User Behaviour on Spotify

BUSA400 - Operations and Analytics Independent Study

Supervised by Prof. Rim Hariss

Written By

Maëlys Boudier

McGill University
Canada

August 29, 2022

1 Introduction

Spotify, a leading music streaming service provider launched in 2008, now boasts over 400 million users [Spotify, 2022]. The platform distributes songs and podcasts for a multitude of artists and implements a revenue-sharing model to redistribute revenue to stakeholders. Revenue-sharing incentivises all the involved parties to perform well to maximize the pay off. However, such models can favor one party over another and in turn lead to unfair distribution of revenue.

This study will provide insights on patterns of consumption and user segmentation across Spotify and other relevant music streaming platforms. First, this paper will explore consumption habits including: How do artists compare to each other in terms of diversity of listeners? And how do users consumer music? Second, a focus will be made on differentiating freemium and premium user segments.

2 Literature Review

Forbes shed light on Spotify users critiquing artists' low pay despite the platform paying multiple billions of dollars in royalties [Dellatto, 2022]. Spotify is a major player in the music industry and leaves very negotiating power to artists. Exposing Spotify's unfair revenue-sharing models brings forth the opportunity to review literature to model Spotify's user and artist base to build a fairer model.

Wlomert and Papies echoed the "controversial debates in the music industry" but focused on the "cannibalization of distribution channels" and the impact on artists' revenues [Nils Wlömert, 2016]. The main issue they highlight is the risk artists face of becoming fully dependent on music streaming platforms with no traditional alternatives such as song downloads and CD purchases. However, the paper concludes that paid streaming results has a positive net impact on market revenue, while free streaming does not. Furthermore, they emphasize that artists must ensure that streaming rights are included as part of their contract (especially as other distribution channels disappear).

A 2013 paper, *Understanding User Behavior in Spotify* modeled session arrivals on Spotify as a non-homogenous Poisson process and found daily consumption patterns [Zhang et al., 2013]. Interestingly, the researchers gained access to the Hadoop Cluster used to "store and analyze Spotify's log data", but one must note that Spotify now uses the Google Cloud Platform [Zhang et al., 2013] [AccelData, 2022]. The session arrivals they modeled included a morning peak, evening peak, weekend effect, and commuting effect on weekdays which represented moments with high consumption levels. Not only do users exhibit daily consumption patterns, but the paper proved that "Spotify users have their favorite times of the day to access the service". Modelling the consumption habits of users allows researchers to create simulations and test their models.

The most recent literature relevant to this paper was published in 2020 by two researchers at the University of Palermo in Italy [Mariangela Sciandra, 2022]. They analyzed the songs audio features to present a new statistical model. They employed a Beta Generalized Linear Mixed Models (GLMM) to analyze Spotify's Data. The data sources they used originated from the Spotify Web API with clear steps detailed 'how-to' use the API on R software via Spotify's Developer Platform. The use of the web API begs the question of the whether the data is as complete as the data from the 2013 User Behavior Paper. The Beta GLMM was determined the most conclusive model to predict track popularity (characterized by the number of streams) using the track's features. Sciandra and Spera, the co-authors, provided a

table with the Spotify audio features, data types, and data description. The audio features in the table include: acousticness, danceability, duration_ms, energy, id, instrumentalness, key, liveness, mode, tempo, time_signautre, amongst other features and resemble some of the data described in the *3 Data Overview* section. This data source, although not documented for lack of a developer account at this stage, is worth exploring for further research.

3 Data Overview

In the hopes of increasing research in the music industry, a crowd-sourcing project in the early 2010's resulted in giant data sets detailing music platform user behavior and music catalogues (including song and artist metadata). However, to use data, one must know what they may find and how it can be opened, read, and analyzed. As such, data has been extracted from the following four resources and documented below:

- Echo Nest
- Last.Fm (MLHD)
- Spotify AI Crowd
- SQL Lite Million Song Challenge

3.1 Echo Nest

In an effort to give back the community and provide music metadata, Echo Nest released the Taste Profile Data Set found at: Echo Nest MillionSongDataset.

The two following TXT files were downloaded: *Taste Profile Usercat 120k* and *Train Triplets*. The *Taste Profile Usercat 120k* is a subset of users but a full data set is also provided on the website. To access information, one must use The Echo Nest API to query users' catalogs.

Taste Profile Usercat 120k includes two columns:

- Catalog Name (string): follows the format userID_tmp_catalog
- Catalog ID (string)

Train Triplets includes three columns:

- User ID (string)
- Song ID (string)
- Play Count (integer)

3.2 Last.Fm (MLHD)

The Music Listening Histories Data Last.FM MLHD provides a "large-scale collection of music listening events". The two following CSV files were downloaded: *MLHD Behavioural Features* and *MLHD Perc Per Unit of Time*.

MLHD Behavioural Features includes thirteen columns:

- uuid (string)
- exploratoryness of album, artist & track (float)

- mainstreamness of album, artist & track (float)
- fringeness of album, artist & track (float)
- genderedness of album, artist & track (float)

MLHD Perc Per Unit of Time includes the following columns:

- uuid (string)
- unit of time: days, weeks... (float between 0 and 1)

3.3 Spotify AI Crowd

The Spotify AI Crowd data cannot be directly linked to the other data sources. Notably, this data set comes with a full data description available upon download. The Spotify AI Crowd data set is particularly relevant when analyzing freemium and premium accounts.

The two following CSV files were downloaded: *Track Features* and *Training Set - Log Mini*. Given the complete documentation already available online, only a brief overview of each file is provided below.

Track Features provides many features associated with each track including:

- Track ID (string)
- duration (float)
- popularity (float)
- acoustic vectors (float)
- ...

Training Set - Log Mini provides many features associated with each session including:

- Session ID (string)
- Track ID Clean (string)
- skips (boolean)
- date (string in format YYYY-MM-DD)
- premium user (boolean)
- ...

3.4 SQL Lite Million Song Challenge

From the SQLite Million Song Challenge the following files were downloaded:

- 2 TXT files: *Unique Tracks* and *Unique Artists*
- 3 DB files: *Artist Term*, *Artist Similariy* and *Track Metadata*

Note: A DB file is a generic database file, the Million Song Data Set provides SQLite files under .db extensions. Visual Studio Code provides a SQLite Viewer to view the files and note the table and column names. Consequently, the files could be opened through the SQLite3 Python Package and converted into a Pandas Data-frame.

Unique Artists includes:

- Artist ID (string)
- Artist MBID (string)
- Track ID (string)
- Artist Name (string)

Unique Tracks includes:

- Track ID (string)
- Song ID (string)
- Artist Name (string)
- Song Title (string)

Artist Term includes:

Definitions:

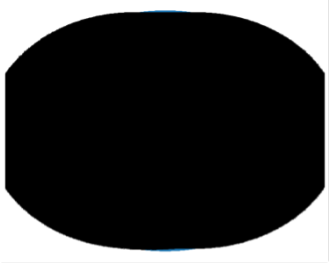
- Term: Echo Nest Tags
- Mbtags: Musicbrainz tags

artist_term.db		Column header		
		artist_id	term	mbtag
Tables (5)	artists	x		
	terms		x	
	artist_term	x	x	
	mbtags			x
	artist_mbtag	x		x

Artist Similarity includes:

artist_similarity.db		Column header		
		artist_id	target	similar
Tables (2)	artists	x		
	similarity		x	x

In the hopes of visualizing the data frame, I created a network using the NetworkX Python Package and used Matplotlib Python Package to plot it. However, the visualization resembled a 'black hole' probably due to the large data set size (figure below).



Track Metadata includes the following columns:

Track_id	title	Song_id	release	Artist_id	Artist_mbid	Artist_name
duration	Artist_familiarity	Artist_hotttnesss	year	Track_7digitalid	Shs_perf	Shs_work

4 Next Steps

Due to the limited time frame, this study provides an overview of the data and available literature on Spotify users' consumption habits. To push the analysis, next steps include creating a network linking artists based on similarity and clustering groups of artists. One must also use the listeners' catalogues to analyze the similarity between the artists they listen to. This can be done by computing a percentage of a complete graph. A complete graph would have all of the artists linked to each other, however, one must weigh the edges as an artist may appear multiple times or songs might be played more than once. This step would allow one to determine how diversified listeners are.

5 Conclusion

Finally, to build the literature review and data documentation I encountered a few challenges. The first obstacle I had to overcome during the study was finding and organizing the relevant data files given the large amount of available data. This also led to some challenges regarding my computer's computing power when dealing with some of the heavier files. Secondly, opening some of the files was very time consuming: to open the db files, one must know the file type (in this case: SQLite) and also know the SQL schema used.

Published research already provides some models to characterize user consumption such as daily patterns, as well as "how-to" access the Spotify Developer Platform web API. However, this paper's literature review and data documentation may serve as a steppingstone providing contextual analysis to build a newer and fairer revenue-sharing model as there is currently a gap in the music industry.

References

[AccelData, 2022] AccelData (2022). Data engineering: How spotify upgraded its data orchestration platform. *AccelData Inc.* https://www.acceldata.io/blog/data-engineering-best-practices-how-spotify.

[Dellatto, 2022] Dellatto, M. (2022). Spotify says it paid $7 billion in royalties in 2021 amid claims of low pay from artists. *Forbes*. www.tinyurl.com/2h2uxztm.

[Mariangela Sciandra, 2022] Mariangela Sciandra, I. C. S. (2022). A model-based approach to spotify data analysis: a beta glmm. *Journal of Applied Statistics*, 49(1):214–229. https://doi.org/10.1080/02664763.2020.1803810.

[Nils Wlömert, 2016] Nils Wlömert, D. P. (2016). On-demand streaming services and music industry revenues — insights from Spotify's market entry. *International Journal of Research in Marketing,*, 33(2):314–327.

[Spotify, 2022] Spotify (2022). About spotify. *Newsroom*. https://newsroom.spotify.com/company-info/.

[Zhang et al., 2013] Zhang, B., Kreitz, G., Isaksson, M., Ubillos, J., Urdaneta, G., Pouwelse, J. A., and Epema, D. (2013). Understanding user behavior in spotify. In *2013 Proceedings IEEE INFOCOM*, pages 220–224.

Note: all the written code and data sources are accessible along with this report