NAME OF THE PROJECT

Product Rating Prediction Using NLP

SUBMITTed by:

Swati Pandey

FLIPROBO SME:

Swati Mahaseth

ACKNOWLEDGMENT

I would like to express my special gratitude to "Flip Robo" team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Swati Mahaseth (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to "Data trained" who are the reason behind my Internship at FlipRobo. Last but not least my family who have been my backbone in every step of my life.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from Towards Data Science, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)
4. Krish Naik YouTube videos on NLP.

Chap 1. Introduction

1.1 Business Problem Framing

Internet is the best source nowadays for any organisation to know public opinions about their products and services. Many consumers form an opinion about a product just by reading a few reviews. Online product reviews provided by consumers who previously purchased products have become a major information source for consumers and marketers regarding product quality. Research has shown that consumer online product ratings reflect both the customers' experience with the product and the influence of others' ratings. Websites prominently display consumers' product ratings, which influence consumers' buying decisions and willingness to pay.

The opinion information is very useful for users and customers alike, many of whom typically read product or service reviews before buying them. Businesses can also use the opinion information to design better strategies for production and marketing. Hence, in recent years, sentiment analysis and opinion mining have become a popular topic for machine learning and data mining.

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.*

1.2 Conceptual Background of the Domain Problem

A recent survey (Hinckley, 2015) revealed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions. More precisely, 54.7% recognized that these *reviews were either fairly, very or absolutely important in their purchase decision making.* Relying on online reviews has thus become a second nature for consumers.

Consumers want to find useful information as quickly as possible. However, *searching and comparing text reviews can be frustrating for users as they feel submerged with information*. Indeed, the massive amount of text reviews as well as its unstructured text format prevent the user from choosing a product with ease. *The star-rating, i.e., stars from 1 to 5 on online platform, rather than its text content gives a quick overview of the product quality.* This numerical information is the number one factor used in an early phase by consumers to compare products before making their purchase decision.

Generally, the ratings and the price of the product are simple heuristics used by the customers to decide over the final purchase of the product. But often, the overall star ratings of the product reviews may not capture the exact polarity of the sentiments. *This makes rating prediction a hard problem as customers may assign different ratings for a particular review. For example,*

<u>For instance, a user may rate a product as good and assign a 5-star score while another user may write the same comment and give only 3 stars.</u> In addition, reviews may contain anecdotal information, which do not provide any helpful information and complicates the predictive task.

The question that arises is how to successfully predict a user's numerical rating from its review text content. One solution is to rely on <u>supervised machine learning techniques such as text classification</u> which allows to automatically classify a document into a fixed set of classes after being trained over past annotated data.

Different users may have different sentiment expression preferences. For example, some users choose to use "good" to describe a just-so-so product, but others may use "good" to describe an excellent product. Beside the user bias, there is also a product bias. We may use different opinion words to review different products, or even use the same opinion word to express different sentiment polarities for different products; for example, the opinion word "long" can be express a "positive" feeling for cell phone's battery life, but may have a "negative" feeling for a camera's focus time. Therefore, it is important to consider the relationship between the review-authors, as well as that of the target products, for review rating prediction.

1.3 Review of Literature

According to the Lackermair, Kailer and Kanmaz (2013), product reviews and ratings represent an important source of information for consumers and are helpful tools in order to support their buying decisions . *They also found out that consumers are willing to compare both positive and negative reviews when searching for a specific product.* The authors argue that customers need compact and concise information about the products. Therefore, *consumers first need to pre-select the potential products matching their requirements. With this aim in mind, consumers use the star ratings as an indicator for selecting products. Later, when a limited number of potentials products have been chosen, reading the associated text review will reveal more details about the products and therefore help consumers making a final decision.*

It becomes daunting and time-consuming to compare different products in order to eventually make a choice between them. Therefore, models able to predict the user rating from the text review are critically important (Baccianella, Esuli & Sebastiani, 2009) .

Chevalier and Mayzlin (2006) [8] also analyse the distribution of ratings in online reviews and come to the same conclusion: the resulting data presents an asymmetric bimodal distribution, where reviews are overwhelmingly positive.

Pang, Lee and Vaithyanathan (2002) [9] approach this predictive task as an opinion mining problem enabling to automatically distinguish between positive and negative reviews. In order to determine the reviews polarity, the authors use text classification techniques by training and testing binary classifiers on movie reviews containing 36.6% of negative reviews and 63.4% of positive reviews. On the top of that, they also try to identify appropriate features to enhance the performance of the classifiers.

Dave, Lawrence, and Pennock (2003) [10] also deal with the issue of class imbalance with a majority of positive reviews and show similar results. *SVM outperforms Naïve Bayes with an accuracy greater than 85% and the implementation of part-of-speech as well as stemming is also ineffective*. Nevertheless, while the previous research led to better results with unigrams, this study shows that bigrams turn out to be more effective at capturing context than unigrams in the specific case of their datasets.

In a slightly different perspective, McAuley and Leskovec (2013) intend to understand hidden dimensions of customer's opinions to better predict the numerical ratings. *They also find out that the user information is a rich source of information to take into account when predicting ratings with review text.*

On the other hand, presented a model which predicts the star rating of a review using sentiment dictionaries. Like most polarity determining approaches, the paper uses the unigram model to represent text. *The unigram model often fails to capture phrase patterns properly which leads to polarity incoherence.* To overcome this drawback, the authors also employ a n-gram model. But this way of representing text vectors leads to the creation of large sparse matrices that are space inefficient known as n-gram sparsity bottlenecks.

To overcome the problem of polarity incoherence, introduces the Bag of opinions model. *This model overcomes the limitations of the unigram and n-gram models*. It has 3 components: root word, modifiers and negation words. Each opinion from the corpora of reviews is assigned a score using ridge regression method. At the end, a final score is calculated by combining all the independent scores of all the opinions.

Memory-based and model-based (e.g., matrix factorization) are the two most popular methods for collaborative filtering. However, both systems take a sparse user-product rating matrix as an input and may fail to capture the quality of the users and their reviews.

In a similar project 'Using Properties of the Amazon Graph to Better Understand Reviews' proposed by Leon, Vasant, Sheldon (2011), the predicted feedback score was represented with a linear combination of the following features: Original Rating, Review Helpfulness Count, Review Unhelpfulness Count, Review Helpfulness Ratio, Reviewer Helpfulness Ratio, Reviewer Bias, Reviewer Expertise, Reviewer Clustering, Reviewer Betweenness, and Reviewer Degree. *However, the goal of their project is to predict the rating of a customer who reviews the same product more than once, while our project is to predict the user's first-time rating score.* Also, they use stochastic gradient descent to minimize the mean square error to capture the weights of such features.

1.4 Motivation for the Problem Undertaken

The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary motivation.

Data needed for this project is require to scrap from E-commerce platform and data cleaning operation over it. Features derived from textual reviews are used to predict its corresponding star ratings. To accomplish it, the prediction problem is transformed into a multi-class classification task to classify reviews to one of the five classes corresponding to its star rating. Getting an overall sense of a textual review could in turn improve consumer experience. However, the motivation for taking this project was that it is relatively a new field of research.

Chap 2 Analytical Problem Framing

1. Mathematical / Analytical Modelling of the Problem

In order to apply text classification, the unstructured format of text has to be converted into a structured format for the simple reason that it is much easier for computer to deal with numbers than text. This is mainly achieved by projecting the textual contents into Vector Space Model, where text data is converted into vectors of numbers.

In the field of text classification, documents are commonly treated like a Bag-of-Words (BoW), meaning that each word is independent from the others that are present in the document. They are examined without regard to grammar neither to the word order. In such a model, the term-frequency (occurrence of each word) is used as a feature in order to train the classifier. However, using the term frequency implies that all terms are considered equally important. As its name suggests, the term frequency simply weights each term based on their occurrence frequency and does not take the discriminatory power of terms into account. To address this problem and penalize words that are too frequent, each word is given a term frequency inverse document frequency (tf-idf) score which is defined as follow:

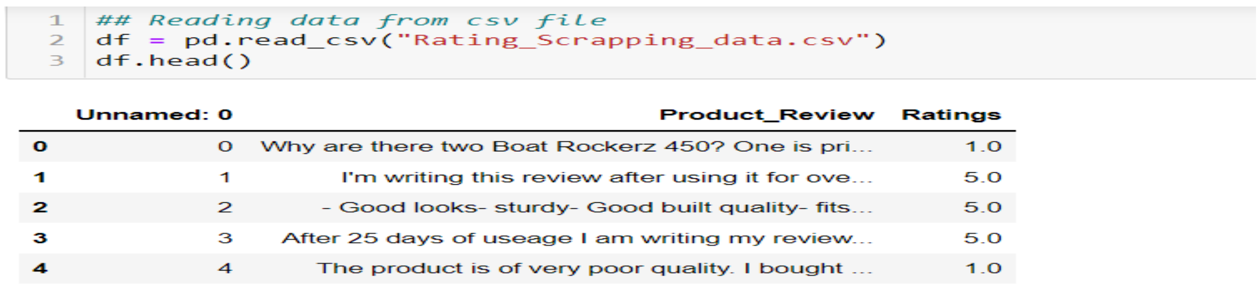
$$tf - \mathrm{id}f_{t,d} = \mathrm{t}f_{\mathrm{t,d}} * \mathrm{id}f_{\mathrm{t}}$$

where:

- $\mathrm{t}f_{\mathrm{t,d}} = \frac{\mathrm{n}_{\mathrm{t,d}}}{\sum_k n_{k,d}}$ with $\mathrm{n}_{\mathrm{t,d}}$ the number of term t contained in a document d, and $\sum_k n_{k,d}$ the total number of terms k in the document d
- $\mathrm{id}f_{\mathrm{t}} = \log \frac{N}{df_t}$ with N the total number of documents and df_t the number of documents containing the term t

2. Data Sources and their formats

Data is collected from Amazon.in and flipkart.com using selenium and saved in CSV file. Around 62000 Reviews are collected for this project.



```
## Reading data from csv file
df = pd.read_csv("Rating_Scrapping_data.csv")
df.head()
```

	Unnamed: 0	Product_Review	Ratings
0	0	Why are there two Boat Rockerz 450? One is pri...	1.0
1	1	I'm writing this review after using it for ove...	5.0
2	2	- Good looks- sturdy- Good built quality- fits...	5.0
3	3	After 25 days of useage I am writing my review...	5.0
4	4	The product is of very poor quality. I bought ...	1.0

This is multi-classification problem and Rating is our target feature class to be predicated in this project. There are five different categories in feature target i.e., The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars.

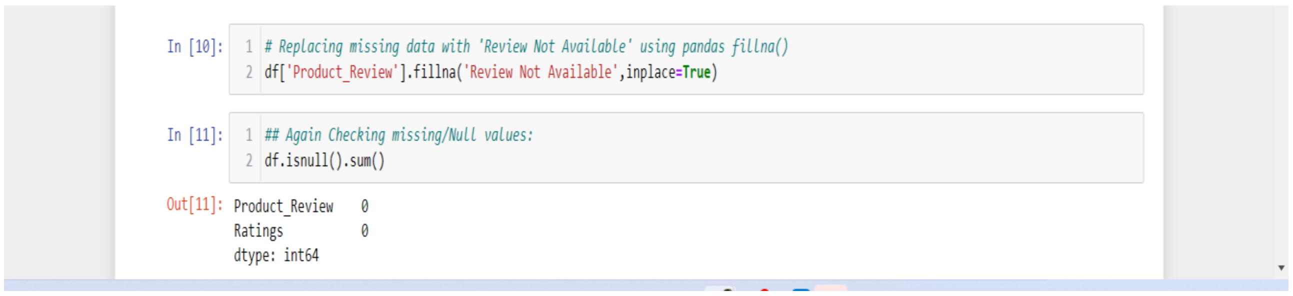
There are some missing values in product review. The datatype of Product review is object while datatypes of Ratings is float.

3. Data Pre-processing

The dataset is large and it may contain some data error. In order to reach clean, error free data some data cleaning & data pre-processing performed data.

- Missing Value Imputation:

Missing value in product reviews are replace with 'Review Not Available'.



```
In [10]: # Replacing missing data with 'Review Not Available' using pandas fillna()
         df['Product_Review'].fillna('Review Not Available',inplace=True)

In [11]: ## Again Checking missing/Null values:
         df.isnull().sum()

Out[11]: Product_Review    0
         Ratings           0
         dtype: int64
```

- Data is pre-processed using the following techniques:

1. Convert the text to lowercase
2. Remove the punctuations, digits and special characters
3. Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
4. Remove the stop words
5. Stemming and Lemmatizing
6. Applying Text Vectorization to convert text into numeric

4. 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. Using word cloud, we can see most occurring word for different categories.

5. Hardware & Software Requirements with Tool Used

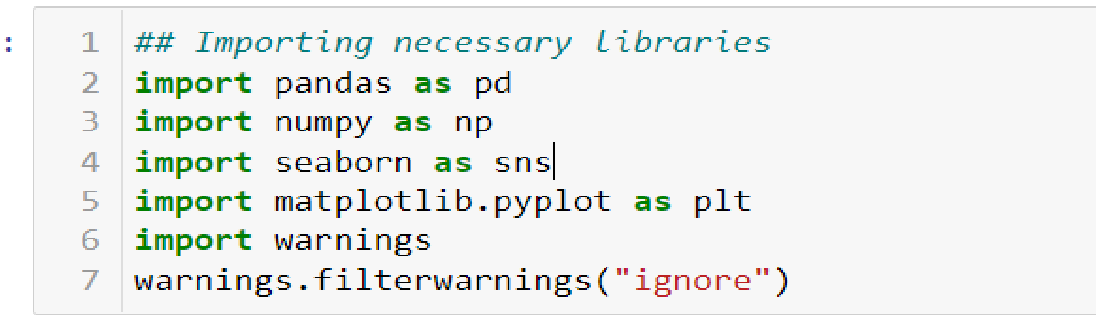
Hardware Used -

1. Processor — Intel i5 processor with 2.4GHZ
2. RAM — 16 GB

Software used-

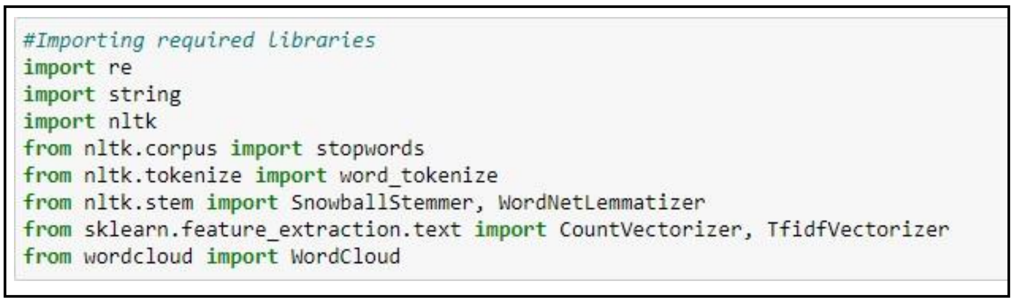
1. Anaconda – Jupyter Notebook
2. Selenium – Web scraping

Libraries Used – General library for data wrangling & visualization



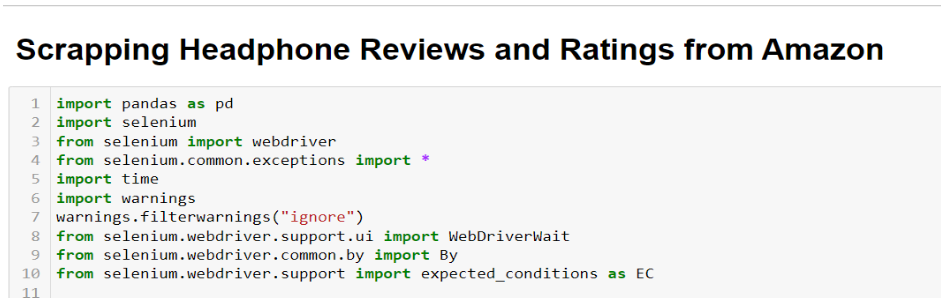
```
## Importing necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

Libraries used for Text Mining / Text Analytics are:



```python
#Importing required libraries
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer, WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from wordcloud import WordCloud
```

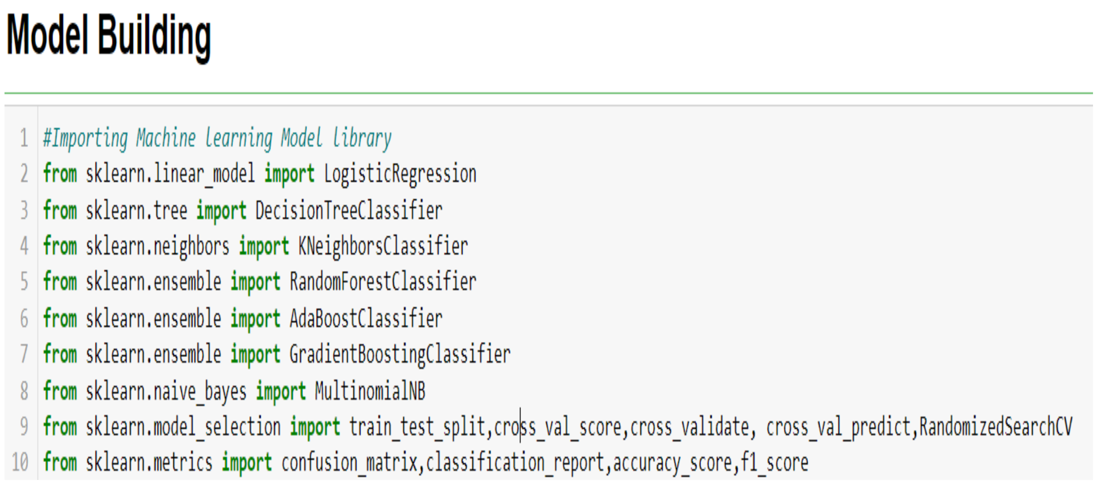
Libraries used for web scraping data from e-commerce website are



Scrapping Headphone Reviews and Ratings from Amazon

```python
import pandas as pd
import selenium
from selenium import webdriver
from selenium.common.exceptions import *
import time
import warnings
warnings.filterwarnings("ignore")
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as EC
```

Libraries used for machine learning model building



Model Building

```python
#Importing Machine learning Model library
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split,cross_val_score,cross_validate, cross_val_predict,RandomizedSearchCV
from sklearn.metrics import confusion_matrix,classification_report,accuracy_score,f1_score
```

Chap. 3 Models Development & Evaluation

1. Identification Of Possible Problem-Solving Approaches (Methods)

First part of problem solving is to scrap data from amazon.in and flipkart.com website which we already done. Second is performing text mining operation to convert textual review in ML algorithm useable form. Third part of problem building machine learning model to predict rating on review. This problem can be solve using classification-based machine learning algorithm like logistics regression. Further Hyperparameter tuning performed to build more accurate model out of best model.

2. Testing of Identified Approaches (Algorithms)

The different classification algorithm used in this project to build ML model are as below:

- ❖ Random Forest classifier
- ❖ Decision Tree classifier
- ❖ Logistics Regression
- ❖ AdaBoost Classifier
- ❖ Gradient Boosting Classifier
- ❖ MultinomialNB Classifier
- ❖ KNeighbors Classifier

3. Key Metrics for Success in Solving Problem Under Consideration

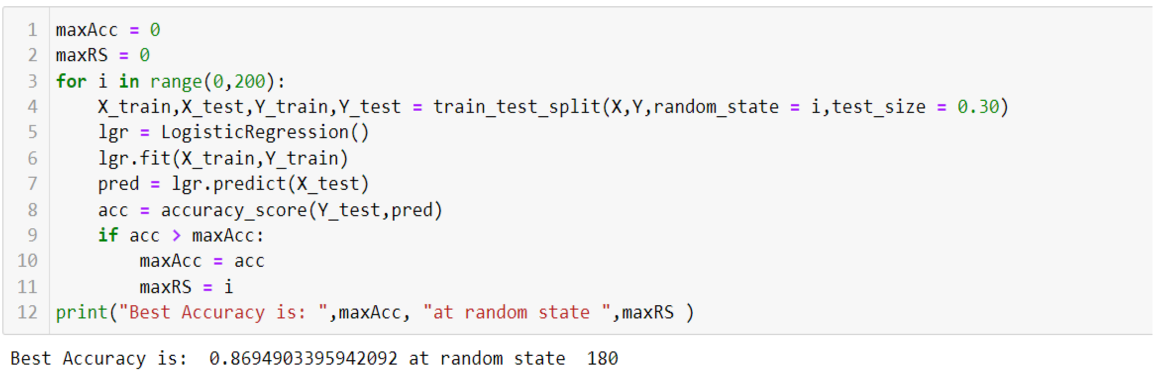
- ▪ Precision can be seen as a measure of quality; higher precision means that an algorithm returns more relevant results than irrelevant ones.
- ▪ Recall is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.
- ▪ Accuracy score is used when the 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.

- Cross validation Score: To run cross-validation on multiple metrics and also to return train scores, fit times and score times. Get predictions from each split of cross-validation for diagnostic purposes. Make a scorer from a performance metric or loss function.
- AUC_ROC_score: ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0
- We have used Accuracy Score and Cross validation score as key parameter for model evaluation in this project since balancing of data is perform.

3. Run And Evaluate Selected Models

1. Logistics Regression

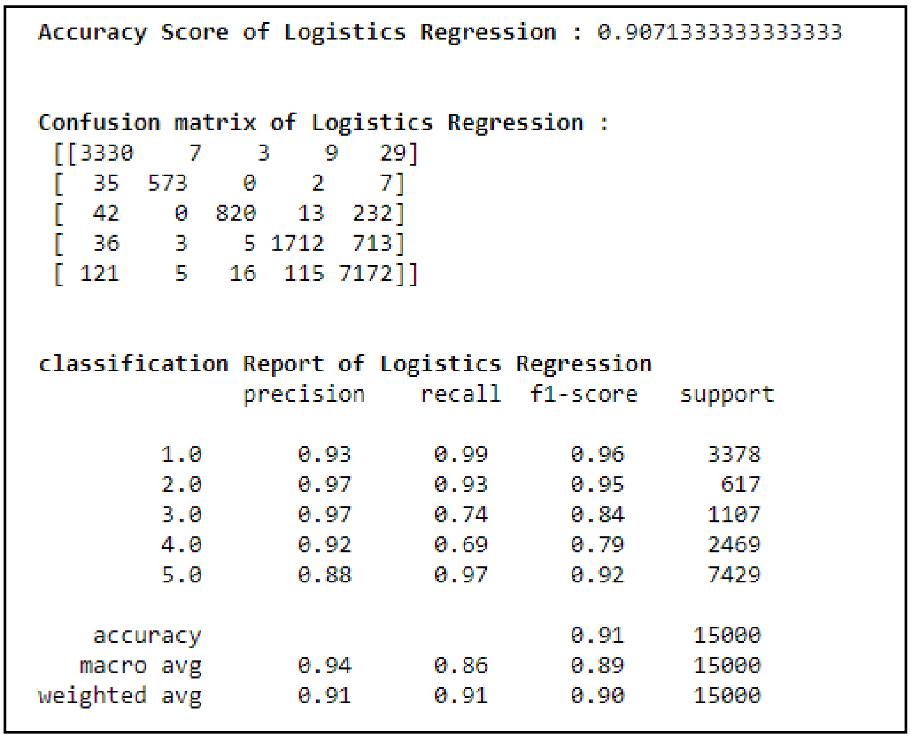
Train-test split is used to split data into training data & testing data. Further best random state is investigated through loop.



```
maxAcc = 0
maxRS = 0
for i in range(0,200):
    X_train,X_test,Y_train,Y_test = train_test_split(X,Y,random_state = i,test_size = 0.30)
    lgr = LogisticRegression()
    lgr.fit(X_train,Y_train)
    pred = lgr.predict(X_test)
    acc = accuracy_score(Y_test,pred)
    if acc > maxAcc:
        maxAcc = acc
        maxRS = i
print("Best Accuracy is: ",maxAcc, "at random state ",maxRS )
```

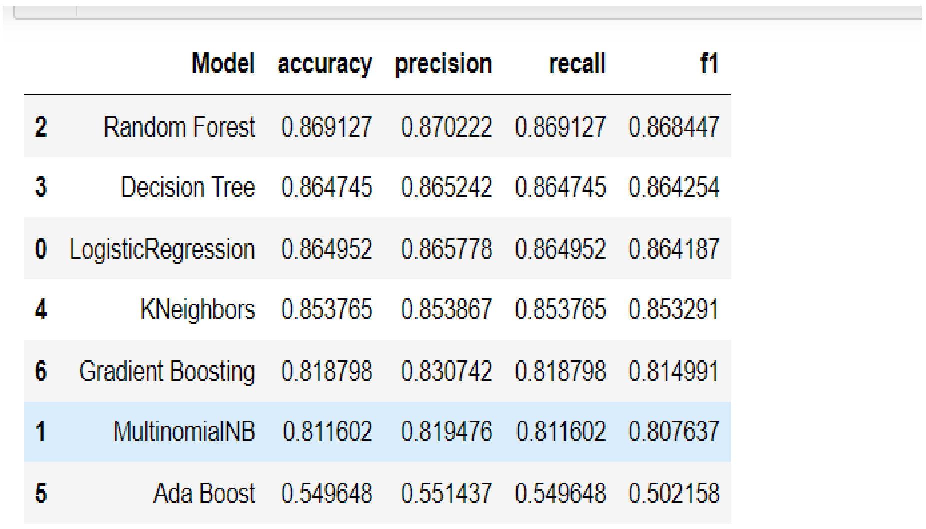
```
Best Accuracy is:  0.8694903395942092 at random state  180
```

Logistics regression evaluation matrix is shown below:



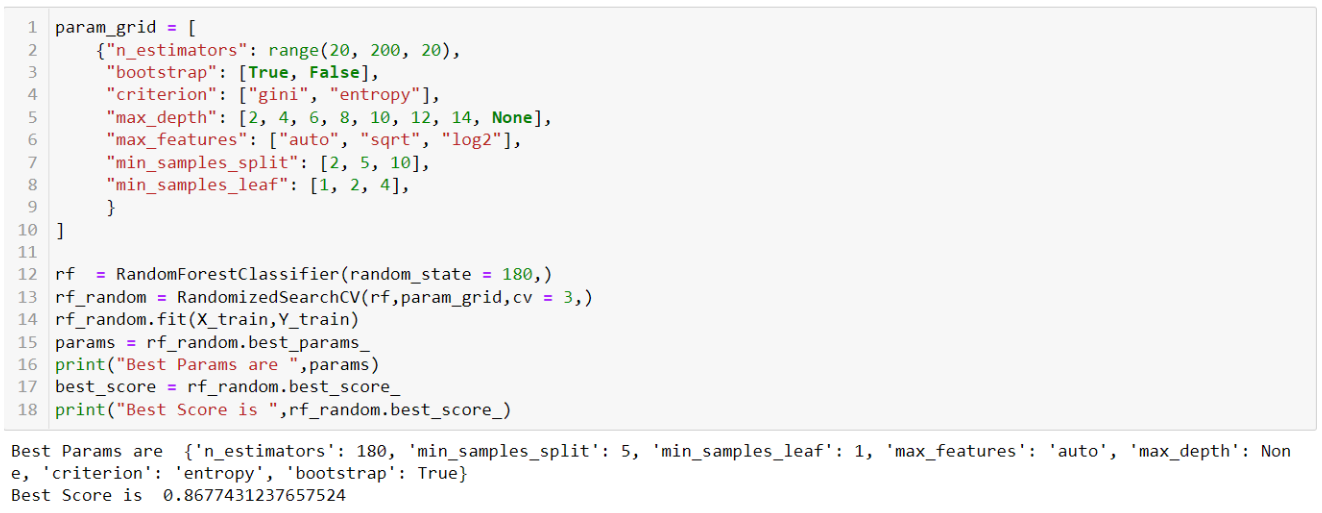
```
Accuracy Score of Logistics Regression : 0.9071333333333333


Confusion matrix of Logistics Regression :
 [[3330    7    3    9   29]
 [  35  573    0    2    7]
 [  42    0  820   13  232]
 [  36    3    5 1712  713]
 [ 121    5   16  115 7172]]


classification Report of Logistics Regression
               precision    recall  f1-score   support

         1.0       0.93      0.99      0.96      3378
         2.0       0.97      0.93      0.95       617
         3.0       0.97      0.74      0.84      1107
         4.0       0.92      0.69      0.79      2469
         5.0       0.88      0.97      0.92      7429

    accuracy                           0.91     15000
   macro avg       0.94      0.86      0.89     15000
weighted avg       0.91      0.91      0.90     15000
```

Evaluation Metrics for different algorithms sorted by F1 score:



	Model	accuracy	precision	recall	f1
2	Random Forest	0.869127	0.870222	0.869127	0.868447
3	Decision Tree	0.864745	0.865242	0.864745	0.864254
0	LogisticRegression	0.864952	0.865778	0.864952	0.864187
4	KNeighbors	0.853765	0.853867	0.853765	0.853291
6	Gradient Boosting	0.818798	0.830742	0.818798	0.814991
1	MultinomialNB	0.811602	0.819476	0.811602	0.807637
5	Ada Boost	0.549648	0.551437	0.549648	0.502158

3-fold Cross validation performed over all model. We can see that Random Forest Classifier gives us good Accuracy and maximum f1 score along with best Cross-validation score. Hyperparameter tuning is applied over Random Forest model and used it as final model.



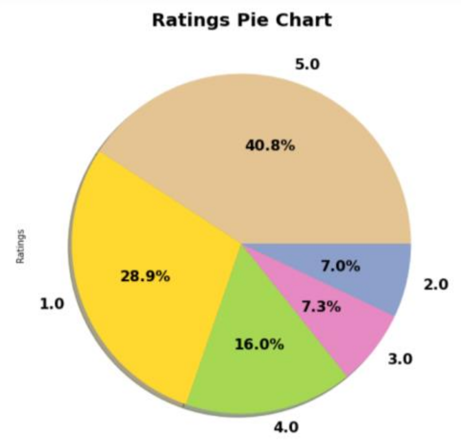
```
param_grid = [
    {"n_estimators": range(20, 200, 20),
     "bootstrap": [True, False],
     "criterion": ["gini", "entropy"],
     "max_depth": [2, 4, 6, 8, 10, 12, 14, None],
     "max_features": ["auto", "sqrt", "log2"],
     "min_samples_split": [2, 5, 10],
     "min_samples_leaf": [1, 2, 4],
     }
]

rf  = RandomForestClassifier(random_state = 180,)
rf_random = RandomizedSearchCV(rf,param_grid,cv = 3,)
rf_random.fit(X_train,Y_train)
params = rf_random.best_params_
print("Best Params are ",params)
best_score = rf_random.best_score_
print("Best Score is ",rf_random.best_score_)
```

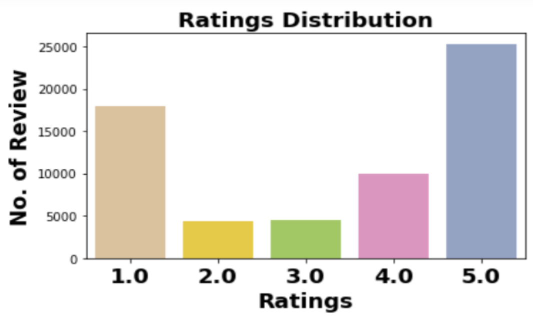
```
Best Params are  {'n_estimators': 180, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': Non
e, 'criterion': 'entropy', 'bootstrap': True}
Best Score is  0.8677431237657524
```

Final model is built using best parameter in hyper parameters tuning. The corresponding evaluation matrix shown below:

```
Final Random Forest Classifier Model
Accuracy Score : 0.8678219686776815


Confusion matrix of Random Forest Classifier : [[4777   85   32   97  383]
 [  74 1052   25   31  123]
 [ 140   18 1081   11   93]
 [ 219   12   30 2333  397]
 [ 453   45   63  125 6882]]


Classification Report of Random Forest Classifier:              precision    recall  f1-score   support

         1.0       0.84      0.89      0.87      5374
         2.0       0.87      0.81      0.84      1305
         3.0       0.88      0.80      0.84      1343
         4.0       0.90      0.78      0.84      2991
         5.0       0.87      0.91      0.89      7568

    accuracy                           0.87     18581
   macro avg       0.87      0.84      0.85     18581
weighted avg       0.87      0.87      0.87     18581
```

5. Visualizations





```
Value Counts of Rating:

5.0    25286
1.0    17895
4.0     9906
3.0     4499
2.0     4350
Name: Ratings, dtype: int64
```

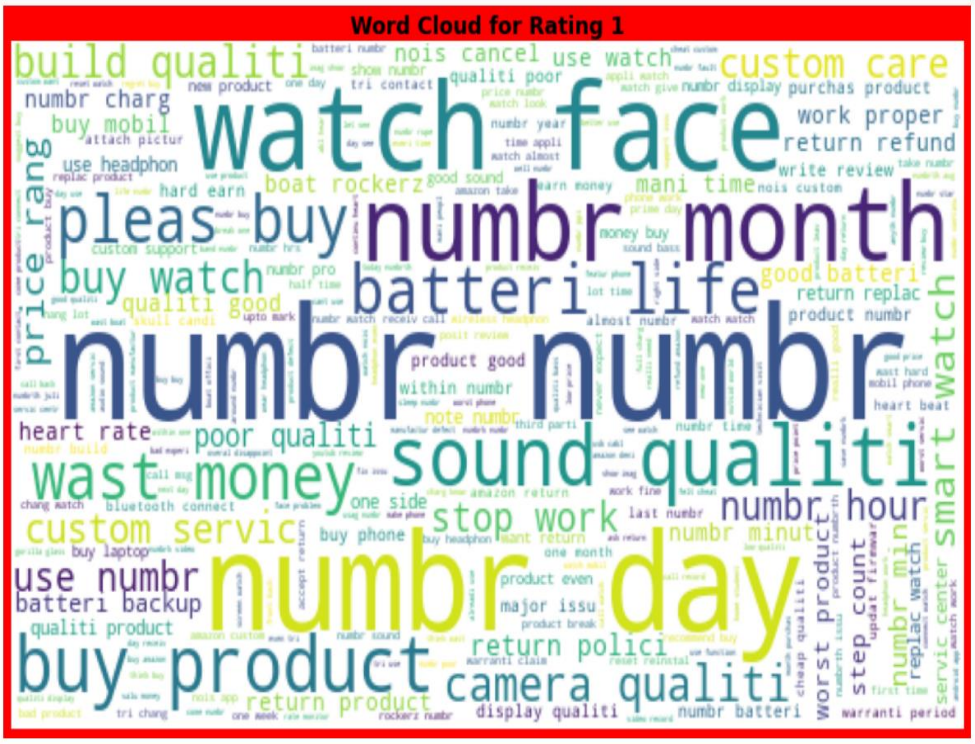
Observations: -> Around 41 % customers have given 5 Rating,followed by around 29% ,which have given 1 rating.

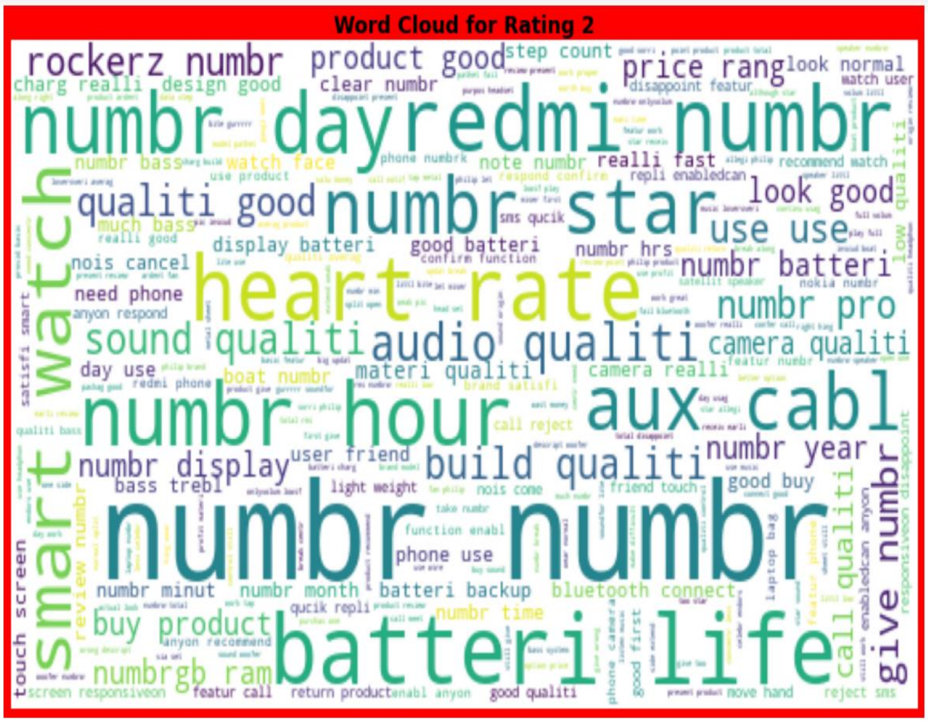
Comment:

1. Around 41% customer given 5- star rating followed by 28.9% customer given lowest 1-star rating.

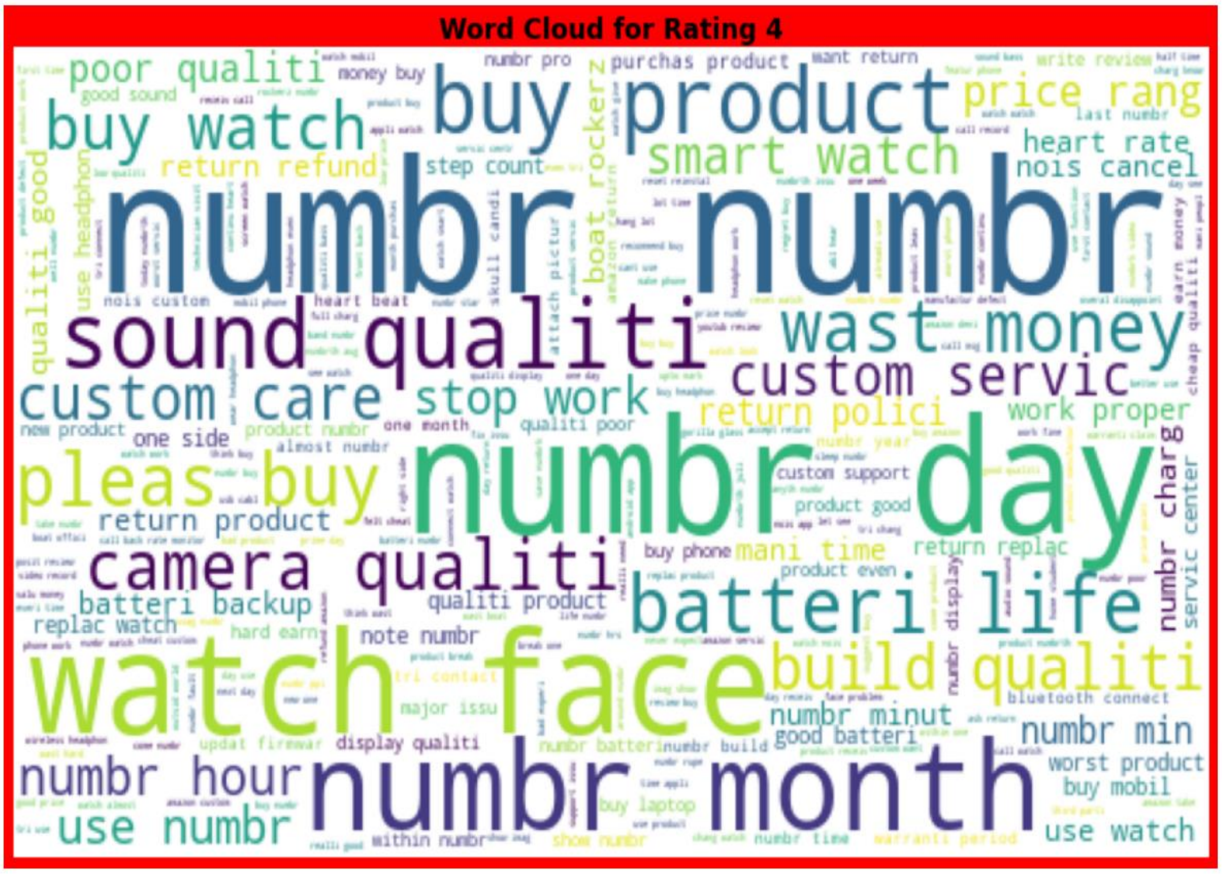
Word Cloud:

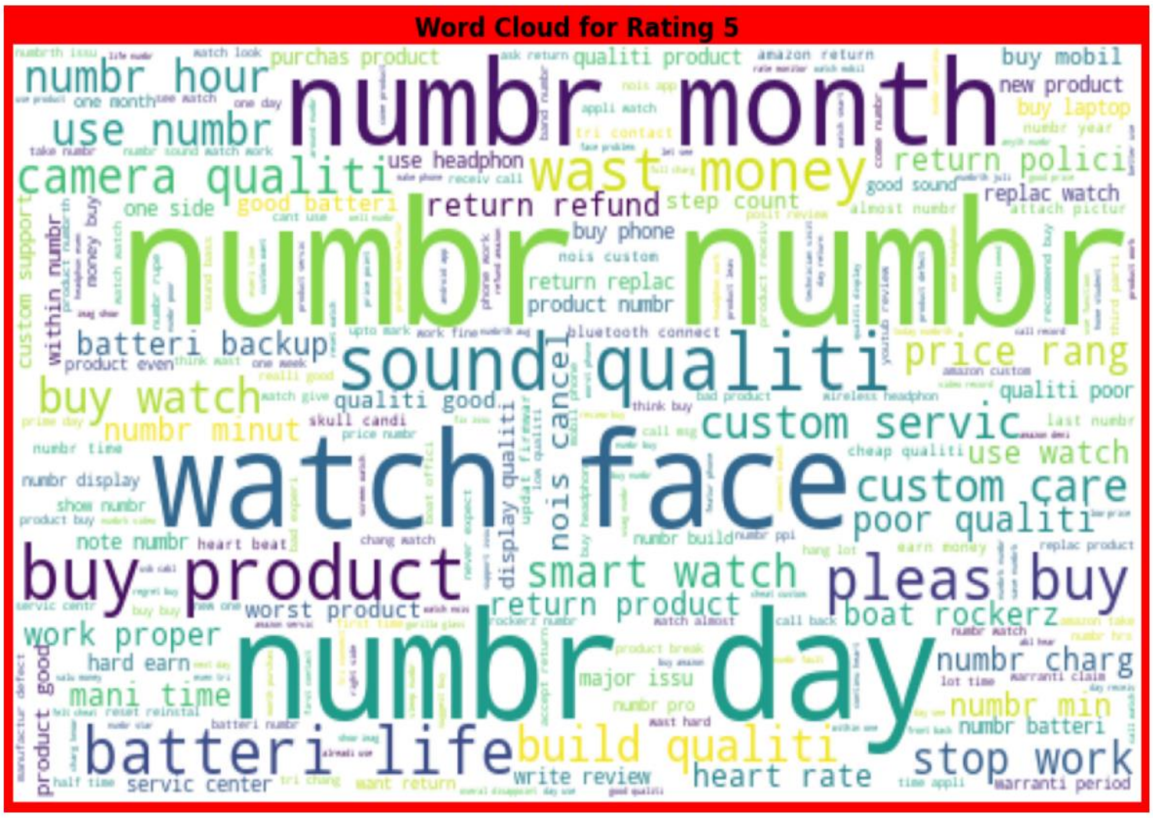
- Word Cloud is a visualization technique for text data wherein each word is picturized with its importance in the context or its frequency.
- The more commonly the term appears within the text being analysed, the larger the word appears in the image generated.
- The enlarged texts are the greatest number of words used there and small texts are the smaller number of words used.





Word Cloud for Rating 3
sound qualiti
build qualiti
numbr day
numbr numbr
batteri life
price rang
qualiti good
qualiti awesom
light weight
nois cancel


Word Cloud for Rating 4
numbr numbr
numbr day
sound qualiti
watch face
buy product
batteri life
numbr month
camera qualiti
wast money
custom servic
custom care


Word Cloud for Rating 5
numbr month
numbr numbr
watch face
numbr day
sound qualiti
buy product
batteri life
camera qualiti
wast money
custom servic
custom care
pleas buy
smart watch
return refund
return polici
buy watch
build qualiti
heart rate
stop work

Chap 4. Conclusion

- Final Model is giving us Accuracy score of 86.7% which is slightly improved compare to earlier Accuracy .

1. Learning Outcomes of the Study in respect of Data Science

- Hands on chance to enhance my web scraping skillset.
- In this project , able to learn various Natural language processing techniques like lemmatization, stemming, removal of Stop words.
- This project has demonstrated the importance of sampling effectively, modelling and predicting data.

2. Limitations of this work and Scope for Future Work

- More input features can be scrap to build prediction model.
- There is scope for application of advanced deep learning NLP tool to enhanced text mining operation which eventually help in building more accurate model with good cross validation score.