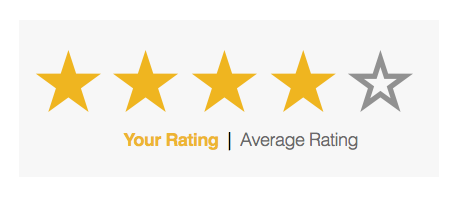


**Customer Ratings Prediction**



**Submitted by:**

**DEEPTHI MALLU**

**ACKNOWLEDGMENT**

I take this opportunity to acknowledge everyone who have helped me in every stage of this project.

Firstly, I am indebtedly grateful to my SME MR. Keshav Bansal sir, who helped me from beginning of my Project. Am also thankful to my Mentor Shankar Gowda Sir and my whole Data Trained team, where I have learnt Analysing the datasets and building the models using Machine learning and making the projects. Finally, am so thankful to my Flip Robo Technologies team, as they provided me the opportunity to work as intern in their company.

I feel pleasure, to make project report on “Malignant comment classifier”. It has been my privilege to have a team of project guide who have assisted me from the commencement of this project. The project is a result of my hard work, and determination put on by me with the help of Wikipedia, You Tube videos related to NLP Concepts, skikit-learn.org, Reffered some old projects on Google.

**INTRODUCTION**

**Business Problem Framing:**

Text communication is one of the most popular forms of day to day conversion. We chat, message, tweet, share status, email, write blogs, share opinion and feedback in our daily routine. All of these activities are generating text in a significant amount, which is unstructured in nature. I this area of the online marketplace and social media. It is essential to analyze vast quantities of data, to understand people’s opinion.

NLP enables the computer to interact with humans in a natural manner. It helps the computer to understand the human language and derive meaning from it. NLP is applicable in several problematic from speech recognition, language translation, classifying documents to information extraction. Analysing, customer reviews is one of the classic examples to demonstrate a simple NLP Bag-of-words model, on customer reviews.

As per the business problem one has to scrape the reviews and ratings given by the customers on products of different ecommerce websites and build machine learning models such that on the given review the model has to predict the ratings.

**Conceptual Background of the Domain Problem:**

When it comes to making decisions, we humans aren’t always independent thinkers. From buying our morning coffee to big-ticket items like a new car, we depend on our network’s opinions, advice, or perspectives for making the right decision.

With the internet, those networks get way, way bigger. We’re no longer confined to just friends, family, and co-workers. we can see what people all around the world think with just a quick Google search.

And these customer reviews hold serious weight with shoppers. Online reviews either positive or negative can impact over 93% of consumers’ decisions.

While reviews can accumulate on their own, they shouldn’t exist in a vacuum. Knowing how to ask for reviews, leverage them to get more business, and respond to less-than-favourable customer testimonials can improve your business image and land you more long-lasting customers.

Based on the reviews and ratings given by the customers one come buy a product from the ecommerce websites. So, Ratings and reviews helps a product to buy in huge number for ecommerce website.

**Review of Literature:**

Customer reviews build something known as social proof for other customers to buy a product, a phenomenon that states people are influenced by those around them. This might include friends and family, industry experts and influencers, or even internet strangers.

Social proof can push customers who are on the fence about buying a product to make a purchase (or consider other alternatives). While there are many different forms of social proof (like influencer campaigns and company partnerships), customer reviews have a special place in shoppers’ hearts.

When considering a purchase, people want first-hand perspectives from other people just like them. While a famous influencer might catch their attention, experiences from peers are also important if you want to convince them to buy.

Depending on the industry your business is in, customer reviews might be especially important because they help offset a generally negative view of certain industries. For example, if you work in advertising, you’ve got a bit of an uphill battle only 23% of respondents surveyed in the [Customer Communications Review](https://www.ringcentral.com/customer-communications-trends-report.html) said that ad agencies provided “very good” or “pretty good” service, and ad agencies rank dead last in comparison to other major industries:

Customer reviews come in many different forms. Including a few good ones on your website can improve your chances of winning more customers.

For nearly 9 in 10 consumers, an online review is as important as a personal recommendation.

Customers are likely to spend 31% more on a business with “excellent” reviews.

72% say that positive reviews make them trust a local business more

92% of users will use a local business if it has at least a 4-star rating

72% of consumers will take action only after reading a positive review

Reliability (27%), expertise (21%) & professionalism (18%) are the most important reputation traits for a local business

Estimated Number Of Online Reviews Customers Read Before Trusting A Business

|  |  |
| --- | --- |
| Estimated number of reviews | %age |
|  |  |
| 0 | 8% |
| 1 | 3% |
| 2-3 | 24% |
| 4-6 | 32% |
| 7-10 | 18% |
| 11-20 | 8% |
| 21-30 | 3% |
| 31-40 | 1% |
| 41-50 | 1% |
| 51+ | 2% |

On average, a one-star increase on Yelp leads to a 5 to 9% increase in a business’s revenue. At the same time, a single negative review can cost a business about 30 customers.

The number of reviews posted every minute by Yelp users is 26,380

The percentage of Yelp users that have made a purchase at a business they found on Yelp is 98%

The percentage of Yelp users that visit Yelp because they intend to make a purchase is 80%

**Motivation for the Problem Undertaken:**

In today's area of internet and online services, data is generating at incredible speed and amount. Generally, Data analyst, engineer, and scientists are handling relational or tabular data. These tabular data columns have either numerical or categorical data. Generated data has a variety of structures such as text, image, audio, and video. Online activities such as articles, website text, blog posts, social media posts are generating unstructured textual data. Corporate and business need to analyze textual data to understand customer activities, opinion, and feedback to successfully derive their business. To compete with big textual data, text analytics is evolving at a faster rate than ever before.

Text Analytics has lots of applications in today's online world. By analysing, customer reviews on ecommerce websites, we can know customers reaction on purchasing a product. Ecommerce websites can understand user feedback or review on the specific product. Book My Show can discover people's opinion about the movie. You-tube can also analyze and understand people’s viewpoints on a video. So, I have taken this project to enhance my skills in the field of data science, to get hands on experience and to know that how the data scientist approaches and work in an industry end to end.

**Analytical Problem Framing**

**Mathematical/Analytical Modelling of the Problem:**

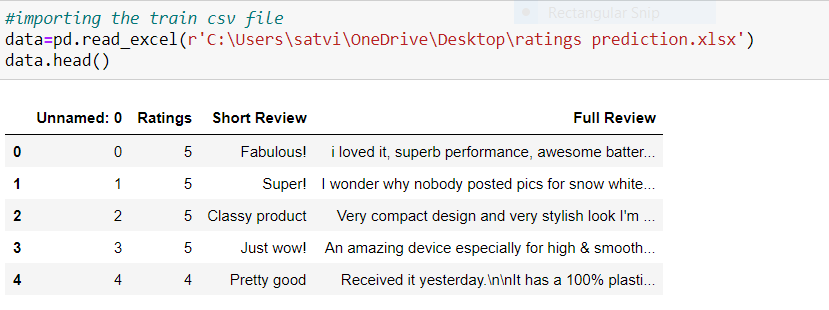
In this project, I have collected the reviews and ratings from ecommerce websites. I have collected the data from Flipkart and Amazon. I have used web-scraping, (selenium) for scraping the details. Then I have scraped the reviews of different laptops, Phones, Headphones, smart watches, Professional Cameras, Printers, Monitors, Home theatre, Router from Amazon and Flipkart. I have collected the columns 1) reviews of the product. 2) rating of the product 3) short reviews. So, by using all the information of the data collected I need to predict the Ratings. After collecting the data, I have put all together in a data frame and saved the data as excel file.

As we are dealing with the reviews, the reviews will be in the form of text. So, this is an NLP based Project and we deal with the textual data. The text data consists of letters, punctuations, strings etc which cannot be understand by the machine so we need to pre-process the text data. So, I used some methods like removing punctuations, numbers, stop words using the lemmatization procedure to convert the complex words into their simpler forms. These processes helped in cleaning the unwanted words form the comments and we were left with only those words which were going to help in our model building. After cleaning the data, we used TF-IDF Vectorizer technique to convert textual data into vector form. This technique works on the basis of the frequency of words present in the document. After training with train dataset, the same I have followed for the test dataset and converted all the text data and applied machine learning algorithms

**Data Sources and their formats:**

This data was collected from ecommerce websites like Flipkart and Amazon using web-scraping selenium and saved the data in a excel file. Using pandas, I have first imported the Excel file and it consists of different columns which includes data in it. Our dataset consists of Features and label. After importing I have checked for shape of the dataset and which consists of rows and columns. Then I checked for null values and need to be treated and then I checked for info () method for knowing the type of the data then I checked for stats using describe method.

Our label is Ratings prediction which is a continuous variable based on the values of independent variables our dependent variable depends



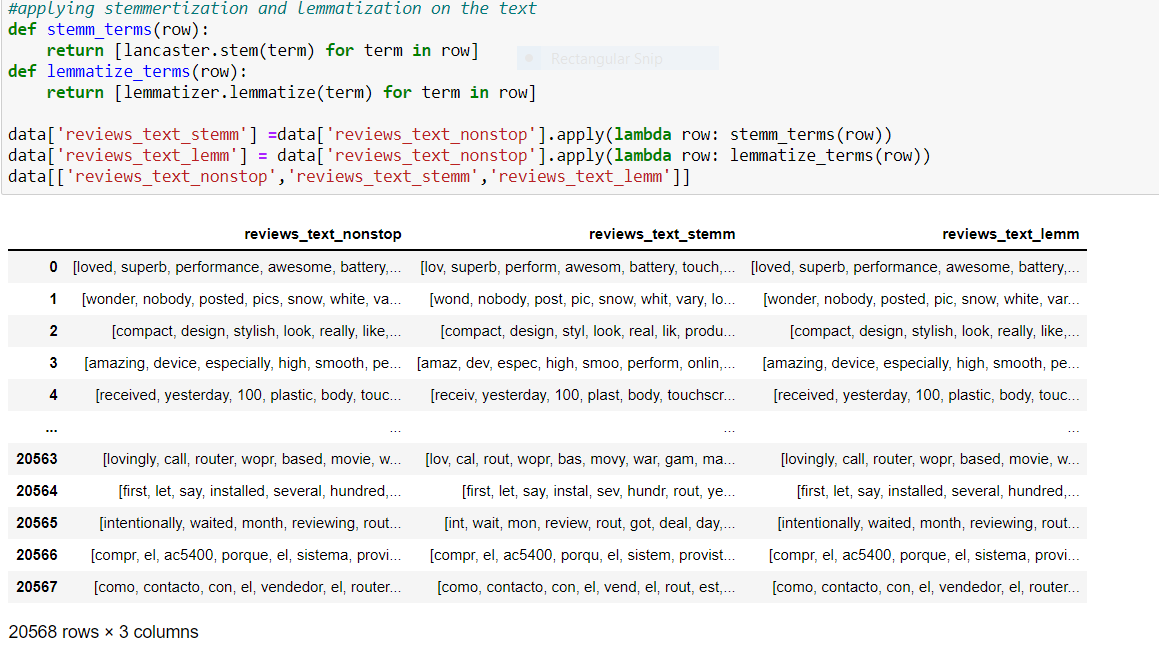
**Data Pre-processing:**

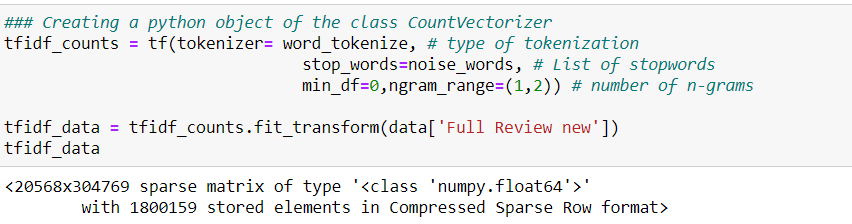
Data pre-processing is the data mining technique that involves transforming raw data into an understandable data format. So, in data pre-processing technique first step is import data and the libraries to be used in model building.

In the dataset there is a column of long review text column, which cannot be understandable by the machine, so by using the Natural language processing techniques (NLP) we need to pre-process the data and apply various techniques to convert the text data and analyse the data. As a data scientist, the aim is to filter out these words from the ecommerce websites and predict the ratings (1 to 5) based on the reviews given by the customers.

Since our feature column, has many stop-words, punctuations and special characters they are to be removed from the reviews text because our machine cannot understand the characters and cannot be able to convert to the machine understandable language. So that need to be pre-processed, to be converted to lower cases and split the text.

Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms.



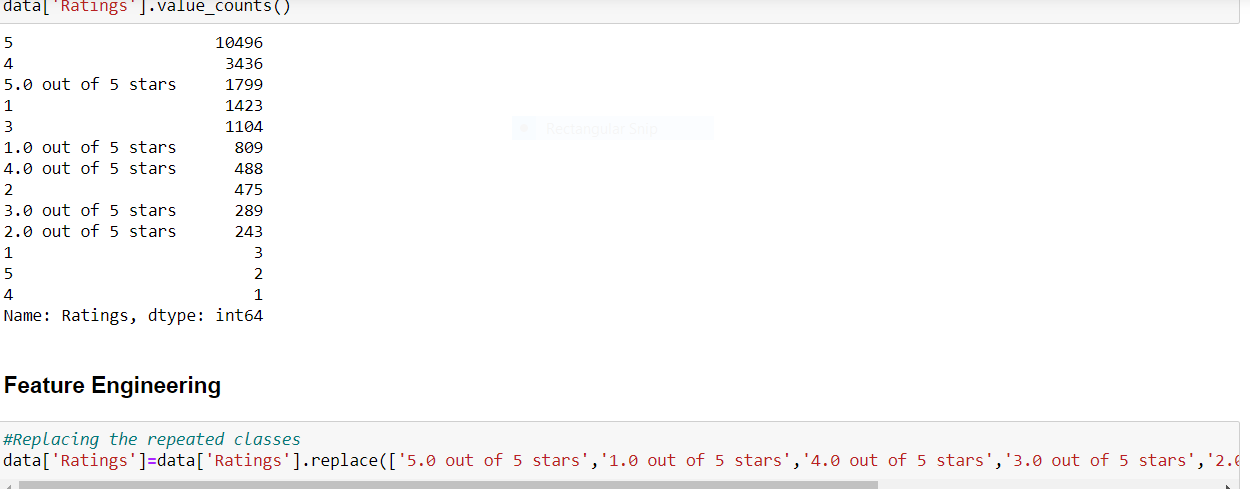


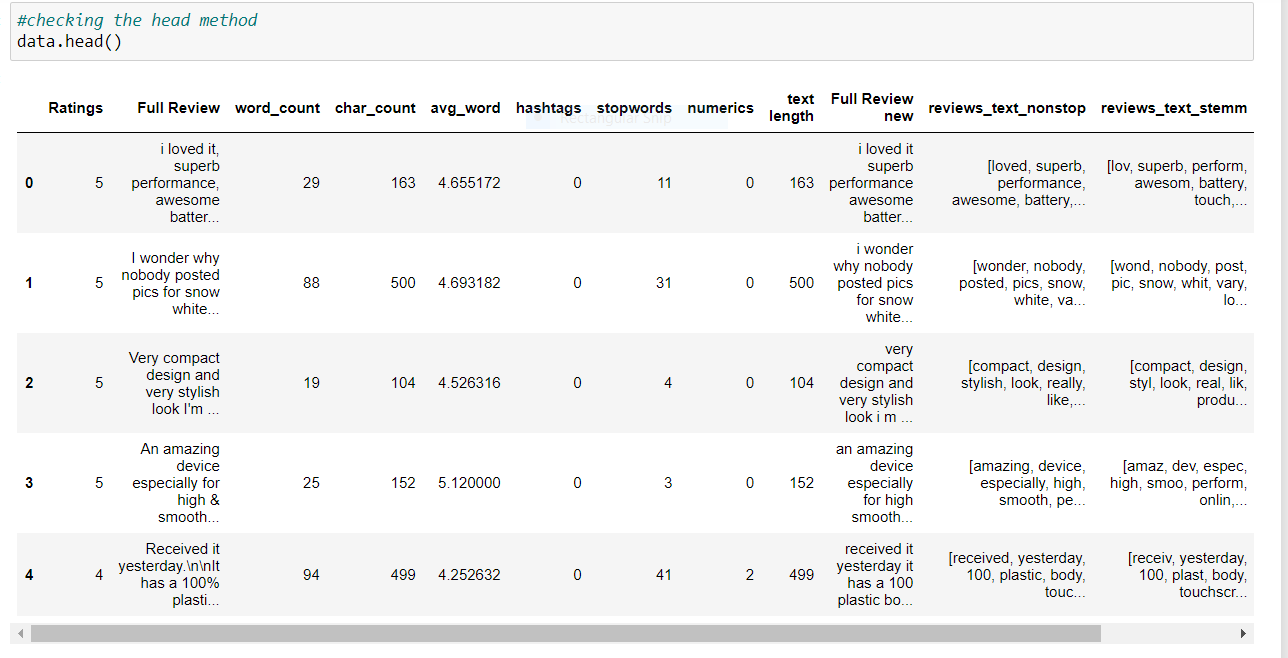
**Data Inputs- Logic- Output Relationships**

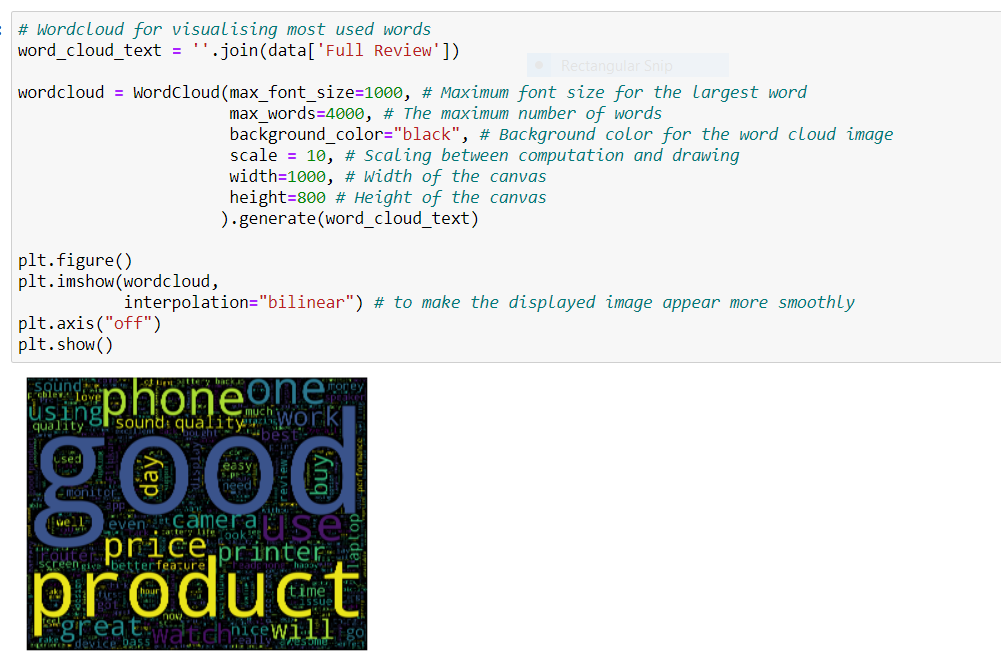
The input data for the processing and getting the output is converted in to the numerical forms or data words, in to the vector form and fed to the model one by one which analyses by the model providing the certain score through the medium of performance metrics.

So, there are some missing values in columns I have filled them using mode. Next step of data cleaning. In data cleaning I have dropped the unknown column and short review columns as it gives no information. After dropping I preprocessed the ratings columns which has various columns which gives the same information.

Now in cleaning the textual data we created a function to remove unwanted space, punctuation, numbers, emails, phone numbers etc and converted upper case letters into lower case and append the result into a new column. After the removal of unwanted notations, we moved to remove stop-words.

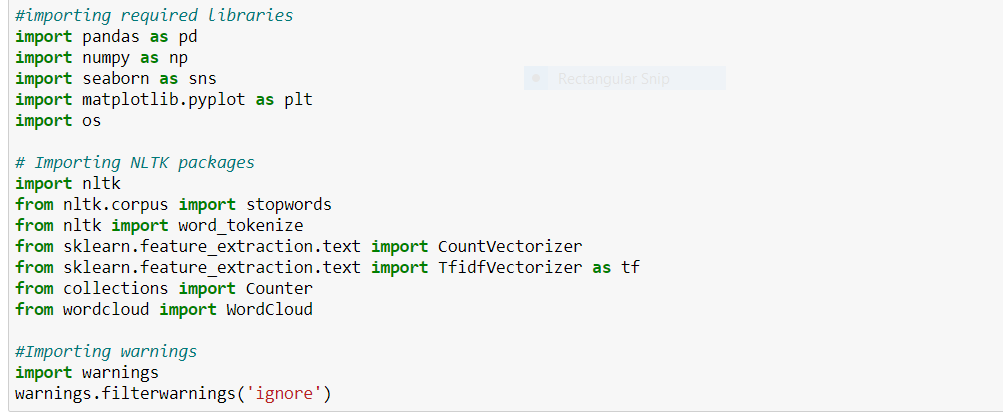


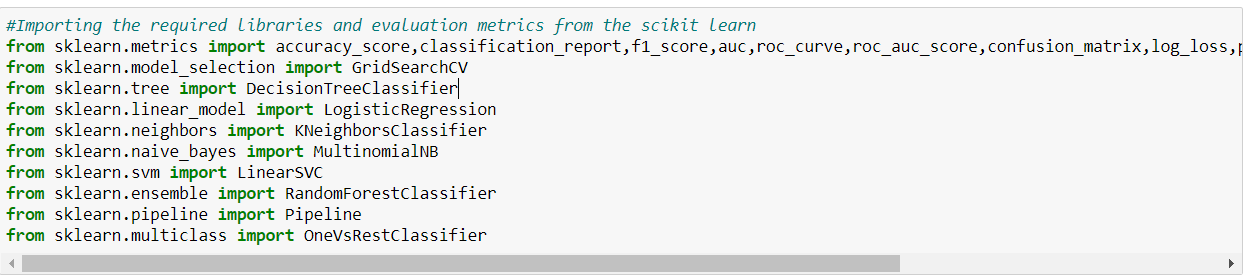




**Hardware and Software Requirements and Tools Used:**

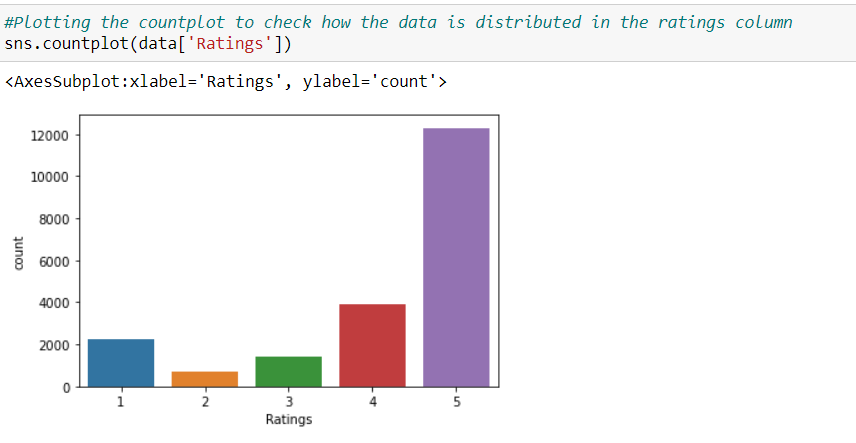
I have used Laptop with i5 processor with 8 GB Ram, Jupyter Notebook, Python libraries for processing, Scikit learn library for using machine learning concepts and NLP concepts for pre-processing the data converting the text to the words using tokenization and then words to vectors using TDIF.



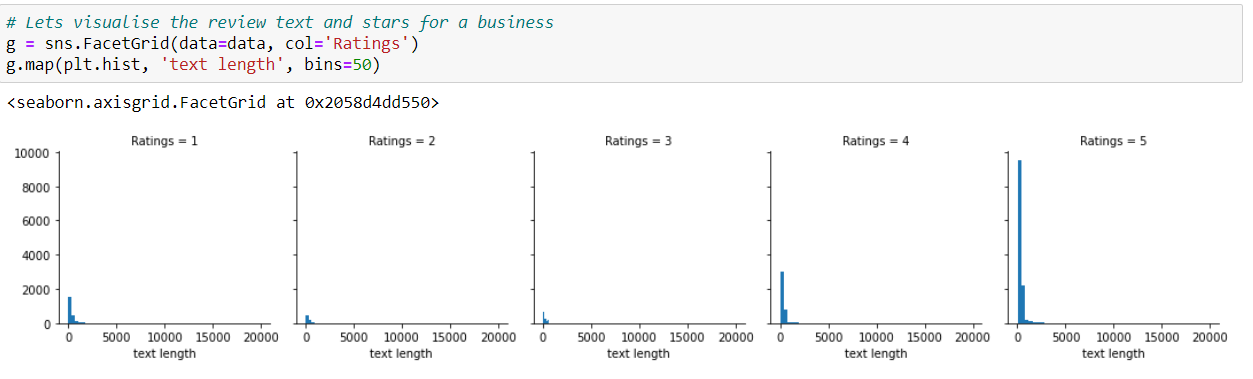


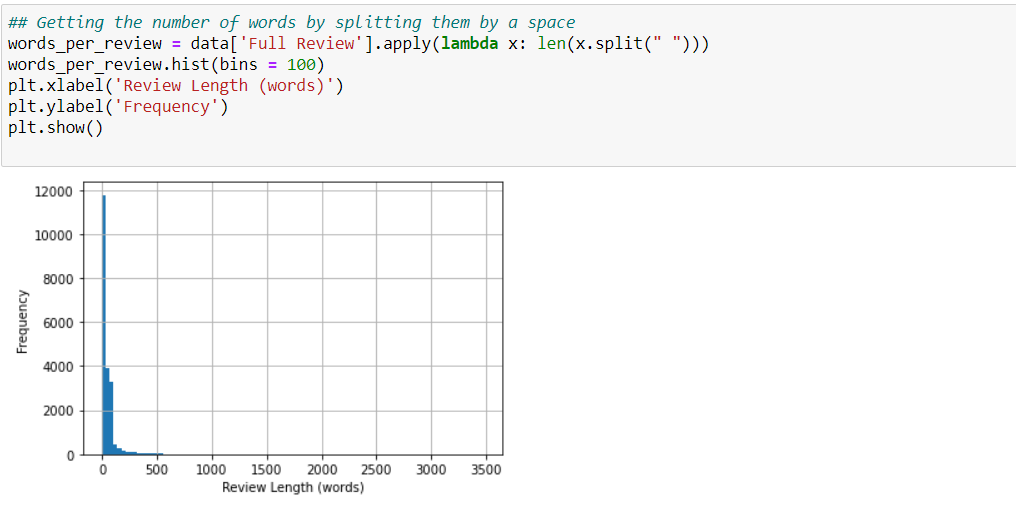
**Visualisation:**

**I have plotted count plots for all the labels which are having the classes as 1 to 5 ratings and checked how the classes are distributed throughout the columns.**

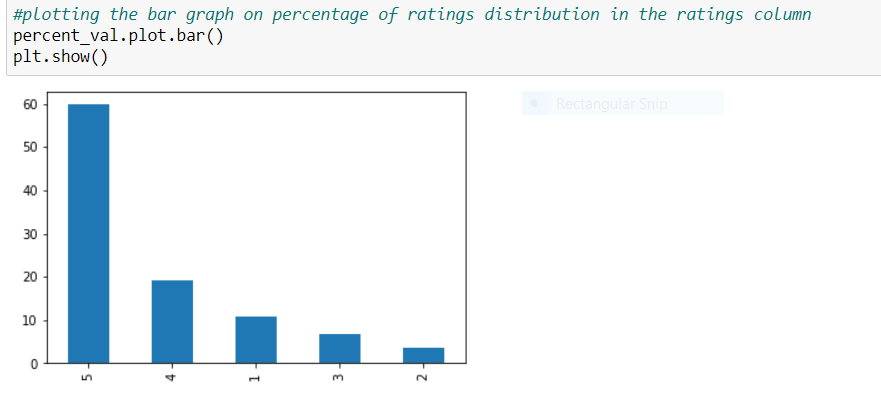


**I have plotted the face grid to visualise between the reviews text and ratings given by the customers.**





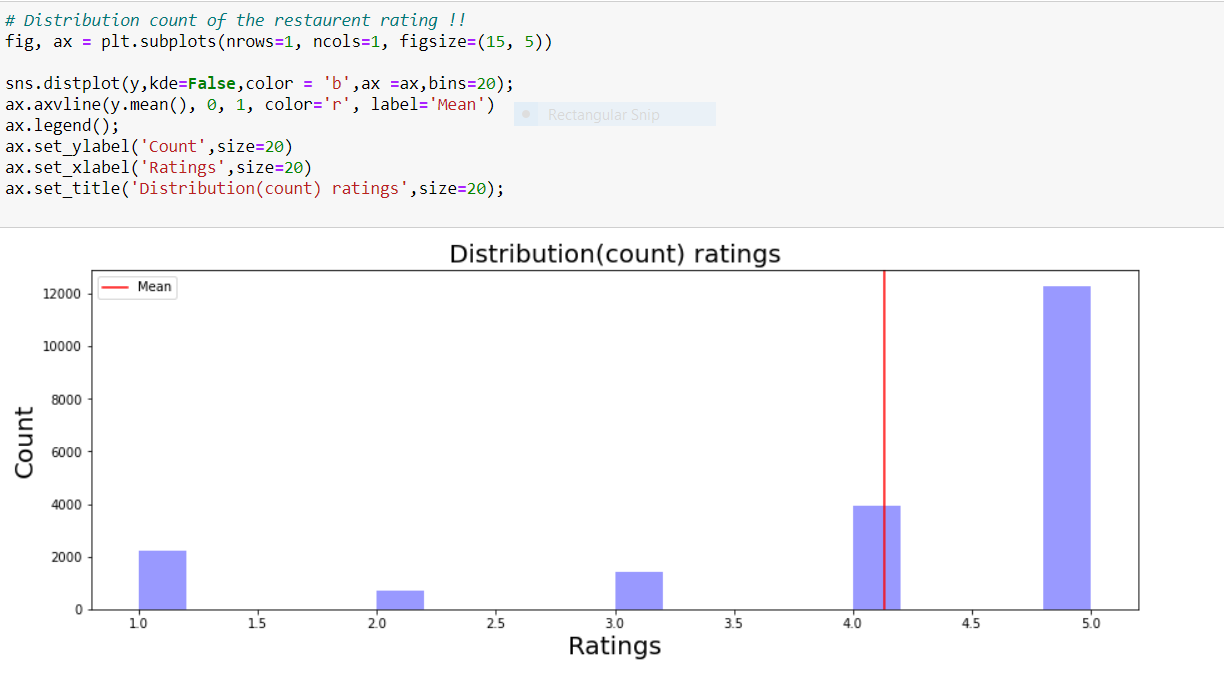
**Plotted bar graph to visualise ratings.**



I have plotted bar graphs and checked the frequent occurances of the classes in the labels.

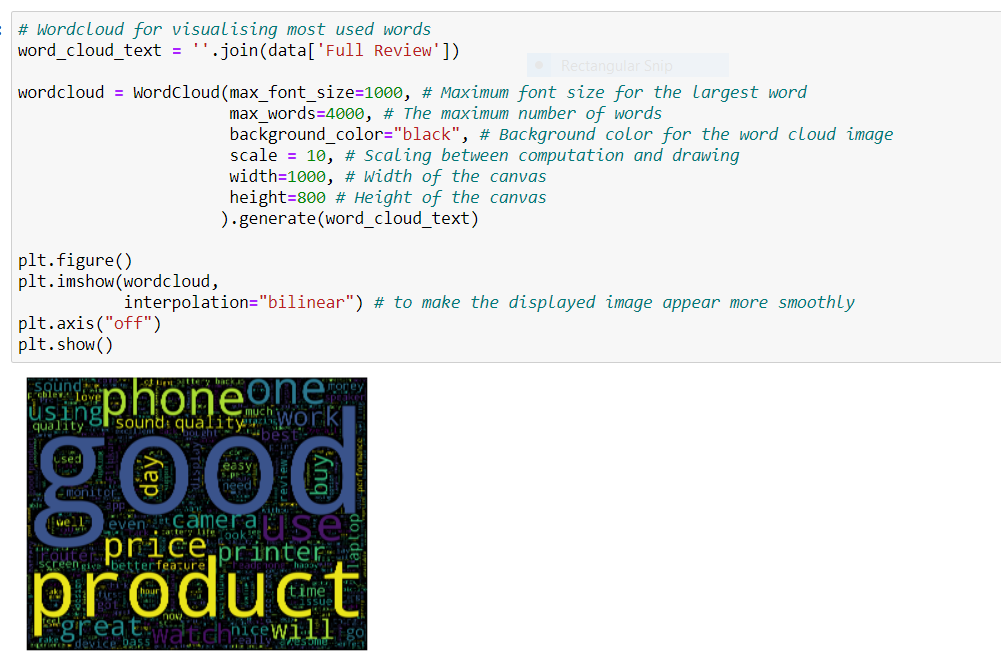
Percentages of the categories of the labels are taken and plotted the graph and checked how the comments are distributed in each label as shown in below screenshot.

**Distribution of ratings based on the reviews of a product**



On average the customers gave 4star ratings on the scraped products.

I have plotted and visualised the frequent used words in the comment texts as shown below using word cloud.



**Identification of possible problem-solving approaches (methods)**

These are the following approaches I have used here:

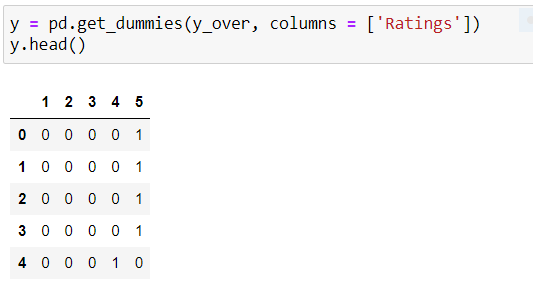
• Importing and drawing insights from the data using python libraries.

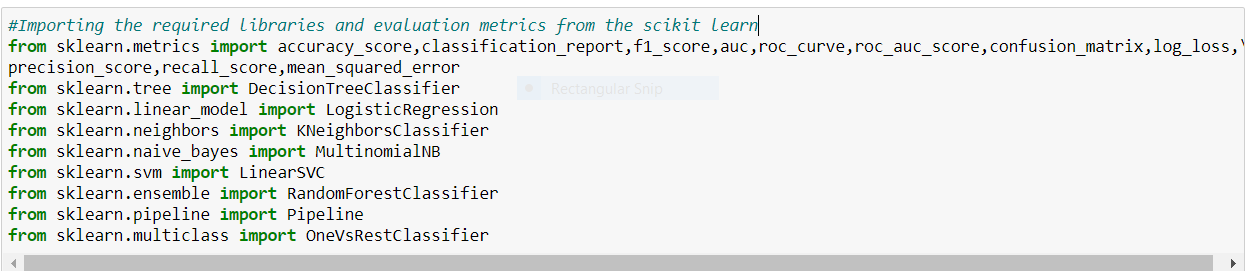
• Analysing the data and used proper pre-processing techniques to extract out important words from the comment text using NLP concepts.

• Applying data operations such as TDIF to convert text to vectorized form.

• Applied different Machine learning algorithms to check the performance using evaluation metrics.

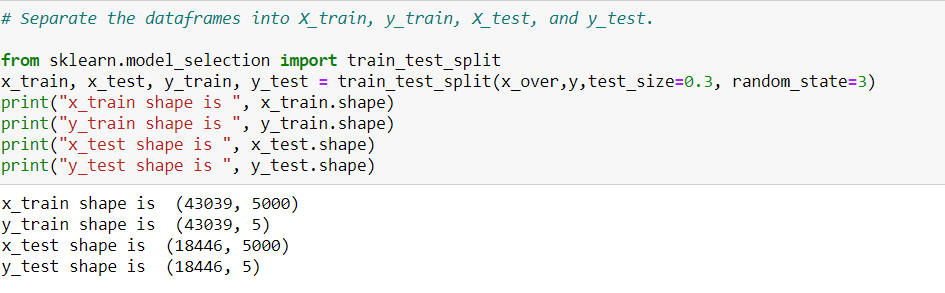
As our target feature is imbalanced. I used SMOTE to balance the label. By using the get dummies method I have converted the ratings label in to 5 labels which defines the type of Ratings like 1star rating, 2star rating, 3star rating, 4star rating, 5star rating because on regression models the model score is very less so I have used get dummies method and converted the label in to multi class, so this type of problem comes under classification problem so I have used classification algorithms for the prediction. We cannot use regular algorithms for prediction because if we use regular SVC or ensemble techniques that may eat up all the memory and there is will be no output at last. So, I have used Linear SVC model rather than regular SVC and imported the algorithms and the methods required for evaluating the models and for splitting our dataset in to train and test data using train test split method.





**Testing of Identified Approaches (Algorithms)**

First, I have splitted the dataset into and x and y variables and then applied TDIF method on the words and converted the words into vectors and then split the x and y variables into train and test data. I have considered 70% train size and 30% test size and took the random state as 3. Then applied Machine learning algorithms on the train data and test data, Checked the performance, evaluated the model and finalised the model.



Following are the algorithms that applied on this dataset.

LogisticRegression

KNeighborsClassifier

DecisionTreeClassifier

MultinomialNB

LinearSVC

RandomForestClassifier

I have used these models because when I tried applying different models for prediction but I have come across with memory related errors, so in order to avoid the errors I have used these models for prediction.

**Logistic Regression**













**Metrics for success in solving problem under consideration:**

For solving the problems and understanding the result of each algorithm, we have used different metrics like

a) Accuracy Score

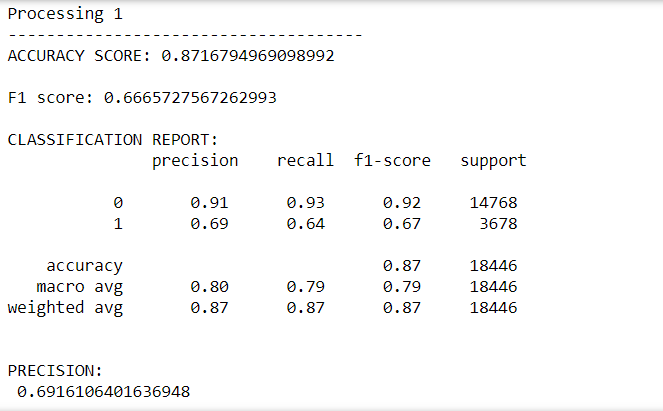
b) Classification Report

c) Confusion Matrix

d) Log Loss

e) Roc-Auc

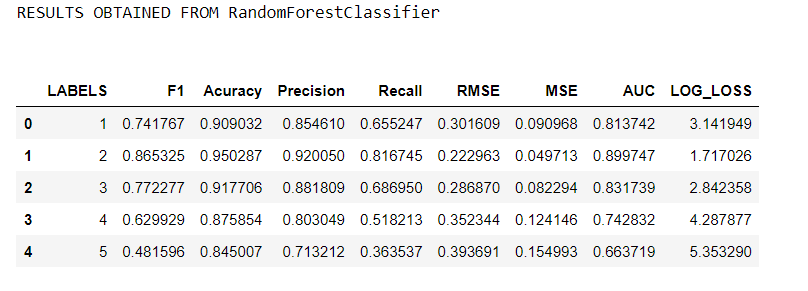
Key metrics are F1-score, Precision, Recall and Roc-Auc Score.



**Interpretation of the Results**

Almost all the algorithms which we applied are giving the good accuracy and the performance is also good. I have considered Random Forest classifier as my final model for building since it is giving best output results and evaluation metrics.

As our target feature is imbalanced which means that alone accuracy score will not give best results but instead accuracy, we are using the classification report log loss score, Roc-AUC-score and confusion matrix to find the best algorithm. So, I have selected Random Forest Classifier which has good ROC\_AUC score, accuracy, RMSE Score is less and log loss is less compared to all the other models.



**CONCLUSION** **Key Findings and Conclusions of the Study**

Some of the findings from data explorations are:

• From displaying the data, it seems there are of special characters present in the data. So, it is better to proceed by filter it out.

• As the reviews data is in text format, so presence of special characters and stop-words will be there.

• After proper cleaning and processing, Random Forest Classifier gives the highest accuracy and ROC AUC score.

From this project I have learnt, new techniques and ways to deal with uncleaned data and find a solution to deal with multiple target variables. Tools used for visualizations, gives a better understanding of dataset. We have used different classification algorithms in which we have two labels. So, after applying many models Random-forest classifier gave better results compared to other models.

**Limitations of this work and Scope for Future Work**

Some of the extension techniques that gives the best results are multinomial Naive bayes algorithm and Random Forest Algorithm. After, analysing for each behaviour separately by creating models. • Random Forest Classifier gave me good results on text data.

As I have used web scraping for data collection the data collection took long time and even tried to scrape the data from other websites but there are no ratings and reviews much. So may be in future all the ecommerce websites should design themselves to get reviews and gain ratings from the customers for the purchase of products. So that we can collect more data which helps us to get better insights and results from the text data in future and predict ratings with minimum errors.