Kieran Knock
BrainStation
Sept. 9, 2023

Predicting Song Popularity

Exploring the Impact of Musical Features on Audience Reception

OVERVIEW

The purpose of this project is to attempt to answer the following question:

"Using machine learning, how might one analyze patterns in top global song releases to identify what makes a song a hit, such that future releases can be tuned to have maximum impact?"

The goal of this project is to understand the degree to which the musical features of a song impacts its success independent of the artist publishing the song.

BACKGROUND

The music industry generated $31 billion in 2022. While artists often collaborate with producers to perfect their sound, the explosion in popularity of machine learning has yet to influence the development of tools for musicians and producers significantly. This project aims to provide a machine-learning solution that would enable the music industry to grow revenue beyond traditional levels by reaching larger audiences.

DATASET

The dataset used for this project was a collection of song data gathered from Spotify & YouTube (the original dataset can be found here). The dataset comprises more than 20,000 songs from approximately 2,000 artists, detailing attributes such as song key, acousticness, danceability, and tempo, among others. A comprehensive list of these features, along with the corresponding data dictionary, is provided in the project submission.

The data was collected on February 6, 2023 and was made available on Kaggle on March 19, 2023. It was collected by a group of 3 Kaggle users for the purpose of analysis. Additional details on the collaborators can be found on the Kaggle page containing the data.

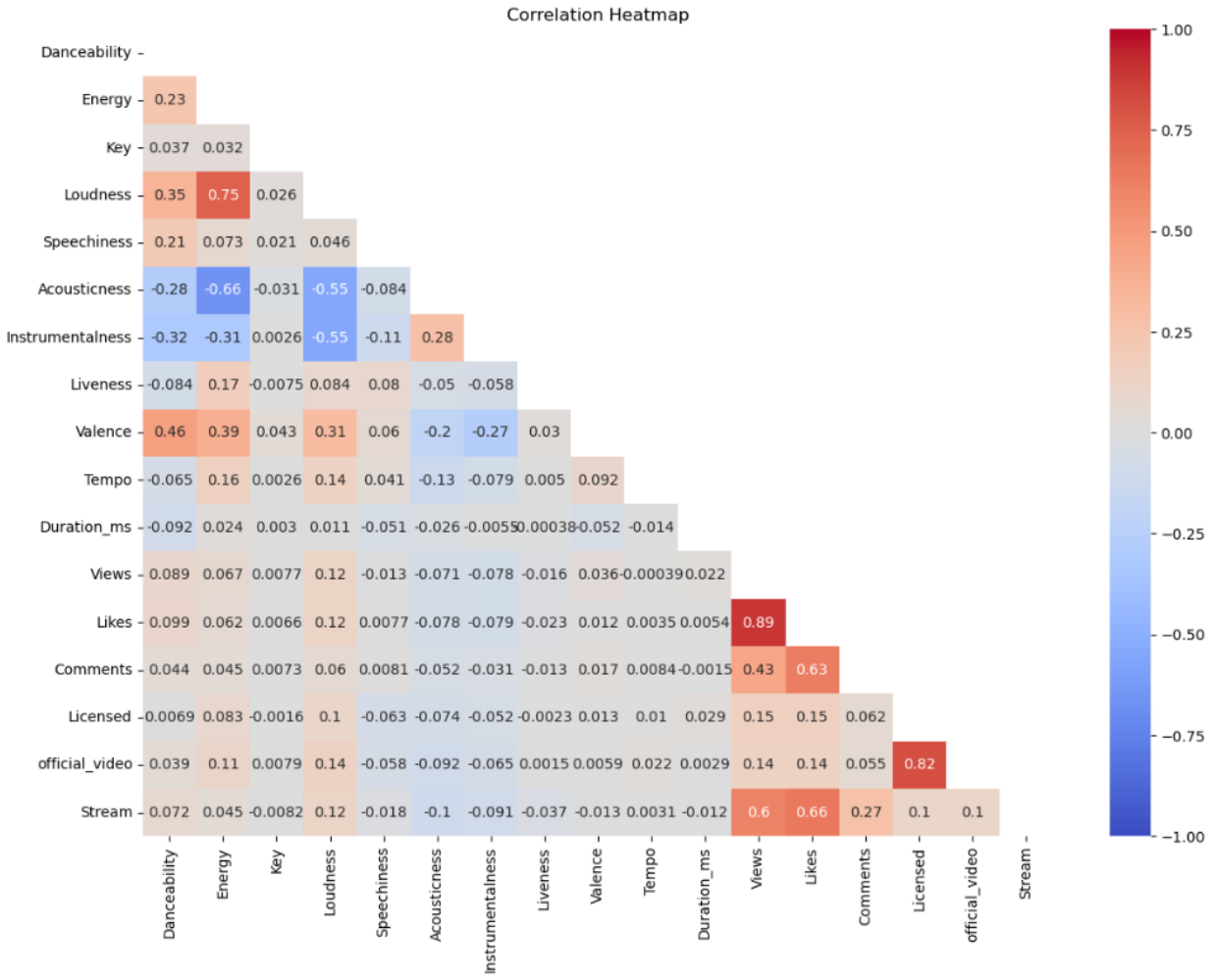
CLEANING & PREPROCESSING

The cleaning process contained some standard cleaning techniques. Some of the data was imputed using a simple linear regression model to replace missing values. Prior to dropping other missing values, a statistical analysis was performed to understand the potential impact on the larger dataset.

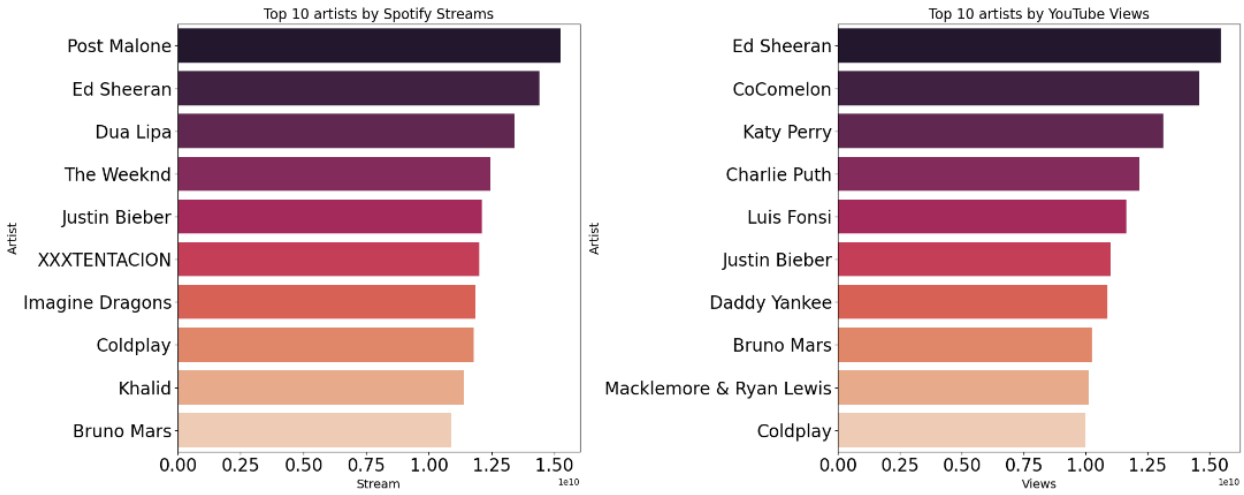
Some additional features were engineered for further exploration of the data and modeling. Most notably, the original dataset was missing observations indicating how long a song had been published, as this would surely impact the number of YouTube views a song would garner. Standard web-scraping techniques were used to extract this data and engineer a feature that would be representative of the length of time a song was published.

The original dataset contained 20,718 rows and 28 columns and the processed data used in modelling contained 19,812 rows and 27 columns.

EXPLORATORY DATA ANALYSIS



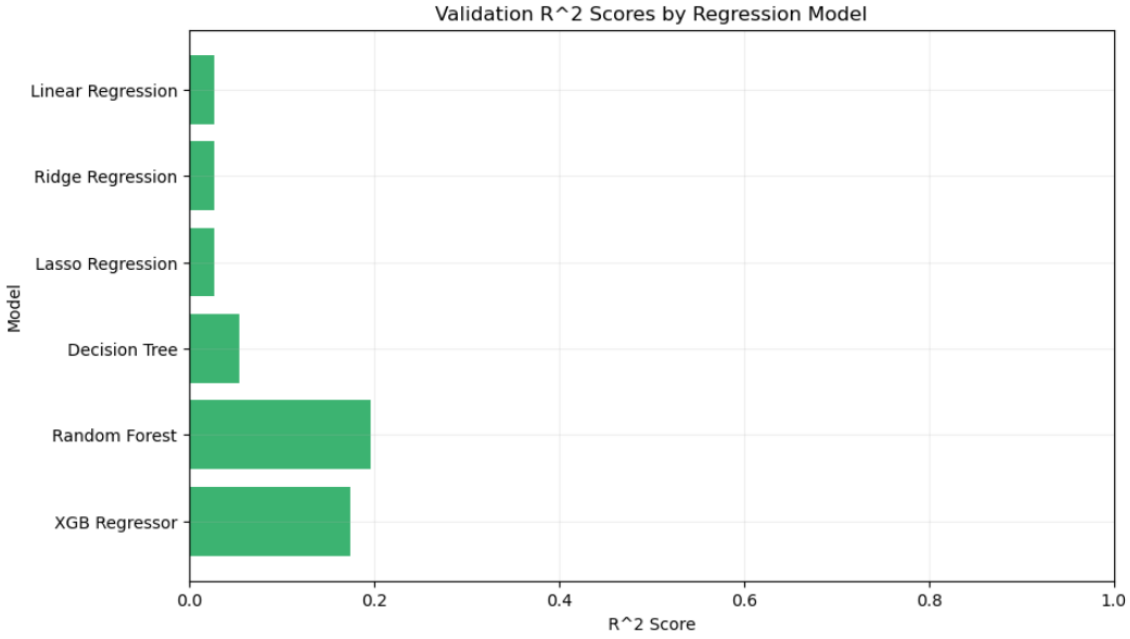
Exploration of the musical features used as predictor variables for the target variable “YouTube Views” showed very weak correlations between the number of views and all musical features. This would suggest that more data points are needed to build a model that could predict the number of views a song would get with a high degree of confidence.



It should also be noted that the number of views (YouTube) and streams (Spotify) that an artist may garner are not immediately comparable. Core audiences are noticeably different depending on the medium. Artists like Post Malone, The Weeknd, and XXXTENTACION dominate streams on Spotify but fail to appear on the top 10 list on YouTube. Conversely, children's artists like CoComelon are more popular on YouTube, as are Latin American artists like Luis Fonsi and Daddy Yankee.

MODELLING

Six different models were fit and tested, 3 different linear models and 3 tree models. The models were optimized by tuning hyperparameters via Grid Search using 5-fold cross-validation.



The models were evaluated on two different metrics:

1. R^2 Score: This measures the proportion of variance of the dependent variable (YouTube Views) that can be explained by the independent (predictor) variables.
2. Mean Absolute Error: This represents the average error between predicted and actual values in the dataset

The optimized Random Forest Regressor model performed the best in the context of R^2 score, which it scored 19.69% on the validation data. This score is quite poor, relative to “good” R^2 scores. The threshold for good may vary depending on the context and organization, but in most contexts, a relatively “good” score may be above 50%. This suggests that the data used is not optimal at predicting the number of views a song may get on YouTube.

The XGB Regressor scored best in terms of Mean Absolute Error with an MAE of 100,892,738. Again, this error is quite poor given how high it is. That said, the value of YouTube views a song may have is sometimes quite high (Ed Sheeran’s top video has 6 billion views on YouTube). So the high MAE is also somewhat expected.

Despite the poor performance of the model, some features were identified to be most important in determining how popular a song would be with the Random Forest model. The top 3 were:

1. Loudness
2. Number of Months Published
3. Acousticness

The lowest-scoring predictor variables were related to the key of the song. This is helpful as it suggests that the key a song is in has little to no bearing on how well it performs.

CONCLUSION

After exploring these models with the dataset independent of the artist releasing the song, these results are to be expected. It would appear that one cannot predict the performance of a song on the musical features alone. The artist themself contributes to the popularity of their songs through their image, individual style, and reputation.

The next steps to optimize these models for high performance would be to evaluate the artist and genre of music to see how that would improve model performance. Higher model performance, backed by more predictive data, would set a benchmark for evaluating song performance. Artists could then use this information to fine-tune their music, maximizing the impact of their future releases.