Car Price Prediction Project

Submitted by
Akhil Vangeti

Acknowledgement

I sincerely thank Mr. Sajid Choudhary for providing me valuable information and guidance for completion of this project. His assistance has helped me a lot during the course of entire project.

I also thank datatrained institute for helping me gain hands on experience about machine learning.

Thanks to the cars24 for allowing their data to be used in this study.

Introduction

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.

The price of the car varies depending on many factors and even the resale price of the car. This study is intended on building a machine learning model which predicts the resale prices of cars.

Since, the sales of cars coming back to pace, this model helps the traders to valuate the resale price offered and make a decision. This model is built with the recently collected data from the popular cities of India.

Analytics and Modelling

The steps followed in this study are as follows:

1) Data collection
2) Data Study and Analysis
3) Data preprocessing
4) Modelling

Data Collection:

The data for this study is collected from the online used car trader CARS24.

The data which is collected from this website are as follows:

a) Location
b) Brand
c) Type
d) Model
e) Variant
f) Year
g) Number of owners
h) Kilometers
i) Fuel
j) Transmission
k) Price of the used car

The Selenium library is used for the scraping process. After the scraping process is complete, we had around 5045 data points.

The data consisted of 10 popular cities which are New Delhi, Noida, Chennai, Ahmedabad, Hyderabad, Bangalore, Pune, Mumbai, Gurgaon and Kolkata. The Types of cars are hatchback, sedan, suv, luxury sedan and luxury suv.

The data then is saved into a csv file. This data is used in the next sections.

The ipynb file for the above is named as cars-data-scrape and can be found using this link: https://github.com/vangetiakhil/FRT3

Data Study and Analysis:

The data used in this study consists of continuous, categorical data.

The format of data is integer and object type. There are a total of 11 variables in this dataset and are also mentioned in Data collection section. These variables are:

a) Location
b) Brand
c) Type
d) Model
e) Variant
f) Year
g) Number of owners
h) Kilometers
i) Fuel
j) Transmission
k) Price of the used car

The dataset contains 5045 rows.

The first step taken in the analysis part is to check for the missing values and it is found that there are 181missing values in the transmission column of the dataset. These values are filled based on the same kind of model, type and brand.

To proceed further with the data analysis part, the strings has to be analyzed. Since, the string may contain white spaces, capital letters and small letters there is a need to bring them under same roof before the analysis because a same model may be written in capital letters and small letters.

From the exploratory data analysis, the following information has been deducted:

a) maruti brand is almost in the entire half of the dataset.
b) maruti, hyundai, honda, toyota, mahindra accounts to approximately 70% of the entire dataset.
c) the luxury cars are very less.
d) re used hatchbacks and suvs are more in number.

e) in maruti -sedan -swift dezire is the top on the list of cars for sale.
f) in hyundai -hatchback -grand i10 is the top on the list of cars for sale.
g) in honda -sedan -city is the top on the list of cars for sale.
h) in toyota -luxury suv -fortuner is the top on the list of cars for sale.
i) in mahindra -suv -xuv500 is the top on the list of cars for sale.
j) While petrol and diesel accounts for majority, there are few cars with petrol +cng and petrol+lpg.
k) manual transmission cars are around 80% and the rest are automatic transmission cars
l) Mostly when the car is 3-8 years old, it will be listed for a resale.
m) petrol+lpg cars are resaled mostly after 10 years of usage.
n) As age of use(No_of_years) increases the resale value decreases, but luxury suv value is decreased very less when compared to other car types.
o) prices of luxury suvs are not very much dependent on the no.of.owners.
p) Factors such as kilometers, number of owners and number of years used are inversely correlated with the resale price of the car.

Data preprocessing:

The Price column is our target variable and the data in it is in integer format but with a value far greater than the predictors values. This column values if are bigger the model may not work as it is intended to be. The errors in predictions will be more.

The columns such as location, type, brand, model, variant, fuel, transmission are categorical variables and needs to be encoded to be fed into the model. This is done by using dummies method.

The continuous variable columns are Kilometers, Years, Owners and Price needs to be normalized before modelling. The skewness of kilometers and price are very high and not under the acceptable range.

Kilometers column have been transformed using cube root transformation and Price column has been transformed using log transformation. Now, the data has been preprocessed and ready for modelling.

Modelling:

The target variable is continuous. Therefore, this is a regression problem.

The different models which have been used for modelling are:

1) Linear regression
2) Decision tree regression
3) Kneighbors regression
4) SVR

The regularization models which have been used are:

1) ElasticNet
2) Ridge
3) Lasso
4) Bayesian Ridge
5) Huber Regressor

The Boosting/Ensemble models which have been used are:

1) Random forest regressor
2) AdaBoost regressor
3) Gradient Boosting regressor
4) XGB regressor

The above mentioned modeling techniques are applied and their r2 scores ranges from 60 – 94%. To check which model performs better, cross validation has been used.

After the cross validation scores have been checked, the following observations are made:

a) Linear regression is not performing well as its mean absolute error values are so high. This may be because of two possible reasons:
    a. Few outliers with high prices for types of luxury sedan and luxury suv.

    b. Many predictors are categorical (can be seen decision trees doing well).
b) Decision Tree, Ridge and Bayesian Ridge have r2_score of above 0.90.
c) Ridge and Bayesian Ridge shows less diff between r2 score and cross val score compared to decision trees.
d) Random forest and XGB boosting techniques are doing good.

To choose the best fit model, 4 metrics have been used to rank the models. They are:

a) Mean absolute error
b) Mean squared error
c) R2 score
d) Difference between r2 score and cross val score

Based on the metrics, the rank of models is as below: -

Considering r2_score: - Random forest > XGB > Ridge > Bayesian Ridge (all > 0.93)

Considering diff b/w cvs and r2: - Bayesian Ridge > Ridge > XGB > Random forest (all < 0.04)

Considering mae - Random forest > XGB > Bayesian Ridge > Ridge

Considering mse - Random forest > XGB > Ridge > Bayesian Ridge

Depending the above ranking order, Random forest regression model has been chosen for this project.

GridsearchCV is used to find the best parameters for the model. A new model has been built using the best parameters found using GridSearchCV. The final model is then saved using Pickle library.

Conclusion

The data has been collected by scraping the website of CARS24. The data has been cleaned, analyzed, encoded and models based on different techniques have been built.

The final model is predicting with an r2 score of more than 90%. This can be further improved by expanding the data collection from many other sources and cities.

Note:- The full details of the project (ipynb file) can be found in this link: https://github.com/vangetiakhil/FRT3