Project Proposal for FMA

Yiyi Chen, Avi Dixit, Sayan Sanyal, Ed Yip
November 20, 2017

Primary Question

In this project we seek to understand what makes a song popular, and specifically if one can predict the popularity of a song based on its audio features and metadata. While the virality of a song might depend on many social factors, such as the effectiveness of its marketing campaign and demographics of its listeners, we hypothesize that the inherent characteristics of a song, such as its name and extracted musical features can also be correlated to and indicative of its popularity.

To further explore this question, we will first define what popularity means within the context of the dataset. We will then build a baseline model to predict a song's popularity, using its audio features and metadata available from the dataset. From there we will expand and refine our model to answer three main questions -

- What's the model's overall performance on the test dataset? What are the most important features in determining a song's popularity?
- How much, if any, does its performance differ across different genres? How do feature importances differ across genres?
- What's the external validity of the model? How does the model perform against songs not in the test dataset, e.g. songs from the Billboard Top 100?
- (If time allows) Can we design better features to improve the overall performance of the model?

Understanding what factors contribute to a song's popularity has practical significance in a few areas. It helps us better understand

- How people evaluate music consciously and subconsciously,
- How people's preferences vary across genres, and
- Any systematic differences in preferences between people who listen to songs on Free Music Archive and the general public.

These insights can be leveraged by music producers to assess the chance of popularity for their songs pre-release, and by future researchers interested in similar problems.

Dataset

We are using the dataset compiled by Defferrard, Benzi, Vandergheynst, and Bresson, sourced from Free Music Archive. The full dataset includes 106,574 tracks. For each track, there are 52 metadata features, e.g. genre, title and date_released, and 518 audio features, e.g. Mel-frequency cepstral coefficients and spectral

contrasts. The creators have also partitioned a small, medium and large sample of the entire dataset for faster processing.

Even though the dataset was intended for benchmarking music genre recognition algorithms, its rich features have enabled us to explore a different question as proposed above. Specifically, we will be using one or some combination of the favorites, interest and listens column in the `tracks.csv` metadata file as our target variables.

On an initial investigation, there seem to be significant amount of null values in the data, e.g. 53.5% of tracks having their root genre missing. Therefore, we expect to spend some time further pre-process the data (handling nulls and removing duplicate rows/columns) before building the baseline model.

Prior Work

There are two prominent sources of previous work done in this particular field of play count prediction.

The first, by Dorien Herremans et al from University of Antwerp in Belgium, focused more on "hit song science" and singled out the genre of dance music to create a prediction algorithm that forecasted whether a song would be a top 10 hit. They used Echo Nest to extract 139 different musical aspects from 3500 songs charted in the top 10 from 1985 to 2014 to analyze each song (e.g. song length, tempo (bpm), time signature, beat, energy, danceability, timbre, tone color, listener feel, etc.). Feature selection and normalization were done through CfsSubsetEval from Weka with GeneticSearch to reduce data to 35-50 attributes. Feature selection was done in order to avoid the 'curse of dimensionality', and allow for thorough testing of model, improved model comprehensibility, and better performance of learning algorithm. 5 models were built for each dataset to determine whether songs would be a hit or not: decision tree variation of a C4.5 tree, RIPPER ruleset, Naive Bayes, Logistic Regression, and SVM using Polynomial and RBF kernel. Logistic regression (with an 83% accuracy score) performed the best.

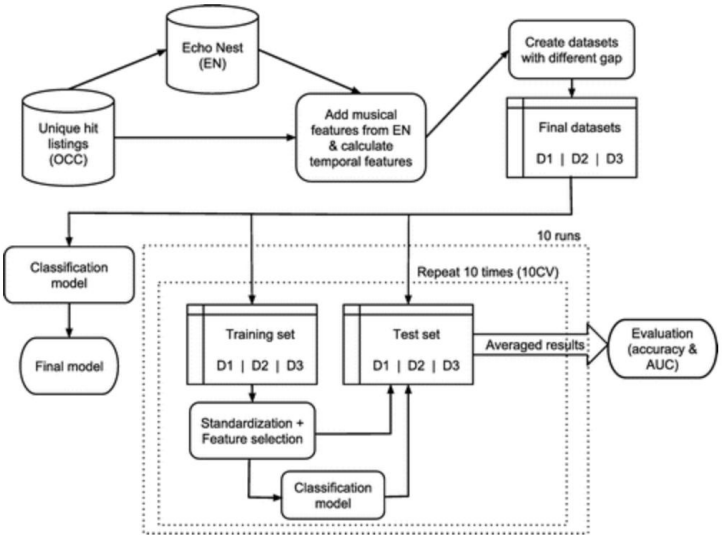


Figure 1: Flowchart of Experimental Setup

Though this research was not directly relevant to play count prediction, the motivations behind and approach toward predicting hit songs are still similar to predicting which songs will receive the most play counts.

The second was done by Matthew Moocarme, who used Apache Spark to implement Alternating Least Squares model to predict song listens for a given user. The process involved determining best values for parameters from ALS, train over several models based on rank (# of columns in user matrix), then select the best model to use for the rest of the predictions. Specifically, Moocarme used collaborative filtering, normally used to predict the rating of an item, to predict the number of plays with the hypothesis that a greater number of plays for a song would equate to a higher ratings. Collaborative filtering was used to approximate user matrices by factorizing the two matrices of properties of a user with properties of each song. ALS was then used to minimize the error in predicting the number of plays by using a fixed set of user factors and the known number of plays to find the best song factors using least squares optimization and then switching out user factors with song factors to find the best user factors. Accuracy for the approach was not very good with an RMSE was around 9 for predicting playcounts of songs excluding a playcount of 1 whereas predicting songs including playcounts of 1 had an RMSE of around 6.

Proposed Analysis

For this particular work, we seek to focus on interpretability of our models while also optimizing for external validity.

To begin with, we plan to turn this problem into a classification problem from a regression problem. We need to operationalize our definition of "popular", and we will do that with a combination of the listens, favourites and interests labels. We shall analyze the distributions of the songs by these columns and manually threshold values at what seems to be a clean break to indicate popularity. We are also considering creating a popularity index which would be a composition of the labels mentioned above. One possible way of getting this index would be to take the first Principal Component of the three mentioned labels to capture the variation across the indicators.

We then plan to use interpretable algorithms like logistic regression and decision trees to try and predict the different popularity measures. After tuning the models to optimize for accuracy and regularizing to avoid overfitting, we will analyze the important features that most inform the models. We shall perform feature engineering over and above the large number of features we already have as and when necessitated by our analysis. The dataset we are working on is probably the largest musical dataset collected, and there has not been much work done with it yet. The large *N* value helps us avoid the "curse of dimensionality" for the most part, but we shall perform ablation tests to understand how each feature contributes to our mode.

If there is variation between the models, we shall report so. We shall also look to train and predict popularity within genres and see whether the difference in feature importances point towards a larger trend in most songs rather than features than are specific to a genre itself.

Lastly, we will acquire data for songs for each of the represented genres from the Billboard Top 100 to see whether our model generalizes to songs that are commercially popular. We shall report our results across genres.

Appendix

Summary Statistics and Figures

The dataset consists of 106,576 track that are split across 16 top level genres.

Total # of observations	106,576 tracks
Subset of Columns for each observation	Album: - Date Created: Date the album was created - Favorites: Number of people who marked as favorite - Listens: Number of people who listen to the album Artist: - Favorites: Number of people who marked artist as favorite - Location: Location of the artists - Name: Name of the artists Track: - Date Recorded: Date the track was recorded - Duration: Duration of the track - Favorites: Number of people who marked the track as favorite - Genre: Top level genre this track belongs to - Listens: Number of people who listen to the song - Title: Title of the track Spectral Audio Features: - Spectral Rolloff: Roll-off frequency - Spectral Bandwidth: P'th-order spectral bandwidth - RMSE: Root-mean-square energy for every frame - ZCR: Zero-Crossing rage of an audio time series

Additional Statistics

Genre Level Characteristics

1. Top 10 genres by number of tracks:
    - Experimental: 38,154 tracks
    - Electronic: 34,413 tracks
    - Rock: 32,923 tracks
    - Instrumental: 14,938 tracks
    - Pop: 13,845 tracks
    - Folk: 12,706 tracks
    - Hip-Hop: 8,389 tracks
    - International: 5,271 tracks
    - Jazz: 4,126 tracks
    - Classical: 4,106 tracks

2. Top 10 genres by track duration:
    - Spoken: 590 seconds
    - Jazz: 381 seconds
    - Experimental: 362 seconds
    - Classical: 310 seconds
    - International: 300 seconds
    - Easy Listening: 289 seconds
    - Soul-RnB: 264 seconds
    - Electronic: 263 seconds
    - Blues: 258 seconds
    - Instrumental: 243 seconds

3. Top 10 genres by tracks per album:
    - Hip-Hop: 8 tracks
    - Electronic: 7 tracks
    - Pop: 6.7 tracks
    - Instrumental: 6.64 tracks
    - Experimental: 6.51 tracks
    - Rock: 6.45 tracks
    - Classical: 6.21 tracks
    - Folk: 6.12 tracks
    - International: 5.96 tracks
    - Soul-RnB: 4.72 tracks

Artist and Track Level Characteristics

1. Top 10 tracks by Favorites, Listens, and Interest:

   Interest: Night Owl (3,293,557), Enthusiast(1,991,344), Siesta(1,563,555), It's Your Birthday!(1,371,526), Epic Song(1,314,156), Hachiko(1,284,065), Fater Lee(1,038,669), Running Waters(876,441), Springish(714,549), Starling(626,592)

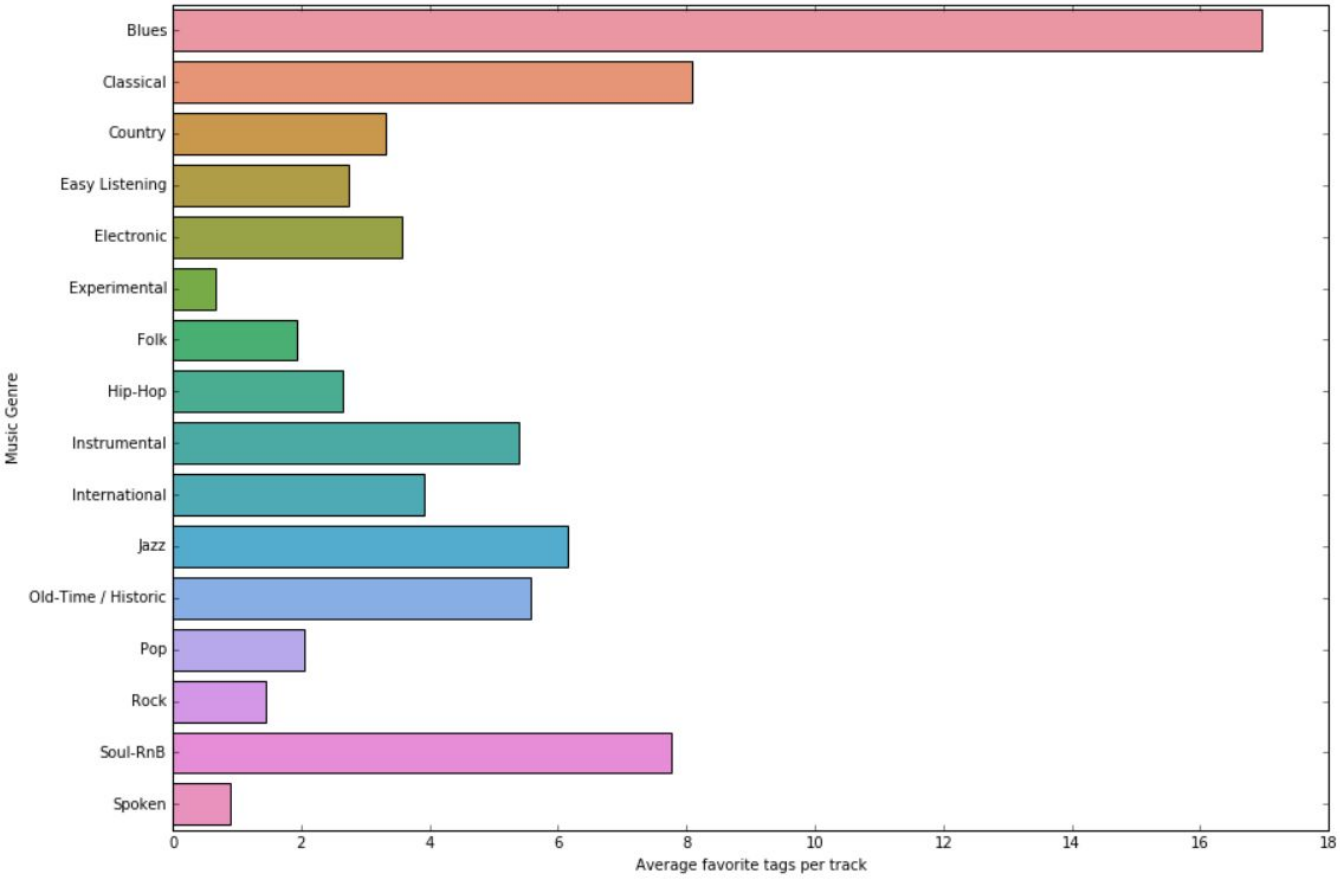
   Favorite: Night Owl(1,482), Something Elated(961), Siesta(796), Enthusiast(765), The Temperature of the Air (633), Hachiko(600), Springish(600), Epic Song(599), Kopeika(554), Requiem for a Fish(537)

   Listens: Night Owl(543,252), Starling(491,235), Springish(468,163), Fater Lee(433,992), Epic Song(429,168), O Tannenbaum(400,404), Hachiko(374,497), Siesta(356,588), It's Your Birthday!(349,903), Enthusiast(335,215)

2. Top 10 Artists by Favorites Count:

   James Kibble (Classical): 319
   Blue Dot Sessions (Instrumental): 230
   The Impossebulls (Hip-Hop): 210
   The Polish Ambassador (Electronic):174
   Cheese N Pot-C (Hip-Hop):159
   Disco Missile (Rock): 144
   Big Blood (Folk): 134
   Obitis (Rock): 126
   FortyOne (Experimental): 125
   Phemale (Pop): 120

Interesting observation: Plot of Average number of Favorite tags for a Track in top level genres



Bar plot that shows the average number of favorite tags for tracks divided by genres. Given that the research question we are looking to explore is the prediction of the popularity of a song, we will have to do a substantial amount of feature engineering to determine what artist, album, or track level feature is relevant in predicting a track's popularity. Further, we will also have to determine what defines a song's popularity, given that the data we have has information on the favorite tags per track/artist, the number of people who listen to a track, and the number of people who are interested in the track. The above bar plot shows that when considering just one of those features, the favorites tag, we see that on average a track belonging to the *Blues* genre tends to have the most favorite tags, followed by *Classical* and *Soul*. At the bottom, we see that *Experimental* and *Spoken* have the lowest average for the number of favorite tags per track.

Papers:

Recognition of music types

Neural Network Based Model for Classification of Music Types

Automatic Genre Classification of Music Content

Music type classification by spectral contrast feature

Articles:

Machine predicting hit dance songs (2015)

Making Music Prediction with Apache Spark