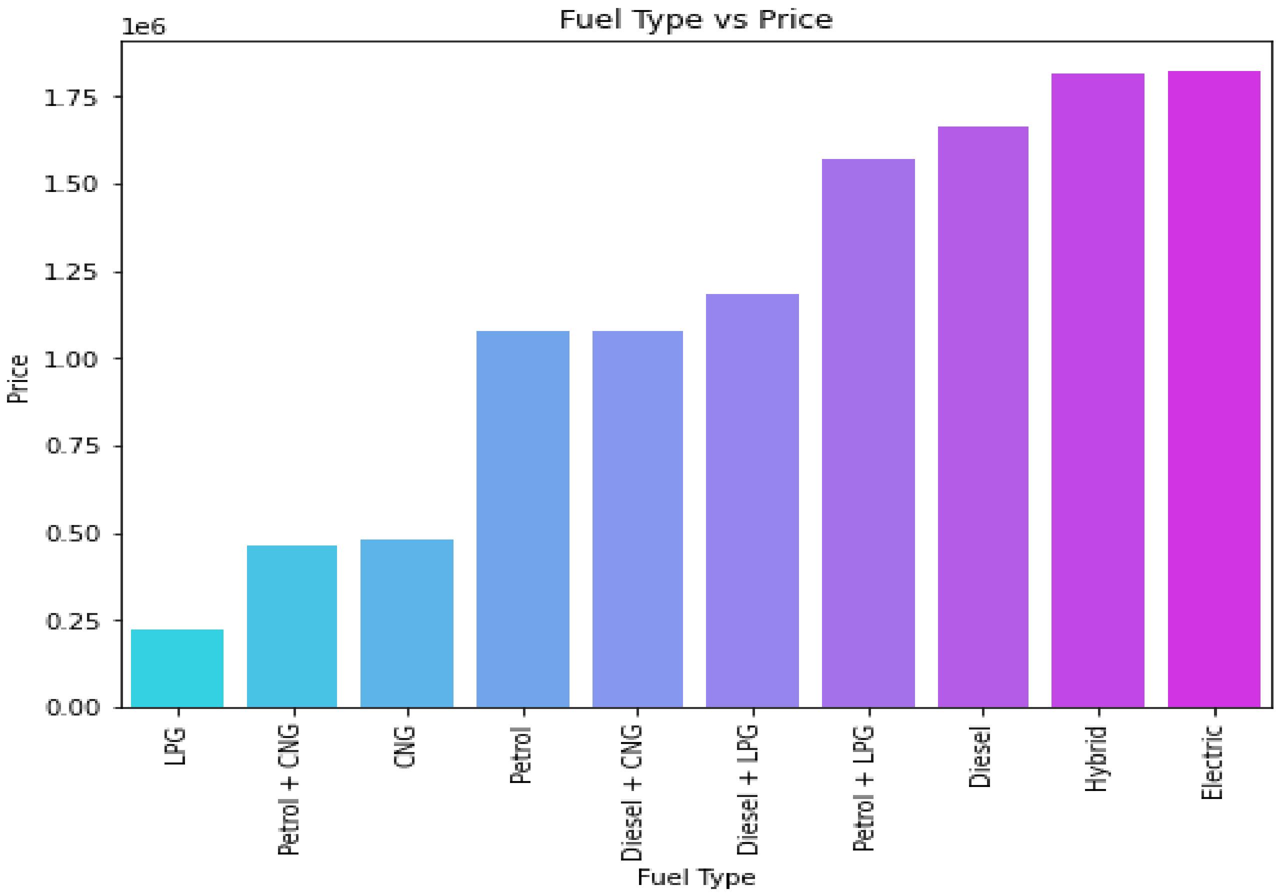


- We see a linear relation in the cars that are up to 23 years old.
- On further analysis we see that there are just 15 cars sold which are higher than 24 years old, which also happen to be mostly exotic cars because of which the avg price has a very high peak.
- We could drop them as it would affect our model adversely.

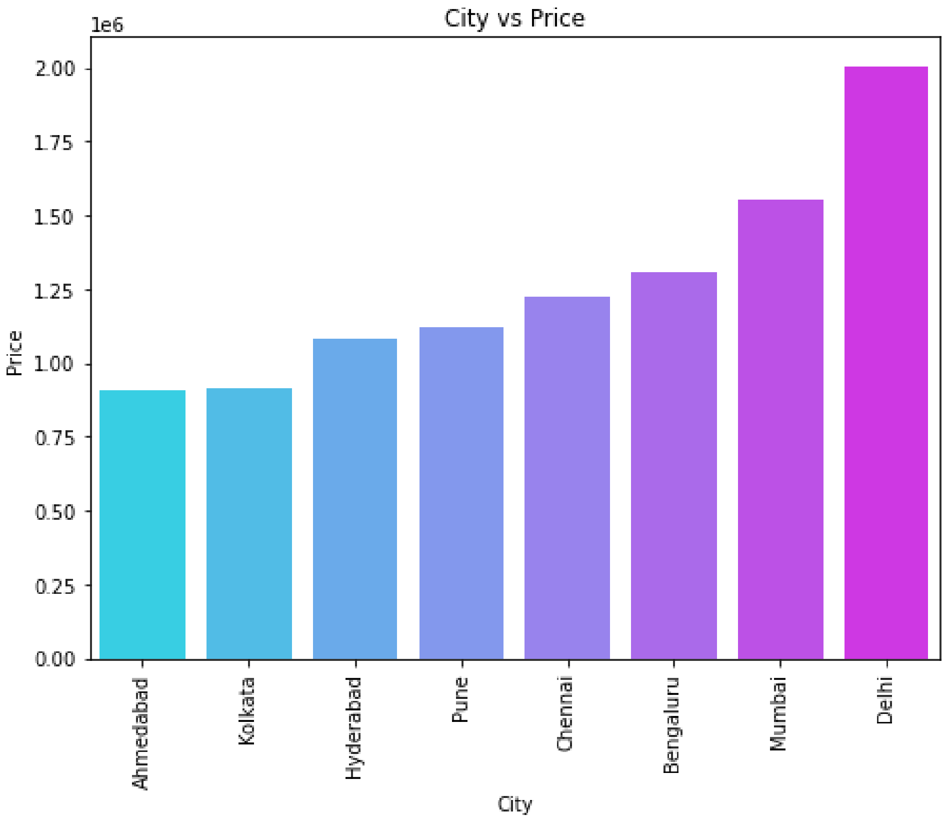

Year vs Price
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Price
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Years Old



- We see that the prices of Automatics are significantly higher than the Manuals.

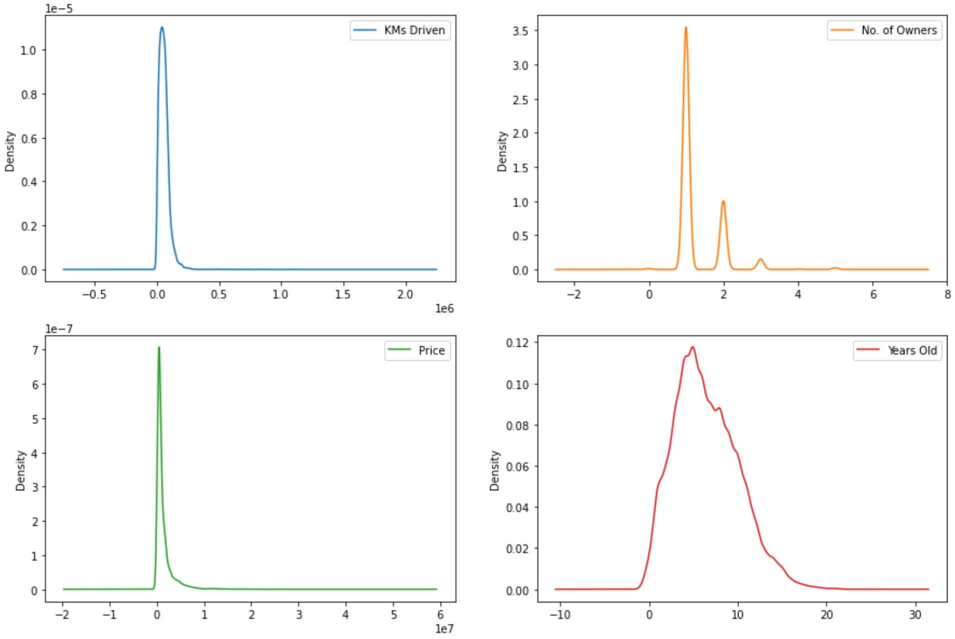


- From above, we see that the Hybrid and Electric vehicles are the most expensive.



- ➢ We see that the vehicle cost is Ahmedabad is the Cheapest and Delhi the highest

Checking Skewness:

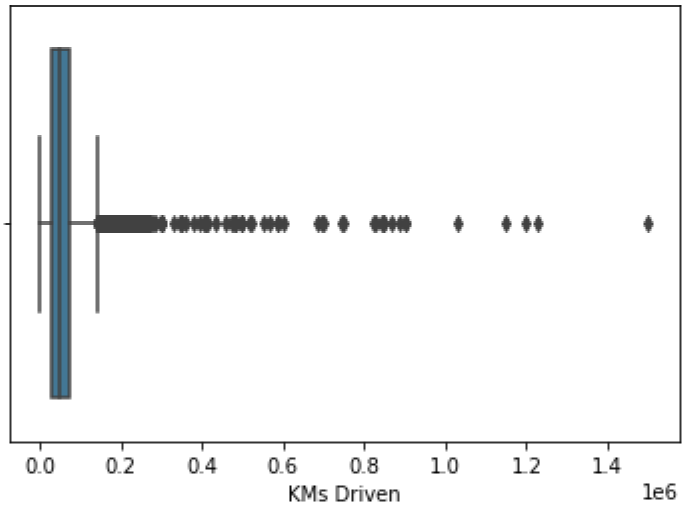


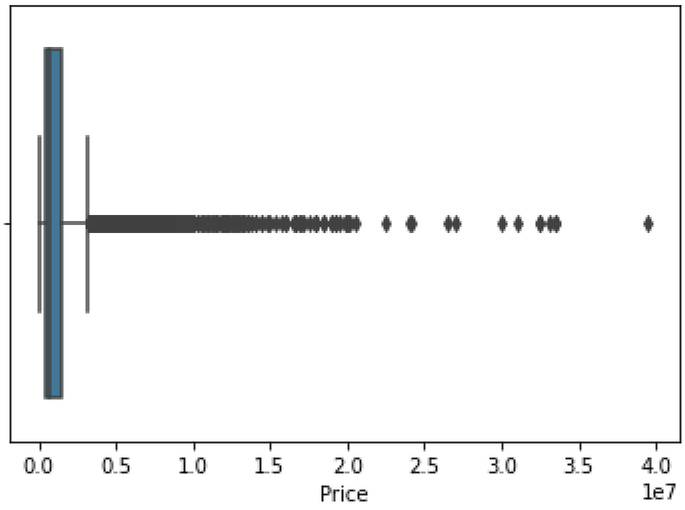
Setting the skewness threshold as +/-0.5. We see right skewness in the data.

- `KMs Driven`
- `No. of Owners`
- `Price`

The above-mentioned features are above the skewness threshold but No. of Owners and Price can be ignored No. of Owners is a categorical variable and Price is Target Variable.

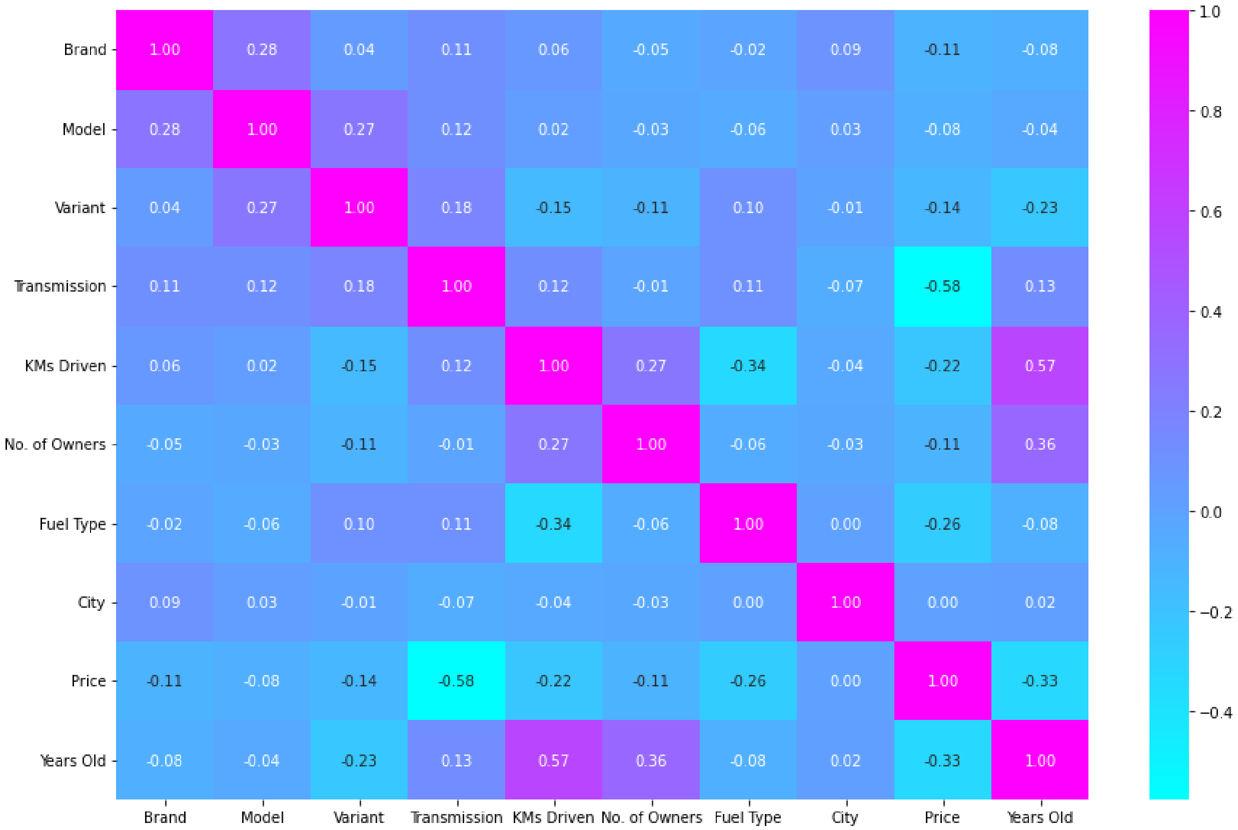
Checking Outliers:





We see a lot of outliers in the price and KMs driven.

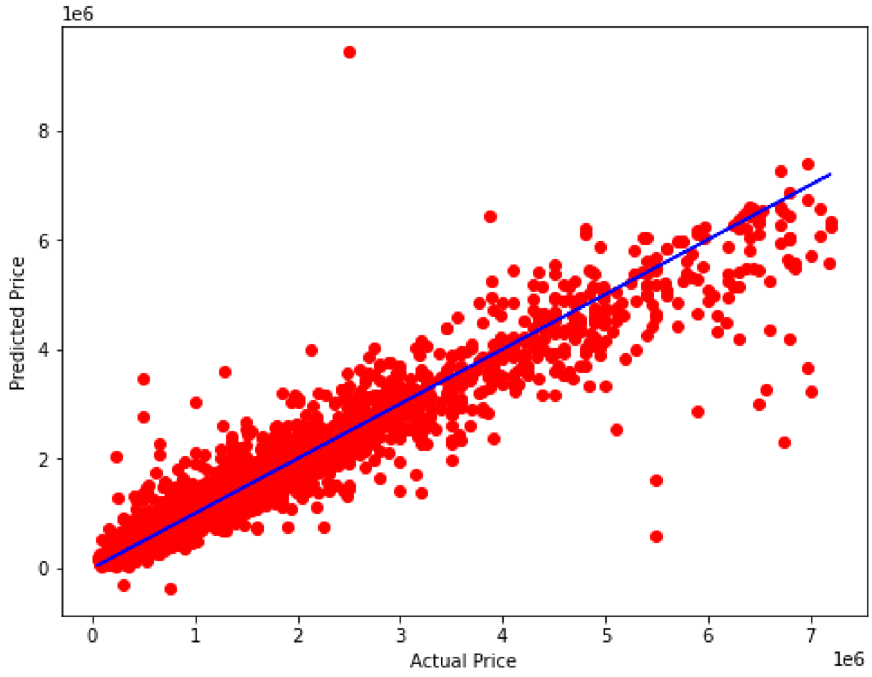
Checking Correlation:



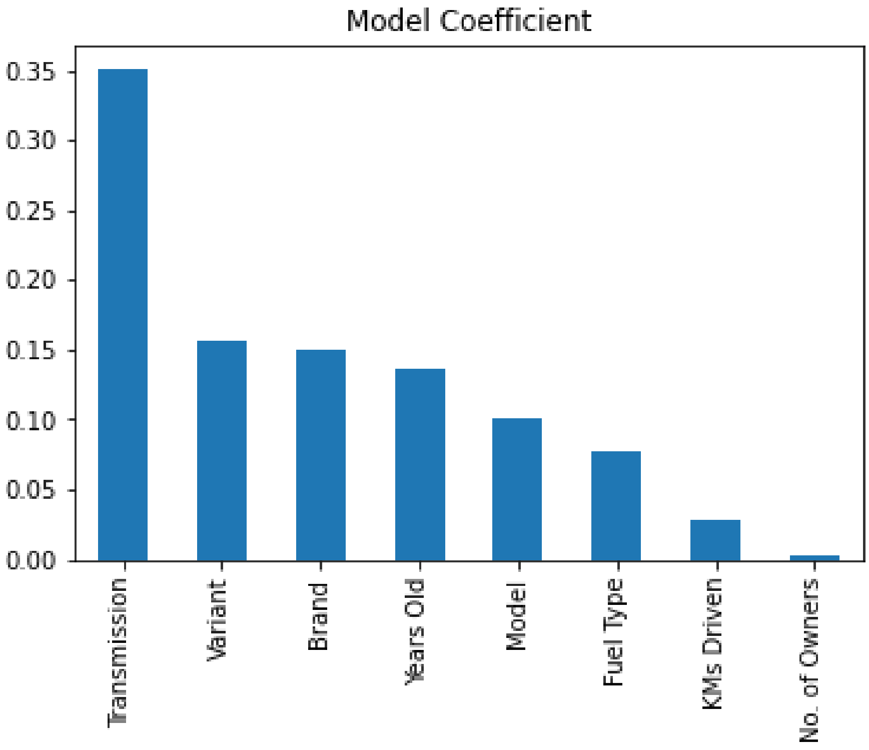
- We see Transmission has the highest negative correlation with the Target.
- Most of the features are negatively correlated and City has 0 correlation, hence we could consider dropping it.

Interpretation of the Results

We have achieved an accuracy of 92% which is pretty good with a difference in cross-validation of 0.44.



The Features that played the most significant role in determining the price are:



CONCLUSION

Key Findings and Conclusions of the Study

Mostly, the customers have the intension of buying the car with best specification and low price. There are certain cases, when the car has high price with all the specifications. May be a colour, variant, transmission type, engine can be key indicator to set price. Price may vary accordingly. With this model built, we can certainly determine car price of a specific brand which deliver high performance

Learning Outcomes of the Study in respect of Data Science

The dataset was not having so much outliers. The only challenge was we needed more data to analyze and train the machine. Price was indicative based on the specifications like torque, number of cylinders, gear box, but we couldn't scrape all the specifications because of some limitations. Although the data we scrapped was Raw and misplaced, data cleaning was very important to get proper prediction. Gradient boosting algorithm was used as it gave the best results.

Limitations of this work and Scope for Future Work

The Actual prices and predicted prices are having some difference which may lead to underfitting. This may be due to the training of model with less specifications or features. So, in future if we need better performance model, we would need to scrape most of the specifications so that the model can give better outcomes.