Problem 2

Predicting the ratings 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

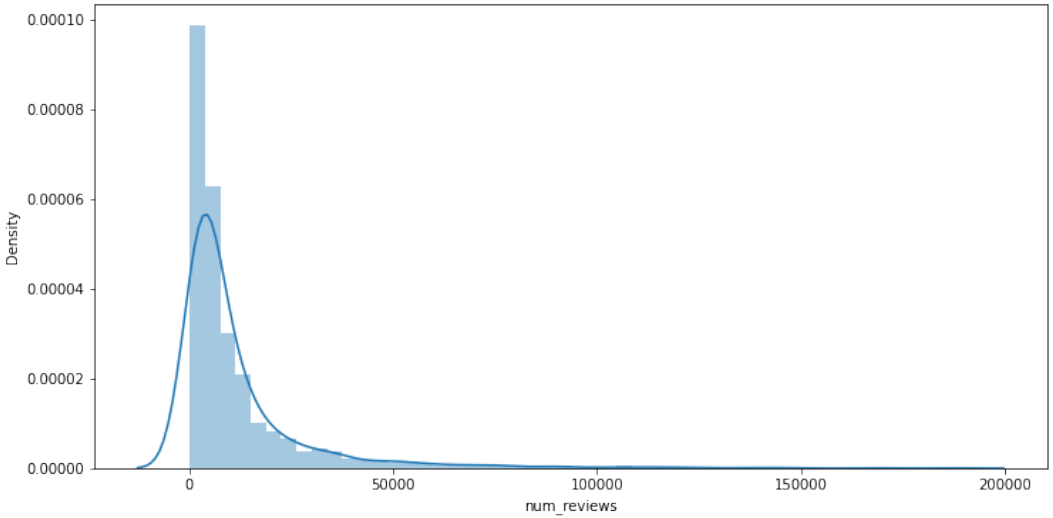
- The numerical features are not normally distributed and have some outliers in the data
- The previous model's data analysis helped to make some key changes in this model like:

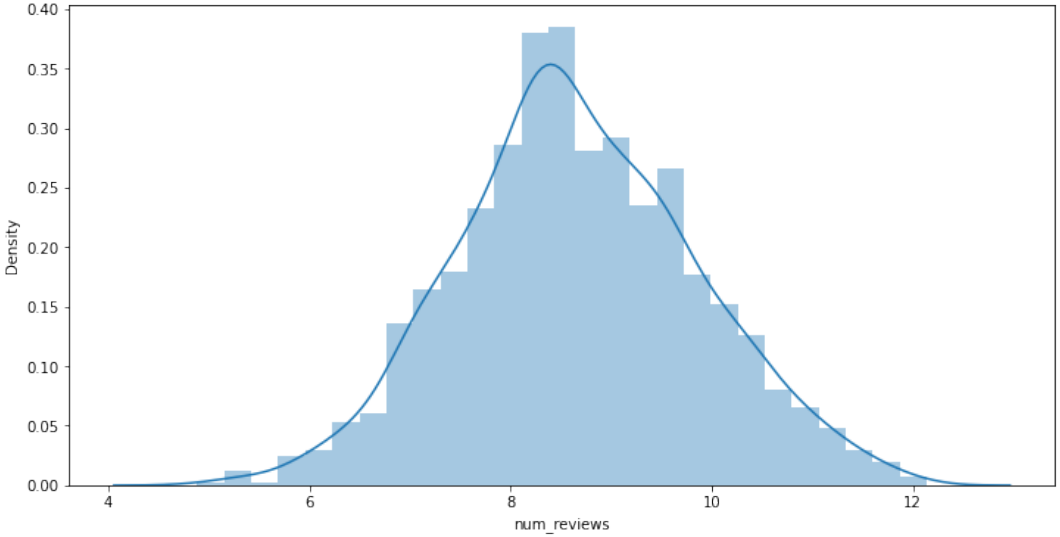
- removing Genre column (since no impact on rating)

- removing no. of ratings feature (multi-collinearity problem with no. of reviews feature)

Data Analysis

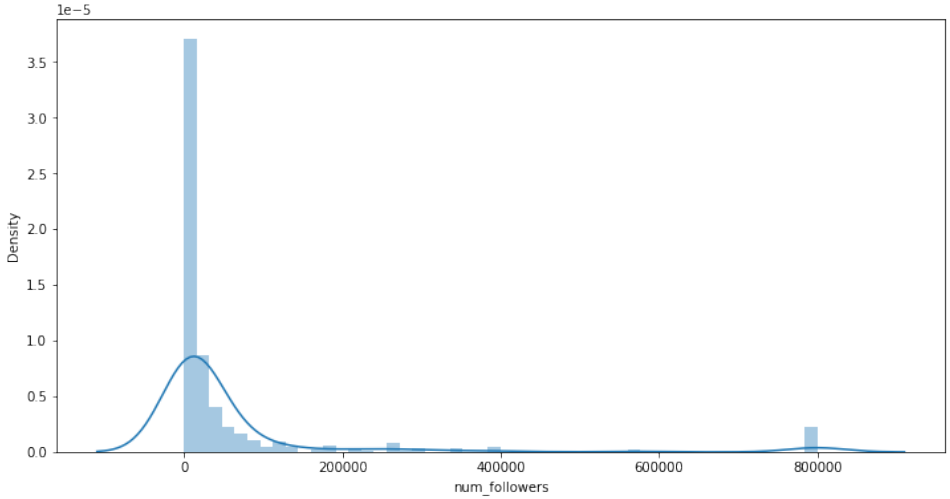
- no. of reviews follow the log normal distribution after applying log transformation

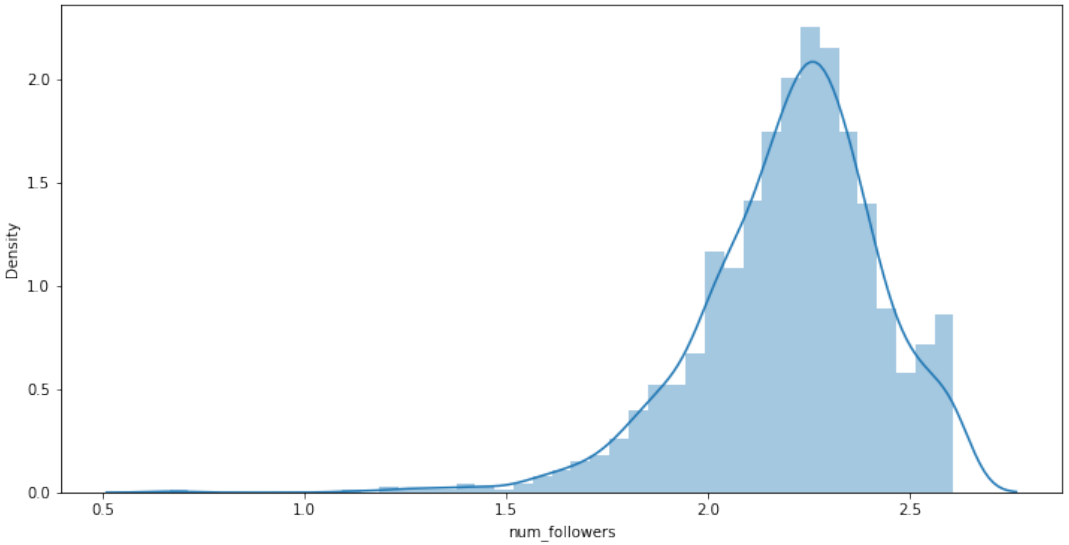




Data Analysis

- no. of reviews follow the log normal distribution after applying log transformation





Model building

- GridSearch CV techniques has been used to find the best parameters for the models used.
- Crossval score has been evaluated

- The models are performing very poor when TF - IDF has been used might be because of so many features

Models Results

- **Elastic net Regression with Spacy**

-r2_score : 0.15906696125727904

-mean squared error : 0.050722263898931 63

-root mean squared error : 0.225216038280872 94

-best parameters : {'alpha': 0.5, 'l1_ratio': 0.001}

Random Forest Regressor with Spacy

r2_score : 0.150042806441113 79

mean squared error : 0.051266570687893 91

root mean squared error : 0.226421224022603 3

best parameters : {'bootstrap': True, 'max_depth': 10, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 20}

Models conclusion

- Elastic net regression with Spacy did very well out of all models