Identifying Features That Make a Song Popular on Spotify

Kevin Van Vorst (kpv23), Manya Walia (mtw62), Saaqeb Siddiqi (ss3759)

October 4, 2020

Background

Spotify is the major audio streaming service of the modern day, and as of July 2020, it boasts of almost 300 million active users all over the world. Through a platform following the freemium structure, it provides access to over 50 million songs. Because of the rise in its popularity, Spotify also has the means to monetarily compensate artists for their contributions to the service according to the success of their music. Owing to this, it would be very beneficial to assess what features and/or combination of audio features determine the popularity of a particular song to what extent, and improve the user's experience by providing recommendations.

Question

Given a song and its features, can we accurately determine whether or not it will be popular, and recommend other similar songs to the user?

Proposal

When studying the arts, it can be observed and analyzed why certain crafts stand out and become popular amongst the masses. While studying music, it could be observed how a certain beat or tone could elicit a particular feeling or mood within the population. If the music is discovered and shared across a network of people, there definitely is a reason beyond just simply an artist's current noticeability. Take Taylor Swift's latest album *Folklore*: there are unique reasons as to why the song "cardigan" has been liked more on Spotify than "betty". Since there are many numerical, categorical, and binary data within the Spotify dataset to list the tempo or the danceability, it can surely be analyzed through the lease of supervised learning to come to a conclusion about which features are correlated and possibly affect the popularity of a song. Given this understanding, an analysis done on this dataset will surely give us a satisfactory conclusion listing the unique characteristics that make a song popular. With this newly found information, a recommendation system can surely be found based off of songs with similar qualities.

Data

The dataset we will analyze consists of over 160,000 songs throughout 1921 to 2020. All the tracks and its information were originally collected directly from the Spotify Web API. Every song has a name, list of artists, release date, key, and is identified by a unique ID generated by Spotify. Each track is then decomposed into 12 numeric audio characteristics, some of them including: acousticness, danceability, popularity, energy, and speechiness. In addition, there are two columns indicating explicitness and key type (major or minor), represented by a boolean variable. In total, the size of the dataset is 169,909 by 19.

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data.csv