Abstract

This is the project proposal for Cultural / Entertainment problem

24S1-C-NYP-ITI105 - MACHINE LEARNING PROJECT

Milestone Report

Lye Suh Jeng (7487427Y)  
Lee Li Neng (6203055B)

Lim Chan Boon (9704541M)

Date: 8 August 2024

Table of Contents

[1. Project Problem 2](#_Toc173441880)

[2. Project Objective 2](#_Toc173441881)

[3. What we’ve accomplished 2](#_Toc173441882)

[3.1 Data Collection 2](#_Toc173441883)

[3.2 Data Exploration 3](#_Toc173441884)

[3.3 Data Preparation 5](#_Toc173441885)

[3.4 Feature Engineering 5](#_Toc173441886)

[3.5 Model Training & performance evaluation 6](#_Toc173441887)

[4. What else we plan to do 6](#_Toc173441888)

[4.1 Continue on feature engineering 6](#_Toc173441889)

[4.2 Deployment of Model 6](#_Toc173441890)

[4.3 Prepare Final Report 6](#_Toc173441891)

[4.4 Prepare Project Presentation 6](#_Toc173441892)

[5. Team’s Involvement 6](#_Toc173441893)

# Project Problem

We choose the following project problem:

4. Cultural/Entertainment

a) The success of the song can often been measured by whether the song is on the Hit Chart such as Billboard Hot 100. It is important for music labels to know what makes a song successful so that they can focus their budget on making songs that has the highest chance of being successful.

# 2. Project Objective

Using the dataset as in <https://www.kaggle.com/datasets/yasserh/song-popularity-dataset>, we would like to use machine learning to train a model that is able to predict the song popularity score (which is range from 0 to 100, with 100 as score of most popular song) base on inputs of some of the parameters as in the dataset.

# 3. What we’ve accomplished

## 3.1 Data Collection

We found 2 dataset that are relevant as follows:

Option 1: https://www.kaggle.com/datasets/yasserh/song-popularity-dataset

* source: <https://www.kaggle.com/datasets/yasserh/song-popularity-dataset>
* Number of sample/records: 13070
* Dataset target: song\_popularity score (range from 0 to 100)
* Features:
  + song\_duration\_ms
  + acousticness,
  + danceability,
  + energy,
  + instrumentalness,
  + key
  + liveness
  + loudness
  + audio\_mode
  + speechiness
  + tempo time\_signature
  + audio\_valence
* This is a regression problem

**Option 2: Popularity of Music Records**

* Source: <https://www.kaggle.com/datasets/econdata/popularity-of-music-records/data>
* Number of sample/records: 7141
* Dataset target: if a song will be top10 of Billboard Hot 100. (1 if it was in the top 10, and 0 if it was not)
* Features:
  + artistname = the name of the artist of the song
  + timesignature and timesignature\_confidence = a variable estimating the time signature of the song, and the confidence in the estimate
  + loudness = a continuous variable indicating the average amplitude of the audio in decibels
  + tempo and tempo\_confidence = a variable indicating the estimated beats per minute of the song, and the confidence in the estimate
  + key and key\_confidence = a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate
  + energy = a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness
  + pitch = a continuous variable that indicates the pitch of the song
  + timbre\_0\_min, timbre\_0\_max, timbre\_1\_min, timbre\_1\_max, . . . , timbre\_11\_min, and timbre\_11\_max = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables)
* This is a binary classification problem.

Both options are supervised machine learning. After discussion, we decided to use Option 1 dataset, because it has bigger dataset (13070 records in option 1 vs 7141 records in option 2).

For option 1 dataset (Song Popularity Dataset), since we are predicting the song\_population score range from 0 to 100, it’s a regression problem.

## 3.2 Data Exploration

We have done the following data exploration:

1. Identify feature that will not contribute to song\_popularity
   * Song\_name will not contribute to song\_popularity and thus is dropped from the dataset.
2. Identify the numeric and categorical features
   * We notice that audio\_mode, time\_signature and key has only 2, 5 and 12 unique values respectively. Hence, we conclude that there 3 features are categorical features
   * The rest of the features are numeric features
3. Data visulisation using histogram

A chart of different types of music

Description automatically generated with medium confidence

* + Base on song\_popularity histogram, the song\_popularity (which is the target of the model that we are going to train) is well spreaded bell curve from 0 to 100. It’s looked like a well balanced dataset.
  + Majority of the song (around 17000+ data) has value 4 in time\_signature features. Time\_signature is most likely not a good feature for prediction of song popularity score. We decided to drop this feature
  + Majority of the song has very low instrumentalness, we suspect that this feature may not be useful as well.

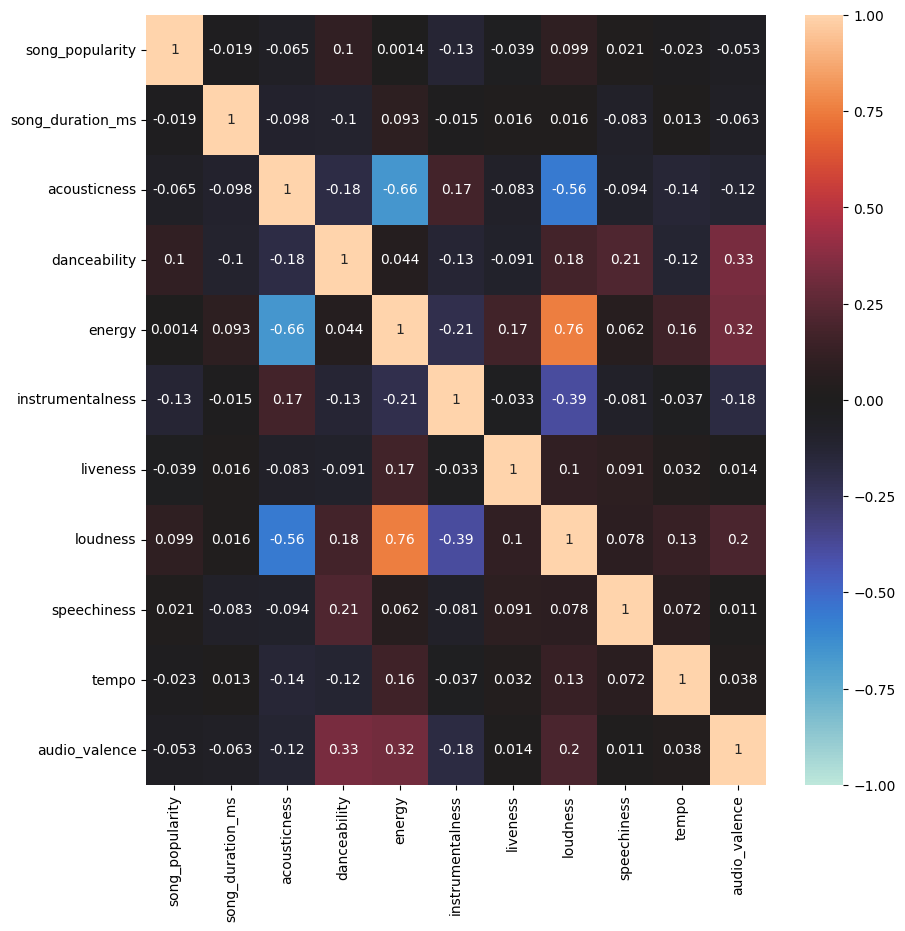
1. Data visulisation using boxplots

A group of graphs with different colored lines

Description automatically generated with medium confidence

* From boxplots above, we observe that song\_duration\_ms, danceability, energy, liveness, loudness and speechiness and tempo seem to have outliers. We shall remove some outliers.
* Intrumentalness feature can’t be ploted to boxplot. We should drop this feature as it don’t contribute to prediction of song\_popularity score.

1. Data visulisation using correlation matrix



Obeservation:

* Top 4 numeric features that have higher correlation to song\_popularity are: danceability, loudness, speechiness, energy
* there are multi-colinearity between the following features:
  + loudness and energy (0.76)

## 3.3 Data Preparation

The following had been done before performing feature engineering:

1. Removing any duplicate records if any
   * There are 3911 duplicate in the dataset.
   * After removing the duplicates, the number of records in the dataframe had been reduced from 18835 to 14924
2. Convert categorical features to numeric using dummy encoding
   * Dummy encoding is used to convert categorical features (audio\_mode, time\_signature and key) to numeric data.
3. Feature scaling
   * MinMaxScaler is used to scale the data into range of 0 to 1

## 3.4 Feature Engineering

## 3.5 Model Training & performance evaluation

# 4. What else we plan to do

1. Continue on feature engineering
2. Deployment of Model
3. Prepare Final Report
4. Prepare Project Presentation

# 5. Team’s Involvement