Spotify Song Data Analysis Using Machine Learning

Harsh Sahay, Srijan Gupta, Sahil Mehta (all UG)

Content

1. Data Preprocessing and Feature Engineering
2. Exploratory Data Analysis (EDA)
3. Model Development and Evaluation

Dataset Source — https://www.kaggle.com/datasets/arnavvvvv/spotify-music
Size — 953 songs
Shape — 24 columns
Partitioning — 80/20 ratio

Pre-processing performed —
Missing Values: Rows with missing values are removed.
Non-Numeric Values: Rows with non-numeric values in the 'streams' column are removed.
Data Type Conversions: The 'streams' column is converted to a float type after removing commas. Several other columns like 'in_spotify_playlists', 'in_spotify_charts', etc., are converted to numeric types, handling non-numeric entries as errors.
Feature Engineering: New features such as 'artist_popularity' (average number of streams per artist) are created.
Encoding: Categorical variables are encoded using OneHotEncoder, though specific categories are not detailed in the snippet.

Section 1

Data Preprocessing and Feature Engineering

Handling Missing Values and Data Cleaning

01

Data Cleaning

The initial steps involved handling missing values and cleaning the dataset to ensure the integrity of the data for analysis.

02

Feature Engineering

New features such as artist popularity and genre diversity were created to enhance the depth of the analysis.

Encoding Categorical Variables and Preprocessing

Categorical Variable Encoding

The process of encoding categorical variables was crucial for preparing the data for machine learning models.

Preprocessing

Standard scaling and one-hot encoding were applied to the data to ensure compatibility with the machine learning algorithms.

Feature Creation and Transformation

01

Artist Popularity

A feature for artist popularity was created, providing insights into the impact of different artists on song streams.

02

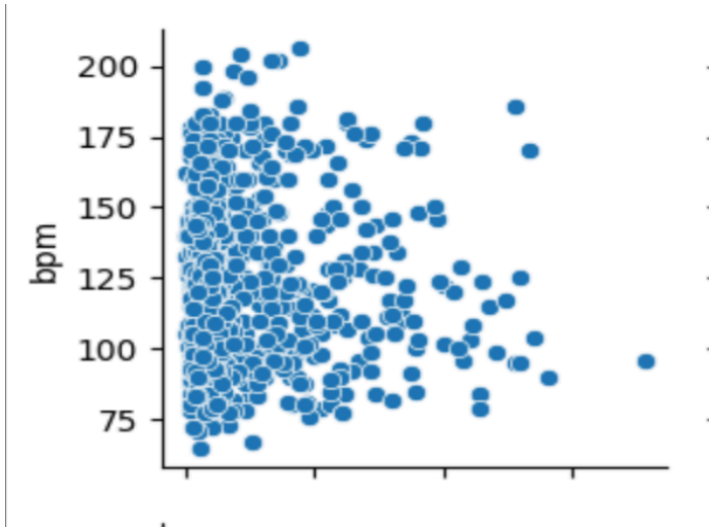
Genre Diversity

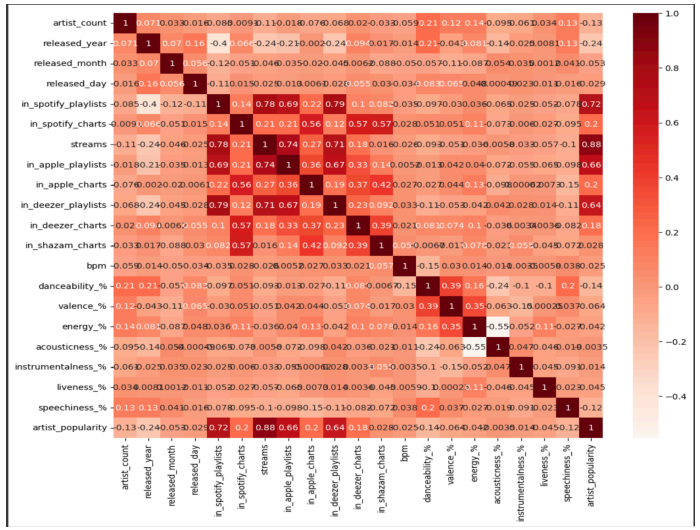
The analysis included the creation of a feature to measure genre diversity, shedding light on the variety of genres represented in the dataset.

Section 2

Exploratory Data Analysis (EDA)

Correlation Analysis





Correlation Matrix

A correlation matrix was utilized to understand the relationships between different variables, providing insights into the interplay of factors influencing song streams,

Visualization

The use of visualizations such as heatmaps facilitated the exploration of correlations within the dataset.

Visualizing Relationships

Pairplot Visualization

A pairplot was employed to visualize relationships between key variables, offering a comprehensive view of the interconnections within the dataset.

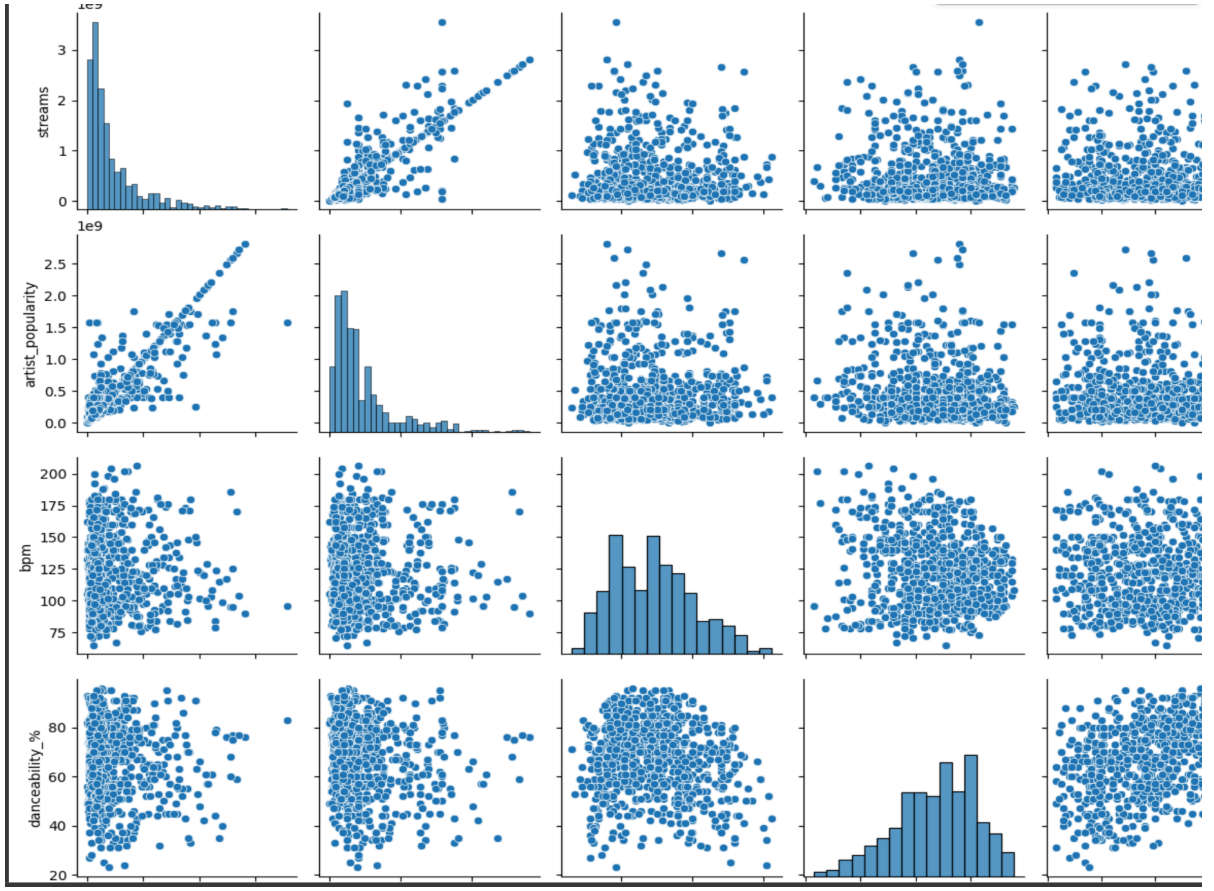
Insights from Visualization

The visualization provided valuable insights into the patterns and associations present in the Spotify song data.

Key Findings from EDA

Insights from EDA

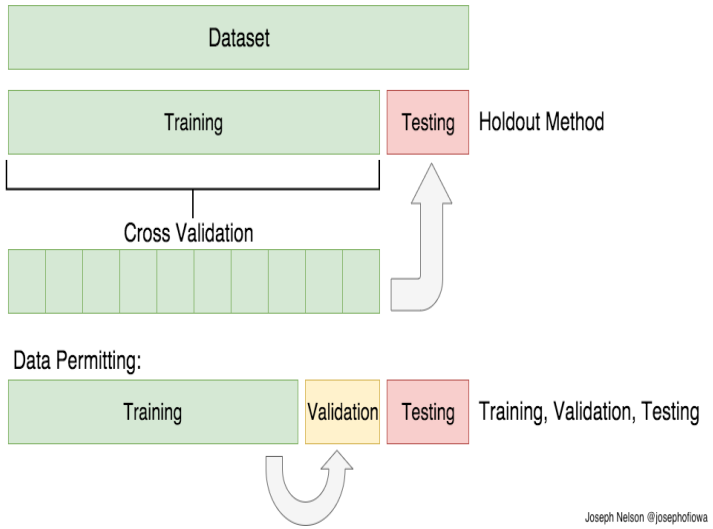
The EDA process yielded significant findings regarding the distribution and relationships of variables, laying the groundwork for subsequent model development and evaluation.



Section 3

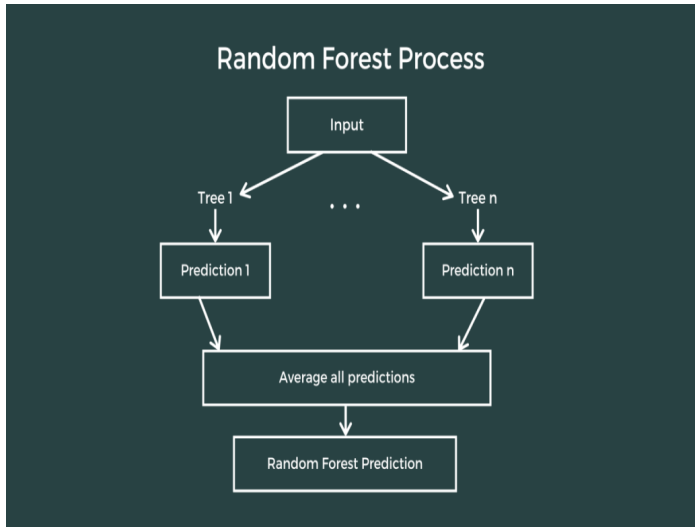
Model Development and Evaluation

Data Splitting and Model Creation



Data Splitting

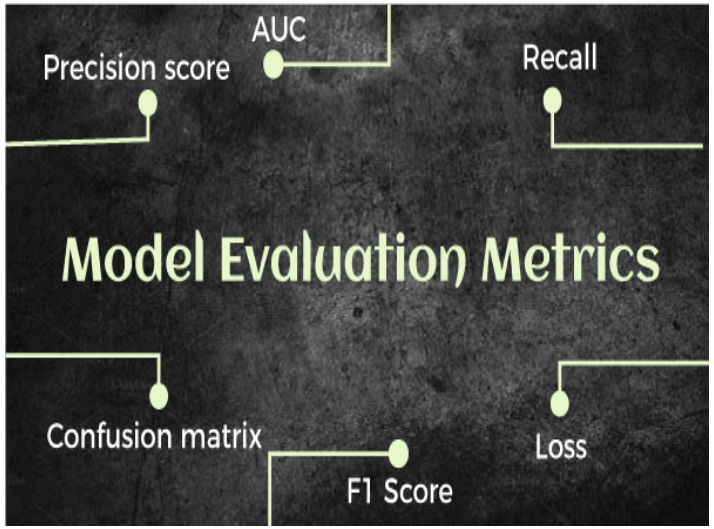
The dataset was split into features and target variables, a crucial step in preparing the data for model development.



Model Creation

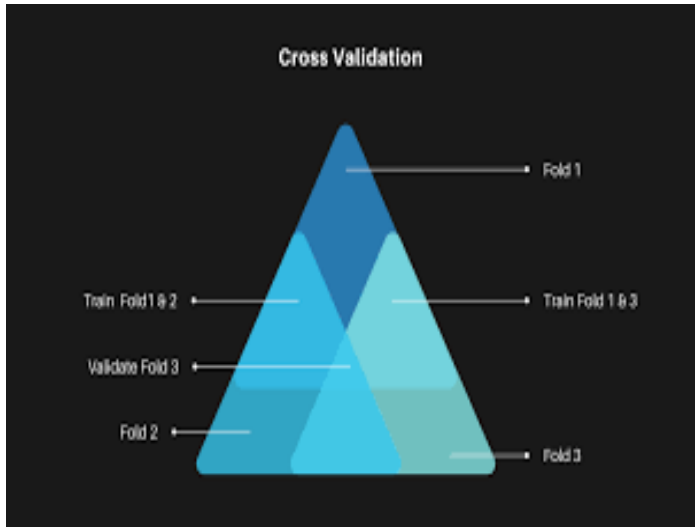
Pipelines for different models, including Linear Regression, Random Forest, and Gradient Boosting, were established to facilitate the modeling process.

Model Performance Evaluation



Model Evaluation Metrics

The performance of each model was assessed using metrics such as RMSE, MAE, and R-squared, providing a comprehensive understanding of their predictive capabilities.



Cross-Validation

Cross-validation techniques were employed to validate the robustness of the models and ensure reliable performance metrics.

Implications and Applications

01

Practical Implications

The analysis and modeling of Spotify song data using machine learning techniques have far-reaching implications for understanding user preferences and behavior.

02

Educational Applications

The insights derived from this analysis can be leveraged in educational settings to illustrate real-world applications of data science in the music streaming domain.

RESULT (1)

Content:

The performance of our predictive models was assessed using RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R-squared metrics to understand how well our models can predict the number of streams—a proxy for song popularity on streaming platforms.

Linear Regression Results:

RMSE: 544,931,461.47 - Indicates the model's predictions deviate from the actual number of streams on average by this value.
MAE: 391,276,191.38 - Suggests that the average magnitude of the errors in the predictions is quite substantial.
R-squared: 0.03 - This low value implies that only 3% of the variance in the song streams is being captured by the model, hinting at a weak predictive power.

Random Forest Results:

RMSE: 573,201,013.35 - Higher than that of Linear Regression, indicating less accurate predictions.
MAE: 408,461,337.68 - Shows that on average, the Random Forest model's predictions are further off than the Linear Regression's.
R-squared: -0.07 - A negative R-squared denotes that the model is worse than a naive mean-based prediction.

RESULT (2)

Gradient Boosting Results:

RMSE: 574,828,398.22 - Similar to Random Forest, indicating inaccuracies in the predictions.

MAE: 407,081,857.85 - Mirroring the pattern of the Random Forest model's MAE.

R-squared: -0.08 - Also negative, suggesting that the model's predictions do not align well with the actual data.

Cross-validation Insights:

The cross-validation process was implemented to evaluate the robustness of our models. It revealed high variance in the RMSE scores for all models, indicating inconsistency in performance across different subsets of data.
Linear Regression had the lowest average RMSE score in cross-validation, making it the most stable among the models, though with considerable variation.
Both Random Forest and Gradient Boosting showed large variances in their cross-validation scores, reflecting substantial inconsistency in their predictive accuracy.

Analysis:

All models performed poorly, with negative R-squared values for Random Forest and Gradient Boosting models, indicating a possible overfitting to the training data or that the models are not well-suited to this prediction task.
The large RMSE and MAE values across models signify that the predictive power of the models is low, and they may struggle with the scale of the 'streams' variable or the features provided are not informative enough.

Contributions to the project

Srijan Gupta

Srijan focused on the data preprocessing and cleaning aspects of the project. This included identifying and handling missing values, correcting data formats, and ensuring data integrity. Srijan also performed the initial data exploration to understand the distribution and characteristics of the various features in the dataset.

Harsh Sahay

Harsh was responsible for the data analysis and visualization part. He created various visualizations to uncover patterns and insights in the data, such as correlation matrices and pairplots. Harsh's analysis helped in understanding the relationships between different musical attributes and their impact on song popularity.

Sahil Mehta

Sahil took charge of the model development and evaluation. He implemented multiple regression models to predict song streams based on other features in the dataset. Sahil also conducted performance evaluation of these models using metrics like RMSE, MAE, and R-squared, and performed cross-validation to ensure the model's robustness.

Thank you, goodbye!