Capstone Project: Music Recommendation System

Problem Definition

Nowadays, the most common way to listen to music is through streaming platforms like Spotify, Soundcloud, Apple Music, Youtube Music, etc. Online music streaming becomes more dominant for people because it gives them freedom to listen to their favorite songs and artists. While there are a lot of online music streaming platforms they are also competing with each other for users. The problem is from time to time people become more busy and searching for good content becomes a distraction. The intended goal here is to figure out how to improve user experience. So the critical challenge here is to figure out ways for users to spend less time searching for good content. The key objective here is to build a good recommendation system to propose the top 10 songs based on the user's preferences and based on search words. This can be done by collecting data from the user's history of songs played and building a This is an approach where the machine learning algorithm collects data based on users' behaviors, activities, and preferences.

Data Exploration:

Data Background:

The Million Songs dataset is a collection of audio features and metadata for contemporary popular music tracks. It started as a collaborative project between The Echo Nest and LabROSA. The Million Song Dataset was created under a grant from the National Science Foundation, project IIS-0713334. The original data was contributed by The Echo Nest, as part of an NSF-sponsored GOALI collaboration. Subsequent donations from SecondHandSongs.com,

musiXmatch.com, and last.fm. The dataset contains two files: count_data and song_data. The count data contains a user_id, song_id and play_count. While the song_data contains a song_id, title, release, artist_name and year.

Observations and Insights:

After exploring the dataset, The song_data has 15 missing values in title and 5 missing values in release(album) replacing it with unknown to clean the dataset. Merging both files to a single dataset results in a user-item interaction data that is very important in the training the model. Creating a new column combining the title of the song and artist_name will be useful for the user-item interaction data.

Proposed Approach:

Since the meta-data of the songs in the dataset are limited like title, release(album), artist_name, and year. A collaborative recommendation machine learning algorithm will be a good solution for the problem. The other approach is content-based recommendation system but the dataset is lacking on song meta-data such as genre of the song or the artist. However, it is not necessary to collect the information for a collaborative filtering approach, once you have user-item(song) interactive histories data, you are ready to build the recommendation model. User-item(song) interaction data can ensure the greatest extent possible to reflect users' preference so then the recommendation based on user-item interaction data can accurately meet users' taste in songs. Using user-item interaction data to train the model is great because it is personalized and useful for a distinct recommender system.

Solution Design:

The solution design is to build a recommendation system based on the Co-occurrence matrix. The goal of the co-occurrence recommendation machine learning algorithm is finding how many times two or more songs appeared together in the user's historical data and calculate the similarity of the songs in the user historical data to all unique songs in the dataset. The algorithm will determine the recommended songs for a certain user by rank and score depends on the user historical data.