

**RATING PREDICTION**

**DOCUMENTATION**

Submitted by:

Pooja Mishra

**ACKNOWLEDGMENT**

I would also like to thank Flip Robo Technologies and Datatrained team who has given me such wonderful opportunity to learn more about Machine Learning and their mentors Shubham Yadav, Sajid Chaudhary for their constant encouragement, valuable suggestions, and cooperation.

This acknowledgement will remain incomplete if I fail to express my deep sense of obligation to my family members and God for their consistent blessings and believe in me.

Thank You.

**INTRODUCTION**

* Business Problem Framing

In this project, online E-commerce websites like Amazon, uses different electronic products reviews to provide different suggestions to different users. Amazon currently uses item-to-item collaborative filtering, which scales to massive data sets and produces high-quality recommendations in real time. This type of filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into a recommendation list for the user. In this project we are going to build recommendation model for the electronics products of Amazon from the data taken on reviews and ratings.

* Conceptual Background of the Domain Problem

In this project, we’ll analyse the various ratings given by customers in online shopping from E-commerce websites. The Exponential growth of internet and its increasing accessibility have accelerated the development of electronic products over the past years. E-commerce is also described as e-business or electronic commerce. It is the purchasing and selling of products through electronic systems over the internet. In simple words, we can say that a person can buy or sell any product without moving out of his/her house. Internet makes this possible by introducing many E-commerce websites which becomes a medium to many businesspersons for selling their products and people sitting at home can get any product as per their need and at desired time. In this project, we have taken reviews and ratings from Amazon website for electronic products.

The exponential growth of the Internet and its increasing accessibility have accelerated the development of electronic

commerce over the past few years. Electronic Commerce is also known as e-commerce or e-business. It is the purchas-

ing and selling of products or services through electronic systems over the Internet.

E-commerce has various advantages as it overcomes geographic limitations and eliminates travel time and cost.

E-commerce allows us to visit the store virtually, with just a few mouse clicks. It is very convenient. An online store is

available all day, every day meaning the customers can visit the store at all times, no matter what their schedule might

be. E-commerce facilitates comparison shopping. Several online portals allow customers to browse multiple e-com-

merce sites and find the best-suited prices.

Both businesses and customers have embraced online sales as a convenient way to shop. But E-commerce has disad-

vantages too. The most significant disadvantage is that one cannot experience the product before pur

* Review of Literature

Amazon is the world’s largest online retailers where millions of customers purchase and review products on its website. However, many Amazon customers do not review and rate products after the purchase, and tons of research projects work on producing better prediction strategies for customer ratings. In this research project, we show a new approach to enhance the accuracy of the rating prediction by using machine learning methods that learn from a graph based on user status, defined by features such as the helpfulness votes and total votes received by the users.

E-commerce has various advantages as it overcomes geographic limitations and eliminates travel time and cost.

E-commerce allows us to visit the store virtually, with just a few mouse clicks. It is very convenient. An online store is

available all day, every day meaning the customers can visit the store at all times, no matter what their schedule might

be. E-commerce facilitates comparison shopping. Several online portals allow customers to browse multiple e-com-

merce sites and find the best-suited prices.

Both businesses and customers have em

E-commerce has various advantages as it overcomes geographic limitations and eliminates travel time and cost.

E-commerce allows us to visit the store virtually, with just a few mouse clicks. It is very convenient. An online store is

available all day, every day meaning the customers can visit the store at all times, no matter what their schedule might

be. E-commerce facilitates comparison shopping. Several online portals allow customers to browse multiple e-com-

merce sites and find the best-suited prices.

Both businesses and customers have em

E-commerce has various advantages as it overcomes geographic limitations and eliminations travel time and cost. It also allows us to visit the store virtually with just with some mouse clicks which is very much convenient. An online store is available all day, every day, 24\*7 which means can visit the store at any time as per their convenience.

* Motivation for the Problem Undertaken

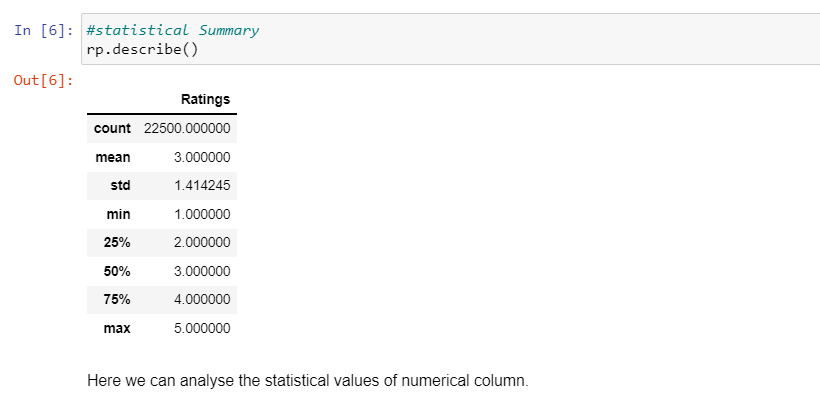
In this challenging and busy environment E-commerce comes out not less than a boon to many of us. The basic objective behind this study to get ourselves aware about the understanding of E-commerce world as well get to know rating prediction with the help of machine learning. By analysing the reviews and ratings, the customer retention can be improved as based on predicting the best rating for respective products will help to suggest similar products to customers. Business must be able to stay ahead of their customers to keep them in a high dynamic market. With help of machine learning they must be able to predict what customers are looking for in their e-commerce store. It gives us an understanding of how a company is going to perform shortly, and accordingly in fact we can change the various business policies to maximize the profit.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

We acquired the dataset from the Amazon site for various electronic products. From this dataset the reviews are based on positive and negative information taken from customer experience.

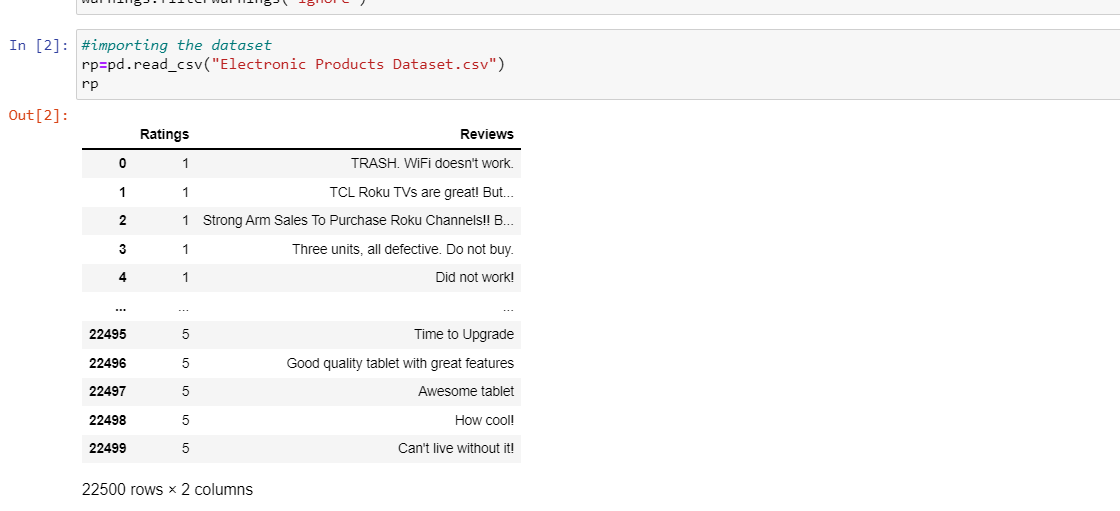
We can observe the count, mean, Standard deviation, min and max range, mean, median, first and third quantile values of each features present in the dataset which will help to perform required data engineering functionalities on the dataset.



The standard deviation here is more than 1 ; the reviews here are with equal count. The data is spread around the mean and rating with 1 are under 25% and rating 2 are above 25% and at 50% rating 3 exist at 75% rating 4 exists and rating 5 at 100%.

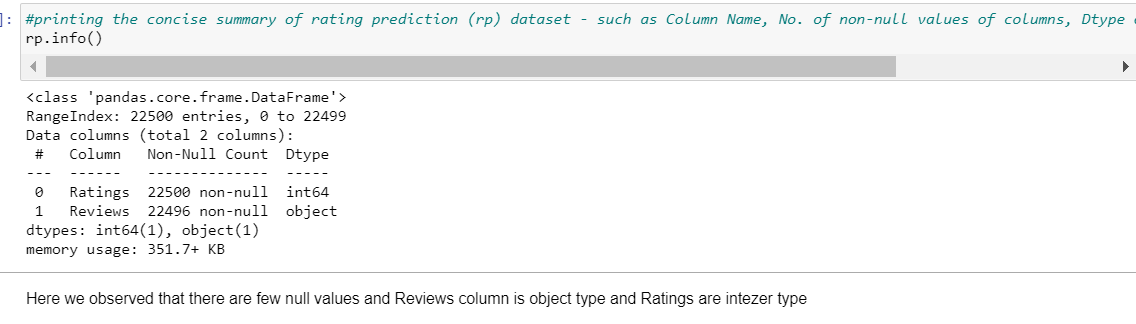
* Data Sources and their formats

This dataset is based on the electronics products available on Amazon. This dataset consists of customer reviews and ratings which have equal count of ratings thus we can say that dataset is balanced. With this analysis we can say that half of the products were associated with positive reviews and half with negative reviews. The dataset contains 22500 records (rows) and 2features (columns).



We can assume that more than 22k customers have shared their reviews for products based on which these will be helpful to improve the products quality and similar products with positive reviews will be suggested to customers who are looking for same quality.

In this dataset only 4 reviews are missing from reviews column.



­

* Data Pre-processing

The dataset here is taken from Amazon website for many electronic products.

Source - Amazon Reviews & Rating. For this case study, we are using the Electronics dataset.

Attribute Information:

● Ratings : Rating star given by users(First Column)

● Reviews : Feedback shared by customers with respect to given ratings. (Second Column)

Data pre-processing is very important to get the dataset into the best format before performing algorithm. This is very important step in Machine Learning that should not be skipped. It involves three stages: Data Cleaning, Data Transformation, and Feature Engineering which converts complicated dataset into quality data.

**Data Cleaning:**

This step, also known as pre-processing, is aimed at cleaning the data. Indeed, text reviews contain unnecessary and redundant characters and have to be normalized. More precisely, the objective of the data cleaning consists in:

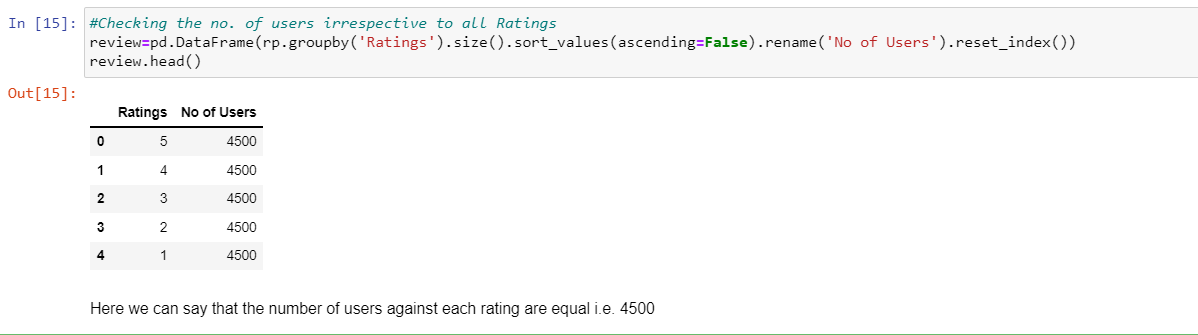
a) Removing punctuation

b) Removing numbers

c) Converting text to lower case (no capital letters)

d) Removing extra whitespace e) Removing stop-words (extremely common words which do not provide any analytic information and tend to be of little value i.e. a, and, are etc.)

Data Resampling – In these datasets, all ratings have equal reviews thus we can say that this dataset is completely balanced.



Most important thing we need to keep in mind that all data pre-processing steps will be performed according the dataset present.

**NLP**

In order to implement the Bag of Words strategy, we will use SciKit-Learn’s CountVectorizer to performs the following:

**Text preprocessing:**

Tokenization (breaking sentences into words)

Stopwords (filtering “the”, “are”, etc)

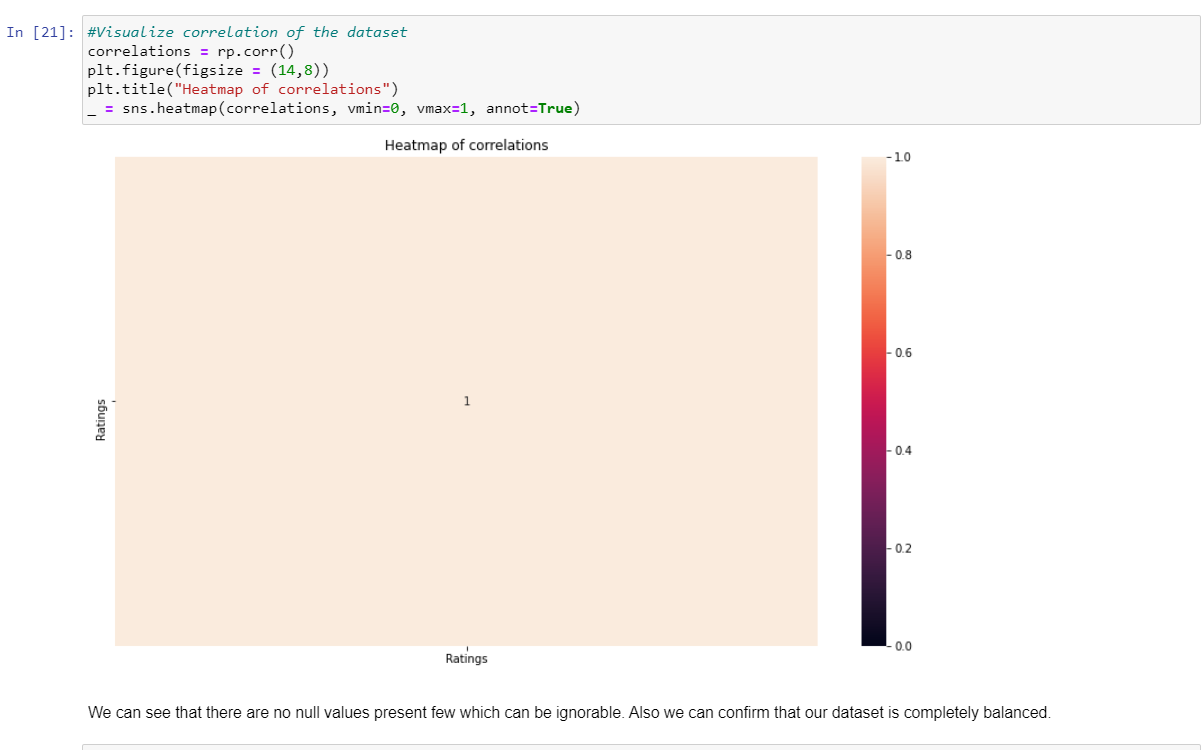
Occurrence counting (builds a dictionary of features from integer indices with word occurrences)

Feature Vector (converts the dictionary of text documents into a feature vector)



**Correlation Matrix:**

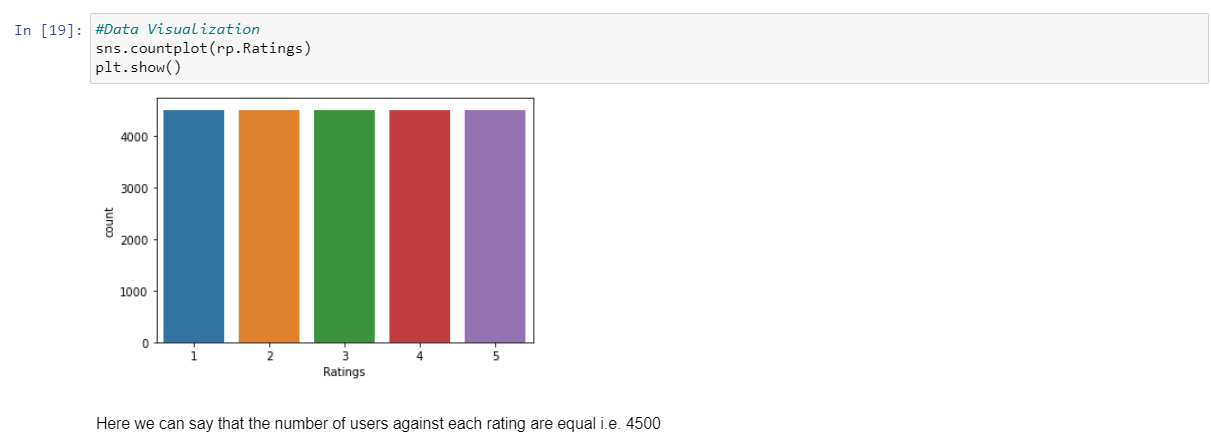
Correlation Matrix is a covariance matrix. A summary measure called the correlation describes the strength of the linear association.



With correlation analysis, we can say that there is relation between ratings and reviews given by customers. Here we also analysed that the dataset is balanced.

* Data Inputs- Logic- Output Relationships

:



With above graph we observed that the count of all present Reviews are same. The rating from 1 to 5 are showing equal number of reviews i.e 4500.

* Hardware and Software Requirements and Tools Used

Machine learning comes with an extensive collection of ML tools, platforms, and software.

Well, in this problem solution we have used following tools that helps to make this project successful as per my possibility.

Jupyter notebook is one of the most widely used machine learning tools among all. It is a very fast processing as well as an efficient platform. Moreover, for this problem solution I have used python programming.

Scikit-Learn is built on top of the three main Python libraries viz. NumPy, Matplotlib, and SciPy. Along with this, it will also help you with testing as well as training your models.

The Libraries are as listed:

1. For Data loading and Visualisation:

Numpy - NumPy is very useful for handling linear algebra, Fourier transforms, and random numbers.

Pandas - Pandas are turning up to be the most popular Python library that is used for data analysis with support for fast, flexible, and expressive data structures designed to work on both “relational” or “labeled” data.

Matplotlib - The library helps to generate histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.

Seaborn – Used for visualization.

train\_test\_split: We use it to perform test train spliting.

**Algorithm Libraries:**

In this problem solution we have used algorithm libraries:

Classifier Models:

Naïve Bayes Classifier

Random Forest Classifier

Logistic Regression Classifier

**NLP Libraries:**

from collections import Counter

import string

import nltk

from collections import Counter

import matplotlib.pyplot as plt

from nltk.corpus import stopwords

from nltk.sentiment.vader import SentimentIntensityAnalyzer

from nltk.tokenize import word\_tokenize



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

In this project, we will build our prediction model: we will choose algorithms for each of the techniques we mentioned in the previous section. After we build the model, we will evaluate its performance and results. In E-Commerce, classifier algorithms can be used to classify ratings of review based on words. The specific words in the language are categorized in advance for their positive or negative sentiments.

* Testing of Identified Approaches (Algorithms)

There are a lot of techniques available for Classifier problems like Random Forest Classifier, Naïve Bayes Classifier, Logistic Regression etc. In simple classifier model, a simple count of positive and negative data points will define overall positive or negative sets. There is a problem with this. For example, in the case of words in sentences, “Great” and “Good” both are positive words. But “Great” has a higher impact than “Good”. We need to train our model to weigh identified data points. Thus, based on words we can clarify the rating with help of classifier algorithms.

* Run and Evaluate selected models

After getting the simplified dataset, we’ll perform test train split to use it on algorithms.

## Splitting the Dataset

As usual for supervised machine learning problems, we need a training dataset to train our model and a test dataset to evaluate the model. So we will split our dataset randomly into two parts, one for training and the other for testing. For that, we will use another function from Scikit-Learn called train\_test\_split():



On this dataset, I performed Naïve Bayes Classifier, Random Forest Classifier and Logistic Regression Classier from which unfortunately not got good accuracy as expected. From all performed Algorithms Logistic Regression given the highest accuracy amongst all.



* Interpretation of the Results

The results in this analysis confirms that there is need of more analysis and Data Pre-processing to get the better accuracy as 455 is not the good accuracy to fit the model. Also, both neutral and negative reviews has large standard deviation with small frequencies, which we would not consider significant as shown by the lower precision, recall and F1 scores in the classification report.

**CONCLUSION**

* Key Findings and Conclusions of the Study

After performing 3 algorithms I can conclude that Logistic Regression seems to be the best model but as the results shows the accuracy is not good thus need to perform the modelling again to get the best accuracy

we created many models: for each model, we searched for good parameters then we constructed the model using those parameters, then trained (fitted) the model to our training data (X\_train and y\_train), then tested the model on our test data (y\_test) and finally, we evaluated the model performance by comparing the model predictions with the true values in y\_test.

* Learning Outcomes of the Study in respect of Data Science

After working on this project, we came to conclude following observations:

To apply NLP data pre-processing and preparation techniques in order to obtain clean data.

To build machine learning models able to predict rating of the products.

To analyze and compare models performance in order to choose the best model.

As This dataset consist of Reviews “words” thus was very helpful to gather more information by performing all functionalities thus will explore more to get best accuracy for this dataset.

* By Limitations of this work and Scope for Future Work

With help of this project data science has given solution to one of the necessary problems in the E-commerce world so that ratings /reviews can be analysed.

.

Thank You.