BA820: Unsupervised Machine Learning
Aditya Sinha (BU-ID: U18474952)

Assignment #2: Don't Grade! Just need feedback.

Overview

Upon receiving the dataset, we first conducted data exploration. During this time, we discovered that the dataset contains numerous missing values for features that could be important to our analysis. Moreover, it also had duplicated values. Therefore, we took several approaches to clean the dataset before proceeding to the next step. Then we performed EDA and sentiment analysis on the clean dataset. After that we cleaned the review column and converted it into meaningful numerical representation. Finally used UMAP for dimensionality reduction and made a rating prediction model. Lastly, we provided the client with some insights on the clusters and provided recommendations based on our findings

Exploratory Data Analysis, Data Cleaning, and Preprocessing

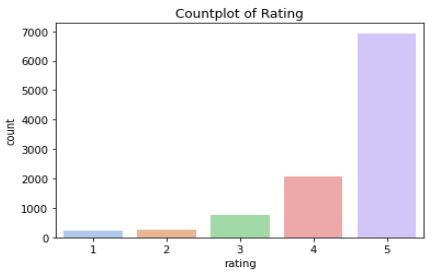
Our dataset consists of 10,261 rows and 2 columns. Each row represents customer feedback and the columns represent review itself and rating for that review. Review was a string column and the rating was a numerical column based on 1-5 scale(1 being worst to 5 being best.) We started by dropping 5 duplicate rows and 2 rows with null value for reviews. When we completed the data cleaning process, we managed to keep over 99.9% of the data. There are now 10254 stocks, 2 columns, and no missing or duplicate values.

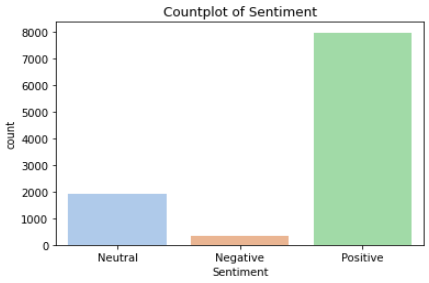
VADER-Sentiment-Analysis

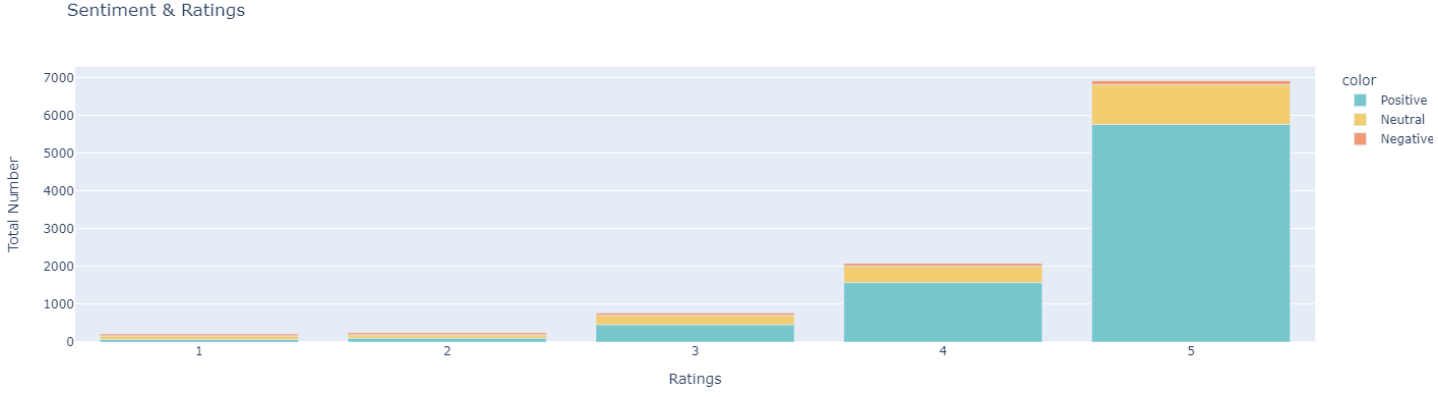
We used Vader to get sentimental scores for the individual review and then converted those scores into 3 categorical Sentiments: Positive, Negative, and Neutral. As a result we added two more columns for Sentimental scores and Sentiment category.

Visual Analysis

Most of the ratings are positive and 5 stars.



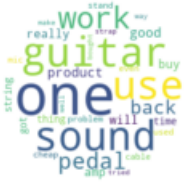




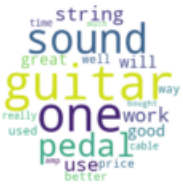
BA820: Unsupervised Machine Learning
Aditya Sinha (BU-ID: U18474952)

Most common words in each rating.

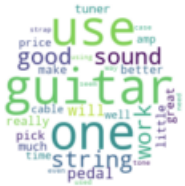
1



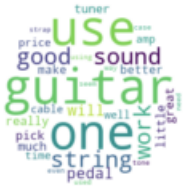
2



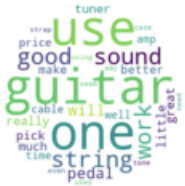
3



4

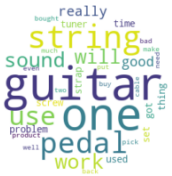


5

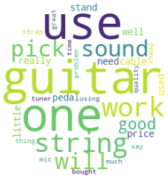


Most common words in each sentiment.

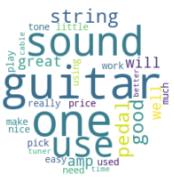
Negative



Neutral



Positive



Prediction Model

We used SnowballStemmer to combine/stem different forms of the same words to prevent getting various forms of words that basically have the same meaning. Then we used RegexpTokenizer to tokenize and keep only words including including. After that, we converted the tokenized word to meaningful numerical representation for ML using TfidfVectorizer. Due to high dimensionality we decided to use UMAP to reduce the dimension to 2 variables.
Next we used 2 UMAP variables with Sentiment_Score to build our ML model to predict rating. We used XGBoost with hyperparameter training to predict the rating with 67% accuracy.

Recommendation

Looking at the top words for the negative and neutral comments, the firm should try to improve in these specific categories.

- Guitar, strings, pedal, quality

The firm should consider combining the 2 step review process into one. Wireframe for the new process is provided as an attachment.

- Comment is an important part of the feedback process and it helps to gain insight about specific areas for improvement.
- On the other hand, rating helps us to quickly identify the overall picture.

Our recommendation is to make the rating available below the comment section and optional for users. Moreover, we developed a ML model that would help to predict rating based on the comment so it could be used if the user does not leave a rating.So, it will keep our database clean with all the information.