

Ratings Prediction Project

Submitted by:

Nirav Mehta

**ACKNOWLEDGMENT**

I scraped Data from Amazon and Flipkart for different products in different websites.

**INTRODUCTION**

* Business Problem Framing

A website has a collection of product reviews for technical items as well as a forum for writing technical product reviews. They are currently adding a new function to their website called.

The reviewer must offer stars (rating) in addition to the review. The evaluation is a 5 stars and there are just five viable choices 1, 2, 3, 4, and 5 stars, respectively. Now they wishing to forecast the ratings for past evaluations for which there is no data rating. It is necessary to use an app to anticipate the rating based on the review in order to be erected. Consequently, an accurate prediction model for user ratings. It is necessary to conduct a review based on the input.

* Conceptual Background of the Domain Problem

The NLTK suite's numerous libraries are used in conjunction with machine learning approaches like predictive modelling and classification algorithms to classify comments. While filtering out words and other noise that have no bearing on the semantics of the comments and reducing the words to their base lemmas for quick processing and precise classification of the comments, the frequencies of harmful words occurring in textual data were estimated and given appropriate weighting using NLTK tools.

* Review of Literature

To gain insight into the significance of contextual information of user sentiments in determining the rating of products, a research paper titled "Review-Based Rating Prediction" by Tal Had ad was examined and studied. Additionally, the role of natural language processing tools and techniques in identifying the user sentiments towards various products based on their reviews and ratings was examined. It is discovered that contextual information about a user's perception of a product can be expressed explicitly or inferred implicitly through a variety of means, including user score ratings and text reviews. Based on a product's quality, features, performance, and cost, a user may indicate in his review his contentment or unhappiness with it. The user may then assign the product a rating score based on their opinion of it.

* Motivation for the Problem Undertaken

Ratings are a crucial indicator in e-commerce applications to assess a product's value, profitability, and consumer demand. Users' rating scores and reviews on products indicate how they feel about those products. This aids determine user perceptions of the product, which provides information on the degree to which consumers accept the product. A significant positive association exists. Between a product's customer demand and rating. Consequently, it's essential to create a prediction model that can reasonably forecast the rating a customer will give a specific product depending on consumer feedback. This aids in interpreting user sentiment. Consumers toward a product to assess the value of and acceptability of the product.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Predictive models were created using a variety of classification analytic techniques in order to comprehend the links between user reviews and the associated user ratings. Reviews from users are gathered, evaluated, and normalised. A forecast of the rating for a particular review can be formed using the context of reviews on other products that have already received equivalent ratings. Models including Logistic regression, Random Forest Classifier Boost Classifier, Extreme Gradient Boost Classifier, Multinomial Nave Bayes Classifier, Complement Nave Bayes Classifier, and Passive Aggressive Classifier were employed to predict scores for user reviews.

* Data Sources and their formats

User reviews and ratings were scraped to create the dataset. The information for different products is gathered from https://amazon.in and https://www.flipkart.com. Under the characteristics, Comment, and Ratings columns, the data was transformed into a Pandas Data frame and stored as an.xlsx file.

Nearly 20469 samples with 2 characteristics make up the data set. This challenge is a multi-classification problem because Ratings is my goal column and it is a categorical column with 5 categories. The Ratings, which indicate how probable the product is to be purchased by the buyer, range from 1 to 5. The data set consists of the following items: Text Content of the Review. Ratings: Star ratings out of 5.

* Data Pre-processing Done

Null values in rows were eliminated.

Columns: Unnamed: 0 (simply a list of numbers) was removed because it didn't help create a strong model for forecasting the values of the target variable.

The contents of the train and test datasets were then changed to lowercase.

Punctuation and other extraneous characters were eliminated, and single words were substituted for currency symbols, phone numbers, web addresses, and email addresses.

As stop words, tokens that had no bearing on the messages' meaning were eliminated. Ultimately, Word Net Lemmatize was used to lemmatize the retained tokens ().

Following that, the string lengths of the original and cleaned comments were compared.

* Data Inputs- Logic- Output Relationships

The classification models take the comment tokens that have been vectorised using Tiff Vectorised as input and predict the corresponding rating based on their context as output.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I separated our feature and labels and used TF-IDF vectorised to turn text into feature vectors. Additionally, I made sure the input data was scaled and cleaned before feeding it into the machine learning models. Simply improving the Reviews will result in fewer words to process and greater accuracy. Removed unneeded spaces, changed phone number, email address, and email keyword, etc. I made an effort to keep reviews brief and pertinent as much as I could.

* Testing of Identified Approaches (Algorithms)

We must predict Ratings, a multi classification problem, in this nlp-based project. I extracted our feature and labels from the text using TFIDF's vectorised, then used One Vs Rest Classifier to generate the model. SGD Classifier, of all the algorithms I used for this, is the best algorithm for our final model because it performs well in comparison to other algorithms. I used the following methods and evaluated them using various metrics.

* Key Metrics for success in solving problem under consideration

To choose the most appropriate algorithm for our final model, I utilised the following metrics for evaluation:

I used f1 score, precision score, recall score, multilevel confusion matrix, and hamming loss.

Precision can be thought of as a quality indicator; greater precision indicates that an algorithm returns more pertinent outcomes than unrelated ones.

Recall is a quantity metric, and a high recall indicates that an algorithm has returned the majority of the pertinent data.

When True Positives and True Negatives are more significant than Accuracy Score, Accuracy Score is used.

When the class distribution is similar, accuracy can be applied.

When the importance of False Positives and False Negatives is critical, the F1-score is used. F1- score is a superior metric when classes are unbalanced.

* Visualizations

For categorical features, I used bar graphs, and for numerical features, I used scatter plots.

Before the data was cleaned up, the majority of the comments ranged in length from 0 to 800. After data cleansing, it was between 0 and 500.

Words like "good," "outstanding," "quality," "worth money," etc. are frequently used in reviews with a 5.0 rating, suggesting extremely high customer satisfaction and high quality products.

Words like "good," "excellent," "performance," "features," "quality," and "price" are frequently used in reviews with a 4.0 rating, indicating a high degree of consumer satisfaction and a high-quality product.

Words like "good," "well," "average," "quality," "problem," etc. are frequently used in reviews with a 3.0 rating to indicate consumer unhappiness and ordinary to below average product quality.

Words like "problem," "awful," "problems," "wasting money," "poor quality," etc. are commonly used in reviews with a 2.0 rating, indicating substantial consumer discontent and below average product quality.

The likelihood that an event will succeed or fail is calculated using a classification technique known as logistic regression. It is used when the dependent variable has a binary value (True/False, Yes/No, 0/1, etc.). It encourages categorising data into separate groups by considering relationships that can be deduced from a set of labelled data. It derives a linear relationship from the given data. Then gives the dataset a nonlinearity in the form of the sigmoid function. It also provides the direction of connection and a measure of a predictor's appropriateness (coefficient size) (positive or negative).

A random forest is a Meta estimator that employs averaging to increase predicted accuracy and reduce overfitting. It does this by fitting a number of classification decision trees to different subsamples of the dataset. A random forest generates accurate predictions that are simple to comprehend. It can effectively handle huge datasets and reduces overfitting. In comparison to the decision tree method, the random forest algorithm offers a higher level of accuracy in outcome prediction.

Complement Naive Bayes Classifier: This classifier uses a variation of the Multinomial Naive Bayes algorithm known as complement naive bays. Work with unbalanced datasets is a specialty of complement naive bays. Instead of calculating the likelihood that an item belongs to a certain class, complement Naive Bayes calculates the likelihood that the item belongs to all classes.

Passive Aggressive Classifier: Passive-Aggressive algorithms are so-called because they do not require a learning rate and maintain the model and do not update it if the forecast is accurate. In other words, the example's data are insufficient to alter the model in any way. Make model modifications if the prediction turns out to be inaccurate. In other words, a model modification could make it right.

Ada Boost Classifier: The primary idea behind this algorithm is to give misclassified observations more weight. Based on the output of the weak classifiers, the meta-learner adjusts, assigning more importance to the incorrectly categorised observations of the previous weak learner. The final model can converge to a strong learner (a learner not influenced by outliers and with a great generalisation ability, in order to have excellent performances on unknown data), even if the performance of each weak learner is worse than random guessing.

**CONCLUSION**

* Key Findings and Conclusions of the Study

I have gathered information on reviews and ratings for various products from flipkart.com and amazon.in for this project.

Based on user reviews, we have attempted to identify the ratings on commercial websites on a scale of 1 to 5. To accomplish this, we used machine learning and natural language processing techniques.

Both often occurring and infrequently occurring words in our data have been examined.

After completing all of these procedures, I created a function to train and test multiple algorithms. Then, using a variety of assessment metrics, I chose the Random Forest Classifier as our final model.

In the end, we achieved the ideal settings for our final model by performing hyper parameter tuning.

* Learning Outcomes of the Study in respect of Data Science

A crucial step in removing null values from the dataset was data cleansing.

The label column's class composition was discovered using data visualisation, as were the terms that appeared most frequently in reviews corresponding to each rating score.

The project's many data pre-processing and feature engineering processes gave consideration to a number of effective approaches for processing textual data. Building classification models and pre-processing text-based data require the NLTK toolkit.

* Limitations of this work and Scope for Future Work

Scraping the data has a number of drawbacks because it is a dynamic operation. Followed by unformatted raw data that cannot be analysed. Additionally, this study won't cover all Regression algorithms; rather, it will focus on the one that was selected, from the most fundamental ensemble techniques to the most sophisticated.