Data Analysis Project Concept

David Baumgartner, Maximilian Danelli, Nico Vergeiner

November 2022

Contents

1 Research question/goal

Can we predict the chart score in Spotify Top 200 by looking at the position of the song in Spotify Viral 50 of the previous week/month?
Can we use Billboard Hot 100 to verify our results?

1.1 Why is this question interesting and challenging?

The music industry has been a very prominent business for a long time already. The appearance of music streaming services has been a significant shift in how music is brought to the consumer. This paradigm shift has affected how charts are formed, since prominent record labels have lost their monopoly position in music advertisement. Users are more self sufficient in music discovery with the help of streaming services.
The shift to music streaming begs the question if popularity of certain songs can be predicted. For this reason, we decided to access data from one of the biggest streaming services **Spotify**. Spotify's Viral Top 50 list is calculated by occurrences in social media platforms and therefore should be a good indicator on how popular a song is getting. This should be reflected in the Spotify Top 200 list. We therefore try to predict the *Top 200 charts* from the *Viral 50 charts* on Spotify.

2 Related work

A substantial body of work can be found in this area of research. In this Section, we will show a few selected bodies of work which are strongly related to our project topic.

2.1 Predicting Music Popularity on Streaming Platforms

In this paper [1], the authors also work with the streaming platform *Spotify*. They attempt to predict whether a "spike in popularity will translate into long term growth" [1]. They achieve this by considering songs appearing on the *Most-Popular* page as having a spike in popularity and songs on the *Viral* page as songs which are experiencing long term growth. Furthermore, they create a data structure for each day containing a list of 50 songs from the *Most-Popular* page and a list containing 50 songs from the *Viral* page. They then check for every song in the *Most Popular* list whether it appears in the *Viral* list k days later, and do the same prediction attempt the other way round; Predicting for each song in the *Viral* list, whether it appears in the *Most Popular* list k days later.
The authors propose their own baseline approach, which considers how many days a song has been featured in the *Most Popular* list, and predicts that it will appear in the *Viral* list in the future if the number of days for that song exceeds the median number of days for all songs. They extract the following features from each song for prediction:

- Mel Frequency Cepstral Coefficients
- Spectral Centroid
- Spectral Flatness
- Zero Crossings
- Tempo

They used these features for prediction in a **Support Vector Machine**.
The authors claim that their predictions using only acoustic features performed the worst, indicating that acoustic features are not necessary for predicting song popularity. Furthermore, they claim past performance of a song predicts future popularity growth the best.

2.2 Using twitter to predict chart position for songs

The authors of this paper [4] collected data from Twitter in order to predict which songs and artists appeared in the *Billboard Hot 100 Charts*.
The authors extracted tweets containing the *nowplaying* hashtag. They performed sentiment analysis on these tweets and created a Database which can be queried by providing a *Track Name* and *Composer*. They combined the data from their Twitter Database with Chart data from *Billboard Hot 100 Charts*, and exported the result as an .arff file. In this file, each data instance represents one specific song for a certain weeks charts. The following attributes are contained in the file:

- Previous Position: Position at previous week
- Chart-weeks: Number of weeks at top 100 chart
- Top Weeks: Number of weeks at chart position 1
- Song play count: number of tweets relating to a song
- Artist play count: number of tweets relating to an Artist, seen over his or her entire body of work (all songs)
- Artist Tweets: Number of tweets related to the Artist who sings this track
- Artist Sentiment Analysis-1: sentiment analysis result using VADER
- Artist Sentiment Analysis-2: sentiment analysis result using SentiWordNet

The authors used 10 fold cross validation when performing their regression Analysis and obtained the best results with **Support Vector Regression** which gave them a mean-absolute error of 4.015.

3 Required data

We need a data set comprised of weekly/monthly *Spotify* Viral 50 and Top 200 track chart data over several years.

Datasets:

- Billboard "The Hot 100" Songs [2]
- Spotify Charts [3]

3.1 Retrieving the data

We will use a dataset of all the "Top 200" and "Viral 50" charts published globally by *Spotify*. Spotify publishes a new chart every day.

We found a dataset on Kaggle that is its entire collection since January 1, 2017 until December 31, 2021.
If we would not have the provided dataset we would have to use the *Spotify API* to retrieve the data ourselves.
The dataset provides links to Spotify where we could gather metadata of the individual songs for further investigations about what kind of songs are more likely to trend and therefore land in the Top 200. Furthermore, we are planning to use the Billboard Hot 100 Charts dataset, retrieved from Kaggle, for verification of our results.

4 Methods

First we want to analyse the data in order to form a model to predict the influence of viral songs on the charts. For this we can use different libraries to plot our data and analyse it, e.g. Seaborn, Matplotlib or Plotly.

After the analysis stage we try to train a model to predict the influence of the viral songs in the top 200. We are not sure which machine learning model will suit this prediction the best, but we have a few options that could fit our needs. For example, Support Vector Machines or Linear Regression could be viable options. Still, we have yet to decide which model to choose in a later stage of the project.

References

[1] Carlos Soares Araujo, Marco Cristo, and Rafael Giusti. Predicting music popularity on streaming platforms. In *Anais do XVII Simpósio Brasileiro de Computação Musical*, pages 141–148. SBC, 2019.

[2] Dhruvil Dave. Billboard "the hot 100" songs. `https://www.kaggle.com/ds/1211465`, 2021.

[3] Dhruvil Dave. Spotify charts. `https://www.kaggle.com/ds/1265407`, 2021.

[4] Eleana Tsiara and Christos Tjortjis. Using twitter to predict chart position for songs. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 62–72. Springer, 2020.