Spotify Analysis: Grouping songs to build custom playlists

Mikhel Semple | Pace University Seidenberg School of CSIS | Mentor: Dr. Christelle Scharff

GitHub: **https://github.com/mnsemple83/Capstone-Project**

Abstract

Spotify is one of the most popular online platforms for streaming music. Based on the music streamed by its listeners, it can create custom playlists with a theme. Therefore, Spotify needs to be able to analyze the metadata collected from the songs in its library to build these playlists. In this study, K-clustering algorithms will be used to cluster songs based on their shared data qualities. Then, a Logistic Regression model will learn from these clusters to add new songs to these playlists. This methodology will contribute to a better experience for Spotify's users.

Research Question

How does Spotify use song data to build custom playlists for its users?

Related Work

Many studies have applied several clustering algorithms to group songs based on their feature data [3,4,5,6,8]; Techniques for feature extraction include (XGBoost) [1], digital signal processing and autoencoding [2], and Pitch Class Profile (PCP) [6]; [3,5] used the Silhouette Score and Davies-Bouldin Index metrics to evaluate the performance of clustering algorithms; [6] used Histogram clustering to measure accuracy; [3] applied t-SNE and PCA to reduce the dimensionality of the dataset; K-Means [3,5], or K-Means++ [6], performed the best when it came to building clusters of songs that share similar characteristics; [4] used the clusters to decide the best playlist to access based on the user's emotional state and the music that they were listening to.

Dataset

The dataset used for this study originates from Spotify and is a collection of 114,000 songs with 20 variables. Each song is classified by one of 114 genres and includes both musical and non-musical data.

Variables (used variables underlined)

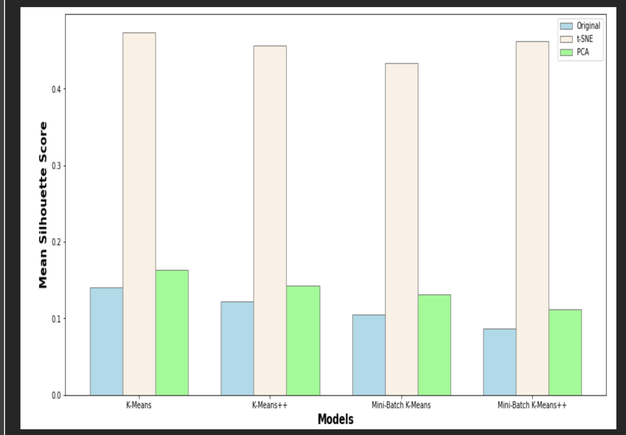
Track ID, Artists, Album Name, Track Name, Popularity, Duration (milliseconds, but converted to minutes), Explicit, Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Time Signature, and Track Genre.

Methodology

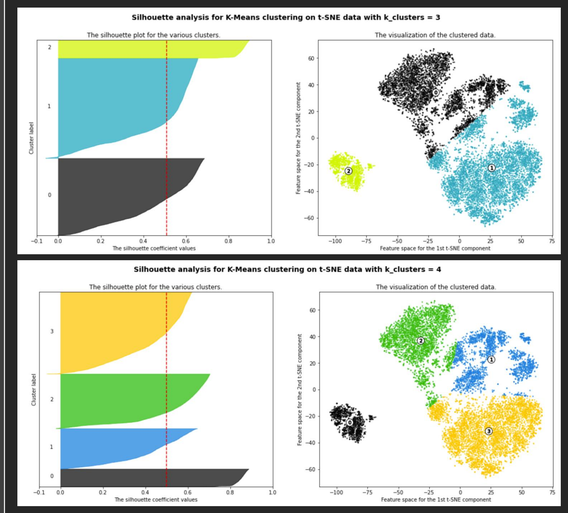
- Sampled 10% of the original dataset.
- Feature selection and standardization was performed.
- Perform dimensionality reduction using t-SNE and PCA.
- K-clustering algorithms used to group songs.
- Best performing algorithm used for cluster analysis.
- Logistic Regression model used to cluster new songs.
- Results of the Logistic Regression model evaluated using Confusion Matrix and Classification Reports.

Results

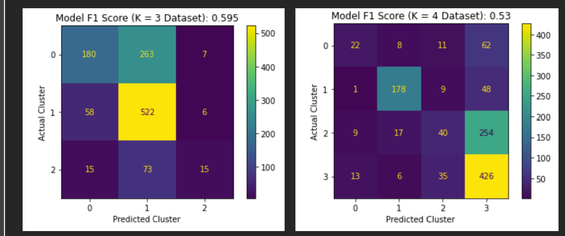
Each of the algorithms performed the best with the t-SNE dataset, and the K-means algorithm had the overall best performance. This is visualized by the following chart that measures the average silhouette scores:



Using the K-Means algorithm, clusters were formed using a range of k values where k represents the number of clusters to be formed from the data. According to the silhouette plots and cluster diagrams, k values of 3 and 4 provided the best clusters.



With each song assigned to a cluster, the datasets for each k value were fitted to the Logistic Regression model for training and testing. The testing set represents the new songs to be added. The results of the Logistic Regression model were evaluated using a Confusion Matrix.



Based on the overall F1 score, the Logistic Regression model performed best with the K = 3 dataset. The performance score is a measure of the model's ability to determine the best playlist for newer songs, which is helpful when making recommendations.

Conclusion & Future Work

K-clustering algorithms have provided a method for Spotify (and any other streaming services) to build custom playlists based on the metadata collected from the songs in its library. It is also possible to add songs to these playlists using a Logistic Regression model that is capable of identifying the best playlist for these songs. However, the performance scores of these models do indicate that there is room for improvement. The quality of these clusters could have been improved if a larger subset of data were used. Also, experimentation was limited by system resources. A more powerful system could have allowed for more extensive tuning of model parameters. Finally, the difficulty of identifying the best clusters could have been influenced by the outliers of the dataset, which were actually natural variations of the songs, and not the result of erroneous data. All this can be considered for future exploration on this subject.

References

[1] H. Tian, H. Cai, J. Wen, S. Li and Y. Li. 2019. A Music Recommendation System Based on logistic regression and eXtreme Gradient Boosting. Retrieved October 8, 2023 from https://ieeexplore.ieee.org/document/8852094

[2] Y. Atahan, A. Elbir, A. Enes Keskin, O. Kiraz, B. Kirval and N. Aydin. 2021. Music Genre Classification Using Acoustic Features and Autoencoders. Retrieved October 8, 2023 from https://ieeexplore.ieee.org/document/9598979

[3] P. N, D. Khanwelkar, H. More, N. Soni, J. Rajani and C. Vaswani. 2022. Analysis of Clustering Algorithms for Music Recommendation. Retrieved October 8, 2023 from https://ieeexplore.ieee.org/document/9824160

[4] M. Bakhshizadeh, A. Moeini, M. Latifi and M. T. Mahmoudi. 2019. Automated Mood Based Music Playlist Generation By Clustering The Audio Features. Retrieved October 9, 2023 from https://ieeexplore.ieee.org/document/8965190

[5] H. Wijaya and R. S. Oetama. 2021. Song Similarity Analysis With Clustering Method On Korean Pop Song. Retrieved October 10, 2023 from https://ieeexplore.ieee.org/document/9617204

[6] R. Sun, J. Zhang, W. Jiang and Y. Hu. 2018. Segmentation of Pop Music Based on Histogram Clustering. Retrieved October 11, 2023 from https://ieeexplore.ieee.org/document/8633060

[7] H. Han, X. Luo, T. Yang and Y. Shi. 2018. Music Recommendation Based on Feature Similarity. Retrieved October 11, 2023 from https://ieeexplore.ieee.org/document/8690510

[8] Michelangelo Harris, Brian Liu, Cean Park, Ravi Ramireddy, Gloria Ren, Max Ren, Shangdi Yu, Andrew Daw, and Jamol Pender. 2019. Analyzing the Spotify Top 200 Through a Point Process Lens. Retrieved October 11, 2023 from https://doi.org/10.48550/arXiv.1910.01445