Can You Predict a Song's Commercial Success?

ORIE 4741 Final Project: Midterm report

Darren Ting, Mario Medina, Joshua Kwasi Agyei-Gyamfi

November 8th, 2020

1 Problem Statement

Often times, when musicians begin the songwriting process, it would be common for them to wonder how well this new song could potentially do. Will it be played on the radio? Will the kids these days enjoy it? Is it a "bop"? Being able to predict the success of a song can potentially estimate its ability to bring revenue for the artist/group as well as go mainstream to make an influence in today's society, or at least a night in a sweaty college town basement. It can show the trends of current musical interests in mainstream culture which can allow musicians to tailor their music accordingly. For this project, we explore this further by asking, "Can you predict a song's commercial success based on characteristics of the song?".

To answer this question, we use supervised machine learning. we build a predictive model using classification of whether or not the song will be successful or not. We will use audio features of songs requested from Spotify's API as our input variables. This data will be used in conjunction with Billboard Hot 100 data that has been shared on Kaggle which will be used as a benchmark for mainstream musical success. This data set will serve as the response variables for our model. We believe that these datasets will allow us to successfully answer our question and give us more insight on what attributes make a song popular.

2 Dataset Description

The Spotify and Billboard datasets, which we used, were obtained from Kaggle resources and created by Daniel DeFoe and Yamac Eren Ay respectively. The music data we used ranged from 1999 to 2019. A Python script was used to recover the data (2 files) from the data source. After converting all the documents into .csv files, the datasets for 1999 to 2019 were combined into one cleaned master .csv file. Some issues were encountered in which the columns were not consistent with the columns from other files. These were handled by manipulating the data in a plethora of ways (on a case by case basis addressed later) to ensure that the final dataset is uniform. The final dataset has 15 columns and 41,900 rows. The columns being used for the classifier have been described in the appendix (See appendix A.1)

3 Data Preprocessing

Our first step in using our data was to incorporate the information from the Billboard dataset into the Spotify dataset. We did this by finding matches in song and artist pairs in the two datasets and marking the Spotify dataset accordingly; however, there were some issues with this process because there are discrepancies between the format of name and artist of the two datasets. First, the artists have different formats in that in the Billboard dataset, artists are shown as a single string (Ex. "Lil Nas X, Cardi B"), while in the Spotify dataset it is denoted as a list (ex. ["Lil Nas X", "Cardi B"]). Another issue involved artist names with accents, as the Spotify dataset has information regarding the artist "Beyoncé", but the Billboard dataset only has information about the artist "Beyonce". In addition, there were discrepancies with the case of different songs. For example, a certain famous Billie Eilish song is represented as "bad guy" in the Spotify dataset, while the Billboard dataset represents it as "Bad Guy".

Taking all of these issues into consideration, we preprocessed these columns in the two datasets by converting all artists and songs into lowercase, stripping all accent marks, and converting the Billboard artist column into a list. We then removed columns that we did not want to use in our data such as genre and popularity.

4 Exploratory Data Analysis

After syntax discrepancies between the Spotify and Billboard datasets were fixed and the two were merged, we were able to then perform some exploratory data analysis. In total, our dataset contains 41,900 song entries, 5719 of which have been classified as commercially successful songs, or 13.6%, which is defined by a song's appearance in the Billboard Hot 100 Chart. To view the data easier, we then use figures (Appendix B). Figure B.1, shows a heat map highlighting correlation between our variables in the feature space. Some variable pairs appear to be significantly correlated such as Energy being highly positively correlated with Loudness, Loudness being negatively correlated with Acousticness and Instrumentalness, and Energy being negatively correlated with Acousticness. Next, Figure B.2 compares the differences between mean song traits of successful and unsuccessful songs, with the difference defined as $(x_{suc} - x_{unsuc})/x_{unsuc}$. Seven of the 14 mean song traits in successful songs are notably different (>10%) from the mean song traits in unsuccessful songs. Figures B.3.1 and B.3.2 show the distribution of song trait values and the presence of outliers.

5 Preliminary Models

In terms of preliminary analyses, we ran our dataset on a few binary classifiers: LogisticClassifier, Decision Tree and SVM using the Scikit Learn library. For each of these classifiers, we ran a grid search that tested multiple values of their different hyperparameters and picked the best combination based on cross validation with the Balanced Accuracy because of the unbalanced percentage of "non successful" songs. The training:test split was 80:20.

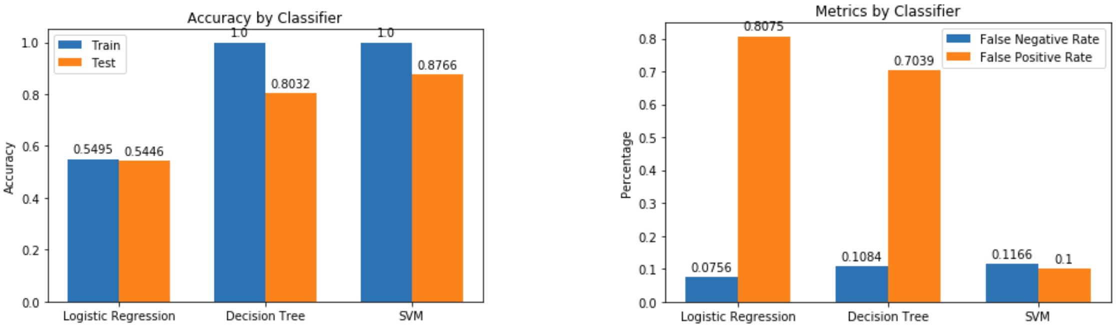


Figure 1: The plot on the left shows the train and test accuracy among the different classifiers. Based on this plot, the SVM did the best, while the Decision Tree performed well. The logistic regression did not perform well. One thing to note is the 100% accuracy on the training set, which may show overfitting, but these models still perform well on the test set. These are still preliminary models, and thus will be expanded upon later. The plot on the right plots the percentage of false positives and false negatives for each classifier on the test set. The decision tree and Logistic regression had very high false positive rates meaning that they classified many songs that weren't on the billboard 100 as songs that were on there. This may either be due to class imbalance, or the difficulty of classifiying these sort of categories. The SVM however seems to have a balanced rate of false negatives and positives.

6 Moving Forward

While our preliminary results look promising, we believe we can do better. To accomplish this, we plan on using RIAA (Recording Industry Association of America) Certification data as another indicator of commercial success to diversify our dataset. Then, we plan on performing additional feature engineering, including PCA to reduce the dimension of our dataset because many variables are shown to not add much valuable information to our model. For our model, we look to incorporate a loss function that penalizes false negatives more than false positives. Our reasoning for this is that if a record company were to use this model to influence business decisions, the opportunity cost of a false negative would be higher than the economic cost of a false positive. Finally, we plan on using more advanced supervised learning methods on our dataset such as gradient boosted trees (XGBoost) to improve our model's accuracy.

A Appendix

A.1 Variables

A.1.1 Numerical:

- acousticness (Ranges from 0 to 1) (The relative metric of the track being acoustic)
- danceability (Ranges from 0 to 1) (The relative measurement of the track being danceable)
- energy (Ranges from 0 to 1) (The energy of the track)
- duration-ms (Integer typically ranging from 200k to 300k) (The length of the track in milliseconds (ms))
- instrumentalness (Ranges from 0 to 1) (The relative ratio of the track being instrumental)
- valence (Ranges from 0 to 1) (The positiveness of the track)
- tempo (Float typically ranging from 50 to 150) (The tempo of the track in Beat Per Minute (BPM))
- liveness (Ranges from 0 to 1) (The relative duration of the track sounding as a live performance)
- loudness (Float typically ranging from -60 to 0)
- speechiness (Ranges from 0 to 1) (The relative length of the track containing any kind of human voice)
- year (Ranges from 1999 to 2019)

A.1.2 Boolean:

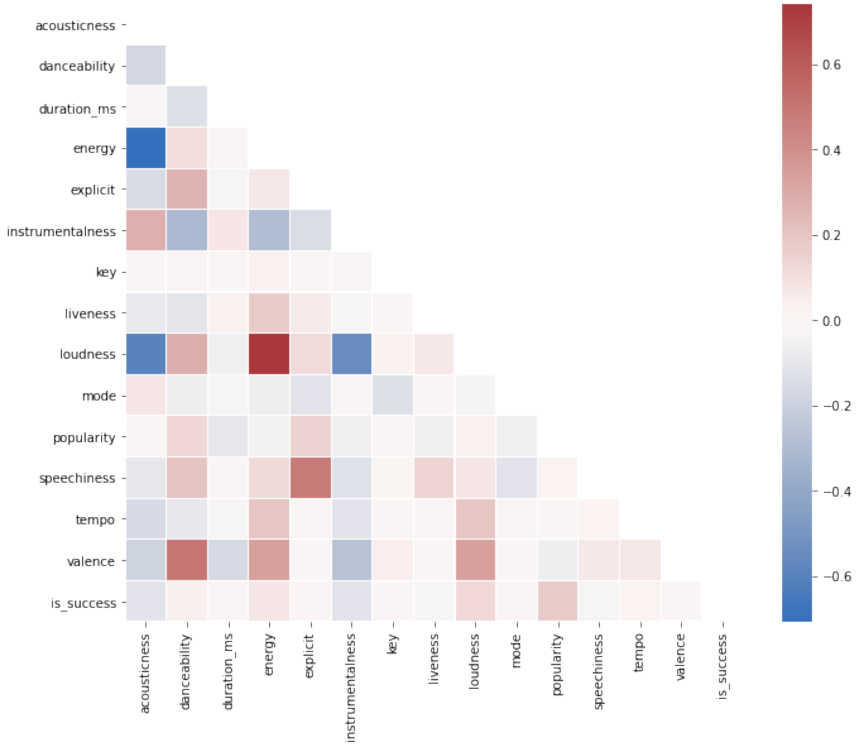
- mode (0 = Minor, 1 = Major)
- explicit (0 = No explicit content, 1 = Explicit content) (The binary value whether the track contains explicit content or not)
- Is-success (0 = not successful , 1 = is successful)

A.1.3 Ordinal

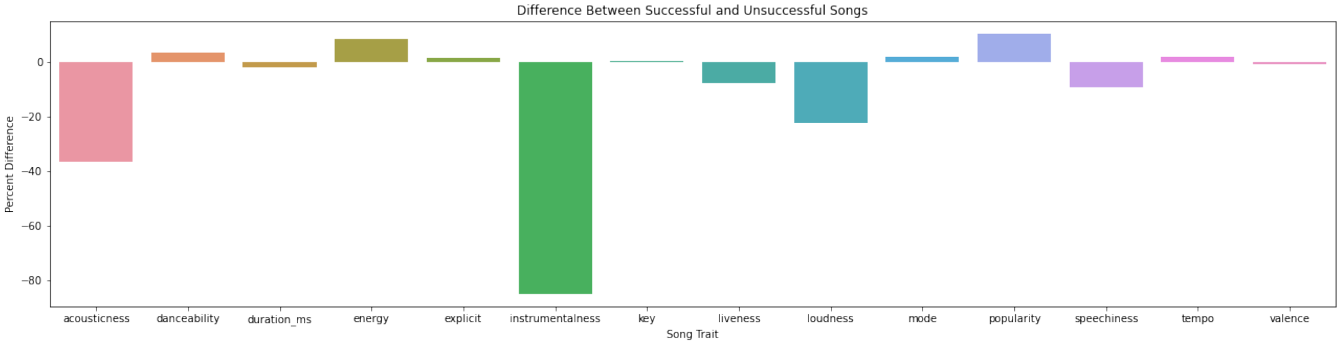
- key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on. . .) (The primary key of the track encoded as integers in between 0 and 11)

B Figures

B.1 Variable Correlation Heat Map

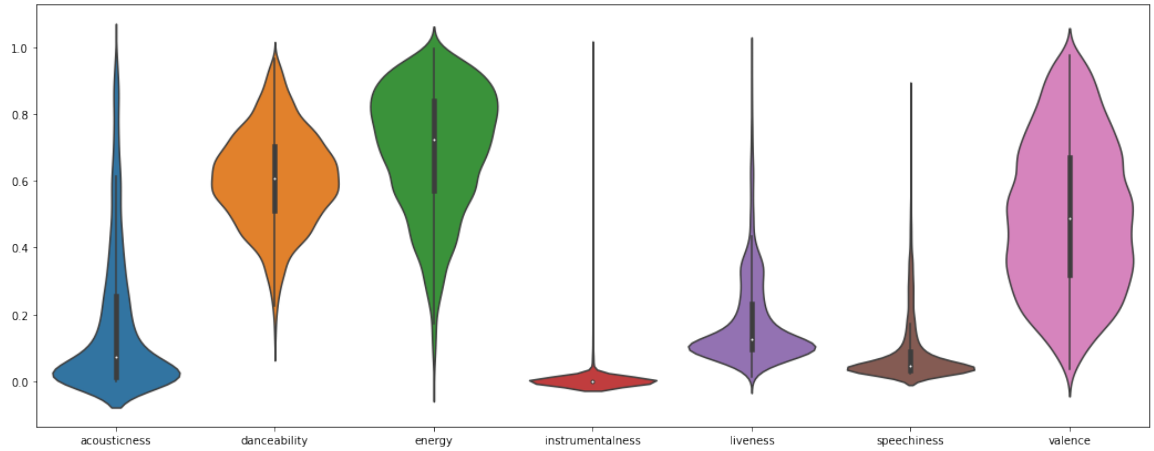


B.2 Mean Differences Barplot

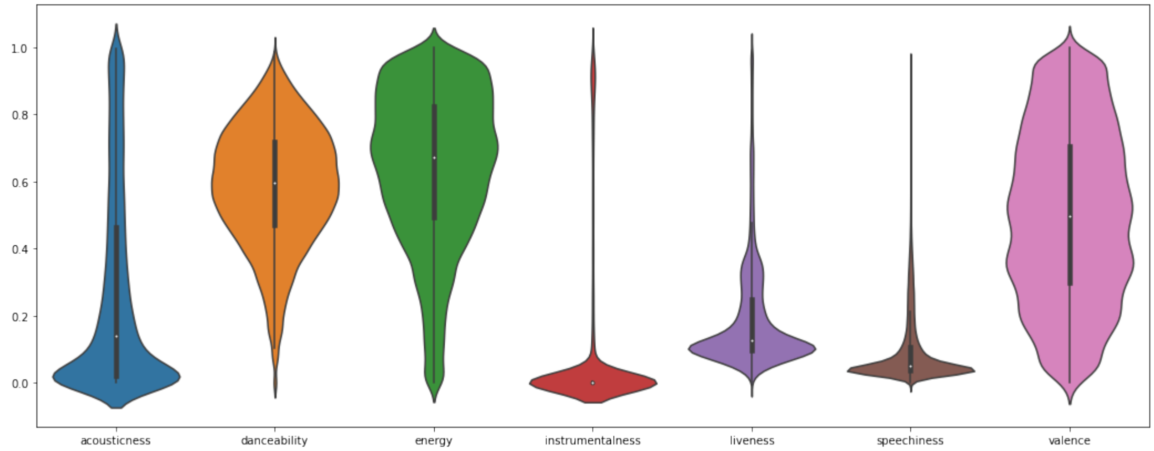


B.3 Song Trait Distrubutions

B.3.1 Successful Songs



B.3.2 Unsuccessful Songs


acousticness
danceability
energy
instrumentalness
liveness
speechiness
valence