

**Rating Prediction Project**

Submitted by:

Arpan Biswas

**ACKNOWLEDGMENT**

Some of the research papers and articles, I find useful for completion of this project.

REFERENCES:

1. Anqing Chen, Jonathan Walsh, Matt MacDonald, Nicholas Chu, Ryed Ahmed, and Saaketh Rao, “Amazon Review Rating Prediction with NLP.” Data Science Lab Spring 2021
2. https://towardsdatascience.com/1-to-5-star-ratings-classification-or-regressionb0462708a4df

.

**INTRODUCTION**

**Business Problem Framing:**

* We have a customer who has a site where individuals compose various surveys for specialized items. Presently they are adding another element to their site for example The commentator should add stars (rating) also with the survey.
* The rating is out 5 stars and it just has 5 choices accessible 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Presently they need to anticipate evaluations for the surveys which were written before and they don't have a rating. Along these lines, we need to assemble an application which can foresee the rating by seeing the survey.
* Here, we want to create a model which can anticipate evaluations utilizing audits of the item.

**Conceptual Background of the Domain Problem:**

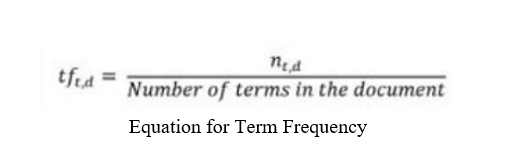
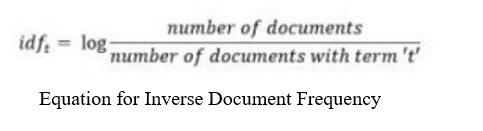
In this project, we want to scrpe surveys and their evaluations of various items and in view of that we will make an AI model. This model is helpful for our customers to anticipate evaluations from surveys.

**Review of Literature:**

* Anqing Chen, Jonathan Walsh, Matt MacDonald, Nicholas Chu, Ryed Ahmed, and Saaketh Rao conduct a study on ratings prediction of amazon reviews using natural language processing. They predict reviews using classification algorithms as well as regression algorithms and then compare both of them.

**Analytical Problem Framing**

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

* **Synthetic minority oversampling technique (SMOTE):** In our dataset target variable is imbalanced and it provides inaccurate results during machine learning part. So, we used SMOTE method to cop up with that issue. SMOTE is an oversampling method and one of the most commonly used oversampling method to solve the imbalance classification problem. This method creates synthetic samples (not duplicate) of minority class. These synthetic records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class.
* **Removing Stop Words:** Stop words are normal words that structure a sentence. Words, for example, "I", "are", and "here" don't add to the sentiment (the rating for our situation) of reviews. Thus, we chose to eliminate stopwords. We utilized NLTK's stopwords bundle to give us the rundown of stopwords.
* **Lemmatization:** Lemmatization is the method involved with changing a word over to its base structure. Lemmatization typically alludes to doing things appropriately with the utilization of a vocabulary and morphological investigation of words, regularly meaning to eliminate inflectional endings just and to return the base or word reference type of a word, which is known as the lemma.
* **Term Frequency-Inverse Document Frequency (TF-IDF):** This way to deal with encoding is likewise genuinely simplistic however conveys more data than simply the quantity of events, similar to Bag-of Words. TF-IDF is a mathematical measurement that likewise reflects how significant a word is to an archive or passage. The plan is parted into two estimations, TF and IDF.
* TF, or Term Frequency, is the number of times a term, *t*, appears divided by the number of terms in the document/paragraph, *d*. 
* IDF, or Inverse Document Frequency, is a measure of how important the term is in the corpus. It’s calculated by taking the log of the number of documents divided by the number of documents with the term *t* in it.
* The TF-IDF encoding scheme outputs TF × IDF for each term in each review.

**Data Sources and their formats:**

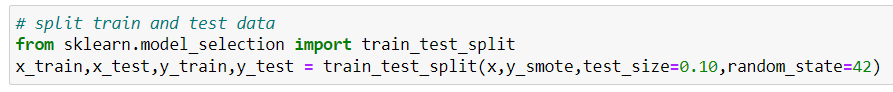
* We scraped the dataset of various products such as mobile phone, printer, smart watches, laptop, earphones etc., from different websites such as amazon, flipkart, myntra.
* The dataset consists of 23883 rows and 3 columns.

**Data Preprocessing Done:**

For cleaning and pre-processing we used some techniques to do so:-

* Dropped Unnamed:0 column, as it was just an index number, which will not help in detecting the ratings of the product.
* We checked for the null values, and there were 56 null values present, so I dropped the null values.
* We checked for the counts of the ratings which were numbered as 1, 2, 3, 4 and 5. But the count of the numbers were not balanced, we used smote technique to balance it.
* We changed our target variable (Rating) datatype from float to integer.
* After data analysis, we prepare our data using NLP (Natural Language Processing). We convert review column in lower case and remove punctuation from that column like +, - , /, etc. because that type of punctuation badly affect out result. We also remove digits from our review column.
* At last, we use stop words, lemmatization, tf-idf and smote techniques for preparation of the data, which we already discussed above.

**Data Inputs- Logic- Output Relationships:**

* Here we used classification algorithm as our target variable is rating and it is a category.
* We then split the dataset into two parts, training dataset and testing dataset where 90% of the dataset is for training and 10% of the dataset is for testing.

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

Hardware Requirements: -

1. Processor: 7th gen core i5 or above.
2. RAM: 4 GB DDR3 or above.
3. HDD/SDD: 128 GB or above.
4. GPU: Intel iris plus graphics 640 1536 MB or above.

Software Requirements: -

Anaconda software must be installed with all the necessary libraries like pandas, numpy, matplotlib, seaborn, scipy, sklearn, NLTK.

**Model/s Development and Evaluation**

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

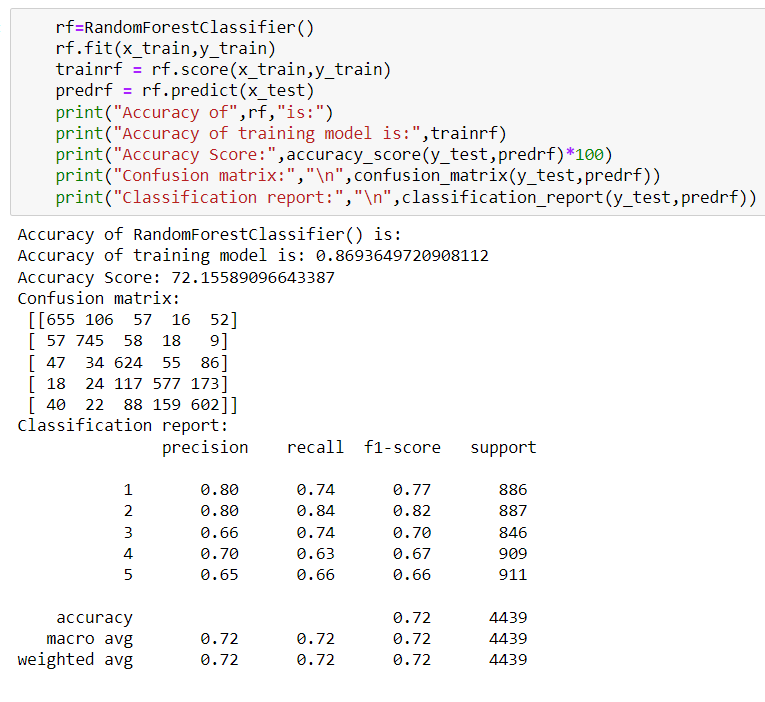
* Below is the method that we use in solve the issue and make the best model:
  + 1. Data reading and understanding
    2. Data cleaning
    3. Data analysis
    4. Data preparation
    5. Over sampling
    6. Train test split
    7. Machine learning algorithms

**Testing of Identified Approaches (Algorithms):**

The classification algorithm that we used is:

1. Logistic Regression
2. Decision tree classifier
3. Multinomial Naive Bayes
4. Random Forest Classifier
5. Support vector classifier

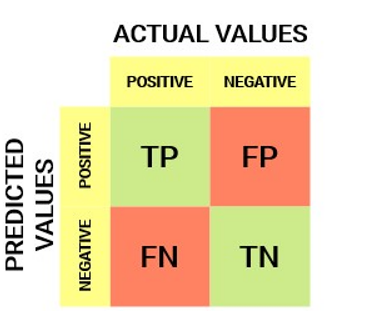
**Run and Evaluate selected models:**

* We used 5 algorithms to test our model and out of which one model turned out to be the best model, we will be describing the best model here.
* Random Forest model turned out to be the best model here. It comes under supervised learning. The algorithm generally generates hundreds or thousands of individual trees.
* The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.
* This classifier produces multiple decision trees and merges them together to produce accurate result. Below is the code of our model with random forest classifier:
* We get the 87% as the training accuracy and 72 % as the test accuracy. And compared to other algorithms we have the best confusion matrix and classification report.

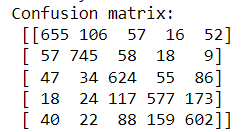
**Key Metrics for success in solving problem under consideration:**

1) Accuracy score: Accuracy score is a measurement for assessing classification model. It is determined by number of right forecasts made partitioned by the all-out number of expectations made.

2) Confusion matrix: A confusion matrix is a nxn network utilized for assessing execution of order model, where n is the quantity of target classes. The matrix contrasts the objective qualities and anticipated qualities which are anticipated by AI model. This matrix provides us with a perspective on how great our grouping model work and what sort of blunders t is making. The following is the basic figure of confusion matrix.

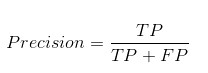


* If the model predicts that a client will enter default, and client really enters into default, then it is true positive situation (TP).
* When the model classifies a case as positive, but it actually negative, then it is a false positive observation (FP).
* If the model predicts that a client will not enter into default, and the client didn’t entered into default in the test set, then it is true negative situation (TN).
* The prediction of the model may be negative, while the client may actually enter into default, then it is false negative classification (FN).
* Below is the figure of our best confusion matrix with random forest classifier:

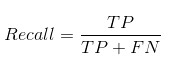


3) Classification Report: A classification report is used to measure the quality of predictions from a classification algorithm. This report shows the metrics like precision, recall and f1- score. The metrics are calculated by using true and false positives, true and false negatives.

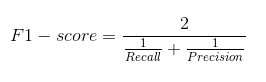
* Precision depicts how many of the correctly predicted cases actually turned out to be positive. Below is the formula of precision.



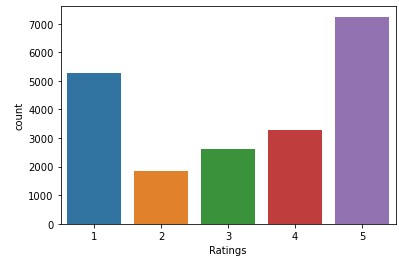
* Recalls depicts how many of actual positive cases we were able to predict correctly with our model. Below is the formula of recall:



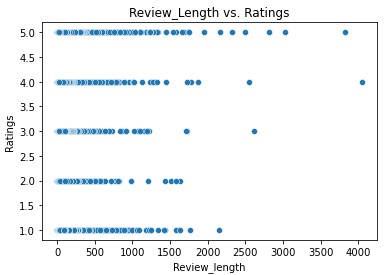
* F1- score gives combined idea about precision and recall metrics. Below is the formula of f1-score:



**Visualizations:**

Let’s look at some visualizations: -

* In this bar plot we can clearly see that the count of rating 5 is the maximum followed by rating 1.
* 2,3,4 rating counts are the less as compared to other two ratings. We can deduce from the plot that people mostly tend to rate at extremities i.e., 5 if the product is good and 1 if the product is bad.



From the scatterplot we can see that the higher ratings have higher review length.

**CONCLUSION**

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

* The reason for this article was make prescient model ready to adequately characterize evaluations dependent on reviews of various items.
* Our target variable is imbalanced, so we use SMOTE procedure to get a decent outcome.
* We utilize numerous classifiers to see as best model and best outcome were noticed for the Random Forest classifier with 72% accuracy.
* By utilizing this model our customer effectively gets the appraisals from surveys.

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

If we can add more features to the dataset then the machine learning will be better than this. And we can use other sampling techniques also to improve the result.