MACHINE LEARNING FOR MUSIC: PREDICTING USER ENGAGEMENT USING SONG CHARACTERISTICS

API 222 - Machine Learning & Big Data Analytics Project, April 2023

Rohini Iyer and Pranav Bhargava

Introduction

- **Problem Motivation –** Using machine learning techniques to help artists and music companies predict the expected views/streams that new tracks might garner (i.e., their popularity), using data on song characteristics like danceability, loudness, valence, tempo, etc. from past songs.
- **Current methods used –** In addition to complex analyses using streaming, sales, and user data; companies currently conduct market research and focus groups, rely on industry experts, and social media analytics to assess user engagement AFTER the songs have been released.
- **Use of ML to augment existing approaches –** ML models can identify patterns and trends in data that may not be immediately apparent to human analysts, making the talent selection process for music companies more data driven. They can also predict user engagement BEFORE songs are released, enabling artists to create content with greater potential for mass appeal.

Models and the rationale behind it

- **Linear regression** – to establish a MSE baseline if there exists a linear relationship
- **KNN** – to take advantage of a non-parametric supervised learning model
- **Decision Tree** –easy to interpret and graph. Not computationally intensive (only 10 predictors)
- **Bagging** – further reduces variance of a decision tree while keeping the same bias
- **Random Forest** – provide an improvement over Bagging
- **Boosting** – another extension of decision trees that might give better performance

Results

Mean Squared Error of Prediction

Model fitted	Views (YouTube) (in 10^{16})	Likes (YouTube) (in 10^{12})	Streams (Spotify) (in 10^{16})
OLS	6.64	2.85	6.69
KNN	6.66 (k = 100)	2.86 (k = 187)	6.78 (k = 72)
Decision Tree	6.59	2.88	6.85
Random Forest	5.68	2.31	5.42
Bagging	5.85	2.37	5.50
Boosting	6.91	2.77	6.29

Predictors used: Danceability, Key, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo

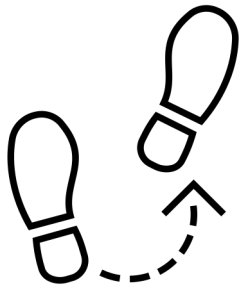
Seed: 222 | **Original N:** 20718

Missing data: 1655 observations had missing data across one or multiple variables and hence were dropped from the data

Training data: Contains 80% of non-missing data observations = 15250 observations

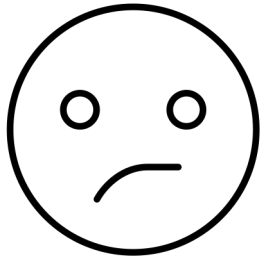
Test data: Contains 20% of non-missing data observations = 3813 observations

Path forward



- **Scope (open questions/things to do)**
  - Classification of upcoming songs as potential "hits" or "misses"
  - Grouping songs into "genres" or "categories" to improve user experience
  - Deep diving into the predictor - are the different predictors correlated?

- **Limitations**
  - Inability to predict hits due to missing song release data
  - Inability to do recommendation analysis due to lack of user data



- **Concerns (with current data) that hinder adoption**
  - Calculating the values of the predictors
  - User engagement could also depend on factors other than song characteristics