RATING PREDICTION

Submitted by:

AVINASH PATEL

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher as well as my intern company FLIPROBO who gave me this golden opportunity to do this wonderful project on the topic RATING PREDICTION, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I want to mention some sites which helped me when I got stuck somewhere while completing my projects.

Those sites are :

http://scikit-learn.org

https://www.w3schools.com

https://www.youtube.com

https://www.kaggle.com

INTRODUCTION

- Business Problem Framing

  Our client does not provide us with data. So we have to scrape data from different websites and we have to scrape more than 27000 rows of data.

  So, In order to improve the prediction of RATING, the client wants some predictions that could help them in predicting the rating of pre existing ratings of the comments.

- Conceptual Background of the Domain Problem

  We only need the rating with there corresponding comments for predicting the rating of any comment.

- Review of Literature

  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.

- Motivation for the Problem Undertaken

  As I am an intern I want to work on as many project as possible for me. So ,I am highly motivated to do this project and to learn new thing and also learn from my mistakes.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  First of all I collected more than 27000 rows of data from amazon.com for my project. Then I removed some duplicate data and some null rows. Then I converted all the rows into lower case characters and applied some regex operation to clean the data and after the cleaning only 64% of the original data is left for processing that means 36% of the data is junk and we saved a lot of processing power and space.

- Data Sources and their formats

  As no Data is provided by our client, I scraped more than 27000 rows of data with 2 column. I scraped data from amazon.com

- Data Pre-processing Done

  For Stars column: This columns is all the stars from one to five for the corresponding reviews people given in amazon.
  For Comments column: This column stores all the comments that is reviews of people on different products.

- Data Inputs- Logic- Output Relationships

  The Remaining Data after all the pre-processing acts as an input data for the Model(in our case SVM) the already trained model gives us an output of an estimated Star rating.

- Hardware and Software Requirements and Tools Used

  Ram:8GB
  ROM:200MB
  Processor : Ryzen 5 3600X(6 cores 12 threads)
  Tool : Jupiter Notebook
  Language Used: Python

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  First I imported all the basic libraries of python used in the problem solving. Then I read carefully the data and the description given to me. After that I imported the provided csv file into the Jupiter notebook. Now first I checked for any nan values in the dataset after that I checked how many number of rows and columns I am dealing with and I also done some more analysis of the raw data to get a general idea and the removed some rows as some of them are duplicate and some of them are null.

  After analysing the data I started cleaning the data as explained above.

  After cleaning I did some visualisation of the data to get some insights from our cleaned data.

  On the completion of all these steps I moved on to the model selection part. Here I tested many models and choose one of the model to proceed.

  After I selected one of the models I did hyperparameter tuning and finally I saved my model.

- Testing of Identified Approaches (Algorithms)
  - Naivebayes MultinomialNB
  - Support Vector Machine

- Run and Evaluate selected models

  After running all the models the lowest difference between the accuracy score and the cross validation is for XGBOOST Regressor.

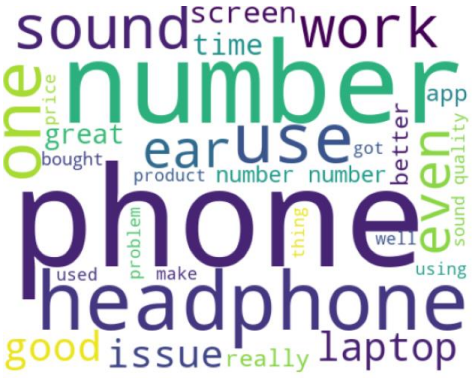
```
y_pred_mnb=MNB.predict(X_test)
print("Accuracy Score MNB:",accuracy_score(Y_test,y_pred_mnb))

y_pred_svm=SVM.predict(X_test)
print("Accuracy Score SVM:",accuracy_score(Y_test,y_pred_svm))
```

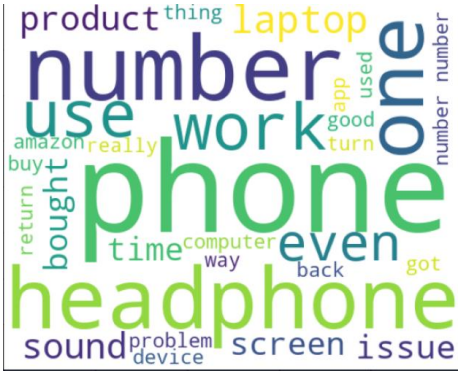
```
Accuracy Score MNB: 0.3423726619636269
Accuracy Score SVM: 0.3604297456475309
```

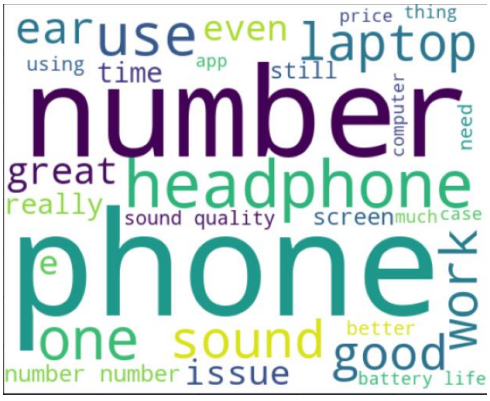
That means SVM is our best model.

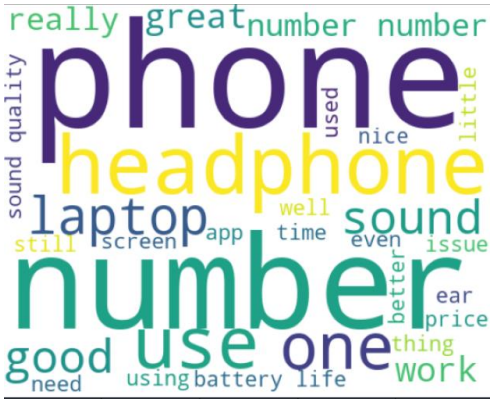
- Visualizations

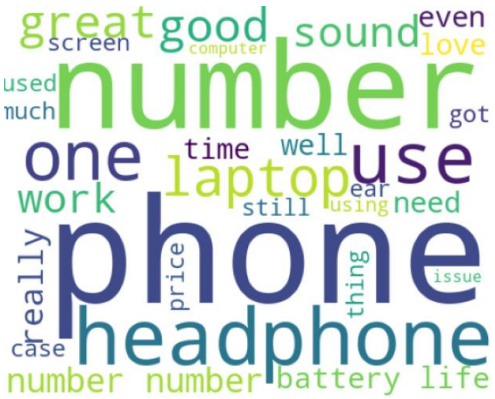
WordCloud:



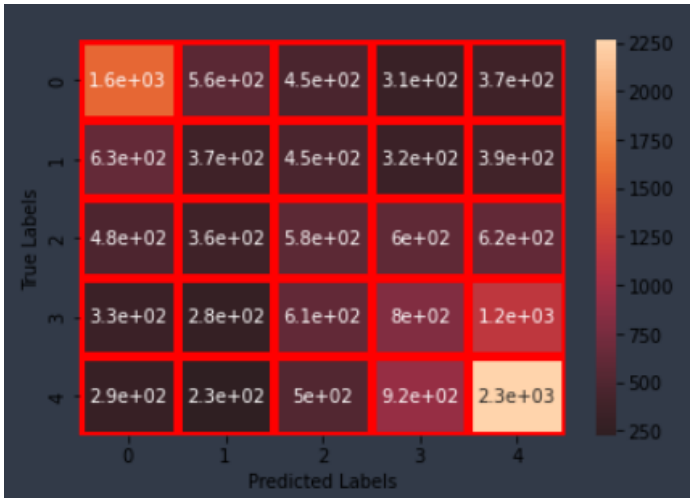








Heatmap OF Confusion Matrix:



- Interpretation of the Results

  After all the pre-processing and then visualising the data I find out that many words in comment column is useless and 34% of the data present in the comments column are useless.

CONCLUSION

- Key Findings and Conclusions of the Study

  After all the pre-processing and then visualising the data I find out that many words in comment column is useless and 34% of the data present in the comments column are useless and we need more data to accurately predict the stars column as we only get an accuracy of 40% after model building.

- Learning Outcomes of the Study in respect of Data Science

  As I have already told that this dataset has a lot of outliers and I have to look into each and every column to get some of the data cleaned. As per the visualisation part I plotted Word Cloud to see the Loud words in all the stars .

  I have also plotted Heatmap of the confusion matrix.

- Limitations of this work and Scope for Future Work

  I could get a better idea of how all the column are working and which outer columns needs to be in the dataset. As per the future scope, yes there is definitely more room to grow as I could get the model efficiency to 40% it could be more on further research.