Predicting Spotify's Popularity Index

Columbia University
Martina Paez Berru, Yuwei Ding, Lydia Ying, Andrea Cuadros Vera, Yongwen Yuan
mp4395, yd2782, lcy2107, abc2225, yy3512

1 Introduction

1.1 Why Statement

Every year, Spotify releases new data on tracks' performances via the Popularity Index. This compelling feature powers everything from music discovery engines and artist recommendations to curated playlists and interactive features like the shuffle function.

This metric profoundly shapes user experience on the platform. Motivated by the mystery surrounding the drivers of an algorithm that underscores our everyday listening habits, we committed to the task of decoding the underlying mechanics of the index. Our research seeks to unravel the key factors that that propel a track to popularity, shedding light on the intricacies that resonate with listeners and dominate streaming charts.

1.2 General Overview

This research study was conducted in a two-part investigation. In Part 1, we employed deductive reasoning with statistical models including regressions, random forests, and neural networks to predict factors influencing song popularity. Once we understood the dynamics of popularity our research took an inductive reasoning approach leading to Part 2, where we analyzed song lyrics similarities across years 2018-2023 and user sentiments to identify recurring themes potentially influencing popularity.

2 Part 1 : Regressions, Random Forest, and Neural Network

2.1 Data Collection

We focused our research on the top streamed songs on the Spotify platform from the years 2018-2023. After trying to source data directly from Kaggle, we found that all of the relevant datasets lacked the number of song streams, which is a key component we wanted to evaluate. Spotify-specific restrictions presented an additional challenge – for example, the Spotify API does not have endpoints for obtaining song streams, and it is impossible to scrape the Spotify website. In the end, we utilized the third-party website Kworb.net, which has data on the song title, artist, total number of streams, and number of daily streams for the 1000 top-streamed songs on Spotify for each year. Using the web scraping library Beautiful Soup, we wrote a Python script to scrape the data from the relevant Kworb.net pages and store it in a Pandas dataframe with columns for song title, artist, number of streams, and year. We then created an algorithm that iterated through each song in the dataframe, used the Spotify API to obtain data on Spotify-specific audio features (i.e. tempo, valence, danceability) and its Spotify popularity score. We initially used the Python Spotipy library, but later found that it was much faster to directly call the API ourselves. We worked around the Spotify API's strict rate limits by batching our requests and including built-in functions that caught 429 server errors and would delay our program for the amount of time specified in the "Retry-After" header.

For our exploration of audio embeddings using YAMNet, we also used the Spotify API to obtain 30-second audio previews for each song. For our later regression analysis, to obtain song genres for songs across the years, since Spotify only stores genres for artists, we used the Last.fm API to obtain all genre tags for individual songs and one-hot encoded occurrences of the top 36 genres for the dataset ; to get the number of followers for artists from 2022-2023, we used the Spotify API.

2.2 Modeling and Analysis

2.2.1 Regression : Spotify Audio Features, Streams, Genre.

Initial Regression. We initially used a multiple linear regression using only Spotify-specific audio features directly obtained from the Spotify API (i.e. key, valence, and danceability) to first isolate the impact of audio features on Spotify popularity score. The resulting model had an extremely low R-squared of around 0.0022, and little correlation between actual and predicted popularity (Figure 2.2.1).

To explore potential improvements, we tested more sophisticated models including random forest and Lasso regression, and performed a grid search to optimize key hyperparameters (e.g., number of trees, maximum depth, and minimum samples per split) for the random forest model.

Although the random forest model slightly improved predictive accuracy compared to linear regression, the overall

performance remained suboptimal (Figure 2.2.2, 2.2.3). The R^2 scores were still low, indicating that the Spotify-specific audio features alone were insufficient for accurately predicting popularity metrics. Feature importance analysis revealed that attributes such as loudness, duration ms, and energy contributed more to the model than others, but their predictive power was limited. Similarly, the Lasso regression model did not yield substantial improvements (Figure 2.2.4). Despite tuning the regularization parameter α, its performance was comparable to that of the random forest model.

Impact of Streams. To test our hypothesis that streams would have an outsize impact on popularity score, we then regressed both Spotify audio features and number of streams on Spotify popularity score. The resulting multiple linear regression model had a significantly predictive accuracy, with a higher R2 of around 0.29, and Figure 2 shows a higher linear correlation between actual and predicted popularity compared to the audio features-only regression. We obtained even better results with random forest regression on the same variables (Figure 2.2.5), and based on this model's performance and feature importances (Figure 2.2.6-2.2.7), it is clear that streaming metrics have an outsize influence on a song's popularity score, which is both expected and could help explain why Spotify is so secretive about releasing streaming data to the general public.

Impact of Genre. We speculated whether genre could relate to popularity score, given that some genres may be more popular overall than others. We converted obtained genre to dummy variables and included them in our earlier linear and random forest regressions. As shown in Figures 2.2.8-2.2.9, the performances of our models improved only marginally in terms of R-squared and mean squared error, if at all, indicating that genre does not have a significant impact. This may be because there are many prevalent genres in today's popular music, probably an effect of the advent of streaming services in the 21st century promoting music discovery and diversity in music taste in today's listeners.

2.1.2 Neural Network Analysis.

Introduction to Yamnet. The initial experiments using Spotify API-provided features did not yield satisfactory prediction accuracy. To address this, we turned to direct audio feature extraction. By employing Yamnet—a pre-trained neural network model designed for audio event classification—we derived 521-dimensional embeddings directly from the audio waveform previews. These embeddings provide a richer representation of a track's acoustic profile than the standard Spotify features.

Audio Analysis from MP3 Files. We obtained 30-second preview clips of tracks via the Spotify Web API and downloaded the corresponding MP3 files for analysis. Due to limitations in preview availability, we focused on a subset of tracks, particularly those from 2022 and 2023. Each audio clip was resampled to 16 kHz and processed with Yamnet, generating a set of scores across 521 audio event classes. These frame-level outputs were averaged to produce a single 521-dimensional vector per track. This embedding was then merged with each track's popularity data to create a new dataset for modeling.

To refine the input features for subsequent modeling, we employed a random forest model to assess feature importance. This step aimed to identify the most relevant acoustic dimensions and reduce the risk of overfitting that could result from using all 521 features.

Neural Network Modeling. With the Yamnet feature embeddings in hand, we next employed a neural network model to predict popularity. The model incorporated both the Yamnet acoustic features and additional temporal features derived from daily streaming dynamics. These time-based features aimed to partially account for temporal trends or shifts in popularity over the observation window. To predict song popularity, we developed a neural network model using TensorFlow and Keras, designed to capture complex nonlinear relationships between input features and popularity scores. The model combined 521-dimensional audio embeddings extracted from Yamnet with temporal features, such as daily growth rate and rolling averages of streaming data. After standardizing the inputs using StandardScaler, the network architecture was built with three fully connected hidden layers of 128, 64, and 64 units, all using ReLU as the activation function. Dropout layers with a rate of 0.3 were added after each hidden layer to reduce overfitting, and the output layer consisted of a single unit with a linear activation function to predict popularity scores. The model was trained with the Adam optimizer and Mean Squared Error as the loss function, using a batch size of 32 over 150 epochs, with 20

Neural Network Result. We achieved a notable improvement in key performance metrics with our neural network model. The out-of-sample R^2 score reached approximately 0.5, which is already a promising result for music popularity prediction. (Figure 2.2.10)

The feature importance analysis revealed that features reflecting strong rhythmic or stylistic characteristics are more influential in driving a track's popularity. Notable examples include : Drum roll, Jingle (music), Roll. (Figure 2.2.11)

3 Part 2 : Dynamics in Lyrics and User and Song Sentiment Analysis

3.1 Data Collection

This study examines lyrical themes and sentiment in the top 25 popular English* songs from 2018 to 2023. Using datasets from Part 1, we identified song and artist names for

the specified years. Lyrics were retrieved via the Lyrics.ovh and Genius APIs, creating a comprehensive dataset.

In addition to the top 25 songs, we analyzed two standout artists in terms of popularity and influence : Taylor Swift and Billie Eilish. For Taylor Swift, we gathered songs from 2019 to 2023, while for Billie Eilish, we included data from 2018 to 2023.

Lastly, we used the YouTube API to collect the top 40–50 comments from music videos on each artist's official YouTube channel, adding user sentiment to our analysis.

3.2 General Overview of Song Sentiment and Similarity (2018 - 2023)

Latent Semantic Indexing (LSI). Latent Semantic Indexing (LSI) was used to uncover dominant lyrical themes and track their evolution across years. Each year's lyrics were processed to extract topics, and cosine similarity was calculated between topics of different years to identify thematic overlaps. The code for this analysis begins with robust text preprocessing, which includes tokenization and removing common stopwords, along with a customized set of stopwords tailored to the domain of song lyrics. This customization eliminates filler words such as "oh" and "yeah," allowing the model to focus on substantive thematic words.

LSI is implemented by constructing a term-document matrix and applying Singular Value Decomposition (SVD) to reduce dimensionality. This reduction uncovers latent themes within the dataset while simplifying the data for meaningful topic analysis. The cosine similarity metric quantifies thematic consistency between years, where values closer to 1 indicate strong similarity and lower values signify thematic divergence. The heatmap in Figure 3.1.1 visualizes these similarities, with diagonal values representing self-similarity (lyrical consistency within the same year) and off-diagonal values reflecting thematic shifts across years.

Insights obtained from the graph :

When it came to the recurrent and relevant themes extracted from the topics, we got the LSI model to identify the dominant topics for each year. The word clouds for the dominant topics in Figures 3.1.4-3.1.9 revealed recurring themes such as love, heartbreak, and introspection, which evolved over the years but remained central to popular song lyrics. The progression of dominant words highlights a shift in lyrical focus over time :

- Earlier years (2018) leaned heavily on themes of romantic idealism and reflection.
- Mid-years (2019–2021) transitioned toward introspection, emotional vulnerability, and themes of perseverance.
- Later years (2022–2023) incorporated more symbolic, physical, and even sensual imagery, reflecting broader and more complex storytelling.

Sentiment Analysis. We employed two types of sentiment analysis for the annual top 25 song lyrics : naive positive/negative polarity score and NRC emotion score. Similar to the LSI similarity analysis, this part begins by retrieving song lyrics from the Genius APIs and appending the result to the songs_with_lyrics_complete Dataframe. A clean_lyrics_genius function then removes irrelevant 'contributor' and 'translator' sections from the lyrics while cleaning, tokenizing, and stemming lyrics into words that are case insensitive. For the naive pos/neg analysis, we started by defining a get_pos_neg_words function to retrieve two separate lists containing positive and negative words from https ://ptrckprry.com/. This is then used as classification criteria as we feed cleaned lyrics into the simple sentiment analysis function, returning a positivity and negativity ratio for each song.

As naive sentiment analysis only provides preliminary emotion sorting, we also included NRC score analysis that further divides positive and negative emotions into groups such as 'sadness,' 'disgust,' 'anticipation,' etc... We repeated the process by getting classification data from NRC-emotion-lexicon, and then looping through each word-based lyrics to retrieve the specific emotion proportion. We visualized the average pos/neg ratio by year in a time-series line chart and the mean NRC score by a stacked bar chart. Key insights from our findings include :

- **Shifted Dynamics of positive and negative ratio in 2018 and 2023**, indicating a gradual increase in lyrics' positivity and decline in negativity (Figure 3.1.10).
- **Average positive ratio reached a peak in 2021,** during the mid-pandemic era. This may be due to the potential inverse relation between social and macroeconomic upheaval and song optimism.
- **Consistency in Naive emotion and NRC score results** as the latter also showcases the highest percentage of 'anticipation,' 'trust,' and 'joy' in 2021, where the general positive ratio of lyrics is the highest (Figure 3.1.11).

3.3 Case Studies : Taylor Swift and Billie Eilish

While applying the same lyrics similarity and sentiment analysis in our focused studies on the two artists, we also extracted listener comments from YouTube API as a direct comparison to the popularity index of each artist's song.

Lyrics Sentiment, Listener Sentiment, and Popularity. We filtered out all Taylor Swift and Billie Eilish songs from our database and stored cleaned lyrics into a new list. We then fetched individual music URLs for each song by sending requests to Youtube API and extracted the unique vi-

deo ID of each song. The getcommentsfromids function takes the video ID and returns the top forty comments for each song. We implemented advanced text processing on those comments by manually compiling a list of emoji and symbol patterns and excluding those from the comments to prevent potential errors.

We employed Textblob text analysis for the comments to retrieve a generic emotion polarity score. We accompanied this result with a VADER sentiment analysis that adds neutral and compound emotion into the category. While comments of both artists are predominantly neutral by the analysis, we can see that the Taylor Swift official channel browsers are more inclined to post positive comments compared to Billie Eilish channel users, whose comments are more negative in 2022 and 2020 (Figure 3.1.12, Figure 3.1.13).

To draw patterns between each artist's popularity index and listener sentiment, we visualized the result using a combined graph of line and bar chart. We noticed that there is tight consistency between Billie Elish's song popularity and her music video reaction, namely a decline from 2018 to 2019 and a continuous increase the year forward. An outlier is observed in 2021, however, for Taylor Swift's case as the song exhibits the lowest popularity index while receiving the highest positive listener comment score (Figure 3.1.14, Figure 3.1.15).

This discrepancy may be due to the fact that Spotify and YouTube are two different platforms and have various targeted users. We also have a limited number of comments data and each artist's songs, resulting in the underrepresentation of true comparison results.

In terms of lyrics sentiment, we adopted the same NRC emotion score approach for the two artists. In general, Taylor Swift has a more even allocation of positive and negative emotion categories in 2019, 2020, and 2022, and more optimism in the 2021 song (Figure 3.1.16). Billie Eilish's songs are much more negative from 2018 to 2020, with an exception in 2021 and 2023. This finding is coherent with the optimism peak for the top 25 popular songs sentient in 2021 (Figure 3.1.17).

Themes and Creative Variability. The case studies provide insights into the artists' thematic evolution using heatmaps. For Taylor Swift, the heatmap in Figure 3.1.2 reveals moderate year-to-year similarities with values such as 0.65, 0.80, and 0.72, indicating thematic continuity but with notable variation. Lower correlations, like 0.30 and 0.54, suggest shifts in lyrical focus in certain years. While her themes remain rooted in love, reflection, and personal growth, the variability implies periods of creative exploration within her otherwise cohesive narrative style.

In contrast, Billie Eilish shows more dramatic fluctuations across the years. The heatmap in Figure 3.1.3, showcases how the early years exhibit low year-to-year similarities, such as 0.29 and 0.01, highlighting significant thematic shifts and experimentation. However, certain consecutive years, such as 0.92 and 0.97, demonstrate high cohesion, particularly during phases of focused storytelling. The overall pattern reflects a mix of exploration and occasional thematic stability, capturing Billie Eilish's evolving artistic approach.

To test these findings, we applied Latent Semantic Indexing (LSI) to extract dominant lyrical themes. Text preprocessing included tokenization, stopword removal, and custom stopwords tailored for song lyrics. Cosine similarity was calculated to measure thematic overlap, and the results were visualized through heatmaps. The heatmaps showcase Taylor Swift's moderate consistency with creative variability and Billie Eilish's contrasting mix of thematic exploration and occasional stability. Additionally, we cross-referenced word clouds and top keywords for the extracted dominant topics. Surprisingly, they had similar topics despite their different "vibes."

4 Conclusion

Our two-part study on Spotify's Popularity Index revealed key insights : while genre and specific Spotify features have minimal impact on song popularity, streaming volume is a significant predictor. This suggests that direct user engagement metrics may be more critical in determining a song's success than its intrinsic musical properties.

Our analysis of lyrical trends from 2018 to 2023 showed little change in themes, affirming that universal emotions like love, struggle, joy, and longing continue to resonate deeply. This persistence of fundamental emotions in trending songs highlights how despite technological advancements, changes in production techniques and cultural shifts, the core ways in which listeners value musical expression remain remarkably consistent.

This project not only illuminates the mechanics of musical popularity but also lays the groundwork for future explorations into the dynamics of musical trends and their implications on cultural narratives. Moving forward, this research will be published on GitHub, allowing music enthusiasts and researchers alike to access our data and methodologies to refine the now more transparent Spotify Popularity Index. Further exploration could enhance our understanding of how evolving trends in production and cultural shifts impact, yet do not fundamentally alter, the way music resonates across generations.