Final Project Data Memo

James Du, Ian Gascon, Annie Huang, and Meng Vong

January 19, 2022

An Overview of Our Dataset

The dataset includes song names and their associated popularity score (as taken from the Spotify Web API). The dataset also includes characteristics of each song such as: genre, artist name, track id, acousticness, danceability, duration, energy, and instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, and valence.

Where/How We'll Obtain It (Plus the Link And Source)

We will obtain the dataset from Kaggle found at Spotify Song Popularity Prediction. The documentation for the features in the dataset can be found from Spotify's Documentation at Spotify Web API Reference.

Number of Observations and Predictors?

There are 228159 unique observations and 8 predictors in this dataset. The predictors include quantified song features such as:

1. ***Genre:*** A list of genres associated with the track.
2. ***Artist Name:*** The artist who performed the track.
3. ***Acousticness:*** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
4. ***Danceability:*** Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
5. ***Duration_ms:*** The track length in milliseconds.
6. ***Energy:*** Measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
7. ***Instrumentalness:*** Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".
8. ***Key:*** The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
9. ***Liveness:*** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
10. ***Loudness:*** The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.
11. ***Mode:*** Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
12. ***Speechiness:*** Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0.

13. ***Tempo:*** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
14. ***Time Signature:*** The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
15. ***Valence:*** A measure from 0.0 to 1.0 describing 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).

Types of Variables We'll Work With

The values of the predictors are mainly floats and integers, while the datatype of our target is an integer representing song popularity.

Quantifying Missing Data and How We Will Handle It

There is no missing data in this dataset since it pulls directly from Spotify's API. However, some songs are listed multiple times under different genres, which we will have to deal with. In addition, some of the observations listed are not actual songs but spoken words such as comedy—however, these tracks may still score relatively high on our predictor variables despite not being traditional songs, so we will see how we deal with that later.

An Overview of Our Research Questions

Variable and Question of Interest

The variable we are mainly interested in predicting is the popularity of the song as determined by Spotify. We will use this to answer the question: ***What features relating to musical composition factor into the popularity of a song on Spotify?***

If time allows, we are also interested in looking at the features that are associated with the genre of the songs.

Outlining and Describing Our Repsonse Variables

Again, our response/outcome variable will be the popularity variable. According to Spotify's API documentation, "popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past."

Approach to Best Answer Our Question

Given the quantitative quality of the target, we will likely approach this question with regression.

Most Useful Predictors

When people talk about what music they like, they often talk about qualities like they like this song's beat or they like the artists, qualities like that. Thus, we believe that the following predictors are especially important: ***Artist Name, Tempo, Danceability, and Loudness.***

Proposed Timeline and Group Work

Division of Work

- **James Du**
  - Regression Model Building
  - Results Interpretation
  - Data Preprocessing
  - Bitcoin mining rig setup
- **Ian Gascon**
  - Data cleaning
  - Exploratory analysis
  - Build the html page
  - Start watchmaking
- **Annie Huang**
  - Data cleaning + wrangling; set up dataset
  - Exploratory analysis
  - Possible classification model for genres
  - Cat wrangling
- **Meng Vong**
  - Explore PCA and higher flexible models
  - Explore interactive visualizations
  - Write interpretations and prose of projects
  - Anxiety eating

Proposed Timeline

Be finished by the project due date.

Week #	Task to Complete
3	Submit data memo and divide work
4	Load and tidy data
5	Run and write up descriptive analysis
6	Build and run the model
7	Write up results
8	Work on project draft
9	Make edits to draft
10	Submit final project

Questions and Concerns

So far, we have no specific questions or concerns but we will make sure to reach out to you/the instructional team as soon as we do!