Project Title: Song/Artist Popularity Analysis

Team Members: Neeraja Kirtane (kirtane3) , Sanil Chawla (schawla7), Omkar Gurjar (ogurjar2), Atharva Naik(annaik2)

Introduction: We aim to study the factors affecting the popularity of a song or artist. Specifically, we aim to develop various supervised ML and data mining models to predict song popularity and identify the properties of the song which determine the popularity.

Motivation: Music has been the source of entertainment, expression and emotion for as long as we can remember. Even today, streaming platforms such as Spotify have over 551 million users[1] making it a goldmine of data related to users' music preferences. Machine learning algorithms coupled with the right data processing and feature engineering can help surface key patterns related to users' audio consumption. The project's potential applications can be extended to various music streaming platforms, where it can help make personalized recommendations and playlist curation. Additionally, it offers a remarkable opportunity to contribute to the ongoing research in both the data science and music communities.

Data: Our data consists of two parts: information about the song and information about the artist. For every song, we have the following attributes: popularity, duration of music, if the song is explicit or not, the artist name, release-date, danceability, energy, key, loudness, mode, spechiness, acousticness, instrumentalness, liveness, valence, tempo, time-signature. For the artist we have the following information: Number of followers, genres and popularity. Our dataset contains information for ∼600K songs and ∼1.1M artists.

Plan of Work: We aim to build data mining models to understand the factors that drive popularity of a song given its many features and attributes. To accomplish this, we divide our plan of work in three broad steps: **(1) Understanding the Dataset:** Given the extensiveness of the dataset, our first goal is to understand its various attributes. We aim to perform a comprehensive EDA including finding inter-feature correlations (to remove redundant columns), calculating central tendancies and dispersion. Next, **(2) Predictive Modelling:** We aim to develop a supervised regression model to predict the popularity of a song. Using the insights from EDA, we will select and engineer fine-grained features for our task. Further, we look to experiment with multiple feature combinations and a variety of models including OLS regression, tree and gradient boosting models (RandomForest, LightGBM), and advanced models such as neural networks. We plan to use standard regression evaluation metrics like MAE and MSE. Finally, **(3) Interpreting and Visualizing Results:** To fully understand the factors leading to song popularity we would study the feature importance of the trained models using methods like SHAP. Moreover, we look to visualize various trends in our dataset at song, artist and genre level to highlight important patterns.

Related Work: Hit Song Science (HSS)[2] is the study of predicting whether a song would get "hit" or popular on music platforms. Works such as [4][5][3] have tried to use song features derived using Spotify API to predict billboard song success and number of streams by training machine learning models. [6] aims to propose statistical tools for identifying features deterministic of a song's success. From a psychology point of view several works such as [2][7][1] have tried to explore the complexities of music preferences based on peoples' personality, mental state and other factors.

Anticipated Challenges: We need to ensure that our model doesn't overfit or underfit. Also we need to add interpretibility to our regression models to better understand our data.

[1] https://newsroom.spotify.com/company-info/
[2] https://en.wikipedia.org/wiki/Hit_Song_Science

References

[1] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. Personality traits and music genres: What do people prefer to listen to? In *Proceedings of the 25th conference on user modeling, adaptation and personalization*, pages 285–288, 2017.

[2] Alinka E Greasley and Alexandra M Lamont. Music preference in adulthood: Why do we like the music we do. In *Proceedings of the 9th international conference on music perception and cognition*, pages 960–966. University of Bologna Bologna, Italy, 2006.

[3] Joshua S Gulmatico, Julie Ann B Susa, Mon Arjay F Malbog, Aimee Acoba, Marte D Nipas, and Jennalyn N Mindoro. Spotipred: A machine learning approach prediction of spotify music popularity by audio features. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 1–5. IEEE, 2022.

[4] Kai Middlebrook and Kian Sheik. Song hit prediction: Predicting billboard hits using spotify data. *arXiv preprint arXiv:1908.08609*, 2019.

[5] Rutger Nijkamp. Prediction of product success: explaining song popularity by audio features from spotify data. B.S. thesis, University of Twente, 2018.

[6] Mariangela Sciandra and Irene Carola Spera. A model-based approach to spotify data analysis: a beta glmm. *Journal of Applied Statistics*, 49(1):214–229, 2022.

[7] Sunkyung Yoon, Edelyn Verona, Robert Schlauch, Sandra Schneider, and Jonathan Rottenberg. Why do depressed people prefer sad music? *Emotion*, 20(4):613, 2020.