Datasheet for Audio Features of Top 100 Greatest Billboard Artists Dataset*

Luca Carnegie

2024-04-24

I present a datasheet detailing the dataset used for my analysis in 'The Anatomy of a Hit'

This datasheet is constructed in the image of an ideal datasheet written about by Gebru et. al in their paper "Datasheets for Datasets" by Gebru et al. (2021). I used the questions extracted from thier paper to motivate the discussion in my datasheet.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to enable analysis of audio features on Spotify and their effect on a song's popularity score on Spotify.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Luca Carnegie, the author of the paper, created this dataset to write the paper "The Anatomy of a Hit".
    - The audio and popularity data itself was collected by Spotify, put into it's API and then collected by Luca to analyze for this paper.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - There was no funding involved with the creation of the dataset used for the paper.
    - Spotify was involved in the collection, processing and hosting of the data in the Spotify API, based on data from their own operations.

*Code and data are available at: https://github.com/lcarnegie/popularity-modeling

4. *Any other comments?*

    - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance represents a song. Each song is by an artist who was ranked in the Greatest Billboard Top 100 Artists, which is a list that compiled the top-performing artists in the Billboard Top 100, since it's inception in 1958.

2. *How many instances are there in total (of each type, if appropriate)?*

    - There are 773 songs in the dataset.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - No. This dataset contains a sample of songs from the larger dataset of all artists/songs in the Spotify API.
    - This dataset was not intended to be representative of the larger dataset, as the point of it's creation was to analyze only the features of the best-performing Billboard artists, that are also on Spotify, with the goal of finding patterns within the audio features that were correlated with a high popularity score.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance consists of a
        - Artist Name
        - Song Name
        - Popularity score (measured from 0-100)
        - Valence (musical positivity, measured from 0-1)
        - Danceability (measured from 0-1)
        - Mode (major or minor)
        - Explicit (yes or no)
        - Loudness (measured in decibels, dB)

  - Duration (measured in seconds)

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Each song is labelled by the artist name and song name.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - No. This was checked and tested using the R programming environment.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Yes. Songs by the same artists are explicitly related by possessing the same artist name within their respective instances. If two individual artists made independent covers of the same songs, both versions would be differentiated by the instance's artist name rather than the song itself.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Possibly. Some of Spotify's metrics, such as valence and danceability are known to be subjective in their collection methods. Conversations by other researchers with Spotify engineers revealed that danceability scores were collected by asking college interns at a music intelligence company to rate each of the songs. From this, other calculated metrics like the valence could also be beholden to errors stemming from the subjectivity of the person making the rating.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is self-contained, but as popularity of different artists changes quickly it is not expected to remain updated for very long, as Spotify updates their popularity data on a near-daily basis.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - You can identify artists by their song name, but

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*

    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- Artist Name: taken from music publisher
- Song Name: taken from music publisher
- Popularity score:
- Valence (musical positivity):
- Danceability: before being acquired by Spotify in 2014, a music intelligence startup used labour by college interns, who rated songs as danceable or not. However, it is unclear if human rating still continues or whether that has been replaced by machine listening techniques.
- Mode (major or minor): machine listening techniques tagged songs as being in a major or minor key.
- Explicit: machine listening techniques tagged songs as being explicit or not.
- Loudness: calculated from taking the average loudness throughout the entire song
- Duration: taken from length of the recording of the song

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected using a combination of manual human curation and machine listening techniques, although specific details on the process were not able to be found due to the proprietary nature of Spotify's technology.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The sampling strategy was mainly based on convenience sampling - I chose only songs whose artists that appeared on the Billboard Hot 100 Greatest of All Time list. -It is of note that using the data collection process, certain songs appear on artists' top ten list did not have audio features that were accessible through the API, so they had to be discounted. As a consequence, there is an uneven amount of songs represented per artist. Therefore, certain artists may be more overrepresented than others.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - College interns were involved in parts of the collection process for variables like danceability, according to Drouillard. However, it is unclear is they were paid or not for their time. It is not clear if any of the songs present in the dataset were subject to this collection technique. It could be that some were rated through machine listening techniques and some through manual human rating. It is also unclear how the machine listening techniques were tuned.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old*

news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

- The timeframe for which the data on the Spotify API was collected is unclear.

- For this analysis, however, the data was collected over 1 hour by querying the Spotify API for each artist, the audio features of their songs, and their top 10 tracks and storing them within a .csv file.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No, none were conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Not applicable, as the data concerns itself with songs and not data about individuals.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - Not applicable

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Not applicable

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Not applicable

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Not applicable.

12. *Any other comments?*

- No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - No.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes. It is available within this repository at popularity-modeling/data/raw_data.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Yes. The data was processed using the RStudio environment, which is a freely available and downloaded software package searchable online.
4. *Any other comments?*
    - TBD

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - It was used to write the paper "The Anatomy of a Hit", which analyzes the determinants of popularity scores of songs on Spotify, using the discographies of top pop music artists.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - It can be accessed at https://github.com/lcarnegie/popularity-modeling
3. *What (other) tasks could the dataset be used for?*
    - It could be used, among other things, to categorize songs based on their characteristics, (e.g. genre, by popularity, etc.)
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide*

a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

- Yes. Since popularity data is by nature quite volatile, this dataset is expected to fall out of date very quickly. This dataset should only be used as a guide to motivate more sophisticated data collection and analysis methods that update as new data is collected.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description*

   - Ideally, this dataset should not be used to cause evil or harm in the world. At the time of writing, no immediately dangerous or immoral acts are considered in this. It is my hope that those, if anyone, who uses it ultimately decides to do good with it in the hope of serving rather than hurting humanity.

6. *Any other comments?*

   - No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - At this stage, there is no plan to do such a thing.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset will be hosted and distributed through the GitHub repository at https://github.com/lcarnegie/popularity-modeling

3. *When will the dataset be distributed?*

   - It is currently available for perusal and analysis, as of April 24. 2024

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - No.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- No. However, through their API terms and conditions Spotify outlines that Spotify content (particularly song recordings, rather than audio data) may not be downloaded, that visual content be kept in its original form and that content attribution.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No.

7. *Any other comments?*

    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset will be hosted within a GitHub repository at: https://github.com/lcarnegie/popularity-modeling

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Luca Carnegie is contactable at luca.carnegie@mail.utoronto.ca

3. *Is there an erratum? If so, please provide a link or other access point.*

    - No.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - It will likely not be updated after the end of writing this paper.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - Not applicable.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - If the dataset is updated, older versions will continue to be supported.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Luca Carnegie is heartily open for collaborations to extend and augment this dataset and it's goals. He is contactable at luca.carnegie@mail.utoronto.ca

8. *Any other comments?*

    - No.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets," no. arXiv:1803.09010 (December). https://doi.org/10.48550/arXiv.1803.09010.