Do Popular Songs Endure?

Prasoon Rai , Sushant Ojal

110921754 ,   110944445

{prrai,sojal}@cs.stonybrook.edu

1   Problem Statement

The aim of this project is to determine if there is a way to predict the long term endurance of a song, and to predict based on a set of recent observations if a song will indeed prevail for a significant span of time. Although its intuitively difficult to know if a song will stay, we have many indicators of what's most likely to stick namely and most notably, enormously popular music that's meant something to somebody. There are other factors such as those of songs which have heralded a huge shift in popular taste, along with the artists who produced them.

But despite all of this, to judge the staying power of a song is still a largely subjective opinion. It's inevitable that most people who encounter the question will declare with certainty that the artists/music they hate will have no staying power at all. We hope to bring some quantification to this debate and predict with reasonable accuracy if a song will prevail. The following sections are a short description of the steps we have taken so far, and largely the path we propose to take henceforth. The current progress of our work can be found at: `https://github.com/prrai/Data_Science_Do_Popular_Songs_Endure`.

2   Data Collection

In order to construct a endurance score for each song, we would require meta data about the song, its artist(s), and most importantly, trend information about it, i.e., factors that indicate the song's popularity over the years. In this view, we have looked into the following resources for obtaining the required data:

1. **Universal Music Database:** We plan to use this to obtain Billboard rankings over the years for popular songs. The web-site provides weekly Billboard top 100 songs along with the number of weeks that song stayed in top 100, its peak position, entry into the Billboard and more. We have already implemented the scraper and collected data from 1959 to 2017 for weekly Billboard top 100 from here.

2. **Billboard.com:** The billboard website provides top 200 albums and artists per week, besides giving information on the number of songs per artist that made to Billboard top 100, 10 and peak (number 1). We have implemented scraper for this collection as well. We additionally plan to use the top 100 artists per week information to create a artist popularity score based on trends in her ranking on billboard.

3. **AZlyrics.com:** We plan to analyze the lyrics of each song to figure out the capacity of that song to endure. The text analysis for instance should provide a higher score to songs that have elements of events that happen periodically, like Christmas, or

patriotic songs or references to famous historical events in the lyrics. Towards the same, we will be collecting the lyrics from azlyrics.com .

4. **Wikipedia:** In order to get meta data about a song, we will use wikipedia which provides original release date, record label, duration of the song and finally, its genre.

5. **YouTube:** To get the online presence further, we will use youtube to get the number of views for top three search results by relevance for each song. We are looking for a way to get the views per year, so that we can comment on the trend of popularity of the song.

3   Approach

This section highlights the features we will be using to build out training set as well as the factors we consider crucial to our final model.

3.1   Features

After scraping the data as highlighted in the previous sections, we have the following set of features:

Features considered: -

Title: The name of the song.

Artist: The band or vocalist who created the song.

Entry Date: The date the song made it to the Billboard Top 100 for the first time.

Peak Position: The highest rank the track made in the charts.

Total Weeks: The number of weeks the song stayed in the top 100 Billboard Charts

TopSongArtist: The number of songs of that artist that have topped the carts.

TopSongArtist10: The number of songs of that artist to have made it to the top 10 of the charts.

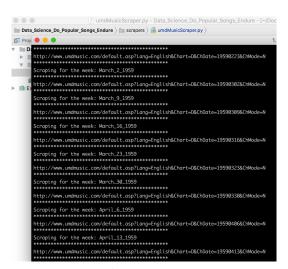TopSongArtist100: The number of songs of that artist to have made it to the top 100 of the charts.

Figure 1: UMD Web scraper, running on console

Figure 2: Sample of the created CSV of features

Below are the features we are planning to add to this:
Release Date: of the song.
Length of Song
Genre of Song
Record Label of the Song
A set of classifier IDs for the lyrics of the song.

3.2 Factors influencing endurance of a song

We consider the following factors which should impact the predictions. The feature set are therefore constructed keeping these in mind:

- The popularity of the song in the initial days of the song: It is more probable that a song which has been popular in its initial days of release to be more enduring than a song which has not. It does not however guarantee this.

- The resurgence of the popularity of a track: If a song has made it back to the Billboard top 100 after falling out of it, there is a strong chance that it will endure. This factor can also be decided to some extent by the online presence of the song long after the song has been released. Scraping YouTube for the number of views for a certain time range can be used to estimate this.

- The impact of the artist creating the song: An artist who consistently produces massive hits is more probable of producing a timeless piece of music than another.

- Demographics: Each generation latches on to the music of its youth, so each era is bound to produce songs that are integral to the memories of millions. We plan to consider the number of people who were in the age range of 15-30 years at the release date of the song.

We propose to create an **Endurance Score** based on the above mentioned factors. This score is an indicator of how well the song has prevailed over time, and is essentially the

output Y in our training model. For any given song, we will create a set of features based on the data for an interval of time in the immediate past. The output from the trained model based on this feature vector will indicate the endurance of the song.

4 Anticipated Challenges

4.1 Data Collection

To study trends, we will also need information on number of copies of the song sold per year on itunes, amazon music, number of plays on Spotify. While we can get the total number of these, more useful would be to get these numbers per year, to encode the trends in our endurance score.

4.2 Scoring recent songs

It is going to be difficult to analyse songs that are relatively recent because there will not be sufficient temporal information to analyse the trends that are being shown by this song. Making a prediction on such songs is going to be challenging.

4.3 Impact of trends and other industries

There is a difference in which music was and is acquired. Sales data is more accurate because the people who bought records were more interested in a song or album. Youtube views are only an indicator of how curious people might have been in a certain song. Furthermore, there is a huge impact of other industries, such as fashion industry, and example of which is the resurgence of retro designs, which inspire not only dressing but also listening to a certain kind of music.