Song Popularity Analysis and Prediction

Group E

Members – Claire Nguyen, Gordon Chong Ren Ong, Nikheil Malakar, Tracey Truong

# Summary

This project will Project will be providing deep analysis and prediction of the popularity of the song. Using more than 200,000 song’s datasets, we will be analysing what makes the song so popular. The dataset includes 15 different columns and category.

# Description

The project “Song popularity analysis and prediction” will go through analysis of songs in depth, providing an answer to why song and which songs will be popular. By using a huge dataset, answers to such questions should be relevant via this project. The dataset is available on Kaggle and is about 17MB in total. In total there is 131k songs for the year 2019 and 116k songs for 2018 with more than 15 features and description of the song.

The attributes or features that will be considered are acousticness, danceability, duration, energy, instrumentals, key, liveness, loudness, etc. We will be closely investigating the key features which will help make a popular song. We will be training two types of dataset where the name of the song is included and another without the name to investigate if the name of the song affects the popularity.

Data set: <https://developer.spotify.com/documentation/web-api/>

## Goals

The goal of this project is to:

1. Analyse the relationship between the popularity of the song and other factors
2. Building clean data
3. Training multiple models
4. Reach high accuracy with both training and validation data

# Dataset Summary

The SpotifyAudioFeature.csv file has 200k+ songs with features listed which has both number and word data. The dataset contains a lot of string specially the name of the song. We will be making a dictionary of word and giving each name or character an ID and then training them. The model will be fed the following data that will be IDs or numbers:

|  |  |
| --- | --- |
| Artist Name | The person/s singing the song. |
| Track ID | Unique ID that is used to track the song for Spotify |
| Track Name | Name of Song |
| Acousticness | How prominent acoustic sounds (values between 0 to 1) |
| DanceAbility | How much can you dance with the song (values between 0 to 1) |
| Duration\_ms | Duration of song in milli seconds |
| Energy | Energy of the song (values between 0 to 1) |
| Instrumentalness | How much instruments are used (values between 0 to 1) |
| Key | The key of the song (values between 0 to 11) |
| Liveness | How live the song is (values between 0 to 1) |
| Loudness | Loudness of the song (values between 0 to 1) |
| Mode | Mode of the song (0 or 1) |
| Speechiness | How likely the song is an audio is speech |
| Tempo | Tempo of the song (values between 0 and 250) |
| Time Signature | How many beats are contained in each measure (values between 0 and 5) |
| Valence | The musical positiveness (values between 0 and 1) |
| Popularity | Popularity of the song (0 to 100) |

# Project Plan

## Techniques

* Data Visualisation using
* Regression training
* RFE Analysis
* K-means
* K nearest neighbour

## Milestones

Establish GitHub: Establish the project which includes version control(GitHub) and project management a

Data Handling: Clean data to suit our research needs removing all unessential data.

Data Modelling: We need to sort 200k rows and ID the names of the songs.

Overfitting: Memorizing training data and not being functional to unseen data.

Regression: analyse data collected and see if there is a linear relationship between song duration and popularity loudness and tempo. compare all the different variables and see which has pair has the best R2 value.

Implementation of K mean and K nearest neighbour.