Spotify Executive Summary

Group 4: Cagan Abney, Jason Cisneros, Joshua Hale and Raheem Yusuff

Introduction

The inspiration of this project was to take a personal look at how the recommender systems in our everyday lives work. When you are watching Netflix, how do they decide what we should watch next? When you buy something online, how do they decide what we also might like?As we all love listening to music and are always looking for that new song we can't stop playing or get out of our head, how can we make finding this song easier? During this project we were able to see how machine learning answers some of these questions we have.

To complete this project we chose a dataset that consisted of Spotify tracks and their respective features. We got this dataset from Kaggle (link found in Works Cited section). This data was originally collected from the Spotify API. Our dataset consisted of the Spotify unique track identifier, the track name, artists, the album name the song belongs to, the song's genre, and how popular each song is. In addition to the details of each track, the dataset also consisted of features for each song. These features included track duration, explicitness, danceability, energy, key, loudness, "speechiness", acousticness, and more.

In order to create insightful Tableau dashboards we also decided to use another Spotify track playlist consisting of the most streamed songs in 2023. This dataset was also from Kaggle. Similarly, this dataset consisted of the song details and features such as track duration, explicitness, danceability, energy, key, loudness, "speechiness", acousticness, and more. In addition, this dataset also consisted of how many Spotify user playlists the song was in, but also other users of other music streaming services such as Apple Music and Deezer. With these two datasets we selected the features of the songs and taught the model.

Machine Learning

For our recommender system, we decided to use the K-Nearest Neighbors (KNN) model. The KNN model is a supervised classification machine learning model which uses proximity to make classifications or predictions about the grouping of an individual data point. The model takes the features inputted and calculates the distance between the query point and the other data points. This distance can be calculated using several distance metrics such as Euclidean, Manhattan, Minkowski, or Hamming. We used the Euclidean Distance and were able to get "okay" song recommendations. Euclidean distance measures a straight line between the query point and the other point being measured. This is usually the most commonly used distance metric used because of its simplicity, however, with simplicity comes less predictability. The K value chosen by us determines how many nearest neighbors the model will look for.

Unlike Euclidean Distance which measures the straight-line distance between two points in Euclidean space. Cosine similarity is a metric we looked into that is used to measure how similar two vectors are in a multi-dimensional space. It calculates the cosine of the angle between two non-zero vectors and provides a measure of the similarities between them.

Tableau

For this project we created 5 tableau worksheets and combined them into 2 dashboards. The first dashboard, called “Playlist” contains visualizations that show what Artists are found in the most Spotify user playlists and what songs of the artists are found the most. In addition, you can see the comparison of the top artists in Spotify vs Apple. From the visualizations you can see it is similar for the two platforms as Taylor Swift and Bad Bunny are the top 2 for both. Lastly, on this dashboard you will find a bar chart of songs’ danceability combined with a line chart of song’s energy throughout the 2000s.

The next dashboard was the “Genre Dashboard". On this dashboard we have 2 visualizations. One being a bubble cart of the most popular genres. Based on the size of the bubbles and the color key, it can be concluded that k-pop, chill and anime are the most popular genres. Once again, this dataset is mostly 2023 released songs, so this is what’s popular right now. On the second visualization, you can see the genre's “Instrumentalness” vs their “Acousticness”.

Conclusions

In conclusion, recommender systems in our everyday life likely use a supervised classification machine learning from features of the product the algorithm is trying to recommend. For this project we used k-Nearest Neighbors, but it would be interesting to see what companies like Spotify, Apple Music, Netflix, and Amazon recommend songs, movies, and products based on users past experiences. There are endless amounts of opportunities to optimize machine learning models to achieve your expected outcome. This is the fun of machine learning because there is no limit for what you can achieve!

Another limitation was the dataset, the data set of course could not obtain every song in Spotify due to size limitations, so the user can not just input any song. There is a bias in our dataset. As this is a sample of songs on Spotify there is a bias to the songs that are available. There may be some older or less known songs that did not get included.