Lye Suh Jeng (7487427Y)

Abstract

This is the project proposal for Cultural / Entertainment problem

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

Proposal

Table of Contents

[1. Project 2](#_Toc171853401)

[2. Formulation 2](#_Toc171853402)

[2.1 Alternative 2](#_Toc171853403)

[2.2 Heuristic 2](#_Toc171853404)

[2.3 Success Metrics 2](#_Toc171853405)

[2 Dataset 3](#_Toc171853406)

[2.1 Methods 4](#_Toc171853407)

[3.1 Deployment 4](#_Toc171853408)

# 1. Project

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.

# Formulation

We would like to create a machine learning solutions that can predict the song’s popularity so that music labels know what makes a song successful. The benefits of the solutions are:

* Cost effectiveness: music labels can focus their budget on market songs that has higest chance of beining successful.
* Increase time to market: music labels can reduce time in determining which song to be market.

if we can gather list of the popular songs that was listed in Billboard Hot 100 and the song’s features such as energy, acoustics, instumentalness, liveness, dancibility and etc, we can use machine learning to create a model that can predict song’s popularity.

## 2.1 Alternative

Without this prediction, traditionally music labels need to hire people who are in touch with youth culture to put their ears to the proverbial ground and listen for what’s new that the young people are listening to. The labels then try and hire the people who are making the song.

## 2.2 Heuristic

## 2.3 Success Metrics

# Dataset

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.

## 2.1 Methods

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.

# Deployment

How do you intend to use the model? Provide an overview of how the model is used in the overall solution. You can illustrate this with a system architecture diagram.