Predicting What Makes a Song Popular

Kat Dizon, Kirtana Krishnan, Regina Rabkina, Karen Vides

Abstract

This project aims to gain a deeper understanding of the ever-evolving landscape of popular music, especially in this age of digital music consumption and the widespread use of streaming platforms such as Spotify—one of the most influential players in the world of music listening and discovery. We hope to shed light on the music trends of 2023 by conducting an in-depth exploration and analysis of Spotify's top-charting songs for the year. The data is derived from the 'Most Streamed Spotify Songs 2023' dataset on Kaggle. A hyper-tuned Random Forest Regressor model will be used to predict the number of streams a song will have. The performance of the model will also be evaluated.

Introduction

Our interest in delving into the factors driving a song's popularity stems from a genuine love for music and a curiosity about what makes certain songs widely appealing. We recognized that there are clear, undeniable reasons why people are drawn to popular music, and we wanted to explore these reasons further. Additionally, we sought to understand the specific elements that resonate with us and attract us to particular pieces of music. This motivated our exploration of Spotify's 2023 top chart data, as we aimed to uncover the underlying characteristics that contribute to a song becoming a hit among listeners.

The problem to be explored involves an analysis of Spotify's weekly top chart data from 2023, to reveal the components and trends contributing to song popularity. Our project revolves around the examination of audio features to discover their correlations with a song's number of streams. This information can be valuable for various contributors in the music industry, including artists, record labels, and music fanatics, and can help answer questions about common audio features of popular songs, how different features correspond to a song's chart performance, seasonal trends in popular music, and patterns in artist popularity and their time spent on the charts.

Ultimately, we want to provide insights into what characteristics truly make a song popular, contributing to stakeholders' decision-making processes and broadening their understanding of the dynamic environment of the music industry.

To solve our problem of understanding the factors contributing to a song's popularity, our goal is to analyze the correlations between possible variables of popularity and the number of streams for the top songs of 2023. To do so, we must receive data from the Most Streamed Spotify Songs of 2023. We must also extract data specific to Spotify only, as we are fixating on this specific music streaming platform. We will clean the data and perform an exploratory data analysis to uncover trends, patterns, and relationships between variables to sort out features that are most influential in determining song popularity. We will then build a model that applies these features to predict the number of streams, and we will evaluate its performance. Overall, we hope to discover and understand what features musicians and other contributors in the music industry should emphasize to promote popularity and thus gain more song streams in the future.

Related Work

We found a project created by the user, Bahdir, on Kaggle, that was most similar to the motivations of our project (https://www.kaggle.com/code/bahadir23/spotify-songs-data-analysis-streaming-prediction). In their project, they performed data wrangling, advanced data analytics, and created a streaming prediction model, using duplicate variables for the Shazam columns to adjust for machine learning processing (they left these columns in their data set for the purpose of their analysis). This project served as a valuable inspiration for our project, as they utilized the same dataset to make predictions on song streams.

Methodology

Data Acquistion

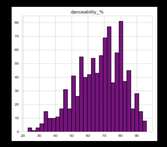
For our data modeling project, we opted for the 'Most Streamed Spotify Songs 2023' dataset on Kaggle as our primary data source. This dataset encompasses a collection of songs that have gained prominence on Spotify, and our focal point is the 'streams' variable, denoting the total number of times a song has been streamed. The objective is to discern the factors contributing to a song's popularity. The dataset was last updated in August 2023, providing a comprehensive snapshot of streaming trends up to that point. Our selection of explanatory variables includes crucial musical features such as bpm, danceability, valence, energy, acousticness, instrumentalness, and speechiness. These variables are instrumental in unraveling the intricate dynamics of what makes a song resonate with the audience.

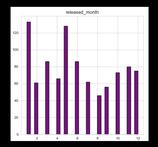
Data Preparation

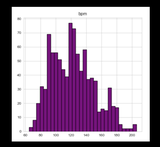
We began our EDA by observing basic characteristics about the dataset such as its range of released years spanned, as well as its variables and their datatypes. We modified our data by removing columns unrelated to Spotify, such as "in_apple_playlists", "in_apple_charts", "in_deezer_playlists","in_deezer_charts", and "in_shazam_charts". This is because we want our data analysis to only focus on Spotify data. The only column with missing values in our dataset was "key". We addressed this through imputation, replacing the missing values in "key" with "-". After this, we converted columns to their appropriate data types, making sure to account for errors by filling them with NaN. We also dropped rows with null values and encoded the "mode" and "key" categorial variables.

Data Analysis

We conducted our analysis by printing statistics for the variables and plotting histograms for each numerical variable. We observed trends in the data, such as that most songs in the data set are performed by a single artist. The most favored months for song releases tend to be 1 (January) and 5 (May), and most songs either are barely in playlists/charts or are in plenty of playlists/charts. There is a wide variety of bpm values, with a noticeable concentration around 80-120 bpm, indicating the popularity of a moderate tempo. In the remaining metrics, there are varying degrees of skewness. However, most, such as danceability and energy, showcase a broad distribution, highlighting diversified musical characteristics.







Which Model and Why

To select our model, we tested 3 regression algorithms: Decision Tree, kNNeighbors, and Random Forest. To train each model, we used features corresponding to the trends found in the data analysis stage, and set "streams" as the target variable. The data was partitioned into a training and test set, with the training set being 70% of the data and the test set being 30%. After partitioning, we fit the model to the data and hypertuned its parameters. We evaluated the performance of the model by calculating its r^2 value, as this would give us insight into the overall goodness of fit and how well the independent variables explain the variability in the dependent variable The model using the Random Forest Regression algorithm had the highest r^2 value of 0.81, indicating strong relationship between the variables. This made it best suited for predicting songs with the highest streams in our dataset. The parameters we hypertuned in our Random Forest Regression model were 'n_estimators', 'max_depth', 'min_samples_split', and 'min_samples_leaf'. After hypertuning, the r^2 value remained the same, indicating our model was already performing close to its maximum potential.

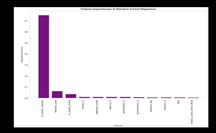
Results & Evaluation

Accuracy of Results: The best n_estimators value for our model was 50. This signifies that 50 decision trees in the random forest produced the best R-squared. The best R-squared was 0.81, which is close to 1, suggesting that our model explains a substantial amount of variance.

Findings: Overall, our model seems to perform well on both training and test sets with a mean training score of 0.964 and a mean test score of 0.798 when looking at the parameters corresponding to row 57. The difference between our mean train score and mean test score is relevant in our model and indicates there may be slight overfitting happening.

As seen in the chart below the most important feature in predicting a song's Spotify streams was whether a song was in Spotify Playlists or not. This feature had an importance score of about 0.78. The following most important features were released_year and in_spotify_charts with important scores of about 0.07 and 0.04. It is evident that whether a song is in Spotify playlists or not is the dominant feature in our model and has the strongest predictive power.

Predictions: When looking at our model's predictions there were some large difference between predicted streams and actual streams, we have noted that is likely due to outliers in our data and temporal trends in music.



Impacts

Through the adoption of our data modeling approach, our project holds the potential to profoundly impact the understanding and prediction of musical trends, ushering in a new era of insights and opportunities within the music industry. By leveraging the dataset on Spotify's top-charting songs for 2023, we aimed to refine and enhance the accuracy of predicting a song's popularity. This advancement has far-reaching implications for various stakeholders. First, artists and record labels stand to benefit from a more nuanced comprehension of the factors that contribute to a song's chart performance. With the ability to identify common audio features associated with popular songs, music creators can strategically tailor their compositions and promotional efforts, potentially increasing the likelihood of widespread recognition and success. This not only empowers individual artists but also contributes to the vibrancy and diversity of the overall music landscape. For music enthusiasts, the impact lies in an elevated and personalized music discovery experience. Through our analysis of trends in musical preferences and artist success, listeners can anticipate a curated selection of songs aligned with their tastes. This not only enhances the enjoyment of individual users but also fosters a deeper connection between artists and their audiences. Additionally, the data-driven insights generated by our modeling have the potential to inform industry professionals, streaming platforms, and even policymakers. Understanding the intricate dynamics of what makes a song popular can influence content curation, marketing strategies, and industry regulations, thereby shaping the future trajectory of the music ecosystem.

Conclusion

Our concerted effort to delve into the correlations between a song's features and a song's stream count, particularly within Spotify's top-streamed songs of 2023, has yielded illuminating insights for stakeholders within the music industry. Our comprehensive analysis revealed compelling findings, notably showcasing an impressive R-squared value of 0.81. This signifies the substantial variance explained by our model, underscoring its efficacy in capturing the nuances influencing a song's popularity. Our meticulous exploration pinpointed that the primary determinant of a song's stream count lies in its inclusion within Spotify playlists, with an importance score of about 0.78. Additionally, variables like the song's release year and presence in Spotify charts also exhibited moderate influence. While acknowledging the model's success, the slight disparity between the mean training (0.964) and test scores (0.798) suggests potential overfitting, necessitating strategies like regularization or model complexity reduction. Moving forward, exploring additional features related to Spotify playlists, refining model hyperparameters beyond Random Forests, and employing advanced validation techniques like k-fold cross-validation could enhance the model's predictive accuracy. Nonetheless, our goals were realized, and our findings provide invaluable guidance to musicians, labels, and industry enthusiasts, shedding light on the pivotal factors steering a song's popularity on Spotify. By highlighting the significance of playlist inclusion and other influential features, our work contributes to a nuanced understanding of what drives song streams, empowering decision-making processes and paving the way for strategic maneuvers within the dynamic realm of the music industry.