CAR PRICE PREDICTION

Tinu Shiby

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my mentor Mr. Sajid Choudhary, Flip Robo Technologies for his inspiration and guidance at every stage of the project. The blessing, help and guidance given by him will certainly help me in achieving better things in future.

I wish to express my profound gratitude to the DataTrained family for providing an opportunity to undertake this internship.

Lastly, I thank the almighty and my parents for their every small support and encouragement which helped me to complete this project successfully.

INTRODUCTION

Covid-19 have impacted the market in various aspects, during the lockdown and after we have seen a lot of changes in the car market. Now some cars are in high demand hence making them costly whereas some are not. Our client is the one who works with small traders, who sell 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 a car price valuation model. The project includes two phases, data collection and model building.

DATA COLLECTION

The details of cars are collected from the website of cardhekho, scrapped the details of almost 5093 cars by giving different locations like Chennai, New Delhi,Ahmedabad, Bangalore,Mumbai , Bangalore , Hyderabad etc.
The following items are scrapped:

- Car name
- Making year
- Kilometers covered
- Fuel type
- Engine
- No.of owners
- Transmission
- Price

Dataframe was created and converted that to csv file for model building.

DATA ANALYSIS

The goal is to find out how price varies according to the variables. To start with the data analysis, i have imported the following:

- Pandas
- Matplotlib
- Seaborn
- Standard Scalar
- Principal Component Analysis
- Train test split
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR
- Gradient Boosting Regressor
- XGBoost Regressor
- GridSearchCV
- Cross-validation score
- RMSE score
- R2 score

EDA

Exploratory data analysis is one of the most crucial steps in data analysis. It provides a better understanding of data set variables and the relationships between them. It can also help in determining if the statistical techniques you are considering for data analysis are appropriate.

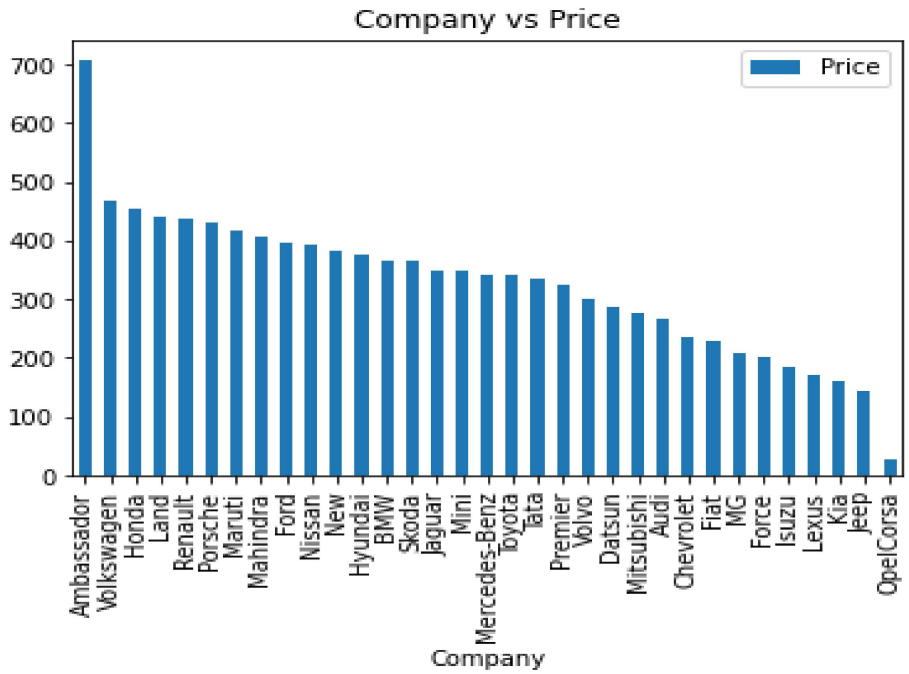
The dataset has 9 variables and 5093 rows, which is splitted into train and test dataset.

The dataset doesn't have null values and the datatype was object.

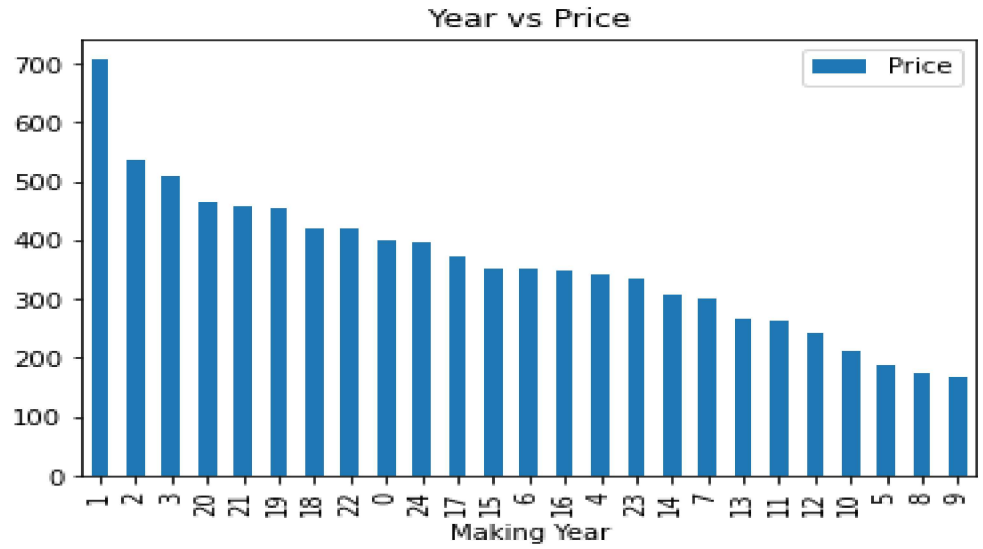
Onehot Encoding Technique is applied for Fuel Type for the rest of the variables LabelEncoding technique is applied.

To find out the relationship between feature variables and target variables, I have plotted graphs.

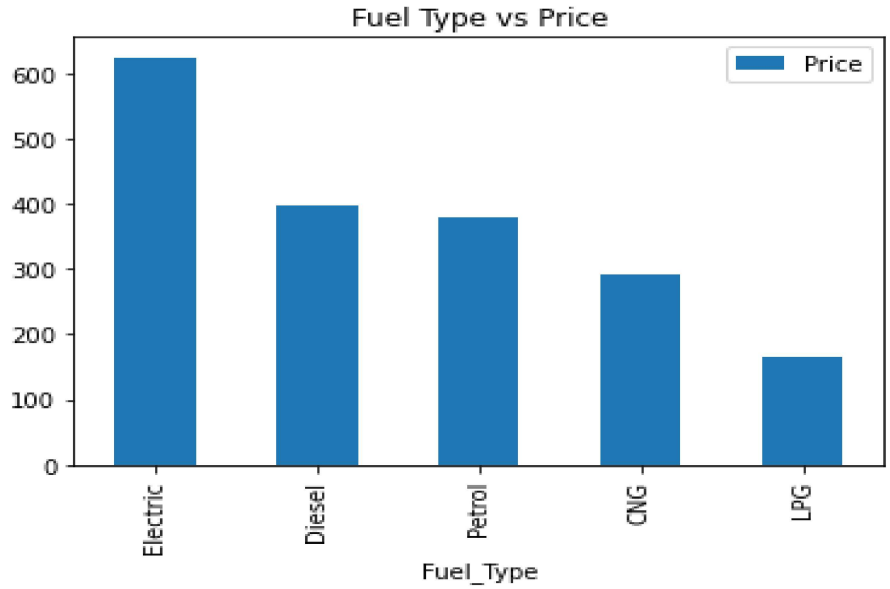
i) Company vs Price



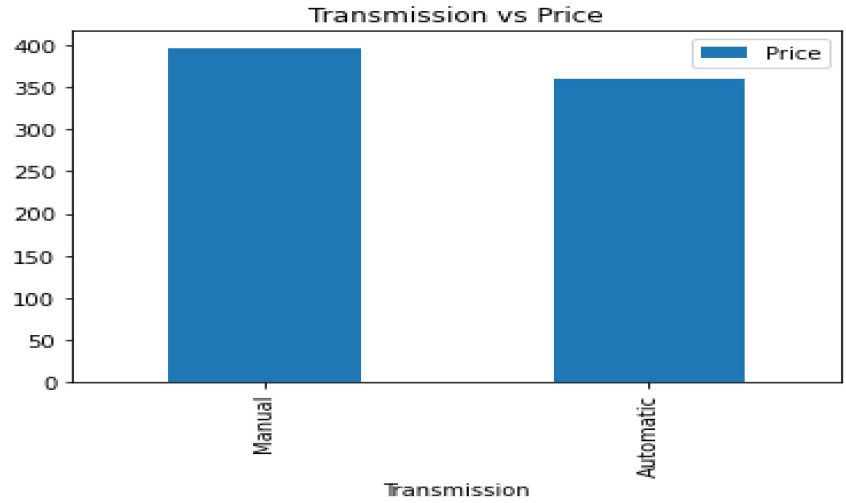
ii) Making Year vs Price



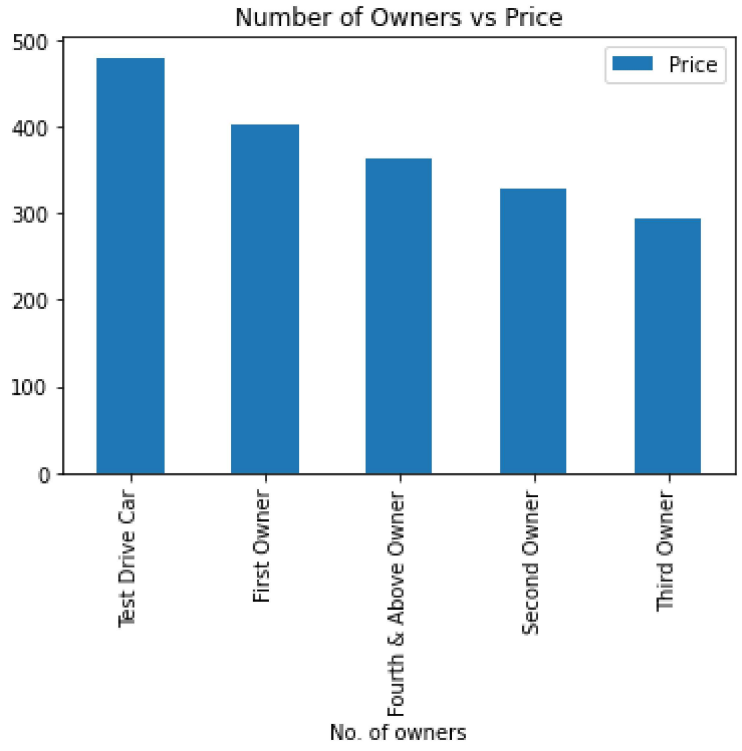
iii) Fuel Type vs Price



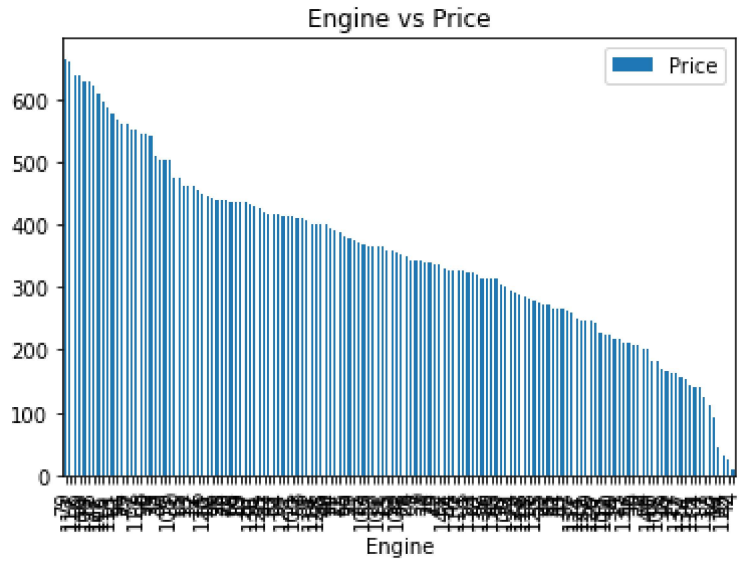
iv) Transmission vs Price



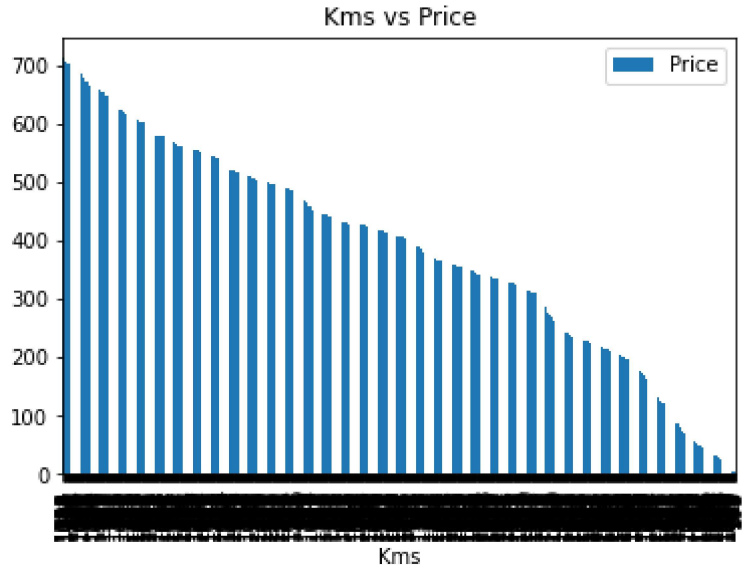
v) No. of owners vs Price



vi) Engine vs Price



vii) Kilometers covered vs Price



From the data analysis part its clear that the following factors plays an important role in determining the price of used cars :

- Making Year
- Engine
- No. of owners

After bivariate analysis I have checked the correlation and it was found out that Making year is having the highest correlation with target variable and Fuel Type Electric is having the least correlation with price. Multicollinearity is not that prominent in the dataset. In order to check for the outliers a statistical analysis is done using the describe method. The difference between Q1,Q2,Q3 and maximum is different, so we came to the conclusion that outliers exist in the data.

DATA PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

The column Unnamed: 0 was dropped as it is same as the index.

Outliers are removed from the dataset using z-score, data loss was almost 4%.

As skewness exists in the dataset, the skewness is removed by applying the power transform method in specific yeo-johnson method. Then the skewness got removed from the dataset.
To standardize the data I have used StandardScalar, the whole idea of using Standard Scalar is that it will transform the data such that its distribution will have a mean value of 0 and standard deviation of 1.

Then find out the best random state and splitted the dataset into train and test.

BUILDING MACHINE LEARNING MODELS

As the problem is a regression problem I have used the following models:

- Linear Regression
- Decision tree Regressor
- Random forest Regressor
- Gradient Descent Boosting Regressor
- SVR
- XGBoost Regressor

Accuracy of Linear regression model ,Decision tree, Random forest,Gradient descent boosting and XGBoost were relatively small.
Cross-validation scores of Linear Regression, Decision Tree, Random Forest, Gradient descent boosting and XGBoost were found out.
Since the difference is least for XGBoost Regressor, that is taken as the best model and applied Hyper parameter tuning using GridsearchCV. The best parameter values are as follows:
Learning rate : 0.06
Max_depth : 10
n estimators : 80
After hyperparameter tuning the score increases to 74%

CONCLUSION

Before entering the market, it is advisable to have a check on various parameters which gives high profit to the company. By careful analysis it is found out that Making Year, Engine,number of owners etc plays an important role in price prediction.
It was found out that XGBoost is the best model to predict the house pricing as it works with an accuracy of 47%. So, the company can use this model and accordingly they can manipulate the strategy of the firm and concentrate on areas that will yield high returns.