Predicting Spotify Song Popularity Using Audio Features

CSC1181 Foundations of Statistical Analysis Machine Learning

November 28, 2025

Contribution Table

Team Member	Contribution
Reeha Althaf (A00050820)	Data preprocessing, EDA, Feature engineering, Model development, report coordination
Rachel Kunjumon (A00049389)	Data preprocessing, Feature engineering, Model development, Ethics analysis, report coordination
Proneel Banerjee (A00050898)	Model development, Feature importance analysis, Error Analysis, Insights beyond popularity
Niket Suresh Ahire (A00049228)	Data preprocessing, Error analysis, Report coordination

Link to Gitlab: Group-38-predicting spotify song popularity

Contents

List of Figures

List of Tables

1 Abstract

This project investigates factors that influence the popularity of songs on Spotify using machine learning models. We utilized a dataset containing metadata, audio features, and genre-related information. This dataset was preprocessed using cleaning, encoding, and feature scaling. Several models including a Random Forest model was trained on around 120 features. This paper analyzes the performance of these machine learning models and discusses possible limitations when it comes to predicting popularity of a song using audio derived features alone.

2 Problem Data

2.1 Project Overview

The aim of this project is to predict the popularity of a song using audio features as well as audio derived features. The performance of the model would then be investigated to understand if the aforementioned features are sufficient to explain whether a song is deemed for success. The analyses focuses on evaluating how well the machine learning models employed capture patterns in audio characteristics and if popularity can be inferred reliably without any additional factors.

Target Variable: Popularity (0-100 scale, where higher values show more listener interest)
Success Metrics:

- Root Mean Squared Error (RMSE) - measures average prediction error
- R^2 score - explains variance captured
- Mean Absolute Error (MAE) - interpretable average deviation

2.2 Dataset Description

The dataset used is sourced from Kaggle's "500K+ Spotify Songs with Lyrics,Emotions More"[1], which contains information about roughly 551443 tracks, each with 39 features .The attributes capture a wide range of audio characteristics that describe the sound of the songs. These include Danceability, Energy, Loudness, Speechiness, and much more. The other musical attributes include Key, Tempo, and Time signature. In addition to audio features, the dataset also provides metadata like track title, artist name, album, and release date. Furthermore, attributes that describe the general mood of the song such as genre, emotion as well as usage recommendations like "Good for Parties," are also included in the dataset.

3 Methods

3.1 Exploratory Data Analysis

In the initial review of the dataset, we saw both numeric features such as danceability, energy, and tempo, and a number of columns either categorical or incorrectly formatted. For example, track duration was represented in mm:ss format and had to be converted into seconds. Genre and emotion fields needed encoding, while loudness needed transformation from text to numeric format for further analysis. Similarity-score columns were all very highly correlated with each other and thus added little information. Finally, correlation analysis showed that most audio attributes have weak relationships with popularity. The popularity variable itself was right-skewed, as one might expect. These findings influenced the subsequent data-cleaning decisions and shaped our approach to preprocessing and model selection.

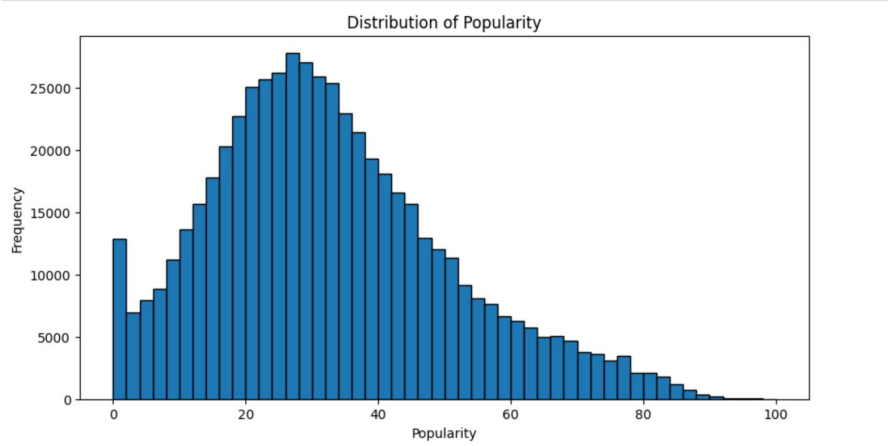

Figure 1: Distribution of Popularity Scores

3.2 Preprocessing

The following preprocessing techniques were applied to our dataset

1. Duplicate Rows: Duplicate rows in our dataset were dropped.
2. Null Values:
    - Time Signature: There were 8 rows with values missing for Time Signature. We decided to find rows with missing Time Signature, see what Genre they map to, and use the mode Time Signature of that Genre.
    - Song, Artist, Similar Song (1, 2, and 3): Filled these columns with the placeholder value "Unknown" as deriving it from pre-existing columns or data would not be possible.
3. Track Length (seconds): parsed from mm:ss text to seconds.
4. Emotion Encoding: Label Encoding
5. Key Encoding: Label Encoding
6. Genre encoding: One - Hot Encoding
7. Loudness: Truncated the 'db' that followed the number.
8. Similarity-score aggregates: Dropped due to potential data leakage.

3.3 Models tested

1. **Dummy Regressor (Baseline Model):** We started with a DummyRegressor model to generate a baseline score by predicting the mean value of the target variable.
2. **Linear Regression:** Further, we applied a Linear Regression model to evaluate how well the target variable can be predicted. However, it performed poorly due to the non-linear nature of our features.
3. **Random Forest Regressor:** A Random Forest Regressor is robust to non-linear relationships, outliers, and in-built 'Feature Importance' [2].

4 Results

4.1 Model Performance

Model	RSME	R^2	MAE
Baseline (Mean)	17.17	-0.0001	13.48
Linear Regression	14.95	0.24	11.72
Random Forest (best)	**12.29**	**0.4871**	**9.38**

Table 1: Test set performance. Best model in bold.

Random Forest Regressor provided the best performance and became the final model.

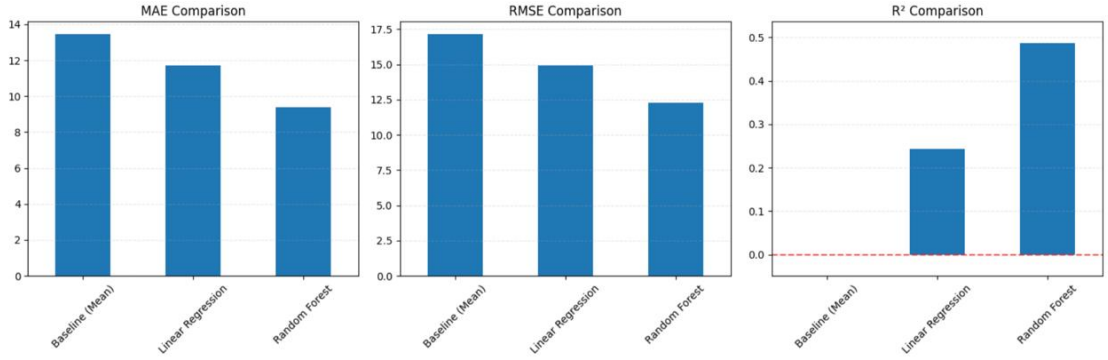

Figure 2: Metric Comparison of models

4.2 Feature Importance

Key features contributing to the track's popularity included being marked as suitable for a party, being louder than average, and having distinctive lengths. Tempo, positiveness, energy, danceability, acousticness combined to indicate listener engagement; while acousticness, liveliness, and key encoding helped distinguish musical characteristics linked to popularity.

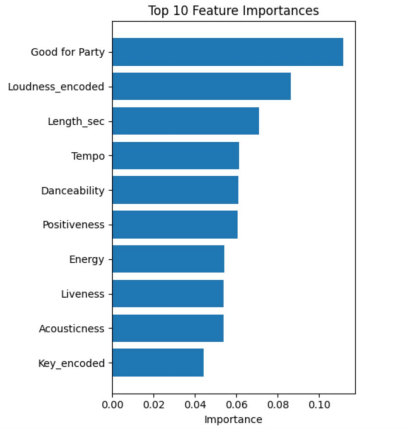

Figure 3: Top 10 Feature Importances from the Random Forest Model

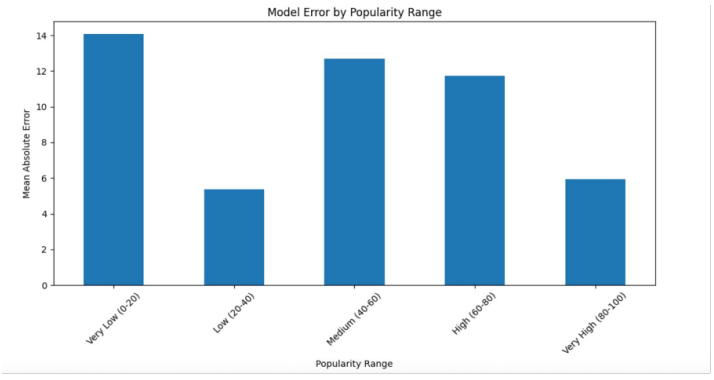

Figure 4: Error Analysis

5 Analysis & Discussion

Interpretation: Audio features explain some of a song's popularity, but other factors like the artist's reputation, marketing, playlist placement, and social trends also have a big impact.

Error Analysis: The model had trouble predicting very popular songs (over 80), often underestimating their popularity, and struggled with very unpopular songs, where random noise and niche tracks caused inconsistent results. Notable errors occurred for a few songs, indicating that factors like release date, artist popularity, and playlist exposure that is beyond the audio itself can affect popularity.

Limitations: The dataset doesn't include time-related or market-context information, and the genre categories are too broad to capture subtle differences. Popularity also depends on platform and social factors that audio features alone cannot explain.
To confirm that our models effectively learns meaningful patterns, we tested its ability to predict an audio feature from other audio characteristics. When predicting energy, the model achieved R^2=0.88, demonstrating strong performance on audio-related tasks. This confirms the models are functioning correctly and can capture audio relationships well.

6 Ethical & Legal Considerations

6.1 Data Protection & Licensing

This project adheres to ethical and legal standards by ensuring responsible use and handling of the dataset. The dataset is present in Kaggle under an open license that permits research and educational use. The data contains only publicly available information about songs and their audio features, with no user-specific listening data or personal information. The artist names present in the dataset refer to public professional identities rather than private personal information, ensuring compliance with data protection principles.

6.2 Bias & Fairness

We acknowledge that Spotify's popularity scores, genre and emotion classifications, as well as certain audio features, may reflect existing cultural, commercial, or algorithmic biases. To address this, we treated the data impartially, followed standardized preprocessing procedures, and refrained from making subjective judgments about the quality of the music or listener tastes. Our study is focused strictly on uncovering statistical links between these audio features and popularity, rather than ranking artists or forecasting commercial outcomes beyond what the data supports. Throughout the project, we remain mindful of the dataset's limitations and take care not to exaggerate the accuracy of our model, promoting a careful and transparent interpretation of the results.

7 Conclusion & Future Work

This project showed that audio features can somewhat predict how popular a Spotify song will be, with the Random Forest model reaching an R^2 of about 0.48. Loudness, tempo, song length, and mood/activity labels were the strongest predictors. However, song popularity is still hard to model because it depends a lot on outside factors like marketing, trends, and artist fame.

Future improvements include adding artist info and release-year data, using lyrics or audio directly, testing other models, improving data balance, and incorporating playlist or chart information to better capture social and commercial influences.

References

[1] 500K+ Spotify Songs with Lyrics,Emotions More (no date). Available at: https://www.kaggle.com/datasets/devdope/900k-spotify (Accessed: November 25, 2025).

[2] Bhuva, L. (2025) "Understanding Feature Importance in Machine Learning," Medium, 20 March. Available at: https://medium.com/@lomashbhuva/understanding-feature-importance-in-machine-learning-d86ec50e0055(Accessed: November 27, 2025).

[3] Feature Importance with Random Forests (15:31:51+00:00) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/feature-importance-with-random-forests/ (Accessed: November 27, 2025).

[4] Feature Selection Using Random Forest (13:18:22+00:00) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/feature-selection-using-random-forest/ (Accessed: November 27, 2025).

[5] One Hot Encoding in Machine Learning (11:43:45+00:00) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/ml-one-hot-encoding/ (Accessed: November 27, 2025).

[6] Random Forest Regression in Python (00:30:53+00:00) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/ (Accessed: November 27, 2025).