

Rating Prediction

Submitted by:

Aamir Rahman

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to FlipRobo, who gave me the golden opportunity to do this wonderful project of Rating Prediction.

Secondly, I would also like to thank DataTrained teachers and mentors who teach me all the basic and professional concept for building the project.

**INTRODUCTION**

* Business Problem Framing

The rise in E-commerce, has brought a significant rise in the importance of customer reviews. There are hundreds of review sites online and massive amounts of revies for every product. Customers have changed their way of shopping and according to a recent survey, 70 percent of customers say that they use rating filters to filter out low rated items in their searches.

The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, companies like Google, Amazon, Yelp and etc.

* Conceptual Background of the Domain Problem

This is rating prediction problem, where we have to develop a model which can predict the rating just by looking at the text of the review. So there is not much domain knowledge needed in this project. He just has to be aware of different Natural Language Processing concept.

* Review of Literature

For me this is my first project of Natural Language Processing. I read many articles on the internet to build an effective model which can predict the rating.

This project is fall under the sentiment analysis where the machine have to learn the sentiment of the review text in order to give the accurate rating.

* Motivation for the Problem Undertaken

Desire to face the challenge in solving the more challenging problems. And also the objective behind taking this project is to improve my skill in statistical as well as in analytical knowledge in machine learning and artificial intelligence.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

These are the various mathematical/ analytical model carried out in this project: -

* **Removing the punctuation: -** We can remove the punctuation from the text by using ‘regular expression’ or just by using the ‘string’ library.
* **Calculating polarity: -** Checking the polarity of the review is most important part in order to analyse the sentient of the text. We can calculate the polarity by using the function ‘polarity’ present in the library textblob.
* **Removing stopwords: -** Stopwords are a set of commonly used words in language, like ‘a’, ‘the’, ‘is’, ‘are’ and etc. Stopwords are commonly used in Natural Language Processing to eliminate words that are so commonly used that they carry very little useful information. Stopwords can be removed by the function ‘stopwords’ present in the library Natural Language Toolkit (nltk).
* Data Sources and their formats

The data source of this is data scraped from different e-commerce websites like amazon, flipkart, myntra. All data then merged together to form a single dataset containing all the review and rating.

* Data Preprocessing Done

In the given data there are some ambiguities present. In order to develop a machine learning model, data pre-processing is the most important step to build an effective machine learning model.

These are some data pre-processing which is done on our data:

* **Null values: -** This is the first thing that we should check in our data. Null values can be treated in two ways either replace it or remove it. But in case of rating prediction we cannot assign randomly any value in place of null value, so it is better to remove that null value.
* **Calculating polarity: -** Checking the polarity of the review is most important part in order to analyse the sentient of the text. We can calculate the polarity by using the function ‘polarity’ present in the library textblob.
* **Removing stopwords: -** Stopwords are a set of commonly used words in language, like ‘a’, ‘the’, ‘is’, ‘are’ and etc. Stopwords are commonly used in Natural Language Processing to eliminate words that are so commonly used that they carry very little useful information. Stopwords can be removed by the function ‘stopwords’ present in the library Natural Language Toolkit (nltk).
* **Removing the punctuation: -** We can remove the punctuation from the text by using ‘regular expression’ or just by using the ‘string’ library.
* Data Inputs- Logic- Output Relationships

In natural language processing the relation between input and output can be defined by using the sentiment analysis. The sentiments are mainly of 3 types positive, neutral, negative. The value of the sentiment can be calculated by using the polarity function present in textblob. It gives the numerical value ranges from -1 to 1. If the value tends towards negative side then the sentiment is negative, if the value tends towards positive side then the sentiment is positive and if the value is 0 then the sentiment is neutral.

* Hardware and Software Requirements and Tools Used

List of tools used in the project:

1. Scientific computing libraries: Pandas, Numpy.
2. NLP Library used: textblob, countvectorizer.
3. Visualization libraries: Matplotlib, Seaborn.
4. Algorithmic libraries: Scikit learn.
5. Programming language: Python.
6. IDE used: Jupyter notebook.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The dataset contains 2 columns. My first approach after data pre-processing data anomaly like: -

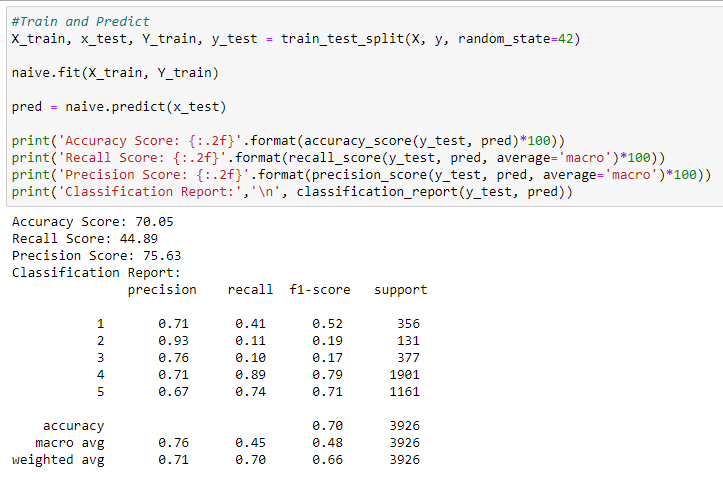
1. Removing null values.
2. Removing punctuation.
3. Calculating polarity.
4. Checking word frequency.

* Testing of Identified Approaches (Algorithms)

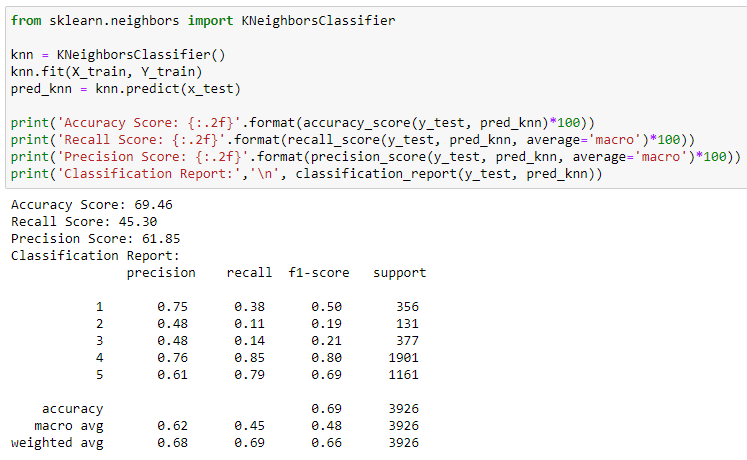
1. **MultinominalNB: -** The Multinomial Naïve Bayes classifier with discrete features (e.g., word counts for text classification). The Multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.
2. **KNearest Neighbours: -** It classifies a sample based on the category of its nearest neighbours. When large samples are involved, it can be shown that this rule has probability of error which is less than twice the optimum error-hence there is less than twice the probability of error compared to any other decision rule.
3. **Decision Tree: -** Decision tree are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as piecewise constant approximation.
4. **Random Forest: -** A random forest is a meta estimator that fits a number of decision tree classifier on various sub-sample of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size in controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.
5. **AdaBoost:** **-** An adaboost classifier is a meta-estimator that begins by fitting a classifier on the original dataset then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifier focus more on difficult cases.

* Run and Evaluate selected models

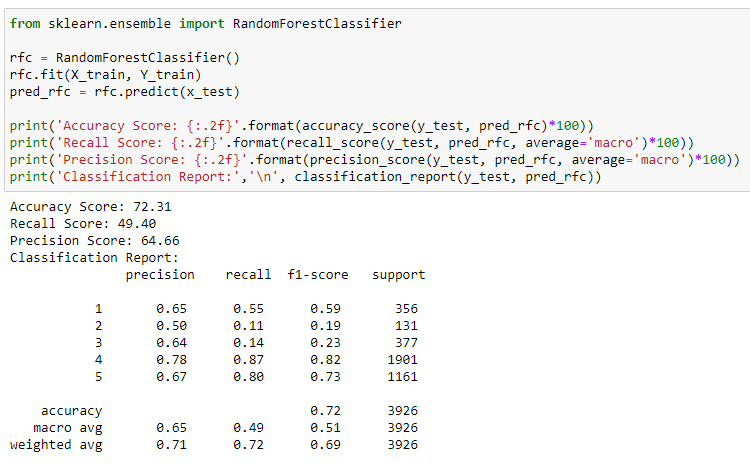
1. **Multinomial Naïve Bayes: -**



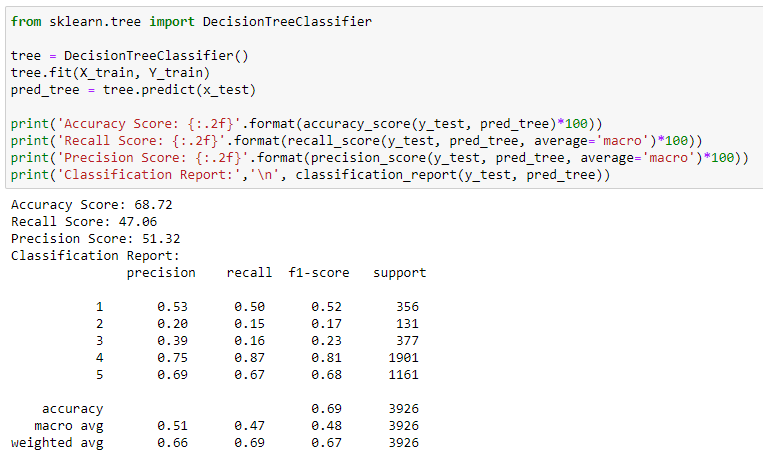
1. **K-Nearest Neighbours: -**



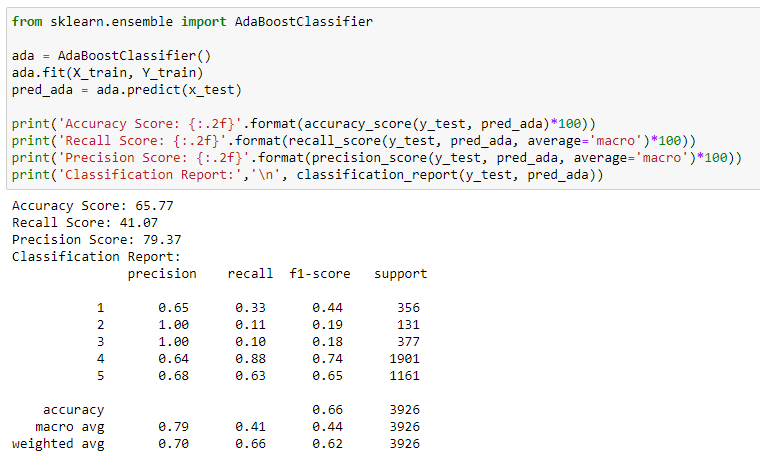
1. **Random Forest: -**



1. **Decision Tree: -**



1. **AdaBoost: -**



* Key Metrics for success in solving problem under consideration

This project is classification problem. These are the metrics used in project: -

1. **Accuracy Score: -** Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

**Accuracy: -** TP+TN/TP+FP+FN+TN

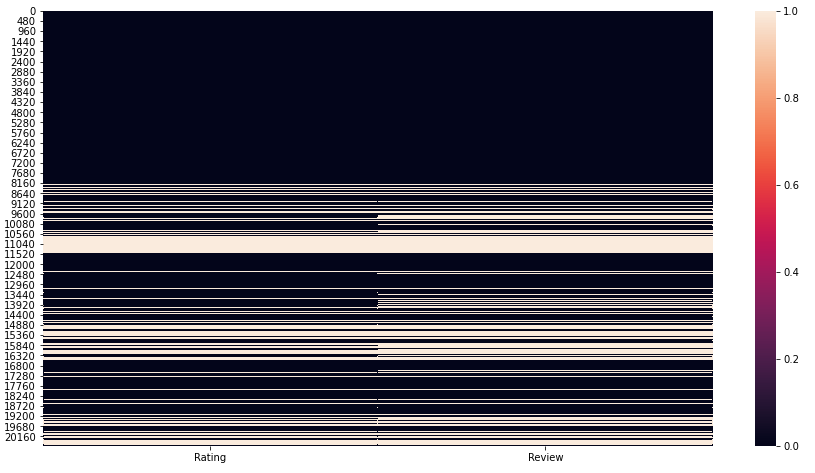
1. **Precision Score: -** Precision is the ratio of correctly predicted positive observations to the total predicted positive observation.

**Precision: -** TP/TP+FP

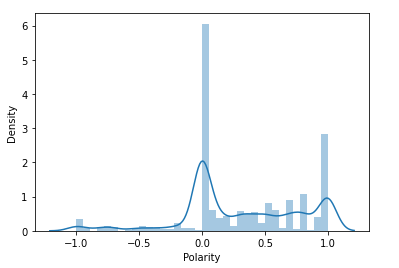
1. **Recall: -** Recall is the ratio of correctly predicted positive observations to the all observation in actual class – yes.

**Recall: -** TP/TP+FN

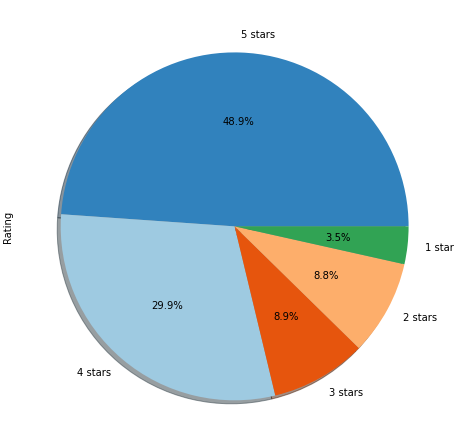
* Visualizations
* Null values present in our dataset.



* Polarity distribution.



* Number of entries present in each star.



* Interpretation of the Results

In the given project, data pre-processing, data visualization and modelling is done.

1. We can conclude from the data pre-processing that there are many null values present in the dataset, if that null value is removed then our model can perform even more accurate.
2. We calculate the polarity, check the distribution of the polarity, removed stopwords, removed punctuation.
3. In modelling phase, we train our data with Multinomial Naïve Bayes, K-Nearest Neighbours, Decision Tree, Random Forest, AdaBoost. And perform a cross validation on these model. Out of these model Random Forest Classifier performs very well. Then we perform the hyper parameter tuning of the Random Forest Classifier model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

This study employs machine learning techniques, to develop a rating prediction model. It uses a rather large publicly available dataset of the review and rating present in different e-commerce website. The classification model performances of the model are compared with one another. The empirical result shows that the Random Forest Classifier algorithm provides superior performances for all metrics under study, the coefficient of determination of accuracy score, and the computational time

* Limitations of this work and Scope for Future Work

The study can be enlarged in a subsequent research by increasing the dataset size so potentially uncovered details and feature of the dataset and of this study can be addressed. An increased dataset would potentially be good enough for employing deep neural networks, which can assure that more in-depth analysis on the rating prediction can be performed. Then, the enlarged housing price prediction problem can be tackled as a classification problem.