Sarah Santiago, Shishi Feng, Simran Regmi, Winnie Lu
IEOR 142
May 12, 2021

Predicting Song Popularity Using Spotify Data

I. Motivation

Why did we choose this project and how is this topic relevant?

A. Background: For many in the music industry, producing the next big hit is a top priority. Currently, artists and music producers have a vague idea of what would make a song a hit - usually catchy songs that are easy to dance along with, such as "Old Town Road" by Lil Nas X or "Party In the USA" by Miley Cyrus are popular amongst the masses. However, it is hard to determine quantitative factors that make a song a hit. This has a significant impact on independent artists who do not have much experience producing and releasing their own songs and may not have a label to help them. Especially with the rise of platforms such as YouTube and SoundCloud where artists can release their own songs independently, it's surprising that there aren't too many resources in this arena.

B. Solution: For our project, we utilized the knowledge we have gained during this semester to determine a way to predict the success of a new song even before it is released. Using Spotify data, we built machine learning models to predict song popularity based on numerous features such as the tempo of the song, the season of the year it was released, and if a song was a collaboration between artists (just to name a few). A model like ours can prove to be a cost effective way for emerging artists to gauge how their song will perform in the mainstream.

C. Additional Uses:

- Audio Streaming + Media Services Provider (Spotify, Apple Music, etc): Spotify currently has a Viral50 chart (measures the popularity of songs based on user sharing + combining analytics on blogs, social media), Top-50 (displays the top 50 streamed songs), and a recommended songs feature where it will provide it's users with songs based on prior listening activity. Using the models from our project, they could implement a fourth chart called Upcoming-50. This chart would run our models on songs published by lesser-known artists and then pick 50 that it predicts would be successful. Through this feature, Spotify would be able to promote upcoming artists and also expand the singing choices it provides to its users.
- Societal Music Preferences: The music that we like as a society has changed drastically throughout time. Rock & roll was extremely prevalent amongst teenagers in the 1950s, disco was at its peak in the late 1970s, and rap is currently one of the more popular genres. All of these genres are associated with social movements and current events so we could use our model to analyze what is most popular currently and why.

II. Data

What is the nature of the data used? How did you collect and process the data?

For our project, we used Spotify data from Kaggle (http://tiny.cc/b5gutz). The 174,389 song dataset was compiled using Spotify's API. The data contained numerous features like: song name, song tempo, danceability, genre, year, etc. Our goal was to make predictions for the "popularity" column in order to gauge the popularity of songs. The provided column had popularity scores ranging from 0 to 100. We changed this column to a binary format where

scores greater than 25 were labeled as 1 and otherwise it was labeled as 0. We chose 25 because that was the average popularity score for all songs.

Our features were obtained from the tracks dataset with a total of 20 columns with 11 numerical variables (i.e. acousticness, danceability, etc) , 6 categorical variables (i.e. key of the song, artist name, release date, etc). With this we were able to get a wide range of features specific to each track that might be correlated with the overall popularity of the artist when a song is released.

In our processing, we wanted to add some useful features to the existing dataset. First, we added a new column indicating if a song was a collaboration between artists. We did this by fixing the format of the existing "artists" column by changing the format into lists. We then could make a binary column indicating if more than one artist was a part of the song. This was done because we found that it was not feasible to one-hot-encode our "artist" column because of how large it would make our dataset. As a compromise, we felt that this could still capture some useful information from the "artist" column. If the song is a collaboration, you can assume more people would listen to it because more than one fan base would be interested in it.

The "release_date" column was also transformed into a "Season" column that stated which season of the year the song came out in. Figure 1 shows the distribution of song popularity by season. The plot verified our initial hypothesis that songs published in a season like summer would have a higher popularity. With this, we could then drop the "release_date" column for simplicity in building our model.

Lastly, we used NLP on the "name" column to see which were the most common words in song titles. We thought that if certain words were used in a song title it could impact a song's popularity so it might make a good feature in our model. We discovered that the top 8 stemmed words in song titles were "live", "love", "mix", "no", "op", "remast", "version", and "year". We then added the vectorized column of these words to our dataset to use for further analysis in our models.

III. Analysis

What methods did you use and what were the results?

To prepare our data for model building, we dropped non-numerical columns such as the name of the artist, the artist id, release date, and name of the song. We also one-hot-encoded our season column so we could use it in our model. Our test and training data split is a 30% and 70% split. Once we have separated the test and training data. We separate the independent variable features from the dependent variable. To reiterate, our independent variables would be features and information that is specific to a song, such as its dancibility, length, and season during which it was released. Our dependent variable would be the song artist's popularity. For all of our models we are encoding popular as a binary variable in that if the popularity of the artist is above the mean of 25, we consider them to be popular (1) and unpopular (0) otherwise. Therefore, what our models predict would be a binary prediction of whether the artist of the song would be popular or not based on the features of a song. For our LDA model, we used the *LinearDiscriminantAnalysis()* function to obtain and train our LDA model using our training data set of song features. Our LDA model had 78% accuracy on the test set with an area under the curve of 83% (Figure 2).

Our logistic regression model had 75% accuracy on the test data and a TPR of 77% (Figure 3). Since it is a logistical regression, the dependent variable is parameterized as a binary field of either the song being popular (1) or not (0) as the result. The features from our original

data set after data cleaning are either continuous or categorical. For the categorical variables I have replaced them in binary coding of 0s and 1s. The scale of the numerical variables in our feature columns vary by a great extent, so the first thing we did was to re-scale the numerical columns using MinMaxScaler from sklearn. Then we ran the model on logistic regression and dropped four independent variables that had a p-value greater than 0.05, which was our cutoff. Then we retrained the model with the updated features and obtained the aforementioned final accuracy of 75%.

Before training the CART model, we first resplit the dataset into training and testing sets with a 70% to 30% split. According to the mean of the popularity (our dependent variable), we converted the popularity score into binary score, which means if the popularity score is greater than the mean, it would be labeled as 1, otherwise 0. And then we applied cross validation to find out the best ccp alpha to build our final CART model. Finally, we got the best ccp alpha as 0.002 and node count as 23 for our CART model, and had an accuracy of 85.9% on our testing dataset.

IV. Conclusion

What did we learn from our project and what can we add in the future?

A. Results: We obtained lower accuracy for the LDA and Logistic Regression model (78% and 75% respectively), which was within expectation since they were less complex than the other models we have tried on this dataset. A general observation was that as we increased the complexity of the model, our accuracy improved by a lot. The accuracy of our CART (86%), Random Forest (88.5%), and boosting (91.5%). It is pretty evident that our model's accuracy substantially increased as we achieved a maximum of 91.5% accuracy with the more complex model of boosting. Curious of how accurate our model can be, we implemented the blended method covered at the end of the semester. Our blended model has a normalized MAE of 0.002002855493548226, a MAE of 0.2002855493548226, a normalized RMSE of 0.03164534320834753, and a RMSE of 0.3164534320834753.

B. Impact: Our models are fairly accurate and give a good overview of how popular a song would be. It is already a useful tool for those in the music industry who would benefit from knowing the potential popularity of certain songs and we are planning to further the features we use to provide more pertinent information. In order to make these models more accessible, we could also create a simple mobile application where users submit the relevant characteristics of their song and in turn, receive an estimated popularity score for the song.

C. Future Additions: Our current features mostly focus on the song itself and while the attributes of the song are quite significant in determining popularity this limits our scope and it is impractical to assume that no other external factors impact popularity. Therefore, in order to make our models more accurate, we would like to add the following features in the future.

1. Song Promotion - Even if someone produces an amazing song, it will not become a hit if the masses do not know about it. Promotion is critical to the success of a song, but it is hard to tell how much of an impact promotion actually has. It would be interesting to measure this impact and if we are successful in doing so, this feature would help music producers understand how much funding they need to allocate for promotion as well as what type of promotion would best benefit them.
    - Amount spent on promotion overall (measured in dollars)
    - Type of promotion (YT Channels such as Buzzfeed or ELLE Song Association, IG, etc)
2. Artist Information - The artist themself also play a large role in the success of a song. They are the face of a song so their actions and background are critical.

- Gender: Most industries, from technology to politics, have a gender bias with male counterparts enjoying more success than their female counterparts. We would like to see if this applies to song popularity. Do the masses have an unconscious bias when it comes to their music taste?
- Culture (Race of Artist, Language of Song): Does race matter when it comes to song popularity? Are songs in foreign languages - Korean, Spanish - less popular than those in English?
- Number of Artists in Song (Solo Artist, Duo, Band, etc): It would be beneficial to also see if solo artists are more successful than bands or duo artists.

We also think it would be helpful to take into consideration any controversies or major events that the artist has been associated with as this may cause bias. For example, when Miley Cyrus released the song "Wrecking Ball," she was surrounded by a lot of controversy from her VMA performance as well as the song's music video. While the song itself was fairly good, a lot of its success can be attributed to the public reaction to Cyrus.

V. Appendix

figures & graphs

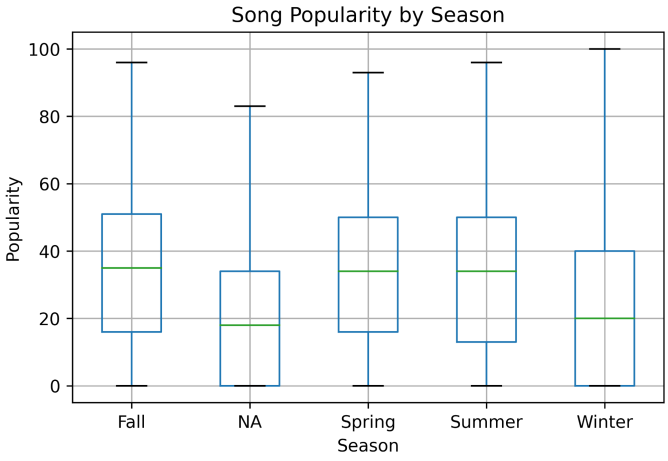

Figure 1: Boxplot of song popularity by season

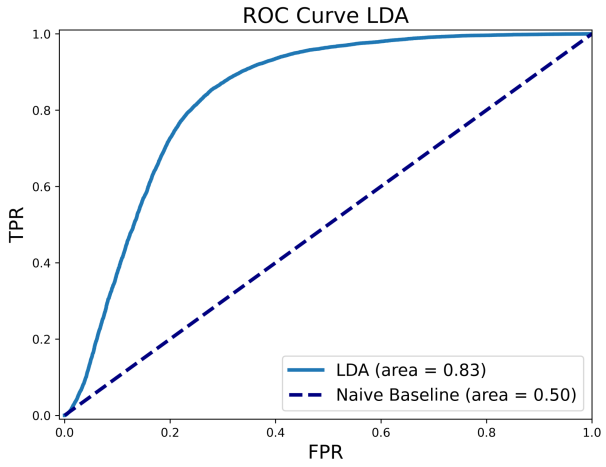

Figure 2: ROC curve of LDA model

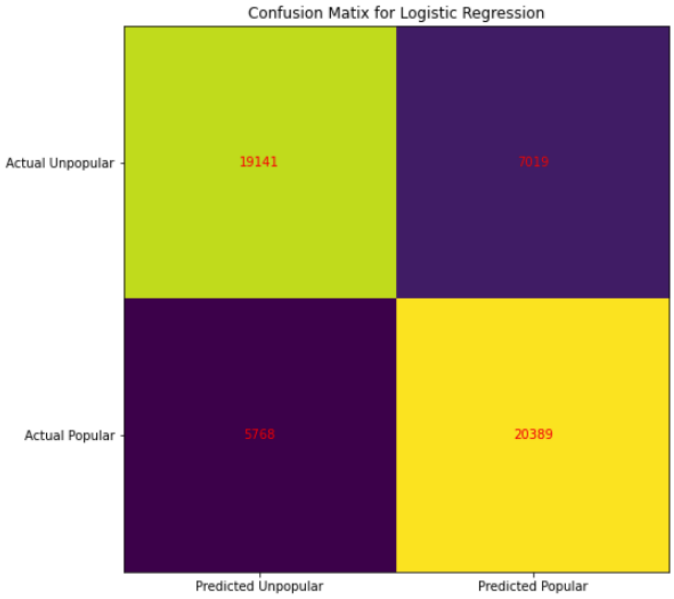

Figure 3: Confusion Matrix of Logistic Regression