IE7280: Statistical Methods in Engineering

Project Report

Topic 2: Automobile: Predicting the price of a used car using Regression Model

Nidhi Saraf

Shruthi Gadkari

Rasagnya Reddy

Northeastern University
College *of* Engineering

Topic 2: Predicting the price of a used car using Regression Model

Introduction

For the second topic, we are going to focus on predicting the price of cars using regression model for dataset car sales. Linear model is to conduct the factors important for sales of car. Coding is done in Python. In python various package such as pandas, matplotlib, and scikit-learn are used to read data, visualize data, train model based on actual price and predicting the price of car.

Dataset Description

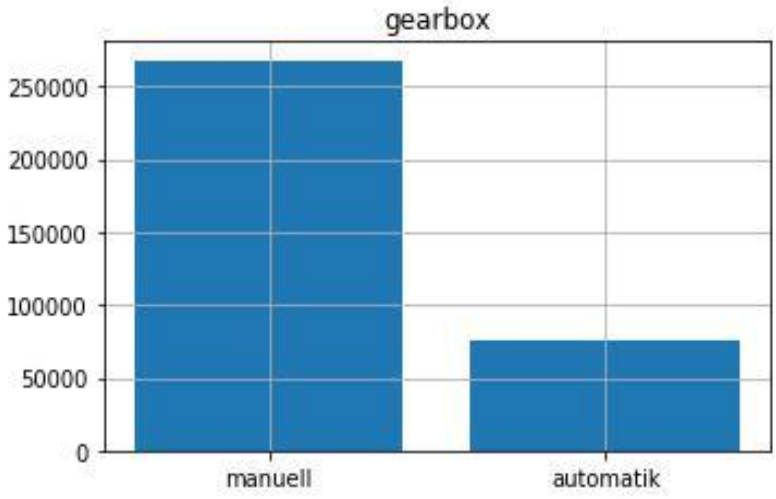
The dataset Cars is picked from Kaggle and has 20 variables. Some of the variables includes kilometer, yearOfRegistration, gearbox, Seller etc. There are 10 categorical variables and 7 numerical variables. 3 columns are dates columns. There are 371539 records in the table. Some of the columns like Seller are not significant in the dataset and will be drooped in data cleaning process

Data Cleaning

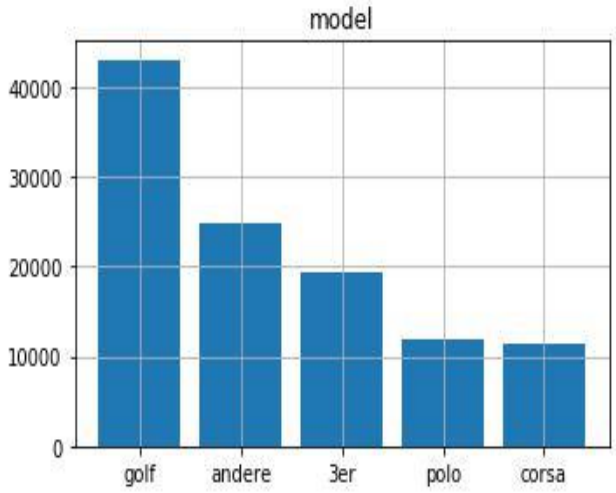
1. Six columns have been dropped from the dataset. 3 columns are categorical- Seller, OfferType, noOfPictures and 3 Dates columns as not much information can be gathered from these columns.
2. There are 5 columns- vehicleType, gearbox, model, fuelType, notRepairedDamage which have Nulls in huge number which can impact the performance of model accuracy. Nulls have been dealt with and replaced with appropriate values (either using most frequency value or using the column value which is relevant).
3. Presence of outliers can also have negative impact on model performance, so they are removed from 2 columns-
    a. yearOfRegistration -In Only 289 registration are there before 1950 and 4000 registration after 2017 so these are outliers and should be removed for better results
    b. Price- only 158 values are there above 2 lakh and 12000 entries below 100 so these should be removed

Data Visualization

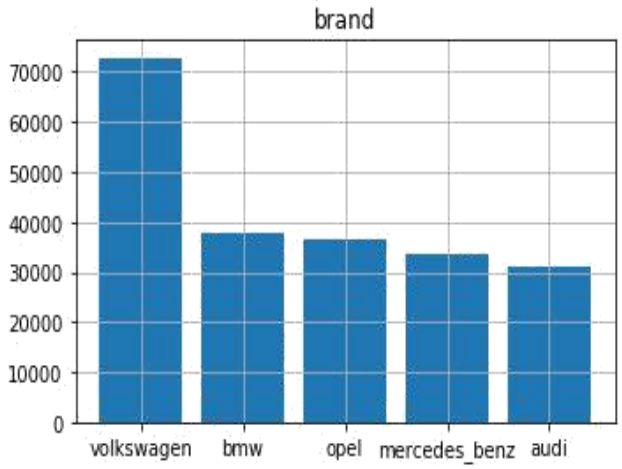
1. **Bar Graphs of various variables**



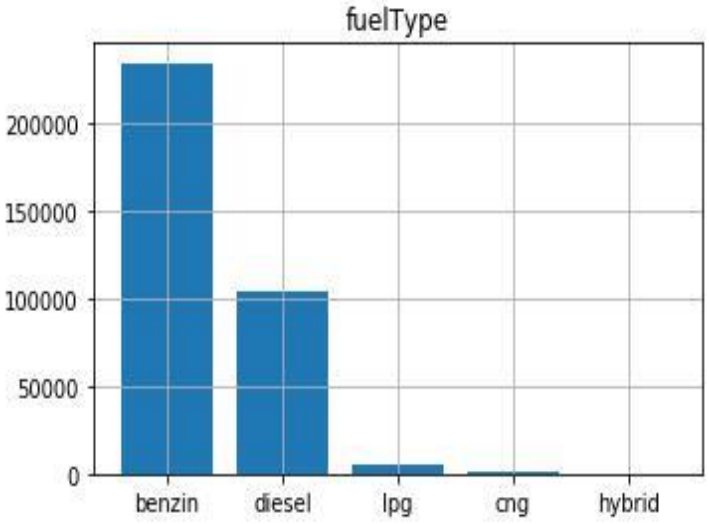
There are 2 categories in gearbox- ‘manuell’(Manual) and ‘automatik’(Automatic). From the above graph values in ‘manuell’ category is much greater than automatic. ‘manuell’ have value at 275000 and ‘automatik’ is around 75000.



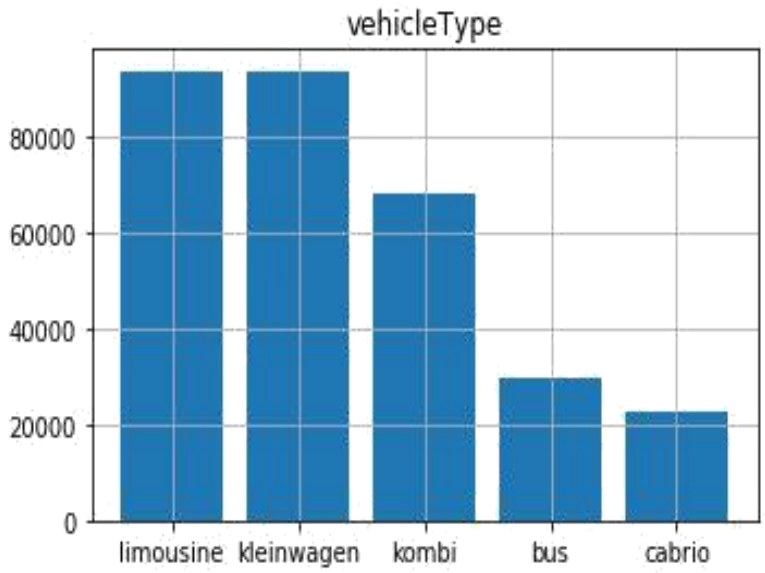
There are 5 categories in Model. The most frequent occurring one among all is ‘golf’ with value at 43000 followed by ‘andere’.



There are 5 categories in brand as well. The most frequent occurring one among all is ‘volkswagen’ with value more than 70000. ‘bmw’ and ‘opel’ are almost at same values.

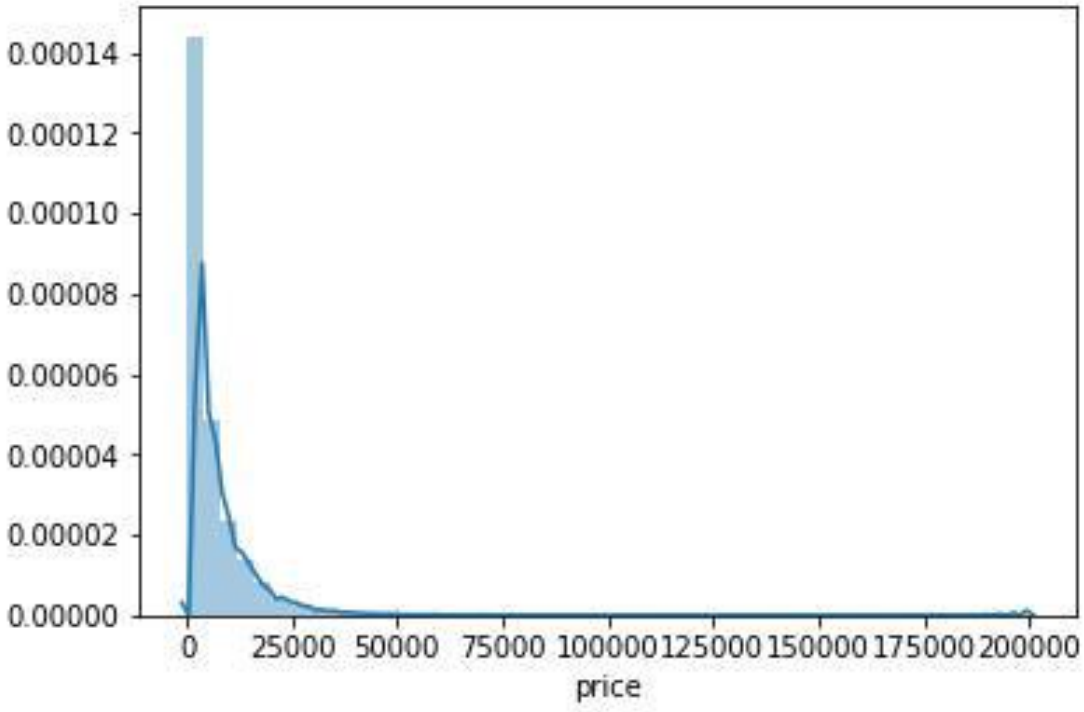


The fuelType has 5 categories though the most used factor is ‘benzin’



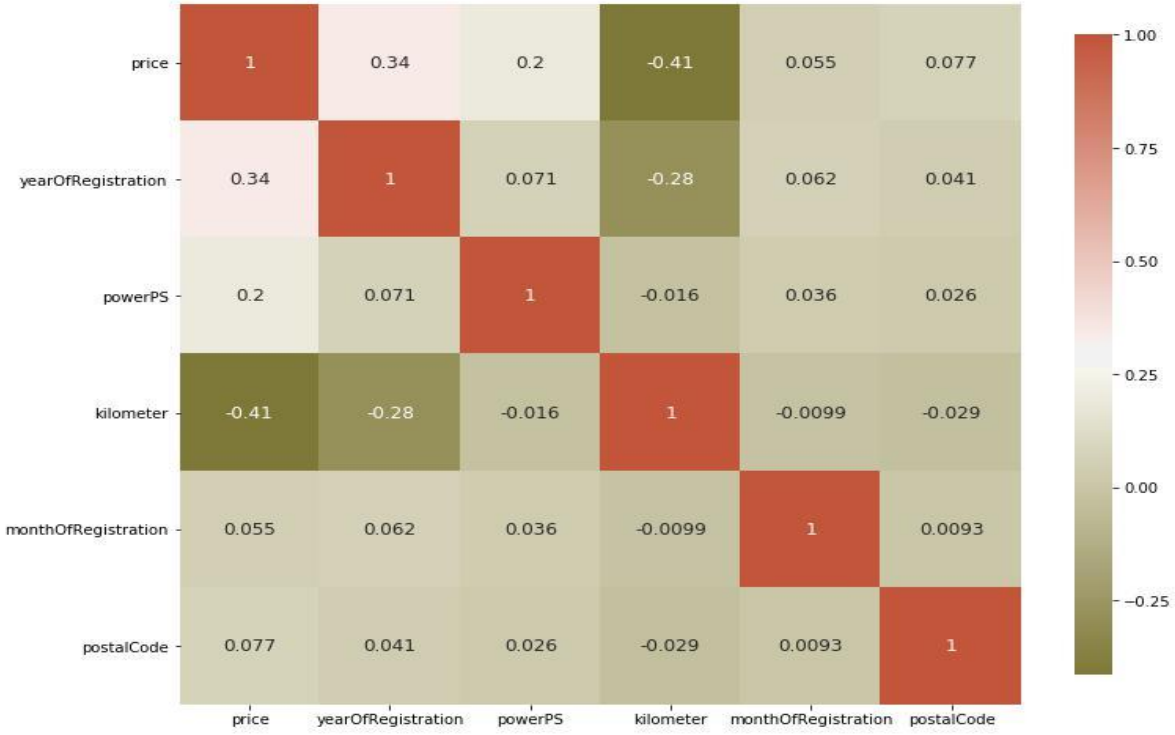
In vehicleType, there are five categories. ‘limousine’ and ‘kleinwagen’ are used at equal number.

2. **Distribution of price**



It is evident from the price distribution graph that price is right skewered. The price is mostly in the range of 0-25000.

3. Correlation Matrix



The above matrix shows the correlation among 6 variables. But we are interested to know that which variables affects the price the most. It is evident form the matrix that price is mainly affected by 3 variables- 'yearOfRegistration', 'powerPS' and 'kilometer'. Since there are 5 categorical categories, these needs to be converted into dummies variables to consider in the model. Pandas "get_dummies" function is used to transform variable type binary. Dummy columns have values as 1 and 0 to indicate yes or no. By converting it to dummies, the number of columns increases from 14 columns to 307 columns.

LINEAR MODEL

To create the model, we have split the dataset to train set and test set. To train the model, dependent variable of the dataset – price is used, and other independent variables are used to have a linear regression model. Price is predicted with the help of dependent variable in test data.

Model performance

To check the performance of model there are 2 ways-

1. Score of the train and test data are calculated and compared. It can be seen from the result below that score of both the datasets are above 50% thus making it a good model.

```
print('Score of train data',lr.score(X_train, y_train))
print('Score of test data',lr.score(X_test, y_test))
```

```
Score of train data 0.5779385246334059
Score of test data 0.5650280725406884
```

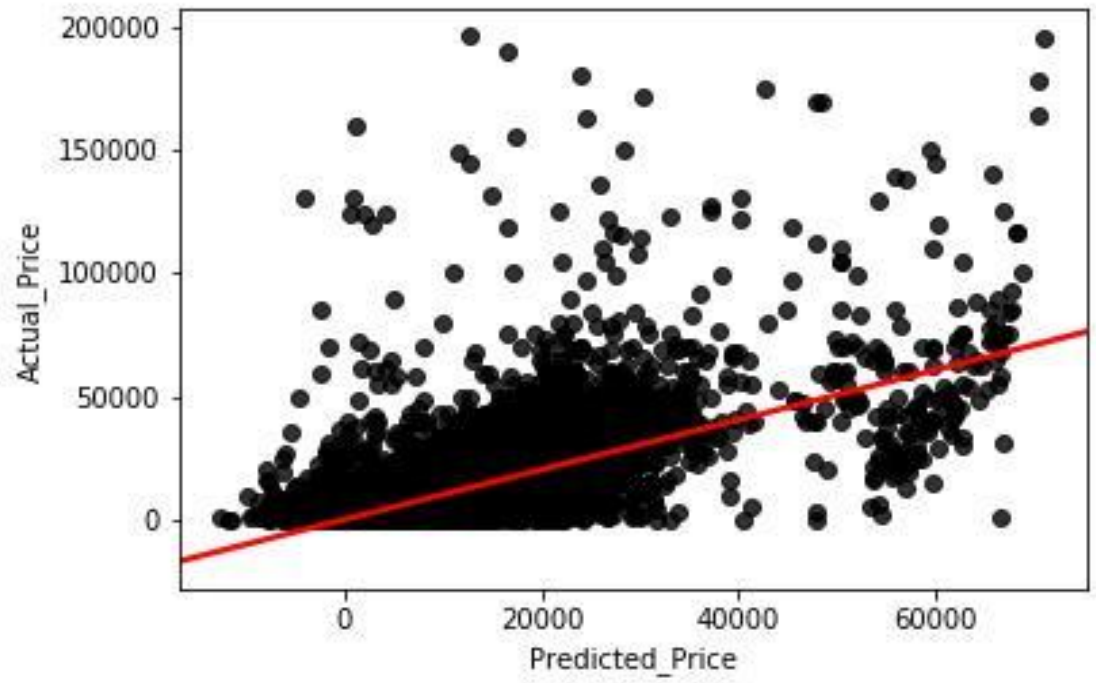
2. MAE, MSE and RSME is calculated for train and test data to compare the accuracy of the model

```
from sklearn import metrics
print('Mean Absolute Error of train data:', metrics.mean_absolute_error(y_train, y_pred_train))
print('Mean Squared Error of train data:', metrics.mean_squared_error(y_train, y_pred_train))
print('Root Mean Squared Error of train data:', np.sqrt(metrics.mean_squared_error(y_train, y_pred_train)))
print('Mean Absolute Error of test data:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error of test data:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error of test data:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error of train data: 2889.2054905009554
Mean Squared Error of train data: 29235811.244860884
Root Mean Squared Error of train data: 5407.015003202866
Mean Absolute Error of test data: 2931.903125683675
Mean Squared Error of test data: 31774791.100226905
Root Mean Squared Error of test data: 5636.913259952375
```

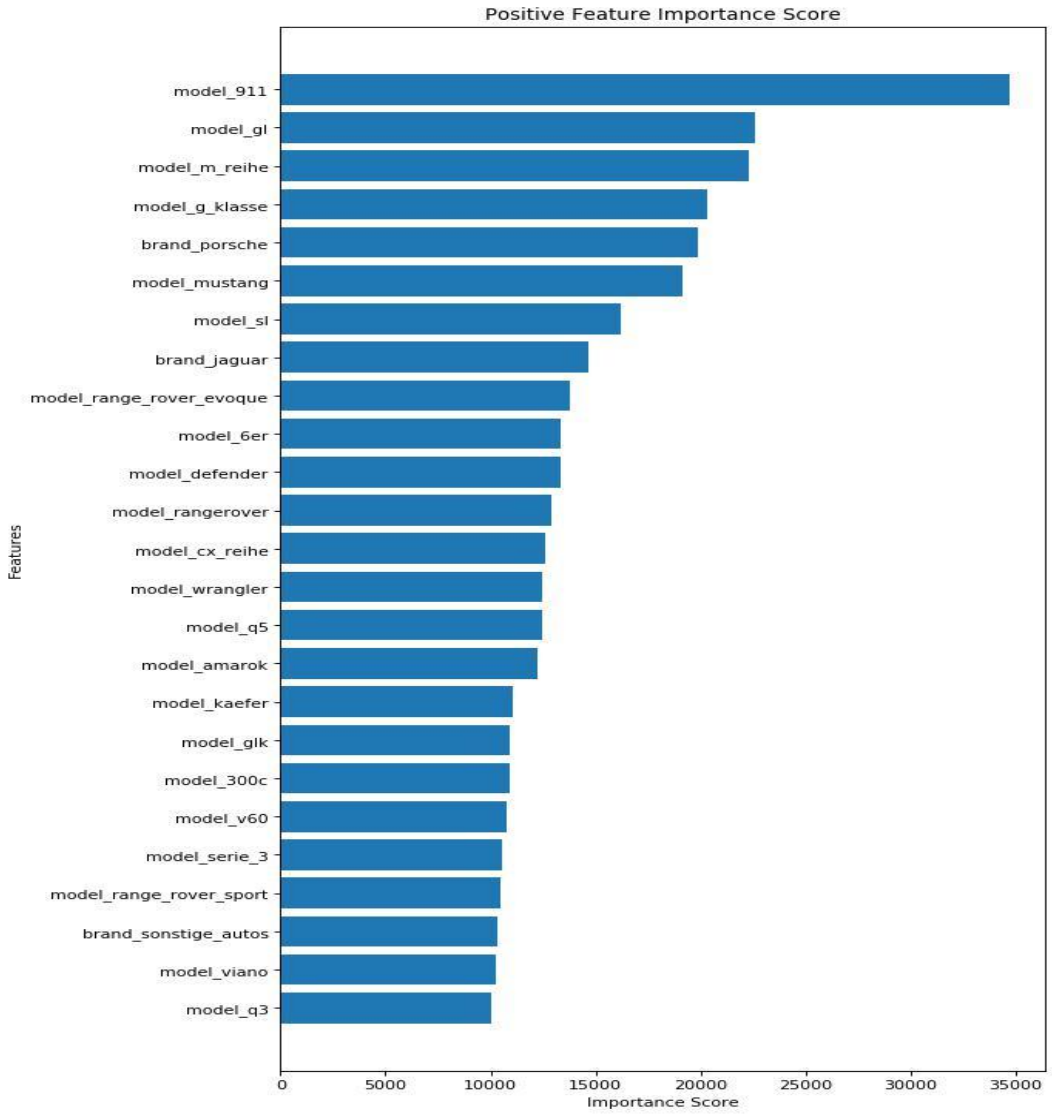
From the above calculation it can be shown that model's performance is good as error of test and train data are in similar range. Thus, price of cars can be predicted with high accuracy.

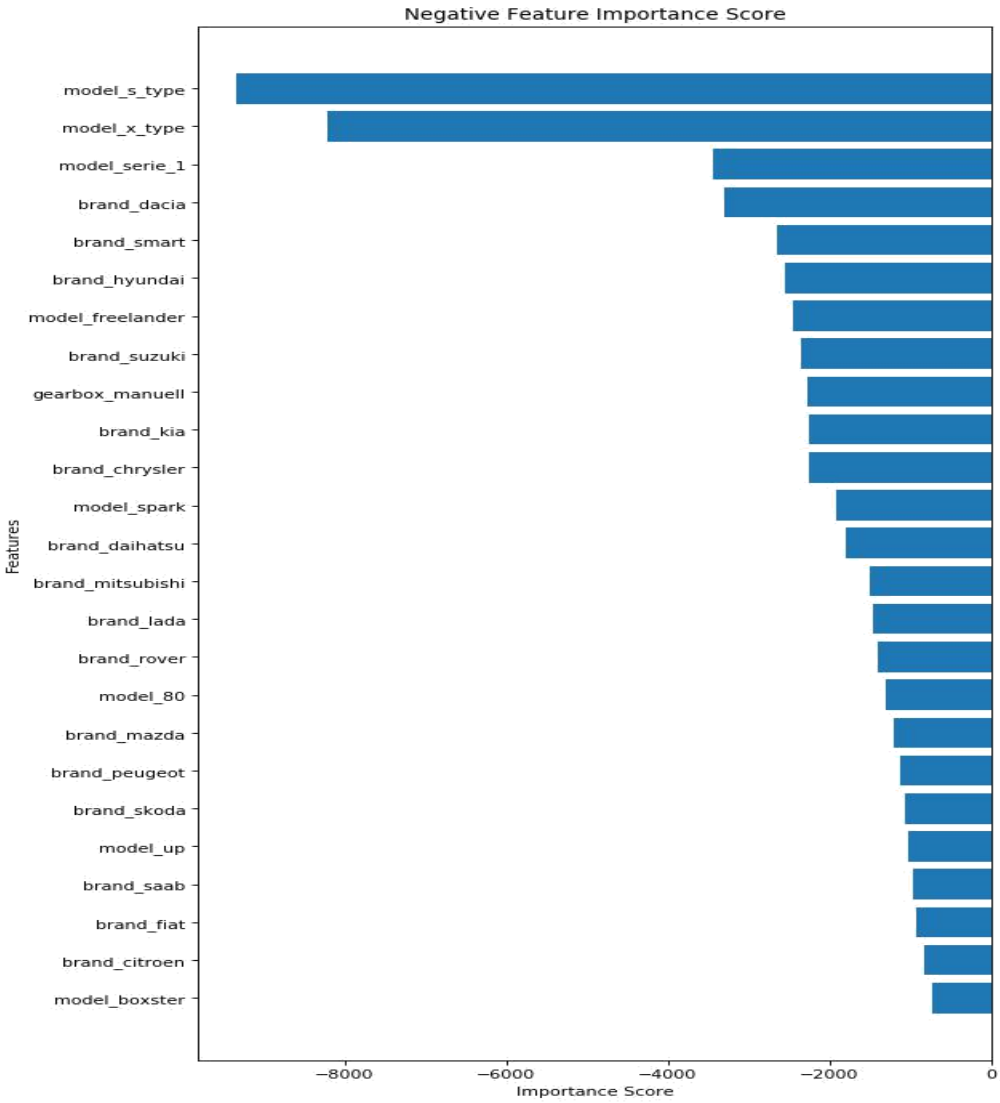
The linear model between actual price and predicted price is shown below with the regression plot graph. The model is underfitting and it can be improved with the help of cross validation technique which split the train into k-fold subsets to have the accuracy of multiple subsets and test the model on average of the training subsets. This technique improves the accuracy of the model.



FEATURE IMPORTANCE

Regular regression coefficients describe the relationship between each predictor variable and the response. The coefficient value represents the mean change in the response given a one-unit increase in the predictor. Consequently, it's easy to think that variables with larger coefficients are more important because they represent a larger change in the response. Hence, we have plotted top 25 positive and negative coefficients of the model that have the more importance as any change in these features will have a larger change in the price of used car




Negative Feature Importance Score
model_s_type
model_x_type
model_serie_1
brand_dacia
brand_smart
brand_hyundai
model_freelander
brand_suzuki
gearbox_manuell
brand_kia
brand_chrysler
model_spark
brand_daihatsu
brand_mitsubishi
brand_lada
brand_rover
model_80
brand_mazda
brand_peugeot
brand_skoda
model_up
brand_saab
brand_fiat
brand_citroen
model_boxster
Features
−8000
−6000
−4000
−2000
0
Importance Score