FLIP ROBO

Car Price Prediction project

Submitted by:

POOJASHREE D S

ACKNOWLEDGMENT

- Ultimate guide to deal with text data by Shubham Jain featured in Analytics vidya
- Visualizing text data using wordcloud by Deepika Singh featured in Pluralsight.

INTRODUCTION

- Business Problem Framing

  The main aim of the project is to predict the price of the used car available in the various website. Due to the Covid 19 we have seen lot of changes in the market.

- Conceptual Background of the Domain 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..

- Review of Literature

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States. Our results show that Random Forest model and K-Means clustering with linear regression yield the best results, but are compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

- Motivation for the Problem Undertaken

  Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately[2-3]. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices

Analytical Problem Framing

- Data Sources and their formats

  Data source: Olx and Car24.

The is mainly collected from the olx and the car 24 websites then i merged both the dataframe to form the new data frame

- Data Preprocessing Done

  First check the all the missing values using isnull() and only the brand column consists of both year and Transformation information. So the year and transformation columns are extracted from the brand column.

And the checked the type of the data, most of the columns are not assigned to the right data type

- Data Inputs- Logic- Output Relationships

  Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

  Here, you can describe any presumptions taken by you.

- Hardware and Software Requirements and Tools Used

  pandas and NumPy are the basic libraries imported. Matplotlib.pyplot and seaborn are used for visualization.

  Sklearn is used for pre-processing and model building steps. All algorithms for processing are imported. Sklearn.metrics will provide the needed evaluation metrics such as accuracy score, classification report and confusion matrix.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The size of the dataset is about six thousand and the dataset does not contain many features, it has nine column of data.

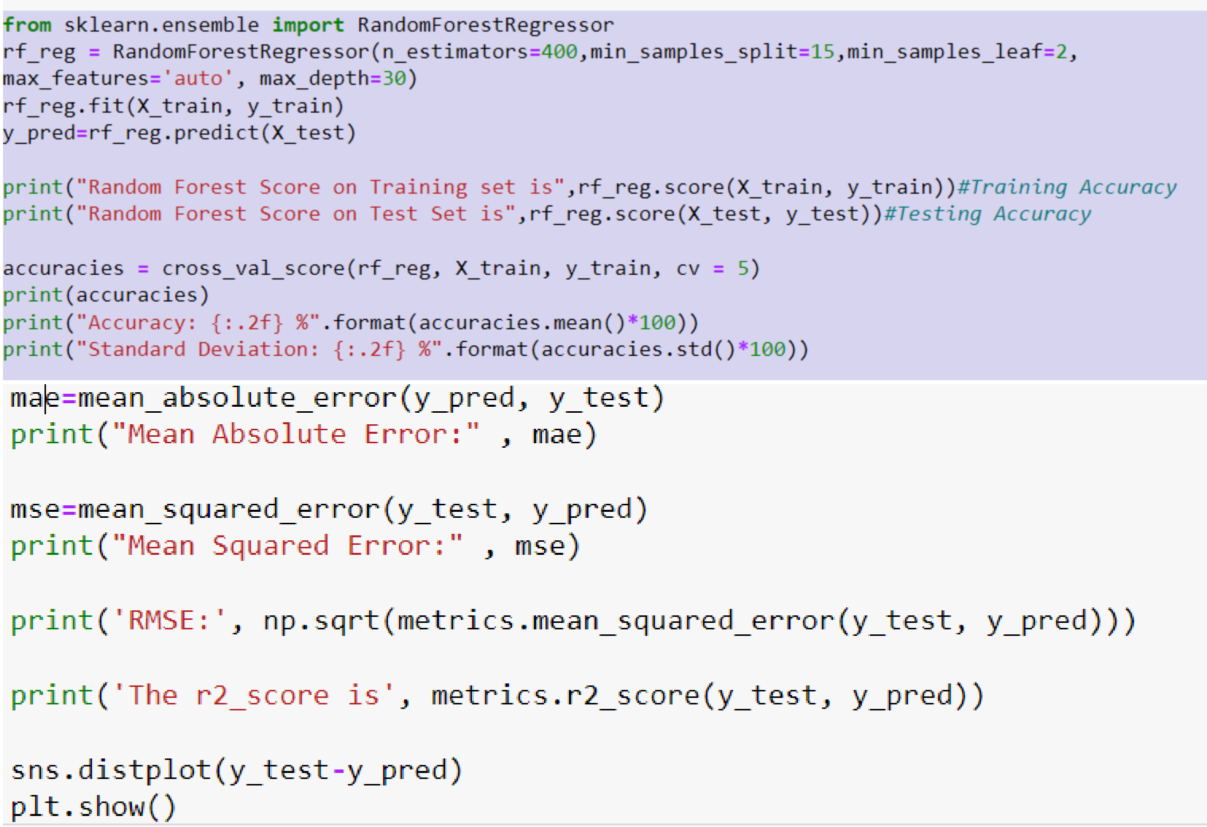
- Testing of Identified Approaches (Algorithms)

  Decision Tree Regressor

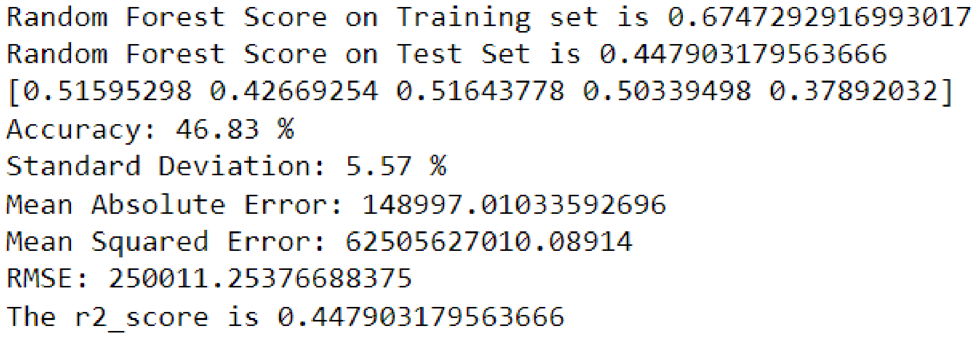
Random Forest Regressor

Voting Regressor

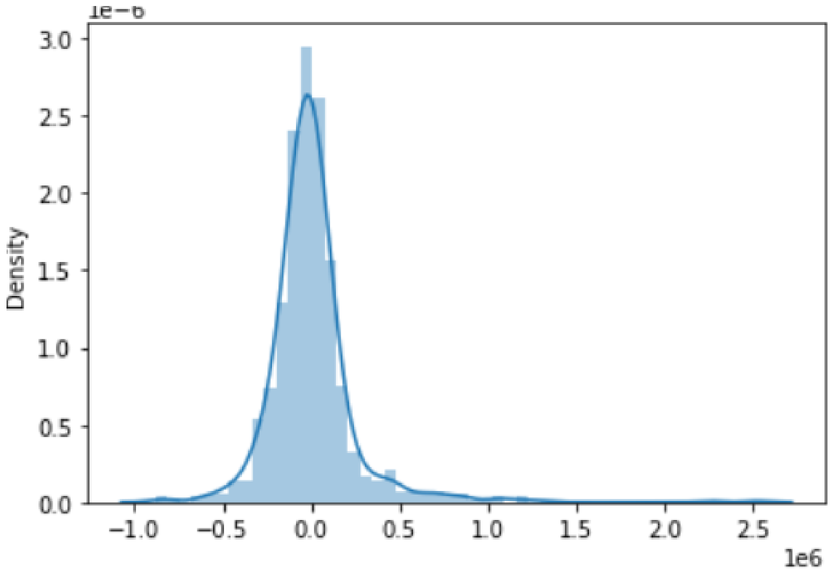
- Run and Evaluate selected models
- The above algorithms are fitted to x_train and y_train which is generated by train test split method. The performance of the model is validated based on the predictions made for x_test.

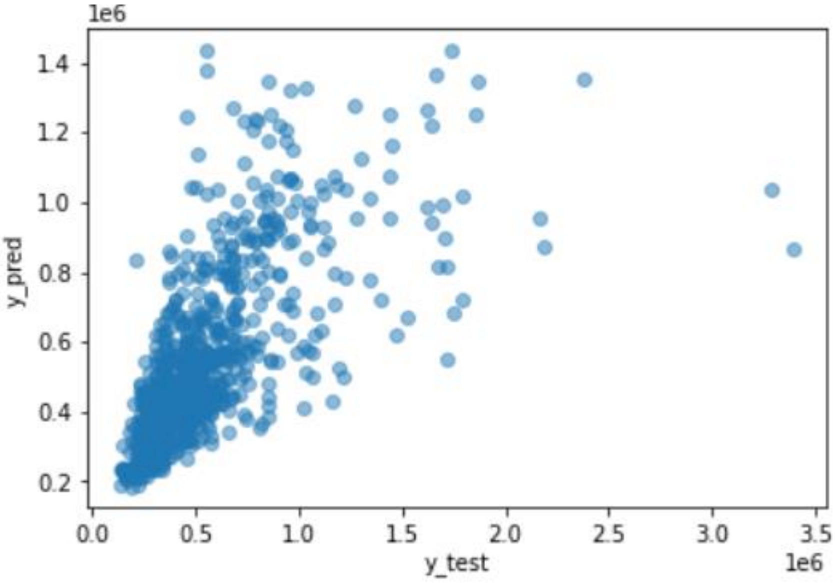


```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor(n_estimators=400,min_samples_split=15,min_samples_leaf=2,
max_features='auto', max_depth=30)
rf_reg.fit(X_train, y_train)
y_pred=rf_reg.predict(X_test)

print("Random Forest Score on Training set is",rf_reg.score(X_train, y_train))#Training Accuracy
print("Random Forest Score on Test Set is",rf_reg.score(X_test, y_test))#Testing Accuracy

accuracies = cross_val_score(rf_reg, X_train, y_train, cv = 5)
print(accuracies)
print("Accuracy: {:.2f} %".format(accuracies.mean()*100))
print("Standard Deviation: {:.2f} %".format(accuracies.std()*100))
mae=mean_absolute_error(y_pred, y_test)
print("Mean Absolute Error:" , mae)

mse=mean_squared_error(y_test, y_pred)
print("Mean Squared Error:" , mse)

print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

print('The r2_score is', metrics.r2_score(y_test, y_pred))

sns.distplot(y_test-y_pred)
plt.show()
```



```
Random Forest Score on Training set is 0.6747292916993017
Random Forest Score on Test Set is 0.447903179563666
[0.51595298 0.42669254 0.51643778 0.50339498 0.37892032]
Accuracy: 46.83 %
Standard Deviation: 5.57 %
Mean Absolute Error: 148997.01033592696
Mean Squared Error: 62505627010.08914
RMSE: 250011.25376688375
The r2_score is 0.447903179563666
```



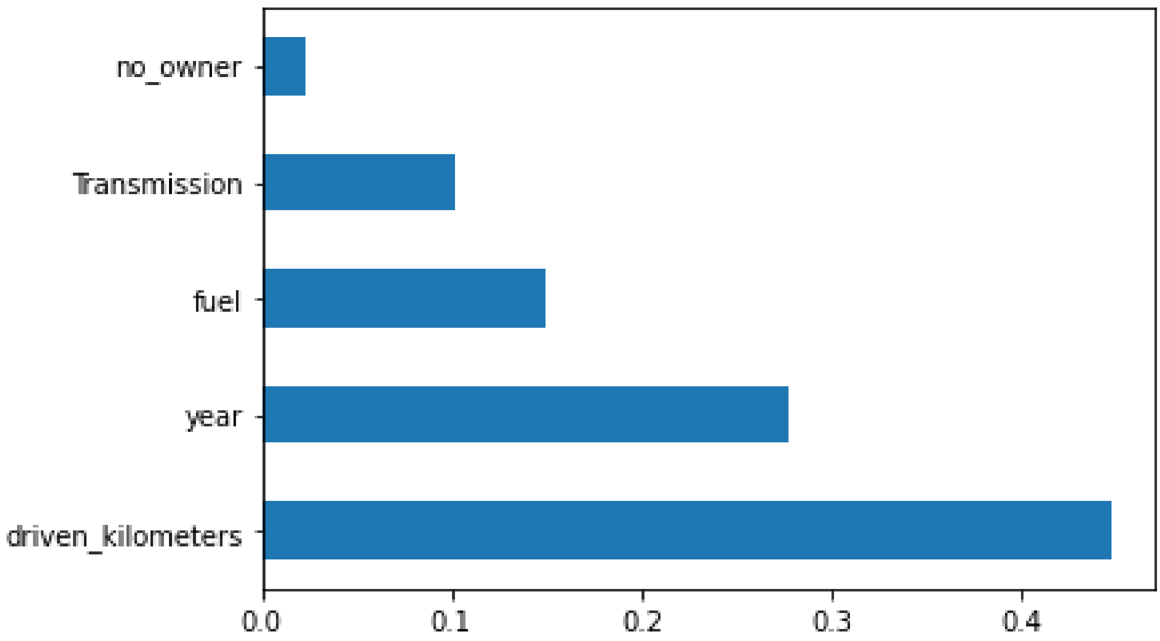


- Key Metrics for success in solving problem under consideration

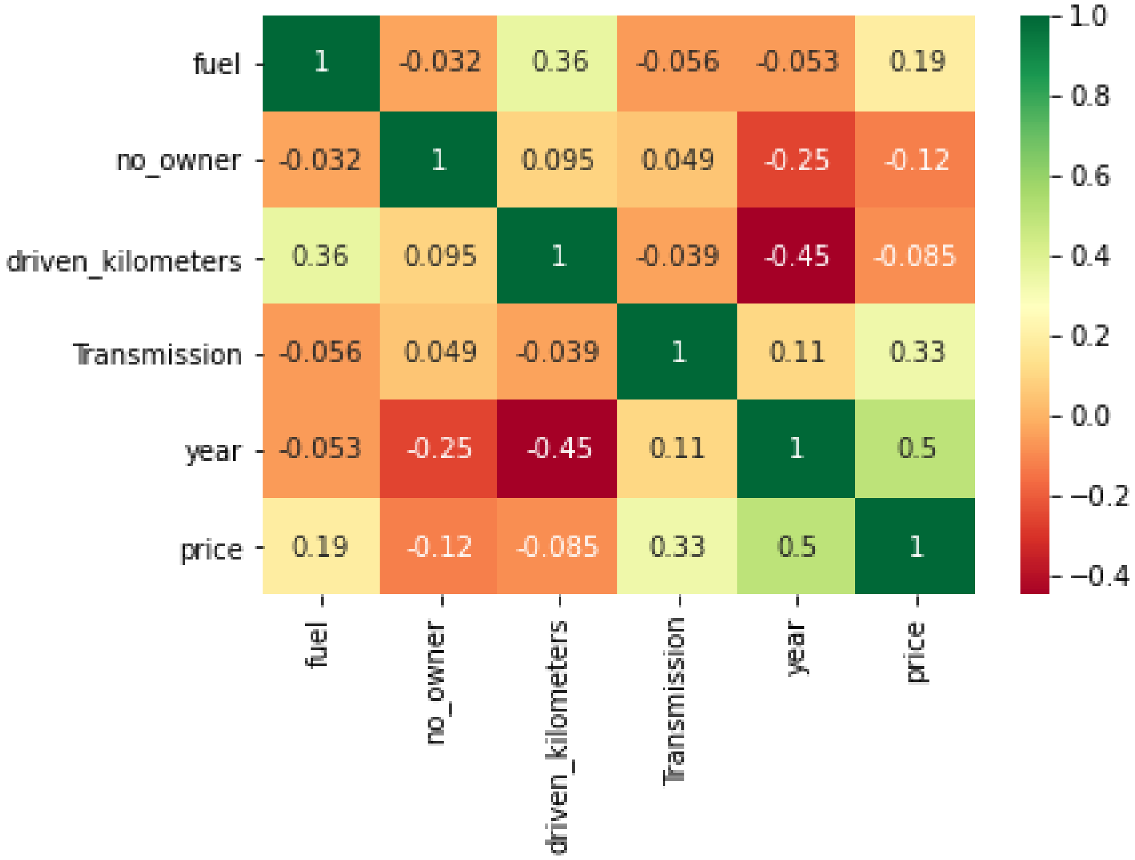
  Random Forest Regressor gives better accuracy than oher two regressor. The metrics used are accuracy score, classification report and confusion matrix. Accuracy score gives the performance of both classes as a whole. Classification report comprising of precision, recall and f1 score help us in understanding the efficiency of model with respect to each class. Confusion matrix places a distinction between false positives and false negatives.

- Visualizations

  Histograms are generated to understand the distribution of word count in different star ratings.



The above graph shows that driven kilometre of the car effect more on the price and the year of manufacture als effect the price of the car



The above picture shows the correlation between the features and the target as per the above picture driven kilometre and year are more negitivey correlated and more positively correlated with the target

CONCLUSION

The aim of our problem was to successfully predict the price of used car. The dataset is obtained by scraping data off Olx and car24 using Selenium. It is regression kind of sample using the Random forest regressor which gives better acuuracy able to predict the price