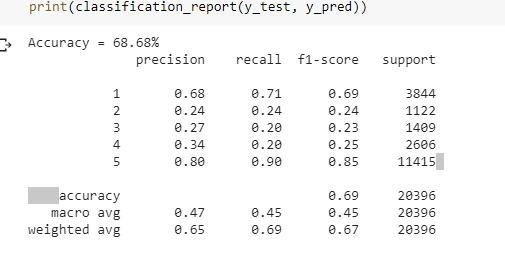
**Ratings prediction**

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!

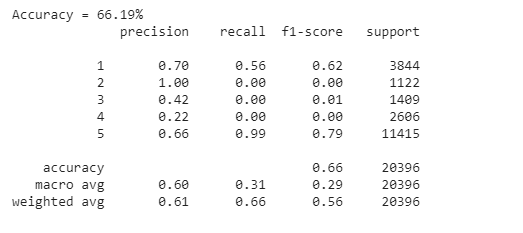
Similar to a typical prediction problem, the data has to be preprocessed before it is fed to the models. The data is cleaned by tokenizing, removing stop words and lemmatizing the words to their roots. The cleaned data or reviews is then split into train and test datasets. The training set is then converted into a document term matrix by passing it through a pipeline containing Count Vectorizer that counts the word frequencies and then through a Tf-idf transformer that transforms the raw counts into TF/IDF values.

A Logistic Regression model is then fit to this document term matrix and is used to predict our response values, Customer reviews. The accuracy, precision recall , f1 score and support are then analyzed to interpret the goodness of the model.



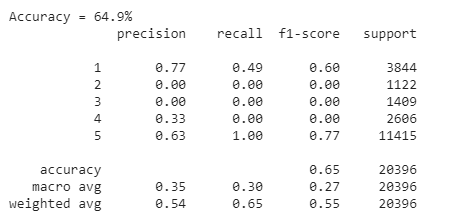
The accuracy for the logistic regression model was found to be 68.68%. The precision for class 1 is 0.68 which means out of all class 1 predictions the number of correct class 1 predictions is 0.61. The recall is found to be 0.71 which means the proportion of correctly predicted class 1 out of total number of actual class 1 cases is 0.71. The F1 score is a harmonic mean of these two figures and the macro avg is the simple mean of f1 scores or precision recall figures over all the classes. Here when we calculate the macro average scores we gave same weightages to all the classes. In weighted scores, we weight the scores from each class by the number of samples from that class. The micro F1 score is same as the overall accuracy of the model. The micro precision score and recall scores are the same. Since our data is highly imbalanced, it is better to look at the weighted f1 score to compare the models. It can be observed that the logistic regression was able to classify the class 1 and class 5 models pretty accurately but the other class predictions were abysmal. Thus we need to explore other models which can overcome this potential issue.

The next classifier model use din the model was Support Vector machines. The classifier, along with the count vectorizer and tfidf transformer were used in the pipeline module. The results obtained are as shown below:



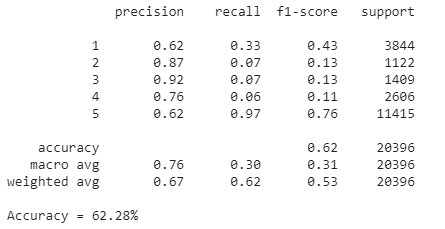
The accuracy of the model was obtained as 66.19% which is same as the micro F1 score for the model. The precision for the different classes in SVM are much better than the logistic regression model. However it comes at the expense of the overall model accuracy. But considering the dataset at hand, and the fact that the reviews would be in general skewed to have a positive sentiment, SVM would be better choice as when compared to Logistic regression.

Another classifier we explored was the Naïve Bayes classifier. The same idea and pipeline module was implemented in calculating the model performance.



Even though at first glance, the model seems okay, it has to be noted that the model performed the worst in predicting the class 2,3 and 4 ratings.

The final model tested in the analysis was the Random Forest model. A slight change was made in how the model was built. While constructing the document term matrix using the tf-idf model, both bigrams and unigrams were considered while building the dtm. Rather than using a count vectorizer , a Tf idf vectorizer was directly utilized in the model to understand the effects on the model accuracy.



It can be observed that the random forest model with the ngram technique was able to improve the precision scores for all the classes. For eg for class 3 ratings, out of all the predicted class 3 ratings, 92% of them were correct. However this reduced the recall score for all the classes consequently. For class 3 ratings, out of the total actual number of class 3 cases, only 7% were predicted correctly by the model.

The model has a micro average F1 score of 0.62 which corresponds to the overall accuracy of the model.