Predicting Song Popularity

Sam Xu *samx@stanford.edu*, Joyce Xu *jexu@stanford.edu*, and Eric Tang *etang21@stanford.edu*

***Abstract*—The project aims to predict the popularity of a music by analysing music features such as tempo, duration, mode, loudness, key - arist information such as artist name, location, popularity - and meta data such as release title and year. The codebase for the project will be located at `https://github.com/inSam/SoundScorer221`**

I. Introduction

The music industry is a multi-billion dollar industry in which over a million song is produced in the United States each year alone. We would like to predict the hotness or success probability of a song through its features prior to release. It will be useful for the music industry as a baseline advisor, and understand what features of songs are currently popular with the public.

II. Related Work

Myra et al 2018 [1] performed an analysis on over 500,000 songs released in UK to correlate success (if it appeared on the top 100 hitlist) using accoustic features and the popstar status of its artist. To perform the study, they used random forest as their predictive method in which a single classification tree is used. Without the super-star feature, they were able to obtain about 70% success rate in classification and about 80% success rate with the super-star feature.

In a previous CS229 project, Pham et al. attempted to predict song popularity using state vector machines, multi-layered perceptrons, and logistical regression in 2015. Pham et al. utilized auditory features such as "danceability" and "energy" derived from The Echo Nest's proprietary algorithms to extract auditory features. Their system performed moderately well with an MSE error of about 0.15 over a popularity index between 0 and 1 [2].

Yang et al. attempted to predict song popularity by using convolutional neural networks (CNNs) to extract complex auditory features from songs. They find that features extracted using CNNs outperform features solely extracted using simple linear-regression techniques on each audio file. [4]

III. Methodology

A. Dataset

Our dataset is a random subset of songs from the past ten years extracted from the Spotify API, which provides acoustic features, author information, and meta data of the songs. More abstract features such as dancebility, energy, and song popularity were pulled from the Echo Next, whose database is incororated into the Spotify API.

As preliminary data, we extracted a small subset of songs from the Million Song dataset, which contains a plethora of pre-extracted features that are also provided by the Spotify and the Echo Nest API. We furthur narrowed the dataset by only using songs that contained the following features:

- tempo, duration, mode, loudness
- artist name,
- song popularity

The features noted above will be the features we use to run our preliminary baseline and oracle tests on.

B. Concrete Example of Input and Output

Here you can see an an Input and Output example for the song *Never Gonna Give You Up*:

$$
\begin{array}{ll}
\text{"}TRAXLZU12903D05F94\text{"} : \{ & \\
& \text{"}features\text{"} : \{ \\
& artist\_name : RickAstley \\
& loudness : -7.75 \\
& tempo : 113.359 \\
& duration : 211.69587 \\
& mode : 1 \\
& \}, \\
& \}
\end{array}
$$

The output is simply the popularity/hotness rating, or in this case 86, out of a popularity rating from 1 to 100.

C. Implementation

We started by importing the spotify dataset [3] which allows us to quickly explore the dataset and to speed up development. We first built a framework that allowed us to extract features through the spotify_api.

As our baseline experimentation, we used Linear Regression to predict song popularity in a test set of 1,500 songs extracted from the Spotify API.

D. Evaluation Metric

One of the things we noticed right away is that is that popularity is extremely time dependent. Therefore to narrow the scope of our project, we decided that our metric and predictor only serve to estimate the current popularity.

We use the L2 norm as our current evaluation matric.

$$\sqrt{\frac{\sum_{i=1}^{n}(\text{Predicted}_i - \text{Actual}_i)^2}{n}}$$

IV. Preliminary results

A. Baseline

For our baseline, we trained a linear regression classifier on seven features already stored in the Spotify database: ”acousticness”, ”danceability”, ”energy”, ”loudness”, ”speechiness”, ”tempo”, ”valence”]. We pulled 10,000 songs from Spotify’s database of 2017 releases, and used a 85-15 train/test split. Our linear regression classifier achieved a mean squared error of 61.1 over a popularity ranging between 1 and 100.

B. Oracle

For our oracle, we simply allowed our program to look at the correct answer and return the true value. As a result, our oracle was able to obtain an total loss of 0 using the L2 norm as our accuracy metric.

V. Next Steps

A. Improve Features

We will extract more features from the Spotify database such as song age, location, audio sections, and signature. We will also try to construct more advanced features from historical data such as musical trends. Using Spotify’s provided links to song previews, we may be able to extract more advanced features from the audio recordings of each song. This step will involve a lot more feature tuning and experimentation.

B. Improve Models

For future models, we will try to tackle this problem by using SVMs and neural networks to predict popularity. In particular, neural networks may be able to extract more meaningful features from the audio tracks of each song.

VI. Challenges

One challenge is to consider the effects of time on the popularity on the song. We may be able to tackle this by simply adding the age of the song as a additional feature.

Another challenge is that there is plenty of noise that influences a song’s popularity, beyond its mere auditory features. While we have assumed a 100% success rate with our Oracle, this may be unrealistic to achieve with a classifier.

References

[1] Myra Interiano, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu, Natalia L. Komarova, *Musical trends and predictability of success in contemporary songs in and out of the top charts* Royal Society Open Sciences, 2006.

[2] James Pham, Edrick Kyauk, Edwin Park, *Predicting Song Popularity* Stanford University CS 229, 2015.

[3] `https://developer.spotify.com/documentation/web-api/`

[4] Li-Chia Yang, Szu-Yu Chou, Jen-Yu Liu, Yi-Hsuan Yang, Yi-An Chen, *REVISITING THE PROBLEM OF AUDIO-BASED HIT SONG PREDICTION USING CONVOLUTIONAL NEURAL NETWORKS* Academia Sinica, Taiwan, 2017.