

Review’s Rating Prediction Project

Submitted by:

Aditya Maurya

**ACKNOWLEDGMENT**

I would like to appreciate FlipRobo for giving me the opportunity to work on such exciting project. It helps us broaden our portfolio and give us the chance to work on different fields of machine learning.

I am grateful for DataTrained which give us the knowledge about Datascience and Machine Learning without which it will be nearly impossible to deal with such Projects. Their technical team really help us to understand every concept and support us with our doubts.

Last but not the least I would like to show my gratitude towards our SME Mohd Kashif , for helping me in every possible way.

**INTRODUCTION**

* Business Problem Framing

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

* Conceptual Background of the Domain Problem:

For this purpose, we need to first collect data, then by using that data we can built the NLP program around. So, our Project basically consist of two parts:

1. **Data Collection Phase**

You have to scrape at least 20000 rows of data. You can scrape more data as well, it’s up to you. more the data better the model.

In this section you need to scrape the reviews of different laptops, Phones, Headphones, smart watches, Professional Cameras, Printers, Monitors, Home theater, Router from different e-commerce websites.

Basically, we need these columns-

1) reviews of the product.

2) rating of the product.

You can fetch other data as well, if you think data can be useful or can help in the project. It completely depends on your imagination or assumption.

1. **Model Building Phase**

After collecting the data, you need to build a machine learning model. Before model building do all data preprocessing steps involving NLP. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like-

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Preprocessing

4. Model Building

5. Model Evaluation

6. Selecting the best model

* Motivation for the Problem Undertaken

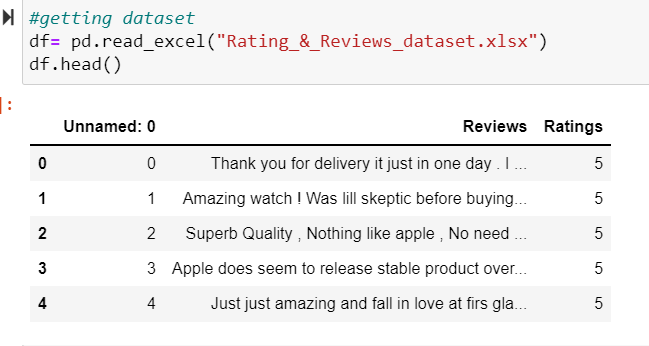
My object is to Predict the rating(1-5) for the given reviews. For that we collect sample data around 30K, which seems to be our training data, on the basis of which entire machine learning model is built.

**Analytical Problem Framing**

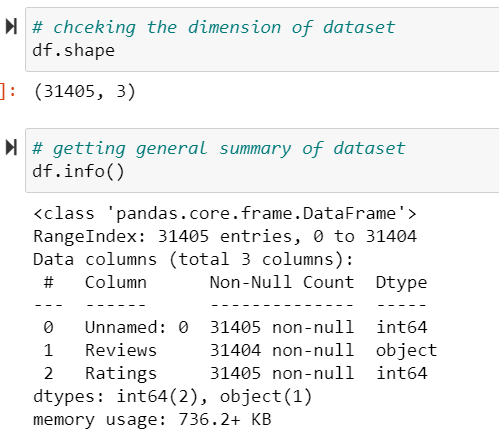
* Mathematical/ Analytical Modeling of the Problem

To start with our Project we prepared a dataset by using WEB SCRAPING, we scrape reviews and ratings of electronics product from e commerce websites, in our case it was flipkart.com.

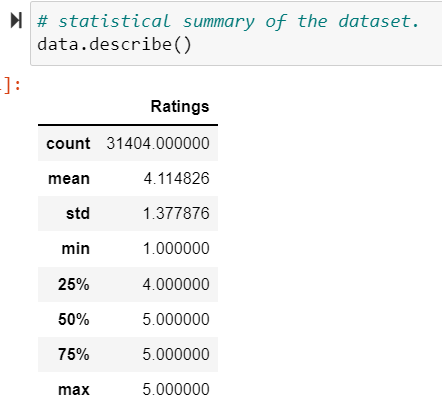
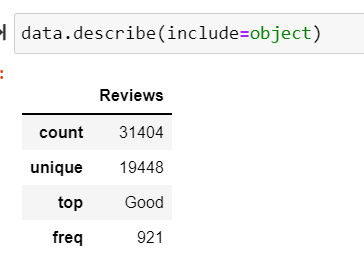
We scrape data and save it in the form of excel sheet, that data is used to build the NLP model.



We perform the dimension analysis, get the detailed summary of the dataset using .info()



This helps us to understand about our dataset, its datatype, number of non-null values, the memory space required by the dataset etc. For statistical analysis, we use .describe()

From above observation we can see that our dataset label is not balanced, as the mean value of the entire dataset is 4. As we Know in rating system, the values are normally near 3 as mean considered to be balanced. Also we notice that certain comments/reviews also kind of repeating. It may be those kind of reviews where user mostly don't elaborate their view points and tired to write using some common words which explains not much about product. We can either keep the duplicates and make our entire model or drop them. But dropping them may result in data loss, while keeping them may result in discrepancy in the accuracy of the model.

We will decide about it later.

* Data Sources and their formats

We collect the data from flipkart.com, using WEB SCRAPING selenium library mainly. Here we collect/scrape basically two things, first is reviews comment and second is their corresponding ratings.

Where review’s is a categorical column of object datatype while rating is our label column of numerical data type.

We store them in the excel sheet, and that excel sheet is used to built the machine leaning/ NLP model.

When we import the excel sheet for model building purpose, we can see that it has 3 columns named, Unnnamed: 0, Reviews and Rating.

The Unnamed: 0 column is merely an index column which we will drop while proceeding further in the model building phase.

* Data Preprocessing Done

For Data Preprocessing, as it a NLP problem, i.e. here our feature column is pure text column, we will perform all the process required to convert this natural language into meaningful machine understandable form.

We perform following steps for data pre processing.

1. Data Cleaning: Here we drop any duplicates present in the dataset, so that we have all unique values in the data, it also helps in keeping the accuracy of the model unbiased and at par.
2. Converting original text to all lower case, then removing or replacing, email address, web address, symbols, punctuation, white spaces, trails etc.
3. Then we remove all the stop words from the text.
4. Then we perform lemmatization/ stemming. (Lemmatizing is the process of grouping together the inflected forms of a word so they can be analysed as a single item.)
5. We did check the original length of the text and also the clean length of the text, this helps us to get clear picture of how effectively our data is pre processed.
6. For making data ready for model building we perform vectorization which is similar to encoding in machine learning.

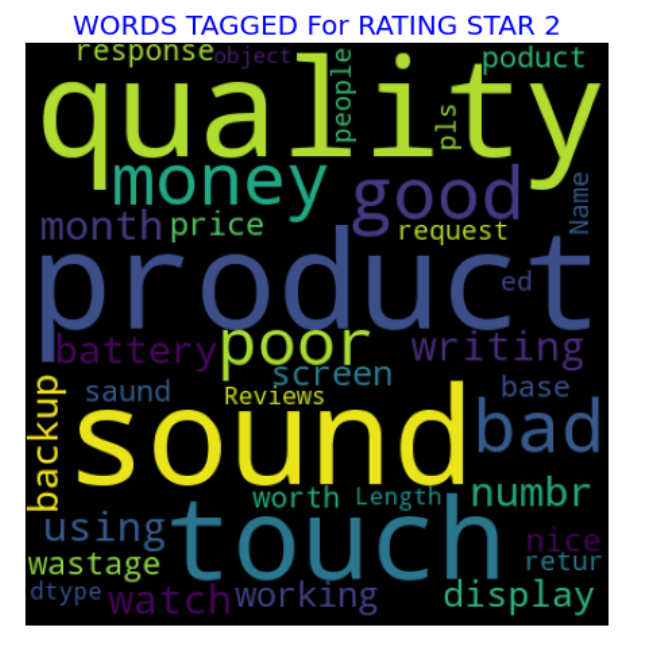
* Data Inputs- Logic- Output Relationships

Here the logical input is the pre processed text, while the logical output is the rating.

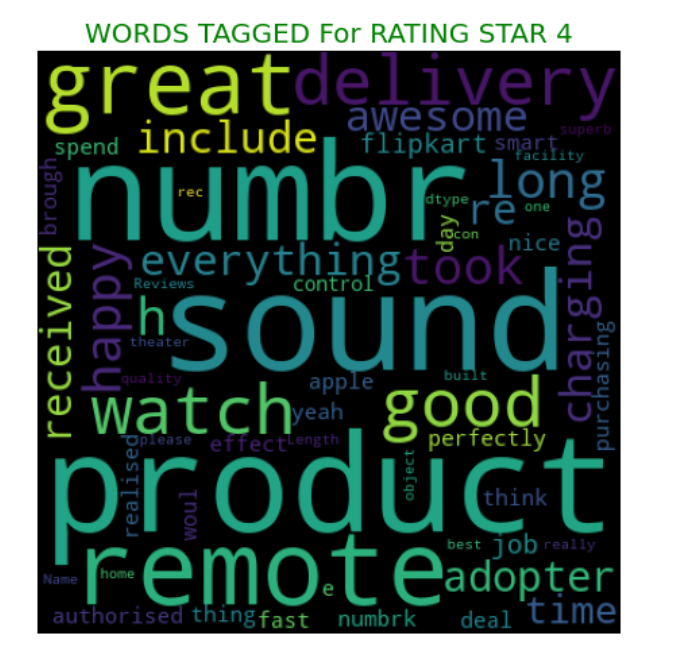
We plot wordcloud to shows this relation, here we got the label which has 5 classes (1,2, 3, 4, 5) where 1 denotes extremely low star rating while 5 denotes extremely high star rating. So we tries to get the words used by user while writing the reviews for different ratings.

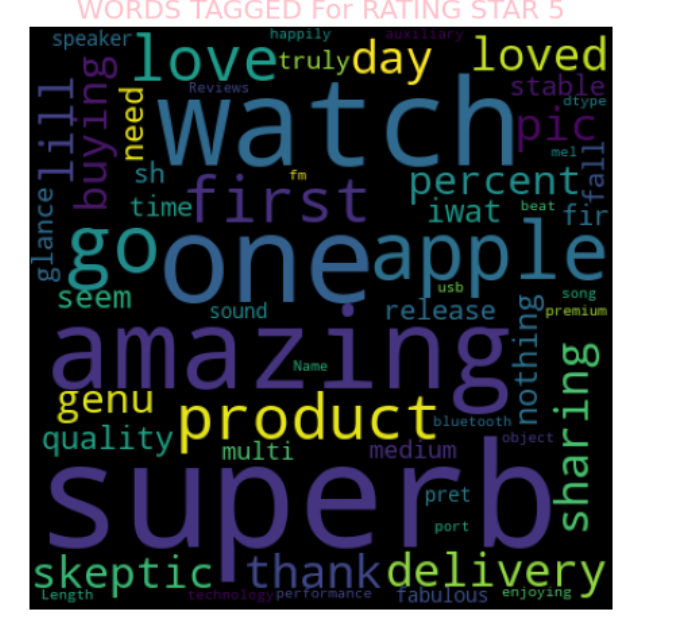


These words in the word cloud are the indicators for corresponding ratings.









From above plots We can see that these words are most occurred word in the reviews from rating 1-5. These words define the rating as per NLP.

* State the set of assumptions (if any) related to the problem under consideration

Here, we assume that Reviews is our feature column on the basis of which we decides about our multiclass target column Rating.

Rating has 5 classes, 1 to 5. We are not going to encode them, we will keep them as it is in our model building.

* Hardware and Software Requirements and Tools Used.

For Data Collection we use following tools/libraries:

1. SELENIUM
2. BEAUTIFULSOUP
3. PANDAS
4. PYTHON

For Model Building we use following tools/Libraries :

1. Python
2. Pandas
3. Numpy
4. Matplotlib.pyplot
5. Seaborn
6. Warnings

NLP libraries:

1. NLTK
2. Re
3. Wordnet
4. Wordcloud
5. Stopwords
6. Ipython
7. Punkt
8. TfidfVectorizer
9. WordNetLemmatizer

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We analyse the data and also do some statistical summarization, then do some data cleaning, data pre processing , encoding, splitting data into train and test dataset using train\_test\_split().

With this pre processed encoded data we starts to train our model using different algorithms we also did some cross validation of each model to cross check the approach and also efficiency.

* Testing of Identified Approaches (Algorithms)

Our is a classification NLP models, so applied different classification algorithm to train and test the model.

Here is the list of all the algorithms used for the training and testing purpose:

1. Logistic regression
2. MultinomialNB
3. Decision Tree Classifier
4. Random Forest Classifier
5. AdaBoost Classifier
6. XGBoost Classifier
7. Bagging Classifier

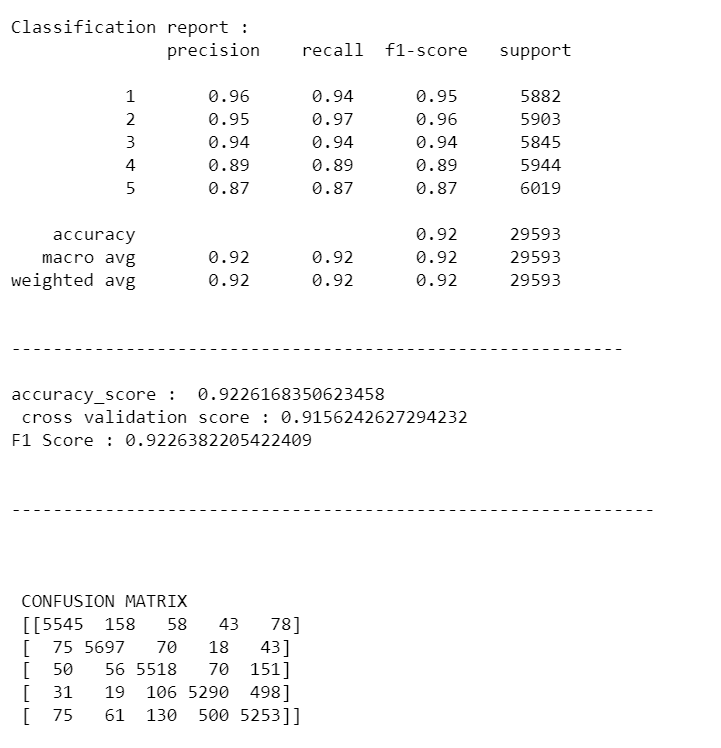
* Run and Evaluate selected models

From all the above approaches, we finalized RandomForest Classifier as the best fitted model for our dataset.

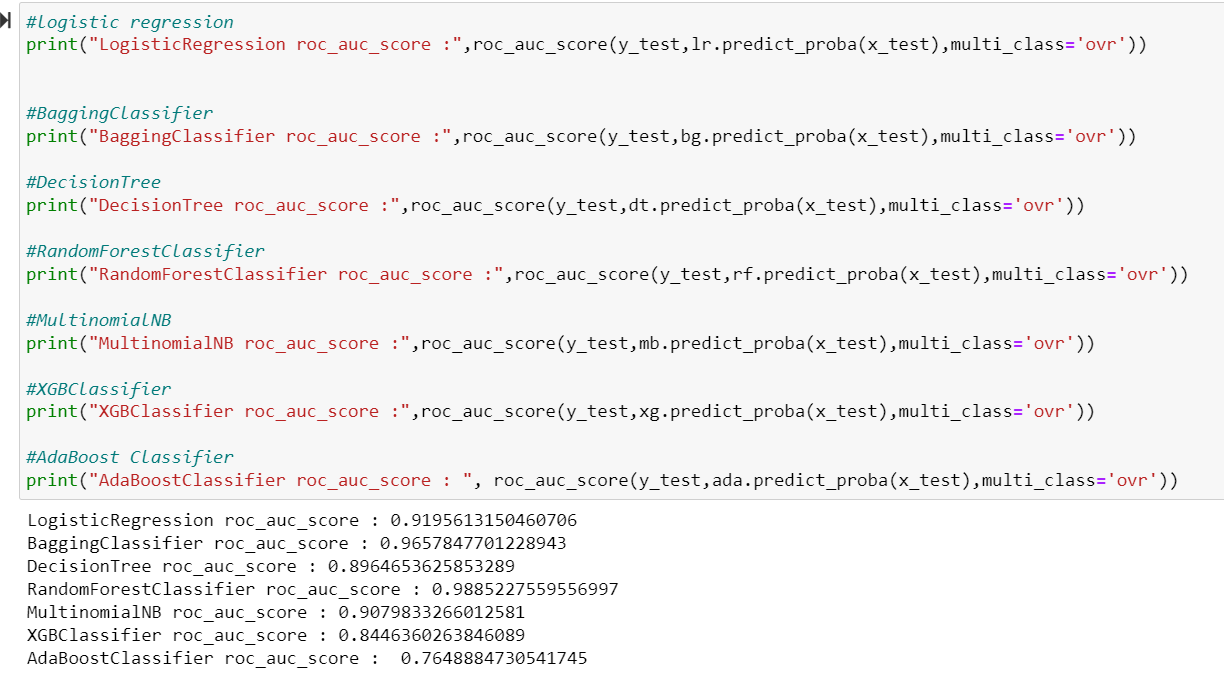


This the code which we use , first we call the model, then fit the data, and then do the prediction.

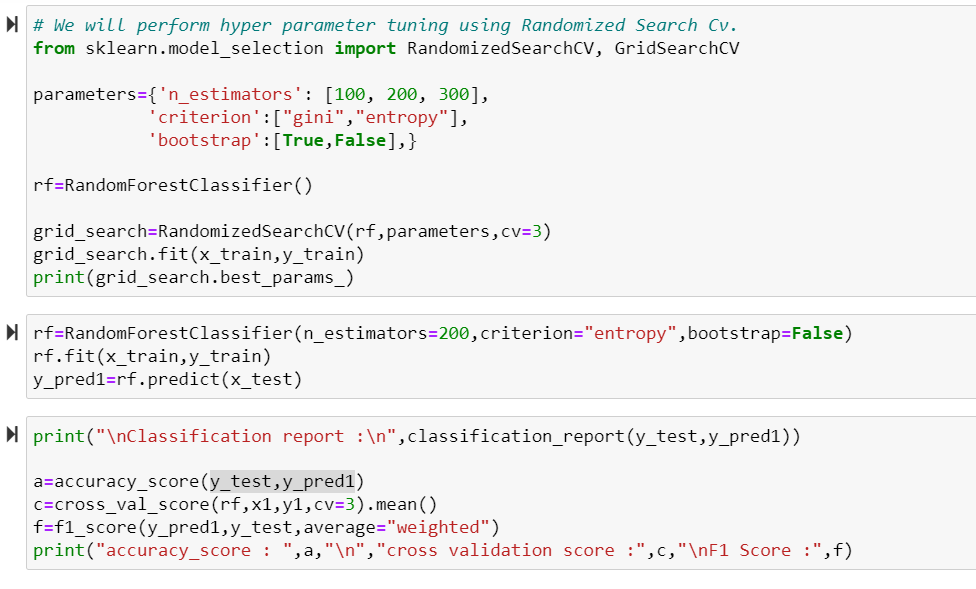
We calculate the classification Report, as it shows clear picture of how our model is predicting the respective output. This will let us observe the individual class f1 score, recall and precision. As well as give as weighted accuracy and f1 score. In short we receive complete report card of our algorithm. Below are the results, how our model performs. We also calculate cross validation score here only.

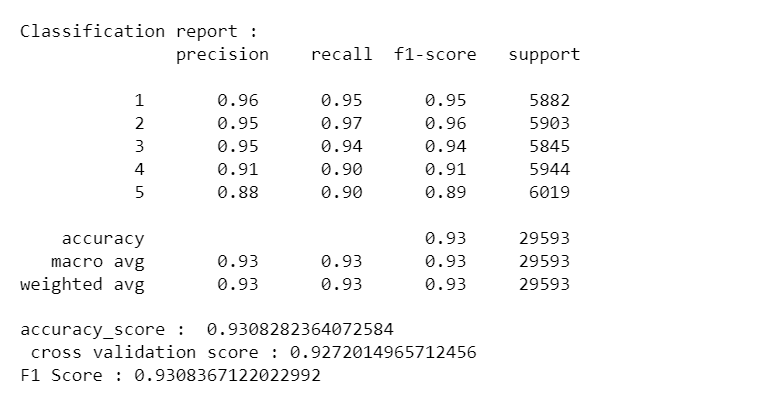


We also calculate the roc\_auc\_score for every single model.



It is clear that RandomForest Model outperform rest of them. We did hyper parameter tuning for randomforest to see if we can increase the performance of the algorithm.





We can see there is slight increase in both the accuracy and f1 score of the model. Even the cross validation score increases. We perform RandomizedSearchCV for hyper parameter tuning.

* Key Metrics for success in solving problem under consideration:

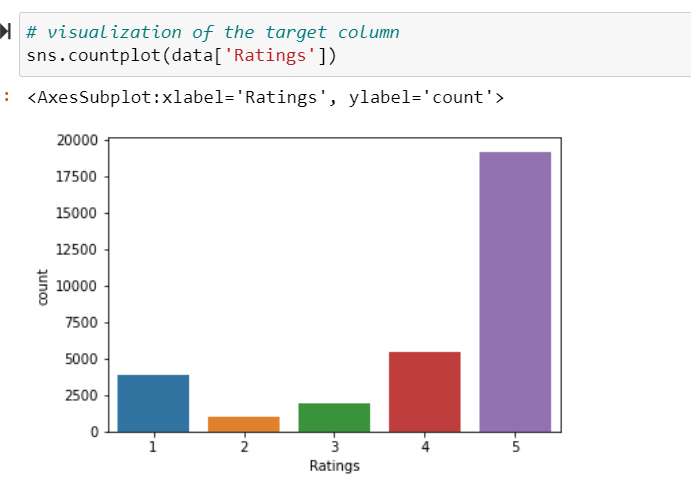
We consider f1 score as our primary key metrics while accuracy score and roc\_auc\_score is considered to be secondary key metrics.

We decide our model which is having high f1 score and high roc\_auc\_score.

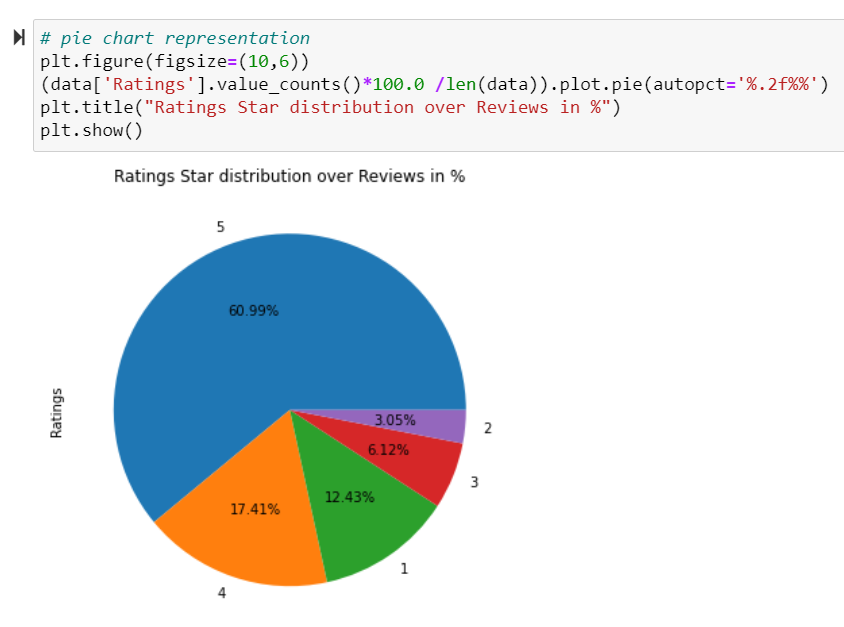
Also we tries to find the difference between cross validation score and accuracy, the model with minimum difference is considered to be best fitted.

* Visualizations:

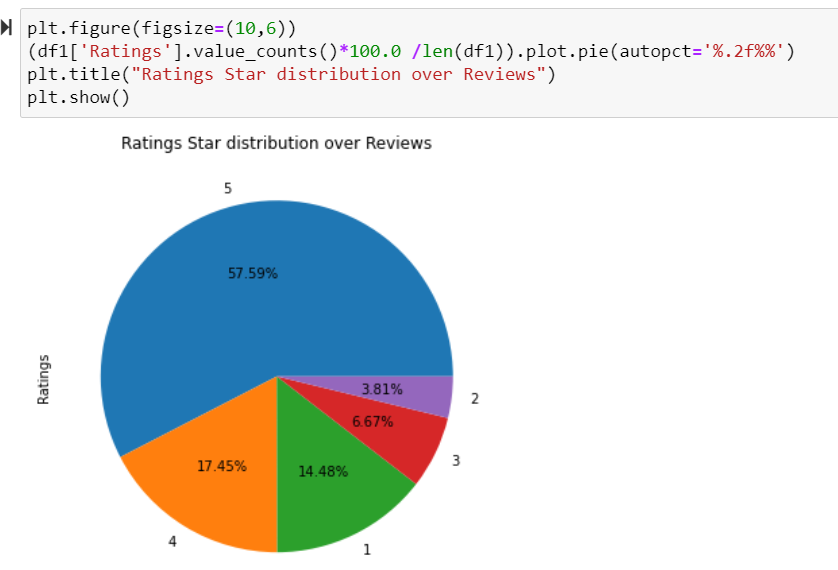
As our dataset consist of merely two columns, in which one is purely text column (reviews column), and other is the target column (rating), we tries to get some visual analysis of the only numerical column.



Above plot shows us the distribution of ratings in our dataset. We can observe that 5 rating reviews are higher in the dataset while 2 star rating reviews are very minimum. We can clearly analyse the imbalance the dataset, so before building the model we need to perform SMOTE technique in order to get balanced dataset. As we don’t want to compensate on the efficiency of the model. Let’s see percentage distribution of various ratings in the entire dataset.



From here also our analysis confirms, that most of the dataset consist of 5 star rating reviews while bare minimum has 2 star, so balancing this before model building is necessity.

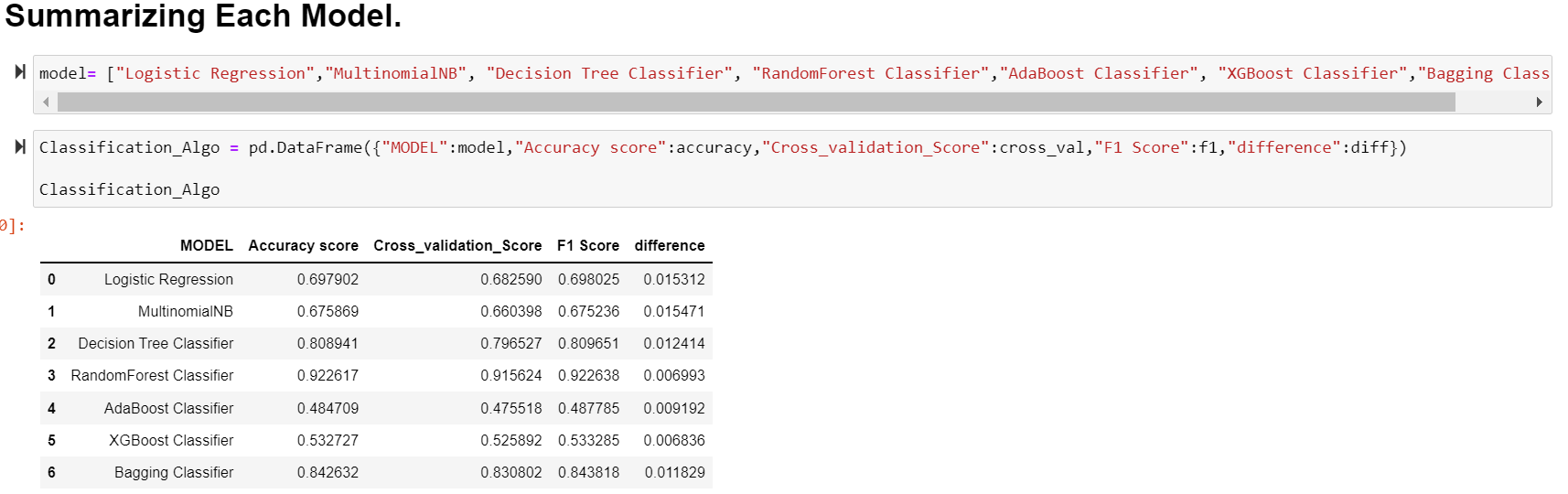


While cleaning the data, we observe that we got certain duplicates in our dataset, so we drop those duplicates and got the final distribution of data in the dataset. There is slight improvement in the distribution pattern, still not the best so balancing the dataset for training and testing model is required.

Other Wordcloud plots are already discussed in the report.

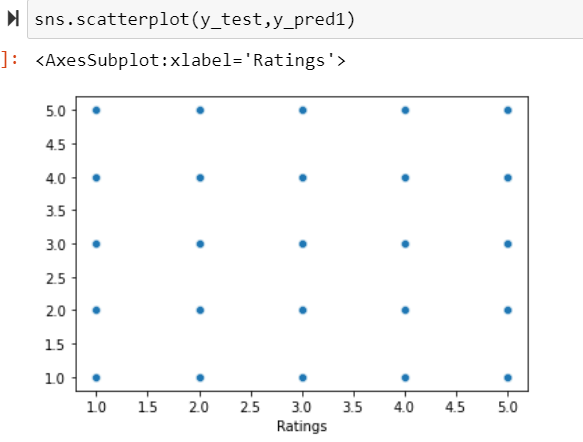
* Interpretation of the Results:

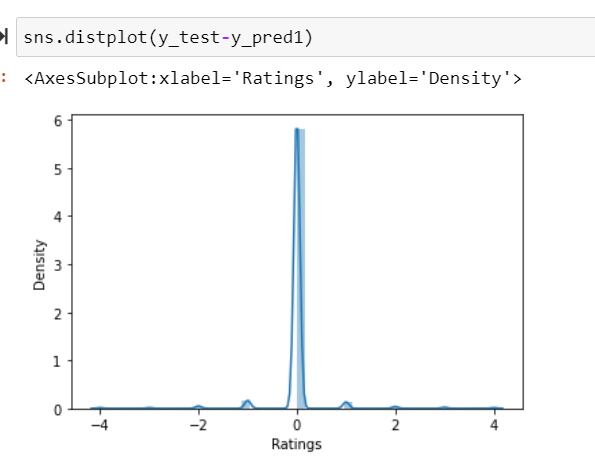
Before finalizing the final model, we look at the summary of each model and here is the result.



From above we can clearly observe that RandomForest Classifier is the best fitted model for our dataset, it has high value for both f1 score and cross validation score. Also the difference between accuracy and cross validation is very less.

Also we can see the roc\_auc\_score of Random Forest and it is confirm that it is the best fitted model for our dataset.

****

****

From above plots it is clear that, RandomForest Classifier algorithm is the best fitted Algorithm for this data set, the predicted value and true values very well align.

**CONCLUSION**

* Key Findings and Conclusions of the Study

While collecting the data, we observe that more than 50% of reviews are for 5-star rating, we can get decent number of reviews for 4 star and 1 star rating but 3 and 2 star rating reviews are very less in number.

It looks like customer normally don’t reviews for average products, they either give high ratings or for poor performance product they give straight 1, no 2 star just 1 star rating. And if they are not satisfied nor dissatisfied, they give 3-star rating.

But to level out we perform smote technique so that the training model has sufficient data for 2-star data as well. We did over sampling here before training and testing the data.

There were some one liner reviews as well, but thing is they consist of all the required words which defines their rating.

The Word Cloud is good techniques to high lights maximum defined words that indicates particular rating.

Random Forest classifier outperforms all the other algorithms which we applied in model building process. We decide on f1 score as we know that our data is imbalanced and it gives the harmonic mean of precision and recall. Good f1 score indicates that our model is working really well for all the classes of our target variable(rating), it did not treat in our case 2-star rating as noise. But as proper class for label which is supposedly required.

We also get good roc\_auc\_score for random forest which double check that we are deciding on right algorithm.

* Learning Outcomes of the Study in respect of Data Science

We can see that all though we start with more than 31K data, but while scraping we came across with some duplication, so for final model we clean our dataset by removing duplicate data, which helps in improving the accuracy of the metrics and algorithm.

Rating system is one of the basic features which now a days every e commerce website is using. It helps the customer to make inform decision on the basis of others experienced, it basically serves as the visual approach for lengthy reviews. As well as it helps the service provider to get the statistics of the products, based on the ratings.

We collect reviews and rating of various electronics products, which help us to see the rating system for wider varieties.

NLP is the fundamental part of text mining. It performs the grammatical and semantic structure analysis to understand the sentiments behind the words. Which is the case for our model. We can rate their sentiments using our NLP model.

* Limitations of this work and Scope for Future Work

There is say in machine learning, “Predicting the future is not magic it’s machine learning”

We can always do more, either more data or more analysis. In our case we can collect more data lower ratings as well, which ultimately helps us to predict better.

---------------------------