

# Music Popularity Predictor

Sam Springer

Computer Science

University of Colorado Boulder

sasp7990@colorado.edu

Sean McCormick

Computer Science

University of Colorado Boulder

semc2429@colorado.edu

Jenna Dean

Computer Science

University of Colorado Boulder

jede4828@colorado.e

Ryan Power

Computer Science

University of Colorado Boulder

ryan.power@colorado.edu

## PROBLEM STATEMENT

This day and age, Spotify, Pandora, and Apple Music have a limitless selection of streaming music that gives listeners the ability to easily and quickly connect with their favorite artists and songs--by just a click of a button. These platforms also generate music recommendations based on the likes and dislikes of each listener. For example, Apple Music generates a playlist of roughly 25 songs each week for a listener based on the type of music the listener listened to the previous week.

Beyond the ingenuity these platforms offer to listeners, these platforms likewise help the music industry understand what listeners prefer, and, as such, create music that is catered to these preferences. In particular, Spotify offers a web API that allows developers to collect various attributes, such as energy, tempo, danceability, of a song. Given the ability to harness these attributes from Spotify, we plan to create a model that can successfully predict the popularity of a song, or even artist, based on the quantitative and qualitative characteristics obtained from a song and/or artist. This model aims to perform the following: (1) predict the popularity of a current song or artist (2) predict the popularity of a new song or artist based on correlating the attributes to those of a popular song or artist, respectively; and (3) predict popularity of a song or artist based on a combination of genres that are popular.

Implementation of this model will be of great benefit to the music industry. It will assist the music industry in better understanding what attributes, e.g., danceability, tempo, energy, it should focus on when creating a new song or even when signing a new artist.

## LITERATURE SURVEY

In the late 1990s, music platforms allowed music discovery through music recommendations, ratings, etc., based on the likes of other listeners. Pandora and the Music Genome Project were the impetus of taking a more analytical approach to music discovery by analyzing the structure of a song so as to discover similar songs that a listener might like. [1]. Following the Music Genome Project launch, many researchers have adopted this analytic approach of analyzing the characteristics of music to provide listeners with songs that are similar to the characteristics of music they have previously listened to. For example, Elbir and Aydin put forth a recommendation engine that classifies music based on acoustic characteristics, which in turn recommends music having similar acoustics to listeners. [2]. A number of other recommendation engines, such as Li and Tzanetakis [3] and Holzapfel and Stylianou [4], also use acoustic characteristics to recommend music to listeners.

## PROPOSED WORK

For data collection, we plan to use several datasets available on Kaggle<sup>1</sup> as well as data collected from the

---

<sup>1</sup> <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>

Spotify API. We will combine the data and use the Pandas Python library to transform the JSON formatted data from the API into easily manipulated and searchable DataFrames. To ensure data quality, we will check for completeness by using attribute means or regression to fill in any missing values where appropriate or deleting the data object if necessary. We will also search for and eliminate all duplicate data and outliers, such as artists and songs with a popularity value of 0. To reduce dimensionality and work only with data that will provide meaningful insights, we will delete those attributes highly correlated with other attributes. To derive data and discover patterns, we plan to use k-means clustering which will allow us to group related data into distinct clusters, and the Apriori algorithm to generate frequent item sets.

Our process for predicting artist and song popularity will differ from prior work done by organizations such as Pandora and its Music Genome Project in that we will not be recommending artists based on users' listening histories. Unlike The Music Genome Project, whose predictive method "responds to each individual's tastes [5]", we will be attempting to predict overall popularity through song characteristics alone. Our process for determining popularity will also differ from Spotify's, which they describe as being "based, in the most part, on the total number of plays the track has had and how recent those plays are [6]." Despite these differences, we will use Spotify's popularity index as a gauge to evaluate our results.

## DATA SET

Our dataset is comprised of fifteen attributes and 28,680 data objects, each of which represents a unique artist. The dataset contains multiple characteristics for every artist available for streaming on Spotify, such as tempo and song length, which were obtained through averaging the song characteristics of an artist's entire body of work. Thirteen of the attributes are numeric, while the remaining two, mode and genre, are binary and nominal, respectively. The mode attribute is asymmetric and skewed towards major with the majority of songs written in major scales. Eleven of the thirteen numeric attributes are floats, while the last

two attributes are represented by integers. Among the numeric attributes, the majority are continuous, ratio-scaled and normalized between 0 and 1. In addition to the dataset described above, we will also be querying the Spotify API in order to obtain more detailed and song specific information such as release date, individual song characteristics, and song popularity. This will be used to supplement our dataset and will allow us to ask and answer meaningful questions about both artists and songs.

## EVALUATION METHODS

Once the data is generated, cleansed and analyzed, our goal is to utilize the song attributes to predict artist and song popularity. We will evaluate our success by comparing the songs and artists that we deem as popular to the ones that are considered popular by Spotify. As mentioned above, Spotify's popularity is "based, in the most part, on the total number of plays the track has had and how recent those plays are", so we will compare our songs and artists to the Spotify algorithm to determine whether our analysis was accurate [7].

## TOOLS

There are many tools that we plan to use for this project. We will use both Kaggle and Spotify to supply our data. Kaggle is a machine learning and data science community where users can publish, explore and build data models [5]. As stated above, the Kaggle dataset contains a multitude of attributes and objects. Alongside the Kaggle dataset, we plan to use the Spotify API in conjunction with a Python script to generate song/artist attributes and features. Once the data is generated, we will put it into DataFrames in order to analyze and manipulate it.

Once we have established our formatted dataset, we plan to use NumPy, Pandas and mlxtend - Apriori Python libraries to begin our analysis. The NumPy and Pandas libraries will assist in performing mathematical operations (such as minimum, maximum, average, etc.) on the attributes, and we will also use them to generate charts on the fly. These charts will be useful in determining whether two attributes are correlated

and viewing the max, min, quantiles, etc., within the various attributes. The mlxtend - Apriori will be used to generate frequent item sets to show general trends within the data.

Lastly, we plan to use Tableau. Tableau is a software used to create dataset visualizations. Our hope is that we can cleanse the data using Python and then upload the cleansed data via a CSV file into Tableau. Once the data is in Tableau, we will create charts and graphs where users can easily view the song and artist popularity based on song attributes.

## **MILESTONES**

As far as milestones go, we plan to make weekly progress. Our first milestone will include converting the data into DataFrames. We will plan to have this milestone done by next week. From there, our second milestone will include analyzing the data to find correlations. This milestone will also include clustering the data and leveraging the Apriori algorithm to generate frequent item sets. Our third milestone will utilize the cleansed data within Tableau to create visualizations for users to easily visualize the results.

## **REFERENCES**

- [1] John Joyce: 'Pandora and the music genome project: song structure analysis tools facilitate new music discovery,' Scientific Computing, 2006, 23:43.
- [2] Elbir, A., Aydin, N: 'Music genre classification and music recommendation by using deep learning,' Electronics Letters, 2020, 56:627-629.
- [3] Tao, Li, Tzanetakis, G.: "factors in automatic musical genre classification of audio signals," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003, 19-22.
- [4] Holzapfel, A., and Stylianou, Y.: 'Musical genre classification using non-negative matrix factorization-based features', IEEE Trans. Audio Speech, Language Process, 2008, 16:424-434.
- [5] 'About The Music Genome Project', Pandora Media, Inc. <https://www.pandora.com/corporate/mgp.shtml>
- [6] 'Web API Reference', Spotify AB. <https://developer.spotify.com/documentation/web-api/reference/#object-trackobject>