An Analysis of Prediction Methods: Song Popularity on Spotify

Shreyas Aswar, Sasha Heslin, Mukul Gharpure, Gummudala Hymavathi

I. Introduction

Music is one of the oldest and most basic forms of human entertainment. It is a universal language that transcends borders, cultures, and religions - bringing people together and fostering emotional connections. While some would consider music to be more amusing than valuable by nature, understanding music and what contributes to its popularity carries many implications. Analysis of song popularity is highly valuable to music industry professionals such as producers and songwriters in giving insight as to what types of music are likely to succeed commercially. Not only can this information drive the business strategy of a major economic industry, but may provide insights into the underlying psychological and social factors that drive human behavior and preference.

Despite its consistent presence in human culture, music has changed significantly as of late. In recent years, there has been a shift in the way that the public consumes music as streaming services have outpaced traditional forms of purchase and consumption, such as pay-per-copy. Now, there is renewed interest in reassessing the way that music becomes popular, and what goes into that phenomena. This has led many researchers to model the relationship between song features and their popularity using machine learning techniques. Several studies have found that certain song features, such as danceability, tempo, and energy, are positively correlated with a song's popularity.

In a study by Su and Lee (2018), researchers used machine learning techniques to predict the popularity of K-pop songs based on their audio features. They found that danceability, energy, and loudness were significant predictors of a song's popularity. Beyond a single genre, Gedikli and Akgun, in their 2017 study, used machine learning algorithms to predict the popularity of songs based on features such as tempo, energy, and acousticness. The authors trained several machine learning algorithms to predict whether a song would be a hit or not and found that random forest and gradient boosting algorithms were the most accurate.

In another study by Cha et al. (2018), the authors analyzed the relationship between song features and popularity using data from the Billboard Hot 100 charts. They found that features

such as danceability, acousticness, and tempo were significant predictors of a song's success. Other studies have additionally focused on the impact of specific genres on song popularity. For example, in a study by Schramm and Van der Meer (2020), the authors analyzed the relationship between musical genres and the success of songs on Spotify. They found that specific genres, such as hip-hop and pop, were more likely to have successful songs.

These studies highlight the potential for using song data and machine learning algorithms to predict song popularity. By identifying the features that are most highly correlated with a song's popularity, these studies provide insight into the factors that contribute to a song's success. While there is evidence to suggest that certain song features and genres can predict a song's popularity using machine learning techniques, there is still much research to be done in this area. This is particularly true in an era of rapidly shifting technology and constantly changing public preferences. As music continues to shift towards streaming platforms, there is a growing need to understand how users interact with music and what factors influence their listening habits.

We therefore were interested in conducting further analysis on data from the most popular streaming platform worldwide: Spotify. Our dataset, which we accessed via Kaggle, contained two datasets pulled from the Spotify web API containing upwards of 250,000 rows. Pulled in November 2018 and April 2019, each set contains the full list of Spotify songs available during that time, alongside measures and features which describe the songs. These features are artist ID, track name and ID, duration, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, time signature, valence, and popularity.

Spotify has created many of these variables, requiring more insight to understand what they mean. Acousticness describes the likelihood of the song being acoustic; Danceability describes how “danceable” a song is based on factors such as tempo and beat strength; Energy describes the perceived intensity and activity of a song; Instrumentalness describes the likelihood of a song being instrumental (the absence of vocals); Key is a categorical variable indicating the key that the song is in, ranging from one to nine; Liveness describes the likelihood of a song being performed live, based on whether there is an audience in the recording. Loudness measures the decibels of the song; Speechiness describes the amount of spoken words in a song; Tempo of the song is measured in BPM (beats per minute); Time signature gives the number of beats in a measure and the type of note that gets the beat and is categorical on a one to five scale; Valence describes how negative or positive a song’s content is; Finally, popularity describes the general

popularity, drawing from factors such as total number of plays and how often the song is added to playlists.

All of these features come together to create a compelling dataset to examine features which may contribute to song popularity. We thus set out to create multiple models to achieve this task in R, a popular statistical software, and then narrow down on a few to compare the most effective methods. In ultimately comparing four popular prediction models against four.

II. Methods

Preprocessing & Initial Evaluation

1. Importing the dataset

In order to create the final dataset of which to run our models, there were a number of preprocessing techniques used to clean and optimize the data. Before importing into R, the two datasets were bound together and stripped of any common values. Due to computational constraints in R, we then took a sample of 20,000 rows from the original dataset to ease the burden of processing.

2. Pre-Processing the data

Upon loading into R, we then removed null and missing values, as well as supplemental or identifying columns that we were not interested in using, such as track name and ID. For our regression analyses, it was necessary to convert any categorical variables to numeric. We therefore used one hot encoding to convert Key, a categorical variable with nine levels, and time signature, with five levels, into a set variables for each level; the level that was present in the song would be labeled with a one, while the rest of the non-present levels would be labeled zero. This way, any key or time signature used would have the same individual weight across songs, as opposed to being weighted according to an arbitrary number.

We can see that songs with popularity “0” do exist, which was almost 10% of our data. We did some research in this regard and found out that Spotify, by default, keeps the popularity as “0” whenever it has data missing about the song. Which is why we decided to drop the instances where popularity is “0”.

3. Data Visualization

We then calculated boxplots and distribution charts for each variable and correlation matrices to determine if there were any initial trends to keep an eye on and to guide our methods. Upon boxplot analysis, we noticed that outliers existed in the song duration variable and variables were on very different scales, which could ultimately affect our model.

4. Dealing with Outliers Normalizing the data

We first plotted the histograms of the continuous variable and we realized that we need to remove the outliers for the feature "duration_ms".

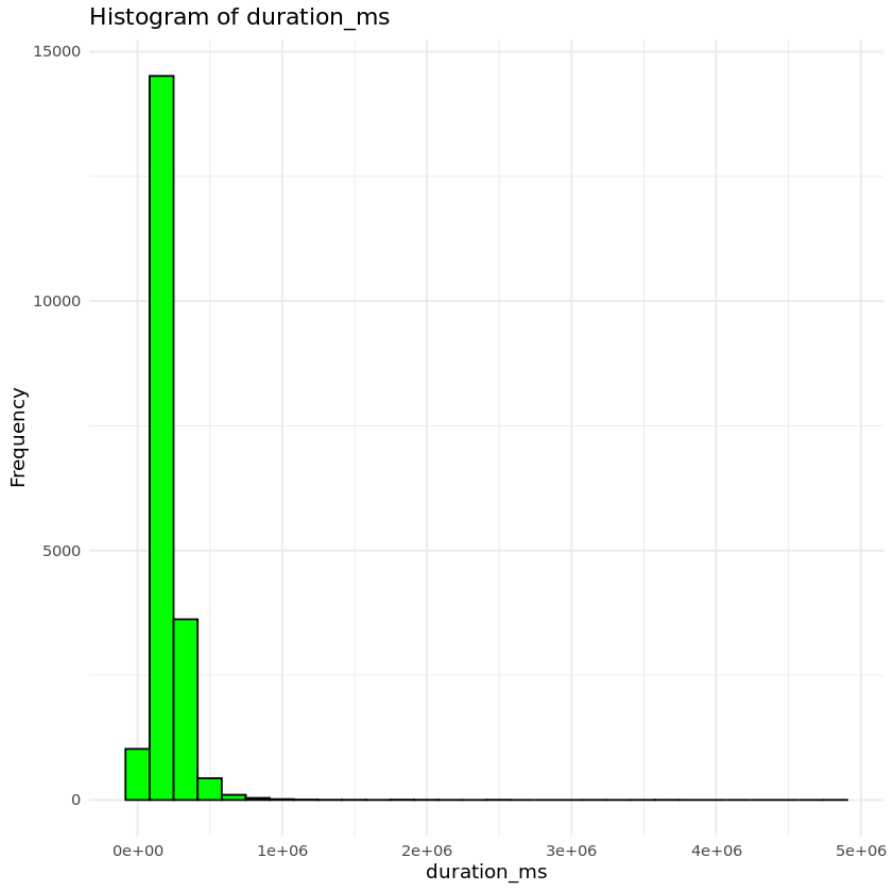

Figure 1: Histogram of Frequency vs Duration_ms

As seen from the above graphs, it is left skewed. We dealt with this by first converting the unit of duration_ms from "milliseconds" to "seconds" by dividing the feature by 60,000. And

then we kept only those instances which were present in the Interquartile range.

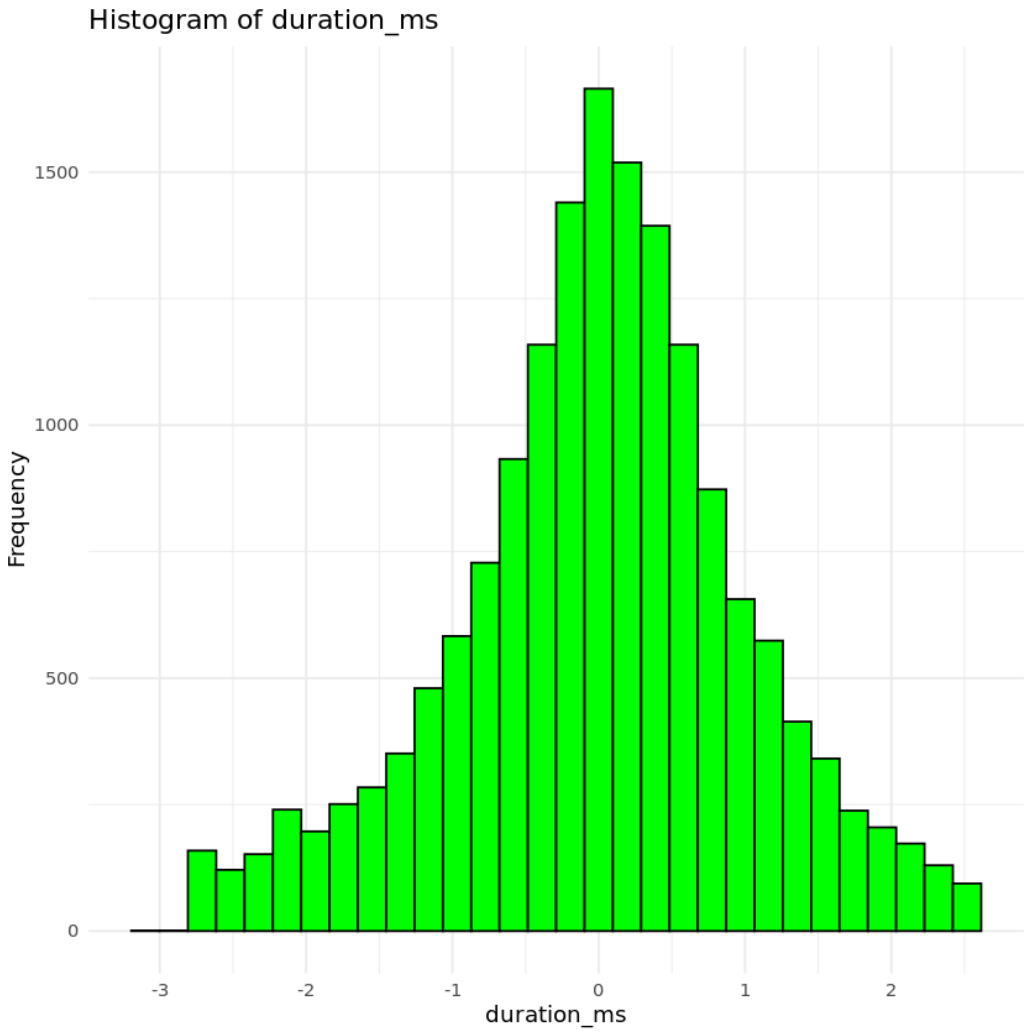

Figure 2: Histogram of Frequency vs Duration_ms

We have used a Standard Scaler technique to standardize the data by removing the median and scaling the data to the interquartile range (IQR). This method is more sensitive to outliers compared to other scalers and doesn't change the distribution of data points, making it more appropriate for our data.

5. Correlation Matrix

To conduct initial assessment of relationships, Pearson and Spearman correlation coefficients were both used. When focusing on the relationship between song popularity and predictors, very few notable relationships stood out. Instrumentalness and loudness of a song each had a correlation coefficient of between .2 and .3 in relation to song popularity, with negative and positive effects, respectively. Following behind, spearman's coefficient matrix

showed danceability, acousticness and energy as very weak (but stronger than most) positive predictors with coefficients of around .1. Beyond that, every variable seemed to have a negligible, if not nonexistent, effect on popularity.

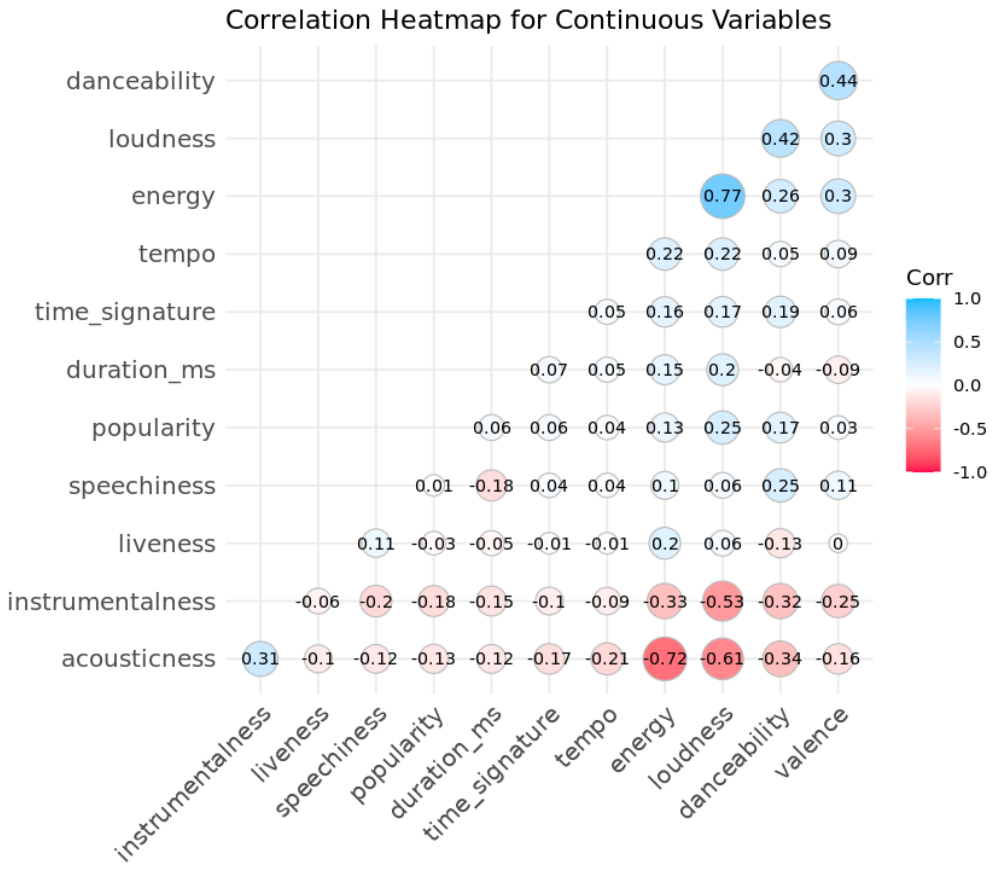

Figure 3: Correlation Heatmap

6. Principal Component Analysis

Finally, we conducted principal component analysis and used the correlation matrix to assess whether there were variables to remove in order to improve the accuracy of our prediction models. Due to the low relationships between each predictor and the response, there were no variables which seemed obvious in terms of removal or inclusion in the model. In addition, our PCA results were somewhat inconclusive. A PCA graph showed a somewhat consistent share of variability in the data across predictors, with none of them jumping out as something we could

remove or “cut off” at. We then looked into the loadings for principal components and found that there were none. This indicates that the variables included in our dataset may not contribute meaningfully to a model.

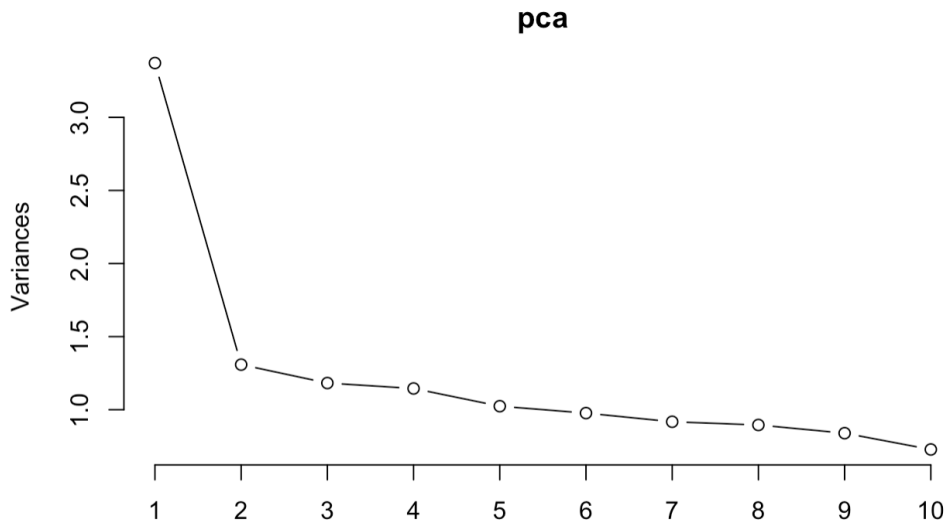

Figure 4: PCA variance graph

Modeling: Regression

Given these findings, we can begin our analysis with limited expectations and hypothesize that, regardless of the most accurate model, there may not be a strong way to predict song popularity from the features provided. Keeping this in mind, we took two general approaches knowing the challenge of accurate prediction: classification and prediction.

We first assessed multiple predictive models to see which performed the most strongly. These models were Linear regression, decision tree, ridge regression, and lasso regression. For each model, data was split using 10 fold cross validation, with a training set consisting of 80% of the data and a testing set for the remaining portion. We then calculated the mean squared error, r squared, and adjusted r squared to evaluate all models against one another.

For linear regression, we split the data into training and testing sets, defined control parameters for cross validation to use 10 folds, and trained the model by using the method “lm”. We then made predictions on the test data using the predict function in R. Finally, we calculated each evaluation metric. Ridge regression was somewhat similar, but required an additional step

in setting tuning parameters. These are used in order to find the optimal balance between fitting the training data well and avoiding overfitting. By varying the strength of the regularization penalty, we sought to control the complexity of the model and improve its performance.

Similar to ridge regression and previous methods, lasso regression required control parameter setting, training the model, making predictions on it, and evaluating. Lasso requires the 'glmnet' method, which then requires us to convert the training data into matrix form. Finally, random forest used the same techniques as previous models, using the 'rf' method.

Modeling: Classification

The results, which are further explained below, were consistently inadequate. As such, we thought it necessary to try another approach, classification models, to see if there was a difference between predicting song popularity by value vs by a more general metric. These models were K nearest neighbor, QDA, LDA, and random forest. This was done first by creating a threshold on which to classify song popularity and then assigning them to each song. We used median popularity as our threshold and classified songs based on whether they were "popular" or "not popular." Due to computational constraints, we were not able to run cross validation on all classifiers. As such, we are aware that there may be an issue of overfitting. However, the validation set method with a 80-20 split was still utilized to train and test most models.

First, for Linear Discriminant Analysis, we created an LDA model object, trained the model, and tested it on the test set. For Quadratic Discriminant Analysis we went one step further and performed Principal Component Analysis on the training data to remove any variables which were beyond the 90% variance cutoff, adjusted the train and test data to remove those variables, and ran the model as was done with LDA. For K-nearest-neighbor, we looped through a function that ran the values k equals one, three, and five, and trained and tested the data for each. Finally, for a random forest, we trained and tested the model. All classifiers were assessed for accuracy, where random forest gave us the most significant results. The results for both our classification and prediction methods can be found below.

III. Results

Model Name	**R Squared**	**Adjusted R squared**	**MSE**
Linear Regression	0.0736586	0.06544585	0.9195855
Ridge Regression	0.07415323	0.06594487	0.9190945
Lasso Regression	0.07388515	0.06567441	0.9193606
Decision Tree Regressor	0.05699978	0.04863934	0.9361228

Table 1: Regression Methods

Model Name	**Accuracy**
KNN (k=1)	55%
KNN (k=3)	56%
KNN (k=5)	57%
QDA	59%
LDA	61%
Random Forest Classifier	63.18%

Table 2: Classification Methods

Classification Report for Best Performing Model - Random Forest :

Classification Report:				
	Precision	Recall	F1-score	Support
Not popular	0.62	0.55	0.58	1540
Popular	0.64	0.7	0.67	1763
Accuracy			0.63	3303
Macro avg	0.63	0.63	0.63	3303
Weighted avg	0.63	0.63	0.63	3303

Confusion Matrix:				
	845	695		
	521	1242		

Table 3: Classification Matrix and Confusion Matrix for the Random Forest Classifier

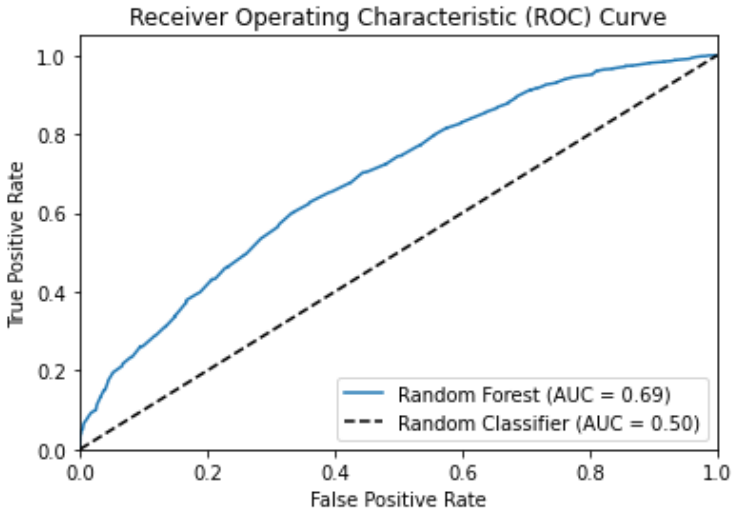

Figure 5: ROC for Random Forest Classifier

According to the prediction models, each performed very poorly in predicting song popularity from the other song features. Linear, ridge, and lasso regression methods all did the "best" compared to the decision tree, but the r squared value of .07 for each is statistically insignificant. In addition, the Mean Squared Error of around .9 for each indicates that the model's predictions are not very accurate, as the average error is relatively large. As such, we did not determine a stand out method for predicting song popularity from production features such as danceability, energy, and loudness, nor did we find any notable predictors.

When it came to classification methods, these methods did not, in most cases, give results that were significantly more useful. Sitting just above 50%, K nearest neighbor methods were not much more accurate than predicting randomly- since we had two groups on which to classify. QDA, LDA, and Random Forest followed behind with accuracy rates of 59%, 61%, and 63%. While these aren't decisively strong models, Random forest stood out as our best performer across all models, both classification and prediction. With an f1 score of .67 for predicting popular songs, and .58 for predicting unpopular songs, it truly did not perform very poorly,

especially when compared to prediction methods. In addition, it's ROC curve indicated a relatively good balance between sensitivity and specificity, measures of the model's ability to correctly identify positive instances and negative instances, respectively.

IV. Discussion

Given the body of previous work on this topic, we were somewhat surprised to find a lack of influence in our prediction model and in many of these predictors in determining song popularity. Additionally, many of our classification models struggled to predict beyond what could randomly be predicted anyways. However, this might support the inconclusiveness present across past studies. Despite this, we were able to conclude that classification models generally performed more strongly than predictive models in their ability to generalize the data, specifically the Random Forest Classifier. This general approach fit the data appropriately, as popularity is not a precise measure, but rather a general ideal to work toward.

Our general lack of conclusiveness shows that, from a production standpoint, there may not be much of a "science" to creating a popular song though the metrics that spotify tracks. While we weren't able to gather many significant findings as it relates to the factors that influence popularity, we can suggest that producers need not get bogged down by the details when it comes to song creation. The components which make up a song interact in different ways to create a unique body of work. Therefore, focusing too technically on the interplay between very specific details, such as noise, acousticness, tempo, and more, will not guarantee a more or less popular song. From a psychology standpoint, this study was sadly, but unsurprisingly, unable to break down any of the mystique behind what drives human preference.

In future analysis, we might consider a dataset with additional features which are less specific to the features of the song itself, but rather describe it more 'externally' (i.e. genre, theme, or money spent to produce it). We might also consider further increasing the number of predictive variables used for greater model accuracy. When it comes to predicting the popularity of a song, it's easy to assume that certain musical features would play a significant role. However, the success of songs like Taylor Swift's "This Love" and "Blank Space" suggests that there are other factors at play. Elements such as marketing strategy, timing of release, and the

artist's personal brand could all contribute to a song's success. Therefore, we should keep an open mind and consider all possible factors when trying to predict the success of a song.

All in all, this indicates that there is a need for new areas of study into music popularity prediction, such as advertising tactics, for example, to inform what truly makes a difference for record labels. Though our models did not prove strong in accurate prediction, we can identify a need for more research into this area to better understand a universal aspect of human culture.

V. Works Cited

Cha, J., Cho, S., Kim, S., & Lee, J. H. (2018). Exploring the impact of audio features on music popularity: A case study of Spotify. Journal of the Association for Information Science and Technology, 69(3), 414-426.

Gedikli, Fatih, and Burak Akgun. "Predicting hit songs with machine learning techniques." 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2017.

Schramm, J. N., & Van der Meer, D. (2020). Spotify and the hitability of music genres. Journal of Cultural Analytics.

Spotify. “Audio Features for 130k Tracks.” Kaggle, 22 Jan. 2021, www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks.

Su, Y., & Lee, J. H. (2018). Predicting popularity of K-pop songs using audio and meta data. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 486-489). IEEE.