

**Ratings Prediction**

Submitted by:

**Abhishek Jain**

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mrs. Khushboo Garg who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

**INTRODUCTION**

* Objective of problem

Review Rating Prediction attempts to infer from the review's content the user's numerical rating (often between 1 and 5 stars). Helping website visitors determine the rating of their reviews is a good usage of the Rating Prediction job. Recognize phoney or dubious online reviews.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data Preliminary data analysis must be performed to gain a deeper understanding of the quality of the data, in terms of outliers and the skewedness of the figures, descriptive statistics, and other factors. Understanding and preparation are essential parts of building a model because they provide insight into the data and what corrections or modifications shall be made before designing and executing the model. To do that, category and numerical variables were statistically analysed. Additionally, it helps to be aware of the key factors that influence how prices are determined. This was accomplished by creating a correlation matrix for each attribute to comprehend the relationships between the various components.

* Data Sources and their formats

The project deals with Indian ecommerce website. Using Selenium, the dataset from flipkart.com and amazon was scraped in order to build the effective intelligent model.



3 Features have been scrapped.

1. Title
2. Review
3. Rating

* Hardware and Software Requirements and Tools Used

Hardware:

Software: Latest Anaconda for Jupyter

Python Libraries:

Pandas , Numpy, seaborn, matplotlib, scikit-learn,

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches

Used NLP for text pre-processing as our data contains emojis,numerics,spaces etc.

* Testing of Identified Approaches (Algorithms)

1. Remove all email addresses
2. Remove all website links if any
3. Capture emojis
4. Remove all special character
5. Convert into lower case
6. Stemmer/Lemmitizer to convert into base word

Then use machine learning tools to find out the best model.

* Run and Evaluate selected models

Logistic Regression

Multinomial NB

Decision Tree Classifier

SVC

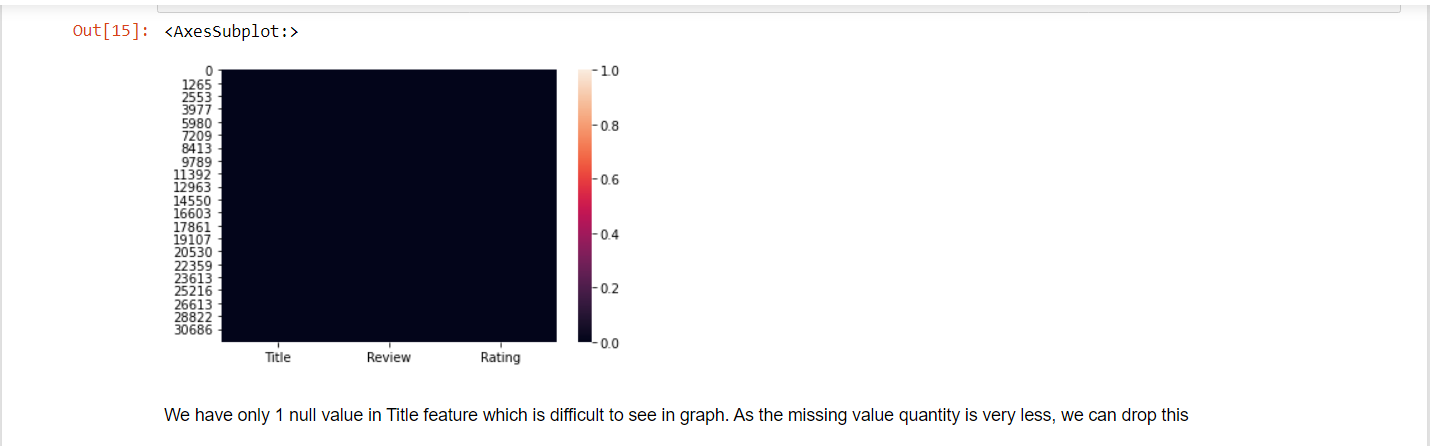
* Key Metrics for success in solving problem under consideration

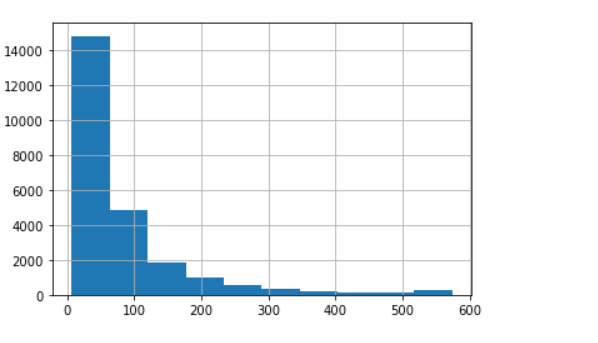
All the data is first converted into string and then data processing is done on it.

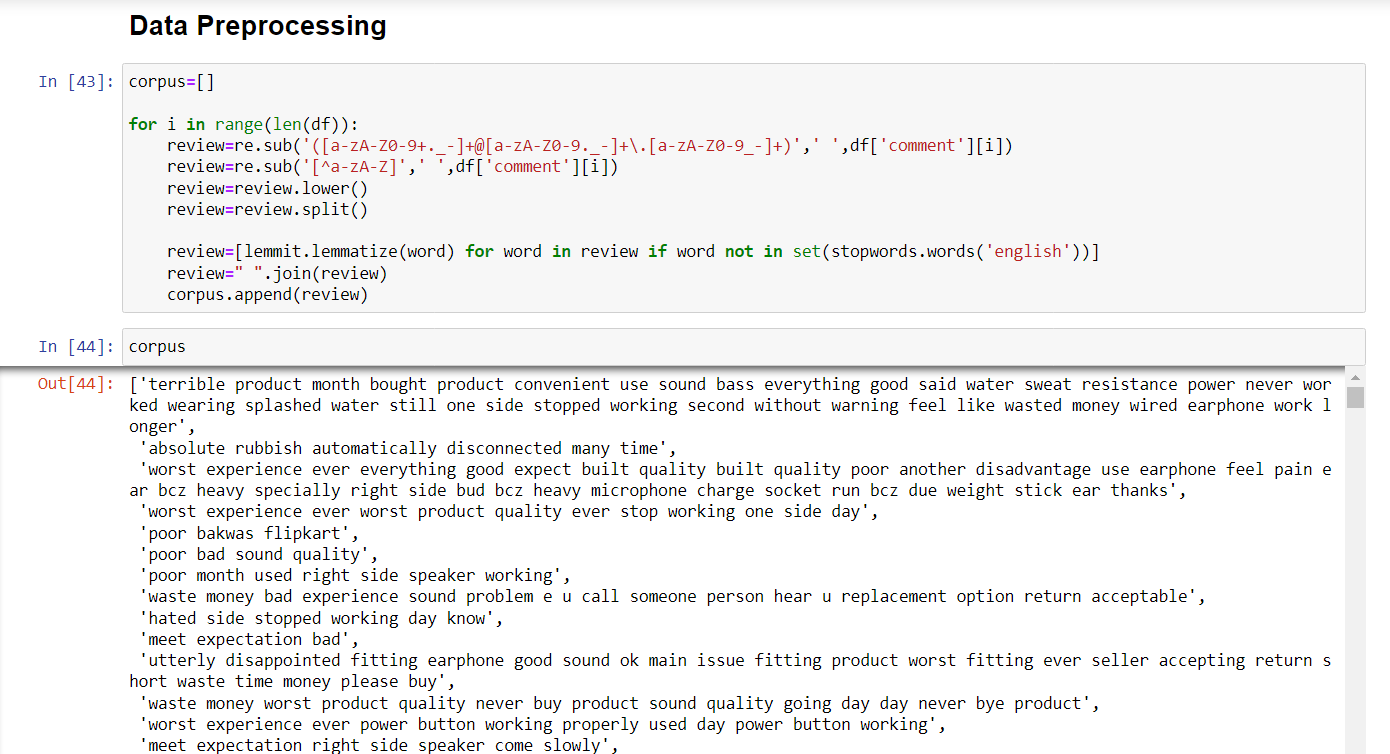
Since the dataset we obtained is irregular , having unequal number of similar data according to the ratings(1,2,3,4,5) so we have to use SMOTE technique to balance data.

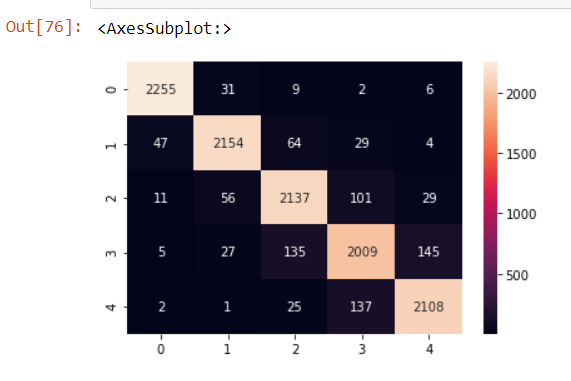
**Visualizations**

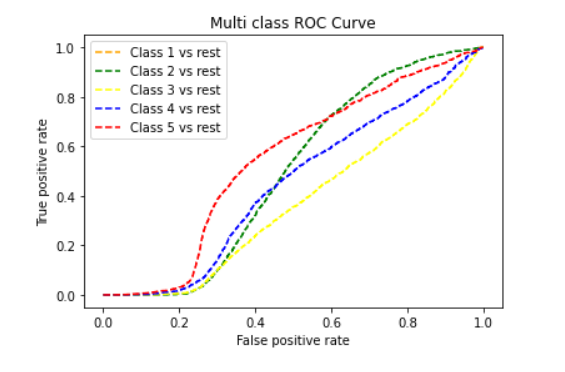












**Conclusion**

* Key Findings and Conclusions of the Study

I used only two websites to scrap data which could be takenas more.

Data may not be properly preprossed even though applying all the necessary alogithms.

Knowledge of NLP algorithms is must before working on this dataset.