CAR PRICE PREDICTION

Submitted by:

Rijul Kumar

ACKNOWLEDGMENT

Reference that i have used are:

- Data Trained Education online video
- Materials provided by Flip Robo
- Data extracted from https://www.cartrade.com/
- Geeks for Geeks
- Stackoverflow

INTRODUCTION

- ## Business Problem Framing

    1. 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.
    2. We are supposed to scrape at least 5000 used cars data from any website and then build a model to predict the price of an old car.

- ## Conceptual Background of the Domain Problem

    The prices of old cars depends on many factors. Such as:

    - Brand, Model, Variant of the car
    - Manufacturing year of the car
    - Driven Kilometers
    - Fuel type
    - Number of owners
    - Location of sale

- Motivation for the Problem Undertaken

  - We are supposed to build a model using Machine Learning in order to predict the price of the old cars as they are currently much different than what they were before COVID.

  - This model will then be used to understand how exactly the prices vary with the variables. This can be used by companies to evaluate the prices of old cars in recent time.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem
    - First of all i imported data from excel file to dataframes using pandas.
    - After that i cleaned the data (like removed substring ‘kms’ from Driven kilometers column data, etc).
    - Afterwards i used .dtypes to know data type of each column of dataframe.
    - Then i changed Driven kilometers and Price column from object datatype to float datatype.
    - After that i used .describe() to know the statistical information (such as max, min value,etc) of dataframe.
    - Then i used .shape to know shape of dataframe.

- Data Sources and their formats
    - Training data has been extracted from ‘https://www.cartrade.com/’ using Selenium.

- Using Selenium, data was extracted to excel file named ‘car_price_data.xlsx’.

- ## Data Preprocessing Done

  - First of all for data preprocessing i checked whether there is a NULL value or not in dataframe using heatmap as well as .isnull()
  - After that NULL values were removed.
  - After that i used count plots from seaborn library to plot all categorical columns for visualisation.
  - Next i used Density plots from seaborn library to plot all continuous columns for visualisation.
  - Then i encoded the dataframe using Ordinal Encoder.
  - After that i checked for correlations using heatmaps, correlation matrix and BAR plot.
  - Finally i confirmed high correlations using VIF and i got that there was no high correlation (i.e. all were less than 10).
  - After that i checked and removed skewness for continuous data columns (except target variable) using yeo-johnson transformation.
  - Finally i checked outliers using boxplot as well as z-score method and afterwards.
  - As removal of outliers from continuous data columns (except target variable column) gave very less loss (1.65%) so i removed them.

- ## Data Inputs- Logic- Output Relationships
  - Data Input :

    These are basically the factors (such as Brand, Model, Variant, Manufacturing year, etc) which affects the prices of old cars.

  - Data Output :

    Our Target variable is ‘Price (in ₹)' which is the price of the old cars and we are supposed to predict it with the help of data input (i.e factors affecting car prices).

- ## Hardware and Software Requirements and Tools Used
  - Hardware used:
    i. Laptop with intel core i5 7th gen
    ii. Internet connection for web scraping

  - Software used:
    i. Jupyter notebook
    ii. Required python libraries such as numpy, pandas, seaborn, matplotlib, etc
    iii. Required libraries for model such as sklearn, etc
    iv. Required libraries for web scraping such as selenium,etc

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    - After preprocessing data (removing NULL, encoding, checking for high correlations, removing skewness and outliers) i separated columns into features and target.
    - As this is a regression problem so we tried 4 models - LinearRegression, SVR, RandomForestRegressor and DecisionTreeRegressor
    - I also tried 4 metrics method - r2_score, mse, rms, mae
    - Then i used Lasso for regularization.
    - Finally i used Ensemble Technique.

- Testing of Identified Approaches (Algorithms)

  As this is a regression problem so we tried following 4 models -
    - LinearRegression
    - SVR
    - RandomForestRegressor

- DecisionTreeRegressor

- ## Run and Evaluate selected models

  I defined a function model and then tried 4 different models using it

In [142]:
```
def model_selection(algorithm_instance,features_train,target_train,features_test,target_test):
    algorithm_instance.fit(features_train,target_train)
    model_1_pred_train = algorithm_instance.predict(features_train)
    model_1_pred_test = algorithm_instance.predict(features_test)
    print("Accuracy for the training model : ",r2_score(target_train,model_1_pred_train))
    print("Accuracy for the testing model : ",r2_score(target_test,model_1_pred_test))

    Train_accuracy = r2_score(target_train,model_1_pred_train)
    Test_accuracy = r2_score(target_test,model_1_pred_test)

    for j in range(2,10):
        cv_score = cross_val_score(algorithm_instance,feature,target,cv=j)
        cv_mean = cv_score.mean()
        print("At cross fold " + str(j) + " the cv score is " + str(cv_mean) + " and accuracy score for training is
        print("\n")
```

Result that i got for each model :

- LinearRegression :
  At random state 7 the training accuracy is : 0.16383619784543546
  At random state 7 the testing accuracy is : 0.1469071923098223

  At cross fold 2 the cv score is 0.1570480439467292 2 and accuracy score for training is - 0.1580800000323732 and accuracy score for testing is 0.1469071923098223

- SVR :
  Accuracy for the training model : - 0.12253247357201102
  Accuracy for the testing model : - 0.16873651665433442

  At cross fold 2 the cv score is - 0.12869919468300006 and accuracy score for training is -0.12253247357201102 and accuracy score for testing is -0.16873651665433442

- RandomForestRegressor :
  Accuracy for the training model :
  0.986697668798722
  Accuracy for the testing model :
  0.8840525978432557
  At cross fold 2 the cv score is 0.8635985105119979 and accuracy score for training is 0.986697668798722 and accuracy score for testing is 0.8840525978432557

- DecisionTreeRegressor :
  Accuracy for the training model :
  0.9981689203007841
  Accuracy for the testing model :
  0.8626950354828585
  At cross fold 2 the cv score is 0.8295530091338674 and accuracy score for training is 0.9981689203007841 and accuracy score for testing is 0.8626950354828585

Finally i concluded that RandomForestRegressor() gives best accuracy and hence i took it as main model.

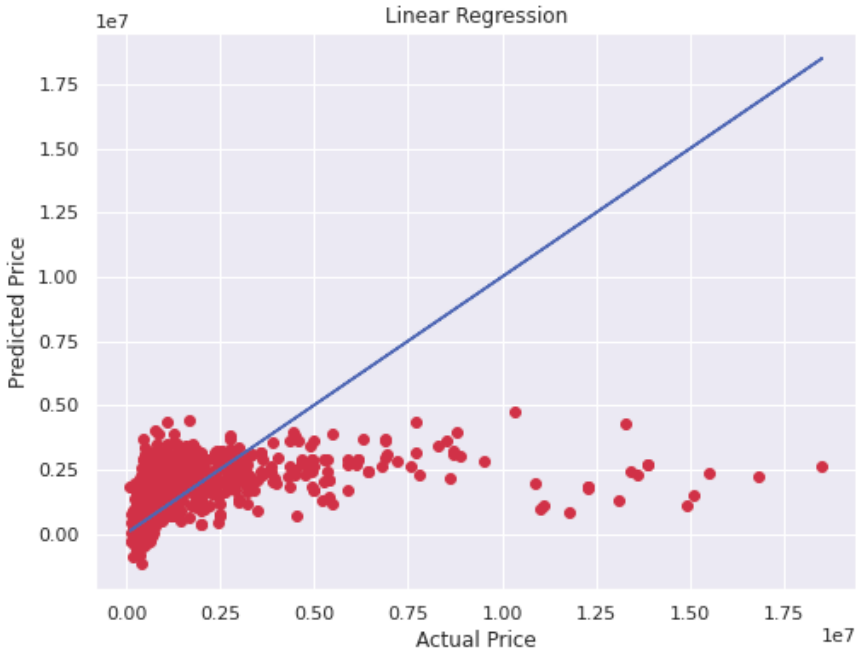
- ## Key Metrics for success in solving problem under consideration

  I tried 4 different metrics method:
  - r2_score
  - mse
  - rms
  - mae

  I got best results from r2_score and hence used it in final model

- ## Visualizations

  We can see that for Linear Regression almost all points lie near best fit curve but not on it.



- ## Interpretation of the Results

  So the results which i got were:

  - Although RandomForestRegressor() is best model but DecisionTreeRegressor() can also be used as they are also very good models.
  - Above point can be seen through r2_score which is the also best among 4 metrics tried.
  - We got our final accuracy (r2_score) as 84.94% after hypertuning.

CONCLUSION

- ## Key Findings and Conclusions of the Study

    - From this study i learnt that sometimes some unrelated factors like COVID can also result in change in prices of old cars in market.
    - Prices of old cars are not highly dependent on any of the factors taken in this project.
    - Almost all of the old cars are sold by their first owner, very few by second owner and others are negligible.
    - Most cars which are sold were manufactured around 2017.

- ## Learning Outcomes of the Study in respect of Data Science

    Some problems faced and their solution (using visualisation and algorithm) used were:

    - While web scraping for data i needed to visit 5000 urls for extracting data from description page of each old car which needed more GPU than i had in my system. So what i did was that i distributed that 5000 urls in a group of 200 each (i.e url[0:200], and so on...) and then executed the code block 25 times as variable data list was already appending.

- After web scraping some data which was supposed to be continuous data column was stored as object datatype in excel file. So what i did was in data cleaning part of model i converted them to float datatype using .astype()

- ## Limitations of this work and Scope for Future Work

  Some limitations are :

  - There are many other factors which are not in the data which may play major role in prices of some old cars such as car colour.
  - Unrelated factors such as COVID may indirectly also affect price of old cars.
  - With evolving technology, prices of cars made with new technology increases while those made with old technologies decreases.

  Scope for future work :

  - This can be made further accurate by taking more and more factors into account.