Project Title: Song/Artist Popularity Analysis

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

1. Abstract

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

2. Introduction

The project aims to perform a comprehensive study of understanding the popularity of songs on online platforms using data driven methods. Specifically, we use a collection of 5.8M songs by 1.1M artists available on a popular streaming platform- Spotify. For every song, we are provided its popularity score along with a set of features determining its musical properties, its genre and the artists. As a part of this report, we perform an EDA of the data and visualize its various aspects to thoroughly understand its strengths, weaknesses and perform appropriate data processing steps. Next, we model a regression task of using song features to predict its popularity, for which we experiment with a variety of models. The code is available at our GitHub repository [1]. As future work, we aim to improve on the robustness of our regression results and generate explainable and qualitative results to identify the factors which make songs popular.

3. 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[2] 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.

4. Related work

Hit Song Science (HSS)[3] 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. On online ML comptetitions forums such as Kaggle [4], several users have independently compiled similar datasets for different time-frames and artists [5][6]. These datasets have been leveraged for other tasks including recommendation[7][8] and classification[9][10].

[1] https://github.com/atharvanaik10/song-popularity-analysis

[2] https://newsroom.spotify.com/company-info/

[3] https://en.wikipedia.org/wiki/Hit_Song_Science

[4] https://www.kaggle.com

[5] https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

[6] https://www.kaggle.com/datasets/jarredpriester/taylor-swift-spotify-dataset

[7] https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset

[8] https://www.kaggle.com/code/ibtesama/getting-started-with-a-movie-recommendation-system

[9] https://www.kaggle.com/competitions/machine-learning-pw4-part2-classification-evd1

[10] https://www.kaggle.com/competitions/machine-learning-pw3-evd2

5. Methodology

We divide our current work into the following three broad steps:
i.Understanding the Dataset: Our dataset contains two sub-datasets (1) Songs data containing some features such as acousticness, danceability and liveliness for 586672 songs along with their popularity which is the target variable. (2) Artist data containing information related to 1162095 artists included in the songs data. As a part of EDA, we compute the descriptive statistics(range, mean, median) for different features, investigate data-quality issues(missing values, any skews in data) and calculate inter-feature correlations. We also visualize the EDA results by plotting feature trends over the years and the Pearson's correlation matrix as a heatmap. Full details of the dataset(description of columns, quality issues etc) are included in the Appendix.
ii. Data Processing: We model the task of predicting the song popularity as a regression task. For data pre-processing, we (1) Merge the song and artist datasets to incorporate artist related features for songs (2) Min-Max[11] normalize the features to the range (0, 1) to eliminate issues due to large ranges (e.g. number of followers). We also experiment with other normalization techniques such as z-score but recieved inferior results. We also scale our target variable popularity to be between 0 and 1 using Min-Max scaling to minimize its range of values from 0 to 99 to 0 to 1.
iii. Predictive Modelling: Next, we perform preliminary modelling and test the sanity of our dataset by training a suite of supervised and unsupervised regression algorithms including linear/lasso regression, elastic net, decision tree, various gradient boosting algorithms and K Nearest Neighbor. We divide our dataset into train and test with an 70:30 ratio and set a particular seed value. Evaluation metrics Mean Squared Error (MSE), Mean Absolute Error(MAE) and coefficient of determination (R^2) are used to compare the performance.

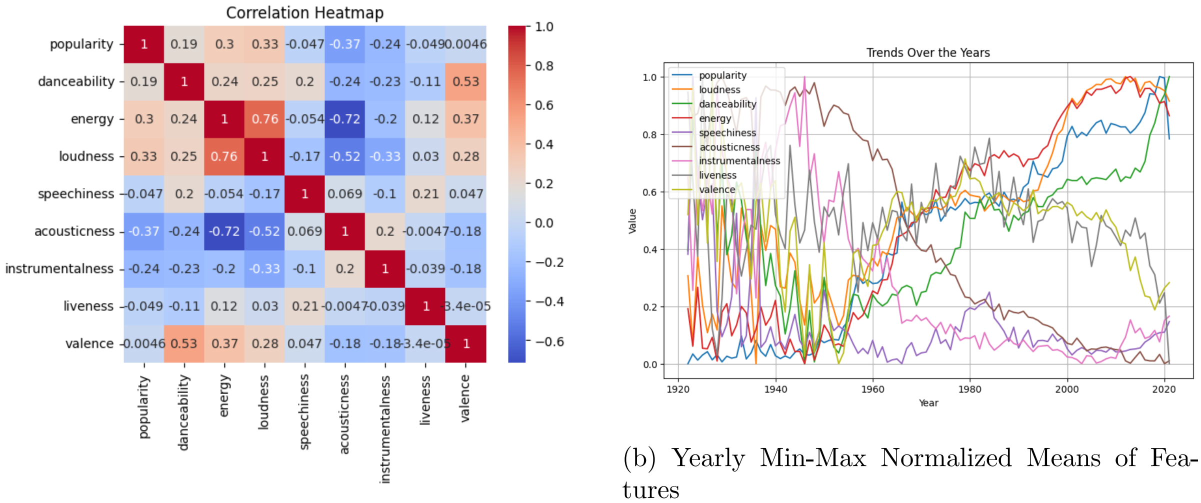
6. Preliminary Results

We pick linear regression as out baseline. Ridge regression is picked to handle multicollinearity observed during EDA. Different tree-based gradient boosting models are tried owning to their ability to handle non-linearity and large number of features. The results of the preliminary modelling experiments are shown in Table 1. As expected, gradient boosting tree models perform better than the rest, in general. Multi-layer perceptron surprisingly performs the worst despite its ability to learn complex-relationships. Note that these results are without any parameter tuning.

Model	**MSE**	**RMSE**	**MAE**	**R^2 Score**
Linear Regression	0.0257	0.1603	0.1300	0.1765
Ridge	0.0257	0.1603	0.1300	0.1765
Lasso	0.0311	0.1763	0.1451	0.0040
ElasticNet	0.0311	0.1763	0.1451	0.0040
Decision Tree Regressor	0.0440	0.2097	0.1622	-0.4090
Gradient Boosting Regressor	0.0235	0.1532	0.1229	0.2485
AdaBoost Regressor	0.0266	0.1630	0.1347	0.1485
K-Neighbors Regressor	0.0342	0.1849	0.1494	-0.0952
MLP Regressor	2.0150	1.4195	0.9874	-63.5507
XGBoost Regressor	0.0225	0.1501	0.1191	0.2787
LightGBM Regressor	**0.0227**	**0.1505**	**0.1199**	**0.2741**

Table 1: Regression Model Metrics.

[11] we utilize MinMaxScaler from Sklearn



(b) Yearly Min-Max Normalized Means of Features

(a) Pearson's Correlation Heatmap of Features

Figure depicting results of EDA and data pre-processing. We observe interesting correlations and temporal trends: (a) Energy and loudness are highly correlated, while acousticness has a negative correlation with both. Loud and energetic songs tend to be more popular. (b) Yearly trends shows an upward trend in danceability and loadness over the years, while instrumentalness has gone down.

7. Future Plan of Work

With the initial analysis showing promising results, we primarily aim to work in following three fronts:

(1) Improving feature selection: To select the best sets of features for our model, we would perform recursive feature elimination. Additionally, we plan to evaluate the impact of dimensionality reduction methods like PCA and tSNE to further refine our features.

(2) Improving Robustness of Models: To select the optimal set of hyper-parameters for our models, we would perform hyper-parameter tuning using techniques like grid search or bayesian opmitzation. While tuning the parameters, we would perform a K-fold cross validation and use the aggregated results from all the folds to enhance the robustness of results.

(3) Adding explanability to results: We aim to perform a SHAP analysis on our top models to understand the most important features which determine popularity of songs. Next, we would create some genre-wise visualizations to qualitatively understand the properties of different genres. Finally, we want to study the impact of artist popularity in determining the popularity of a song. For this, we would use unsupervised clustering methods like KMeans and Hierarchical clustering to group similar songs together and use hypothesis testing to study whether songs by popular artists are significantly more popular than the rest .

8. Conclusions / discussions

In this study we analysed the spotify popularity data and built models to predict the popularity of songs given various parameters. We showed different insights through our visualizations, did the necessary pre-processing steps and analyzed our regression models. From our preliminary work we have the following important finding that normalization of parameters is an extremely crucial step while training any regression model. We see the boosting models perform the best whereas MLP performs the worst in our given set of models.

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.

A Appendix

1. **Dataset Details and summary**
   The dataset that we have chosen contains the following attributes -

   Track ID Unique identifier for the track.

   Track Name Name of the track.

   Popularity Popularity of the track in the range 0 to 100.

   Duration (ms) Duration of the song in milliseconds.

   Explicit Indicates whether the track contains explicit content or not.

   Artists List of artists who created the track.

   Artist IDs Unique identifiers of the artists who created the track.

   Release Date Date of release of the track.

   Danceability A measure of how danceable a song is, ranging from 0 to 1.

   Energy A measure of how energized a song is, ranging from 0 to 1.

   Key The key the track is in. Integers map to pitches using standard Pitch Class notation ranging from 0 - 11

   Loudness How loud a song is in decibels (dB) normalized to the range of 0 to 1

   Mode The modality of the track, where 0 indicates minor and 1 indicates major.

   Speechiness The presence of spoken words in the track, ranging from 0 to 1.

   Acousticness How acoustic a track is, ranging from 0 to 1.

   Instrumentalness The absence of vocal sounds.The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

   Liveness Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

   Valence A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

   Tempo The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. Values scaled to 0 - 1.

   Time Signature An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures

   Followers The followers for an artist having an artist id. Values were scaled to between 0 to 1

   Artists Popularity The popularity for an artist having an artist id. Values were scaled to between 0 to 1

Metric	Mean	Std	Min	25%	50%	75%	Max
Popularity	28.39	17.49	0.00	15.00	28.00	41.00	99.00
Duration (ms)	226,738	114,759	3,344	175,647	214,787	261,516	5,621,218
Explicit	0.03499	0.18375	0.00000	0.00000	0.00000	0.00000	1.00000
Danceability	0.56591	0.15908	0.00000	0.46000	0.57700	0.68200	0.99100
Energy	0.55545	0.24411	0.00000	0.36700	0.56000	0.75500	1.00000
Key	5.22310	3.52034	0.00000	2.00000	5.00000	8.00000	11.00000
Loudness	0.76658	0.07065	0.00000	0.72674	0.77801	0.81879	1.00000
Mode	0.66443	0.47219	0.00000	0.00000	1.00000	1.00000	1.00000
Speechiness	0.10076	0.17809	0.00000	0.03320	0.04280	0.07130	0.96900
Acousticness	0.42915	0.33930	0.00000	0.08880	0.39400	0.74600	0.99600
Instrumentalness	0.09376	0.24188	0.00000	0.00000	0.00002	0.00512	1.00000
Liveness	0.21400	0.18524	0.00000	0.09800	0.13900	0.27800	1.00000
Valence	0.56433	0.25263	0.00000	0.36100	0.57600	0.77800	1.00000
Tempo	0.48391	0.12056	0.00000	0.39111	0.47910	0.55724	1.00000
Time Signature	3.87966	0.45668	0.00000	4.00000	4.00000	4.00000	5.00000
Release Year	1988.85	21.28	1900.00	1976.00	1992.00	2006.00	2021.00
Followers	0.01350	0.04881	0.00000	0.00016	0.00119	0.00755	1.00000
Artists Popularity	0.50311	0.19689	0.00000	0.38000	0.52000	0.65000	1.00000

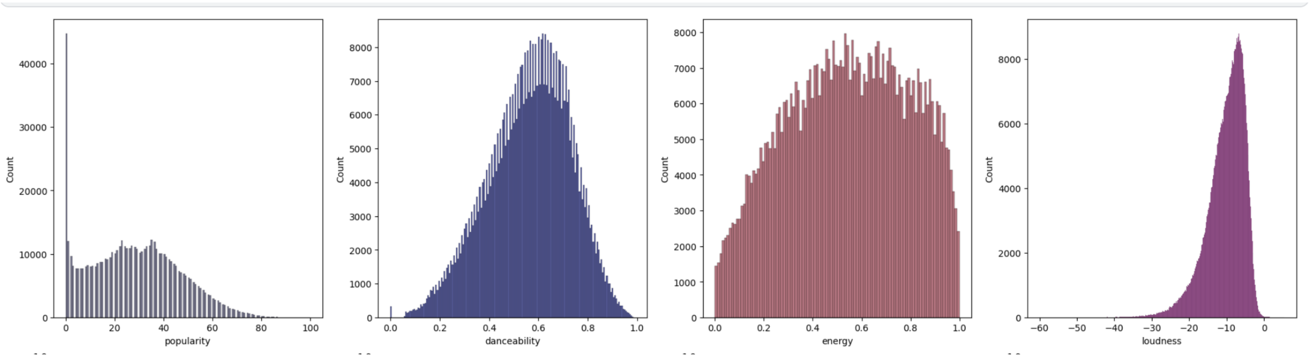
Table 2: Summary Statistics of the Data

The dataset contains 586,672 rows of popular tracks, but after merging the tracks data with artists we minimze the dataset due to missing artist ids, thereby having a total of 470,038 rows. It contains Null values only in the following column -

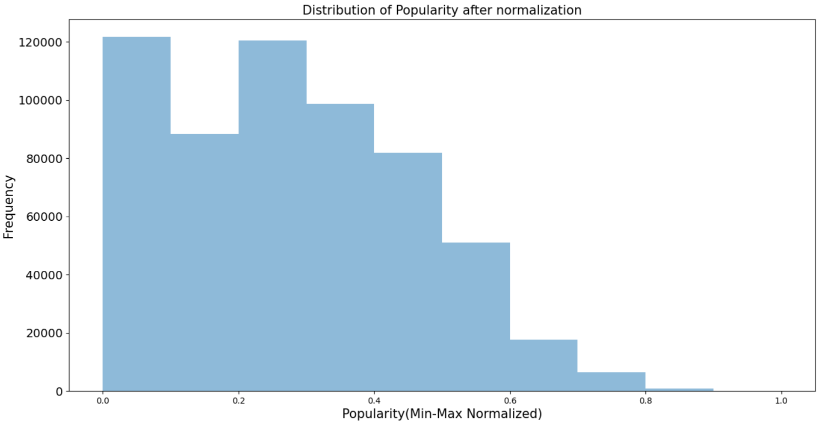
(a) Track Name - 71 rows

(b) Followers - 11 rows

2. **Data Visualization**



Distribution of datapoints for the following features: popularity, danceability, energy and loudness



Distribution of popularity after min-max normalization