PROJECT REPORT

Submitted By

Aswini Arun Patil

NAME OF THE PROJECT

Ratings Prediction

ACKNOWLEDGMENT:

- Primarily I would like to thank God to being able to complete this project with success. Then I would like to express my special thanks of gratitude to my SME,

- And I am thankful I am part of flib rob technology of employee, who given me the golden opportunity to do this wonderful project on the given topic which is also help me in doing a lot of research and I came to know about so many new things, I am really thankful to flip robo.

DATE:08/08/2022

ASWINI A. PATIL
Data Science course
Institute: Data trained education
Internship: Flip Robo technology
@Bangalore

INTRODUCTION:

Business Problem Framing:

- Rating prediction is a well-known recommendation task aiming to predict a user's rating for those items which were not rated yet by her. Predictions are computed from users' explicit feedback, their ratings provided on some items in the past.
- Another type of feedback are user reviews provided on items which implicitly express users' opinions on items. Recent studies indicate that opinions inferred from users' reviews on items are strong predictors of user's implicit feedback or even ratings and thus, should be utilized in computation.
- As far as we know, all the recent works on recommendation techniques utilizing opinions inferred from users' reviews are either focused on the item recommendation task or use only the opinion information, completely leaving users' ratings out of consideration.
- The approach proposed in this paper is filling this gap, providing a simple, personalized and scalable rating prediction framework utilizing both ratings provided by users and opinions inferred from their reviews.
- Experimental results provided on a dataset containing user ratings and reviews from the real-

world Amazon Product Review Data show the effectiveness of the proposed framework.

- We have a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their 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. Now they want to predict ratings for the reviews which were written in the past and they don't have a rating. So, we have to build an application which can predict the rating by seeing the review.

Conceptual Background of the Domain Problem:

- Compared to the traditional systems which mainly utilize user's rating history, review-based recommendation hopefully provides more relevant results to users. We introduce a review-based recommendation approach that obtains contextual information by mining user reviews.
- The proposed approach relates to features obtained by analyzing textual reviews using methods developed in Natural Language Processing (NLP) and information retrieval discipline to compute a utility function over a given item. An item utility is a measure that shows

how much it is preferred according to user's current context.

- all the recent works on recommendation techniques utilizing opinions inferred from user's reviews are either focused on the item recommendation task or use only the opinion information, completely leaving user's ratings out of consideration.
- The approach proposed in this report is filling this gap, providing a simple, personalized and scalable rating prediction framework utilizing both ratings provided by users and opinions inferred from their reviews.
- Recommendation systems are an important unit in today's e-commerce applications, such as targeted advertising, personalized marketing and information retrieval. In recent years, the importance of contextual information has motivated generation of personalized recommendations according to the available contextual information of users.
- Before customer buys any product, they look into the ratings of the product. How many people has given the review, ratings? Whether the product is good or not. Then they will decide to buy the product or not. So, rating the product is very important in business. Most of business firm

focused on customer service. Better the service more the customer. It also helps the business to understand what customer sees in the product, whether the customer likes the quality of product, cheaper price, long durable, offer on product etc. Based on this business can be focus in those area to give the better customer service.

Review of Literature:

- In real life, people's decision is often affected by friends' action or recommendation. How to utilize social information has been extensively studied. Yang et al.
- Propose the concept of "Trust Circles" in social network based on probabilistic matrix factorization. Jiang et al. propose another important factor, the individual preference. some websites do not always offer structured information, and all of these methods do not leverage user's unstructured information, reviews, explicit social networks information is not always available and it is difficult to provide a good prediction for each user. For this problem the sentiment factor term is used to improve social recommendation.
- The rapid development of Web 2.0 and e-commerce has led to a proliferation in the number of online user reviews. Online reviews contain a wealth of sentiment

information that is important for many decision-making processes, such as personal consumption decisions, commodity quality monitoring, and social opinion mining.

- Mining the sentiment and opinions that are contained in online reviews has become an important topic in natural language processing, machine learning, and Web mining.

Motivation for the Problem Undertaken:

- The dataset was provided by the Flip Robo Technologies. Models able to predict the user rating from the text review are critically important. Getting an overall sense of a textual review could in turn improve consumer experience. My motivation on this as this new data set that involves using the Natural Language Pre-processing technique used on the project. It helps me to work on removal of irrelevant words from the review.
- It is interesting to work on the project how customer gives reviews and ratings for the product. It is excited to build the model that predicts the ratings based on the review.
- Customer also shows their anger, frustration in the review if they didn't like the product of, they have expected. This kind of review will decrease the sell in business.

- The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary objective. Many product reviews are not accompanied by a scale rating system, consisting only of a textual evaluation.

• Data Pre-processing Done:

Data pre-processing is the process of converting raw data into a well readable format to be used by Machine Learning model. Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be

derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process our data before feeding it into our model.

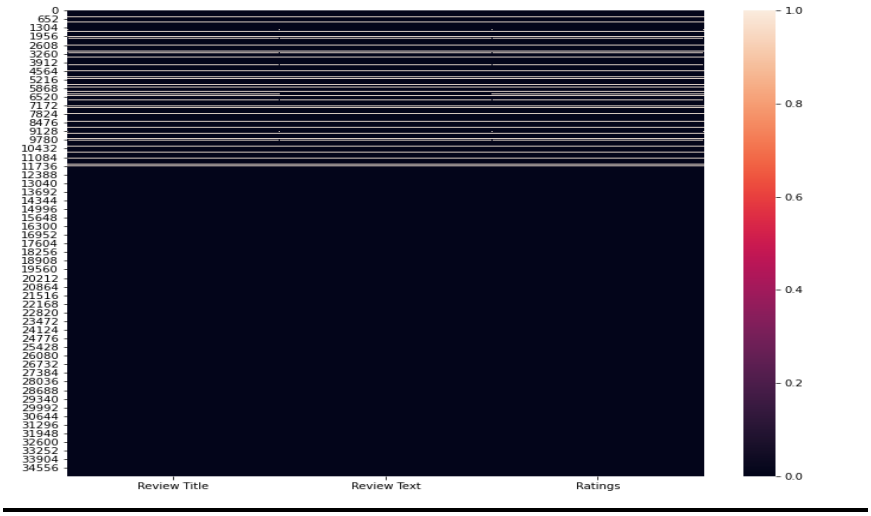
I have used following pre-processing steps:

- ✓ Importing necessary libraries and loading dataset as a data frame.
- ✓ Checked some statistical information like shape, number of unique values present, info, null values, value counts etc.
- ✓ Checked for null values and I replaced those null values using imputation method. And removed Unnamed: 0.

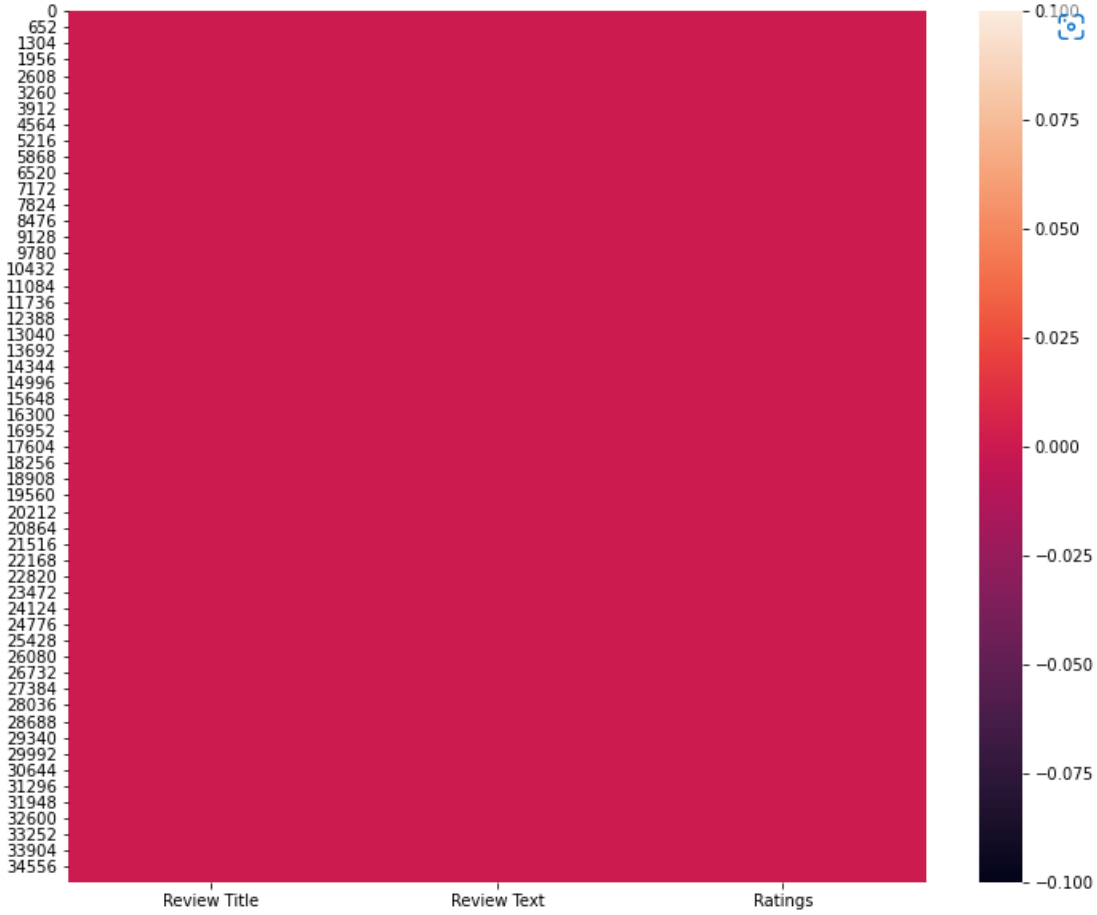
- ✓ Visualized each feature using seaborn and matplotlib libraries by plotting distribution plot and word cloud for each rating.
- ✓ Done text pre-processing techniques like Removing Punctuations and other special characters, Splitting the comments into individual words, Removing Stop Words, Stemming and Lemmatization.
- ✓ After getting a cleaned data used TF-IDF vectorizer. It'll help to transform the text data to feature vector which can be used as input in our 6 modelling. It is a common algorithm to transform text into numbers. It measures the originality of a word by comparing the frequency of appearance of a word in a document with the number of documents the words appear in.
- ✓ ü Balanced the data using SMOTE method.

Before removing the null value:

```
plt.figure(figsize=(12,10))
sns.heatmap(df.isnull())
plt.show()
```



Replace the null value with mode:



• Data Inputs- Logic- Output Relationships:

- The dataset consists of 2 features with a label. The features are independent and label is dependent as our label varies the values(text) of our independent variable's changes.
- I checked the distribution of skewness using distribution plots and used count plots to check the counts available in each column as a part of univariate analysis.

- Got to know the frequently occurring and rare occurring word with the help of count plot. And was able to see the words in the Review text with reference to their ratings using word cloud.

• Key Metrics for success in solving problem under consideration:

In this project I have used the Mean Squared Error, Mean Absolute Error, Hamming Loss, Classification Report, Confusion Matrix, Accuracy Score.

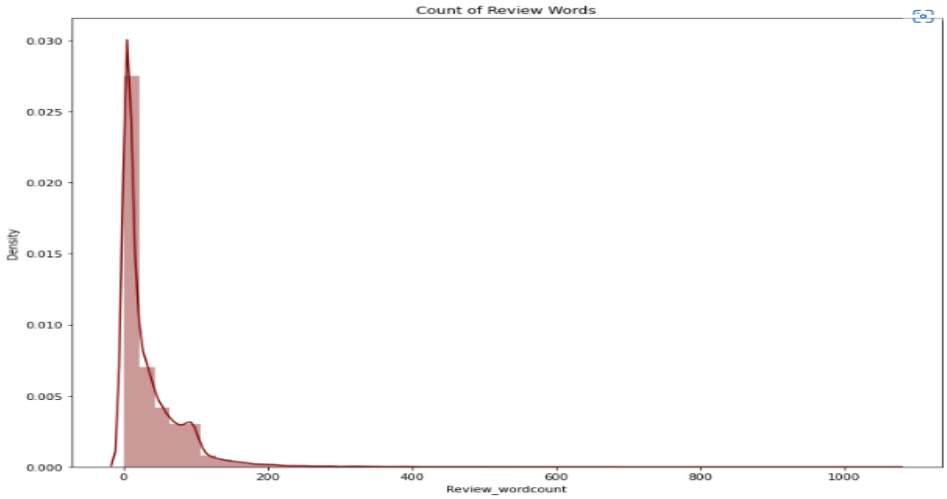
- **Precision** can be seen as a measure of quality; higher precision means that an algorithm returns more relevant results than irrelevant ones.
- **Recall** is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.
- **Accuracy** score is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.
- **F1-score** is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.
- The Hamming loss is the fraction of labels that are incorrectly predicted. Read more in the User Guide. Ground truth (correct) labels. Predicted labels, as returned by a classifier. Sample weights. New in

version 0.18. Return the average Hamming loss between element of y_true and y_pred.

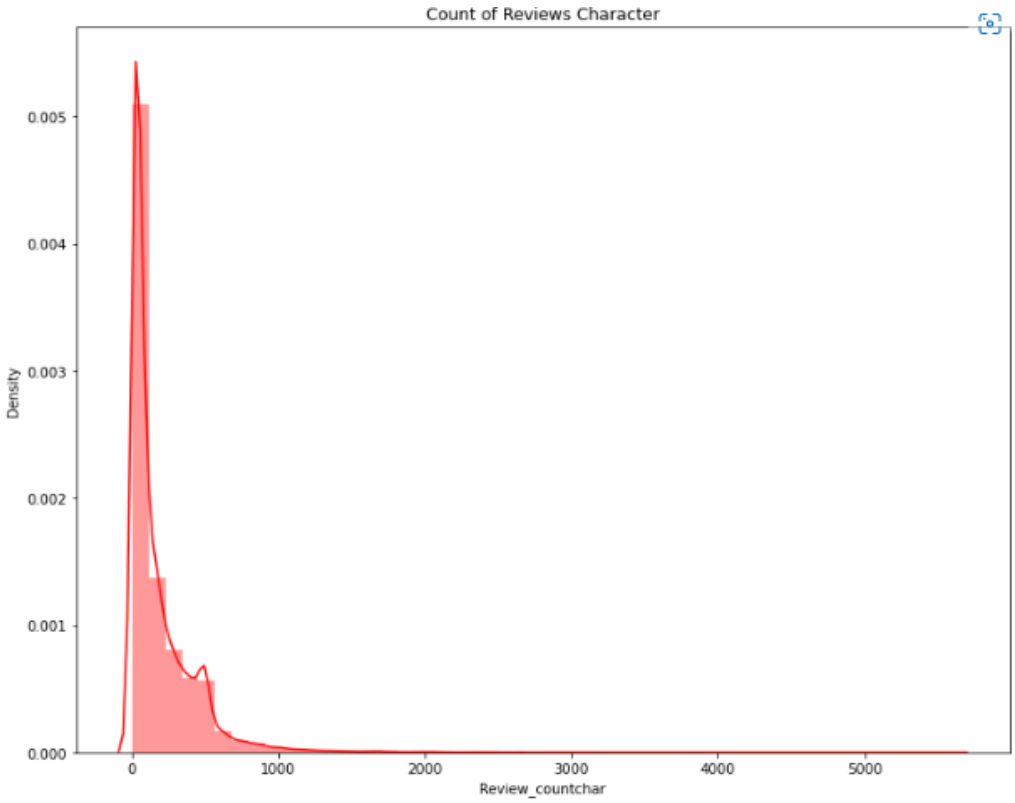
- **Mean squared error** (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).
- **Mean Absolute Error** or MAE. We know that an error basically is the absolute difference between the actual or true values and the values that are predicted.

• Visualizations:

➢ Distribution plo

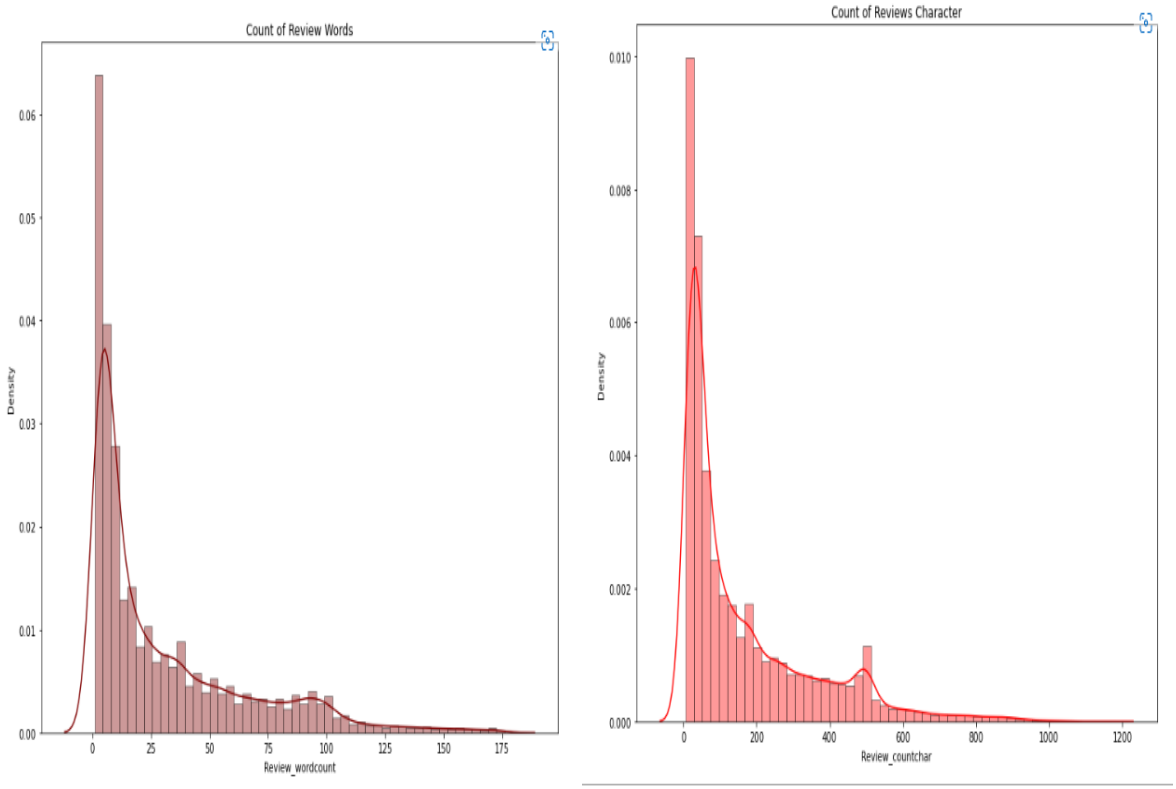


Distribution plot for Character count:



- By observing the histogram, we can clearly see that most of our text is having the number of words in the range of 0 to 200, But some of the reviews are too lengthy which may act like outliers in our data.
- Above plot represents histogram for character count of Review text, which is quite similar to the histogram of word count.
- As we know that some of the reviews are too lengthy, so i have to treat them as outliers and remove them using zscore method. After removing the outliers, the word count and character count looks as below.

Distribution plot after removing the outliers:



CONCLUSION:

▪ Key Findings and Conclusions of the Study

➢ I have collected the data using the web scraping tool selenium from Amazon and Flipkart website. ➢ Loaded the dataset and analyse the data.

➢ Data set contains the null value, replace the null value with mode. Done the data cleaning, removed the column Unnamed: 0 from the dataset.

➢ Using the nltk tool done the text pre-processing, removed the unnecessary word, punctuation, stop words, phone number, url from the reviews.

➢ Checked the count for word and character from review

➢ Using the distribution list, we found that there is a outliers present. Removed the outliers using the zscore.

➢ Used various visualization tool to analyse the data on the scale of rating 1 to 5.

➢ Converted the text data into vector using the TDIDF Vectorizer.

➢ Balanced the data using the over sampling technique SMOTE.

➢ Splitted the data into training and testing.

➢ Used various machine learning model to analyse and check the accuracy score and its metrics. Chosen the best model that performed well and gives a good accuracy score. In our project Random Forest Classifier gives a good accuracy score 86%.

➢ Then done the hyper parameter tuning for our final model. Finally saved the model using the pickle and predicted the ratings value.

➢ The predicted and actual value is almost similar. As our model has performed well.

• Learning Outcomes of the Study in respect of Data Science:

➢ I have scrapped the reviews and ratings of different technical products from flipkart.com and amazon.in websites. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed.

➢ New analytical techniques (NLP) of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps remove unrealistic values, punctuations, URLs, email address and stop words.

➢ This study is an exploratory attempt to use 6 machine learning algorithms in estimating Rating, and then compare their results.

➢ To conclude, the application of NLP in Rating classification is still at an early stage.

➢ We hope this study has moved a small step ahead in providing some methodological and empirical contributions to crediting institutes, and presenting an alternative approach to the valuation of Ratings.

• Limitations of this work and Scope for Future Work:

➢ As we know the content of text in reviews is totally depends on the reviewer and they may rate differently which is totally depends on that particular person.

➢ So, it is difficult to predict ratings based on the reviews with higher accuracies. Still, we can improve our accuracy by fetching more data and by doing extensive hyperparameter tuning.

➢ While we couldn't reach out goal of maximum accuracy in Ratings prediction project, we did end up creating a system that can with some improvement and deep learning algorithms get very close to that goal.

➢ As with any project there is room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result.

➢ This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others.

➢ Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.