Fliprobo

Customer Ratings Prediction Model

Report

Submitted by:

Arjun Verma,

Intern Data Scientist

ACKNOWLEDGEMENT

I would like to express my greatest appreciation to the all individuals who have helped and supported me throughout the project. I am thankful to Fliprobo team for their ongoing support during the project, from initial advice, and encouragement, which led to the final report of this project.

A special acknowledgement goes to my institute Datatrained who helped me in completing the project and learning concepts.

I wish to thank my parents as well for their undivided support and interest who inspired me and encouraged me to go my own way, without whom I would be unable to complete my project.

Below following are the other references:

www.towardsdatascience.com

www.medium.com

www.stackoverflow.com

Datatrained lectures

INTRODUCTION

➢ Business Problem Framing

We got 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

Companies such as flipkart.com, amazon.com etc which are technical sites where most of the people buys technical products. But before buying a product we have to find out good technical products for the client site. Similarly in the given task we have to build a model that can predict best technical product for their site by checking their ratings from the extracted dataset from old buyers.

➢ Review of Literature

Data has been collected from website flipkart.com. We collected most of the important technical products rating and import into the dataset that can impact the model building precisely. Model is created using the data by splitting the data as dependent and independent variable. These dataset are further split into test and train. The train data is trained through various classification algorithms. The algorithm having the lease difference between accuracy score and cross val score will be used for hyperparameter tuning. The best parameters are used to tune the model. This model is given to the client in further using to visualise data for future car price prediction.

➢ Motivation for the Problem Undertaken

Genuinely it's a need of the any site to complete their goal with higher revenue and low expenditure with customer satisfaction. Hence this model can brings higher revenue because we can predict ratings of the customer by taking their reviews.

➢ Mathematical/ Analytical Modeling of the Problem

Data is statistically analysed through TFIDF vectorization techniques.. Graphical modelling done through seaborn and matplotlib to understanding how different features impact dataset.

Statistical models used
- ➢ Logistics Regression
- ➢ Naïve Bayes
- ➢ Bernoulli
- ➢ SGD Classifier

➢ Data Sources and their formats

Datasets are extracted by site flipkart.com for building machine learning model to predict rating of the product based on given parameter.

Dataset is having 30450 rows and 4 columns including target.

The information about features are as follows

```
'Unnamed: 0', 'Ratings', 'Review',
'Product'
```

Dataframe Description:

- **Ratings : Ratings of the customer**
- **Review : Review of the customers**
- **Product : Products**
- **We make a addition columns by taking length of the review**

	Ratings	Review	Product	length
0	4	A bit expensive when we compare with today's i...	Laptop	498
1	5	Fantastic value for money machine!! Absolute b...	Laptop	499
2	5	The best you can get, looks and performance bo...	Laptop	334
3	5	Ultimate machine, best laptop I have ever used...	Laptop	183
4	5	For everyone, who is planning to buy MBA M1- ...	Laptop	500

We have done feature engineering for preprocessing dataset and get below information

Dataset Information

'Ratings', 'Review', 'Product' all are objects columns

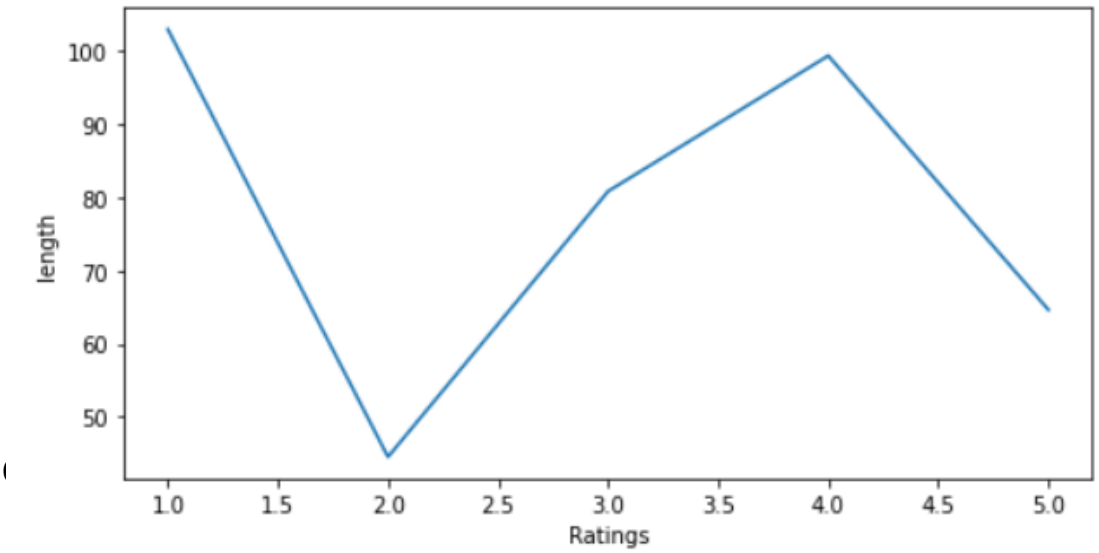
Checking Null Values of the dataset

Dataset having no null values.

Dataset having 13127 duplicated values which have been removed

Visualization of important features for understanding

Customer ratings values

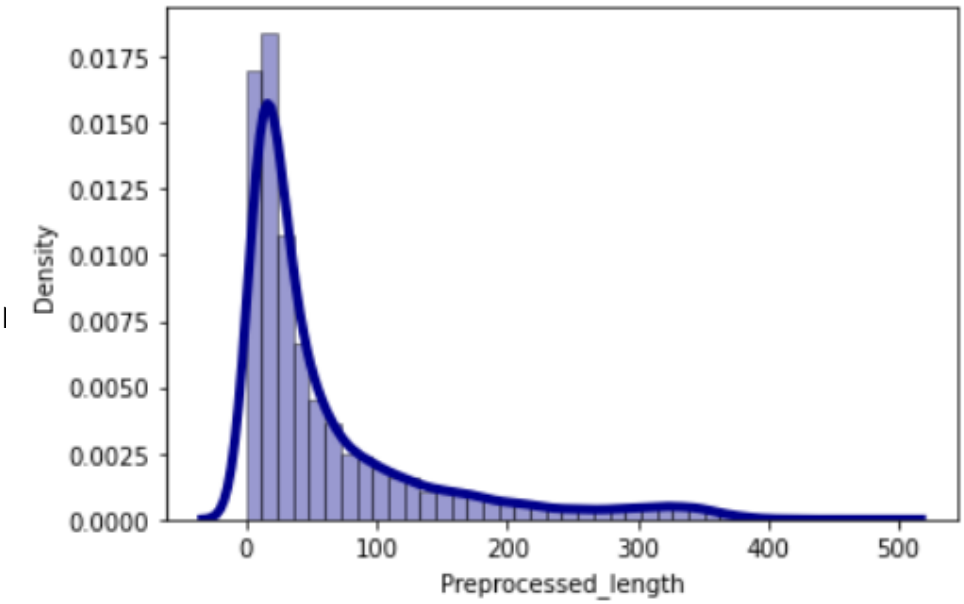


Length of the rating is highest for 4 and lowest for 2

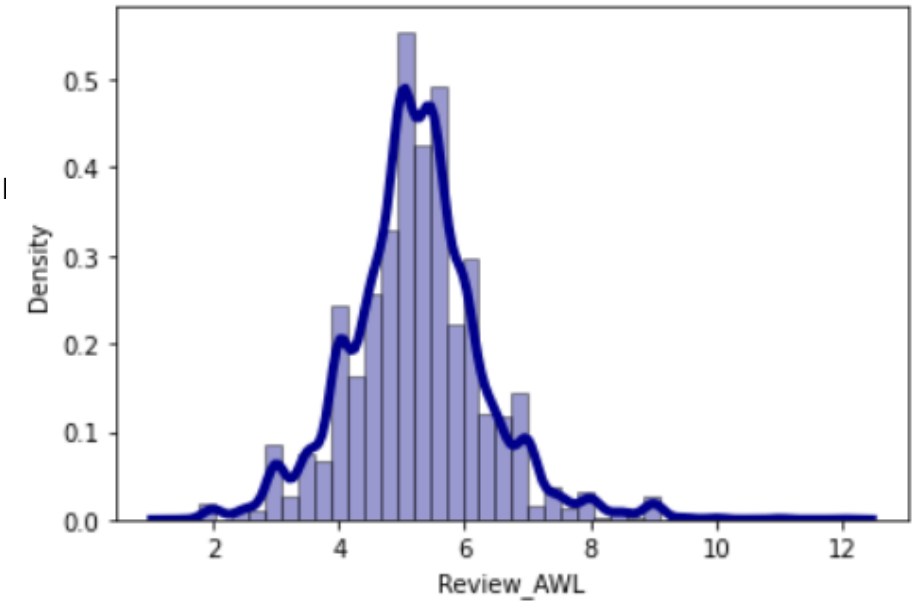
.

EDA

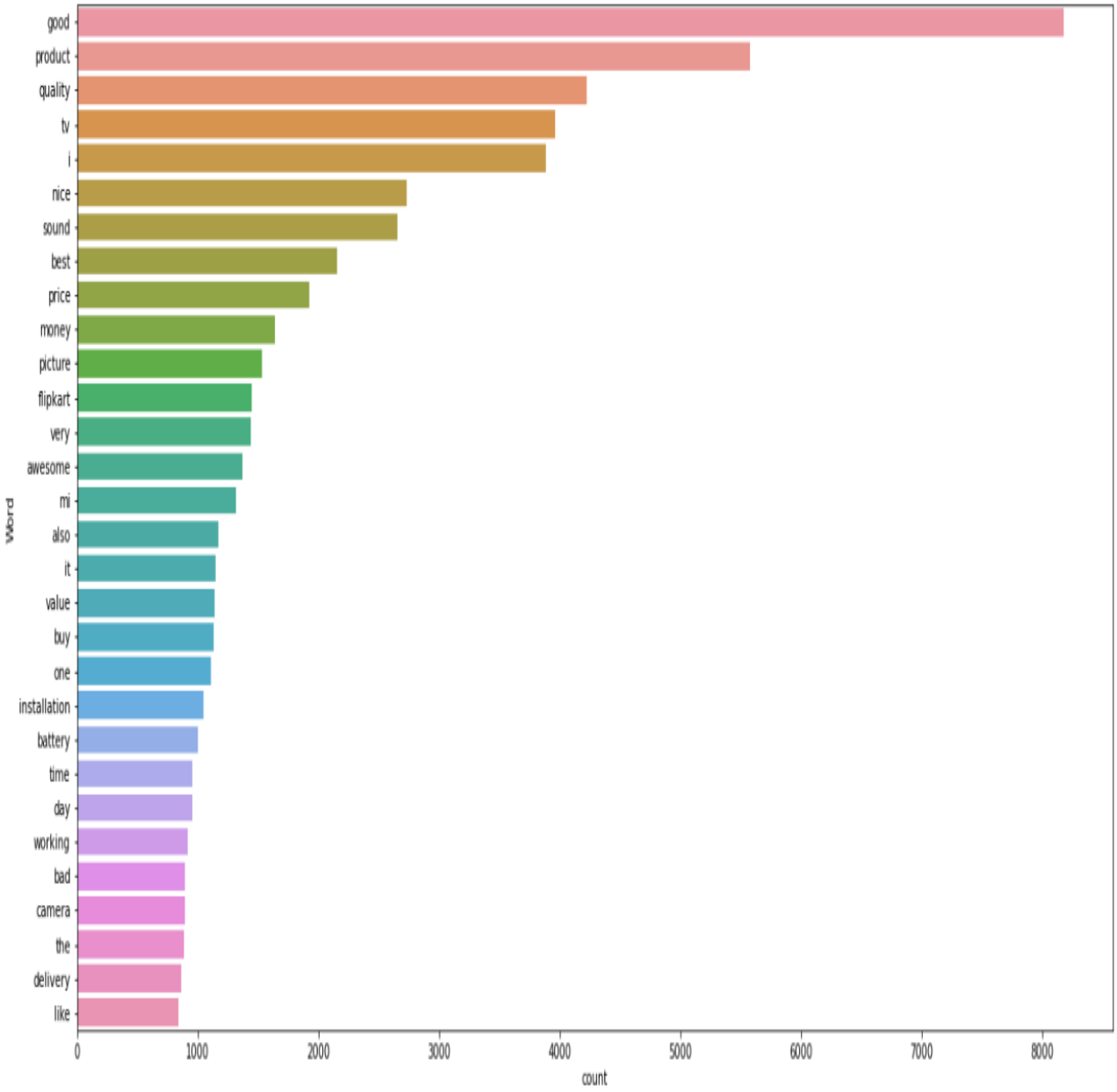
Character Count Density Plot



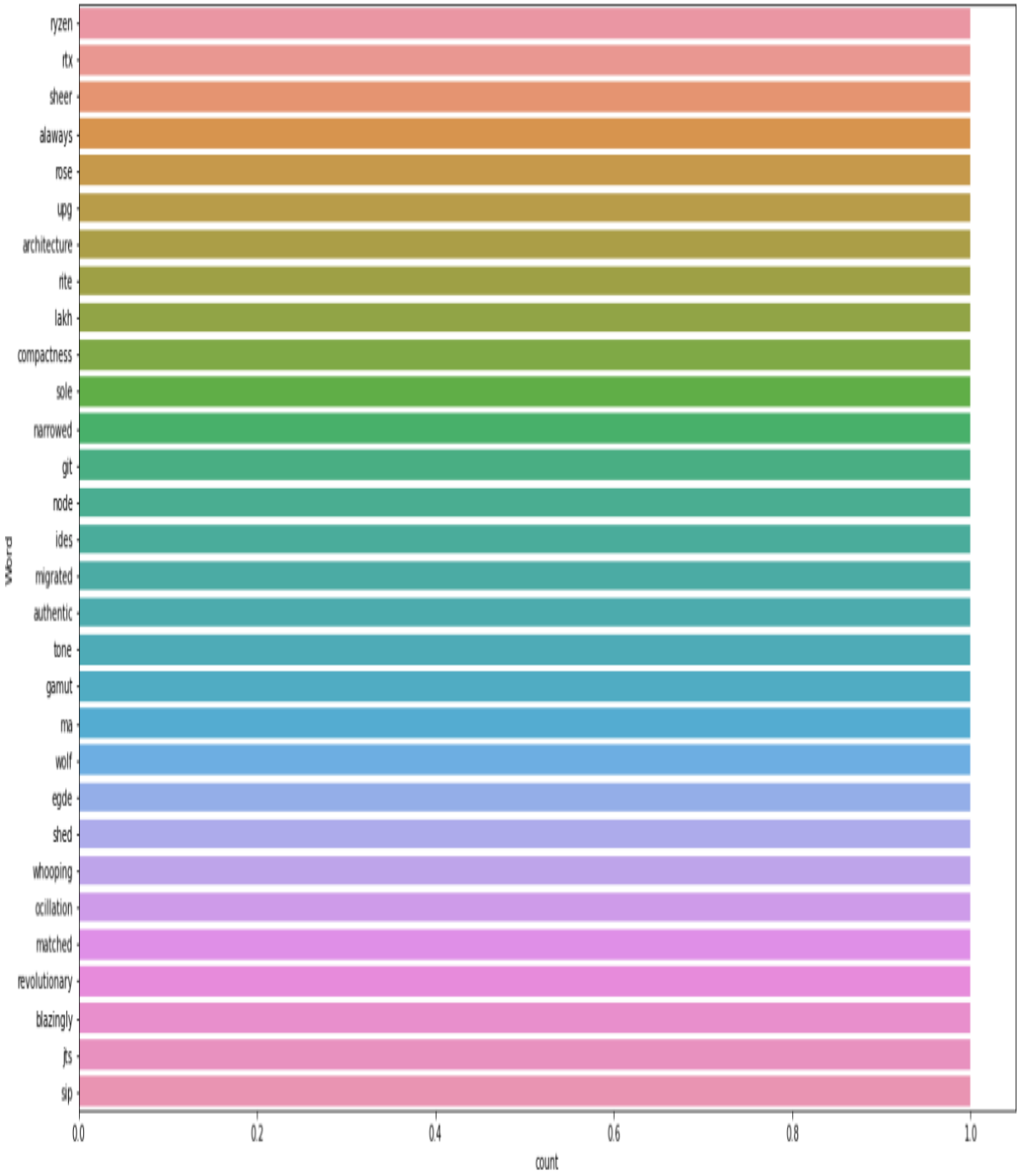
Average values of each text in reveiw



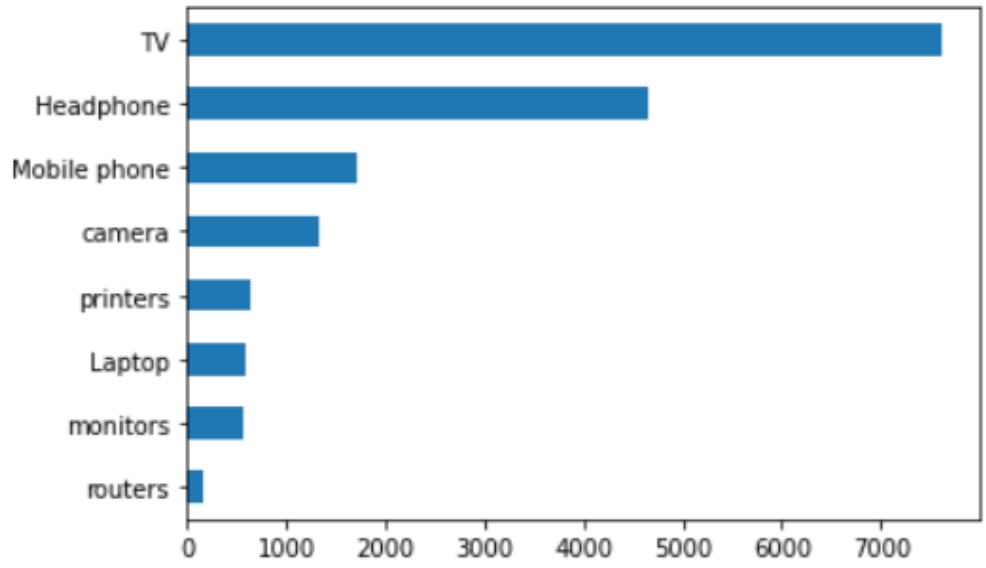
30 Most Frequent words



30 Least used words

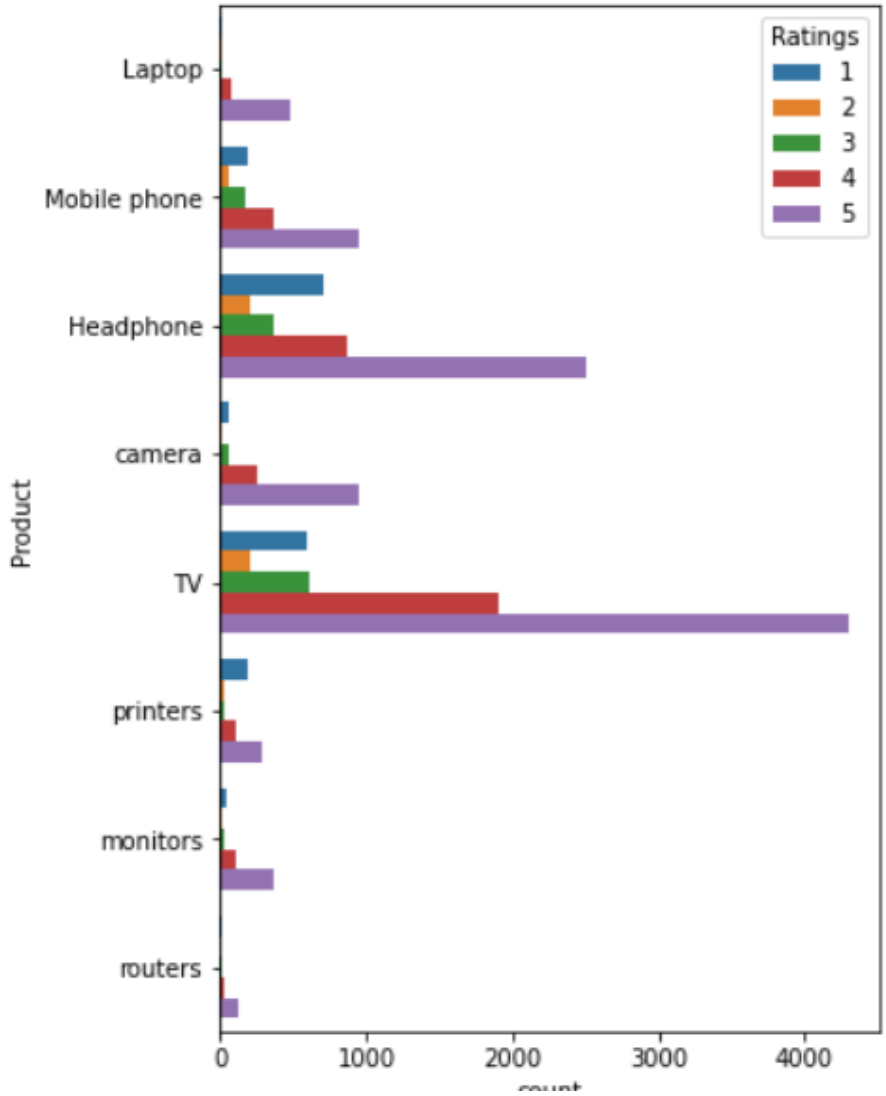


No of Products



The highest reviews are of TV and lowest is for routers

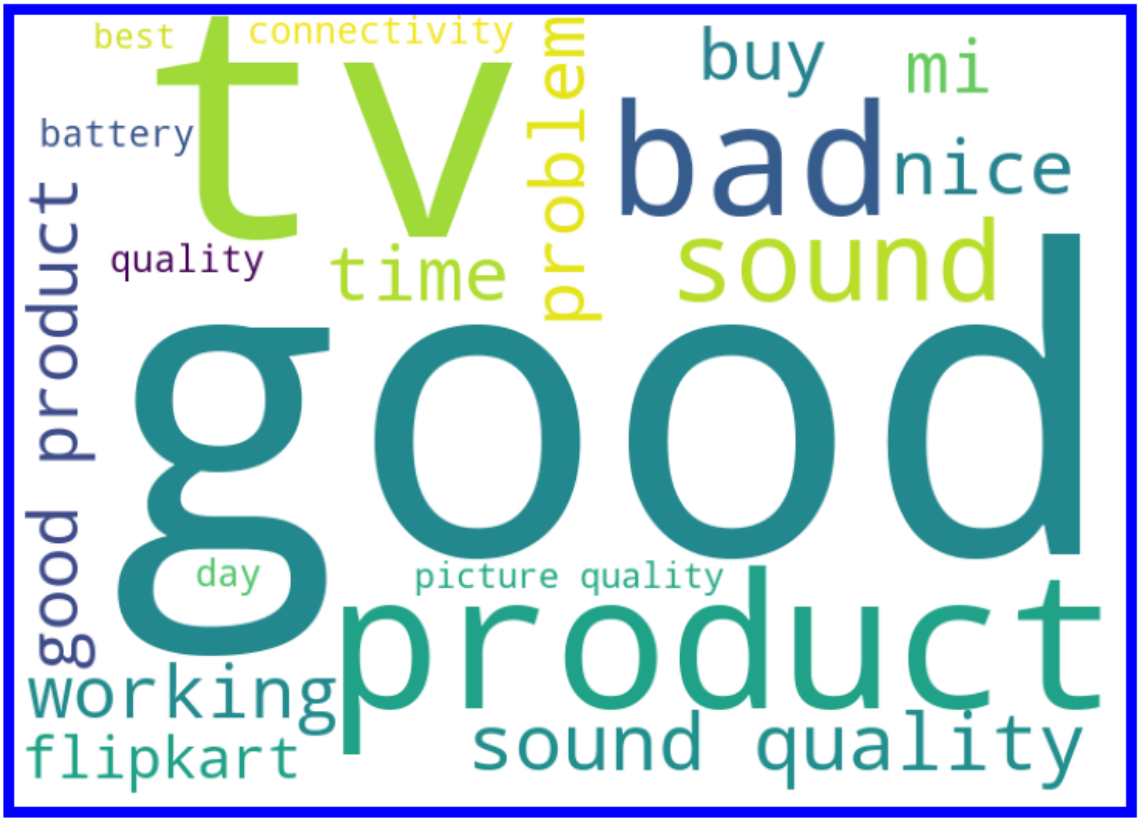
Products ratings in a single view



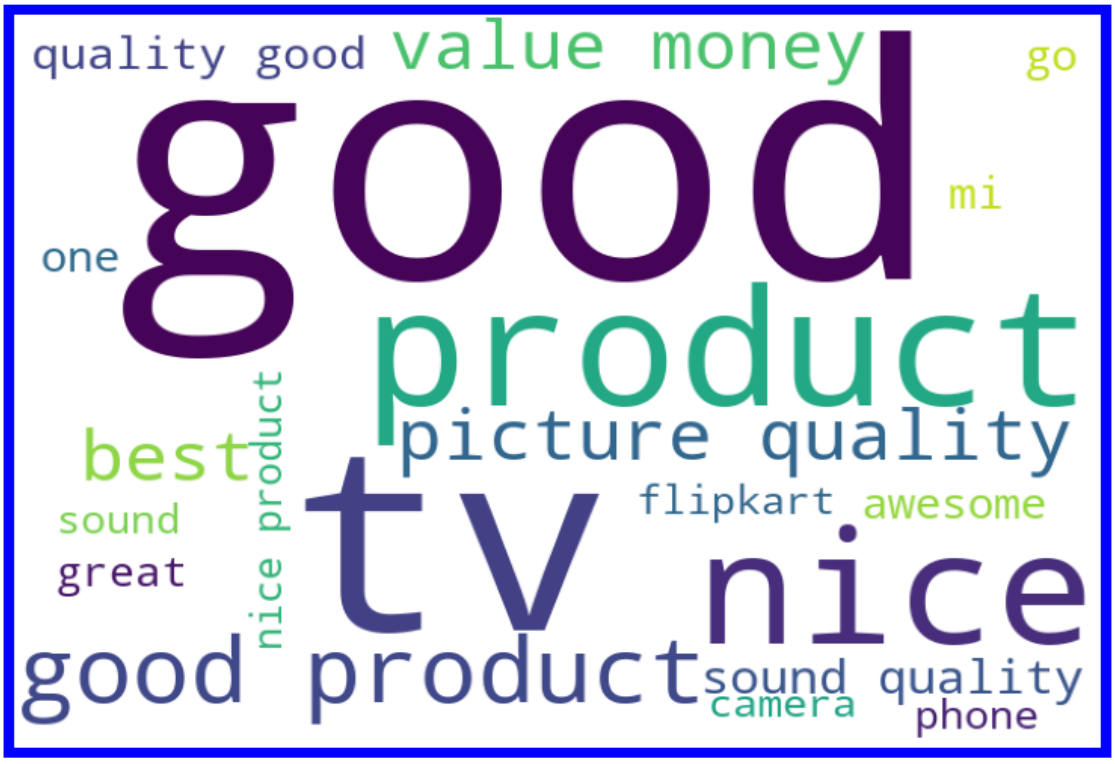
Processed Reviews Texts by wordcloud

Rating 1 Reviews

Rating 2 Reviews

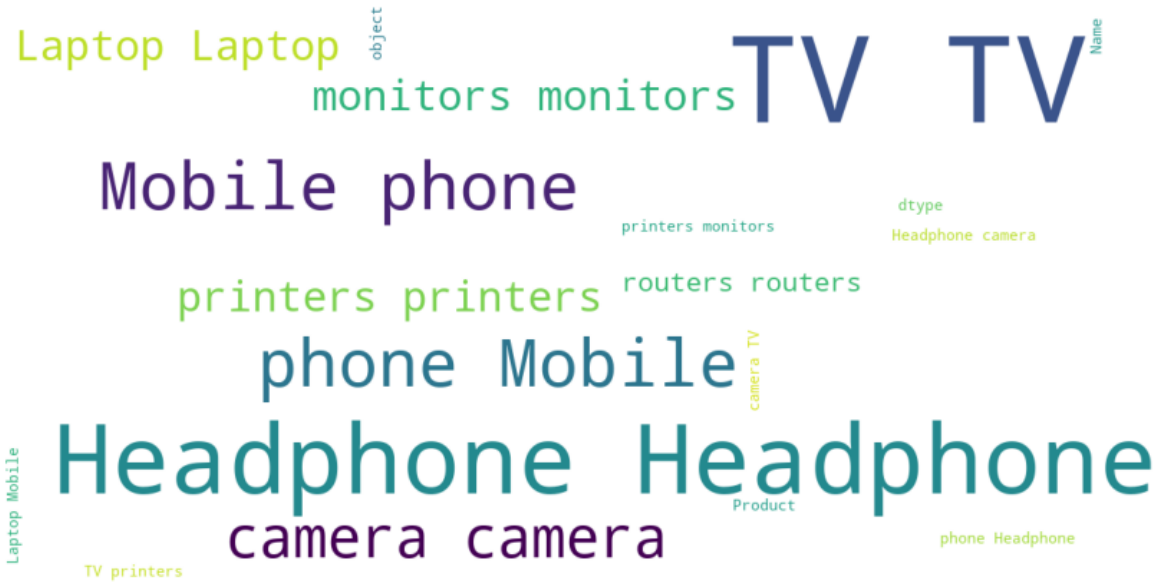


Ratings 3 Reviews



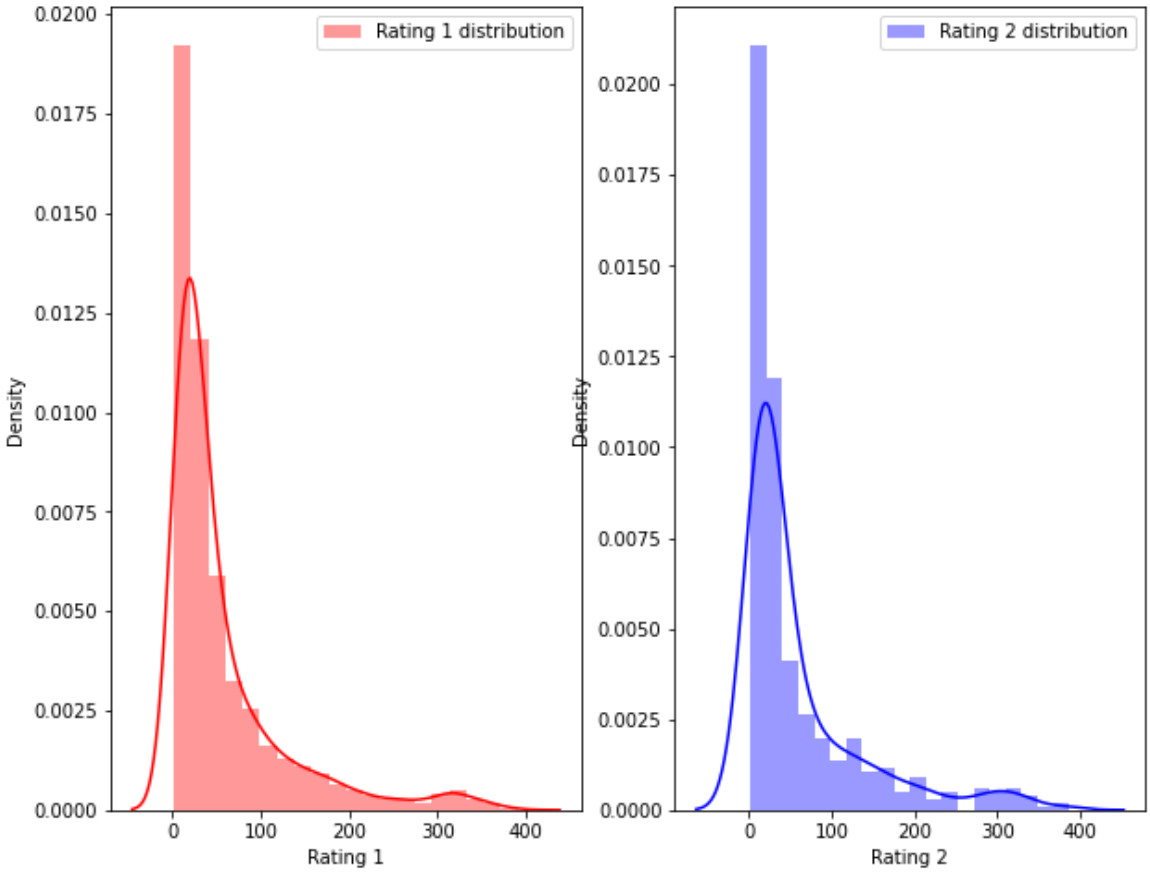
Ratings 5 Reviews

Product

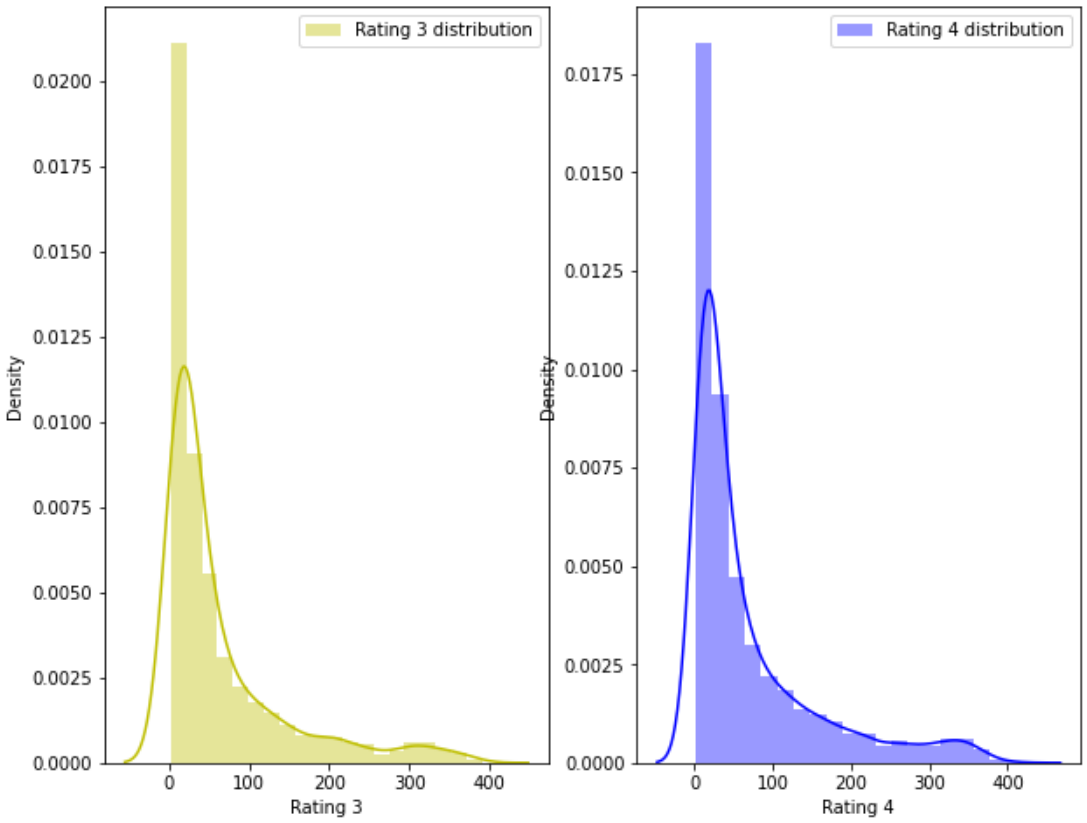


Ratings Distribution

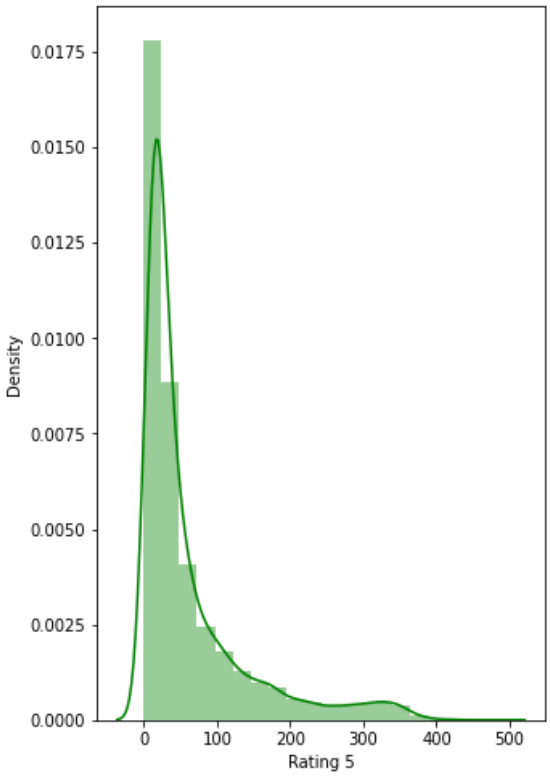
Ratings 1 and 2



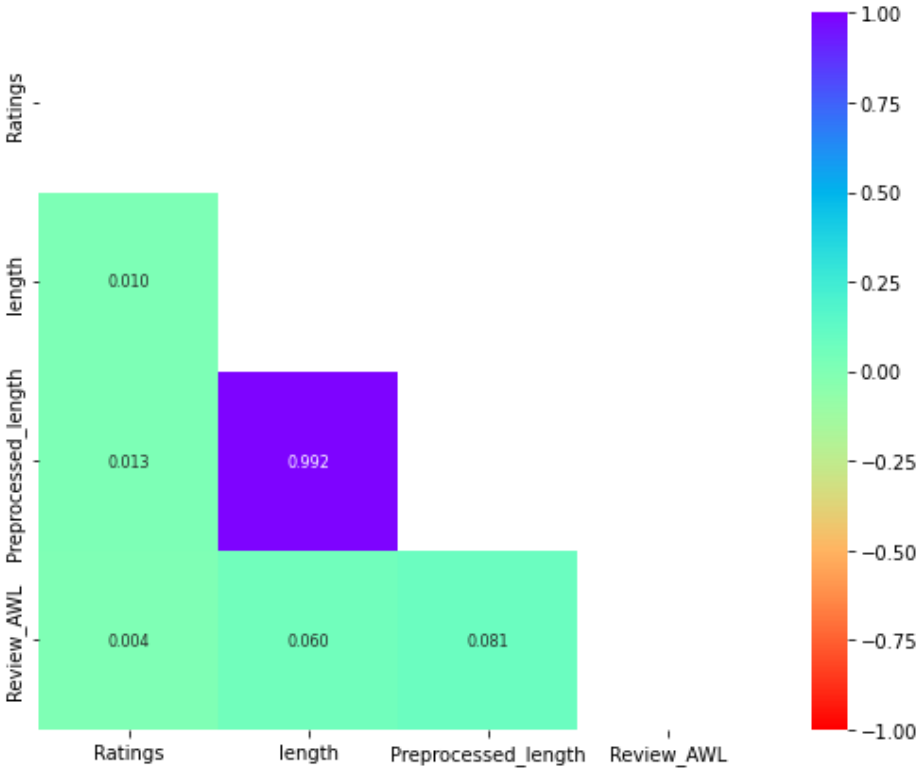
Ratings 3 and 4



Ratings 5



Correlation of the Dataset



Model Building and Predictions

LogisticRegression

```
LogisticRegression()
Training accuracy is :  0.8377105666156203
Testing accuracy is : 0.7436355515855293
--------------------------------------------------------------------------------
Classification Report:
               precision    recall  f1-score   support

           1       0.78      0.81      0.79      2178
           2       0.94      0.76      0.84      2762
           3       0.87      0.73      0.79      2633
           4       0.59      0.71      0.65      1839
           5       0.55      0.69      0.61      1783

    accuracy                           0.74     11195
   macro avg       0.74      0.74      0.74     11195
weighted avg       0.77      0.74      0.75     11195

Confusion Matrix:
 [[1762   26   52  140  198]
 [ 170 2106  163  161  162]
 [ 155   87 1917  217  257]
 [  62   17   42 1307  411]
 [ 116    7   39  388 1233]]
--------------------------------------------------------------------------------
Cross value score
cv score 0.7272152414296273 at 2 cross fold
cv score 0.7480385613196159 at 3 cross fold
cv score 0.7544160362281654 at 4 cross fold
--------------------------------------------------------------------------------
```

Multinomial NB

```
MultinomialNB()
Training accuracy is :  0.7058575803981624
Testing accuracy is : 0.6284948637784725
--------------------------------------------------------------------------------
Classification Report:
               precision    recall  f1-score   support

           1       0.59      0.73      0.66      1838
           2       0.89      0.63      0.74      3136
           3       0.73      0.59      0.66      2748
           4       0.44      0.70      0.54      1379
           5       0.49      0.53      0.51      2094

    accuracy                           0.63     11195
   macro avg       0.63      0.64      0.62     11195
weighted avg       0.67      0.63      0.64     11195

Confusion Matrix:
 [[1345   46  103  152  192]
 [ 317 1986  235  281  317]
 [ 260  132 1626  339  391]
 [  81   27   55  967  249]
 [ 262   52  194  474 1112]]
--------------------------------------------------------------------------------
Cross value score
cv score 0.6050925844771623 at 2 cross fold
cv score 0.6330712518888822 at 3 cross fold
cv score 0.6384037436886434 at 4 cross fold
--------------------------------------------------------------------------------
```

BernouliNB

```
BernoulliNB()
Training accuracy is :  0.5592266462480857
Testing accuracy is : 0.5212148280482358
--------------------------------------------------------------------------------
Classification Report:
               precision    recall  f1-score   support

           1       0.56      0.74      0.64      1714
           2       0.50      0.95      0.66      1195
           3       0.44      0.82      0.57      1186
           4       0.38      0.65      0.48      1289
           5       0.72      0.28      0.40      5811

    accuracy                           0.52     11195
   macro avg       0.52      0.69      0.55     11195
weighted avg       0.60      0.52      0.49     11195

Confusion Matrix:
 [[1265   87  108  119  135]
 [  18 1131    6   18   22]
 [  15    0  975   74  122]
 [  58    0   35  839  357]
 [ 909 1025 1089 1163 1625]]
--------------------------------------------------------------------------------
Cross value score
cv score 0.5069014797046475 at 2 cross fold
cv score 0.5171926507385952 at 3 cross fold
cv score 0.5203008961799894 at 4 cross fold
--------------------------------------------------------------------------------
```

SGD Classifier

```
SGDClassifier()
Training accuracy is :  0.8370980091883614
Testing accuracy is : 0.7358642251004913
--------------------------------------------------------------------------------
Classification Report:
               precision    recall  f1-score   support

           1       0.83      0.73      0.78      2552
           2       0.93      0.77      0.84      2716
           3       0.85      0.75      0.80      2500
           4       0.61      0.71      0.65      1897
           5       0.47      0.70      0.56      1530

    accuracy                           0.74     11195
   macro avg       0.74      0.73      0.73     11195
weighted avg       0.77      0.74      0.75     11195

Confusion Matrix:
 [[1875   84  134  174  285]
 [ 177 2079  117  158  185]
 [  90   62 1876  182  290]
 [  49   13   62 1340  433]
 [  74    5   24  359 1068]]
--------------------------------------------------------------------------------
Cross value score
cv score 0.7379076224792375 at 2 cross fold
cv score 0.7457334516350387 at 3 cross fold
cv score 0.7467780991539137 at 4 cross fold
--------------------------------------------------------------------------------
```

Model Building Results

Best models

- Logistics : Model shows low accuracy score in training and testing accuracy hence we cannot consider it.

- MultinomialNB: Model shows very much difference in training and testing accuracy hence we cannot consider it. Model becames underfit.

- Bernouli: Model shows low accuracy score in training and testing accuracy hence we cannot consider it.

- SGDClassifier : Model shows good results as it gives similar testing and CV score hence we are going to consider it.

Final Model SGDClassifier

```
the training accuracy is : 0.8370597243491578
the testing accuracy is : 0.7381866904868245
```

```
Classification Report:
               precision    recall  f1-score   support

           1       0.82      0.75      0.78      2496
           2       0.94      0.76      0.84      2769
           3       0.85      0.75      0.80      2515
           4       0.60      0.72      0.65      1861
           5       0.48      0.69      0.56      1554

    accuracy                           0.74     11195
   macro avg       0.74      0.73      0.73     11195
weighted avg       0.77      0.74      0.75     11195

Confusion Matrix:
 [[1862   92  122  166  254]
 [ 177 2107  130  157  198]
 [ 109   32 1886  191  297]
 [  36   12   46 1332  435]
 [  81    0   29  367 1077]]



Cross value score
cv score 0.7360853148154669
```

Model Deployment

Deploy the model

```
import pickle
filename = 'rating_project.pkl'                # model name
pickle.dump(final_model, open(filename, 'wb'))
```

Loading Model

```
load_model =  pickle.load(open('rating_project.pkl', 'rb'))    # loading deployed model
result = load_model.score(x_test, y_test)
print(result)
```

```
0.7381866904868245
```

```
original = np.array(y_test)
predicted = np.array(load_model.predict(x_test))
# convert columns in to np.array
```

```
print(predicted.shape)
print(original.shape)
print(x_test.shape)
print(y_test.shape)
```

```
(11195,)
(11195,)
(11195, 205548)
(11195,)
```

```
conclusion = pd.DataFrame({'Original ': original, 'Predicted': predicted}, index = range(len(original)))
# Dataframe creation
```

```
pd.set_option('display.max_rows', None)  # To maximize the rows
conclusion.sample(5)
```

	Original	Predicted
3570	3	3
10990	3	3
689	4	1
6257	1	1
5033	5	5

➢ Hardware and Software Requirements and Tools Used

Operating System: Window 11

RAM: 8 GB

Processor: i5 10th Generation

Software: Jupyter Notebook

Python Libraries: Mainly

Pandas: This library used for dataframe operations .

Numpy: This library gives statistical computation for smooth functioning .

Matplotlib: Used for visualization.

Seaborn: This library is also used for visualization.

Sklearn: This library having so many machine learning module and we can import them from this library.

Pickle: This is used for deploying the model.

CONCLUSION

➢ Key Findings and Conclusions of the Study

This project has built a model that can predict ratings of the customer reviews.vFor this company can find their ratings of products.

➢ Learning Outcomes of the Study in respect of Data Science

Data cleaning is the most important part in this model building as we see above there are so many stopwords, punkt and symbols values we remove it from the dataset dataset for better observations. This project has gives so much information about parameters that how a single parameter can hit customer ratings.

➢ Limitations of this work and Scope for Future Work

Model work with similar parameters as we build the whole model if some of the parameters missed then we need to train model with remains parameter after that we can predict upcoming ratings of customer reivews for client hence we need to up to date all the parameters as per training dataset.