

MICRO CREDIT DEFAULTER PROJECT



Submitted by:

**Bhushan Kumar Sharma,**

**Intern Data Scientist**

**ACKNOWLEDGMENT**

**\*\*\*\*\***

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with this dataset and it has helped me to improve my model building skills.

A huge thanks to my academic team “Datatrained” who is the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank to many other persons who has helped me directly or indirectly to complete the project.

**Following are the external references which I used:**

[www.geeksforgeeks.org](http://www.geeksforgeeks.org)

[www.stackoverflow.com](http://www.stackoverflow.com)

[www.w3school.com](http://www.w3school.com)

[www.google.com](http://www.google.com)

Datatrained Lectures

**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 reviews 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 and Yelp.

There are two main methods to approach this problem. The first one is based on review text content analysis and uses the principles of natural language process (the NLP method). This method lacks the insights that can be drawn from the relationship between costumers and items. The second one is based on recommender systems, specifically on collaborative filtering, and focuses on the reviewer’s point of view.

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 rating. So we, we have to build an application which can predict the rating by seeing the review.

* **Conceptual Background of the Domain Problem**

Recommendation systems are an important units in today's e-commerce applications, such as targeted advertising, personalized marketing and information retrieval. In recent years, the importance of contextual information has motivated generation of personalized recommendations according to the available contextual information of users. Compared to the traditional systems which mainly utilize user’s rating history, review-based recommendation hopefully provide more relevant results to users. We introduce a review-based recommendation approach that obtains contextual information by mining user reviews. The proposed approach relate to features obtained by analysing textual reviews using methods developed in Natural Language Processing (NLP) and information retrieval discipline to compute a utility function over a given item. An item utility is a measure that shows how much it is preferred according to user's current context. In our system, the context inference is modelled as similarity between the user’s reviews history and the item reviews history. As an example application, we used our method to mine contextual data from customer’s reviews of technical products and use it to produce review-based rating prediction. The predicted ratings can generate recommendations that are item-based and should appear at the recommended items list in the product page. Our evaluations (surprisingly) suggest that our system can help produce better prediction rating scores in comparison to the standard prediction methods.

* **Review of Literature**
  + Some websites do not always offer structured information, and all do not leverage user’s unstructured information, i.e. reviews, explicit social networks information is not always available and it is difficult to provide a good prediction for each user. For this problem the sentiment factor term is used to improve social recommendation.
* **Motivation for the Problem Undertaken**

The project was first provided to me by FlipRobo as a part of the internship program. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary objective.

Many product reviews are not accompanied by a scale rating system, consisting only of a textual evaluation. In this case, it becomes daunting and time-consuming to compare different products in order to eventually make a choice between them.

Therefore, models able to predict the user rating from the text review are critically important. Getting an overall sense of a textual review could in turn improve consumer experience. However, the motivation for taking this project was that it is relatively a new field of research. Here we have many options but less concrete solutions.

The main motivation is to build a prototype of online hate and abuse review classifier which can used to classify hate and good comments so that it can be controlled and corrected according to the reviewer’s choice..

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

The data was collected by using web scrapping for extracting review data. In web scrapping I have used selenium. In this dataset problem the Ratings can be 1, 2, 3, 4 or 5, which represents the likely ness of the product to the customer.

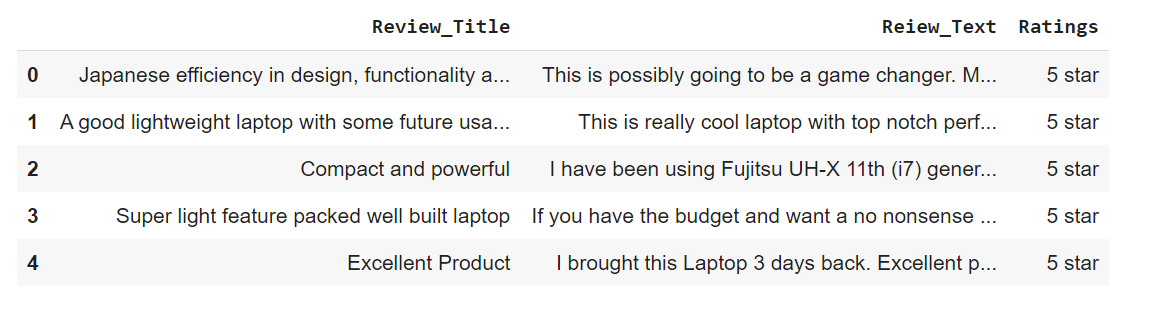
As it is containing multi class So clearly it is a multi-classification problem and I have to use all classification algorithms while building the model. We would perform one type of supervised learning algorithms: Classification. Here, we will only perform classification. Since there only 1 feature in the dataset, filtering the words is needed to prevent over fit.

In order to determine the regularization parameter, throughout the project in classification part, we would first remove email, phone number, web address, spaces and stops words etc. In order to further improve our models, we also performed TFID in order to convert the tokens from the train documents into vectors so that machine can do further processing. I have used all the classification algorithms while building model then tuned the best model and saved the best model.

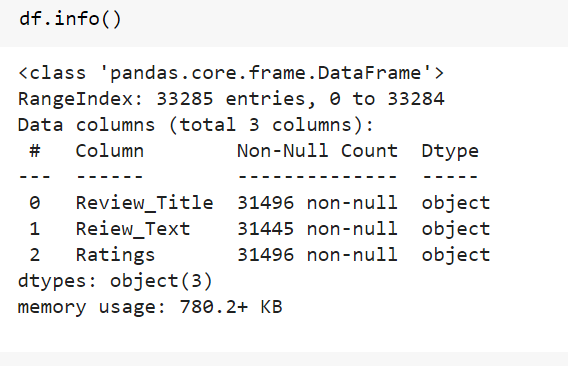
* **Data Sources and their formats**

Scraped dataset is having 33285 records and 3 columns, in this dataset “Ratings” column is my target column and I have to apply machine learning algorithm accordingly.

* **Review\_Title**: Title of the Review.
* **Review\_Text**: Review content
* **Ratings**: Rating of review

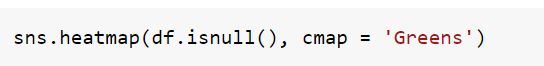


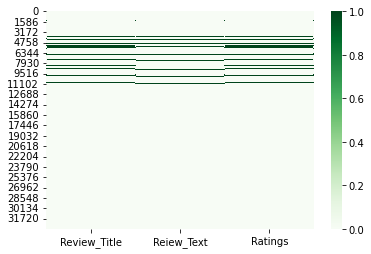
**Data Information**



As we can see in the above output not-null values are not equal for every column, which is indicating that dataset is having null values.

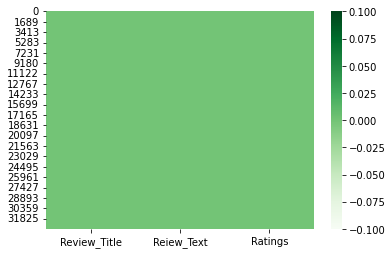
* **Data Pre-processing:**
* **Null values identification:**

****

****

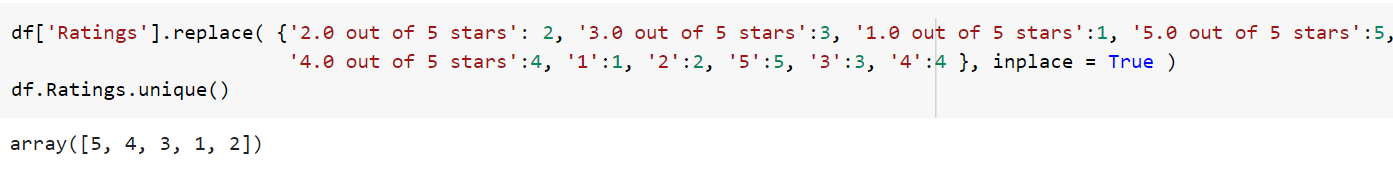
After found null values, I dropped the all null values from the dataset



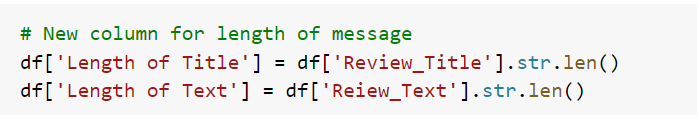


Note: **As heatmap is now clear, which indicating Now, no null values are present in the dataset.**

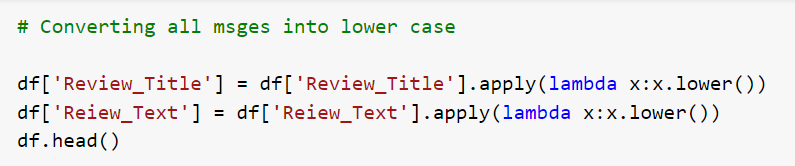
* *Converted all target values into [5,4,3,2,1] class only*

****

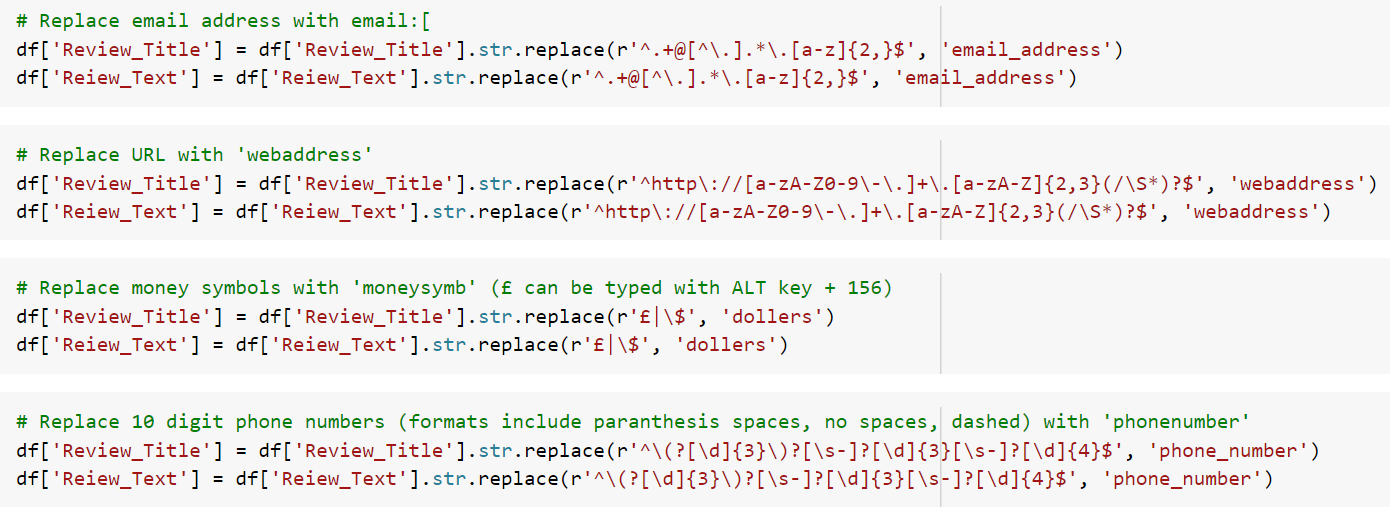
* *Created New columns for length to get amount of cleaned data records*



* *Converted all text into lower case:*

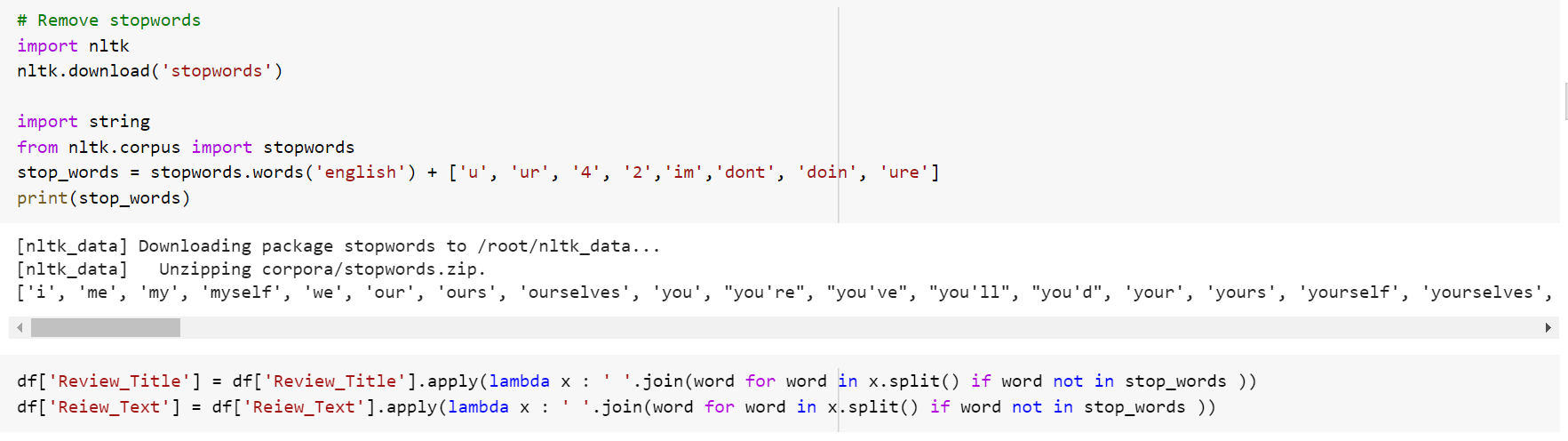


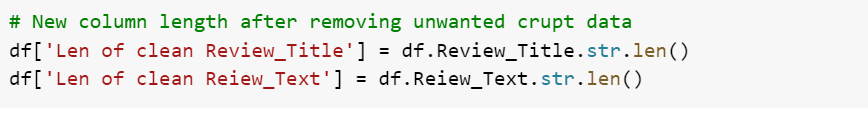
* *Various operations perform to clean the review content and review title content*





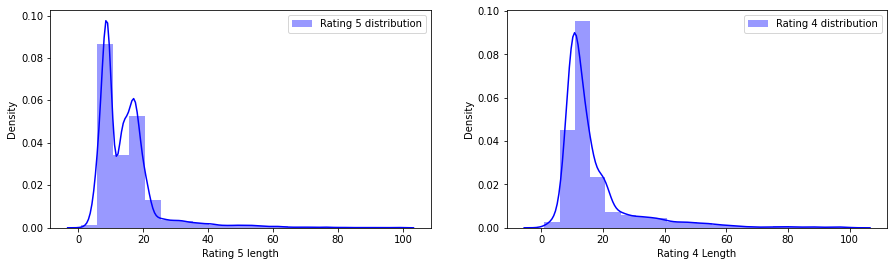
* *Now, Applied operation to remove stopwords*

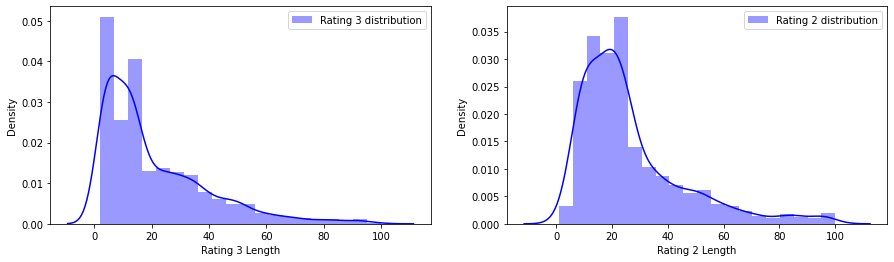
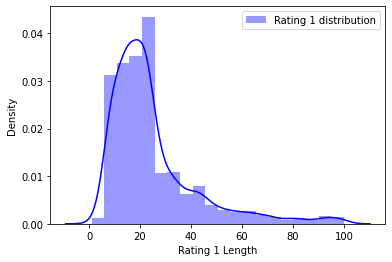




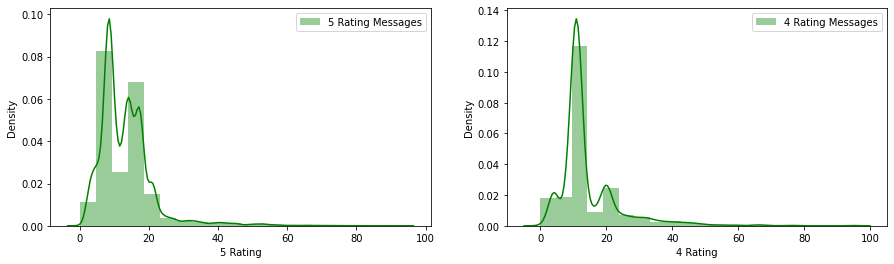
# By finding this Len of clean Review and Text we can see the content weightage, how much data cleaned by processing various operation to convert into this desired form.

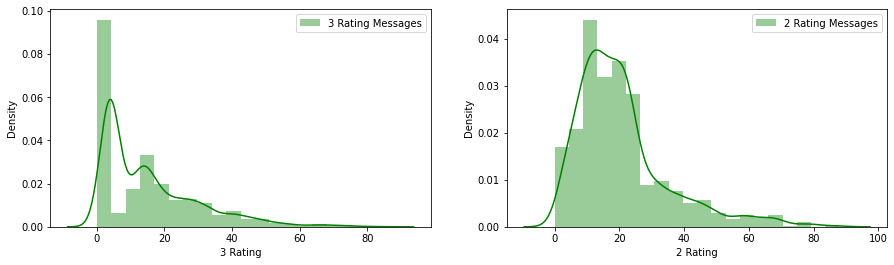
* ***Review text Distribution before cleaning***

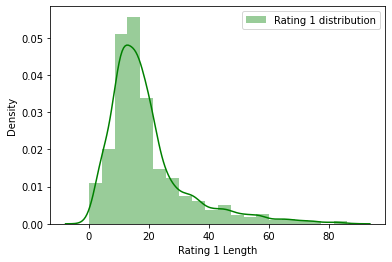


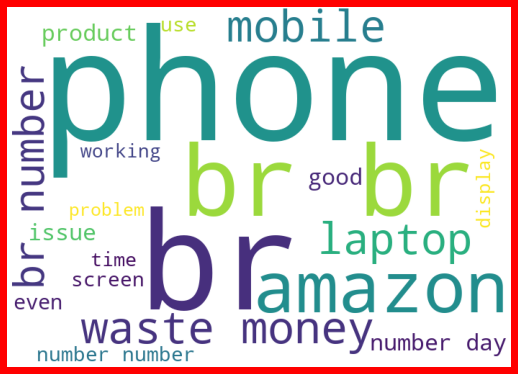
* ***Review Distribution after cleaning***



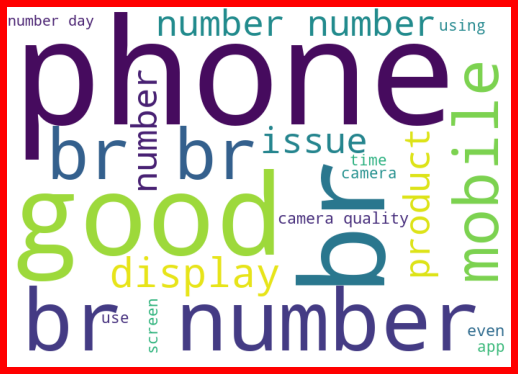




* ***Visualization*:**
* **Word Cloud for Rating 1:**

****

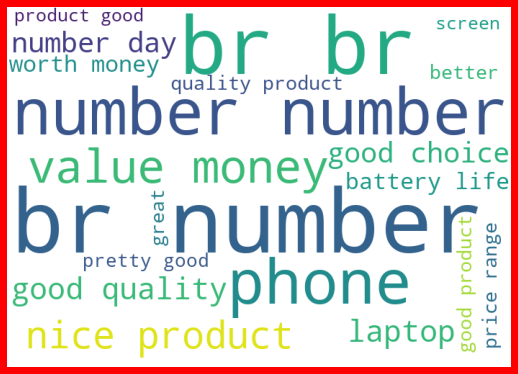
* **Word Cloud for Rating 2:**

****

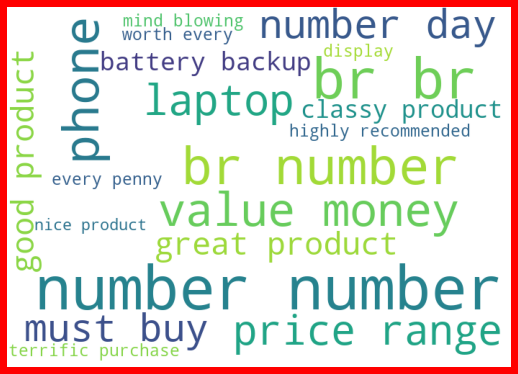
* **Word Cloud for Rating 3:**

****

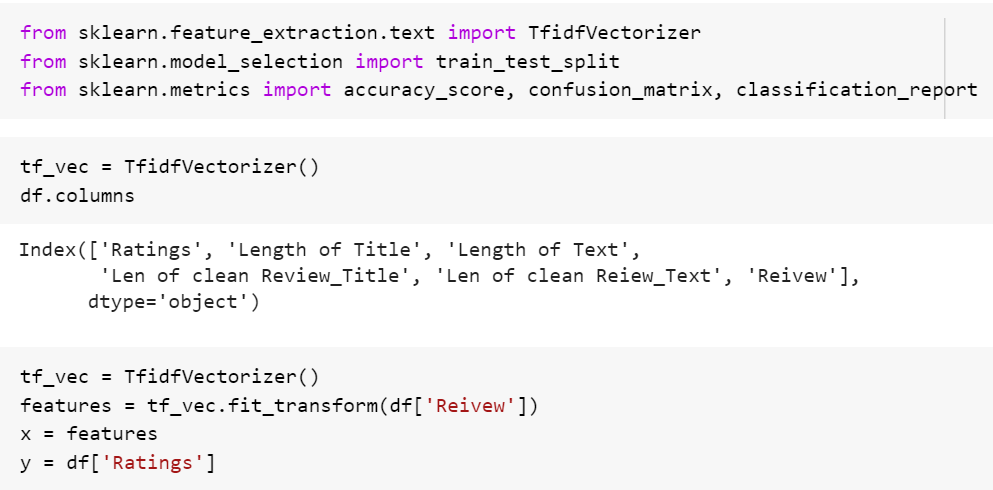
* **Word Cloud for Rating 4:**

****

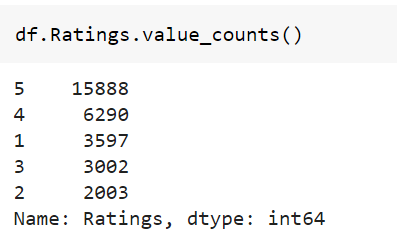
* **Word Cloud for Rating 5:**

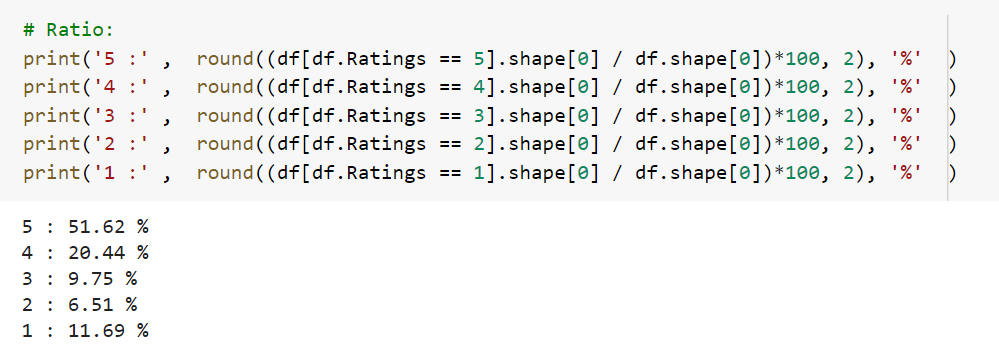
****

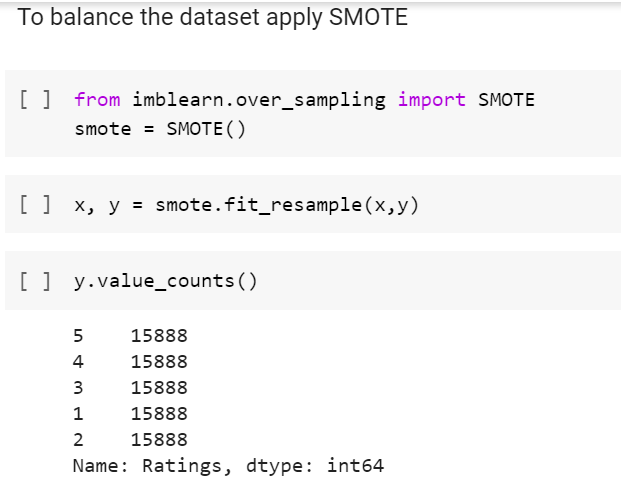
* **Apply TfidVectorizer to the Combined Review column**

****

* **Dataset is highly imbalanced applied SMOTE technique to overcome from it:**

****

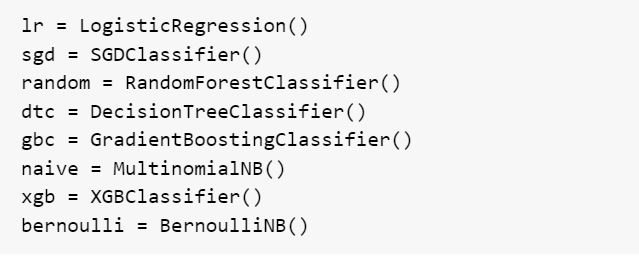
****



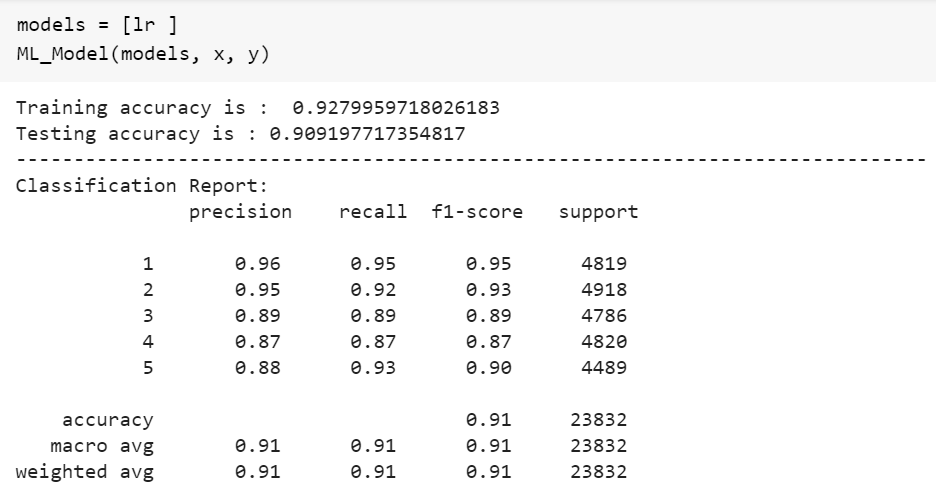
* **Machine Learning:**

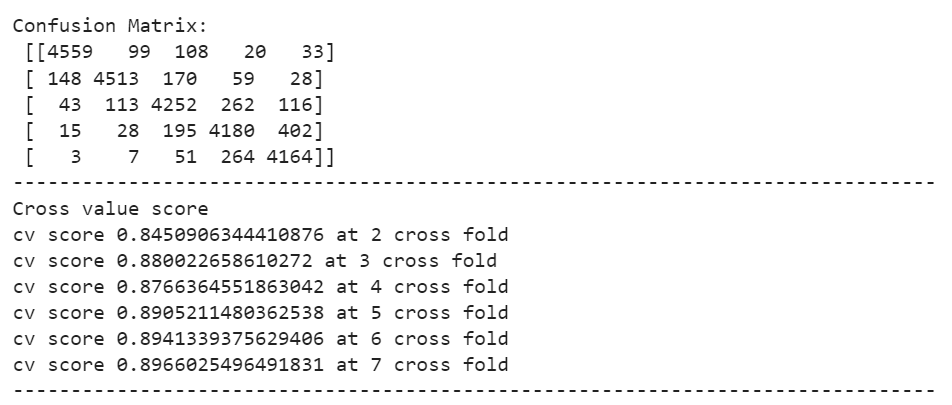


Note: ML\_Model is a function which give accuracy and metrics with CV values at difference cross fold.

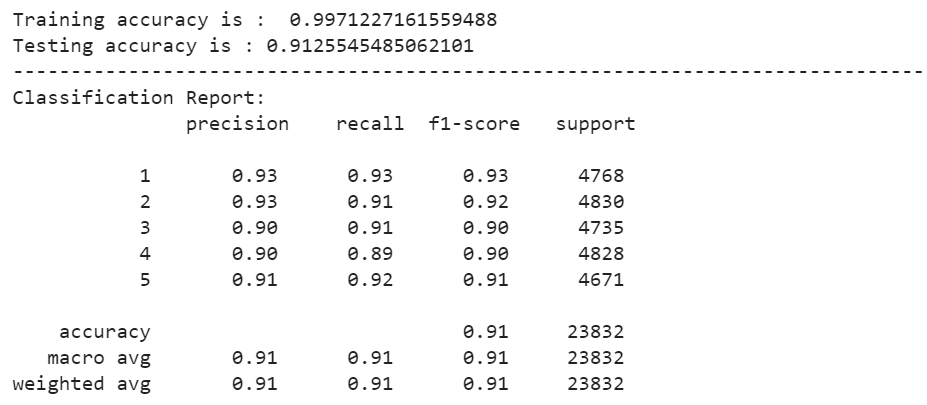


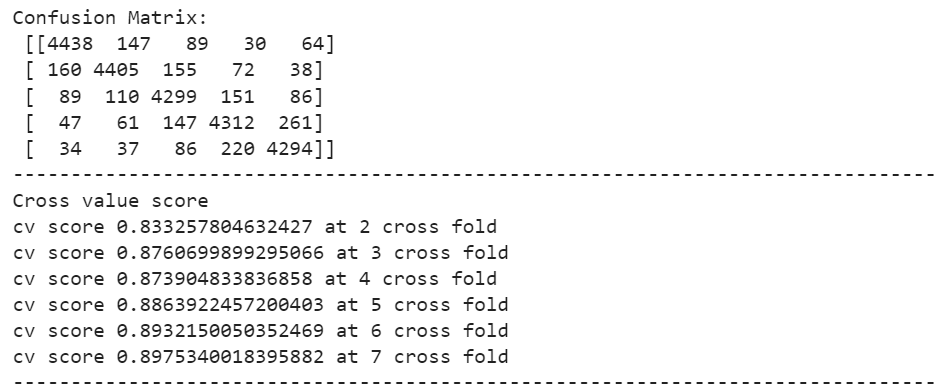
**# Linear Regression Model Performance:**

****

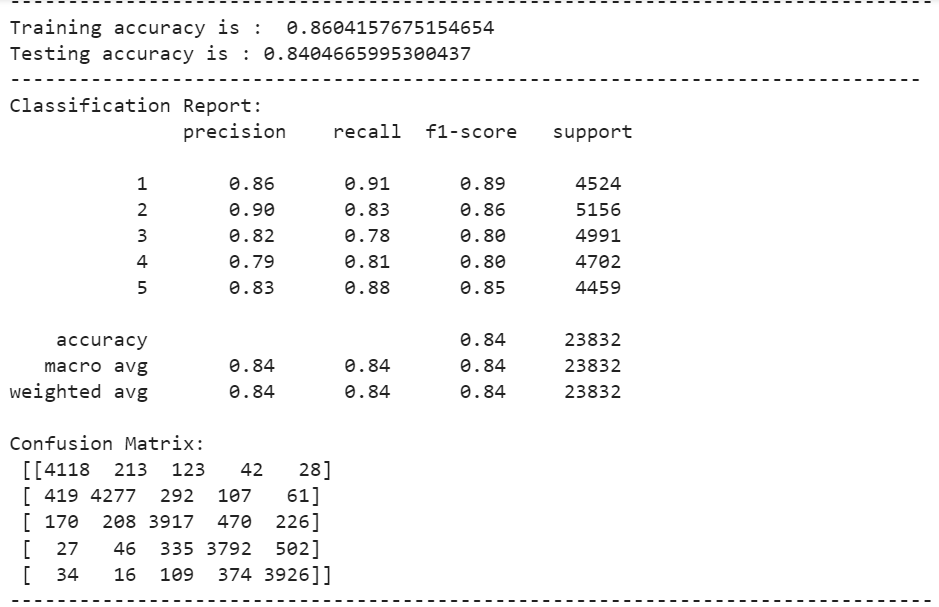
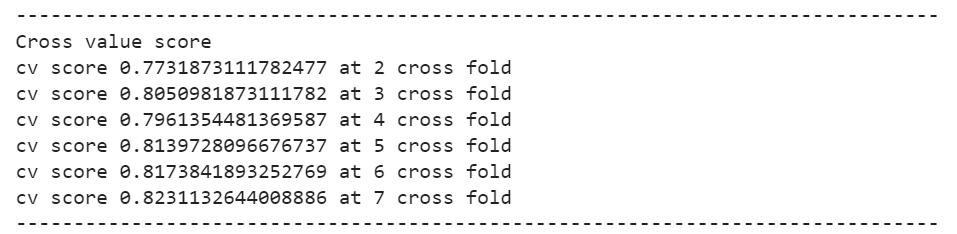
****

**# DecisionTreeClassifier:**

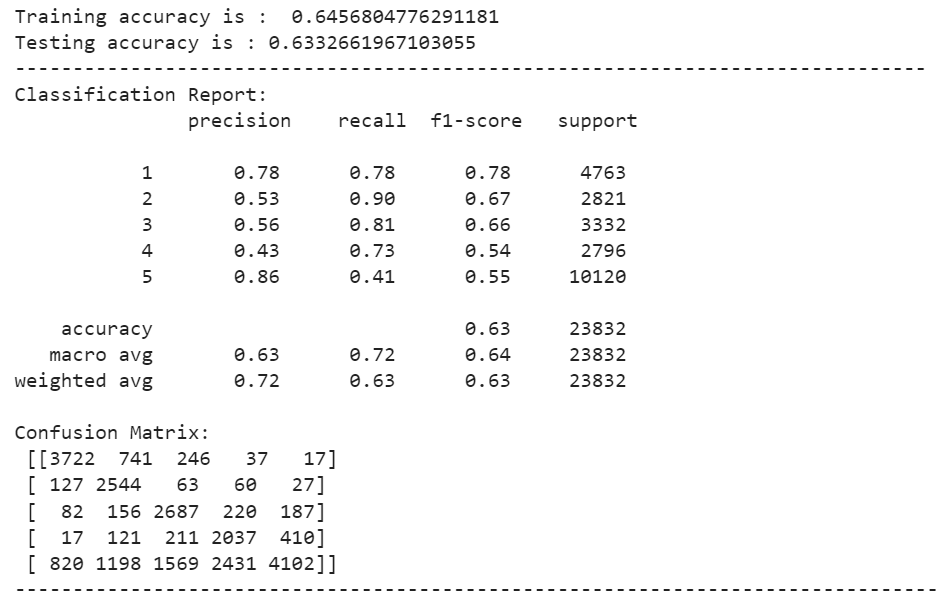
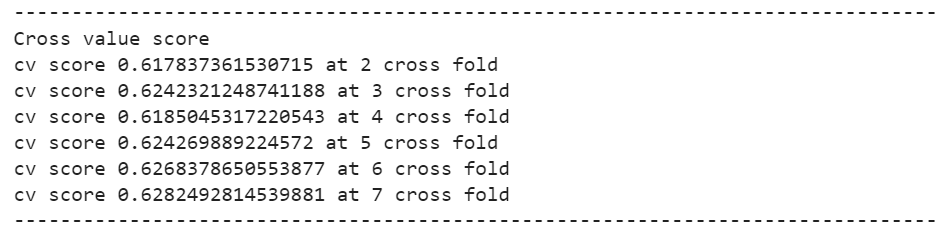
****

****

**# Naïve Bayes**

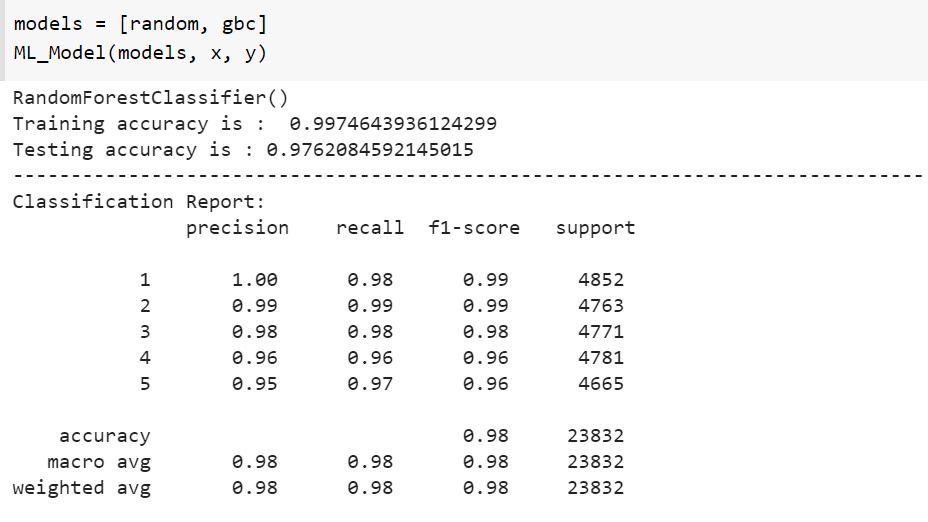
**** ****

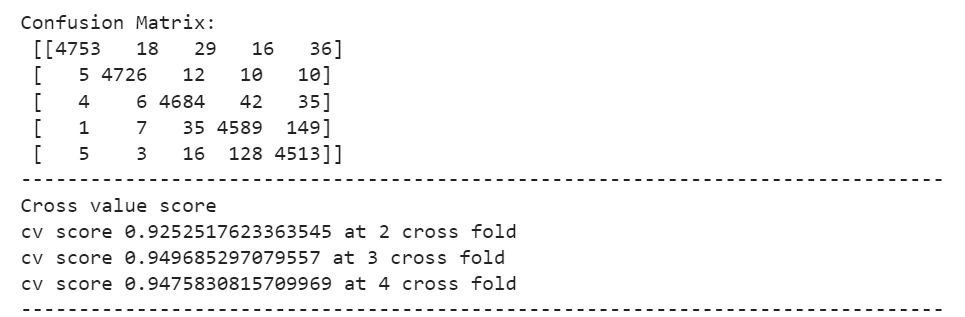
**# Bernoulli**

**** ****

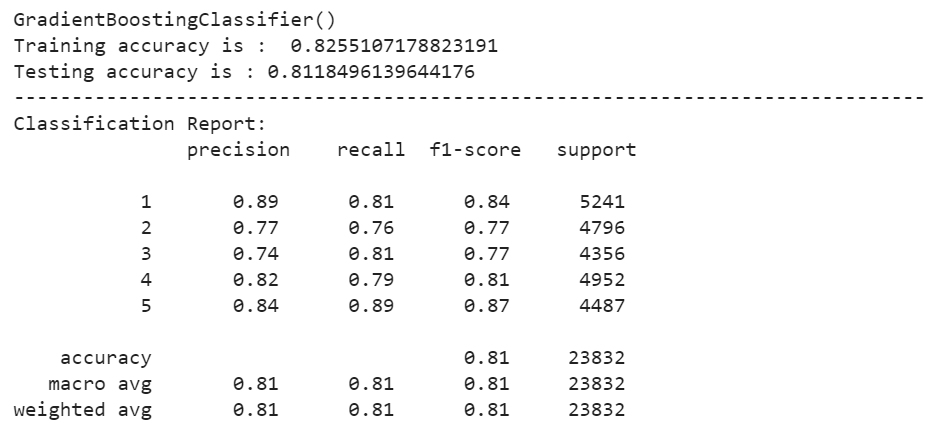
**# Applied Boosting Techniques:**

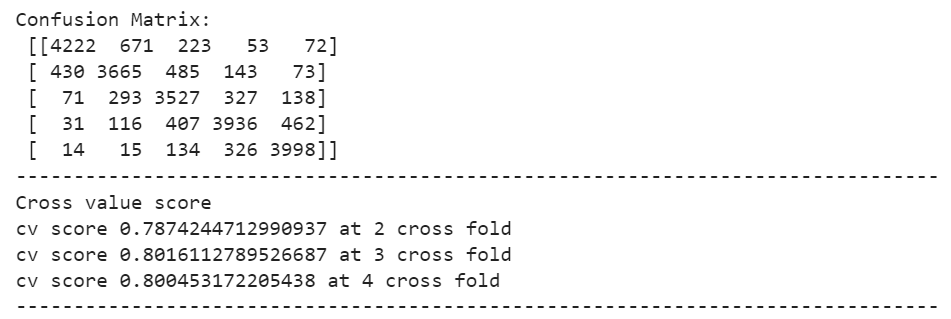
**# RandomForestClassifier**

****

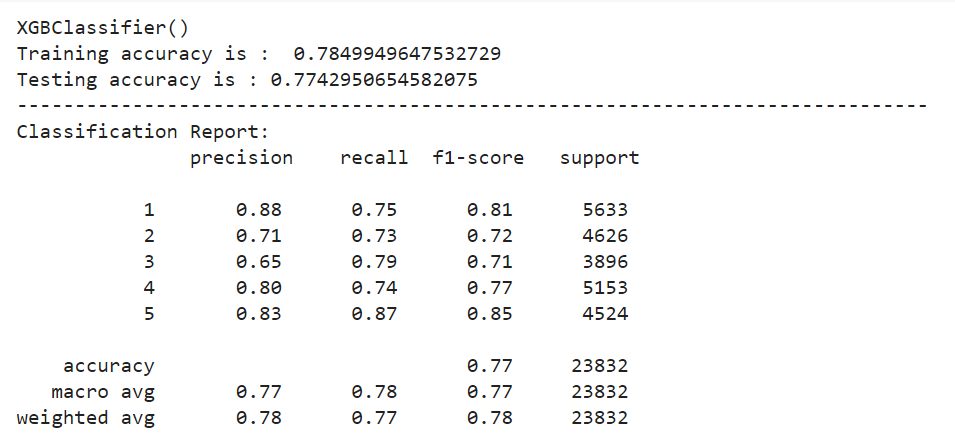
****

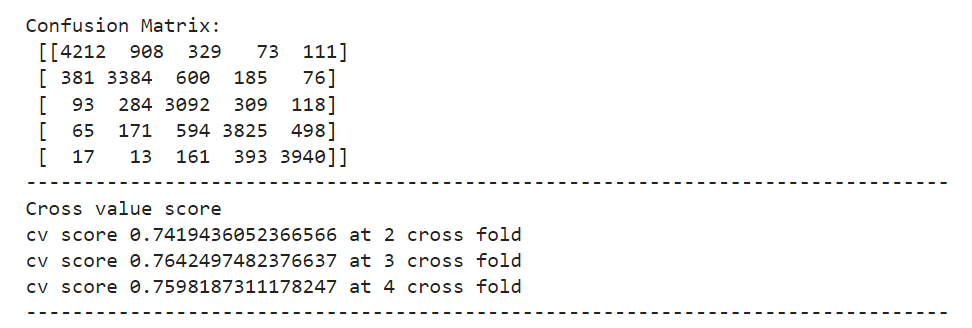
**# GradientBoostingClassifier**

****

****

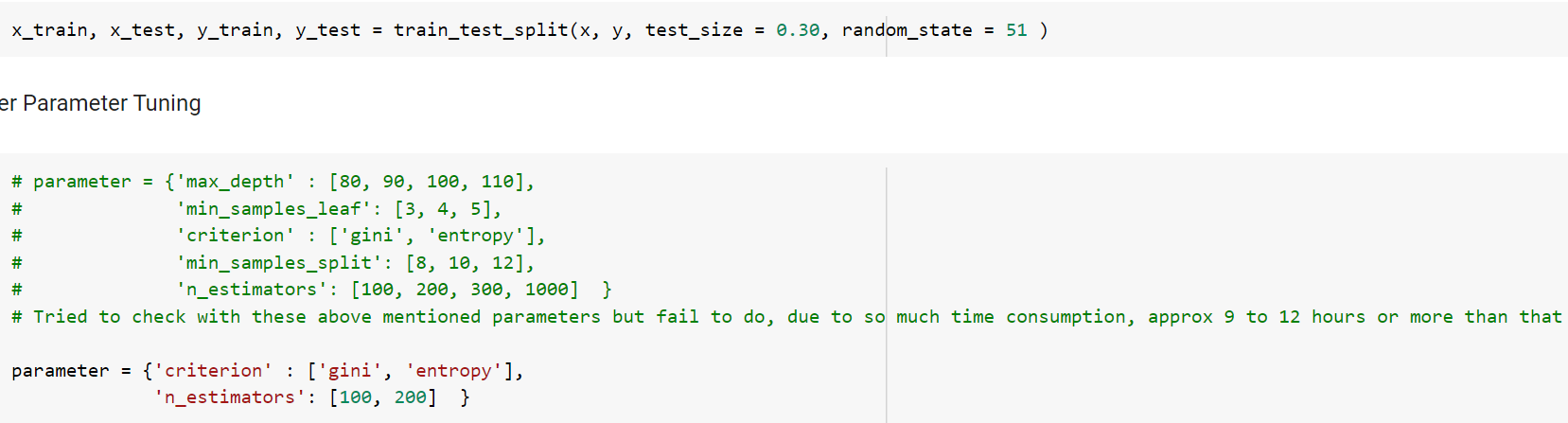
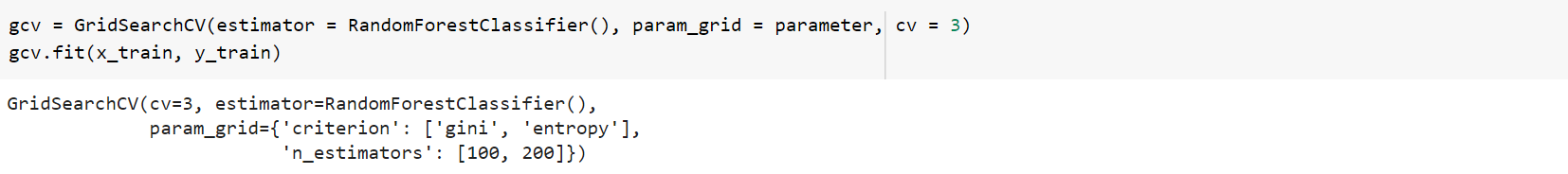
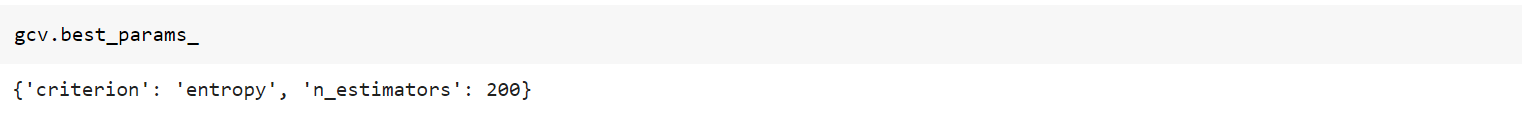
**#XGB Boost:**

****

****

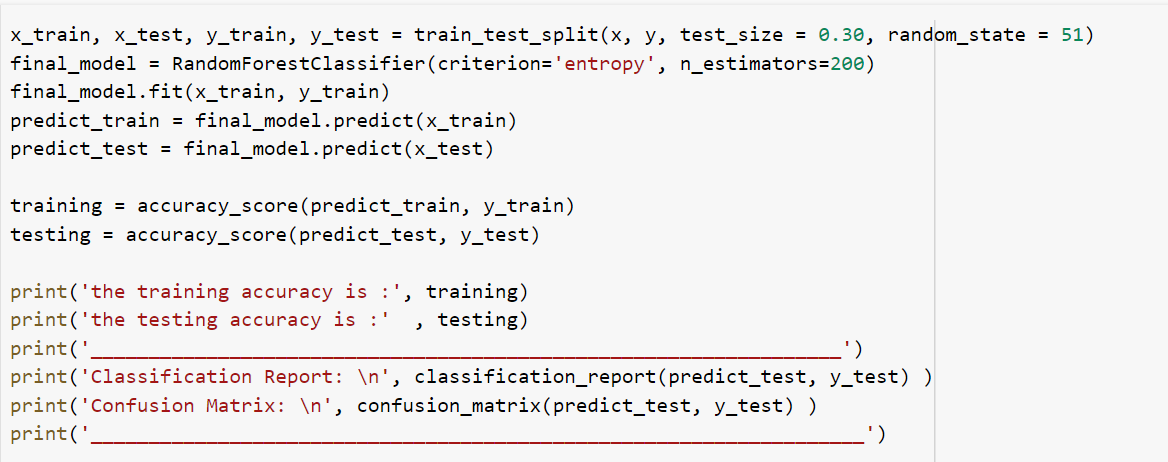
**Note:** By observation various outputs of the difference machine learning algorithm, random forest is giving best result, In random forest training and testing accuracy is also very close to each other and also having very close value of CV, therefore randomforest algorithm selected as final machine learning algorithm to train the dataset for final model

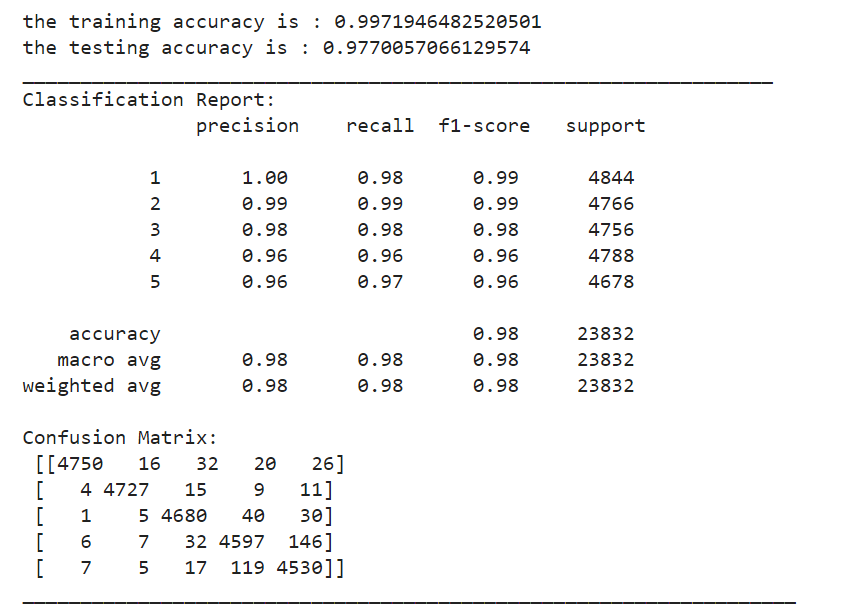
* **Ensemble Technique of RandomForestClassifer:**

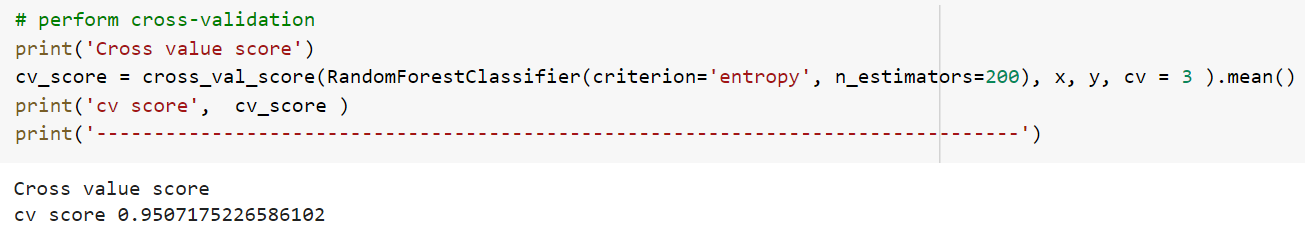
**** **** 

**Note**: I have applied ensemble technique by using various parameters but it was taking so much time, approx 18 to 20 hours, due to which I have reduced parameters.

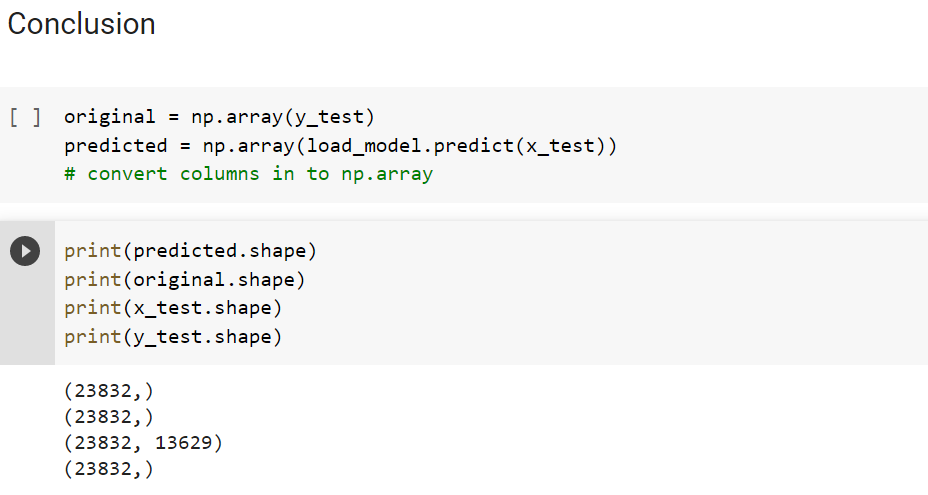
* **Final Model (RandomForestClassifer)**

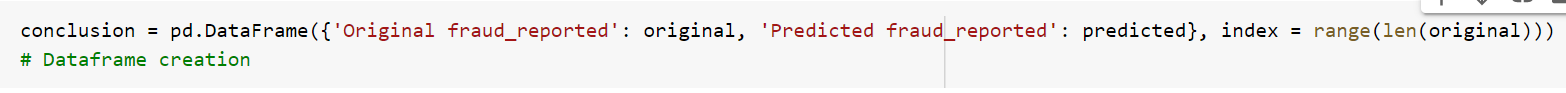
****

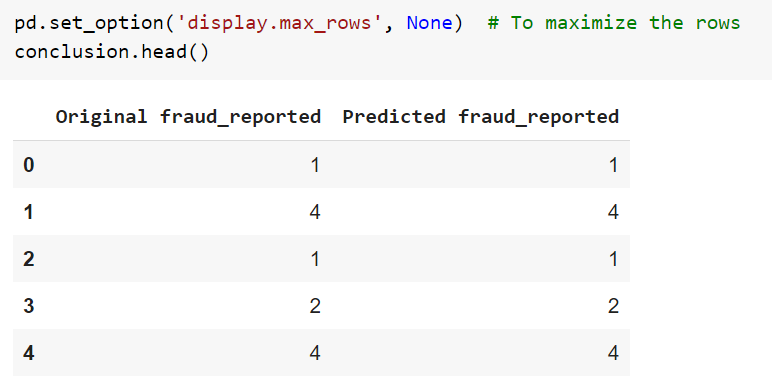
****

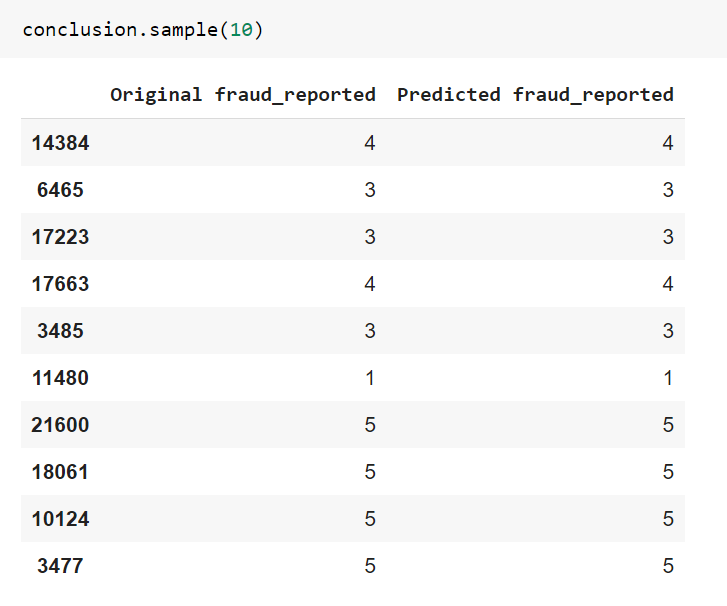
****

****

****

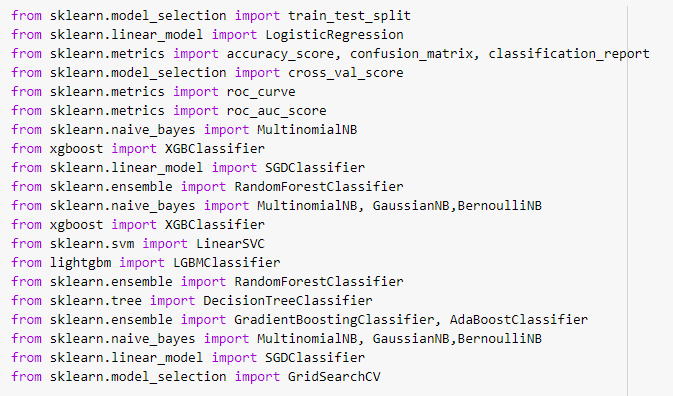
****

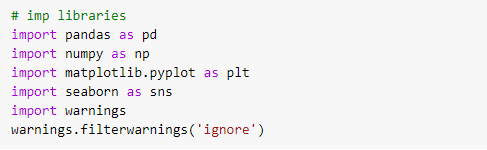
****

****

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

All used libraries:





**Pandas**: This library used for dataframe operations

**Numpy**: This library gives statistical computation for smooth functioning

**Matplotlib**: Used for visualization

**Seaborn**: This library is also used for visualization

**Sklearn**: This library having so many machine learning module and we can import them from this library

**Pickle**: This is used for deploying the model

**Imblearn**: This library is import to get SMOTE technique for balance the data

**Scipy**: It is import to perform outlier removing technique using zscore

**Warning**: To avoid unwanted warning shows in the output

I am giving this requirement and tool used, based on my laptop configuration.

**Operating System: Window 11**

**RAM: 8 GB**

**Processor: i5 10th Generation**

**Software: Jupyter Notebook,**

* **Observations from the whole problem.**

1. Review column is combine column of Review\_Title and Review\_Text.
2. Dataset was having null values.
3. Tfvectorizer applied to review column to convert this column into machine learning language.

* **Learning Outcomes of the Study in respect of Data Science**

My learnings: - the power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say,

Various algorithms I have used in this dataset and to get out best result and save that model. The best algorithm is RandomForestClassifier.

Ensemble operation was giving biggest challenge which I have faced while working and as this dataset is very large which have leads to take lot of time for machine learning.

# for scraping the data, it had taken almost 7 to 8 hours’ time for execution.

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

We can train machine learning model with more data, which should be in lacs but this model is also working well.