

Ratings Prediction

Submitted by:

MAHESH AGASHIMANI

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledgement the allowance by Data Trained Education and permission by Flip Robo. I extend whole hearted thanks to them I worked and learned a lot and sharing me the knowledge and experience.

Data Trained Education and Flip Robo provided training is the very important to completion of project.

**INTRODUCTION**

* Business Problem Framing

It describes the application of a range of supervised and unsupervised machine learning models to a dataset of Amazon product reviews in an effort to predict rating value. The desired output to an input of a text review is a “star” rating on a continuum from 1 to 5. Because we wish to produce a continuum of sentiment rather than just polarity, this project could be categorized as a regression variation of sentiment analysis.

* Conceptual Background of the Domain Problem

Our goal is to produce the most versatile and accurate model that handles the widest range of mixed and polarized sentiments expressed in reviews.

* Review of Literature

The article will first describe the curation of the dataset utilized to train and test the models, describing all review selection methodologies, pre-processing steps, and feature engineering.

* Motivation for the Problem Undertaken

It concludes with overall results, major findings, and recommendations for improvement and future work.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This dataset contains information about Rating Prediction listed on train dataset. This data can be used for a lot of purposes such as Rating prediction to exemplify the simpler encoding options like Bag of Words and TF-IDF in Machine Learning.

* Data Sources and their formats

The columns in the given dataset is as follows:..

The data set contains the training set, which has approximately 32,352 samples.

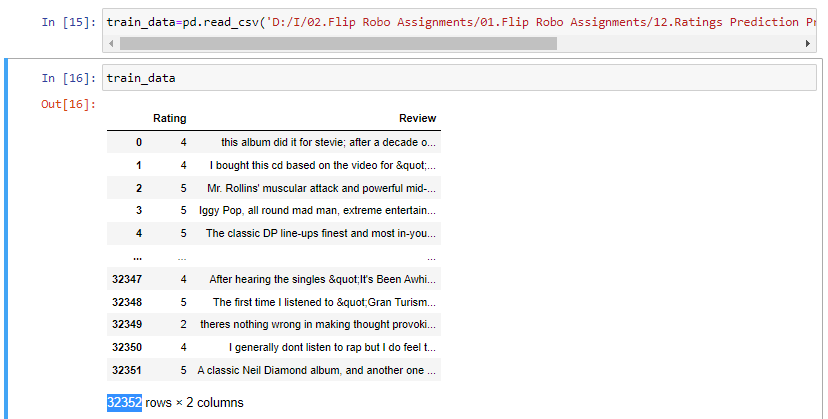
The data set includes:

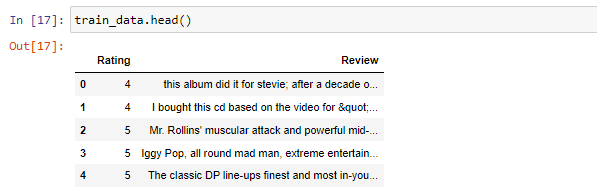
* **Ratings:** It contains the rating of amazon data.
* **Review:** It contains the review of amazon data.

Importing the Libraries:

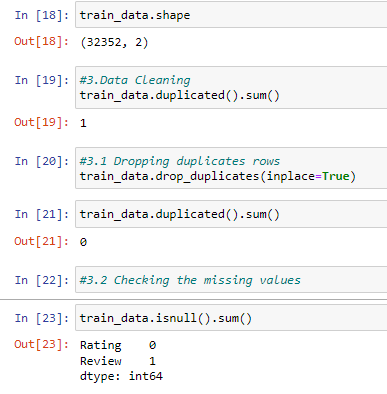


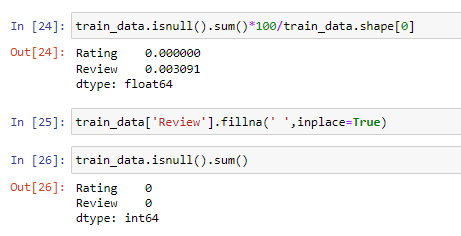
. Loading the Dataset:





After that we are checking if Null values are present in datasets.





In checked above data sets, there is Null values present, so removed null values.

* Hardware and Software Requirements and Tools Used

The needed time to train the model depends on the capability of the used system during the experiment. Some libraries use GPU resources over the CPU to take a shorter time to train a model.

|  |  |
| --- | --- |
| Operating System | Windows 10 |
| Processor | Core i7 |
| RAM | 16GB |
| Graphics card | 1080 TI OC |

Also we are using Jupiter notebook for running the code

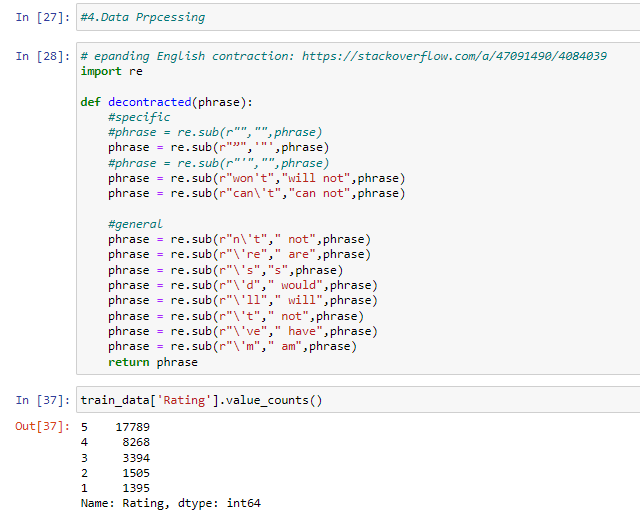
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

consider reviews written by verified purchasers to decrease the risk of fraudulent reviews with dubious ratings. Only star\_rating, review\_headline, and review\_body columns were considered to reduce feature complexity.

# Data Preprocessing

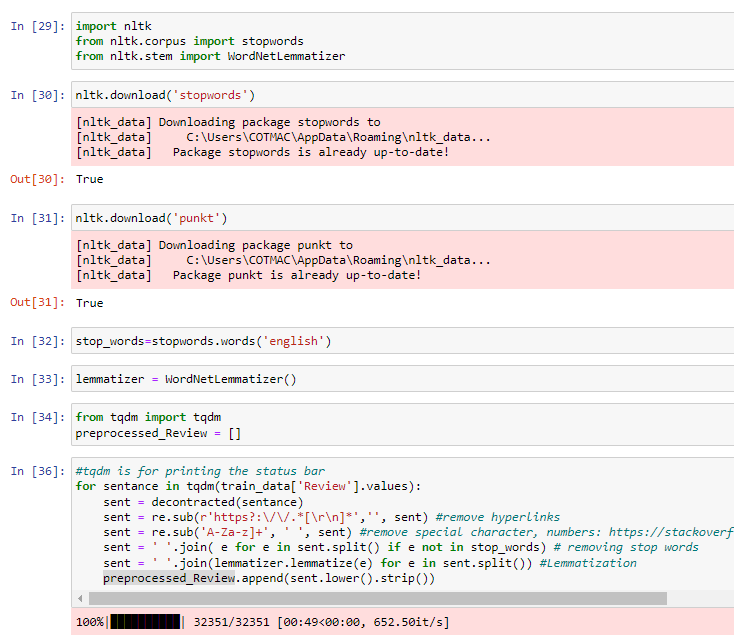
Text preprocessing is an important step for any natural language processing task. It transforms text from its raw format into something that is more processable for computer algorithms.



As above image we encoded the data processing.

# Removing Stopwords

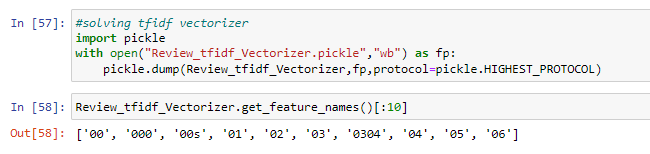
Stop words are common words that structure a sentence. Words such as “I”, “are”, and “here” do not contribute to the sentiment — the rating in our case — of reviews. Hence, we decided to remove stopwords to further denoise the input. We used NLTK’s stopwords package to provide us with the list of stopwords. Here, we made an adjustment to avoid the removal of certain negation stop words, namely“not” and “no”, since they do indeed influence sentence meaning. A product that is “not good” is certainly different from a “good” product.



* Testing of Identified Approaches (Algorithms)

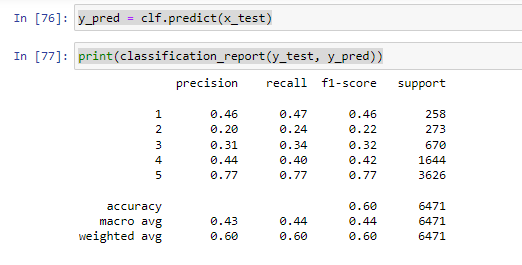
We chose to consider only a small subset of the dataset, as the amazon dataset reviews. We performed an 80:30 train/test split uniformly sampled reviews based on rating, ensuring a stratified split to preserve a balanced dataset.

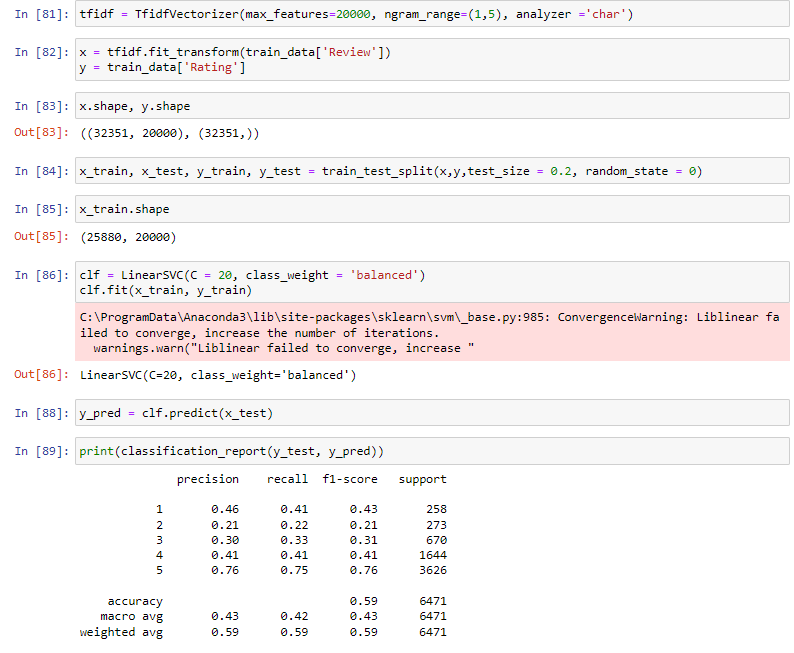


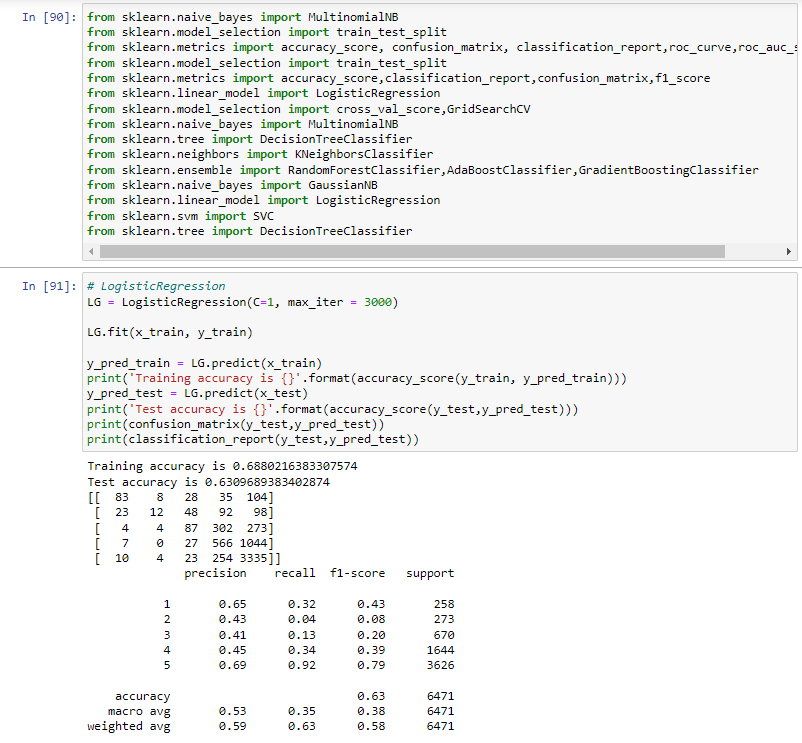


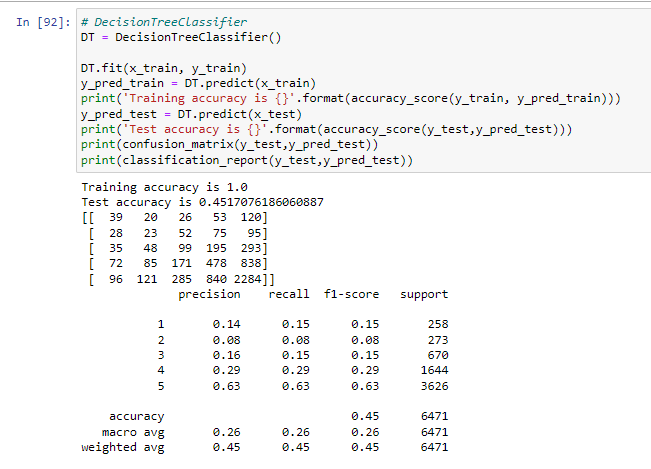
Using SVM

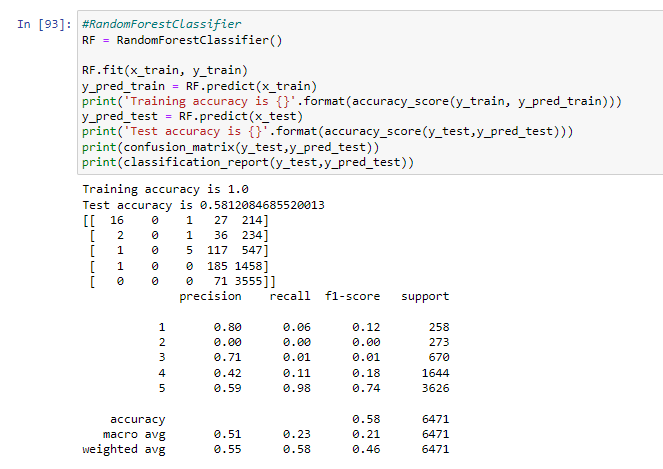


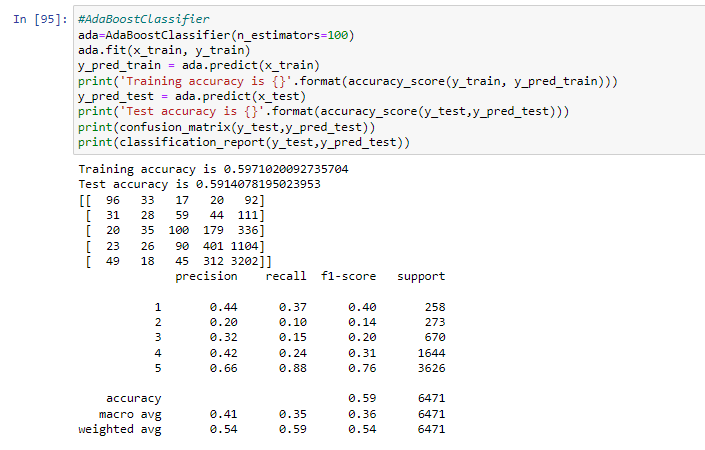


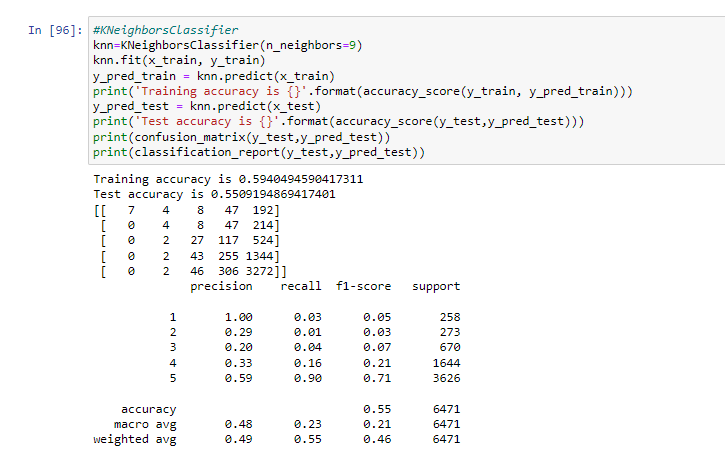








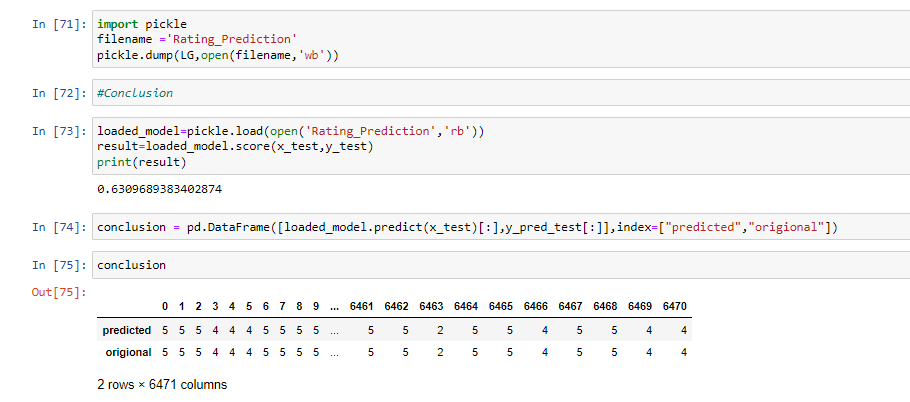




**CONCLUSION**

* Key Findings and Conclusions of the Study

As compare multiple model Logistic Regression is better accuracy 0.63.



* Learning Outcomes of the Study in respect of Data Science

As we checked in dataset present null values & the removed null values from dataset.

Pre-processing is done.

After all this split the data into train & test split.

* Limitations of this work and Scope for Future Work

The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combination of pre-processing methods to achieve better prediction accuracy.

Make available features in combining the features and predict it may be improved performance.