Intermediate Project Report : Song Predictive Analysis

Team Members: Adarsh Agrawal (A20517609)
Abhishek Sharma (A20506258)
Dheeraj Goud (A20500290)

Abstract:

Music preferences have evolved over time. With easier than ever access to new music, music platforms are using machine learning to identify particular trends and traits in order to outperform rivals and draw in more users. In this project, we want to delve into the field and see for ourselves how music platforms classify music into micro-genres, how they use the data they gather to provide better music recommendations to their users, and much more. One inspiring example is the way the Spotify mobile app has divided its music library into thousands of subgenres.

Problem Statement:

In our project, we are attempting to provide answers to the queries below. Additionally, if it's feasible, we might be able to glean some additional insights from the data. However, we have referred to various statistical sources. With the dataset that does contain the features we need to forecast, we are considering addressing these queries.
1. Determine the song's year based on various details like the record cover, etc.
2. Can we determine a song's popularity based on the characteristics it has?
3. Given a music, what other selections would be appropriate for the same user?

Methodology:

We imported data from 2 datasets to address each question as part of the strategy used to address our problem statement. We had to use various datasets to address different research questions because each one needed features or information that could not be found in a single dataset. After the data was imported, it was cleaned before being used for analysis and visualization. As we worked on the study question, we were also able to find answers to other questions in addition to the original one.

Depending on the need, the dataset was split into training and testing sets as needed, and analysis was then performed using regression, classification, or clustering. For example, to determine the year of the song based on various characteristics, classification was used.

Data Sources:

Dataset 1:
The dataset is from Kaggle (link). This dataset has audio features of over 500,000 songs. This is an open source and can be downloaded as a CSV file from Kaggle.

Metadata: This dataset has two subsets of data, artists & tracks. Here are the details for each dataset:

Tracks: It consists of 20 columns and has 586,672 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	Type	Brief Description
ID	String	A unique identifier for the song
name	String	The name of the song
popularity	Numeric	Defines the popularity of the song. The value is between 0 to 100
duration_ms	Numeric	Defines the duration of the song in milliseconds
artists	String	The name of the artist
id_artists	String	A unique identifier for the artist
danceability	Numeric	Defines the danceability of the song. The value is between 0 to 1
energy	Numeric	Defines the energy of the song. The value is between 0 to 1
loudness	Numeric	Defines the loudness of the song. The value is between -60 to 6
speechiness	Numeric	Defines the speechiness of the song. The value is between 0 to 1
acousticness	Numeric	Defines the energy of the song. The value is between 0 to 1
liveness	Numeric	Defines the liveness of the song. The value is between 0 to 1
tempo	Numeric	Defines the tempo of the song. The value is between 0 to 250

Artists: It consists of 5 columns and has 1,104,349 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	Type	Brief Description
ID	String	A unique identifier for the artist. This column can be joined with the id_artist from the Tracks dataset
followers	Numeric	The number of followers the artist has
name	String	The name of the artist

Dataset 2:

Core information: The dataset is from Kaggle (link). This dataset again has audio features of different songs. This is an open source and can be downloaded as a CSV file from Kaggle.

Metadata: This dataset has one file that is genre_music.csv. Here are the details for the dataset

genre_music.csv.: It consists of 20 columns and has 41,099 rows. Each column defines a specific feature of the song. Some of the relevant columns for our purpose are as follows:

Field Name	**Type**	**Brief Description**
ID	String	A unique identifier for the song
genre	String	Defines the genre of the song
track	String	Defines the name of the song
artist	String	Name of the Artist
danceability	Numeric	Defines the danceability of the song. The value is between 0 to 1
energy	Numeric	Defines the energy of the song. The value is between 0 to 1
key	numeric	Defines the key of the song, value is between 0 to 11
loudness	Numeric	Defines the loudness of the song. The value is between -35 to 4
mode	numeric	Define the mode of the song, values lies between 0 and 1
speechiness	Numeric	Defines the speechiness of the song. The value is between 0 to 1
acousticness	Numeric	Defines the acoustic-ness of the song. The value is between 0 to 1
liveness	Numeric	Defines the liveness of the song. The value is between 0 to 1
valence	numeric	Defines the valence of the song,value is inbetween 0 to 1
tempo	Numeric	Defines the tempo of the song. The value is between 55 to 220
duration_s	numeric	Defines the length of the song in millisecond
time_signature	numeric	Defines time signature of the songs, value between o to 5
chorus_hit	numeric	Defines the chorus hit of the song, value between 0 to 433
sections	numeric	Defines the sections of the song, value between 0 to 169
popularity	numeric	Defines the popularity of the song, value is 0 or 1
decade	string	Defines the decade in which song release

Data Preprocessing:

All the three datasets are CSV files and the size of all the files are approximately 180 MB combined. This dataset consists of string and numeric values. Some of the features in the dataset will not be relevant to our analyses and will be removed from the dataset. Before carrying out the analysis we:

- Made sure that we did not have any duplicate rows.
- Removed any outliers.
- For any missing value, we were not considering that feature for that row in the analysis.

Data Preprocessing has been completed.

Implementation:

We used linear regression on the processed data to predict the year the song was released based on different factors. As of now it’s predicting decades, not exactly the year. But if we have more data with specific years, and some more relevant columns like genre, artist names, artist basic information, that will be significant features to predict a year.

After cleaning the data and did a 0.3 train-test split and fed the data to a linear regression model. After training, we get an accuracy of 70%.

Future Work:

Next we will try to improve the accuracy of the model by using the Random Forest algorithm. And then work towards the next 2 queries.