Aaron Munoz
Project McNulty

Classifying Hip Hop vs Pop

Goal

Music streaming platforms such as Spotify and Apple Music have become incredibly popular, reaching over 85 million subscribers. Reliable genre classification is important for some of these companies' most popular features, such as user targeted playlists and 'similar artist' lists.

The purpose of this project was to create a model that could identify hip hop songs from pop songs. A particular product use case was not developed for this project, so F1 score was used as the metric for evaluating models since it's a nice balance between precision and recall.

Data

1 million songs worth of music data ("time signature", "key", "tempo", etc) and 250,000 songs worth of song lyrics were obtained from "The Million Song Dataset" and Musixmatch. Genre information was stored as a many-to-one relationship to an artist, so an artist's most popular genre tag was assigned as their main genre for the purpose of this project.

Modeling

All feature and model selection test runs were conducted on 80% of the data, with 20% set aside for final testing. Each "test run" was conducted with 5-fold cross validation and returned the mean score between all 5 folds.

Baseline modeling with just music data didn't show much promise, producing a ROC AUC score of 0.52. Given the weak signal from these features, it was determined that lyrics should be brought into the modeling.

10,226 pop and 3266 hip hop songs with lyrics were selected from the pool of songs. 4000 unique words were found in the combined 13,490 dataset. Given the unbalanced distribution of the classes, 2:1 class weighting was used with all model testing.

A testing pipeline was created to test both the viability of multiple models, as well as finding the optimal number of world features. Starting with the 10 most common words in the dataset and increasing the feature space with the next 10 most common words each iteration—Accuracy, Precision, Recall, and F1 scores were recorded with Logistic Regression, Gaussian Naive Bayes, Decision Tree, and Random Forest models. Ultimately a Random Forest model with 100 estimators was found to be the best performing model (F1 ~0.57) while using the top 300 most common lyrics.

Training a Random Forest model with validation data and scored with the held out 20% test data on the top 300 most common words produced a 0.602 F1 score, which is a score that indicates the model is picking up on a meaningful relationship between the lyrics and genre.

Conclusions

It should be noted that this model scores higher on Precision(0.69) than Recall(0.53). This score could indicate that the model is very good at classifying particular sub-genres of hip hop, but misses out on others. The n-word, the word 'rap', and several forms of profanity consistently appearing in the top 15 most important features in Random Forest modeling probably indicates this model is mostly picking up on 'harder' forms of rap.