Project report on

Ratings Prediction Project

Submitted By

ANIRUDH BHARDWAJ

# ACKNOWLEDGMENT

I express my sincere gratitude to Flip Robo Technologies for giving me the opportunity to work on this project on Ratings Prediction using machine learning algorithms.

It is my sensual gratification to present this report on “RATINGS PREDICTION” project which is a

NLP project. Working on this project was a good experience it helped me to enhance my knowledge and skills in machine learning.

**Business Problem Framing**

# INTRODUCTION

The rise in E-commerce has brought a significant rise in the importance of customer reviews. There are hundreds of review sites online and massive amounts of reviews for every product. Customers have changed their way of shopping and according to a recent [survey](https://www.reviewtrackers.com/online-reviews-survey/), 70 percent of customers say that they use rating filters to filter out low rated items in their searches.

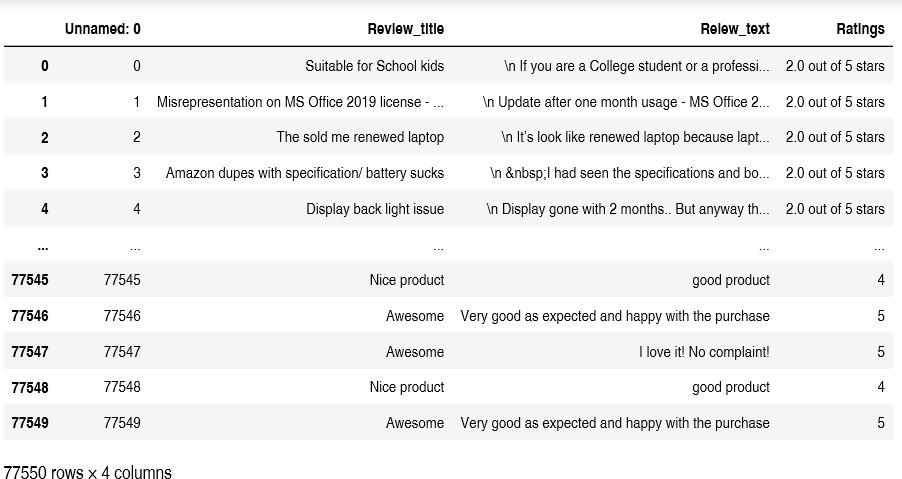
The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, companies like Google, Amazon and Yelp!.

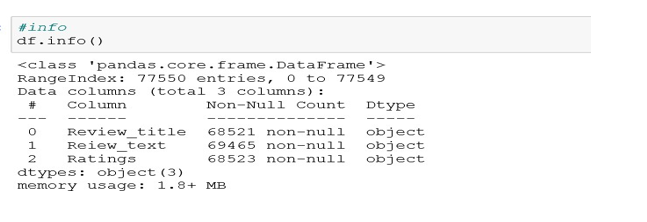
There are two main methods to approach this problem. The first one is based on review text content analysis and uses the principles of natural language process (the NLP method). This method lacks the insights that can be drawn from the relationship between costumers and items. The second one is based on recommender systems, specifically on collaborative filtering, and focuses on the reviewer’s point of view.

## Analytical Problem Framing

For rating prediction first I have scraped reviews and ratings from well-known e-commerce sites. This is then saved into .csv format.

Then I imported the Rate prediction file in jupyter Notebook. And performed some need full steps before processing the Data.





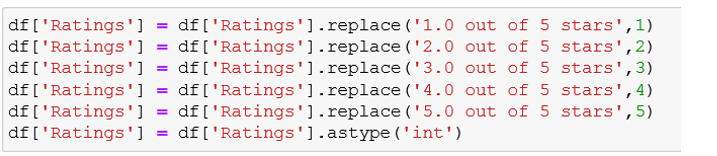
Looking at above both figures we can see that our data set contains 77550 different rows and 4 columns among which I have removed unwanted column(Unnamed:0). And for this project Ratings is our target column. There are some missing values in our dataset which have been removed from the dataset.

## Data Processing:

To start with, I have joined both columns Review\_title and Review\_text into a new column as Review.



Then all the entries from Ratings columns have been converted to respective integer values

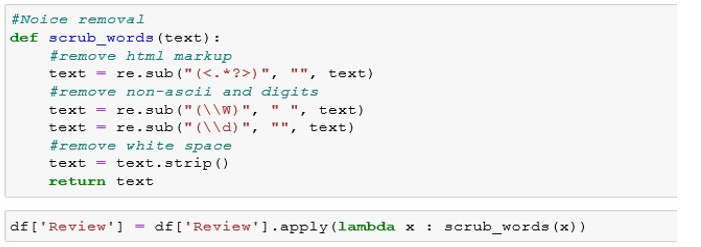


### Text processing

****

For text processing I have defined a function to replace some words with proper words. All text has been converted to lowercase and individual punctuation marks have been removed from the text in the review column.

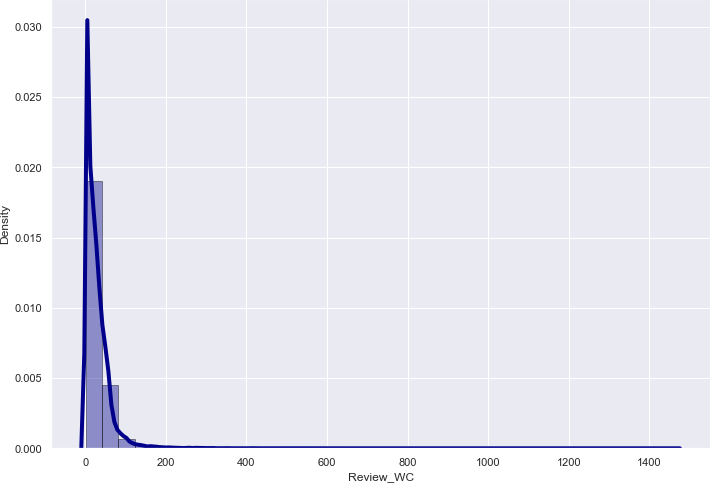
### Text Normalization – Standardization

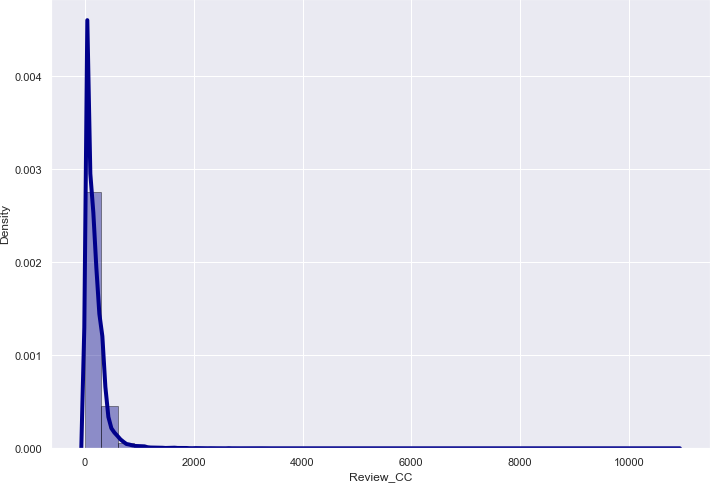
****

Finally to standardize our test and to remove numbers from it I have defined a function as scrub\_words as you can see above and also applied it to the review column.

## Exploratory Data Analysis:

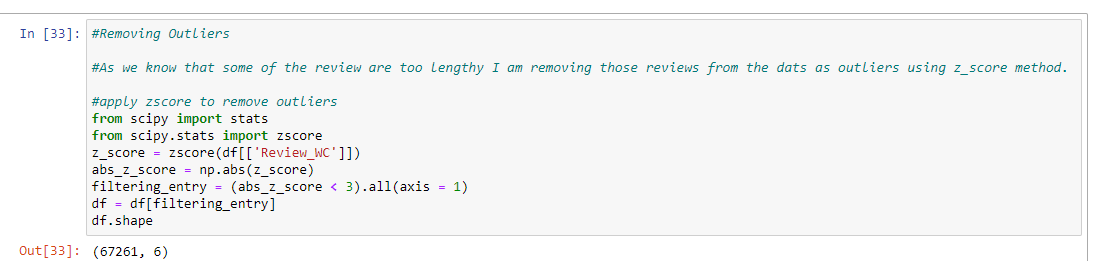
#### Word\_count of review





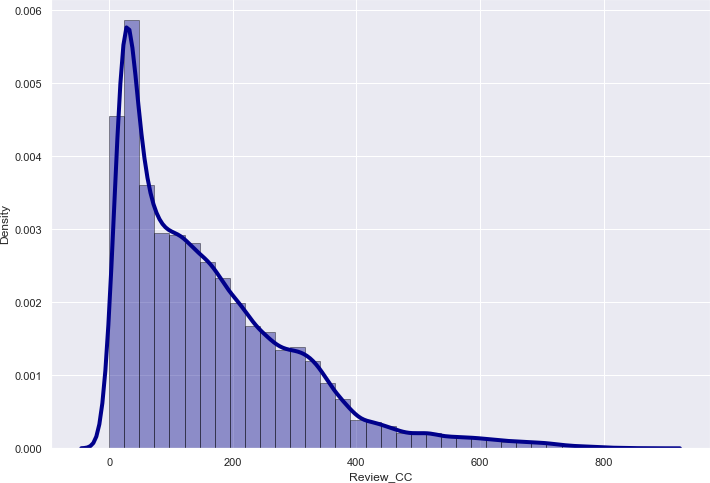
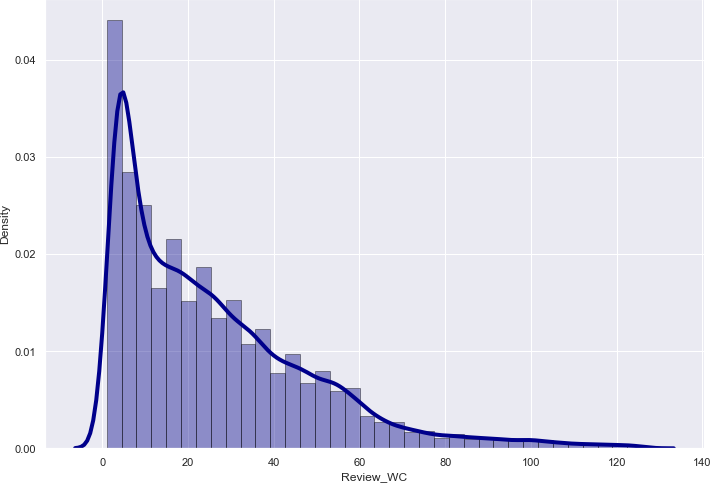
After looking at the graphical presentation of the Data. I have decided to remove the outlier from the Data.

**Removing Outliers**



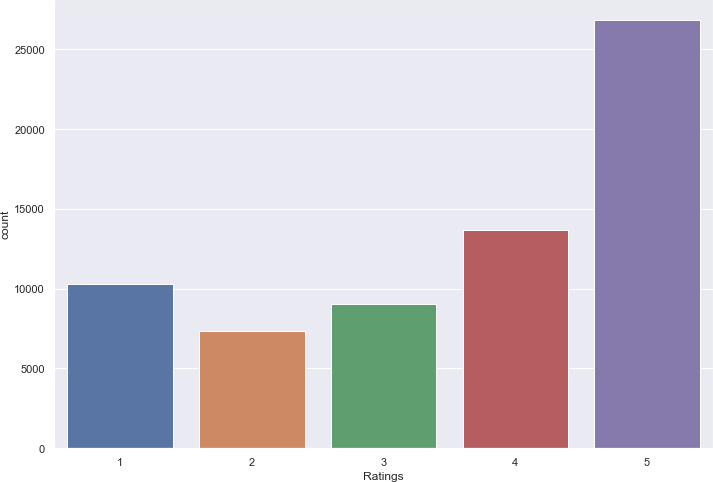
By removing outliers, we are not losing much of the data. So, it seems to be good to remove the outliers.

**Plotting histograms for word count and character counts again after removing outliers**



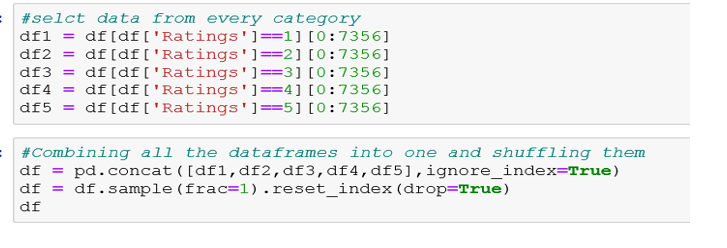
After removing outliers, we can now see that we are getting decent range of number of words and characters.

#### Ratings (Target Variable):

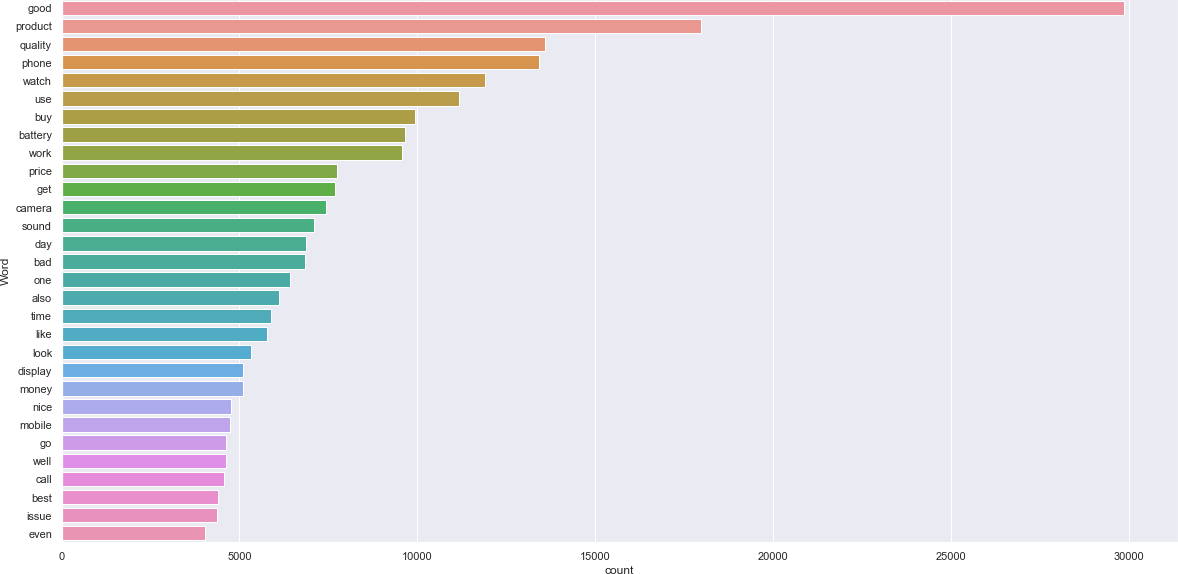


The above figure is representing a count-plot for our target variable which is "rating". Looking at the plot we can say that the number of reviews is as high as 5 stars as compared to others. And the reviews which are rated 2 stars are very few in number. This will cause imbalance problem for our model.

After considering the findings, I have decided to select equal number of reviews from Every class. As observed for 2 Stars rating there are only 7356 reviews. Which is the least value among all. So to keep the model balances I am taking only 7356 reviews for all the ratings.



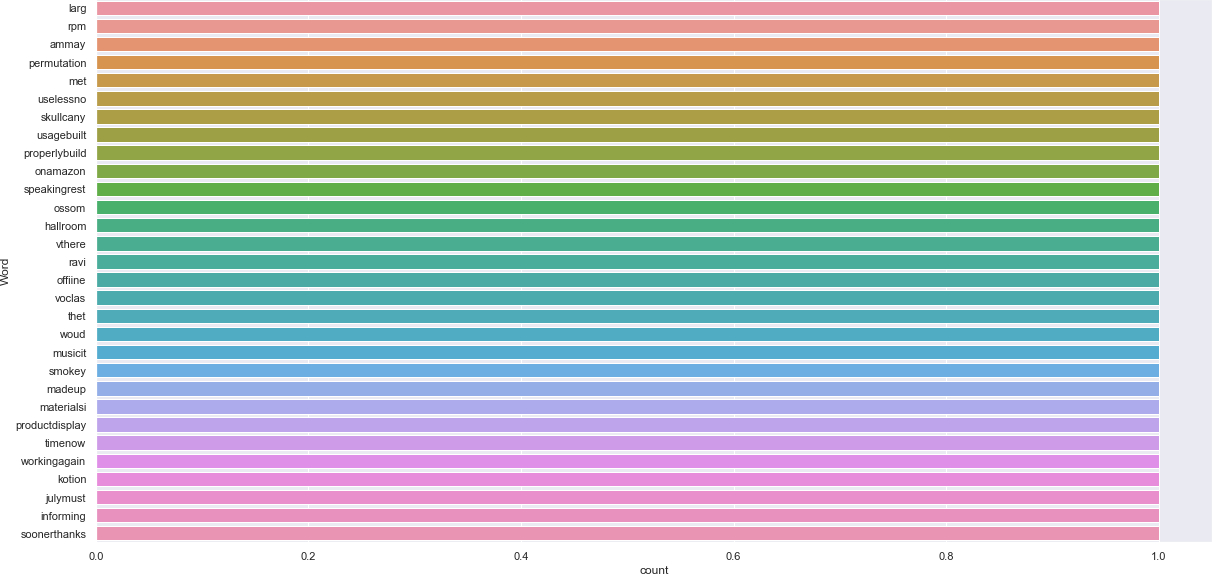
#### Top 30 most frequently occurring words:



The above bar plot is showing top 30 most frequently occurring words in our reviews. We

can see the words like ‘good’, ‘product’, ‘quality’ etc. are occurring more frequently.

#### Top 30 Rarely occurring words:



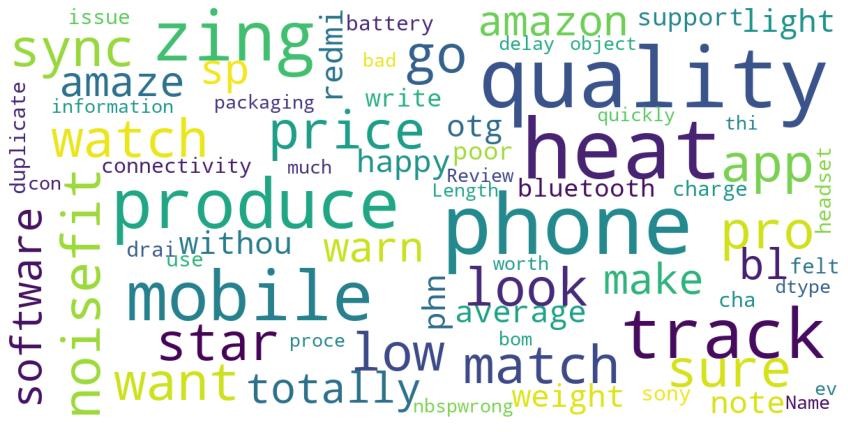
Above figure is representing bar plot for top 30 rarely occurring words. Many of which are spelled incorrectly that’s why these are occurring only once.

Now with the help of word cloud we will try to identify the various word which were used by the customers for individual ratings.

#### Words for rating = 1:



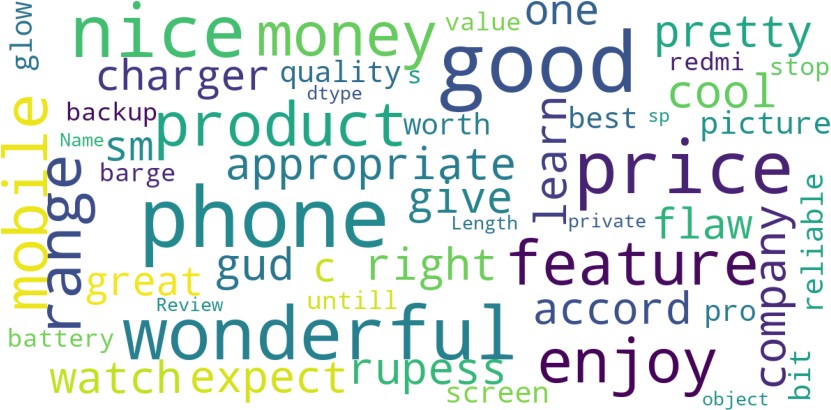
**Words for rating = 2:**



#### Words for rating = 3:



**Words for rating = 4:**



**Words for rating = 5:**



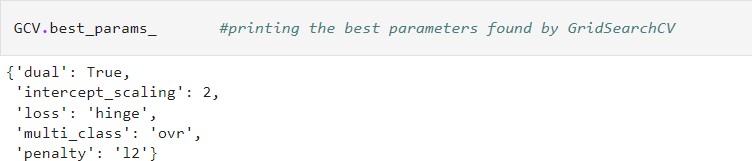
## Model Development and Evaluation

As this one is the rating based project So have accumulated the Data from various renewed online shopping sites. Below the are algorithm which I have used to get the best result.

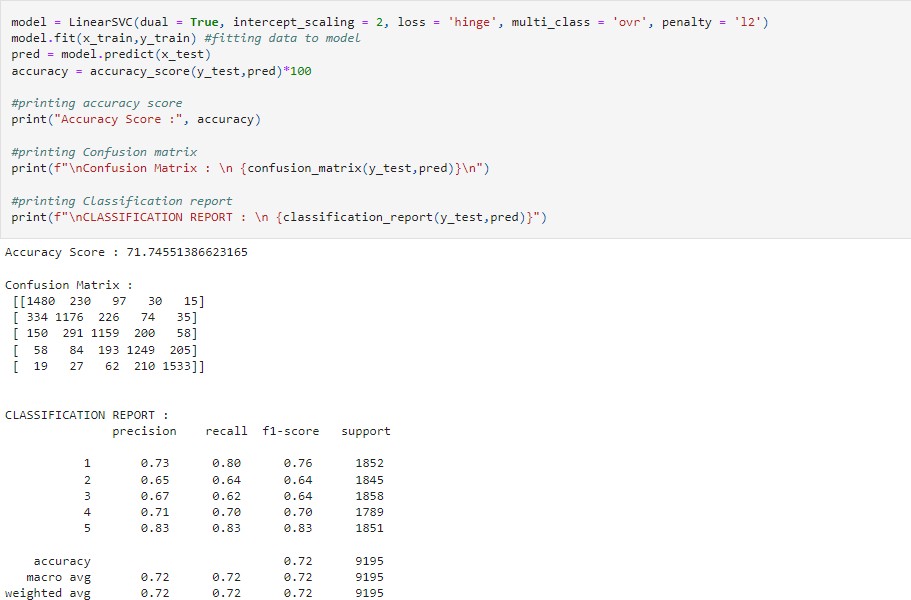
* RandomForestClassifier
* LinearSVC
* LogisticRegression
* MultinomialNB
* XGBClassifier
* BernoulliNB
* LightGBMClassifier
* SGDClassifier
* From all of these above models LinearSVC was giving me good performance.

### Hyperparameter Tuning

I did hyperparameter tuning for Linear SVC for various parameters. After tunning, below are the parameters which I found suitable for our final Model



## Final Model:



After performing hyperparameter tunning we are getteing good results.

**Conclusion:**

### Key findings of the study

In this project, I have collected reviews Data and Ratings from amazon.in and flipkart.com. I then did separate text processing for the review column and selected an equal number of texts from each rating class to keep the model, Balance. I have analyzed the text by doing various EDA steps. We also checked the frequently occurring words as well as the less frequently occurring words in our data. After all these steps I have created function to train and test different algorithms and using different evaluation metrics I have selected Linear SVC for my final model.

We got good accuracy after performing the hyperparameter tunning.

### Limitations of this work and scope for the future work

As we know, the text content in reviews is completely up to the reviewer and they can rate

differently, which is totally up to that particular individual. Therefore, it is difficult to predict

ratings based on reviews with great accuracy. However, we can improve our accuracy by acquiring

more data and extensive hyperparameter tuning.