Most Streamed Spotify Songs of 2023: An Analysis

Meera Sriram

Introduction

Spotify has released a list of all of their most popular songs that have been streamed, saved, or have made charts in 2023. Information on the musical qualities of each song are mostly quantitative, with some variables, such as mode and key, being categorical.

The goal of this project is to understand how influential musical qualities are in making a song popular, with number of streams being used as a measure of popularity. Some important secondary questions considered for this analysis are what individual predictors or collection of predictors are the best to use when identifying a likely number of streams and whether the same variables can also be used to predict the mode of a song.

Given the structure of the dataset, it was clear from basic exploratory data analysis that these variables are independent of each other, therefore, none of these variables were strongly correlated enough to be further explored. However, the secondary questions serve the purpose of uncovering other ways these predictors can identify the likelihood of a song gaining popularity.

Overall, this analysis aims to use machine learning models on an unpredictable dataset, such as clustering and Random Forest regression and classification, to answer the questions posted and draw conclusions about the influence of these variables on bringing these songs to the top best.

The Dataset:

Titled '[Most Streamed Spotify Songs 2023]', this dataset consists of 943 unique values and 24 columns. Missing values, especially from the 'key' column, were removed, and categorical variables like mode and key were encoded into numeric values for future analysis. Most importantly, values within the column 'streams' were converted from objects to integers; this column was the only numeric column with values that weren't automatically in the dataset as integer values.

Other than 'track_name' and 'artist_name', the only other categorical variables in the loaded dataset were keys and mode. Variables like 'artist_count', 'release month', 'release_day', and values measuring performance on platforms other than Spotify were disregarded for this study.

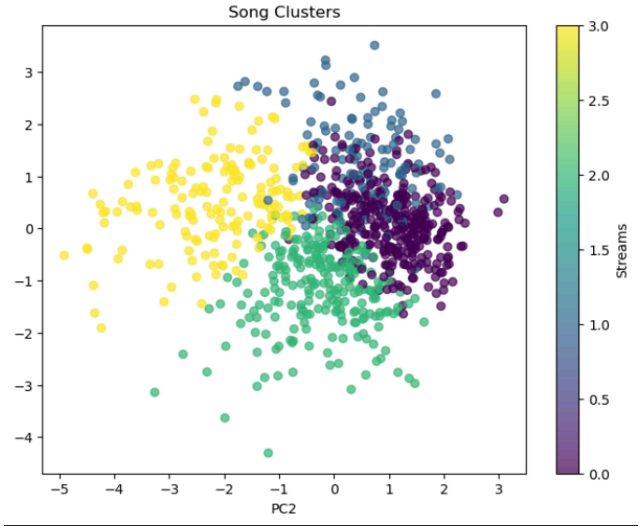
Variables observed for this analysis were: 'streams', 'bpm', 'Key_Type', 'Mode_Type', 'danceability_%', 'valence_%', 'energy_%', 'acounsticness_%', 'liveliness_%', and 'speechiness_%'. The variable, 'release_year' was initially included as a necessary predictor, however, it was quickly discovered that 'release_year' was a confounding variable, so it was removed from our analysis. Variables like 'in_spotify_charts' and 'in_spotify_playlist' were used for exploratory data analysis, but were not included as necessary variables for the models.

Results

Three models were used to evaluate the primary and secondary questions for this study: Clustering, Random Forest Regression, and Random Forest Classification:

Question 1: What individual predictors or collection of predictors are the best to use when understanding a song's likelihood of gaining popularity?

Principal Component Analysis (PCA) / Clustering



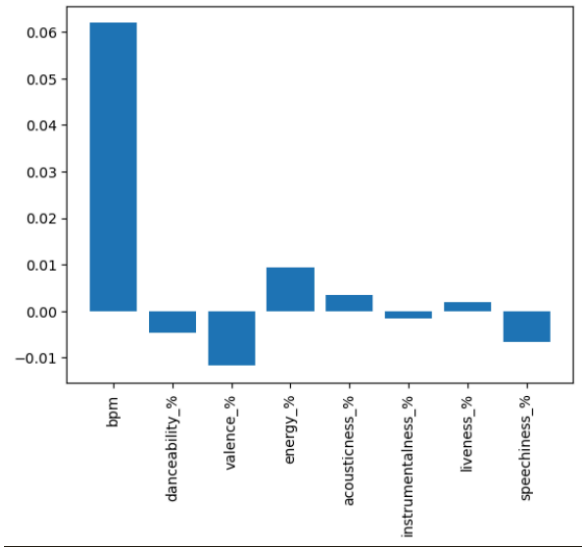
The x- and y-axis are organized based on PC1 and PC2, with PC1 representing the highest variance among predictors and PC2 represented by the second largest variance among predictors and is orthogonal to PC1. PC1 is most significantly impacted by high danceability, high valence, and high energy while contrasting with predictors like low bpm, low acousticness, and low instrumentalness. PC2, the category with the next greatest variance that is orthogonal to PC2, is defined by moderate acousticness, high speechiness, and also high danceability, with predictors like low bpm, low energy, low instrumentalness, and low liveliness contrasting against PC2 predictors.

Above is a scatter plot of all clusters after performing k-Means clustering our data. Once applying our hue, 'streams', the graph shows a relationship between number of streams and the combination of predictors that are associated with these streams.

As expected, these clusters are overlapping, so we can infer these predictors are not entirely dependent on each other, however, we can make out clusters like yellow that are slightly aggregated towards a lower PC2 and moderately PC1. When mapping ‘streams’ as a hue onto this scatter plot, we can infer that songs positively associated with qualities from PC1 and less associated with qualities from PC2 are moderately strong predictors for a song’s popularity based on streams. Therefore, we can conclude that songs associated with a combination of high danceability, valence, energy, and speechiness while being less associated with a combination of high acousticness, speechiness, and danceability are capable of predicting popularity based on number of streams. However, based on the aggregation of clusters, it is also important to note that these predictors do not strongly correlate with the number of streams a song could get. This cluster plot indicates that different combinations of predictors are not indicative of a song’s popularity, indicating that there exists diversity of musical qualities in popular songs, even if songs with high energy, valence, danceability, and speech are likely to do better in garnering numbers.

Now that we know what collection of predictors can infer about popularity, how do individual predictors compare?

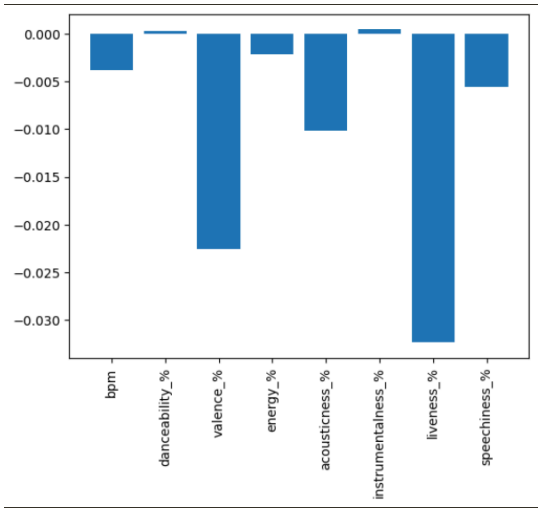
Random-Forest Regression



After splitting the data into testing and training sets, running a Random Forest Regressor, fitting the data, and making predictions on the test dataset, the accuracy score returned was about -0.0382. When including 'released_year' as one of the predictors, the accuracy score increases to about 0.507, further proving how impactful this confounding variable is on the model. Without 'released_year', there appears to be some significant underfitting happening with the data, which is expected given how weakly correlated the predictors are to the number of streams, further proving that predictors are entirely independent and are not strongly influential on popularity. However, when calculating permutation importance (shown above in bar graph), bpm (beats per minute) appears to be the strongest individual predictor when predicting number of streams, with predictors like valence, speechiness, and danceability, all strong predictors identified in the cluster, have weaker predictive power when evaluated individually.

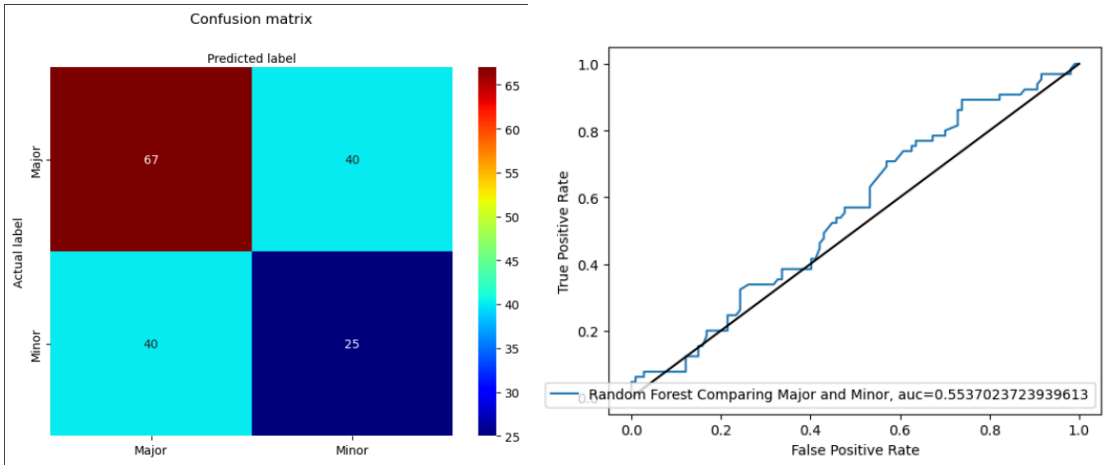
Question 2: What is the relationship between individual predictors and mode of a song?

Random Forest Classifier



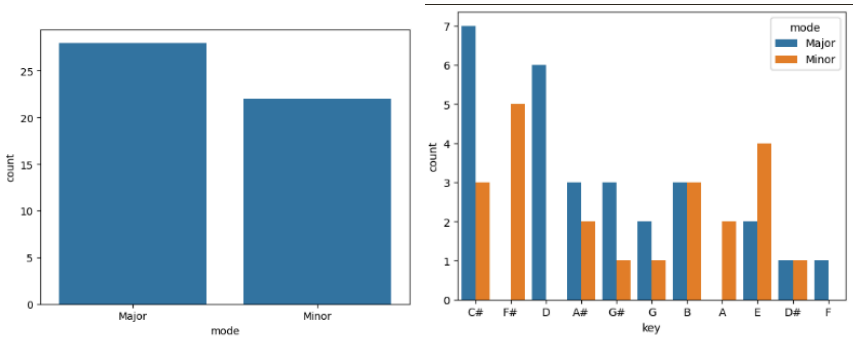
The y-value for this classifier is ‘mode’, Major or Minor, and the goal of this model is to identify whether it is possible for a song to either be identified as major or minor based on the predictors analyzed in the first question and whether bpm (beats per minute) or a different individual predictor has strong predictive power over this categorical variable. After splitting the dataset again into a different set of training and testing data, running a Random Forest Classifier, fitting the data, and making predictions about the test data (similar to steps taken during regression), the accuracy score returned is 0.535. This is considerably better than the accuracy score calculated for the regressor, which indicates that the data fits much better when evaluating these predictors based on mode. This data also indicates that there is still a weak correlation between predictors and mode, however.

The same trends are observed when running a confusion matrix and mapping the results onto a heat map and ROC curve. All three methods of evaluating accuracy indicate a close 50/50 chance that this classification model can generalize to other datasets.



Based on the performance of this Random Forest Classifier, we can conclude that the predictors we've previously analyzed do not help with predicting whether a song could be major or minor. While there are very, very miniscule predictors, such as danceability and instrumentalness, that do have a small degree of predictability, we can ultimately conclude that we cannot identify what individual variables are the most influential in determining the likely mode of a song.

However, when further exploring the data for trends regarding the most common key among the list of top songs, it can be inferred that C# Major is the key of most songs in this dataset, still implying that there is a preference for keys even when there isn't a definitive relationship between musical factors and streams or musical factors and mode.



Discussion

Do musical factors influence the popularity of a song to a certain degree, not necessarily. While there are a certain combination of related factors, such as danceability, valence, energy, and speech and individual predictors like bpm that can make moderately strong predictions about streams, the dataset ultimately proves that there is no one type of song that increase the likelihood of popularity: As previously mentioned, regardless of musical factors, any song is still guaranteed an equal likelihood of becoming popular. Questions like the relationship between predictors and modes make strong secondary questions that can lead to further investigation into the data, such as the relationship between streamed songs and keys based on the exploration of these categorical variables. While the numerical predictors explored may not generalize to other datasets, these results may also lead to future predictions and questions regarding musical trends and additional factors that could influence popularity of future top songs.