Ratings Prediction Project

Submitted By:

Shivanchal Asthana

ACKNOWLEDGMENT

It is my sensual gratification to present this report. Working on this project was an incredible experience that will have a tremendous impact on my career. I would like to express my sincere thanks to the company Flip Robo Technologies for a regular follow up and valuable suggestions provided throughout. They always been an origin of spark and direction. I also thank all the respondents who have given their valuable time, views and valid information for this project.

Shivanchal Asthana

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:

Rating prediction is a well-known recommendation task aiming to predict a user's rating for those items which were not rated yet by customers. Predictions are computed from users' explicit feedback i.e., their ratings provided on some items in the past. Another type of feedback are user reviews provided on items which implicitly express users' opinions on items. Recent studies indicate that opinions inferred from users' reviews on items are strong predictors of user's implicit feedback or even ratings and thus, should be utilized in computation. As far as we know, all the recent works on recommendation techniques utilizing opinions inferred from users' reviews are either focused on the item recommendation task or use only the opinion information, completely leaving users' ratings out of consideration. The approach proposed in this project is filling this gap, providing a simple, personalized and scalable rating prediction framework utilizing both ratings provided by users and opinions inferred from their reviews. Experimental results provided on dataset containing user ratings and reviews from the real-world Amazon and Flipkart Product Review Data show the effectiveness of the proposed framework.

- ## Review of Literature:

What is rating?

Rating is a classification or ranking of someone or something based on a comparative assessment of their quality, standard or overall performance. This project is more about exploration, feature engineering and classification that can be done on this data. Since we scrape huge amount of data that includes five stars rating, we can do better data exploration and derive some interesting features using the available columns

The goal of this project is to build an application which can predict the rating by seeing the review. In the long term, this would allow people to better explain and review their purchase with each other in this increasingly digital world.

- ## ·Motivation for the Problem Undertaken:

Every day we come across various products in our lives, on the digital medium we swipe across hundreds of product choices under one category. It will be tedious for the customer to make selection. Here comes 'reviews' where customers who have already got that product leave a rating after using them and brief their experience by giving reviews. As we know ratings can be easily sorted and judged whether a product is good or bad. But when it comes to sentence reviews, we need to read through every line to make sure the review conveys a positive or negative sense. In the era of artificial intelligence, things like that have got easy with the Natural Language Processing (NLP) technology. Therefore, it is important to minimize the number of false positives our model produces, to encourage all constructive conversation. Our model also provides beneficence for the platform hosts as it replaces the need to manually moderate discussions, saving time and resources. Employing a machine learning model to predict ratings promotes easier way to distinguish between products qualities, costs and many other features.

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.

Analytical Problem Framing

- ## Mathematical/ Analytical Modelling of the Problem:

  As per the client's requirement for this rating prediction project I have scraped reviews and ratings from well-known e-commerce sites. This is then saved into CSV format file. Also, I have shared the script for web scraping into the GitHub repository. • Then loaded this data into a data frame and did some of the important natural language processing steps and gone through several EDA steps to analyse the data. After all the necessary steps I have built an NLP ML model to predict the ratings. • In our scrapped dataset our target variable column "Ratings" is a categorical variable i.e., it can be classified as 1, 2, 3, 4 and 5 stars. Therefore, we will be handling this modelling problem as a multi class classification project.

- ## Data sources are provided internally by the enterprise.

This project is done in two parts: ▪ Data Collection Phase ▪ Model Building Phase

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 theatre, router from different e-commerce websites. Basically, we need these columns: 1) reviews of the product. 2) rating of the product. Fetch an equal number of reviews for each rating, for example if you are fetching 10000 reviews then all ratings 1,2,3,4,5 should be 2000. It will balance our data set. Convert all the ratings to their round number as there are only 5 options for rating i.e., 1,2,3,4,5. If a rating is 4.5 convert it 5. Model Building Phase: After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing 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 mentioned below: 1. Data Cleaning 2. Exploratory Data Analysis and Visualization 3. Data Pre-processing 4. Model Building 5. Model Evaluation 6. Selecting the Best classification model We collected the data from difference e-commerce websites like Amazon and Flipkart. The data is scrapped using Web scraping technique and the framework used is Selenium

- ## Data Pre-processing:

- Checked the ratings column and it had 10 values instead of 5 so had to clean it through and ensure that our target label was updated as a numeric datatype instead of the object datatype value. Made sure that the string entries were replaced properly

  Lemmatizing is the process of grouping together the inflected forms of a word so they can be analysed as a single item. This is quite similar to stemming in its working but differs since it depends on correctly identifying the intended part of speech and meaning of a word in a sentence. As well as within the larger context surrounding that sentence such as neighbouring sentences or even an entire document. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

- ## Data Inputs- Logic- Output Relationships:

To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. From this describe function that also known as Five-point summary analysis if there are any outliers are present for a particular column. Also five point summary analysis was done for the target variable to explore & understand the data in a better way.

- ## State the set of assumptions (if any) related to the problem under consideration:

  By looking into the target variable/label, we assumed that it was a multiclass classification type of problem. Also, we observed that our dataset was imbalance so we will have to balance the dataset for better prediction accuracy outcome. The 5-star rating system allows respondents to rank their feedback on a 5- point scale from 1 to 5. The more stars that are selected, the more positively your customer is responding to the purchased products. People tend to overlook businesses with a less than four-star rating or lower. Usually, when people are researching a company, the goal is to find one with the highest overall score and best reviews. Having a five-star rating means a lot for your reputation score and acquiring new customers

- ## Hardware and Software Requirements and Tools Used:

For this particular dataset the Hardware is used Windows as operating system, and the software used are mainly Jupyter notebook for model building and various internal packages that are defined in the anaconda/jupyter notebook.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods):

For this particular project I have used different classification models to predict the outcome of this dataset. After the model implementation Random forest classifier method predicted the best outcome out of all the models in terms of accuracy score and also I have used cross validation to flag the problem related overfitting or selection bias for the dataset and hence we can use this model for further evaluation.

- ## Testing of Identified Approaches (Algorithms):

I have used mainly different classification methods to get the outcome of the house price prediction and 80% data used for training purpose and rest 20% are used for testing the prediction of the accuracy score for this machine learning model building process.

- ## Run and Evaluate selected models:

To predict the result of this dataset below are machine learning models used for evaluations.

Out of all the machine learning models used I have selected Random forest classifier model for further evaluation of this project.

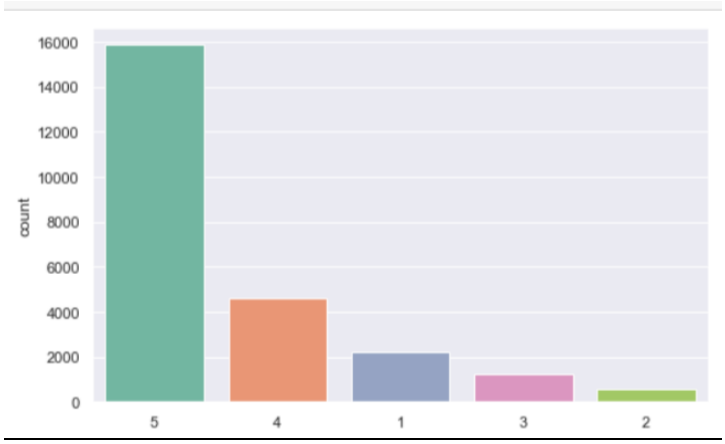
- ## Key Metrics for success in solving problem under consideration

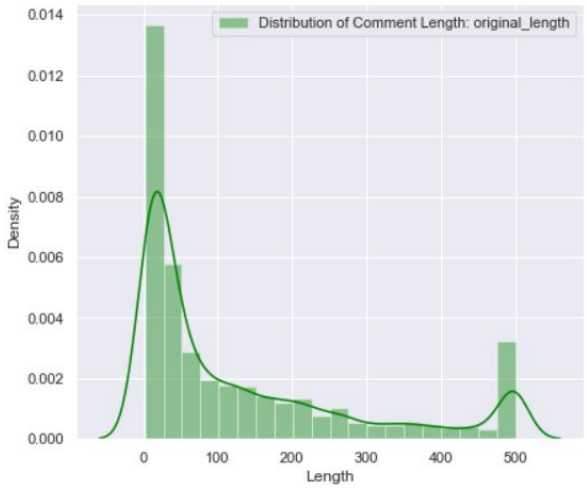
The key metrics that were mainly taken into consideration were the followings:

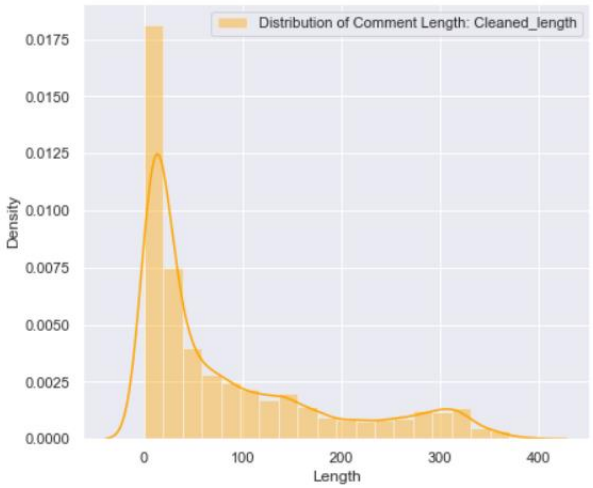
- Stars
- Short review
- Full Review

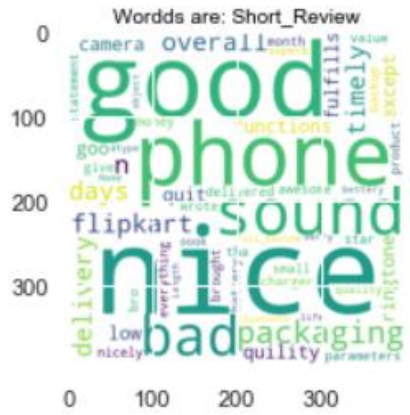
These are all the prime metrics under consideration.

Visualizations:

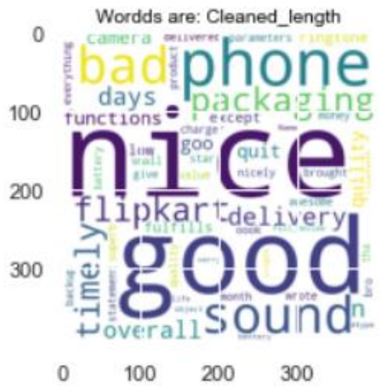




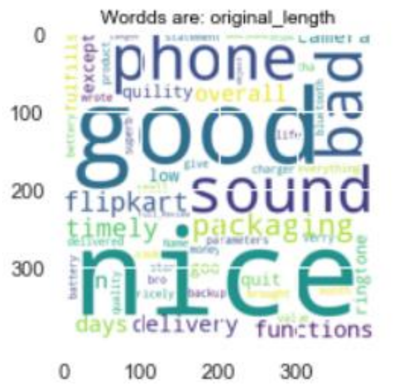



Wordds are: Short_Review
good
phone
sound
nice
bad
packaging
flipkart
delivery
camera
overall
timely
fulfills
except
days
low
nicely
quility
0
100
200
300


Wordds are: Full_Review
bad
phone
delivery
flipkart
nice
packaging
good
sound
timely
overall
functions
ringtone
camera
days
low
0
100
200
300


Wordds are: Cleaned_length
bad
phone
days
packaging
functions
nice
flipkart
delivery
good
sound
timely
overall
camera
0
100
200
300


Wordds are: Full_Review
bad
phone
delivery
flipkart
nice
packaging
good
sound
timely
overall
functions
ringtone
camera
days
low
0
100
200
300


Wordds are: original_length
phone
good
bad
sound
flipkart
timely
packaging
nice
days
delivery
functions
overall
quility
except
low
0
100
200
300

1. We can see here, our columns are imbalanced.
2. We can see, original length and cleaned length both are imbalanced.

ACCURACY SCORE PERCENTAGE: 92.14003541751805
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.96	0.96	0.96	656
2	0.98	0.91	0.94	174
3	0.99	0.68	0.81	367
4	0.97	0.72	0.82	1327
5	0.90	0.99	0.94	4817
accuracy			0.92	7341
macro avg	0.96	0.85	0.90	7341
weighted avg	0.93	0.92	0.92	7341

**********LinearSVC**********
ACCURACY SCORE PERCENTAGE: 93.28429369295736
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.97	0.96	0.96	656
2	0.97	0.96	0.96	174
3	0.99	0.77	0.87	367
4	0.96	0.75	0.84	1327
5	0.92	0.99	0.95	4817
accuracy			0.93	7341
macro avg	0.96	0.89	0.92	7341
weighted avg	0.94	0.93	0.93	7341

**********BernoulliNB**********
ACCURACY SCORE PERCENTAGE: 80.07083503609861
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.46	0.96	0.62	656
2	1.00	0.37	0.54	174
3	0.76	0.49	0.60	367
4	0.92	0.49	0.64	1327
5	0.88	0.90	0.89	4817
accuracy			0.80	7341
macro avg	0.80	0.64	0.66	7341
weighted avg	0.84	0.80	0.80	7341

**********MultinomialNB**********
ACCURACY SCORE PERCENTAGE: 84.62062389320256
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.91	0.83	0.87	656
2	1.00	0.29	0.45	174
3	0.97	0.31	0.47	367
4	0.95	0.53	0.68	1327
5	0.82	1.00	0.90	4817
accuracy			0.85	7341
macro avg	0.93	0.59	0.67	7341
weighted avg	0.87	0.85	0.83	7341

**********SGDClassifier**********
ACCURACY SCORE PERCENTAGE: 92.80751941152432
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.96	0.96	0.96	656
2	0.96	0.94	0.95	174
3	0.98	0.77	0.86	367
4	0.97	0.73	0.83	1327
5	0.91	0.99	0.95	4817
accuracy			0.93	7341
macro avg	0.96	0.88	0.91	7341
weighted avg	0.93	0.93	0.92	7341

**********RandomForestClassifier**********
ACCURACY SCORE PERCENTAGE: 93.13445034736412
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.96	0.96	0.96	656
2	0.97	0.96	0.96	174
3	0.98	0.77	0.86	367
4	0.96	0.75	0.84	1327
5	0.92	0.99	0.95	4817
accuracy			0.93	7341
macro avg	0.96	0.89	0.92	7341
weighted avg	0.93	0.93	0.93	7341

**********XGBClassifier**********
ACCURACY SCORE PERCENTAGE: 92.52145484266448
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
1	0.97	0.95	0.96	656
2	0.97	0.96	0.96	174
3	0.98	0.75	0.85	367
4	0.97	0.72	0.82	1327
5	0.91	0.99	0.95	4817
accuracy			0.93	7341
macro avg	0.96	0.88	0.91	7341
weighted avg	0.93	0.93	0.92	7341

1. We check our accuracy by using 6-7 models.
2. Out of which RandomForestClassifier is giving us 93% Accuracy.

CONCLUSION

• Key Findings and Conclusions of the Study

This research evaluated the rating of a product classification using machine learning and deep learning techniques. Using real data, we compared the various machine learning algorithms' accuracy by performing detailed experimental analysis while classifying the text into 5 categories. Generally, Random Forest Classification machine learning algorithms have shown a better performance with our real-life data than others, and the most performing models are all ensemble classifiers.

• Learning Outcomes of the Study in respect of Data Science

In this project we were able to learn various Natural Language Processing techniques like lemmatization, stemming, removal of Stop Words, etc. This project has demonstrated the importance of sampling effectively, modelling and predicting data. Through different powerful tools of visualization, we were able to analyse and interpret different hidden insights about the data. The few challenges while working on this project are: 1. Imbalanced Dataset. 2. Lots of Text data. The dataset was highly imbalanced so we balanced the dataset using smote technique. We converted text data into vectors with the help of TfidfVectorizer.

• Limitations of this work and Scope for Future Work

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens. This forecast also predicts broader applications for sentiment analysis – brands will continue to leverage this tool, and so will individuals in the public eye, governments, non-profits, education centres and many other domain organizations. Sentiment analysis is getting better because social media is increasingly more emotive and expressive. A short while ago, Facebook introduced "Reactions," which allows its users to not just 'Like' content, but attach an emoticon, whether it be a heart, a shocked face, angry face, etc. To the average social media user, this is a fun, seemingly silly feature that gives him or her a little more freedom with their responses. But, to anyone looking to leverage social media data for sentiment analysis, this provides an entirely new layer of data that wasn't available before. Every time the major social media platforms update themselves and add more features, the data behind those interactions gets broader and deeper. Negative reviews can carry as much weight as positive ones. One study found that 82% of those who read online reviews specifically seek out negative reviews. Research indicates that users spend five times as long on sites when interacting with negative reviews, with an 85% increase in conversion rate. Customers like to see lots of reviews. A single review with a few positive words makes up an opinion, but a few dozen that say the same thing make a consensus. The more reviews, the better, and one study found that consumers want to see at least 40 reviews to justify trusting an average star rating. However, a few reviews are still better than no reviews.