Bernard Mulaw (U24401658, username: ***bmulaw***)

CS 506 - Data Science

Professor Galletti

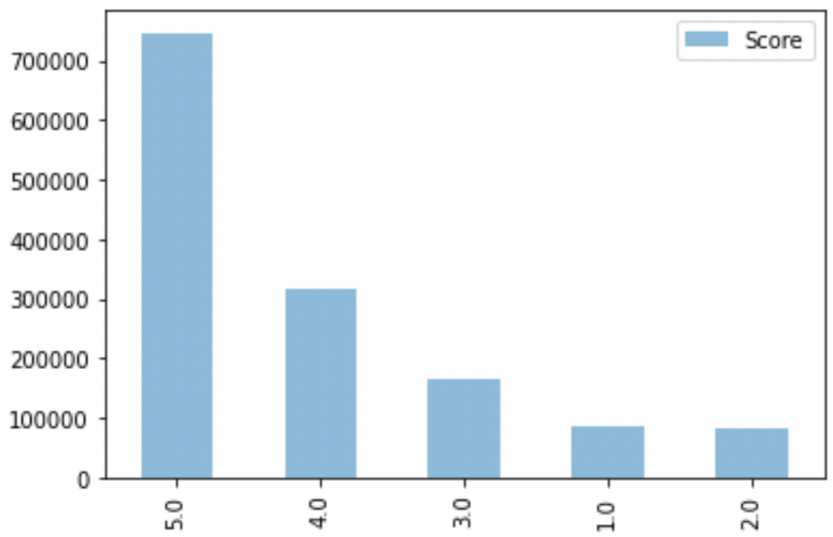
10/27/2021

CS 506 Midterm Essay

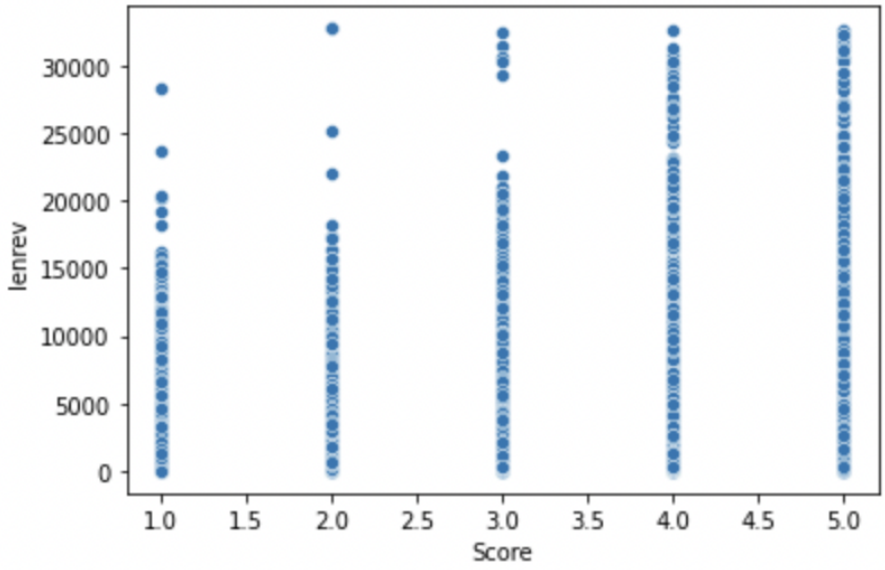
For this project, I choose Native Bayes as my model in predicting user's rating based on their reviews. I picked Native Bayes for two main reasons. One, I have had experience with using it on a project for my CS 111 class (project github link) and I feel most comfortable with it. The second reason is because Native Bayes works really well with analyzing texts and classifying it with some class. I was aware that Native Bayes has been used by GMail to detect scam emails based on the texts so I figured this would be a good model for my project.

My solution/code is in "506-midterm-bmulaw.ipynb". First, I did some preliminary analysis of the *train.csv* file by looking at the top 5 data entries and examining their column types. I counted each rating in the "Scores" column and clearly saw that there is an overwhelming amount of 5-rating compared to 1-rating or 2 ratings (see diagram below).

/>



This warns me that there is bias in my dataset towards more positive reviews and I should be careful not misclassify negative reviews as high ratings. Also, I tried to see if there could be some kind of relation between the Score and the length of the review. From first glance, I see that higher ratings tend to have longer reviews than lower ratings but there are a few outliers (see diagram below).

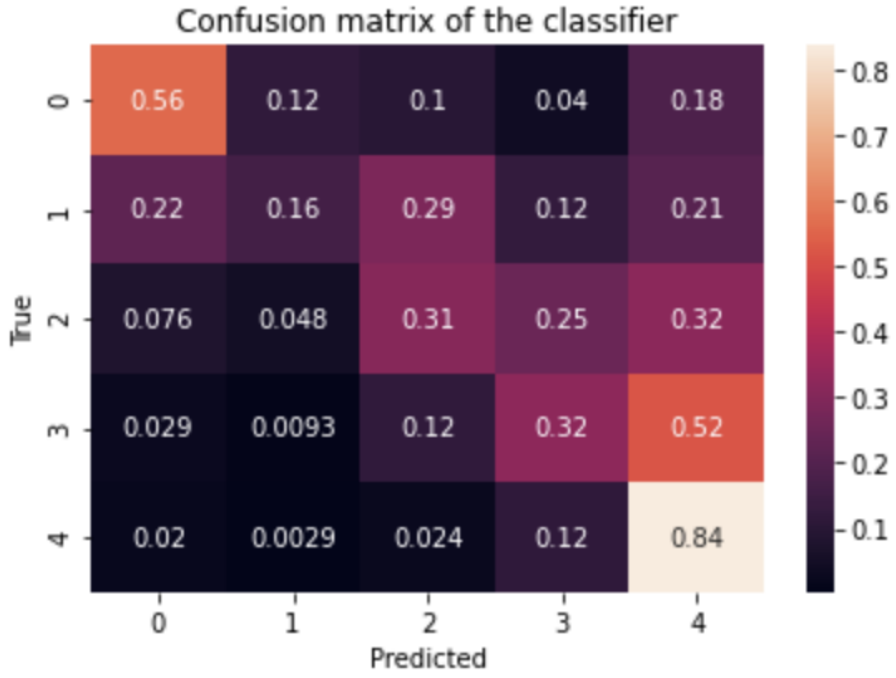


I decided not to use this information in my classification because I wanted to focus on the text.

When cleaning my dataset, my "trick" was that I used **nltk** library to get an array of stopwords, which I used to filter out stopwords from Text. This allows me to have a more intuitive view of the review (what is it's main point) and classify it as good or bad rating faster. Following that, I split the data into a training and testing set and vectorized them; then, I trained the dataset and used inspiration from this web article to use MultinomialNB https://www.ritchieng.com/machine-learning-multinomial-naive-bayes-vectorization/.

Lastly, I evaluated the performance of the algorithm by replicating the professor's given code to get the accuracy score, RMSE, and a confusion matrix (see diagram below).

```
Accuracy score:  0.6017815127974674
RMSE on testing set =  1.1311089230205573
```



Overall, I had a 60% accuracy and RMSE of around 1.13, which I thought was acceptable given that I only used Native Bayes without weights or tfidf transformations. In retrospect, I would like to explore using the length of a review as a parameter in classifying which rating a review leaves, and also explore using SVM or even logistic regression models to compare with Native Bayes model.