Popularity Measurement Verification and Analysis of Features of the Top 100 songs

Project Report

Foundations of Computational Social Systems

WS 2022

Valentin Forster, 11823171

Clemens Farnleitner, 11803491

Deborah Huber, 01636255

Hanna Kröll, 51908594

www.github.com/ValentinForster/Foundations-Project

Introduction

With almost 500 million active users, Spotify is the most popular and widely used streaming platform for music, podcasts, and audiobooks (Statista, 2023). Therefore, it might be of interest for artists how they can grow their fanbase and make songs that get popular. One way to achieve this would be to evaluate and understand the popularity measurement Spotify generates – there are separate ones for songs, artists, and albums. The platform provides the popularity factor with a value between 0 and 100, with 100 being the most popular. Decisive for the value is a combination of two factors: the total number a track has been played and the recency in which these streams occurred. Based on the song popularity, an artist and an album popularity measurement are calculated (Spotify for Developers). Understanding the mechanism behind the platform could help artists maximize their listener-range.

The interest of this report therefore lays on the validation of the popularity measurement, as output by Spotify, or more precisely on the relationship between this measurement and the number of the followers an artist has on the platform. For the second part and further analysis, the relationship between a song's features and its popularity will be analyzed. Furthermore, the 100 most popular songs of 2022 will be compared with the dataset as whole. The features given by Spotify are following: acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, and valence.

If there are major correlations between the various features and the popularity of a song, there could be a "recipe" for a Top-100 Song – making it interesting again for artists and songwriters. This project is based on the following two research questions:

RQ1: Is there a correlation between the Spotify-popularity measurement of an artist with the number of their followers?

Thinking logically, the more popular an artist is, the higher their following should be. Keeping in mind on how Spotify generates the popularity measurement, we expect a positive correlation between the two factors.

Our hypothesis for the first part of the study, which concerns the validation of the popularity measurement, is following:

> H_1: The popularity measurement of an artist correlates positively with the number of their followers.

RQ2: Is there a correlation between the different features of songs and their popularity?

The second research question will be tested exploratively. We expect correlations and anticipate that certain features will be more important than others for the popularity of a song.

Data Retrieval

To verify our research questions and hypothesis we used the Spotify Web API to retrieve our data. To access the API, we used the Python Package Spotipy, which makes the processing of the requests much easier. Spotipy is a lightweight Python library for the Spotify Web API. With Spotipy it is possible to get full access to all music data provided by the Spotify platform. Our source code is available at our GitHub repository. A link to this repository is provided at the cover page of this project report. Our data consists of three datasets and each consisting of 1000 rows, respectively: artist_spotify_data.csv, track_spotify_data.csv, feature_spotify_data.csv. Due to limitations of the API itself we were only able to filter for tracks out of the year 2022 without any additional filters, like top tracks or the top 100 of 2022. Even after studying the documentation of the API, we were unable to understand the reasoning of why these songs were output. Therefore, the dataset consists of 1000 tracks which are presumably random sampled by Spotify itself.

Data Processing

Since we requested only data useful to us from the API, very little data cleaning had to be done. We joined the three tables into one and removed songs of which data was not present in all tables. The retrieved features and songs are visible in our dataset feature_spotify_data.csv. The following are brief explanations of each feature according to Spotify's documentation (Spotify for Developers):

<u>Acousticness</u>: A measure from 0.0 to 1.0 that detects acoustic sounds in a track.

<u>Liveness:</u> A measure from 0.0 to 1.0 that detects the presence of an audience in the recording.

<u>Speechiness</u>: A measure from 0.0 to 1.0 that detects the presence of spoken words in a track. The more speech sounds are heard in the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the value will be.

<u>Instrumentalness</u>: A measure from 0.0 to 1 that reflects the extent to which a track does not contain vocalizations. The closer the instrumentalness value is to 1.0, the greater the likelihood that the track does not contain vocal content.

<u>Energy:</u> A measure from 0.0 to 1.0 that reflects the intensity of a track. Energetic tracks are usually fast, loud, and noisy.

<u>Loudness:</u> A measure from -60 and 0 that represents the overall loudness of a track. This value is measured in decibels (dB). Loudness values are averaged across the entire track.

Danceability: A measure from 0.0 to 1 that indicates how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, and beat strength. A value closer to 0.0 indicates that a track is less danceable, and value closer to 1.0 is indicates that a track is more danceable.

Valence: A measure from 0.0 to 1.0 that reflects the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Key: A measure from 0 to 11 that indicates the key of a track.

Mode: A measure that indicates the key in the music of the track (1 is major, and 0 is for minor

Results

To determine if there is a relationship between the number of followers of an artist and their popularity score, the Pearson correlation coefficient was computed. The resulting correlation coefficient of ≈ 0.591 indicates that Spotify's popularity score does indeed correlate with the number of followers an artist has on Spotify. To visualize this relationship, we plotted the data into a scatterplot. Looking at the plot, there is a clear association between the two values. However, below ≈ 10.000 followers, the number of followers seems to have only a small impact on the popularity score. This can be due to the specific dataset, since our dataset is limited to artists with a popularity above 47. It's possible that the correlation looks similar for lesser-known artists and our dataset contains only those outliers, where artists have unusually low number of followers compared to their popularity.

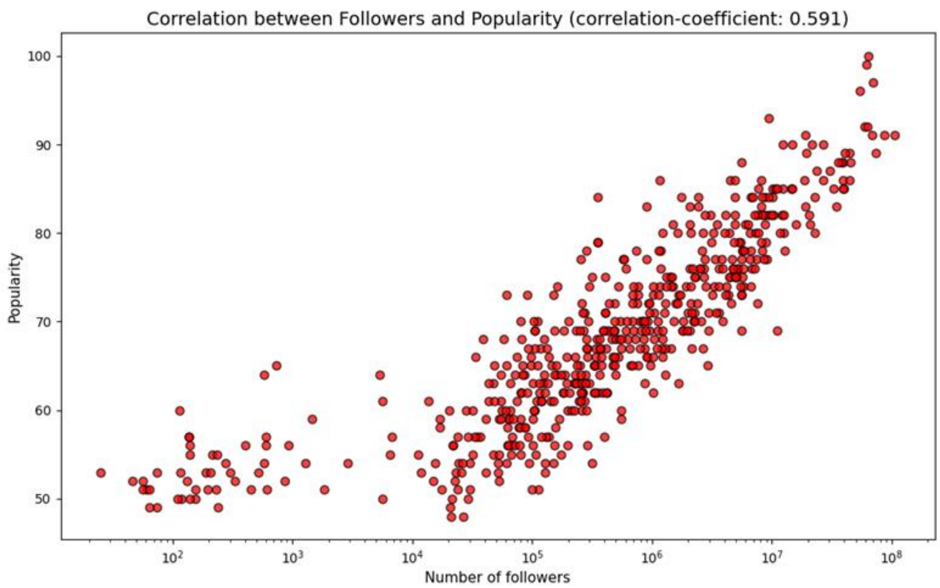

Figure 1: Scatter-Plot

For the second research questions we again used the Pearson correlation coefficient. The correlation coefficients between each feature of a song and its popularity as well as the corresponding p-values are displayed in Table 1. Overall, there is no feature which correlates highly with popularity. At a significance level of 0.05, the only significant correlations were with loudness, danceability, speechiness and mode. Speechiness and mode were negatively correlated with popularity. Loudness has the highest correlation, with a coefficient of ≈ 0.127.

	track_popularity	p-value
loudness	0.127374	0.000055
danceability	0.110342	0.000486
speechiness	-0.105866	0.000819
mode	-0.064995	0.040286
energy	0.038129	0.229269
liveness	-0.036942	0.244098
valence	0.031728	0.317164
instrumentalness	-0.031615	0.318885
key	0.030016	0.343995
acousticness	0.017162	0.588512

Table 1: Correlation Coefficients of features

To further investigate the role features play in the popularity of a song, we trained both a random forest regressor and a neural network on our dataset. Both were unable to accurately predict the popularity of a song based on its features alone. For the random forest regressor, the r-squared value amounted to ≈ 0.058. For the neural network, it even was negative. Looking at the SHAP values in Figure 2, the most important features for the prediction of the models again were danceability, speechiness and loudness.

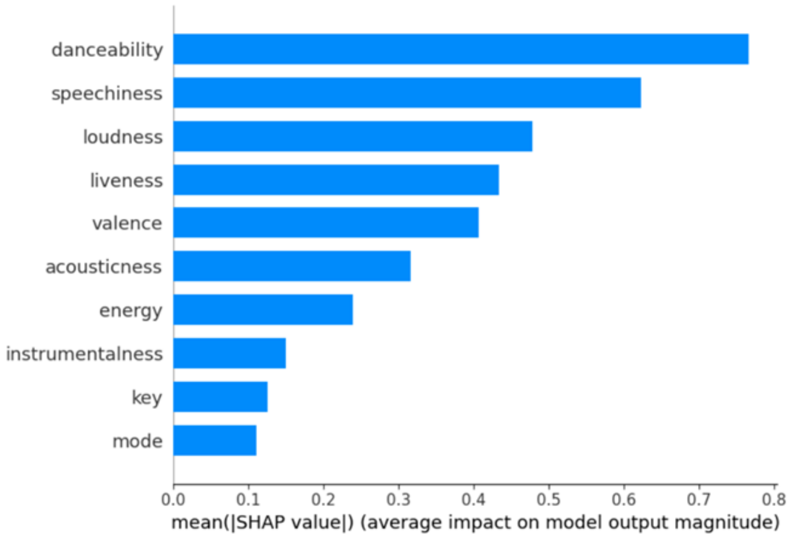

Figure 2: SHAP values for each feature

Lastly, we compared the 100 most popular songs in our dataset to the dataset as whole. For numerical attributes like tempo and features, we compared the median values. For the key and (musical) mode, we compared the mode values. All in all, the differences were negligible. The biggest difference was loudness, with the top 100 songs having a higher median loudness than the whole dataset.

```
All data: Median tempo: 123, Median duration min: 3.08, Mode key: 1, Mode mode: 1
Top 100: Median tempo: 125, Median duration min: 3.22, Mode key: 1, Mode mode: 1
------------------------
```

All data:

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	track_popularity
1	0.649	0.6345	5.0	-6.444	1.0	0.05885	0.1685	0.0	0.123	0.4785	71.0

Top 100:

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	track_popularity
1	0.6745	0.6925	5.0	-5.8035	1.0	0.0533	0.152	0.000002	0.12	0.4625	83.0

Figure 3: Comparison between the top 100 songs and the whole dataset

Conclusion

RQ1:

There is a strong correlation between the popularity score of an artist and their followers on Spotify. Although Spotify doesn't explain the details of how the popularity score is computed, our analysis suggests that the number of followers may be one component of its calculation.

RQ2:

Comparing the 100 most popular songs in our dataset to the dataset as whole revealed very little differences. This is in accordance with our analysis of the relationship between each feature and a songs popularity. The most notable features with regard to popularity are loudness, danceability and speechiness. It is important to note that these features don't necessarily cause a song to be popular. Rather, it could also be the other way around. As an example, popular songs may be louder on average, because they are made by successful artists, who are able to employ more skilled audio-engineers.

Limitations

There are several limitations that should be taken into consideration when interpreting the results of this project:

- Data limitations: The data used in this project was obtained through the Spotify Web API and is limited to 1000 songs released in the year 2022. This may not be representative of all tracks on the platform and could limit the generalizability of the findings.
- Correlation does not imply causation: While we may find correlations between Spotify popularity measurements and certain features of a song, it is important to remember that correlation does not imply causation. Therefore, it would be inappropriate to assume that changing certain features of a song would necessarily lead to increased popularity on the platform.
- Bias in data collection: The data used in this project is limited to what is available through the Spotify Web API, which may introduce bias in the sample. For example, the sample may be biased towards certain genres or artists that are more popular on the platform.
- Limitations of the Spotify popularity measurement: The popularity measurement generated by Spotify is based in large part on a combination of total number of streams and recent streams, which may not fully capture the popularity of a track or artist. Additionally, the popularity measurement may be influenced by factors outside of the control of the artist, such as placement on playlists or exposure on the platform's homepage.
- Interpretation of features: The evaluation of the features provided by Spotify may be specific to the algorithm used, with different song analysis methods potentially leading to different values. Additionally, there may be other features that are not provided by Spotify that could be important for predicting popularity on the platform.

Therefore, it is important to interpret the results of this project with caution and to consider these limitations when drawing conclusions.

References

Statista (2023). Number of Spotify monthly active users (MAUs) worldwide from 1st quarter 2015 to 4th quarter 2022. https://www.statista.com/statistics/367739/spotify-global-mau/

Spotify for Developers. Web API Reference – Get Several Tracks. https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-tracks