Problem 1

Predicting the Genre of the book

Primary motive

- synopsis is the key feature to predict the genre of the book
- its also better to combine the name of the book and writer with the synopsis
- it has been noted that few fetures are in string format so those needs to convert into integer format
- need to use word to vector techniques and various ML models for each combination

Text Preprocessing

- book title, author and synopsis has been combined into a single column description and a function has been made

*data_pre is a function to

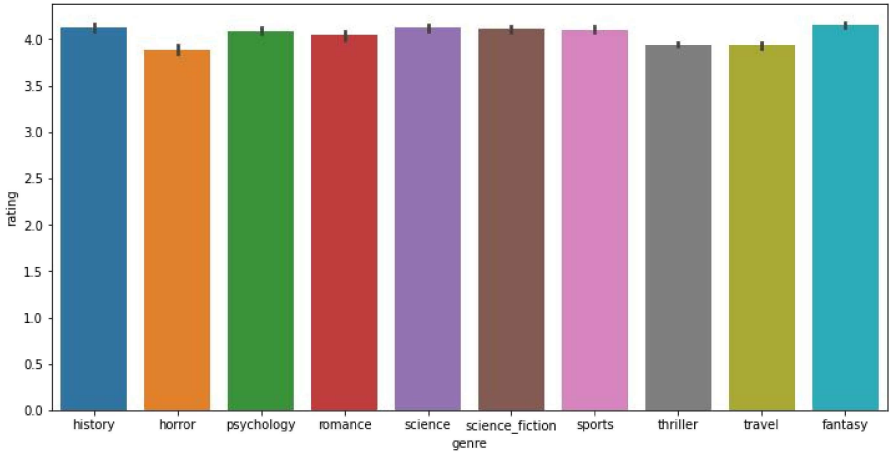
- clean the description
- to remove stop words in description
- to lemmatize the words in description.

* convert1 is a function to convert string values of no. of followers, no. of ratings and reviews to numerical values

Data Analysis

- Genre vs Rating

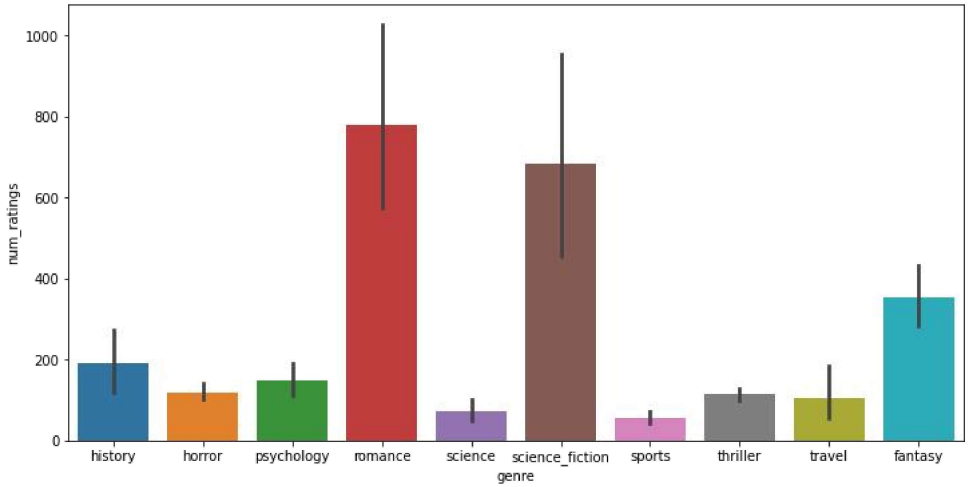
It has been observed that there is no importance of rating in predicting the genre, we can omit this feature



Genre vs No. of ratings

It has been observed that number of ratings is a key feature in predicting the genre of the book.

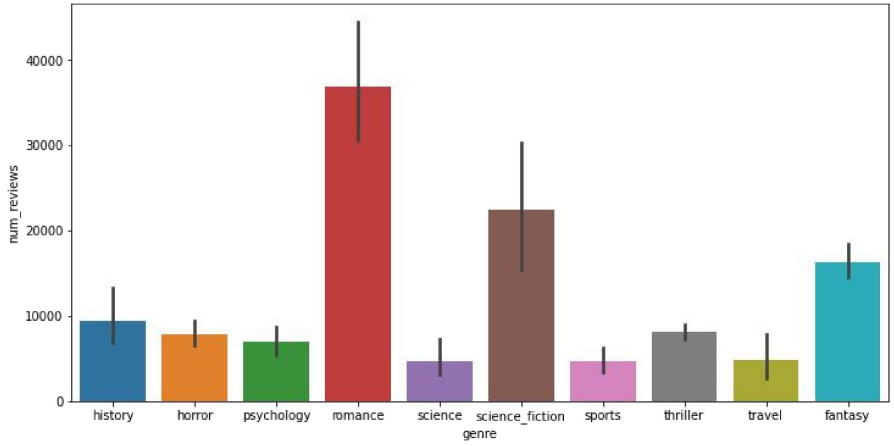
no. of ratings is very high for romance and science fiction genres



Data Analysis

- Genre v/s Number of reviews

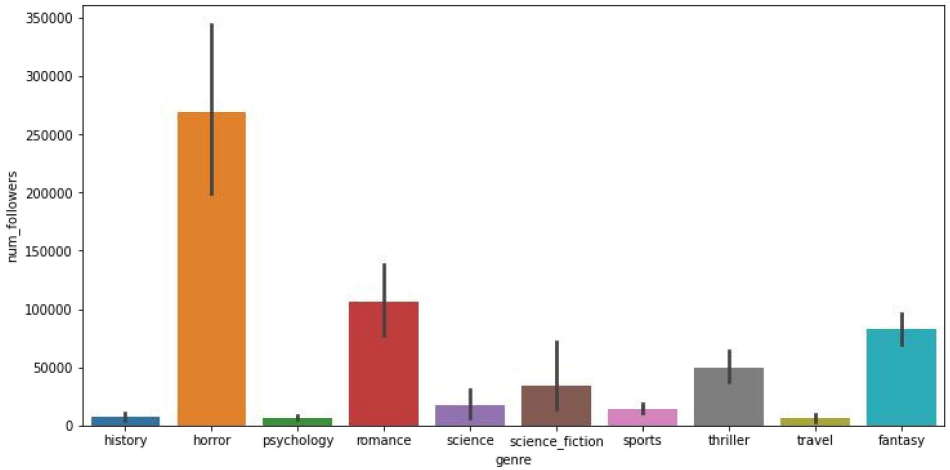
The role of Number of reviews in predicting the genre is almost identical to that of 'Number of ratings'



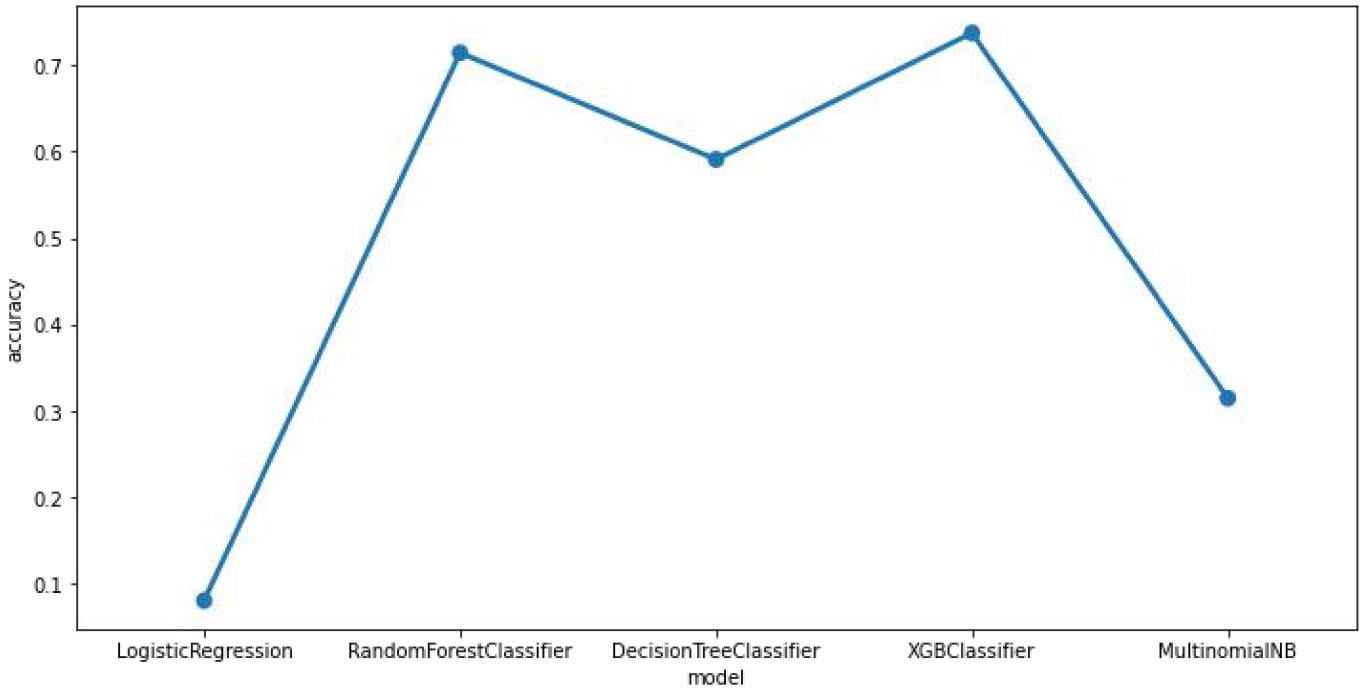
Genre v/s Number of followers

It has been observed that number of followers is a key feature in predicting the genre of the book.

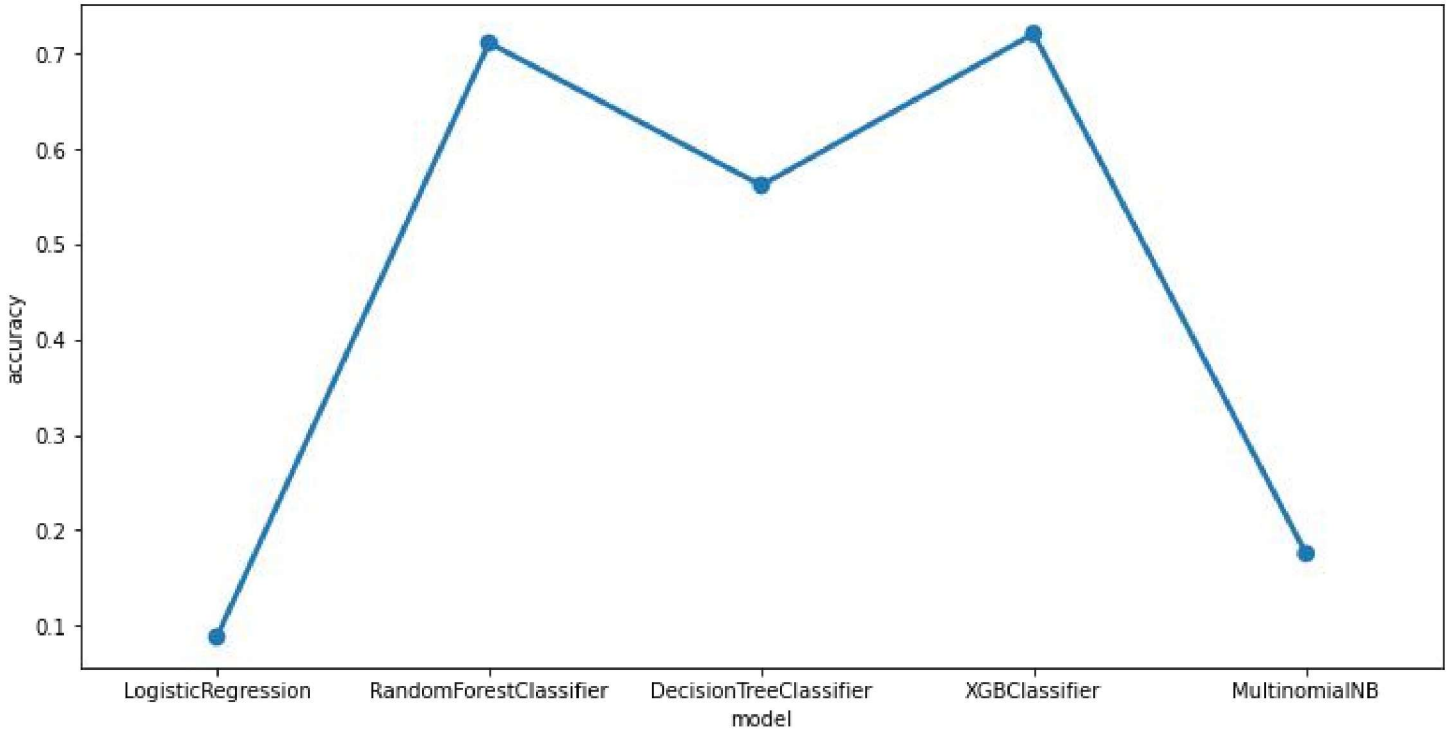
no. of ratings is very high for horror genre followed by fantasy genre



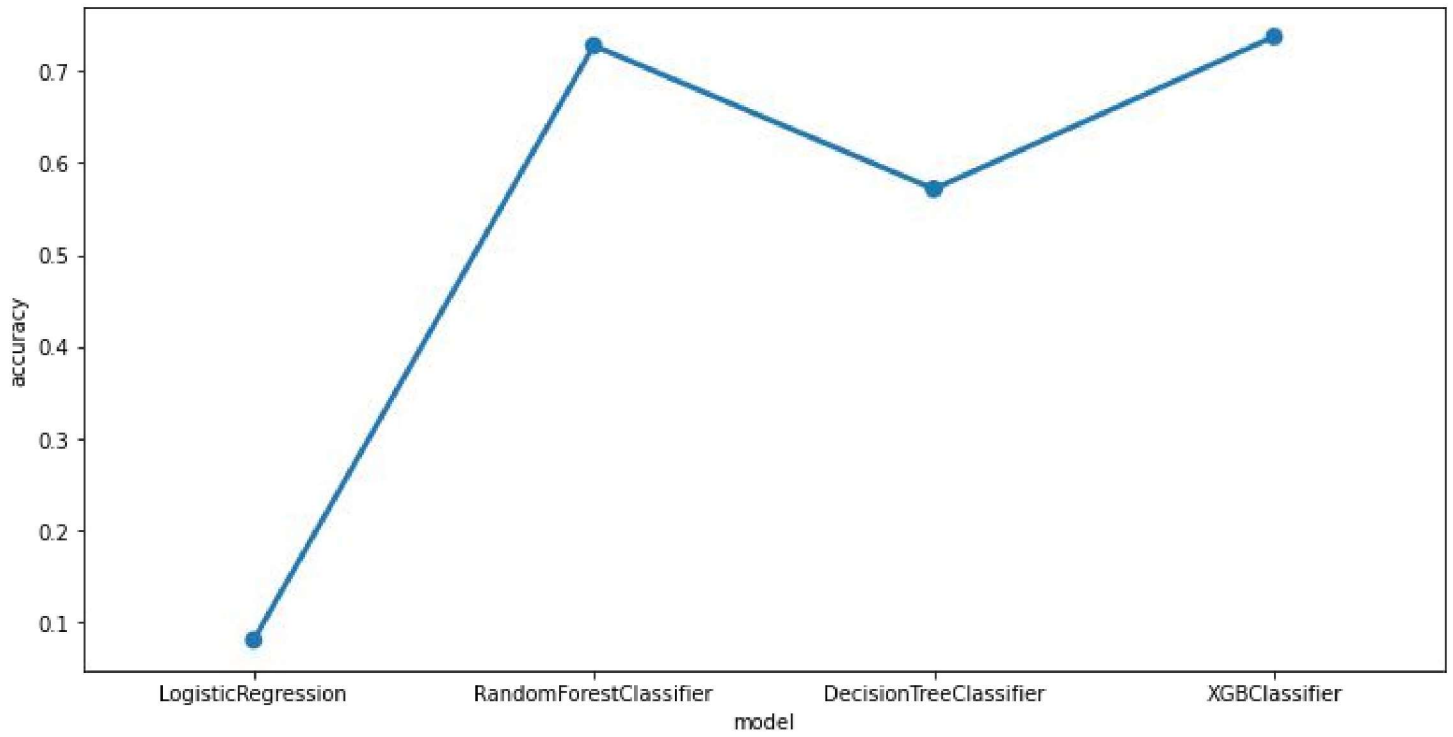
Using Countvectorizer and imbalanced data



Using TF-IDF vectorizer and imbalanced data



Using SPACY and imbalanced data

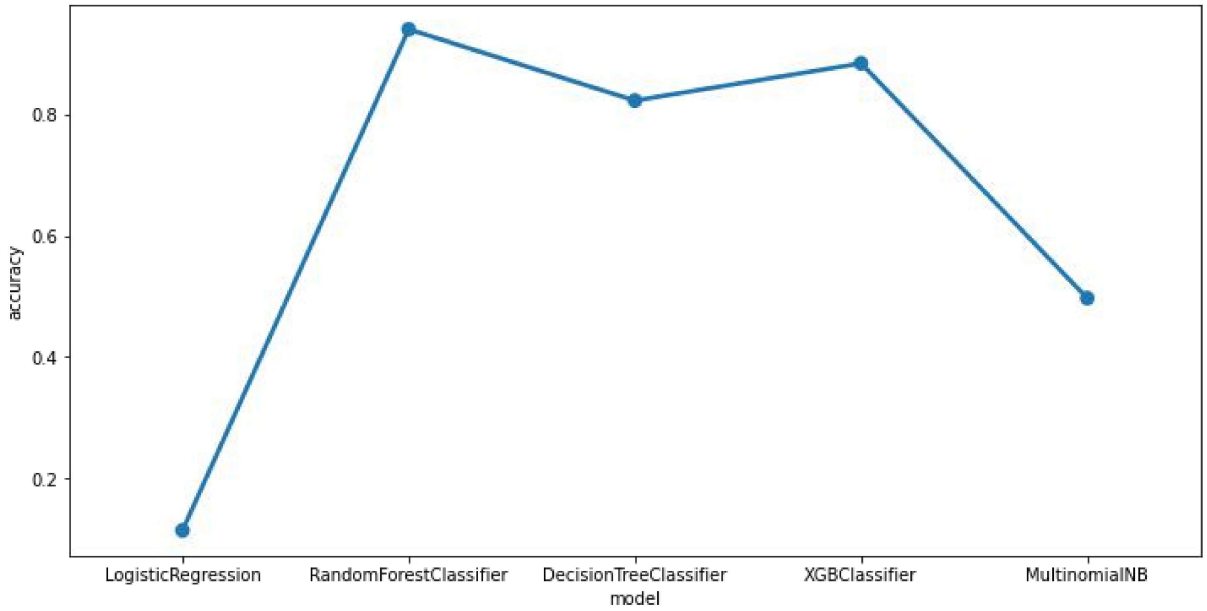


Balancing the imbalanced data set

- SMOTE technique has been used to balance the data set
- All sets of minority class records has been undergone SMOTE itertively which generates and provides the required data for the model.

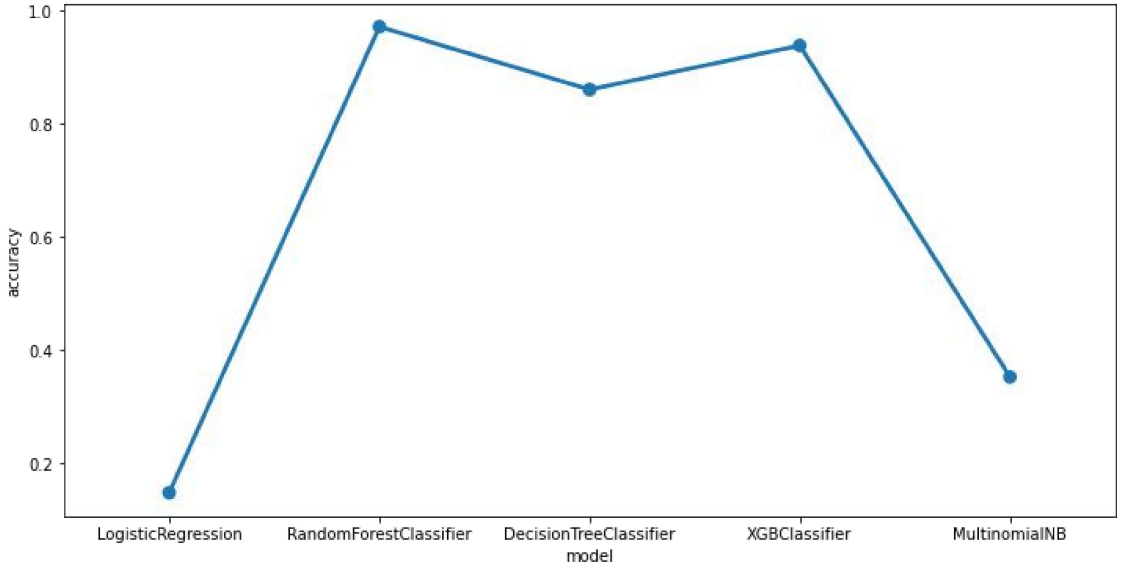
Using Countvectorizer and balanced data

- LogisticRegression 0.11434511434511435
- RandomForestClassifier 0.9397089397089398
- DecisionTreeClassifier 0.822453222453222
- XGBClassifier 0.8835758835758836
- MultinomialNB 0.4968814968814969



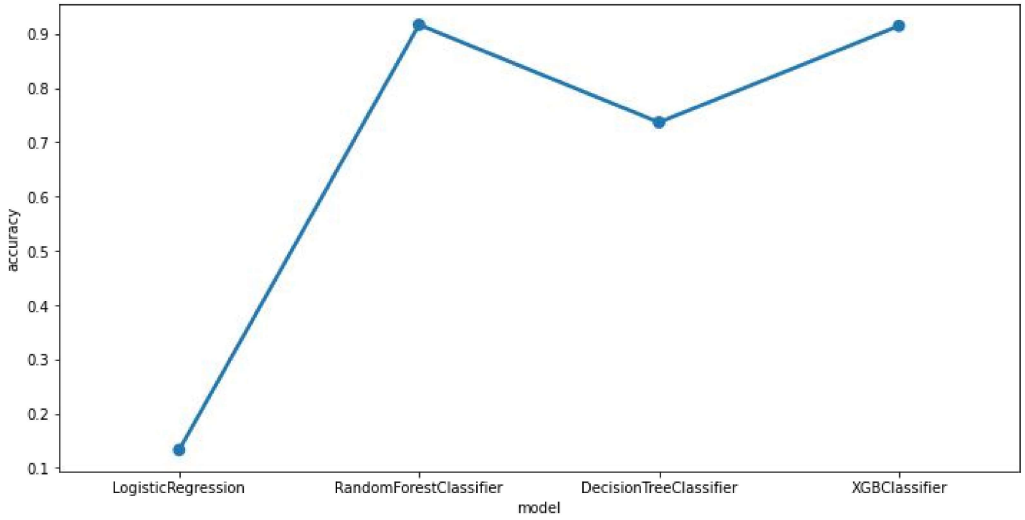
Using TF-IDF vectorizer and balanced data

- LogisticRegression 0.1476091476091476 2
- RandomForestClassifier 0.9719334719334719
- DecisionTreeClassifier 0.8607068607068608
- XGBClassifier 0.9386694386694386
- MultinomialNB 0.3523908523908524



Using SPACY and balanced data

- LogisticRegression 0.1330561330561330 6
- RandomForestClassifier 0.9168399168399168
- DecisionTreeClassifier 0.737006237006237
- XGBClassifier 0.9147609147609148



Conclusion

- RandomForest Classifier and XGB classifier with balanced data with TF-IDF vectorizer can be used to predict the genre.

- We can tweak the hyperparameters for further more accuracy.