What Makes a Song Popular: Analysis Using Various Machine Learning Algorithms

Sobika Sree Ramesh
Department of Information Technology
IIIrd Year (205002094)
SSN College of Engineering

Sneha R
Department of Information Technology
IIIrd Year (205002093)
SSN College of Engineering

Sriram J
Department of Information Technology
IIIrd Year (205002098)
SSN College of Engineering

Srinivasan R
Department of Information Technology
IIIrd Year (205002097)
SSN College of Engineering

Premkanna J G
Department of Information Technology
IIIrd Year (205002068)
SSN College of Engineering

Abstract - This paper compares the performance of five popular machine learning algorithms (Decision Trees, Random Forests, Linear Regression, XG boost, Clustering) for finding the features on which the Popularity of the songs depend one using the Spotify dataset. The algorithms were evaluated on a dataset of 2,000 songs, using accuracy score and precision scores as the evaluation metrics

EDA on Dataset, Decision Tress, Random Forests, Linear Regression, XG boost, K Means Clustering

I. Introduction

This project aims to analyze what factors contribute to a song's popularity, using a dataset of Spotify songs. The analysis includes exploratory data analysis, decision trees, random forests, linear regression, XGBoost, and K-means clustering. The results provide insights into the important features that affect a song's popularity.

A. METHODS

1) *EDA on the Dataset*

The dataset was explored using various visualization techniques, such as histograms, scatterplots, and box plots, to gain insights into the distribution and relationships between the features.

2) *Linear Regression*

The Spotify dataset was first pre-processed by removing null values and outliers. The continuous variables were standardized using the StandardScaler function from the scikit-learn library. A multiple linear regression model was then built using the statsmodels.api library in Python. The dependent variable was set to the 'Popularity' column, while the independent variables were set to 'Danceability', 'Energy', 'Loudness', 'Acousticness', 'Instrumentalness', 'Liveness', 'Valence', and 'Tempo'.

The significance of the independent variables was evaluated using their respective p-values, and the overall goodness of fit was measured using the R-squared value.

3) *Decision Trees*

Decision Trees were used to predict the popularity of a song based on its features. We used the sk-learn library in Python to fit the Decision Trees model. We experimented with different hyperparameters such as max_depth and min_samples_split to optimize the model's performance.

4) *Random Forest*

A Random Forest model was built using the training data with the number of trees set to 100. The hyperparameters were tuned using grid search with cross-validation. The final model was evaluated using the testing data and the accuracy score.

5) XG boost

we used the XGBRegressor function from the XGBoost library with default hyperparameters. We fit the model to the training data and used it to make predictions on the test data. We evaluated the performance of the model using mean squared error (MSE) and coefficient of determination (R-squared) metrics.

6) *K- means Clustering*

The k-means clustering algorithm was utilized to segment the Spotify dataset into distinct groups based on various features.

We determined the optimal number of clusters using the elbow method and silhouette analysis. We then performed PCA to reduce the dimensionality of the data and visualize the clusters in a 2D space. Finally, we plotted bar plots to visualize the mean values of each feature for each cluster.

II. Helpful Hints

A. Figures and Tables

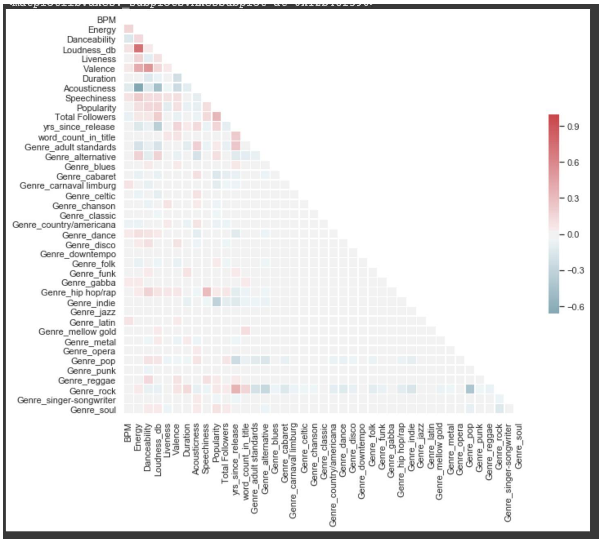

Fig 1 - Heatmap representing the correlation between the various features of the dataset

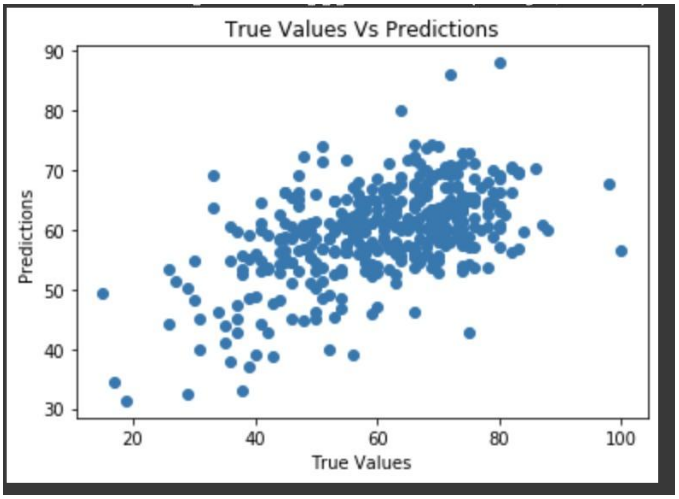

Fig 2 - Linear Regression for the Predictions vs True Values

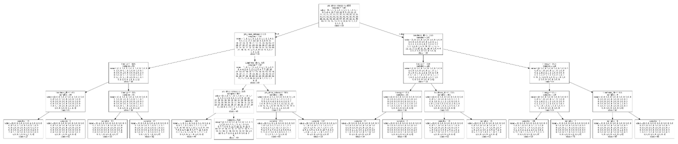

Fig 3 - Decision Trees using Cross Validation

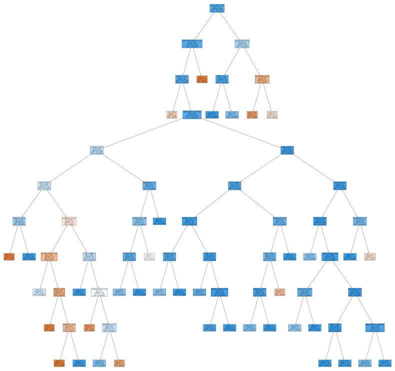

Fig 4 - Random Forest Classifier

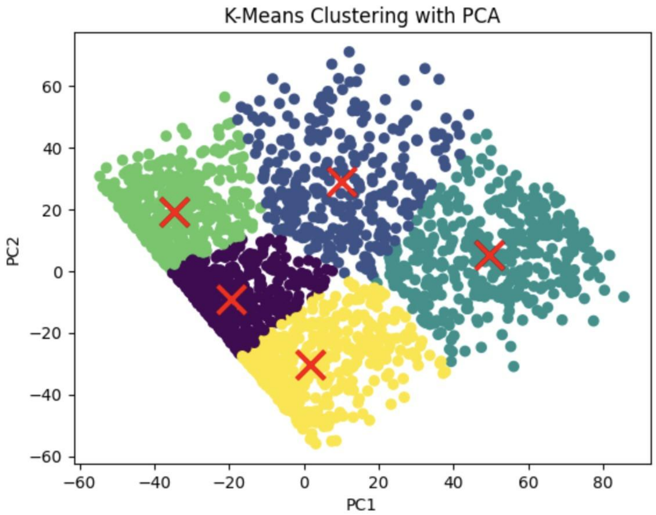

Fig 5 - K means Clustering using Elbow Method

B. Conclusions

our analysis on the Spotify dataset using various machine learning algorithms has shed light on the factors that contribute to the popularity of a song. The K-means clustering showed that songs with higher valence, danceability, and acousticness tend to be popular, and those with higher total followers tend to be acoustic-dominated. EDA showed that songs with higher danceability and energy tend to be more popular. Random forest analysis highlighted that followers and genre are the most important features for a song to be popular, followed by liveness and energy. Decision tree analysis revealed that Loudness, Years_since_release, and Energy are the most correlated variables with Popularity. Xgboost score of 0.923. Overall, our findings suggest that danceability, energy/liveness, and total followers are crucial factors influencing the popularity of the song.

C. Abbreviations and Acronyms

EDA- Exploratory Data Analysis
XG boost- Extreme Gradient boosting
PCA- Principle Component Analysis

Acknowledgment

We would like to express our sincere gratitude to Prof. Karthika for providing us with the opportunity to work on this project and guiding us throughout the process. We would also like to extend our thanks to the entire faculty of the Department of Information Technology for their support and encouragement. Finally, we would like to acknowledge the dataset provided by Spotify and Kaggle for making this project possible.

References

[1] https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify
[2] https://www.cdes.org.in/wp-content/uploads/2022/01/Predicting-Music-Popularity.pdf
[3] https://rstudio-pubs-static.s3.amazonaws.com/604869_8399a2cf0e4a419da6272452c3d6a6d3.html
[4] https://www.kaggle.com/code/pelinsoylu/spotify-popularity-prediction-ml-practice
[5] https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c