Project Status Report
May 16, 2016
Eric Hao ezh825, Dylan Ong dco668, Max Schuman mas608, Siyu Zhang szw910

I. Project Objective

Our task is to predict the winner of and nominees for the 2017 GRAMMY award for Record of the Year. The output of our project will be a ranking of songs based on their probabilities to win the GRAMMY award, given a list of the current top 100 songs eligible for the award and their relevant attributes. Currently taking 19 attributes into consideration, we are aiming to find a deeper, predictive connection between the Record of the Year and the characteristics of songs. Our results will hopefully yield insights into subtle differences between popular expectations and actual winners, especially in upset years, and can help provide context for identifying shifts and trends in popular music.

II. Dataset for the Project

We spent a lot of time writing python scripts, using various APIs, and scouring the web, scraping data from various sites to obtain various attributes. The data set we have compiled thus far includes roughly 5400 songs from 1958 to 2015. All of these songs were part of the Billboard Year-End Top 100 list. We can improve this data set by adding songs not necessarily on that list in the near future. It also includes all past Record-Of-The-Year winners and nominees.

The complete list of attributes:

Identifiers:

song_title, spotify_id, artist, year (from 1958 to 2015), is_winner (0 for No and 1 for Yes),
is_nominee (0 for No and 1 for Yes)

15 numeric attributes:

popularity (0-90), danceability (1-10), energy (0-10), loudness (-30 to -1), speechiness (0-1), acousticness (0-1), instrumentalness (0-1), liveness (0-1), valence (0-1), tempo (30-220), duration_ms (90000-2000000), word_count (1-1400), reading_ease (-2 to 150), polarity (-1 to 1), subjectivity (0 to 1)

4 nominal attributes:

genre (Disco, Hip-Hop, RNB, blues, classic_rock, country, electronic, folk, funk, indie, instrumental, jazz, latin, oldies, pop, punk, rap, reggae, rock, soul), key (0, 1, … ,11), mode (0 for minor and 1 for major), time_signature (1, 2, … , 5)

When we make our predictions for next year's award winner and nominees, we plan to weigh more recent songs more heavily in our model, as we feel that more recent trends in Record of the Year winners will be more useful in predicting future winners. As of now, however, we have decided not to consider the impact of song release year in our modelling, as we are more concerned with the overall predictive ability of each model at this stage and will later factor in time.

III. Preliminary Models and Results

As our task involves ranking a group of songs based on their likelihood of being nominated for and winning the Record of the Year award, we focused our preliminary modeling efforts on models that would output a continuous range of values representing this likelihood that we could compare between songs in the same year, such as logistic regression and naïve Bayes models.

Using a logistic regression model with 20-fold cross-validation we found that 66 of the 286 nominees in our data set are correctly classified as nominees and 7 of the 58 winners are correctly classified as winners. Similarly, using a naïve Bayes model, 80 nominees were correctly classified as nominees and 28 winners were correctly classified as winners. We found that the naïve Bayes model has a much higher recall when predicting the winners and nominees in the set.

However, classification recall is a flawed metric for our purposes, as the nature of our task involves ranking songs within years to determine the most worthy Record of the Year winner in a given year, not identifying the winners from all years in a batch of songs released over the course of sixty years. Therefore, we chose to look at the classification probabilities outputted by our logistic regression model for each song in our training set as a "rating" of each song, and within each year we checked to see if the eventual winner was among the highest-"rated" songs eligible for the award that year.

By this analysis, our logistic regression model rated 15 of the true winners as the strongest songs in their years and had 28 total winners rated among the year-by-year top-fives (the usual cutoff for award nominees in a year is five), an encouraging result. In particular, our model was particularly good predicting winners in the 60s, 70s, and 90s, and weaker in the 80s and in the last 10 years, indicating that weighing more recent data more heavily in our modeling could be wise as we look to predict future years. We plan on evaluating the naïve Bayes model and other models by the same criteria moving forward.

IV. Next Steps

In the future, there are modifications we would like to make to our data set in order to both capture a wider range of music and improve our predictive capabilities. Although we feel our current set is sufficient for our purposes, we will look to add more songs to our training set, particularly those which are not part of the Billboard Year-End Top 100 list. This list is not very comprehensive because it does not include songs that may have been a top 100 song during the year, but was knocked off the list by the time the year came to a close. Also, including songs that perhaps never made the top 100 list is an interesting idea we will explore. Aside from expanding the data set, we will look towards obtaining attributes for songs that help paint a fuller picture of the song's qualities, such as information about artist age and ethnicity as well as additional measures that might help better describe a song's style and character.

As stated above, we have ignored the influence of time in our present models. However, the trends of music change as time passes; as such, we plan to determine the best way to weigh more recent results more heavily in our predictions. Moreover, we will refine our ways to train the data and try different validation methods. As we believe performance in predicting recent winners is an important sign of potential success in future predictions, our potential plan going forward is to train using all songs older than 2010, weighing recent year data more heavily, and then to validate by examining the song ranking by year in the set of most recent songs. We are open to feedback and suggestions about this plan.

We will complete our project by presenting a final ranking of current songs eligible for next year's Record of the Year award that we predict to be most likely to win the award if the GRAMMYs were held today.