

RATING PREDICTION PROJECT

Submitted by:

SURJEET SINGH

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to **FlipRobo** who gave me the golden opportunity to do this wonderful project on the topic **Rating Prediction** project. The data used in this project was scraped from the website of **Amazon**.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain Problem

The problem we are going to solve is a classification problem. As we have five different classes of prediction. But since the rating is scaled from 1 to 5 we can also solve this problem as regression problem.

* Motivation for the Problem Undertaken

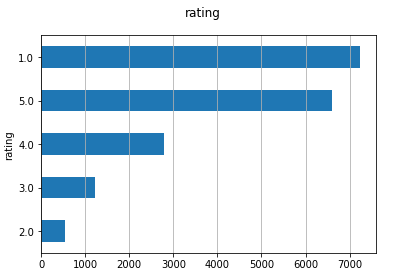
This machine learning model will be helpful for the clients to predict the rating of the previous review given to the company but has no rating. Company will be able to analyze those review as well.

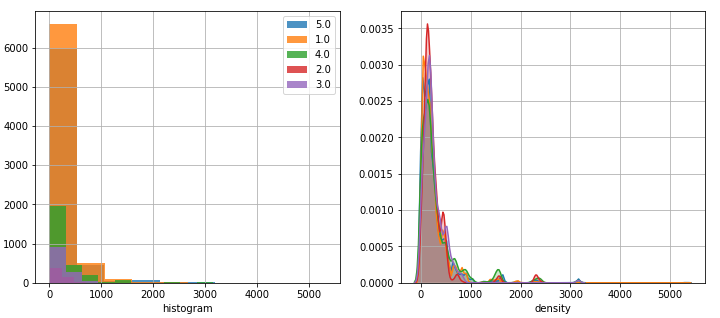
**Analytical Problem Framing**

* Data Sources and their formats

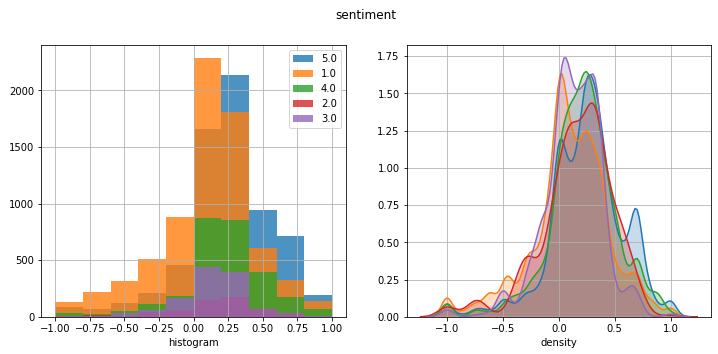
Data used in this project have been scraped from the website of Amazon. The data has three columns, review, main review and rating. The review column has just summary of the review, main review has complete text of the review and rating column has the rating given by the user along with the review.

* Data Preprocessing Done

Since the data we have scraped is a text data, we have to convert that data into proper format so that machine learning algorithms can accept the data. The graph shown in the right is the count plot of each rating. The data is imbalanced and we can use undersampling or oversampling to balance the data. Since user can give review in any of the language known by user so we have to detect the language of the each review. It is found that the language of each review is English. We have count the characters, words, sentences of the review and then we have analyzed the data.



The above graph belongs to character counts of the review. It is clear from the plot that most of the review has character length between 0 and 1000. Some of the review has very large number of character which we can filter as outliers. We have removed all the stopwords because stopwords do not have a strong impact on the prediction of the rating of the review. Then we have added a new column with the sentiment. In this column we have predicted the polarity of the review with the help of a pretrained model TextBlob.



It is clear from the graph that the reviews with low rating has negative sentiment and the reviews with high rating has positive sentiment. It is also important for us to analyze the words that are important in prediction. We can plot a word cloud to find the most frequently used words in the data.



Since the machine learning models can’t accept the data in the text format, we need to convert the given data into numbers. We have vectorize the text using tfidf vectorizer.

* Data Inputs- Logic- Output Relationships

The words present in the review having good sentiment will increase the rating while the words having negative sentiment will reduce the rating in the prediction. For example, words like “good”, “best”, “value for money”, “fantastic” will increase the rating while the words such as “bad”, “worst”, “don’t buy” will decrease the rating.

* Hardware and Software Requirements and Tools Used
* All the codes are written in python 3. Jupyter notebook used to create all visualizations, training, and testing of the machine learning models. Device used in this project : hp laptop with intel i5 process and 4 GB RAM.

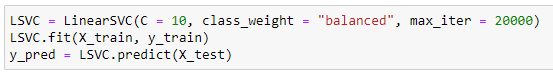
**Model/s Development and Evaluation**

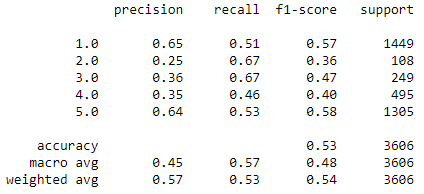
* Testing of Identified Approaches (Algorithms)

Since it is a problem which can be solved by both, regression as well as classification. We will try both type of models and will select the model with the best accuracy.

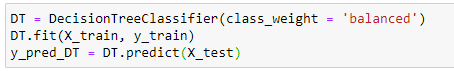
* Run and Evaluate selected models

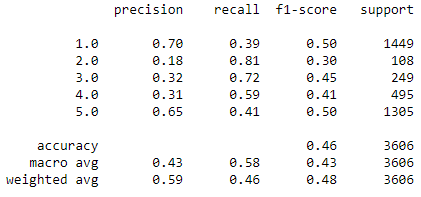
Linear SVC



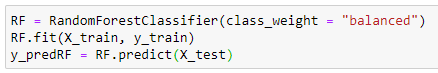


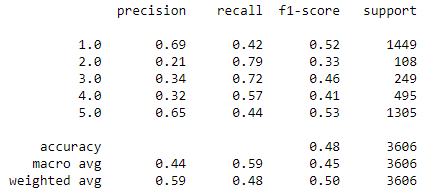
Decision Tree Classifier



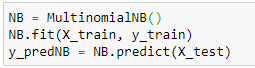


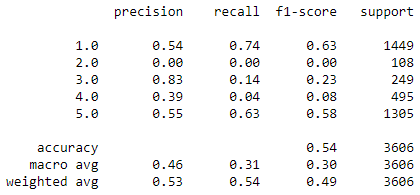
Random Forest Classifier





Naïve Bayes





**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

While working in this project, I have learned how to handle the text data, stemming, stopwords, vectorization of the text data.

* Limitations of this work and Scope for Future Work

We have applied several machine learning algorithms but it was hard to get accuracy beyond 54%. We have also tried to fine tune some models but no significant improvement noticed. In future work, I will try to apply deep learning models so that we can find some hidden pattern within the text data which simple machine learning failed to detect.