RHYTHMIC PULSE: TAKING A BEAT ON SONG STREAMS USING DATA SCIENCE

Sakshi Shastri
Department of Computer Science
Bowling Green State University
Bowling Green, Ohio, 43402

Kieran Delaney
Department of Computer Science
Bowling Green State University
Bowling Green, Ohio, 43402

Ryan Renken
Department of Data Science
Bowling Green State University
Bowling Green, Ohio, 43402

Introduction

The dynamics of the music industry have changed significantly as a result of the enormous growth of the streaming music space. This project undertakes a thorough investigation of streaming habits and predictive modeling approaches in an effort to comprehend the minute details that influence song popularity and consumption patterns on Spotify. Using a carefully selected dataset from Kaggle that includes songs published through July 2023, this project aims to identify the fundamental trends that determine song popularity and shape listeners' choices.

The primary objective of this analysis is to delve deep into the 2023 Spotify, Apple music, Deezer streaming trends for top songs. Utilizing a linear regression model, our aim is to forecast this year's total song streams by harnessing the power of diverse libraries and methodologies, including NumPy, Pandas, Seaborn, Matplotlib, and Scikit-learn models. The iterative testing approach adopted in this project involves the systematic evaluation of regression models through logical attribute combinations. By iteratively refining the models, we aim to decipher the crucial attributes that significantly impact song popularity and streaming metrics.

Our methodology revolves around meticulous data curation, exploratory data analysis, and the application of predictive modeling techniques. The Kaggle-sourced dataset serves as the cornerstone of our analysis, encompassing a comprehensive repository of songs released until July 2023. Leveraging this dataset, we apply rigorous data preprocessing techniques to clean and prepare the information for analysis. Subsequently, we utilize the power of linear regression models and polynomial models to predict the total song streams for the year, employing a range of libraries and methodologies to enhance the accuracy and interpretability of our predictions.

This project is extremely important for deepening our comprehension of the complex dynamics that exist inside the music industry. This investigation establishes the foundation for predictive modeling methodologies in predicting music consumption patterns, while also offering industry stakeholders useful information by identifying the factors that influence song popularity on Spotify. By being aware of these nuances, musicians, record companies, and streaming services may be better able to make judgments, plan promotions, and adjust to the changing preferences of music fans.

Problem Definition and Algorithm

Task Definition

In this study, it is our goal to conduct an exploratory analysis to determine the features to consider while predicting streaming counts of individual songs. The data in this CSV file is highly structured where each row represents a single song, and there are several describing features, including the track name, artist(s) name, number of artists, release date, and the cumulative number of times the song has been streamed on Spotify. There are also fields that measure the number of times the song has been added to a playlist or was present in the 'charts' of different streaming platforms– Spotify, Apple, Deezer, and Shazam (playlists are not available in Shazam, so only Shazam charts was included). On top of these metrics, the dataset also has features that describe the song itself, like the beats per minute (bpm), key, mode, and various percentages the coincide with the style of the music, including danceability, valence, energy, acousticness, instrumental, liveliness, and speechness. In other words, a song that has a fast paced beat and might be played at a dance club might have high energy and danceability percentages.

We found that there were several outliers in our data set that could cause disruption in our predictive modeling efforts. Instances where individual tracks were outside 3 standard deviations from the mean were removed from our models altogether.

In this project, we were particularly interested in the stream counts for songs in this data. To us, the raw total number of streams that a song has is a good indicator of the song's overall popularity. Regardless of the time period the song was released, it is available to users the same as any other song. If stream count is the most important metric for music creators on Spotify, are there any features that could help predict popularity for individual songs? This is the question we hope to explore.

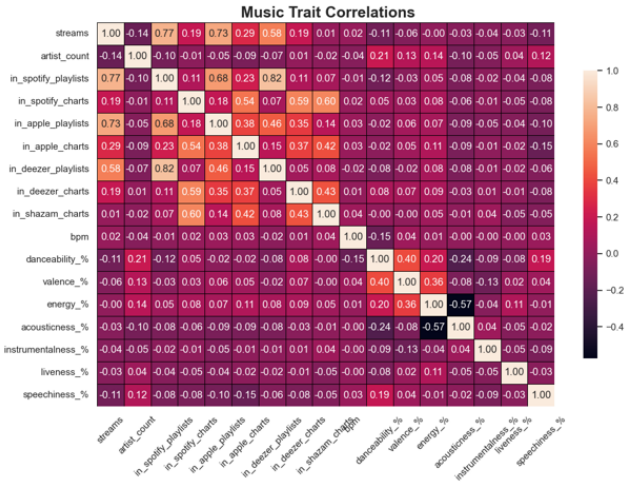

Figure 1: Distribution of Streams and Playlist input variables before and after log transformation.

Algorithm Definition

We initially planned on pursuing a multiple regression model with hopes of finding features that have a strong correlation with the stream counts of songs. After analyzing a correlation matrix, to our surprise, none of the music-style-percentages had any strong relationship with the number of streams (see Figure 1 for a visual). The only features that had a seemingly strong linear relationship were the playlist metrics–how many times a song gets added to a playlist on the three different platforms, respectively. Once we did some scatter plots to see the trends graphically, we would move to building the linear regression model using the Linear Regression function from the sklearn linear model library. We fit this model to predict the total number of streams for individual songs based on the number of times the song was added to playlists on Spotify, Apple, and Deezer. When the model is fit to these features, an equation is created that represents a line that best fits and represents the data we are looking. This equation assigns various weights to the different independent variables (playlists fields). For any new data points that we are trying to predict based on our model, we multiple each of the features by the corresponding weights to get a predicted output of streams.

Since our data is heavily skewed, we can perform some kind of transformation to help each of the features have a more normal distribution. The variations of each of the features in this data are very dramatic, so we wanted to use a method that accounted for drastic differences in the data. For this use case, we chose to use a logarithmic (log) transformation where we take the log of all the values in each of our model's fields. After transformation, we could perform the same linear model steps as before and expect more reasonable results. See Figure 2 for histograms showing the variable distributions before and after the log transformation.

After we complete our analysis using a linear regression model, we can also consider implementing a polynomial regression model. The idea here would be that using a polynomial model could account for non-linear rates of change throughout the graph. Mathematically, we say that the rate of change between incremental input variables is not consistent. Graphically, the shape of a scatter plot may show a trend with curvature rather than a straight line.

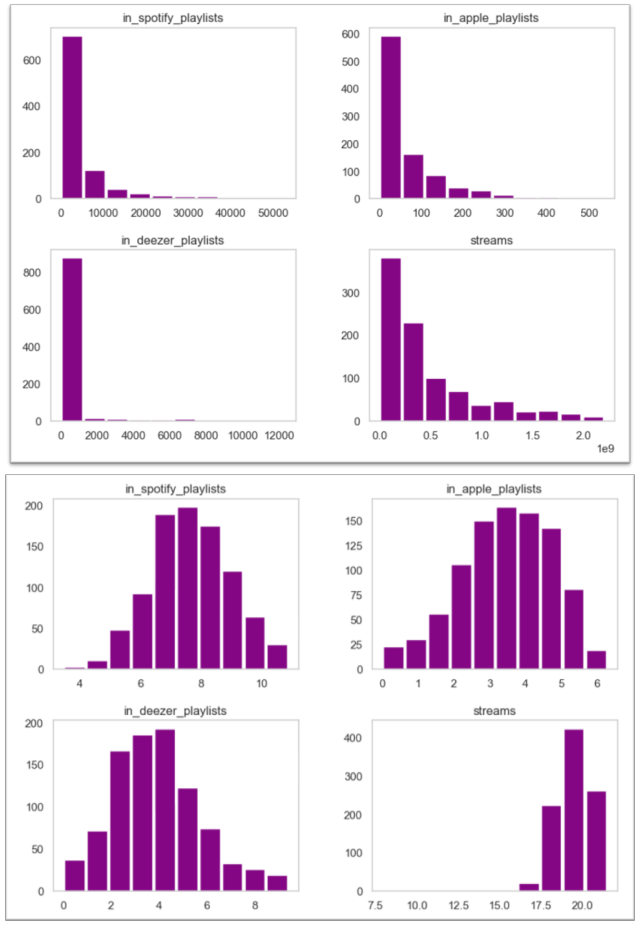

Figure 2: Distribution of Streams and Playlist input variables before and after log transformation.

Experimental Evaluation

Methodology

We chose to utilize a k-fold cross validation approach to calculating the Root Mean Squared Error (RMSE) and the Root Mean Absolute Error (RMAE) for evaluating the performance of our models. Our goal is to predict the number of streams a song should realize based on the number of times it was added to playlists on three popular streaming platforms. These evaluation metrics are indicators for our models' accuracy.

First, we set our parameters for how we wanted to partition and validate our results. Using a k-fold method, our algorithm randomly extracted 20 percent of our records to hold out for testing–we call this the validation set. We trained our model on the remaining 80 percent and used that model to predict the results of the validation set. In doing

so, we can use the trained model to predict stream counts of songs we already have an output value for. Once we have a set of predictions for this validation set of data, we can calculate the RMSE and RMAE to show the overall accuracy of predictions on the unseen data points. This process is considered one single fold; we chose to iterate this process 4 additional times to compile the evaluation metrics a total of 5 times using different samples of data to validate across.

Similarly, if we wanted to use the same process for a polynomial model, we would also like to test various degrees of polynomials to know what fits our data the best. So, we used the same k-fold process with 5 folds. However, we chose to take the average evaluation metrics of each of the 5 folds and iterate the process for polynomial degrees 1 through 10. In other words, we wanted to see which polynomial degree would have the lowest RMSE and RMAE values; so, we processed the k-fold method for each degree, took the average values of the folds, and repeated the process for the next degree. Once completed, we would be left with a data frame consisting of the 5-fold average RMSE and RMAE metrics of each polynomial degree 1 through 10.

Results

After completing each of the aforementioned steps in our methodology, we got increasingly better results. In other words, using a log transformation to fit the polynomial regression model helped lower both of the error evaluation metrics. See Table 1 and Table 2 for the K-fold results using the raw and transformed data.

For the raw-data fitting, the mean RMSE over 5 folds was just over 275 million and mean RMAE value of 13,916.

Table 1: K-Fold cross validation results of liner regression model for raw music data.

R^2	RMSE	RMAE
0.602072	3.599420e+08	15986.961469
0.661575	2.220942e+08	12670.646195
0.643093	2.730765e+08	14331.136412
0.544251	3.436387e+08	14810.845568
0.409864	1.785353e+08	11780.918980

Furthermore, using the same log-transformed playlist features and streaming counts, the polynomial model we chose ended up resulting with even better results than the linear model. See Table 3 for the cross validation values of each degree polynomial 1 through 10. Here, we found that using a polynomial of degree 3 would have the lowest error metric values and fit our use case the best.

Table 2: K-Fold cross validation results of liner regression model for log transformed music data.

	R^2	RMSE	RMAE
0	0.654430	0.763740	0.767716
1	0.693804	0.553570	0.659671
2	0.606631	0.589860	0.678899
3	0.665868	0.529639	0.651837
4	0.112938	0.644246	0.713392

Table 3: Average K-Fold cross validation results of polynomial models with degrees 1 through 10.

Degs	R^2	RMSE	RMAE
1	0.112938	0.644246	0.713392
2	0.253516	0.590997	0.686475
3	0.281530	0.579801	0.678459
4	-1.996362	1.184056	0.926885
5	0.029061	0.674017	0.710168
6	-0.545757	0.850444	0.744572
7	-0.622175	0.871212	0.768523
8	-51.851124	4.972813	1.518621
9	-193.058405	9.528867	1.574795
10	-7116.758165	57.709412	3.705862

After fitting a polynomial with degree of 3 to our data, we calculated the overall RMSE, RMAE, and graphed the predicted results against each of the playlist features. The average RMSE of this model was just under 232 million, which is 44 million smaller than the RMSE of the original linear regression model we built; similarly, the RMAE was 12,395, which is roughly 1,500 smaller than the original linear regression model output.

It is clear that there are still some variations between predicted and actual values on a 1:1 comparison; however, the overall shape of the predicted values mimics the shape of the actual data more accurately than a straight line. See Figures 3-5 for the resulting graphs of actual streaming totals versus predicted values based on each of the playlist features used to build the model.

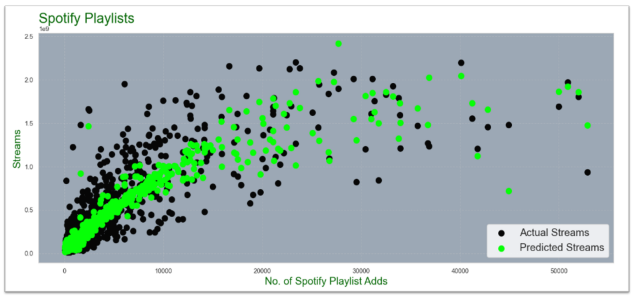

Figure 3: Scatter plot of Actual vs. Predicted streaming counts for songs based on the number of times they were added to a playlist on Spotify.

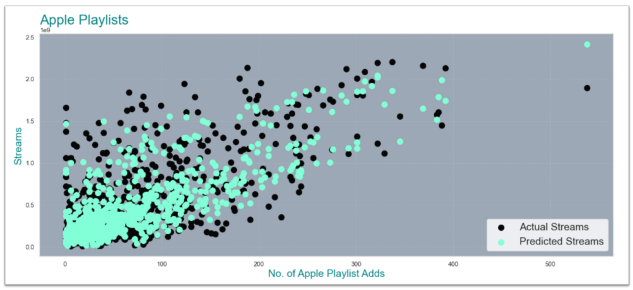

Figure 4: Scatter plot of Actual vs. Predicted streaming counts for songs based on the number of times they were added to a playlist on Apple.

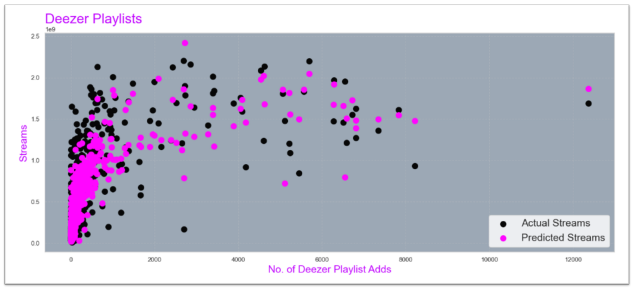

Figure 5: Scatter plot of Actual vs. Predicted streaming counts for songs based on the number of times they were added to a playlist on Deezer.

Discussion

The study presents an investigation into the dynamics of the music industry, focusing on the growing popularity of streaming music that has significantly changed the music industry. Using a dataset from Kaggle which encompasses songs released until July 2023, the research employs linear regression and polynomial models to understand the complex details influencing song popularity and consumption patterns on Spotify. The analysis delves into 2023 Spotify streaming trends, aiming to predict total song streams for the year. Surprisingly, the musical style percentages showed no strong correlation with the number of streams, leading the study to prioritize playlist metrics. Using a polynomial model, particularly of degree 3, proved to be more effective than a linear regression model in predicting song performance metrics, showing its enhanced accuracy. Platform specific dynamics were observed, with variations in the correlation between playlist adds and streams on Spotify, Apple Music, and Deezer. This study emphasizes the important role in playlists in driving song popularity and offers insights for industry stakeholders to tailor strategies and optimize song research and engagement. The study points toward the evolving landscape of music streaming analytics, integrating advanced predictive models, data sources, and personalized recommendation systems for a more insightful understanding of music consumption patterns in the future.

Conclusion and Future Work

The analysis delving into playlist adds and subsequent streams across Spotify, Apple Music, and Deezer unveiled key insights. A strong positive correlation between playlist adds and streams underscores their pivotal role in driving song popularity. The employed polynomial model adeptly captures this relationship, revealing its efficacy in predicting song performance metrics. Across platforms, distinctions emerged: Spotify and Deezer exhibit a trend where stream increases slow at higher playlist adds, while Apple Music demonstrates a more linear correlation, suggesting predictable growth. Leveraging the model's predictive power, stakeholders in the music industry gain foresight into potential streaming metrics, aiding strategic decisions. Understanding these nuanced relationships is crucial, empowering industry players to tailor marketing strategies and playlist placements to platform-specific user behaviors, ultimately optimizing song reach and engagement. This analysis highlights the varied nature of playlist-to-stream dynamics, offering actionable insights for industry stakeholders to navigate and capitalize on these distinct platform behaviors.

The integration of advanced predictive models, enriched data sources, and personalized recommendation systems signifies the evolving landscape of music streaming analytics. These future endeavors aim to deepen our understanding of user behaviors, refine predictive accuracy, and empower stakeholders in the music industry to tailor their strategies, curate content, and engage with audiences more effectively. These efforts pave the way for a more insightful and adaptive approach towards understanding music consumption patterns, thereby fostering a more enriching experience for both artists and listeners alike.

In the future, further research could delve into refining predictive models by exploring additional features that may contribute to a more in-depth understanding of music consumption patterns. Diversifying datasets and incorporating real time data could enhance the accuracy and the applicability of the models. Additionally, investigating the impact of song popularity may provide valuable insights. Furthermore, exploring advanced machine learning techniques and algorithmic approaches could contribute to a more accurate predictive model. The integration of user feedback and preferences into the analysis could refine personalized recommendation systems, creating a tailored music streaming experience. Lastly, expanding the study to include a broader range of streaming platforms could offer a more holistic view of the evolving landscape of music consumption.

To reference our specific code, please access our Jupyter Notebook file (.ipynb) in our public GitHub repository here: `https://github.com/SakshiShastri12/RHYTHMIC-PULSE-TAKING-A-BEAT-ON-SONG-STREAMS/tree/main`