**Project Status Report**

**Team member: ezh825, dco668, mas608, szw910**

**(I) Project Objective**

Our task is to predict the winner and nominees of the GRAMMY award for Record of the Year 2017. The output of our project will be a ranking of songs based on their probabilities to win the GRAMMY award. Taking 20 attributes that may exert any influence on the possibility of a song to win the GRAMMY into consideration, our project presents a detailed analysis of connections between the GRAMMY award and characteristics of songs. The result will not only help artists and producers understand how to work toward winning the award, but also yield insights to subtle differences between popular expectations and actual winners, especially in upset years.

**(II) Dataset for the Project**

We spent a lot of time writing scripts, using various APIs, and scouring the web to scrape data from various sites to obtain various attributes. The big data set we have compiled so far includes 5800 songs from 1958 to 2015. All of these songs were part of the Billboard Year-End Top 100 list. We can easily improve this data set by including songs not necessarily on that list in the near future. They also include all Record-Of-The-Year Grammy winners and nominees. Some example attributes out of the 20 we have collected include popularity, genre, word\_count, and even danceability.

**The list of attributes is as following:**

*4 attributes not included in the arff file:*

song\_title, spotify\_id, artist, year (from 1958 to 2015)

*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)

*6 nominal attributes:*

genre (Disco, Hip-Hop, RNB, blues, classic\_rock, country, electronic, folk, funk, indie, instrumental, jazz, latin, oldies, pop, punk, rap, reggae, rnb, rock, soul)

key (0, 1, … ,11), mode (0 for minor and 1 for major), time\_signature (1, 2, … , 5),

is\_winner (0 for No and 1 for Yes), is\_nominee (0 for No and 1 for Yes)

We plan to weigh data from recent years more when training our data by applying different weight in Logistic Regression or modify K-nearest neighbor later. At the initial period, we have decided not to consider the influence of years in order to check whether our model works. So we have scrambled the data to exclude the influence of time. When we use cross-validation, it will not take chunks of songs that are close together chronologically.

**(III) Models and Preliminary Results**

In order to get the ranking of probabilities, we have decided to use Logistic Regression Model and Naïve Bayes Model.

We first tried Logistic Regression Model to run the scrambled data. We used 20-fold cross-validation and the result is that 66/286 nominees are correctly classified as nominees and 7/57 winners are correctly classified as winners. Then we tried Naïve Bayes Model. The result is that 80/286 nominees are correctly classified as nominees and 28/57 winners are correctly classified as winners. We found that Naïve Bayes Model has a much higher accuracy when predicting the winners.

However, our performance metric is not the accuracy percentage but the probabilities and rankings because we plan to present the probabilities of songs to win the GRAMMY award and give a list of the most possible winners. So we used the scrambled data as training data and ran it on the Logistic Regression Model to check our accuracy about predicting the winner of each year. We first excluded the is\_nominee field. For each year, we did: (1) use the data from this year as testing data, (2) use the data left as training data to build the model, (3) run testing data on our model and make predictions, (4) compare our predictions with the actual results. Our result for this year-by-year predictions with the new model is: 15/58 years predicted the winner correctly and 28/58 years had the eventual winner in top 5 songs based on their probabilities to win the award.

Furthermore, according to Logistic Regression Model, we found our that Genere=RNB is the most negatively related factor to winning the award, and Time\_signature =1 (which means 1 beat per measure) is the second, Genre=PUNK is the third.

Though those results are not as high as expected, they are acceptable for our initial models. Since we plan to add more attributes and take the year factor into consideration, we think we can get higher accuracy toward the ranking of possible winners.

**(IV) Future Plan**

*(i) About Dataset:*

Regarding the data set, there are many modifications we would like to make in order to both expand our data set and better our model. The most obvious is to include more songs, 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 definitely work on obtaining more attributes. We are currently trying to scrape data about the ethnicities of the artists as well as their ages. For our particular model, having as many different attributes as possible is beneficial because we can then select the best combination of them, as well as exclude those which don't help the model.

*(ii) About Models:*

For our present models, we have ignored the influence of time. However, the trend of music changes as time goes. So we will figure out how to weigh more recent results more heavily in our predictions, in particular, how to assign different weights to data from different years in Logistic Regression Model. Moreover, we will refine our ways to train the data and try different validation methods. Since we are more concerned with more recent years’ performance, we plan to train on the older years set, weighing more recent years’ data more heavily and then validate by performance on the most recent years’ set. But our performance metrics will not be things like accuracy since our output will be the rankings of songs with their probabilities to win the GRAMMY award.

*(III) Predictions*

We will finish our project by presenting the ranking of songs, each with the probability to win the award and making the predictions that which songs are the most possible nominees.