Explaining Assignment Rapyder

The problem statement was to improve the reading book platform where I need to build the models which can predict the Genre of the book and the rating of the book.

1.) Prediction of Genre:

First, I did the **Data Analysis**, the data had shape of (1539, 8) with no null values. Then proceed with checking data types of columns, and checking the unique values. Interestingly all titles were unique. Genres were 10 and Thriller was the most repeating Genre. Then, I convert num_ratings, num_reviews, num_followers into numeric by removing the string parts.

I did the **Data Visualization** part, by observing that most ratings were around 4 to 4.25. I also tried to find some hidden information e.g., which title receive the max number of ratings, and top titles with max number of reviews visually.

Then I begin with Genre prediction. I predict the genre with the help of synopsis. I understood that this is a NLP Problem where I used the NLTK library to **pre-process** the Text.

I begin with converting whole synopsis part into some meaningful information in one go. First, I removing the special char part from text using regex. then converting into lower case and splitting into tokens so that I can remove the stopwords such as a, an, I etc.

I also used Stemming on the text. It is a process of extracting the root word and removing the rest. Here we used **Porterstemmer** library from **nltk**. After done with pre-processing I did the append them into a whole single list named ‘corpus’. I also mapped the genre into numeric values.

Since ML couldn’t understand the text, we need to convert it into numeric data. I used **TF-IDF vectorizer** for this purpose. TF-IDF is used to assign a value according to the importance of the document in corpus which removes natural occurring words in English.

TF(term frequency) = number of times appear in a document / total no of times

IDF(inverse-document frequency) = log_e(total no of documents/ no of doc with terms in it.

The train feature was the vectorized data with max features of 10000, and the target feature was the mapped values of Genres.

After splitting the data into train and test, I used **Multinominal Naïve bayes** model and **Linear SVM**. With the help of Hyperparameter tuning I received better accuracy from MultinominalNB which is around 77% compare to Linear SVC which is 74%.

I did my final prediction with MultinominalNB model, with alpha value of 0.05.

However, the accuracy can be increased if we could acquire some more data because as I discussed earlier that 2-3 genres were making more than half genres so data was kind of imbalanced. Also, if we could use any deep learning method which could increase our accuracy and prediction ability of model.

Also, the model couldn’t able to understand the difference between Thriller, horror and Romantic, Fantasy since they contain very similar words.

2.) Prediction of Rating:

To Predict the Ratings, I used the num_ratings, num_reviews, num_followers and genre as training features. I did **pre-processing** of data by removeing the **outliers** from these features by observing them visually with the help of **boxplot** and remove them. I also **onehot encode** the genres with the help of **get_dummies** function in **pandas** library. So, these were the training features and the target feature was rating.

After splitting them into train and test I use the **Linear Regression** and **Random Forest Regressor**. Where Random Forest outperformed Linear Regression. I also showed them Visually on notebook as well.

Metrics of Linear regression were –

```
MAE: 0.1529693779157134
MSE: 0.04085601719829251
RMSE: 0.2021287144328893
R2_value: 0.17542059178705938
```

And Metrics of Random Forest Regressor were –

```
MAE: 0.14517398119122257
MSE: 0.03668871332288401
RMSE: 0.19154298035397696
R2_value: 0.2595274920448615
```