FLIP ROBO

REVIEWS&RATING PROJECT

SUBMITTED BY
Raghavulu Patnala

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot.

Some of the reference sources are as follows:

- Internet
- Coding Ninjas
- Medium.com
- Analytics Vidhya
- Stack Overflow

INTRODUCTION

BUSINESS PROBLEM FRAMING

☐ This is a Machine Learning Project performed on customer reveiws. Reviews areprocessed using common NLP techniques.

☐ Millions of people use **Amazon and Flipkart** to buy products. For every product,people can rate and write a review. If a product is good, it gets a positive review and gets a higher star rating, similarly, if a product is bad, it gets a negative reviewand lower star rating. My aim in this project is to predict star rating automatically based on the product review.

☐ The range of star rating is 1 to 5. That means if the product review is negative, then it will get low star rating (possibly 1 or 2), if the product is average then it will get medium star rating (possibly 3), and if the product is good, then it will get higherstar rating (possibly 4 or 5).

☐ This task is similar to Sentiment Analyis, but instead of predicting the positive and negative sentiment (sometimes neutral also), here I need to predict the star rating.

AIM OF THIS PROJECT

Our goal is to make a system that automatically detects the star rating based on the review.

CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

☐ The advent of electronic commerce with growth in internet and network technologies has led customers to move to online retail platforms such as Amazon, Walmart, Flip Kart, etc. People often rely on customer reviews of products before they buy online. These reviews are often rich in information describing the product. Customers often choose to compare between various products and brands based on whether an item has a positive or negative review. More often, these reviews act as a feedback mechanism for the seller. Through this medium, sellers strategize their future sales and product improvement.

☐ There is a client who has a website where people write different reviews for technical products. Now they want to add 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.

REVIEW OF LITERATURE

☐ This project is more about exploration, feature engineering and classification thatcan be done on this data. Since we scrape huge amount of data that includes five stars rating, we can do better data exploration and derive some interesting featuresusing the available columns.

☐ We can categorize the ratings as:

1.0, 2.0, 3.0, 4.0 and 5.0 stars

☐ The goal of this project is to build an application which predict the rating by can
seeing the review. In the long term, this would allow people to better explain and reviewing their purchase with each other in this increasingly digital world.

MOTIVATION OF THE PROBLEM UNDERTAKEN

☐ Every day we come across various products in our lives, on the digital medium we swipe across hundreds of product choices under one category. It will be tedious forthe customer to make selection. Here comes 'reviews' where customers who have already got that product leave a rating after using them and brief their experience by giving reviews.

☐ As we know ratings can be easily sorted and judged whether a product is good orbad. But when it comes to sentence reviews, we need to read through every line to make sure the review conveys a positive or negative sense. In the era of artificial intelligence, things like that have got easy with the Natural Language Processing (NLP) technology. Therefore, it is important to minimize the number of false positives our model produces, to encourage all constructive conversation.

☐ Our model also provides beneficence for the platform hosts as it replaces theneed to manually moderate discussions, saving time and resources. Employing amachine learning model to predict ratings promotes easier way to distinguish between products qualities, costs and many other features.

ANALYTICAL PROBLEM FRAMING

MATHEMATICAL/ANALYTICAL MODELLING OF THE PROBLEM

☐ In our scrapped dataset, our target variable Rating " is a **categorical** variable i.e., it can be classified as ‘1.0’, ‘2.0’,’3.0’,’4.0’,’5.0’. Therefore, we will be handling this modelling problem as classification.

☐ This project is done in two parts:

Data Collection Phase:

☐ You have to scrape at least 20000 rows of data. You can scrape more data as well,it’s up to you. More the data better the model.

☐ In this section you need to scrape the reviews of different laptops, Phones, Headphones, smart watches, Professional Cameras, Printers, monitors, hometheatre, router from different e-commerce websites.

☐ Basically, we need these columns

1) reviews of the product.
2) rating of the product.

☐ Fetch an equal number of reviews for each rating, for example if you are fetching
10000 reviews then all ratings 1,2,3,4,5 should be 2000. It will balance our data set.

☐ Convert all the ratings to their round number, as there are only 5 options for rating i.e., 1,2,3,4,5. If a rating is 4.5 convert it 5.

Model Building Phase:

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps involving NLP. Try different models with different hyper parameters and select the best model. Follow the complete life cycle of data science. Include all the steps like-

1. Data Cleaning
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building
5. Model Evaluation
6. Selecting the best model

DATA SOURCES AND THEIR FORMATS

In this phase, we scraped nearly 36000 of reviews data from Amazon of different products like laptop,phone and camera etc. and it is collected by using Webscraping and Selenium.

```
#reading the dataset file using pandas
df=pd.read_csv('Rating_Reviews.csv')
#Checking the dataset
df
```

	Unnamed: 0	Product_Review	Ratings
0	0	I own a beyerdynamic headphone & If you have a...	5.0
1	1	Well one says we should Compare apple to apple...	5.0
2	2	For the price excellent headphones. Great soun...	5.0
3	3	I don't understand what was all the hype assoc...	2.0
4	4	I have thoroughly used for 1 week So I will sa...	4.0
...	...	...	...
37995	37995	Recently I ordered it, while recording video f...	1.0
37996	37996	Very good display and in hand feel is great. G...	5.0
37997	37997	Best mid range 5G smartphone by OPPO.Pros :1. ...	5.0
37998	37998	My usecase - I replaced by secondary phone Red...	5.0
37999	37999	I am writing my open review of this phone afte...	2.0

38000 rows × 3 columns

☐ In the end, we combined all the data frames into a single data frame and it looks like as follows:

☐ Then, we will save this data in a csv file, so that we can do the pre-processing and model building.

DATA PRE-PROCESSING

☐ Handling missing data using fillna and checking the datatypes

```
In [6]: #Checking the dimensions of the dataset
        df.shape

Out[6]: (38000, 2)

In [5]: #checking information of the dataset
        df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38000 entries, 0 to 37999
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Product_Review  37280 non-null  object
 1   Ratings         38000 non-null  float64
dtypes: float64(1), object(1)
memory usage: 593.9+ KB
```

In information we observed many things like shape of the dataset 38000 columns and 2 columns,data types of a columns product review is onject type and ratings are float type and there are some null values present in product review column we need to fill null values.

Data pre-processing

```
In [9]: #checking for null values
        df.isnull().sum()

Out[9]: Product_Review    720
        Ratings             0
        dtype: int64
```

we have total 720 null values present in a product review column so we need to do fill that null values

```
In [10]: #We can handle missing data by filling them with 'No Review' using fillna()
         df['Product_Review'].fillna('No review',inplace=True)

In [11]: #Checking after filling null values
         df.isnull().sum()

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

now happy to see this there are no null values in a dataset

☐ Checking average rating and value counts of each rating present

```
print('Rating counts','\n',df.Ratings.value_counts())

Rating counts
 5.0    13562
1.0    13304
4.0     5773
3.0     3513
2.0     1848
Name: Ratings, dtype: int64
```

Pre-processing using Natural Language Processing (NLP):

☐ We cleaned the data using regex, matching patterns in the comments and replacing them with more organized counterparts. Cleaner data leads to a moreefficient model and higher accuracy. Following steps are involved:

1. Removing Punctuations and other special characters
2. Splitting the comments into individual words
3. Removing Stop Words

☐ There is a corpus of stopwords, that are high-frequency words such as "the", "to" and "also", and that we sometimes want to litter out of a document before further processing. Stop-words usually have little lexical content, don't alter the general meaning of a sentence and their presence in a text fails to distinguish it from other texts. We used the one from Natural Language Toolkit a leading platform for building Python programs to work with human language.

☐ The code is attached below:

```
def clean_text(df, df_column_name):

    #Converting all messages to lowercase
    df[df_column_name] = df[df_column_name].str.lower()

    #Replace email addresses with 'email'
    df[df_column_name] = df[df_column_name].str.replace(r'^.+@[^\.].*\.[a-z]{2,}$','emaila

    #Replace URLs with 'webaddress'
    df[df_column_name] = df[df_column_name].str.replace(r'^http\://[a-zA-Z0-9\-\.]+\.[a-zA

    #Replace money symbols with 'dollars' (£ can by typed with ALT key + 156)
    df[df_column_name] = df[df_column_name].str.replace(r'£|\$', 'dollars')

    #Replace 10 digit phone numbers (formats include paranthesis, spaces, no spaces, dashes
    df[df_column_name] = df[df_column_name].str.replace(r'^\(?[\d]{3}\)?[\s-]?[\d]{3}[\s-]

    #Replace numbers with 'numbr'
    df[df_column_name] = df[df_column_name].str.replace(r'\d+(\.\d+)?', 'numbr')

    #Remove punctuation
    df[df_column_name] = df[df_column_name].str.replace(r'[^\w\d\s]', ' ')

    #Replace whitespace between terms with a single space
    df[df_column_name] = df[df_column_name].str.replace(r'\s+', ' ')

    #Remove leading and trailing whitespace
    df[df_column_name] = df[df_column_name].str.replace(r'^\s+|\s+?$', '')

    #Remove stopwords
    stop_words = set(stopwords.words('english') + ['u', 'ü', 'â', 'ur', '4', '2', 'im', 'd
    df[df_column_name] = df[df_column_name].apply(lambda x: ' '.join(term for term in x.sp
```

```
#Calling the class
clean_text(df, 'Product_Review')
df['Product_Review']
```

```
0        beyerdynamic headphone idea premium headphones...
1        well one says compare apple apple went ahead c...
2        price excellent headphones great soundstage ma...
3        understand hype associated headphones found av...
4        thoroughly used numbr week say pros cons headp...
                               ...
37995    recently ordered recording video front camera ...
37996    good display hand feel great good camera works...
37997    best mid range numbrg smartphone oppo pros num...
37998    usecase replaced secondary phone redmi note nu...
37999    writing open review phone using approximately ...
Name: Product_Review, Length: 38000, dtype: object
```

```
# Lemmatizing and then Stemming with Snowball to get root words and further reducing charac
stemmer = SnowballStemmer("english")
import gensim
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text,pos='v'))

#Tokenize and Lemmatize
def preprocess(text):
    result=[]
    for token in text:
        if len(token)>=3:
            result.append(lemmatize_stemming(token))

    return result
```

```
#Processing review with above Function
processed_review = []

for doc in df.Product_Review:
    processed_review.append(preprocess(doc))

print(len(processed_review))
processed_review[:3]
```

```
38000
```

- Tokenizing the data using RegexpTokenizer

```
#Calling the class
clean_text(df, 'Product_Review')
df['Product_Review']
```

```
0        beyerdynamic headphone idea premium headphones...
1        well one says compare apple apple went ahead c...
2        price excellent headphones great soundstage ma...
3        understand hype associated headphones found av...
4        thoroughly used numbr week say pros cons headp...
                                ...
37995    recently ordered recording video front camera ...
37996    good display hand feel great good camera works...
37997    best mid range numbrg smartphone oppo pros num...
37998    usecase replaced secondary phone redmi note nu...
37999    writing open review phone using approximately ...
Name: Product_Review, Length: 38000, dtype: object
```

```
#Tokenizing the data using RegexpTokenizer
from nltk.tokenize import RegexpTokenizer
tokenizer=RegexpTokenizer(r'\w+')
df['Product_Review'] = df['Product_Review'].apply(lambda x: tokenizer.tokenize(x.lower()))
df.head()
```

	Product_Review	Ratings
0	[beyerdynamic, headphone, idea, premium, headp...	5.0
1	[well, one, says, compare, apple, apple, went,...	5.0
2	[price, excellent, headphones, great, soundsta...	5.0
3	[understand, hype, associated, headphones, fou...	2.0
4	[thoroughly, used, numbr, week, say, pros, con...	4.0

Stemming and Lemmatizing:

• **Stemming** is the process of converting inflected/derived words to their word stem or the root form. Basically, a large number of similar origin words are converted to the same word. E.g., words like "stems", "stemmer", "stemming", "stemmed" as based on "stem". This helps in achieving the training process with a better accuracy.

```
# Lemmatizing and then Stemming with Snowball to get root words and further reducing charac
stemmer = SnowballStemmer("english")
import gensim
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text,pos='v'))

#Tokenize and Lemmatize
def preprocess(text):
    result=[]
    for token in text:
        if len(token)>=3:
            result.append(lemmatize_stemming(token))

    return result
```

```
#Processing review with above Function
processed_review = []

for doc in df.Product_Review:
    processed_review.append(preprocess(doc))

print(len(processed_review))
processed_review[:3]
```

```
38000
```

• **Lemmatizing** is the process of grouping together the inflected forms of a word so they can be analysed as a single item. This is quite similar to stemming in its working but differs since it depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighbouring sentences or even an entire document.
• The **wordnet library in nltk** will be used for this purpose. Stemmer and Lemmatizer are also imported from nltk.
☐ Processing the review and assigning the updated review in the data frame

df['Product_Review'] = df['clean_review'].apply(lambda x:' '.join(y for y in x))

```
df['clean_review']=processed_review  #Assigning this to the dataframe
df.head()
```

	Product_Review	Ratings	clean_review
0	[beyerdynamic, headphone, idea, premium, headp...	5.0	[beyerdynam, headphon, idea, premium, headphon...
1	[well, one, says, compare, apple, apple, went,...	5.0	[well, one, say, compar, appl, appl, go, ahead...
2	[price, excellent, headphones, great, soundsta...	5.0	[price, excel, headphon, great, soundstag, may...
3	[understand, hype, associated, headphones, fou...	2.0	[understand, hype, associ, headphon, find, ave...
4	[thoroughly, used, numbr, week, say, pros, con...	4.0	[thorough, use, numbr, week, say, pros, con, h...

```
df['Product_Review'] = df['clean_review'].apply(lambda x:' '.join(y for y in x))
df.head()
```

	Product_Review	Ratings	clean_review
0	beyerdynam headphon idea premium headphon must...	5.0	[beyerdynam, headphon, idea, premium, headphon...
1	well one say compar appl appl go ahead compar ...	5.0	[well, one, say, compar, appl, appl, go, ahead...
2	price excel headphon great soundstag may bassi...	5.0	[price, excel, headphon, great, soundstag, may...
3	understand hype associ headphon find averag go...	2.0	[understand, hype, associ, headphon, find, ave...
4	thorough use numbr week say pros con headphon ...	4.0	[thorough, use, numbr, week, say, pros, con, h...

☐ Getting sense of words for all ratings using WordCloud

Word Cloud is a data visualization technique used for representing text data in which the size of each **word** indicates its frequency or importance.
Similarly, we found the sense of words for ratings 2.0 – 5.0 and the output will be as follows:

```
df['clean_review']=processed_review  #Assigning this to the dataframe
df.head()
```

	Product_Review	Ratings	clean_review
0	[beyerdynamic, headphone, idea, premium, headp...	5.0	[beyerdynam, headphon, idea, premium, headphon...
1	[well, one, says, compare, apple, apple, went,...	5.0	[well, one, say, compar, appl, appl, go, ahead...
2	[price, excellent, headphones, great, soundsta...	5.0	[price, excel, headphon, great, soundstag, may...
3	[understand, hype, associated, headphones, fou...	2.0	[understand, hype, associ, headphon, find, ave...
4	[thoroughly, used, numbr, week, say, pros, con...	4.0	[thorough, use, numbr, week, say, pros, con, h...

```
df['Product_Review'] = df['clean_review'].apply(lambda x:' '.join(y for y in x))
df.head()
```

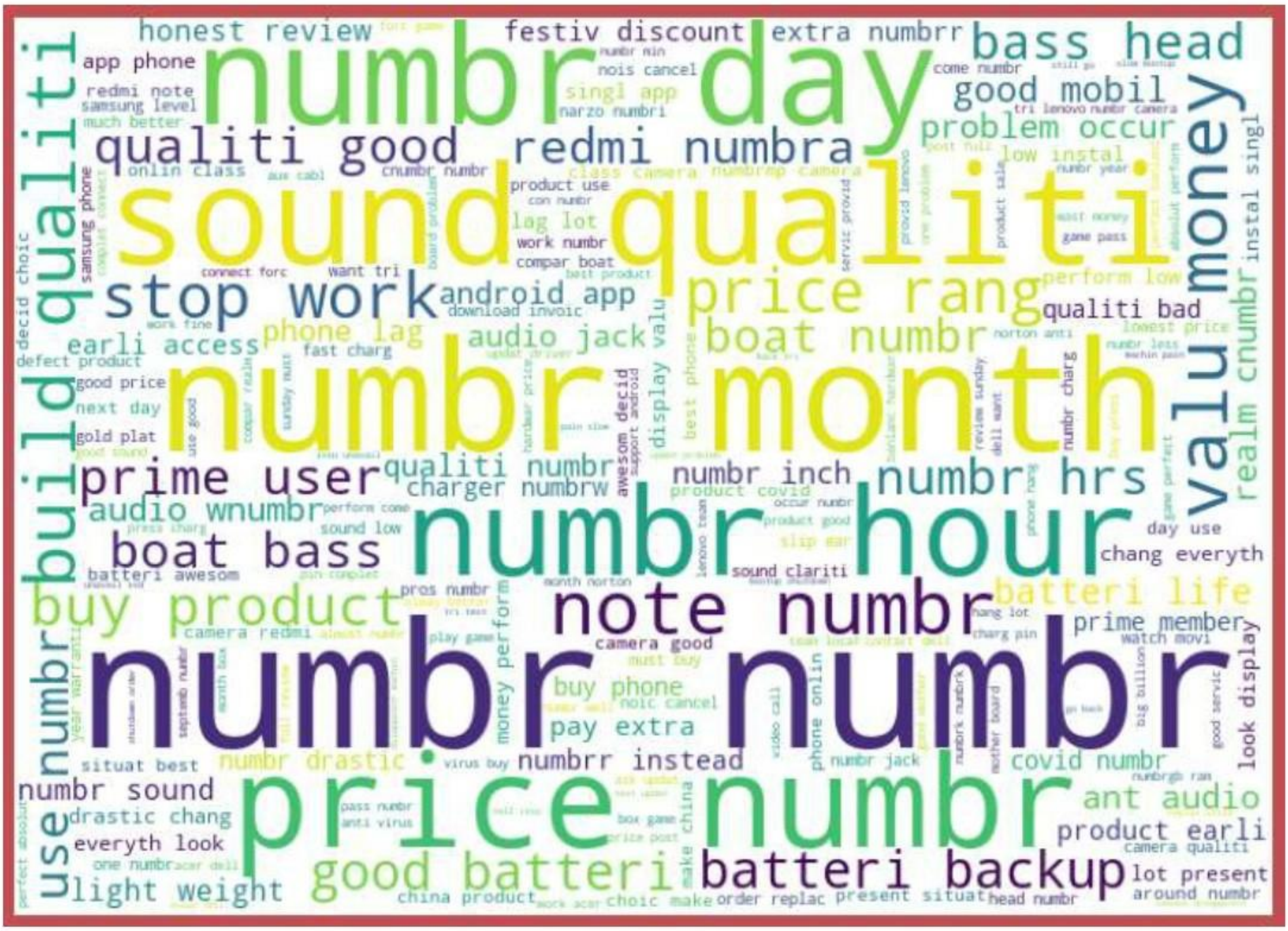
	Product_Review	Ratings	clean_review
0	beyerdynam headphon idea premium headphon must...	5.0	[beyerdynam, headphon, idea, premium, headphon...
1	well one say compar appl appl go ahead compar ...	5.0	[well, one, say, compar, appl, appl, go, ahead...
2	price excel headphon great soundstag may bassi...	5.0	[price, excel, headphon, great, soundstag, may...
3	understand hype associ headphon find averag go...	2.0	[understand, hype, associ, headphon, find, ave...
4	thorough use numbr week say pros con headphon ...	4.0	[thorough, use, numbr, week, say, pros, con, h...

For rating 5.0:

```python
#Getting sense of words in Rating 5
one = df['Product_Review'][df['Ratings']==5.0]

one_cloud = WordCloud(width=700,height=500,background_color='white',max_words=200).generat

plt.figure(figsize=(10,8),facecolor='r')
plt.imshow(one_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```

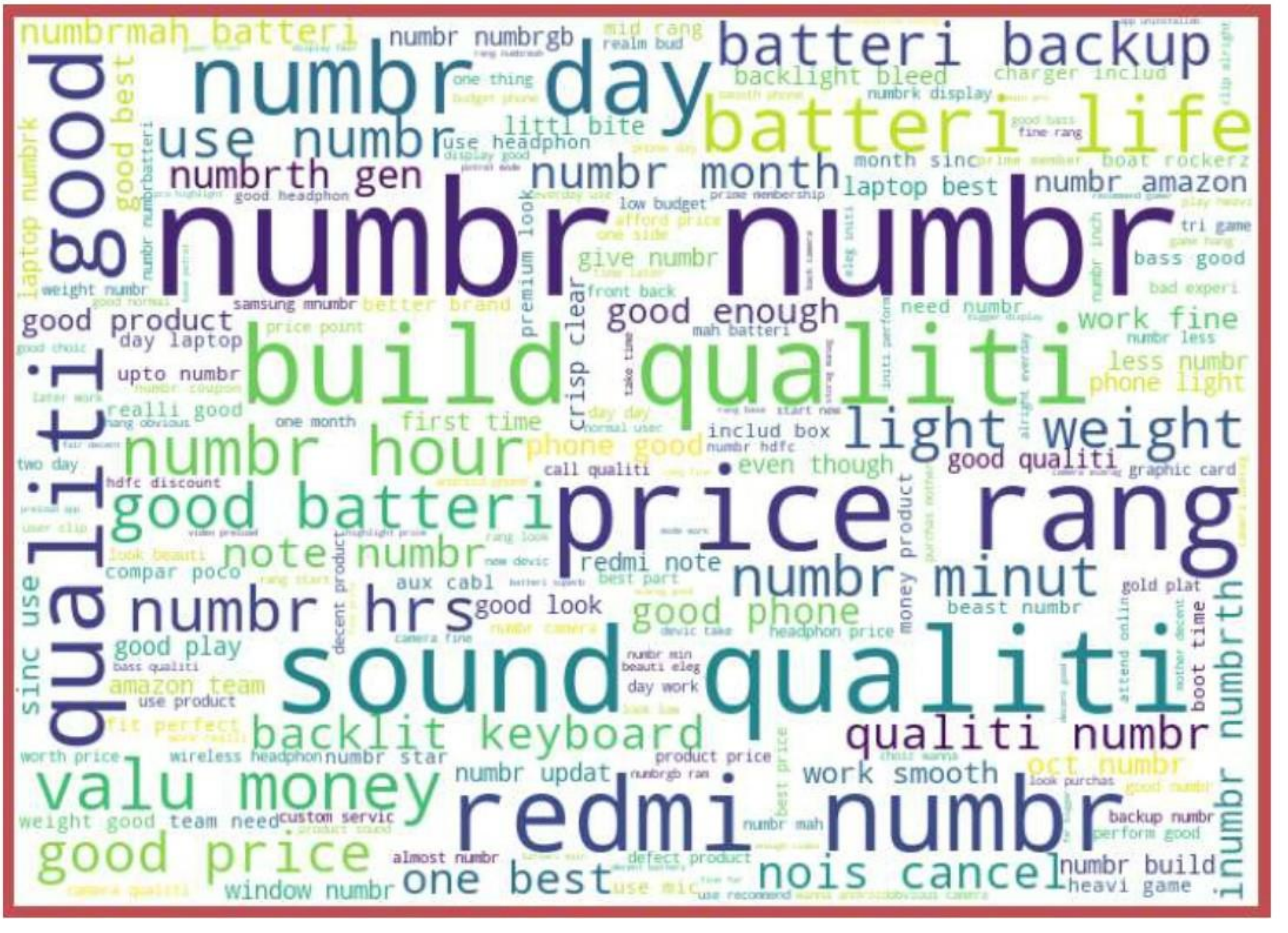
For rating 3.0:

```
#Getting sense of words in Rating 3
one = df['Product_Review'][df['Ratings']==3.0]

one_cloud = WordCloud(width=700,height=500,background_color='white',max_words=200).generate

plt.figure(figsize=(10,8),facecolor='r')
plt.imshow(one_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```



For rating 4.0:

```
#Getting sense of words in Rating 4
one = df['Product_Review'][df['Ratings']==4.0]

one_cloud = WordCloud(width=700,height=500,background_color='white',max_words=200).generat

plt.figure(figsize=(10,8),facecolor='r')
plt.imshow(one_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```



For rating 2.0:

```
#Getting sense of words in Rating 2
one = df['Product_Review'][df['Ratings']==2.0]

one_cloud = WordCloud(width=700,height=500,background_color='white',max_words=200).generate

plt.figure(figsize=(10,8),facecolor='r')
plt.imshow(one_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```

Observations:
The enlarged texts are the most number of words used there and small texts are the less number of words used.
It varies according to the ratings.
Feature Extraction:
Here we can finally convert our text to numeric using Tf-idf Vectorizer.
Term Frequency Inverse Document Frequency (TF-IDF):
This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

Feature Extraction

```
#Converting text into numeric using TfidfVectorizer
#create object
tf = TfidfVectorizer()

#fitting
features = tf.fit_transform(df['Product_Review'])
x=features
y=df[['Ratings']]

x.shape
```

```
(38000, 3274)
```

```
y.shape
```

```
(38000, 1)
```

HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

HARDWARE:

HP ENVI X360AQ105X

SOFTWARE:

Jupyter Notebook (Anaconda 3) – Python 3.7.6

Libraries Used:

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

MODEL/S DEVELOPMENT AND EVALUATION

☐ Listing down all the algorithms used for the training and testing.

Model building

```
#Importing model building libraries
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.metrics import confusion_matrix,classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
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.metrics import accuracy_score
```

```
#Initializing the instance of the model
LR=LogisticRegression()
mnb=MultinomialNB()
dtc=DecisionTreeClassifier()
knc=KNeighborsClassifier()
rfc=RandomForestClassifier()
abc=AdaBoostClassifier()
gbc=GradientBoostingClassifier()

models= []
models.append(('Logistic Regression',LR))
models.append(('MultinomialNB',mnb))
models.append(('DecisionTreeClassifier',dtc))
models.append(('KNeighborsClassifier',knc))
models.append(('RandomForestClassifier',rfc))
models.append(('AdaBoostClassifier',abc))
models.append(('GradientBoostingClassifier',gbc))
```

```
#Creating train_test_split using best random_state
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=56,test_size=.20)
```

☐ Running and evaluating the models

Model building

```
#Importing model building libraries
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.metrics import confusion_matrix,classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
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.metrics import accuracy_score
```

```
#Initializing the instance of the model
LR=LogisticRegression()
mnb=MultinomialNB()
dtc=DecisionTreeClassifier()
knc=KNeighborsClassifier()
rfc=RandomForestClassifier()
abc=AdaBoostClassifier()
gbc=GradientBoostingClassifier()

models= []
models.append(('Logistic Regression',LR))
models.append(('MultinomialNB',mnb))
models.append(('DecisionTreeClassifier',dtc))
models.append(('KNeighborsClassifier',knc))
models.append(('RandomForestClassifier',rfc))
models.append(('AdaBoostClassifier',abc))
models.append(('GradientBoostingClassifier',gbc))
```

```
#Creating train_test_split using best random_state
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=56,test_size=.20)
```

```
#Making a for loop and calling the algorithm one by one and save data to respective model 
Model=[]
score=[]
cvs=[]
rocscore=[]
for name,model in models:
    print('******************************',name,'****************************')
    print('\n')
    Model.append(name)
    model.fit(x_train,y_train)
    print(model)
    pre=model.predict(x_test)
    print('\n')
    AS=accuracy_score(y_test,pre)
    print('accuracy_score: ',AS)
    score.append(AS*100)
    print('\n')
    sc=cross_val_score(model,x,y,cv=5,scoring='accuracy').mean()
    print('cross_val_score: ',sc)
    cvs.append(sc*100)
    print('\n')
    print('Classification report:\n ')
    print(classification_report(y_test,pre))
    print('\n')
    print('Confusion matrix: \n')
    cm=confusion_matrix(y_test,pre)
    print(cm)
    print('\n\n\n')
```

```
****************************** Logistic Regression ****************************

LogisticRegression()

accuracy_score:  0.7682894736842105

cross_val_score:  0.3528157894736842

Classification report:

              precision    recall  f1-score   support

         1.0       0.73      0.86      0.79      2684
         2.0       0.80      0.65      0.72       378
         3.0       0.79      0.57      0.66       657
         4.0       0.81      0.65      0.72      1213
         5.0       0.79      0.80      0.79      2668

    accuracy                           0.77      7600
   macro avg       0.78      0.71      0.74      7600
weighted avg       0.77      0.77      0.77      7600


Confusion matrix:

[[2298   14   32   78  262]
 [  61  247    0   16   54]
```

```
 [ 179    1  372   23   82]
 [ 237    3   12  794  167]
 [ 376   44   53   67 2128]]


******************************* MultinomialNB ***************************

MultinomialNB()

accuracy_score:  0.7392105263157894

cross_val_score:  0.355

Classification report:
              precision    recall  f1-score   support

         1.0       0.71      0.83      0.76      2684
         2.0       0.79      0.63      0.70       378
         3.0       0.71      0.56      0.62       657
         4.0       0.81      0.60      0.69      1213
         5.0       0.75      0.77      0.76      2668

    accuracy                           0.74      7600
   macro avg       0.75      0.68      0.71      7600
weighted avg       0.74      0.74      0.74      7600


Confusion matrix:

[[2217   16   52   73  326]
 [  60  239    0   14   65]
 [ 165    4  368   29   91]
 [ 270    3   14  728  198]
 [ 418   41   87   56 2066]]


******************************* DecisionTreeClassifier ***************************

DecisionTreeClassifier()

accuracy_score:  0.7684210526315789

cross_val_score:  0.34031578947368424

Classification report:
              precision    recall  f1-score   support

         1.0       0.73      0.85      0.79      2684
         2.0       0.80      0.68      0.73       378
         3.0       0.77      0.58      0.66       657
         4.0       0.79      0.67      0.72      1213
         5.0       0.80      0.79      0.79      2668

    accuracy                           0.77      7600
   macro avg       0.78      0.71      0.74      7600
weighted avg       0.77      0.77      0.77      7600


Confusion matrix:

[[2286   18   34   86  260]
 [  52  256    0   26   44]
 [ 174    1  380   28   74]
 [ 232    3   16  812  150]
 [ 376   44   61   81 2106]]



***************************** KNeighborsClassifier **************************

KNeighborsClassifier()

accuracy_score:  0.7218421052631578

cross_val_score:  0.3266578947368421

Classification report:

              precision    recall  f1-score   support

         1.0       0.75      0.74      0.74      2684
         2.0       0.73      0.60      0.66       378
         3.0       0.51      0.55      0.53       657
         4.0       0.73      0.64      0.68      1213
         5.0       0.75      0.80      0.77      2668

    accuracy                           0.72      7600
   macro avg       0.69      0.67      0.68      7600
weighted avg       0.72      0.72      0.72      7600


Confusion matrix:

[[1989   28  182  137  348]
 [  40  225   13   13   87]
 [ 142    9  364   34  108]
 [ 197    9   59  776  172]
 [ 301   37   91  107 2132]]
```

```
***************************** RandomForestClassifier **************************

RandomForestClassifier()

accuracy_score:  0.7685526315789474

cross_val_score:  0.36115789473684207

Classification report:
              precision    recall  f1-score   support

         1.0       0.73      0.85      0.79      2684
         2.0       0.80      0.68      0.73       378
         3.0       0.77      0.58      0.66       657
         4.0       0.78      0.68      0.73      1213
         5.0       0.80      0.79      0.79      2668

    accuracy                           0.77      7600
   macro avg       0.78      0.71      0.74      7600
weighted avg       0.77      0.77      0.77      7600


Confusion matrix:
[[2278   18   34   86  268]
 [  52  256    0   26   44]
 [ 174    1  380   28   74]
 [ 230    3   16  825  139]
 [ 368   44   61   93 2102]]


***************************** AdaBoostClassifier **************************

AdaBoostClassifier()

accuracy_score:  0.4425

cross_val_score:  0.3286578947368421

Classification report:
              precision    recall  f1-score   support

         1.0       0.43      0.71      0.54      2684
         2.0       0.72      0.14      0.24       378
         3.0       0.55      0.20      0.29       657
         4.0       0.36      0.04      0.08      1213
         5.0       0.45      0.46      0.45      2668

    accuracy                           0.44      7600
   macro avg       0.50      0.31      0.32      7600
weighted avg       0.45      0.44      0.40      7600


Confusion matrix:
[[1902    3   50   39  690]
 [ 179   54    4   27  114]
 [ 386    0  129    2  140]
 [ 586    2   12   54  559]
 [1358   16   40   30 1224]]


***************************** GradientBoostingClassifier **************************

GradientBoostingClassifier()

accuracy_score:  0.7459210526315789

cross_val_score:  0.3471842105263158

Classification report:
              precision    recall  f1-score   support

         1.0       0.72      0.82      0.77      2684
         2.0       0.83      0.59      0.69       378
         3.0       0.85      0.50      0.63       657
         4.0       0.90      0.56      0.69      1213
         5.0       0.71      0.84      0.77      2668

    accuracy                           0.75      7600
   macro avg       0.80      0.66      0.71      7600
weighted avg       0.76      0.75      0.74      7600


Confusion matrix:
[[2189    8   23   39  425]
 [  64  222    0    4   88]
 [ 171    1  329   19  137]
 [ 263    2    8  685  255]
 [ 349   36   26   13 2244]]
```

```
#Finalizing the result
result=pd.DataFrame({'Model':Model, 'Accuracy_score': score,'Cross_val_score':cvs})
result
```

	Model	Accuracy_score	Cross_val_score
0	Logistic Regression	76.828947	35.281579
1	MultinomialNB	73.921053	35.500000
2	DecisionTreeClassifier	76.842105	34.031579
3	KNeighborsClassifier	72.184211	32.665789
4	RandomForestClassifier	76.855263	36.115789
5	AdaBoostClassifier	44.250000	32.865789
6	GradientBoostingClassifier	74.592105	34.718421

Key Metrics for success in solving problem under consideration

The key metrics used here were accuracy_score, cross_val_score, classification report, and confusion matrix. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

1. Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces where each part being 20% of full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

2. Confusion Matrix:

A **confusion matrix**, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a **matching matrix**). Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e., commonly mislabelling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

3. Classification Report:

The classification report visualizer displays the precision, recall, F1, and support scores for the model. There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative**: the case was negative and predicted negative
2. **TP / True Positive**: the case was positive and predicted positive
3. **FN / False Negative**: the case was positive but predicted negative
4. **FP / False Positive**: the case was negative but predicted positive

Precision: Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive

and false positive. It is the accuracy of positive predictions. The formula of precision is given below:
Precision = TP/ (TP + FP)

Recall: Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. It is also the fraction of positives that were correctly identified. The formula of recall is given below:
Recall = TP/(TP+FN)

F1 score: The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. The formula is:

F1 Score = 2*(Recall * Precision) /
(Recall + Precision)

Support: Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

4. Hyperparameter Tuning:

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.** These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. You must select from a specific list of hyperparameters for a given model as it varies from model to model.

We are not aware of optimal values for hyperparameters which would generate the best model output. So, what we tell the model is to explore and select the optimal model architecture automatically. This selection procedure for hyperparameter is known as **Hyperparameter Tuning. We can do tuning by using GridSearchCV.**

GridSearchCV is a function that comes in Scikit-learn (or SK-learn) model selection package. An important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

Hyper Pramater Tuning

```
#RandomForestClassifier
parameters={'n_estimators':[1,10,100]}
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
rfc=RandomForestClassifier(random_state=76)   #Using the best random state we obtained
rfc=GridSearchCV(rfc,parameters,cv=3,scoring='accuracy')
rfc.fit(x_train,y_train)
print(rfc.best_params_)      #Printing the best parameters obtained
print(rfc.best_score_)       #Mean cross-validated score of best_estimator
```

```
{'n_estimators': 100}
0.7746053470683214
```

```
#Using the best parameters obtained
rfc=RandomForestClassifier(random_state=56,n_estimators=100)
rfc.fit(x_train,y_train)
pred=rfc.predict(x_test)
print("Accuracy score: ",accuracy_score(y_test,pred)*100)
print('Cross validation score: ',cross_val_score(rfc,x,y,cv=3,scoring='accuracy').mean()*1(
print('Classification report: \n')
print(classification_report(y_test,pred))
print('Confusion matrix: \n')
print(confusion_matrix(y_test,pred))
```

```
Accuracy score:  76.84210526315789
Cross validation score:  40.23429361746957
Classification report:

              precision    recall  f1-score   support

         1.0       0.73      0.85      0.79      2684
         2.0       0.80      0.68      0.73       378
         3.0       0.77      0.58      0.66       657
         4.0       0.79      0.66      0.72      1213
         5.0       0.80      0.79      0.79      2668

    accuracy                           0.77      7600
   macro avg       0.78      0.71      0.74      7600
weighted avg       0.77      0.77      0.77      7600

Confusion matrix:

[[2286   18   34   86  260]
 [  52  256    0   21   49]
 [ 174    1  380   28   74]
 [ 232    3   16  806  156]
 [ 376   44   61   75 2112]]
```

After applying Hyperparameter Tuning, we can see that RandomForestClassifier Algorithm is performing well as the scores constant, i.e., accuracy score is 76% and cross_val_score from 36% to 40%. Now, we will finalizeRandom ForestClassifier algorithm model as the final model.

Final the model and save the model

```
rfc_prediction=rfc.predict(x)

#Making a dataframe of predictions
rating_prediction=pd.DataFrame({'Predictions':rfc_prediction})
rating_prediction
```

	Predictions
0	5.0
1	5.0
2	5.0
3	2.0
4	4.0
...	...
37995	1.0
37996	5.0
37997	5.0
37998	5.0
37999	2.0

38000 rows × 1 columns

Saving the Model

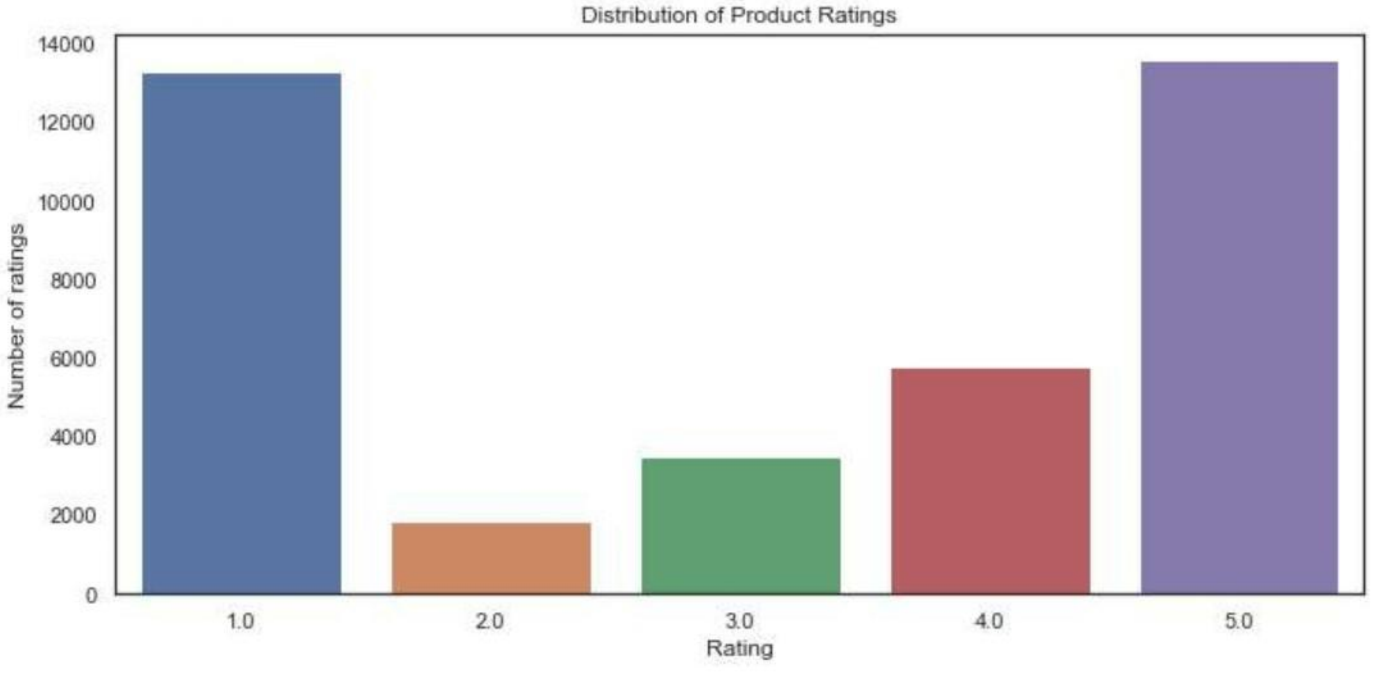
```
#saving our model

import joblib
joblib.dump(rfc,'Rating&Reviews_prediction.csv')
```

```
['Rating&Reviews_prediction.csv']
```

```
#Saving predicted values
rating_prediction.to_csv('Rating&Reviews_Prediction_Results.csv')
```

DATA VISUALIZATION

```
#let's visualize the count of Ratings variable using Seaborn
sns.set_theme(style = 'white')
plt.figure(figsize = (10,5))
ax = sns.countplot(x = df['Ratings'])
ax.set(title="Distribution of Product Ratings", xlabel="Rating", ylabel="Number of ratings'
plt.tight_layout()
```



Summary

After the completion of this project, we got an insight of how to collect data, preprocessing the data, analyzing the data and building a model.

1.we collected the reviews and ratings data from e-commerce website Amazon it was done by using Webscraping. The framework used for webscraping was Selenium, which has an advantage of automating our process of collecting data.

2.We collected almost 38000 rows of data which contained the ratings from 1.0 to 5.0and their reviews.

3.then, the scrapped data was combined in a single dataframe and saved in a csv fileso that we can open it and analyze the data.

4.We did the preprocessing using NLP and the steps are as follows:

a.Removing Punctuations and other special characters
b.Splitting the comments into individual words

c.Removing Stop Words

d.Stemming and Lemmatising

e.Applying Count Vectoriser

f.Splitting dataset into Training and Testing

5. After separating our train and test data, we started running different machine learning classification algorithms to find out the best performing model.

6. We found that RandomForest is performing well, according to their accuracy and cross val scores.

7. Then, we performed Hyperparameter Tuning techniques using GridSearchCV for getting the best parameters and constant the score. In that, RandomForestClassifier performed well and we finalised that model.

8. We saved the model saved the predicted values in a csvformat.

9. The problems we faced during this project were:

a.More time consumption during hyperparameter tuning model, as the data was large.

b.Less number of parameters were used during tuning.

c.Scrapping of data from different websites were of different process and the length of data were differing in most cases so I sticked to Amazon and Scrapped data which are famousin the site.

d.Some of the reviews were bad and the text had more wrong information about the product.

e.WordCloud was not showing proper text which had more positive and negative weightage.

10. Areas of improvement:

a.Less time complexity

b.More accurate reviews can be given

c.Less errors can be avoided.