Predicting Songs Popularity

Mengqi Fan[1] (fanm3@rpi.edu)

[1]Rensselaer Polytechnic Institute 110 8th St., Troy, NY, 12180 United States)

Introduction

Music has been an integral part of our culture all throughout human history. In 2017, the music industry generated $8.72 billion in the United States alone. Of this $8.72 billion, the majority of the revenue is generated by popular, mainstream songs. Popular songs secure the lion's share of revenue. Therefore, having a fundamental understanding of what makes a song popular has major implications to businesses that thrive on popular music, namely radio stations, record labels, and digital and physical music marketplaces.

Every song has key characteristics including lyrics, duration, artist information, temp, beat, loudness, chord, etc. The aim of this project is to investigate the following questions:

- Are there certain characteristics for popular songs?
- What are the largest influencers on a song's success?
- Is it possible to predict popular songs with machine learning using audio features? What is the best performance of our machine learning models?

Data Description

We used data from Kaggle datasets collected from Spotify web API. Spotify for Developers offers a wide range of possibilities to utilize the extensive catalog of Spotify data. One of them are the audio features calculated for each song and made available via the official Spotify Web API. Combined all the information we have got, this dataset met our criteria (good and "professional" data source, audio features with high variety, most recent data, etc.), and hence is chosen for my project.

The 16 audio features in the data include attributes about the music track itself (duration, key, etc.) and more abstract features (danceability, energy, etc.). Our project uses a subset of this data.

Analysis

1. Basic Summary of Data

- Original dataset has dimension of 116372 * 17.
- We use a subset of the original data as our training (5000 records) and testing (5000 records) datasets.
- Three factors: artist_name, track_id, track_name.

2. Feature Distribution

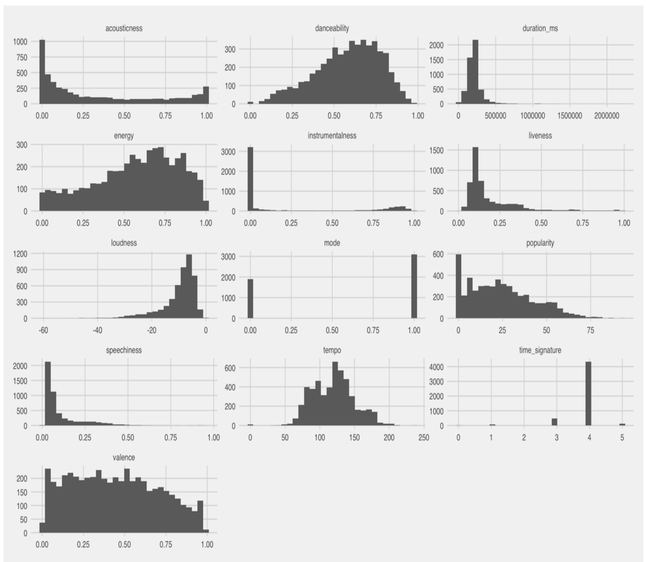

Figure 1. Distribution of Different Features

- Most of the songs are not acoustic (classical music).
- Most of the songs are not recorded during concert.
- Songs pretty fast: BPM distribution has high mean.

3. Features Correlations

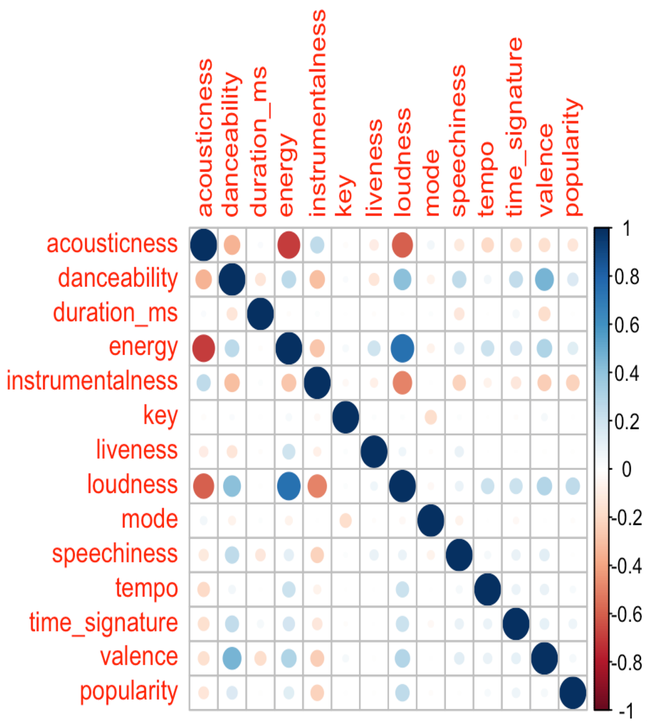

Figure 2. Correlation Matrix

- loudness & energy/ energy & acousticness.
- Not much dependence of popularity with predictors.
- Top 5: Loudness, Instrumentalness, Energy, Danceability, Acousticness.

4. More on Feature Engineering

- Drop the energy variable.
- Convert duration to seconds.
- Transformed key/time_signature to factors.
- Mode variable: "major" and "minor".
- Popularity_new: "like", "dislike", "neutral".
- Scale the numerical variables.
- Create dummy variables.

5. Outliers

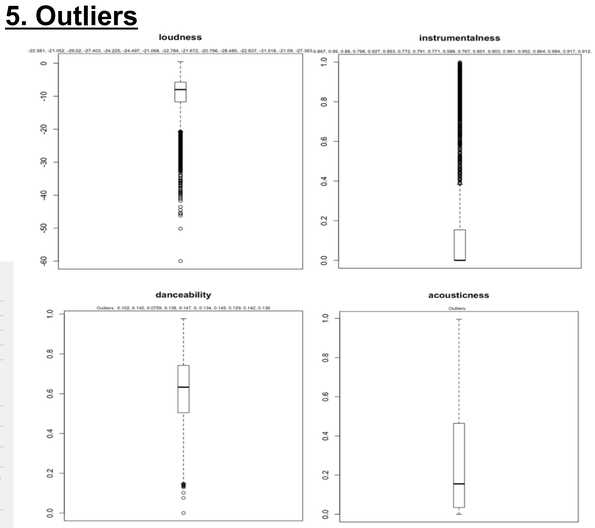

Figure 3. Boxplots of Predictors

2. Model Development & Evaluation

- Algorithms: Linear Discriminant Analysis, Support Vector Machines, Decision Trees, Random Forest, k- Nearest Neighbor.
- Good mixture of simple linear (LDA), nonlinear (CART, kNN) and complex nonlinear methods (SVM, RF).
- First Trial: all independent variables as input.
- Second Trial (Optimization): Part of predictors.
- Overall, LDA has the best performance.

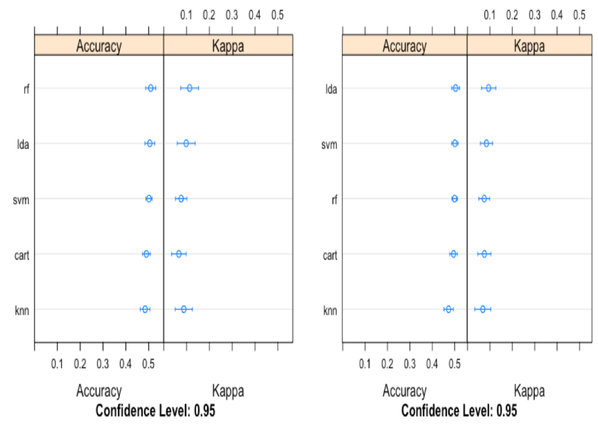

Figure 4. Accuracy Comparison

3. Making Predictions

- Use LDA to predict songs popularity in original testing dataset.
- 52.42% accuracy with 95% confidence. The accuracy is close to accuracy in validation data.

4. Modeling Conclusion

- Overall the accuracy of these 5 models are not good (less than 60%).
- The biggest possible reason might be all predictors have low correlation with target variable.
- Since we chose to use a dataset containing several different genres, the dataset might have been too complex to generalize and classify.
- Another reason is that we randomly choose our cutoffs to divide the previous popularity value into three classes, which may lead to different prediction results.

5. Text Analysis

Figure 5. Popular Artist Names vs Popular Track Names

Modeling & Text Analysis

1. Preparation

- Use 10-fold cross validation to estimate model fit.
- Model Tuning: Grid search (tuneLength=5).
- Metric: "Accuracy" - ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate).

Future Work

One improvement would be to use more features than we did in our research. We believe that including more metadata such as genre, label, lyrics, and artist popularity would improve the accuracy of the models considerably. Another improvement would be to do feature selection and model optimization. Lastly, it would be favorable to define what a hot song is in a more explicit way.

Poster: MT15A-08

Glossary:

RPI – Rensselaer Polytechnic Institute

TWC – Tetherless World Constellation at Rensselaer Polytechnic Institute

Acknowledgments:

Person1 and Person2 from the Tetherless World Constellation at RPI

Resources:

Spotify Audio Features: https://www.kaggle.com/tomigelo/spotify-audio-features
Your First Machine Learning Project in R Step-By-Step (tutorial and template for future projects):
https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
Predicting Hit Songs with Machine Learning: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1214146&dswid=6570
Song Popularity Predictor: https://towardsdatascience.com/song-popularity-predictor-1ef69735e380
Predicting Song Popularity: http://cs229.stanford.edu/proj2015/140_report.pdf
Lyric Analysis with NLP & Machine Learning with R: https://www.datacamp.com/community/tutorials/R-nlp-machine-learning