CS109a Final Project: Milestone 3

Katherine Deng, Jess Eng, Anna Li, Lizzy Yang

November 20, 2019

Revised Project Statement

After some consideration of how we could best evaluate the performance of our models, we decided to refocus our project statement on identifying the audio features that characterize the most popular songs on Spotify for a given year. By querying tracks from Spotify and training our model on a subset of popular and (relatively) unpopular songs from 2018, our project seeks to classify and predict whether new songs (for instance, from Spotify's weekly "New Music Friday" playlists) will or will not become popular, and even rank them based on their expected popularity.

Exploratory Data Analysis

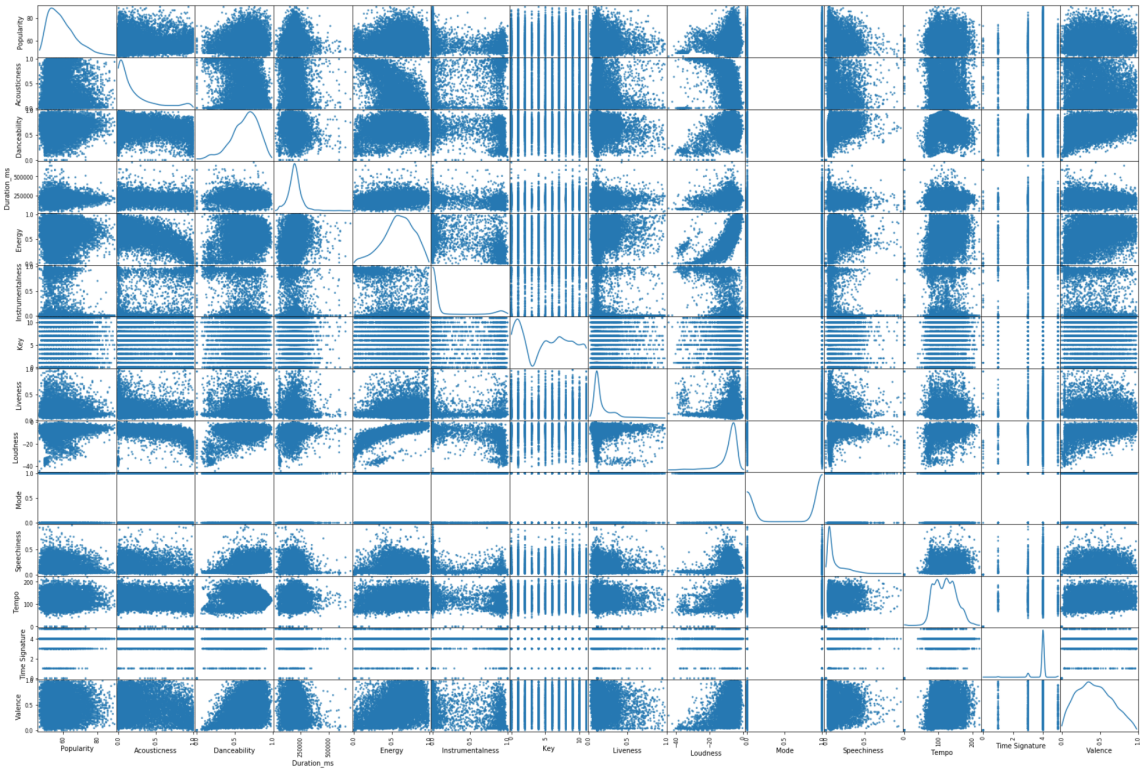
First, we used the Spotify API to query the 10,000 most popular songs on the platform in 2018. In hopes of modeling a diverse range of audio features, we decided on the following quantitative predictors auto-generated by Spotify for each track: acousticness, danceability, duration (in milliseconds), energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, and valence. We saved the queries as a `.csv` file that we then loaded into Jupyter Notebook as a 10,000 x 17 dataframe for exploratory data analysis. Each of the 10,000 rows corresponds to an individual song, and the 17 columns consist of artist ID, track name, track ID, popularity (a numerical measure determined by Spotify), and the 13 audio features mentioned earlier.

Next, we peeked into and cleaned the data. We examined the first several rows of the dataset, calculated summary statistics for each of the features, and printed out the data types of each of the features, looking for any unusual patterns or results. Everything appeared to reconcile with what we'd expect—just in case, we also selected a few of the songs at random and searched them on Spotify, to ensure that the track ID, artist, and artist ID were correct and that there were no misalignment issues during the querying process. We also checked for missing values and did not find any.

Now that our data was cleaned and reconciled, we created a new column for our binary response variable, `tophit`. As discussed in Milestone 2, our project will include both quantitative response and classification frameworks. In the first set of models, we'll follow Spotify's lead and use a quantitative popularity metric between 0 and 100 as our response variable.

This allows us to later *rank* unfamiliar songs by their expected popularity. In the second set of models, we'll create a binary response variable indicating whether a song is popular or not, where popularity is initially defined as having a quantitative popularity score greater than 60. We decided on this cutoff because out of the 10,000 songs in our 2018 dataset, 60 is approximately the median popularity score, and we would like an even split of levels for our new binary response variable.

Finally, we began to look for relationships between the variables. We created a matrix of pairwise scatterplots between the variables. It was difficult to identify clear associations for most of the variable pairs (especially since many of the plots had non-constant variance), but we did notice a positive association between energy and loudness and a negative association between energy and acousticness. This suggests the presence of multicollinearity; we should explore techniques for dimensionality reduction, variable selection, and variable importance (such as principal component analysis, stepwise regression, and permutation feature importance) in later models. To anticipate which predictors might be significant in modeling popularity, we also included the quantitative response variable `popularity` in the scatter matrix. From an initial glance at the first row of the scatter matrix below, it seems that there are very slight positive associations between popularity and the features loudness and time signature, and very slight negative associations between popularity and the features liveness and speechiness. For some of the audio features, we notice more of a quadratic relationship. Popularity tends to increase as song duration increases up to a point, before deteriorating as songs become too long. Similar patterns hold for danceability and energy. The nonlinearity of these relationships suggests that we should explore including quadratic and higher-order polynomial terms in our later models.



Baseline Model

We decided to first tackle the classification portion of our project before moving on to the regression approach, so our baseline model uses the binary response variable `tophit`. We created a predictor dataframe with only the 13 audio features from Spotify, then applied an 80-20 train-test split, stratifying by popularity. As a baseline classification model, we fit a single decision tree on the training data, using cross-validation to determine the best maximum tree depth. From the cross-validation procedure, we obtained a best maximum tree depth of **3**, which yielded a classification accuracy of **0.645** on the training set and **0.6395** on the test set. Since the classification rate decreases only marginally from the training set to the test set, our current model is most likely not overfitting, and we have room to explore more complex models.