AL/ML – DA – ASSIGNMENT - RAPYDER

According to the Problem Statement I understood that to improve the online reading book platform they want to add some features to the platform for readers who can easily get the book they want according to their requirements. I have observed that I need to build a model which can predict the ratings for sorting out the books according to the highest rating. Also, prediction of genre of the book.

I have written the code using Jupyter Notebook by installing and importing numerous libraries which helps I building the effective code.

According to the Data Science cycle I have performed the below steps:

1. Business Understanding
2. Data collection
3. Data Understanding
4. Data Preparation
5. Model Building
6. Model Training
7. Model Testing
8. Model Evaluation

Business Understanding:

We need to Build a model in order to predict the Ratings and genre of the book.

So, here by using the synopsis we can predict the genre of the book also building the model with best accuracy which can predict the ratings of the book.

Imported Necessary Libraries to build the code.

Pandas for the data manipulation, NumPy used for numerical operations.

For, Data Visualization imported Matplotlib to plot the graph used pyplot and used seaborn for the beautification of the graph.

Imported nltk (natural language tool kit which provides various text processing libraries in order to remove stop words and performed Lemmatization and stemming on synopsis for genre prediction.

For Model building imported form the sklearn library.

Data Collection:

The Data Frame consists of title, rating, name, num_ratings, num_reviews, num_followers, synopsis, genre.

AL/ML – DA – ASSIGNMENT - RAPYDER

Data Understanding:

Initial Data Analysis

The book dataframe consists of 1539 rows and 9 columns. Checked whether the data has null values. As, they are no null values present in the data.

There is an unnecessary column unnamed. So removed the column by using iloc method.

Described the data and get to know the count, mean, median, max. min values of ratings.

I want to check whether the book titles all are same or unique. So, I checked using a unique () method and came to know that they are no unique book titles.

Checked the same for the genre to know how many unique genres are there and came to know that they are 10 unique genres which are

```
'fantasy', 'history', 'horror', 'psychology', 'romance', 'science',
      'science_fiction', 'sports', 'thriller', 'travel'].
```

To know more I looked into no of books each genre has and thriller has the hi ghest books 481 and science_fiction has the lowest 45.

checked the datatypes of the book dataframe. Observed that num_ratings, num_ reviews, num_followers datatype shows object but, in the data, we can see num bers and there is no string contain.

In the data num_ratings and num_reviews there are commas where it identifies as object and in the num_followers we have both float and string in the data. Ne ed to convert them into Integer int64.

So, we need to clean the data and prepare for visualization and model building.

Data Preparation:

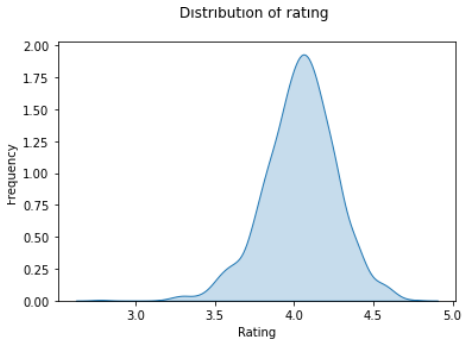
Removed commas by using replace method. Here created a function to convert num_followers into int. As, it contains both float and int. By importing re (whic h is regular expression which checks the string matches the regular expression) and s [-1] which address the last index k in the value. Function returns to the val ue and multiplied with 1000 where k values handle which converts into integer. Also, by using to_numeric converted num_ratings and num_reviews into Intege r.

AL/ML – DA – ASSIGNMENT - RAPYDER

Checked the Datatypes and now the data is cleaned for building the model. No n ull values present and removed the unneccasary columns.

Data Visualization performed for better Understanding of the data.

Distribution to ratings throughout the plot check the average ratings. The plot



It ranges from approximately 3.2 to 4.7 and the highest average rating given is 4.0.

To know how many numbers of books each with their respective rating. In whic h it can observed that the average rating. They are approximately 25 books with 4.0 rating. Similarly, we can observe that for different ratings and number of bo oks.

To know more about the popular book depends on the number of ratings. High r ating books. In the plot we can see that the harry potter and the sorcerer's stone.

Also, checked the book which has higher reviews. By observing plot, we can se e that hunger games have highest number of reviews. Similarly checked the boo ks with number of followers and top 20 books published. Also, customers who f requently read the books of particular author.

Using synopsis to predict the genre. Data Pre-processing done for Cleaning the synopsis, performed stop words removal, lemmetization and stemming. In lemm etization used wordnet lemmatizer to extract base form of word. Used Porterste mmer to extract root word.

Model Building:

Fitting the models by training and testing. At, first splitted the model 80-20% an d performed on dataset. Tf-idf performed on synopsis for both xtrain and xtest. Using logistic regression fitted the model on train data. Making predictions and

calculating accuracy score also generated classification report. The accuracy sco re obtained is 61.3%.

Also used svm(support vector machine kernel ad linear) accuracy score obtaine d is 76.6 %.

We, again split the model with 80- 15% where logistic regression generated acc uracy score of 63.6 % and svm linear generated 73.5 %. As, we can see that whe n the model splitted to 80 – 20 % the svm gave the highest accuracy. Logistic re gression obtained different accuracy score depending on the randomness and by splitting of model.

Hence, by using the above models generated the actual and predicted genre of th e book.

As, you can see result that the model got confused by some words 2-3 books ge nerated with different predicted values as it may contain similar words. To han dle this more data of genres required.

In the Model prediction of Ratings used two models Linear Regression and Ran dom Forest. Linear Regression gave the best score which is 100 % in which mo del can predict the ratings. As, well Random Forest best score is 99.8 %. Predict ions done by using two models. Checked the model score on test data by using metrics residual sum of squares, mean squared error and accuracy score. In whic h linear regression gave the best accuracy score 1.0.

In linear regression the residual sum of squares is less and model fitted accuratel y. Hence the model shows train score and test score of linear regression and ran dom forest and plotted accordingly.

Hence, these are my understandings in which I performed in order to predict the genre of book and ratings.

AL/ML – DA – ASSIGNMENT - RAPYDER