Understanding Music Popularity Utilizing Machine Learning Techniques

Kevin Nyquist KEVIN.NYQUIST@UGA.EDU
Rory Hibbler RORY.HIBBLER@UGA.EDU
School of Computing, The University of Georgia
Athens, GA 30601 USA

Abstract

Today's music business is a multi-billion-dollar industry, with thousands of new songs coming out every year, generating $15.9 billion of revenue in 2022 [11]. Record labels are always on the lookout for hit songs to top the charts. What makes a song a "hit"? In this study, we aim to answer this question and identify the exact qualities of the ideal "popular song". To do this, we gather audio characteristic data on the top 100 most-streamed songs on Spotify, which we define as "hits". We analyze the characteristics of these songs, along with the characteristics of an equal number of less-popular songs, to try and pinpoint the prevailing audio characteristics of a "hit" song on Spotify. Using a two-part approach of Support Vector Machine and Random Forest analysis, we identify the most important qualities, such as a song's energy, tempo, and danceability, to see what makes for a chart-topping hit. Using the models we have built, we are able to predict, with 68% accuracy, the probability that a song belongs on the all-time list, based on its audio characteristics alone.

1. Introduction

Spotify boasts itself as one of top players in the entertainment streaming industry providing millions of music tracks, audio books, and podcasts to more than 574 million users in 180 markets [20]. Spotify's success in the music industry can be partially attributed to its effective use of recommendation algorithms to suggest songs to a user based on their listening history and a variety of other factors. The platform also attracts a flock of inspiring artists who seek to gain popularity and reach new audiences. As opposed to physical or digital music sales, Spotify's artists earn most of their money off of the revenue generated from the platform distributed according to the popularity of their songs in relation to all other streams. In this way artists are very interested in creating songs that can benefit from the accelerated growth the platform can offer. Although the specific details of their algorithms are a mystery, many researchers have attempted to make sense of how the platform generates user recommendations and groups similar songs together.

A growing research discipline called Hit Song Science (HSS) attempts to study the factors that most influence the growth and popularity of certain songs. This presents interesting applications that aim

to predict the popularity of a song based on a variety of physical audio characteristics derived from audio analysis methods. There are many ways to assign attributes to an audio track, and Spotify makes a few of these attributes publicly available through their web API platform. Attributes such as danceability, energy, valence, acousticness, and tempo were used in this project to create a custom dataset with a variety of popular songs from the Top 100 All Time Most Streamed Songs and 2022, 2020, and 2019's Top 50 tracks on Spotify as well as other songs from the same albums as those songs that were attributed as being less popular. Using this dataset we performed feature extraction to determine the attributes that most influence a certain song's popularity.

2. Literature review/discussion of related work

In this section we discuss the current findings in similar research involving the use of audio attributes in music data mining applications, specifically focusing on algorithms and techniques used by researchers to best design accurate music recommendation and genre and popularity classification algorithms.

Many of the most popular machine learning models have been used to tackle these problems such as Naïve Bayes, K-Nearest Neighbors, Decision Trees, Linear Regression, Random Forest, and SVM. The wide variety of audio data available through online sources has influenced researchers to attempt to narrow down the most influential attributes that link to songs' popularity and genre category. For example researchers have performed pre-processing on datasets to achieve a balance between accuracy, computation time, and model complexity by selecting the most influential features towards a song's popularity [2]. Likewise, other researchers have taken raw audio files in an attempt to extract the most influential audio features in improving the accuracy of their machine learning and deep learning models and generalize the applicability of audio features for a variety of uses such as genre classification and music recommendation [3, 8]. This can be used to generate large data sets that accurately analyze and group related music that can greatly improve the listener's experience with music streaming services, adding value to their platform. In this process, user interaction with the assigning of these attributes through active learning could greatly benefit the machine learning algorithms' ability to accurately judge music and group it into genre categories based on similar audio features.

Predicting the popularity of songs based on their audio characteristics provides tremendous value to the stakeholders in the music industry. Research into this area is concerned with developing the most efficient models for identifying trends in the similar features that popular songs possess. The approaches that utilize historical data on the most popular songs in an attempt to project the likelihood that a certain song will be popular in the future proposed innovative ways to predict the position of a song on the top tracks music charts [6, 7]. In testing the particular machine learning models that could be used for

analyzing audio attributes, models such as Random Forest (RF) and Support Vector Machine (SVM) yielded much higher accuracy than other models because of their ability to deal with high-dimensional data [1, 5]. These approaches are generally limited however by the availability of large data sets that have data for specific time periods and geographical regions [10]. Other external factors such as social media visibility and user preferences can also have a large impact on the popularity of a song.

Deep learning has the potential to improve the effectiveness of these applications over the traditional machine learning applications because of their ability to deal with large datasets, highly-dimensional data, and learn features in the data previously unseen using traditional methods. There are many benefits to using Convolutional Neural Networks (CNNs) to generate predictions and identify relevant data through feature extraction [3, 9].

3. Data Summarization

The Spotify Song Attributes dataset is composed of the Top 100 All-time streamed songs on the Spotify platform as well as less popular songs from the same albums as the top 100 songs. Additionally, the dataset contains popular songs collected from Spotify playlists of top tracks from 2019, 2020, and 2022 [16 -19]. The dataset contains information that Spotify provides such as the track and album names used to identify each of the songs as well as their corresponding audio characteristics. The audio attributes were generated from the publicly available Spotify Web API which allows users to access metrics relating to a song's audio qualities. The audio analysis features that Spotify has available includes: acousticness, danceability, duration, energy, instrumentalness, key, liveliness, loudness, mode, tempo, time signature, and valence. Many of these metrics are already normalized on a scale of 0.0 - 1.0. To properly label the data, we assigned a popular category which reads true if the song was present on one of Spotify's top track playlists and false otherwise. The partition of this dataset approximately follows the ratio 9:20 for popular to unpopular songs. Additional songs from 2019, 2020, and 2022 were added in order to provide a more even sampling of each class and avoid a class imbalance between positive and negative predictions in training our models. Our assessment of accuracy, precision, recall, and F1-score in the results section goes into further detail on the reason for these decisions.

Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Danceability	A combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity describing how suitable a track is for dancing. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

Instrumentalness	The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
Loudness	Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
Tempo	The overall estimated tempo of a track in beats per minute (BPM).
Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative.

Figure 1. Audio Attributes Generated From Spotify's Web API [22]

4. Data Preprocessing/Transformation/Feature Selection

Before conducting experimental analysis on the dataset, the first step was to clean the data. We conducted deduplication on the entire dataset to ensure there were no duplicate songs in the dataset. Then, we performed feature subset selection to eliminate features that did not contribute to increased model accuracy. In this process we visualized the data with a variety of methods to try and understand attributes that make a song popular versus unpopular as well as the correlations between certain features.

After visualizing the data, we found that the standard distribution of danceability as well as valence among popular songs tended to be higher than that of unpopular songs. This indication led us to choose them for our model in distinguishing whether or not a song is popular.

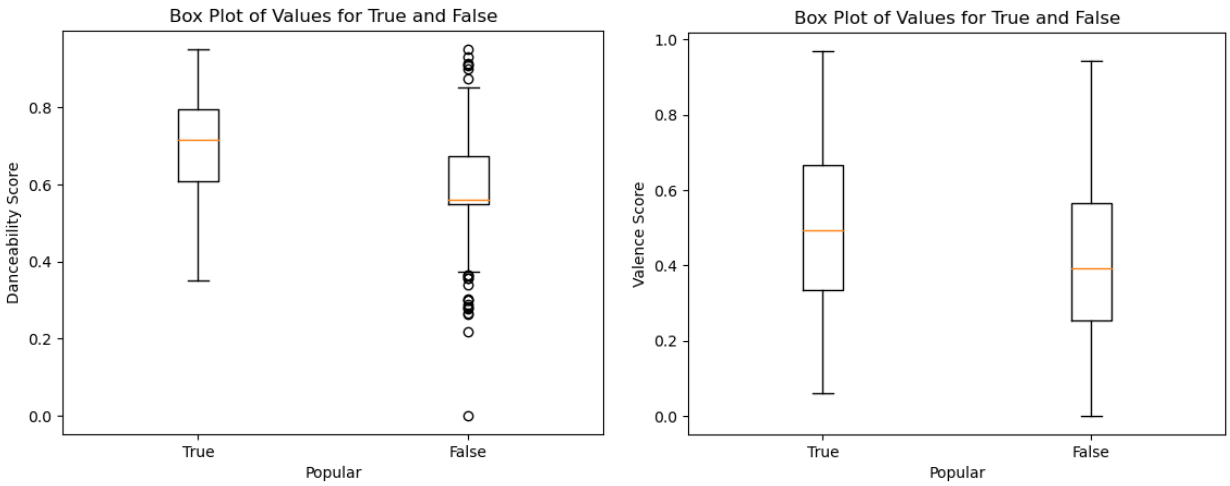


Figure 2. Standard Distribution of Danceability and Valence among Popular and Unpopular Songs

We also tested different combinations of the nine features to assess the performance of our machine learning algorithms, taking into account the accuracy of their predictions as well as the overall runtime. Narrowing down the features from nine to three greatly improved the runtime of each of the algorithms with minimal impact on prediction accuracy. The three features decided upon that yielded the greatest balance between model complexity and performance were danceability, energy, and valence. For Random Forest, the four features that yielded the best results were danceability, energy, loudness, and speechiness.

5. Discussion of Experiments/Model Development

Two models, Random Forest(RF) and Support Vector Machine (SVM) were created to compare the prediction accuracy of different machine learning models on our dataset. Ten-fold cross validation was conducted on each of the models to assess the cross validation score and select the best model for our dataset. We compare each model's accuracy, precision, recall, and F1-score to determine the best model for predicting popular songs using Spotify's generated music attributes.

Random Forest is an ensemble learning technique in which an arbitrary amount of decision trees of a chosen size are generated through choosing a random subset of attributes. For each classification task, the forest of trees takes a majority vote on what the class should be. This is known as the average prediction. The strength of the random forest model as a classifier depends on the diversity of the forest. In the process of generating the forest, the algorithm ensures that different subsets of attributes are chosen at each splitting point. Ideally, no two trees in the forest are exactly alike, which means the model explores a large variety of attribute combinations to come to a consensus on classification [14].

Support Vector Machine is a classification model that separates classes of data by learning a hyperplane, or a border, to separate instances of two classes. This hyperplane can be either linear or nonlinear. In our case, we use a nonlinear hyperplane to differentiate the two classes we are classifying. The method gets its name from the utilization of Support Vectors, which are data points near to the border between two classes. Using these support vectors, the algorithm learns where the hyperplane should be placed in order to meet the desired threshold for accurate classification [15].

Both models were chosen for their ability to deal with high-dimensionality data, because of the large number audio attributes in our dataset, and their ability to deal with outliers. Utilizing the popular sklearn Python library, we implemented the Random Forest with 100 trees and the splitting criterion of Gini score. We settled on n=100 for the forest, because beyond that number gave negligible improvement in prediction accuracy. As a side note, the Random Forest algorithm runs extremely quickly, building a classifier in under 5 seconds with a forest size of 100.

Additionally, we implemented the SVM classifier using the polynomial kernel type with the regularization parameter C=2 to produce a decision boundary that provides the best classification accuracy. This lends the advantage of being able to identify patterns from high-dimensional inputs by separating the data with non-linear decision boundaries. As compared to other kernel types, this allows the algorithm to better map each of the input songs to the feature space and determine the appropriate class.

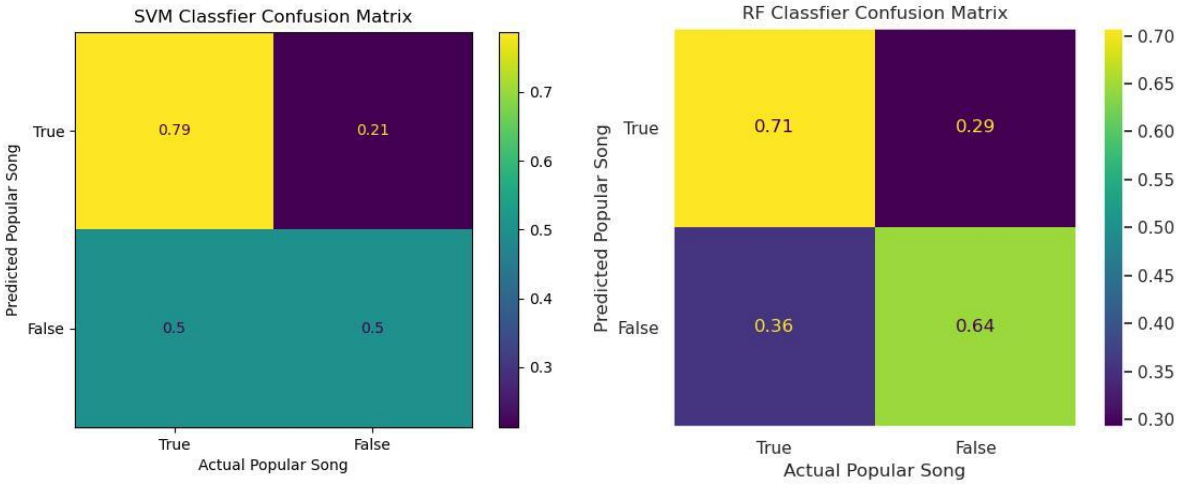
By reducing the dimensionality of our data during the preprocessing phase, we were able to significantly improve the runtime of our selected models without negatively affecting the accuracy. This step eliminated unnecessary features that in practice did not support increased model accuracy on our chosen metrics.

6. Analysis of Results

From running the models on the data set, we calculated the following results:

Machine Learning Algorithm	Precision	Recall	F1-Score	Accuracy
SVM	0.66	0.67	0.66	0.66
Random Forest	0.70	0.71	0.68	0.68

Figure 3. Classification Statistics

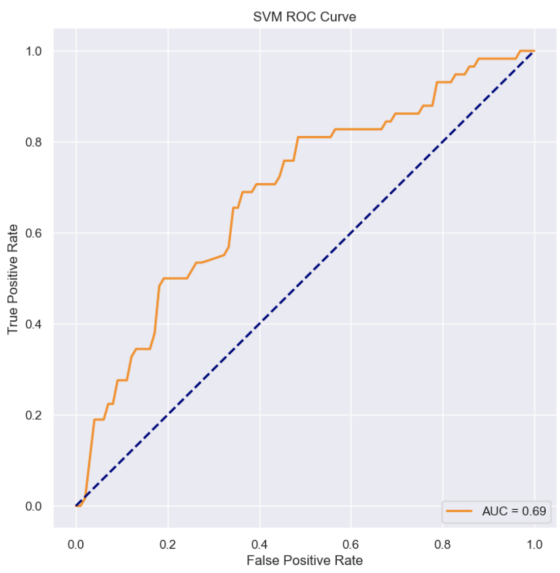


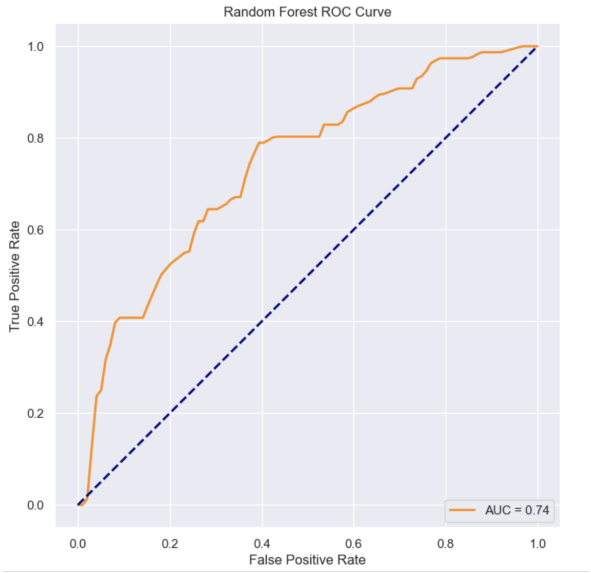
Figures 4 and 5. Confusion Matrices

First, we will look at the classification statistics. We can see that both models were able to reach an accuracy of at least 66%. This satisfies our initial prediction that we would get at least 60% accuracy on this model, which we are happy with, but we understand that important metrics such as the f1-score should also be considered to assess model accuracy. We can see that the random forest model performs slightly better than SVM.

The precision and recall of both models is above 65%, with both yielding a slightly higher score in the Random Forest model. This shows in the F1 score, where RF is 0.03 points higher than SVM. This result shows how the models perform at predicting true positives of the target class, giving insight to the potential use in predicting popular songs.

Looking at the confusion matrix reveals some interesting patterns with the two classification models. Looking at the SVM confusion matrix, it has the highest rate for true positives at 0.79, beating the Random Forest model by 8%. However, for SVM, false predictions had 50% accuracy, which brought down the accuracy score. In contrast the RF model has 64% accuracy on a false prediction. This goes to show that RF yielded a more balanced classifier overall, but SVM was better at identifying true positives in the data set in practice.





Figures 6 and 7. ROC Curves

The Area Under the Curve (AUC) score shows a measure of the true positive rate versus the false positive rate in predicting our target class. In our analysis we found that random forest had a higher AUC of 0.74 versus SVM which had a score of 0.69. Our ROC curve provides a graphical way of understanding the accuracy of our models in predicting a song's popularity as compared to random chance. Both models prove moderately effective at the task.

7. Future Work: Discussion of limitations and possible extensions of work

Our model, while effective in predicting top hits with moderate accuracy, falls victim to certain limitations concerning its data set. First and foremost: the "all-time" sound. Our model is simple; it takes the sounds of the Top 100 most streamed songs of all time on Spotify as well as other popular tracks and isolates the characteristics of those songs. Looking deeper, these songs vary across several genres (Hip-hop, Pop, Rock, Folk) as well as more than a decade of variation between the songs labeled as popular. Since the mix is so eclectic, there is a lot of noise present in our data set, which limits our ability to accurately predict what songs are popular. Nonetheless, some important characteristics shine through which are present in all popular songs, namely danceability, energy, and valence (positivity).

Additionally, a further limitation of our research is its usefulness for artists. Spotify does not make its algorithms for calculating features like danceability, energy, or valence open to the public. Due to this, it is hard for an artist to pinpoint the perfect mix of a danceable, energetic, and positivity when they are producing it. Much like in sports, these qualities are the "intangibles" of a good song; an artist

can only guess at what these qualities signify, “aim for them”, and hope for the best when creating a song. It would likely benefit this model to add more concrete measurements of songs, such as lyrical content, as well as other factors that contribute to a song’s success, such as marketing statistics, pre-existing artist success, or information about the intended audience. In future works, these additional characteristics could make for a very robust prediction algorithm.

Lastly, one pattern that our model did not account for is the “star effect”. Multiple artists, such as Taylor Swift, Ed Sheeran, Justin Bieber, amongst others, have much higher presence in the Top 100 list than any other given artists. By nature of these artists already being stars, it is obvious that their music is more likely bound for success than lesser known artists. A comprehensive predictor of a song’s success must, at very least, take into consideration the size of an artist’s fanbase. Otherwise, a pop song that resembles something recorded by Taylor Swift, might be predicted to be a smash hit when, in reality, the “star factor” is not there to help the song succeed. Future work should consider this factor in creating an accurate popularity predictor.

8. Statement of Division of Work

- Kevin:
  - Introduction, Literature review, dataset generation and cleaning, SVM model experiments, dataset visualization, ROC curve analysis
- Rory:
  - Abstract, Random Forest experiments, Results analysis, Future Work

9. References

[1] Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify

[2] Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms

[3] Music Genre Classification and Recommendation by Using Machine Learning Techniques

[4] Combined Transfer and Active Learning for High Accuracy Music Genre Classification Method

[5] Music Popularity: Metrics, Characteristics, and Audio-Based Prediction

[6] Predicting Music Popularity Using Music Charts

[7] Musical track popularity mining dataset: Extension & experimentation

[8] Music Feature Extraction and Classification Algorithm Based on Deep Learning

[9] Music Stream Analysis for the Prediction of Song Popularity using Machine Learning and Deep Learning Approach

[10] Predicting Song Success: Understanding Track Features and Predicting Popularity Using Spotify Data

[11] RIAA 2022 End-of-Year Revenue Report

[12] Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify
[13] Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms
[14] Introduction to Data Mining. Pang-Ning Tan. pp. 512-513
[15] Introduction to Data Mining. Pang-Ning Tan. pp. 478-486
[16] Spotify Top 200 Songs All Time (# of streams)
[17] Spotify Top 50 Songs of 2022 (# of streams)
[18] Spotify Top 50 Songs of 2020 (# of streams)
[19] Spotify Top 50 Songs of 2019 (# of streams)
[20] Spotify Newsroom
[21] Spotipy Python Library
[22] Spotify's Web API Audio Attributes Documentation