# STAT 451 Project Proposal: Predicting Spotify Song Popularity with Machine Learning

Anantha Rao
anrao3@wisc.edu

Xingrong Chen
xchen942@wisc.edu

Desmond Fung
dfung2@wisc.edu

## 1. Introduction

Do we know a popular song when we hear one? While we may be able to approximately tell for ourselves what songs are popular or catchy, how might we statistically predict which songs will be popular? One of the many facets of human life where there is an abundance of data is in music. In this project, we use a Spotify data set with 19,000 songs to uncover the relationship between the features of a song and the song's popularity. From this analysis we will create a model which can predict a song's popularity.

Previous work has found some features and attributes about songs that do and do not predict popularity. One such study examining the impact of lyrical content on popularity found that lyrics have little power in predicting popularity[4]. This same study seemed to suggest though, that the song's genre is an important predictor in how popular it will be.

Other studies have found time to be a relevant factor. For example, a study at Stanford found seasonal effects; discovering that while summer months(June- August) are not necessarily more likely to produce popular music than holiday months(November - January) it appears the features of popular music are slightly different between the two seasons[3]. They also discovered that popular songs in roughly the same time period are more similar than songs before and after that time period; suggesting that popular songs tend to change together over time.

Our research will examine spotify-defined features related to the musicality and sound qualities of songs. Features related to the sound such as song duration, acousticness, loudness, tempo, etc. are the things we are interested in here; as well as the connection they have (if any) to song popularity. If there is a connection between these features and popularity, we hope to identify a model or an ensemble of models to which will both predict the next big hit, but also bring light to the attributes of a song that humans find most attractive.

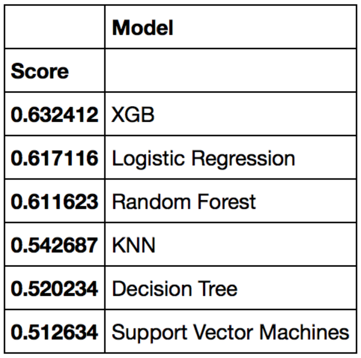
## 2. Motivation

According to Billboard [2], the music industry alone generated \$11.1 billion in revenue in the United States in 2019. Having the ability to predict how popular a song will be, in this fast-paced and competitive industry, will not only allow producers to determine which factors are most important in determining the next greatest hit, but from a monetary perspective, it could also help investors to decide which musician and songs to invest in for the greatest return. Before releasing a body of music, record labels or artists could consult data experts in music for advice on gauging how much money to spend on advertising and promotion of -say- an album based on the predicted attention a track would get.

Additionally, as a side-effect of this study, we learn something more generally about human listening behavior in the realm of music consumption. We'd learn what kind of attributes of a song appeal to people in aggregate, and gain some insight into the sounds in music that naturally attract positive attention. Musicians may even gain insight into what the consumer listens to most intently and this could inform music production as well as distribution. If our model is interpretable, we could tell the precise effects of each feature on determining the popularity of a song. This would greatly help in forming general theories or 'rules of thumb' about music when it comes to writing music.

Thanks to the growing popularity of streaming platform such as Spotify and Apple Music, data can now be easily reached and user behaviors are accessible for research. Using a variety of machine learning techniques including Decision Trees, Random Forest, and the kNN algorithm, we can now address the questions of what makes a song popular and examine the largest component of a song's success.

The main motivation of our project is to create a model that uses musically relevant features to accurately predict song popularity. We will be implementing many machine learning algorithm, for example, kNN or Naive Bayes classifier, etc and see which model return the highest prediction accuracy. If implemented correctly, we will be able to use this model to determine whether a new song will be popular based only on the key features of a song. [5]



|  | **Model** |
|---|---|
| **Score** |  |
| **0.632412** | XGB |
| **0.617116** | Logistic Regression |
| **0.611623** | Random Forest |
| **0.542687** | KNN |
| **0.520234** | Decision Tree |
| **0.512634** | Support Vector Machines |

Figure 1. There have been many attempts of using various machine learning algorithm to predict if a song is a hit song or not, including logistic regression, random forests, kNN, and more. [5]

## 3. Evaluation

The dataset we will use is the ”19,000 Spotify Songs” dataset provided by Edwin Ramirez on Kaggle [1]. It contains 19,000 songs that were released before 2018 as well as 15 unique features like loudness, dance-ability, and acousticness etc that we will be analysing. We will split our data into training, validation, and testing sets using the ”train test split” method from the sci-kit learn package. While the values for label ”song popularity” in our initial dataset ranged from 0 to 100, we will likely convert this to a binary variable with the arbitrary threshold of popularity $\geq 60$ labelled as 1 and otherwise 0.

| Feature | Value range |
|---|---|
| song popularity | [0,100] |
| song duration(ms) | [12.0k,1.80m] |
| acousticness | [0,1] |
| danceability | [0,1] |
| energy | [0,1] |
| instrumentalness | [0,1] |
| liveness | [0,1] |
| key | [0,11] |
| loudness | [-38.8,1.58] |

Table 1. List of Value Range of features extracted from the dataset

A successful project here would be finding a model or an ensemble of models which accurately predict the popularity of a song. We hope that we can even supersede the accuracy of the models and methods used in the standard kaggle notebook shown above. If we can randomize our training, validation, and test sets each time and achieve perfect training accuracy, good validation accuracy, and ultimately a low test error, we will consider our model/ensemble of models to have attained our goal.

An added success from this investigation would be to have some level of interpretability in our model so as to gain a more human understanding of the relevant features that make a popular song. Then, rather than having a black box predictor which only tells us ”popular” or ”unpopular”, we have some level of understanding as to why a song isn’t popular on a feature level. In this way, the creation of an interpretable model would be useful to consumer scientists/behavioral analysts because the models could be leveraged to develop theories of listening behavior. With more data and expertise in those areas, a model produced here could inspire prediction of overall music trends over the years.

Before we formally define success in our model, we should first re-emphasize that we’ll be converting the ’song popularity’ variable to a class label with 1 and 0 as defined as songs with ratings $\geq 60$ and $< 60$ respectively. Then to quantify success in the aforementioned areas, we would first examine training accuracy between our predictions and class labels. Then we’d look at the validation error of our model to see if it generalized well. We could then run a test set through our model to evaluate its prediction error.

Various model accuracy and feature importance (using XGBoost) will be computed, where the accuracy is the number of accurate predictions we make compared to the true label of the track. The F score computed from XGBoost allows us to determine which feature is promising to the model accuracy. We may consider our project to be successful if we achieve more than 80% test accuracy on any of the machine learning models and the F score is at least 80.

## 4. Resources

As stated above, we will use the dataset mentioned before and we will likely only require the computational power of our personal laptop computers.

In terms of data preparation, we will use a combination of R and python to visualize the data. The R package ggplot2 has many useful data visualization tools and will likely be leveraged for getting a sense of the data before all the machine learning components. We will then be using python and sci-kit learn to partition the data into training, validation and test sets, train a model and then eventually make a pipeline which includes more trained models if we see fit.

We may also want to do more investigation into what other features make a popular song. It remains to be seen if we will use the python package ’spotipy’ in order to get more data from the Spotify API. If we end up doing this, we will use that in order to gather more data on other features like artist who wrote the song and year the song was written. This may open up new avenues for predicting song popularity as it may become relevant to conduct an investigation on whether music trends are different in the last two years as they've been in the original time scope of our investigation. This may identify and give weight to the hypothesis that popular songs in different time periods have different characteristics.

## 5. Contributions

For the implementation of the project, Desmond will be responsible for building the machine learning model, ensuring the functionalities of the models align with our goals. Xingrong will be responsible for removing incorrect or corrupted data within the dataset while ensuring the data being collected was accurate. Anantha will be responsible for helping with coding up the models as well as producing visualizations of the data. Overall, however, everyone will be responsible for ensuring the successful development of the project. Everyone will collaborate to some effect on all parts and this will be roughly tracked in GitHub commits.

When it comes to the report write up, Anantha will be structuring the paper and creating figures to display. Desmond's focus will center on evaluating the model's accuracy and conducting analysis on features that make a song popular. Lastly, Xingrong will focus on the conclusion and analyzing key takeaway from the project.

Again, everyone will help out where it is needed on any of the fronts of this investigation. Whether it is model building, interpretation, making figures, or writing the report.

## References

[1] "19,000 Spotify Songs Dataset". In: (). URL: https://www.kaggle.com/edalrami/19000-spotify-songs?select=song_data.csv.

[2] "Billboard - US Recorded Music Revenue". In: (). URL: https://www.billboard.com/articles/business/8551881/riaa-music-industry-2019-revenue-streaming-vinyl-digital-physical#:~:text=The%5C%20U.S.%5C%20recorded%5C%20music%5C%20business,billion%5C%20it%5C%20reached%5C%20in%5C%202018.%7D.

[3] "HITPREDICT: PREDICTING HIT SONGS USING SPOTIFY DATA". In: (). URL: http://cs229.stanford.edu/proj2018/report/16.pdf.

[4] "Predicting A Song's Commercial Success Based on Lyrics and Other Metrics". In: (). URL: http://cs229.stanford.edu/proj2014/Angela%20Xue,%20Nick%20Dupoux,%20Predicting%20the%20Commercial%20Success%20of%20Songs%20Based%20on%20Lyrics%20and%20Other%20Metrics.pdf.

[5] "Song Popularity Predictor AUC scores". In: (). URL: https://towardsdatascience.com/song-popularity-predictor-1ef69735e380.