ISYE 7406 – Spring 2024

Project Proposal

Team members: Hong Yee GAN, Fangfei LI

Project name: Predictive Classification for Song Popularity

Project description:

Individuals have established strong connections with songs and music, recognizing their potential to enhance mood, alleviate pain and anxiety, and provide avenues for emotional expression. Extensive research indicates that music can contribute significantly to both physical and mental well-being.

Recent efforts in various studies have been dedicated to comprehending the popularity of songs, exploring specific influencing factors. These studies involve the analysis of song samples, breaking them down into distinct parameters that are meticulously documented for analysis.

To produce a popular song, strategic planning and extensive market research are required to analyze competitor strategies, audience demographics and potential gaps. These methods often involve a significant investment of time, resources, and expertise. Considering these factors, utilizing modeling to predict song popularity emerges as an efficient approach.

The objective of this project is to build classification models using song attributes to predict whether a song is popular. These models, which include Naïve Bayes, Logistic Regression, Random Forest and K-Nearest Neighbors (KNN), will be developed. To accomplish this, a repeated cross-validation methodology was applied, using 10 folds and 3 repetitions on the training set. The success of the models will be evaluated based on classification accuracy. The optimal model will exhibit minimal disparity in classification accuracy between the training and test datasets, indicating the ability to generalize to unseen data. Subsequently, the models will be further interpreted using confusion matrix parameters.

The project timeline spans approximately one and half months, structured into distinct phases. The initial stage involves one week dedicated to data cleaning and exploratory data analysis. Following this, two weeks are allocated for prediction modeling. Subsequently, one week is designated for the preparation of presentation, and the final week is reserved for compiling the comprehensive final report.

Dataset

The Song Popularity dataset, widely utilized for analyzing song popularity, is available on Kaggle at https://www.kaggle.com/datasets/yasserh/song-popularity-dataset/data. This dataset comprises of 18,835 records, providing detailed insights into song attributes such as duration, energy, key, liveness, loudness, and danceability. After eliminating duplicate entries, the dataset consists of 13,070 observations, featuring one response variable, song popularity and fourteen predictors.

To facilitate classification analysis, a binary variable, 'popular', was introduced. This variable categorizes songs into two classes: '1' denotes popular songs, indicating a popularity level above the median, while '0' represents less popular songs, falling below the median.

After this categorization, the result is a balanced dataset with 6,571 observations categorized under the less popular class and 6,499 under the high popularity class. Furthermore, the dataset was partitioned into training and testing sets, with 30% of the observations randomly assigned to the test set, and the remaining 70% allocated to the training set. Summary of the dataset breakdown can be found below:

	Train Set	**Test Set**	**Total Count**
Class '0': Less popular songs	4,600	1,971	6,571
Class '1': High popularity songs	4,549	1,950	6,499
Total	9,149	3,921	13,070
Dataset Percentage	70%	30%	100%

Scientific Research Questions

Primary research question: "What specific musical attributes significantly influence the popularity of songs and to what extent do they contribute to a song's popularity?"

Supporting research questions:

1. "To what extent does danceability contribute to the appeal and categorization of songs as popular or less popular?"
2. "Is there a significant correlation between high energy level songs and classification of a popular song?"