FLIP ROBO

RATING PREDICTION PROJECT

Submitted by:

Rijul Kumar

ACKNOWLEDGMENT

Reference that i have used are:

- Data Trained Education online video
- Materials provided by Flip Robo
- Data extracted from www.amazon.com and www.flipkart.com
- Geeks for Geeks
- Stackoverflow

INTRODUCTION

- ## Business Problem Framing

  1. There is website where people write different reviews for technical products. Now there is a new feature on the website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars.

  2. Now we want to predict ratings for the reviews which were written in the past and we don't have a rating. So, we have to build an application which can predict the rating by seeing the review.

  3. We are supposed to scrape at least 20000 rows of data for the model.

- ## Conceptual Background of the Domain Problem

  The rating can be derived from reviews with the help of some factors. Such as:

  - Some data with both review and ratings
  - Certain recurring words in reviews such as good, bad, etc
  - And so on...

- ## Motivation for the Problem Undertaken

  - We need to scrape the reviews of different laptops, Phones, Headphones, smart watches, Printers, etc.
  - We have to scrape at least 20000 rows of data.
  - This model will then be used to predict ratings for the reviews which were written in the past and for which we don’t have a rating.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem
  - First of all i imported data from excel file to dataframes using pandas.
  - I used .dtypes to know data type of each column of dataframe.
  - After that i used .describe() to know the statistical information (such as max, min value,etc) of continuous data columns in dataframe.
  - Then i used .shape to know shape of dataframe.

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

  - Using Selenium, data was extracted to excel file named ‘Ratings_prediction_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 i removed punctuations and converted reviews into lower case alphabets.
  - Next i tokenized the reviews and removed the stop words from them.
  - After that i applied both Stemming and Lemmatization method.
  - Afterwards i focused on lemmatized output and interpreted them into Bag of words model.
  - Next i used Density plots from seaborn library to plot all continuous columns for visualisation.
  - After that i checked for correlations using heatmaps and correlation matrix.
  - After that i checked and removed skewness for continuous data columns (except target variable) but as almost all columns were skewed.
  - Finally i checked outliers using boxplot as well as z-score method and discovered lots of outliers.
  - As there was too much skewness and outliers so i chose to use tree based algorithms in model building phase rather than removing all those because it would lead to too much data loss.

- ## Data Inputs- Logic- Output Relationships

  - Data Input :

    These are basically the reviews which are to be rated by the machine learning model.

  - Data Output :

    Our Target variable is ‘Rating' which is the rating of the reviews.

- ## 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

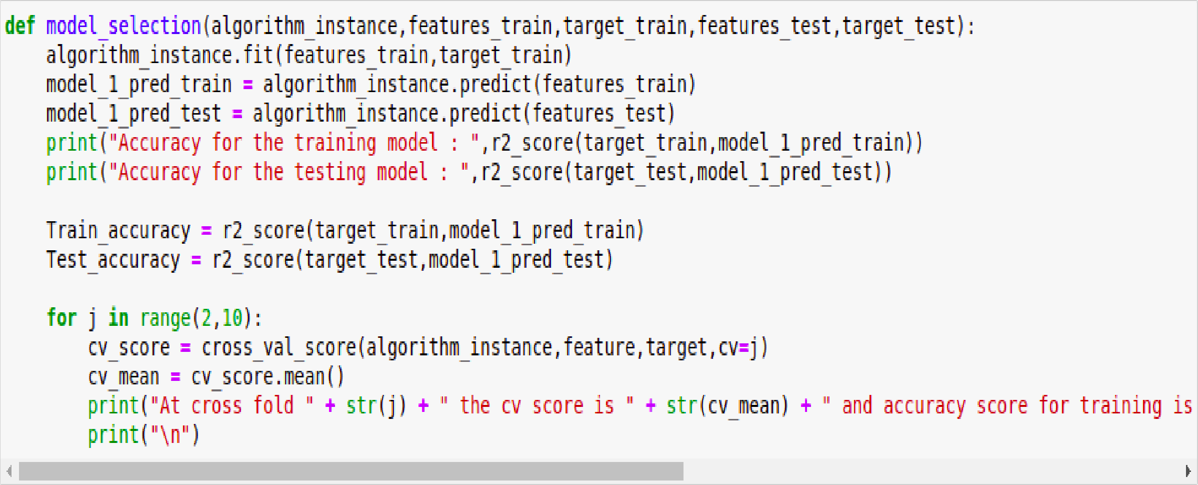
- Identification of possible problem-solving approaches (methods)
  - After preprocessing data (removing NULL, encoding, checking for high correlations, removing skewness and outliers) i separated columns into features and target.
  - As this is a regression problem so we tried 4 models - LinearRegression, XGBRegressor, 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
  - XGBRegressor
  - RandomForestRegressor
  - DecisionTreeRegressor

- <u>Run and Evaluate selected models</u>
  I defined a function model and then tried 4 different models using it



```
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 :

- <u>LinearRegression</u> :
  At random state 3 the training accuracy is : 0.10499536558166533
  At random state 3 the testing accuracy is : 0.09688953257873034

  At cross fold 2 the cv score is -7.148313218511615e+24 and accuracy score for training is -0.11419990928700341 and accuracy score for testing is 0.09688953257873034

- <u>XGBRegressor</u> :
  Accuracy for the training model : 0.17113317595280508
  Accuracy for the testing model : 0.17444362379688338

  At cross fold 2 the cv score is -0.12343403113445128 and accuracy score for training is 0.17113317595280508 and accuracy score for testing is 0.17444362379688338

- RandomForestRegressor :
  Accuracy for the training model : 0.18998468920871847
  Accuracy for the testing model : 0.17938940743021847

  At cross fold 2 the cv score is - 0.10493400123583363 and accuracy score for training is 0.18998468920871847 and accuracy score for testing is 0.17938940743021847

- DecisionTreeRegressor :
  Accuracy for the training model : 0.20051157005475173
  Accuracy for the testing model : 0.13323484216975012

  At cross fold 2 the cv score is - 0.11696279466012482 and accuracy score for training is 0.20051157005475173 and accuracy score for testing is 0.13323484216975012

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

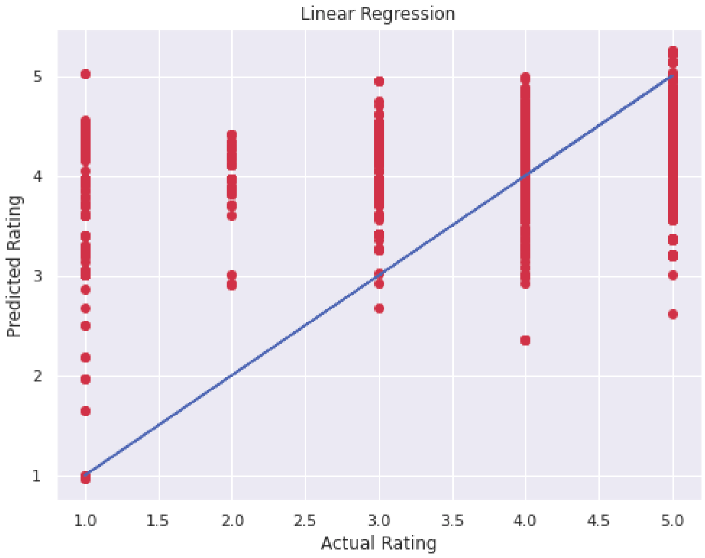
I tried 4 different metrics method:

- 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 do not lie near best fit curve.



- ## Interpretation of the Results

  So the results which i got were:

  - Although RandomForestRegressor() is best model but DecisionTreeRegressor() can also be used as it is also very good model.
  - 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 18.12% after hypertuning.

CONCLUSION

- ## Key Findings and Conclusions of the Study

  - From this study i learnt that many users use certain words a lot while writing reviews such as good, worthless, etc.
  - There can also be many fake reviews whose authenticity we can not check among all the data gathered.
  - Many users have given different ratings with the same review such as 2 people have given 3 star and 4 star for a certain product and the review which they have given for the same product is ‘good product’.

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

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

  - There was skewness and outliers in almost all columns removing which would have created huge data loss (88.1%). So i chose to use tree based algorithm in model building phase rather than removing skewness and outliers as tree based algorithms are not affected by them.

- ## 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 review ratings such as authenticity of review, etc.
  - With evolving technology, there may be increase in bot reviews which are also required to be taken care of.

  Scope for future work :

  - This can be made further accurate by taking more and more factors as well as authenticity of user / review into account.