Rachael Enders
Faith Mathew
Cameron Schneider
Shannon Wasson
CIS 635
May 17, 2024

Project Proposal

Overview

In the rapidly evolving realm of digital entertainment, online music streaming platforms like Spotify play a pivotal role. Spotify, with its vast user base, allows users to discover and enjoy music effortlessly. Our project delves into this dynamic landscape by analyzing Spotify tracks using a dataset that encompasses diverse genres and their corresponding audio features. Our approach involves using various data mining techniques like clustering to group Spotify tracks based on different features like tempo, danceability, loudness, mode to obtain the accuracy and personalization of music recommendations. The motivation emanates from the growing importance of personalized recommendations in the music streaming industry, aiming to enhance user experience and engagement on the Spotify platform. The goal of this project is to apply all the contents from this course to analyze Spotify tracks and extract relevant features and also evaluate the effectiveness of our clustering approach in improving music recommendations.

Related Work

Spotify is one of the most popular music streaming applications available and its algorithm is quite well known. As a result, quite a few projects have been done on this topic. Since there are only so many changes that can be made to an algorithm, they all do hold some similarities to our project at their core. However, there are differences in the goals and manipulation of the data that make the projects different. A few examples of these include:

1. This study aimed to see how changing Spotify's interface and user interaction impacts the effectiveness of the music recommendation algorithm. They built two different interfaces, one that allowed the user to slide on how much they liked certain attributes of music (for example:

danceability slid to 80) and another that was shaped like a radar that allowed the user to select how much they liked certain attributes of music on a radar graph. This is very different from our project because they are testing the algorithm by changing the user interface whereas we are just building our own algorithm and comparing it to Spotify's own. [1]

2. This study used data mining techniques similar to ours (t-tests) but the data gathering and goals of the study were different. The authors of this project sent out a survey and gathered songs from the survey taker's public Spotify playlists. From these playlists, they used similar data mining techniques as us but used them to try to predict personality traits, demographic characteristics, and random habits (for example: smoker versus nonsmoker). [2]

These are just a few examples of similar projects that use Spotify's algorithm as the basis of their projects. Overall, ours is different because most are trying to determine how someone's playlist is an indicator of who they are as a person. Instead, we are working to build and potentially better an existing algorithm.

Data Plan : Spotify Personalized Music Recommendation System

Primary Data Source: Spotify Tracks Dataset from Kaggle (pre-existing)

Contains extensive information on tracks including track IDs, artists, genres, and various audio features.

Acousticness	**Instrumentalness**	**Key**	**Liveness**
Valence	**Loudness**	**Tempo**	**Time Signature**
Danceability	**Mode**	**Speechiness**	**Energy**

Secondary Data Source:

Spotify Web API

Provides real-time access to Spotify's music database.

Can be used to get up-to-date track information, user listening history, and playlists.

Data Gathered By:

The Kaggle dataset was gathered by a third-party user, aiming to provide a comprehensive dataset for music analysis and machine learning projects.

The Spotify Web API data is provided by Spotify itself, primarily for developers to build music-related applications and services.

Data Preprocessing:

Data Cleaning: remove duplicates, ensure each track is unique, identify missing values, remove records with significant data gaps.

Feature Engineering: normalization (numerical feature e.g. danceability, energy) to create uniformity. Encoding (convert genres to a numerical format). Composite features (combine existing features like tempo, energy, danceability into one feature like "rhythm").

Data Transformation: aggregation (avg. features per artist / album).

Data Split: divide the dataset into training, validation, and test sets to get a proper evaluation of the recommendation system.

***User Profiles*:** Create user profiles by analyzing history and preference. Create collaborative filtering techniques to find similar users and improve the recommendation accuracy.

Project Goals and Evaluation:

***Objective*:** Develop a personalized music recommendation system leveraging Spotify's dataset.

Problem Solving: Compare the effectiveness of our recommendation system with Spotify's existing recommendation engine.

Study and Analysis:

Analyze the dataset to understand the key factors influencing music preferences. Experiment with different recommendation algorithms.

Implementation Plan

For our data mining pipeline, we will first follow through with the data plan. First, we will do data preprocessing. During the data preprocessing, the Kaggle data set will have duplicates removed and any missing or incorrect values will be removed. This will allow for the data to be accurate so the next steps can be carried out correctly. From there, we will do the feature engineering, data transformation, and data split.

Once all data preprocessing has been completed and analysis can be done, we can now look at Spotify's API and do A/B testing. We would do A/B testing by completing a t-test and using the p-value to determine if the test was successful.

We plan to run tests on the data set using Python as the primary software language and numpy, pandas, scikit-learn, and matplotlib as the libraries. Pandas will be used to build a DataFrame to analyze and edit the data. Numpy will be used for mathematical analysis. Scikit-learn will be used for A/B testing and other statistical tests that may arise. Matplotlib will be used to plot or graph any data for easier visualization if needed.

Evaluation Plan

Use metrics such as precision, recall, and F1-score to evaluate the recommendation system. Conduct A/B testing to compare user satisfaction between our system and Spotify's native recommendations.

The goal is to understand the dynamics of music recommendation systems and potentially improve upon existing models with new data mining and machine learning techniques.

Plan for Group Collaboration

The plan for collaborating as a group is asynchronously using discord to share updates, feedback and ask questions as they arise. We are using Google Docs to collectively add our contributions to the project proposal.

To collaboratively implement the data mining pipeline we will split up the steps data preprocessing, feature engineering, data transformation, EDA, model development, a/b testing.

Timeline

Week 1	Group Formation and brainstorming on a possible topic.
Week 2	Meeting, discussing, and documenting the group proposal. Submission 5/17/2024
Week 3	Analyzing the dataset Working on the progress report.
Week 4	Compile the progress report. Submission 5/31/2024
Week 5	Work on the Final Report
Week 6	Compile our final report. Submission 14/6/2024.

References

[1] Millecamp, M., Htun, N., Jin, Y., & Verbert, K. (2018). Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces. In Proceedings of the 26th Conference on User Modeling, Adaption and Personalization, *(UMAP '18)* (101-109). Association of Computing Machinery. https://doi.org/10.1145/3209219.320922

[2] Tricomi, P. P., Pajola, L., Pasa, L., & Conti, M. (2024). "All of Me": Mining Users' Attributes from their Public Spotify Playlists. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)* (634-636). Association for Computing Machinery. https://doi.org/10.1145/3589335.3651459 1