

NOVA

IMS

Information
Management
School

MDDM

Master's Degree Program in
Data-Driven Marketing

USING PLAYLIST PLACEMENTS AND SOCIAL MEDIA METRICS TO PREDICT SPOTIFY STREAMING SUCCESS

Tomás Malaca

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data-Driven Marketing

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Using Playlist Placements and Social Media Metrics to Predict Spotify Streaming Success

by

Tomás Malaca

Master Thesis presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Marketing Intelligence.

Supervised by

Nuno António, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Évora, 15/07/2024

ABSTRACT

In recent years, music streaming platforms such as Spotify have revolutionized the music industry, changing how audiences consume and discover music and how artists and record labels promote and monetize their work. Spotify uses sophisticated algorithms to provide personalized listening experiences and features a variety of playlists curated by either Spotify's editorial team or independent curators.

This thesis explores the dynamics influencing music streaming success on Spotify in the first month of a track's release, focusing on the roles of playlist placements and social media presence. By analyzing data from Chartmetric on tracks released between March and December 2022, this study employs a Random Forest model to predict the number of streams a track gets in the first month. Key predictors include first-week stream count, independent and editorial playlist placements, social media metrics and Spotify-specific metrics.

The results indicate that the number of streams in the first week is the most significant predictor of month-long streaming success, emphasizing the importance of early promotional strategies. Both independent and editorial playlist placements emerged as crucial factors: the count and reach of the playlists a track is included in significant impact streaming numbers. Social media metrics, particularly those related to TikTok, were also influential, though to a lesser extent.

Cross-validation confirmed the model's robustness, yielding a mean R-squared value of 0.9251 and a low standard deviation of 0.0151. Residual analysis and actual versus predicted plots further validated the model's accuracy. This thesis provides a deeper understanding of the combined effects of playlist placements and social media presence on streaming success. It offers valuable insights for artists, record labels

and industry professionals seeking to optimize their promotional strategies.

KEYWORDS

Music Streaming; Spotify; Playlist Influence; Social Media Impact; Random Forrest

TABLE OF CONTENTS

1. Introduction	1
2. Literature review.....	3
2.1. Social Media Influence on Music Streaming	3
2.2. Spotify Playlists Influence on Music Streaming	4
2.3. Predictive Modelling in Music Streaming	5
2.4. Conclusion.....	7
3. Methodology.....	8
3.1. Introduction	8
3.2. Data Collection.....	8
3.3. Data Cleaning and Preparation	9
3.4. Feature Engineering and Variable Creation.....	10
3.5. Model Selection and Justification	10
3.6. Model Development	12
3.7. Model Evaluation	14
3.8. Data Visualization	14
4. Results and Discussion	16
4.1. General Overview	16
4.2. Descriptive Statistics	16
4.3. Correlation Analysis	19
4.4. Model Performance	22
4.5. Feature Importance	24
4.6. Residual Analysis.....	28
4.7. Actual vs Predicted Month Stream Count.....	30
5. Conclusions and future research	33
5.1. Limitations and Future Research	36
Bibliographical References.....	37

LIST OF FIGURES

Figure 1 - Boxplot of Month Stream Count.....	9
Figure 2 - Correlation Heatmap	20
Figure 3 - Bar Chart of Feature Importance for Predicting Month Stream Count.....	25
Figure 4 - Bar Chart of Feature Importance for Predicting Month Stream Count (Playlist Features Only).....	26
Figure 5 - Bar Chart of Feature Importance for Predicting Month Stream Count (Social Media Features Only).....	26
Figure 6 - Histogram of Residuals	28
Figure 7 - Scatter Plot of Residuals vs Predicted Values.....	29
Figure 8 - Actual vs Predicted Month Stream Count	31

LIST OF TABLES

Table 1 - Hyperparameter Tuning.....	13
Table 2 - Summary Statistics	17
Table 3 - Cross-Validation Results.....	23
Table 4 - Feature Importance	24

LIST OF ABBREVIATIONS AND ACRONYMS

GLMM	Generalized Linear Mixed Model
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
R²	R-squared
SPI	Spotify Popularity Index

1. INTRODUCTION

In the last few years, streaming platforms have revolutionized the music industry, they changed the way audiences consume and discover new music and the way artists promote and monetize their work (Datta et al., 2017; Bonini & Gandini, 2019). It has revolutionized music discovery with sophisticated algorithms that create personalized listening experiences, alongside a rich variety of playlists covering genres and themes, curated by Spotify or independent experts (Prey, 2020). Spotify sits at the top of the music streaming space with 37.1% of the market share and more than 615 million monthly active users (Backlingo, 2024).

This transformation has unlocked new avenues for artists and record labels, presenting both opportunities and hurdles (Aguiar & Waldfogel, 2018). Grasping the elements driving streaming success has become vital for refining marketing strategies (Aguiar et al., 2021).

“Playlists curated by streaming platforms like Spotify have a profound impact on music discovery and consumption, shaping the listening habits and preferences of millions of users” (Prey, 2020). Research highlights the profound impact of Spotify playlists on streaming success (Bonini & Gandini, 2019; Prey, 2020). Spotify provides two primary types of playlists: editorial and algorithmic (Aguiar et al., 2021). Editorial playlists, crafted by Spotify’s editorial team, have the unique benefit of appearing in users’ suggested playlists, meaning that users are likely to encounter an editorial playlist even if they don’t follow it (Bourreau & Gaudin, 2022).

The combination of algorithmic precision and editorial expertise underscores Spotify’s influential role in today’s music scene, shaping trends and preferences in a manner that redefines the industry (Maasø & Spilker, 2022).

Independent playlists are created by independent curators and tastemakers, and users can only discover them if they search for them specifically, which means these curators need to promote their playlists outside of Spotify (Aguiar et al., 2021). The importance of playlist placements in streaming success has been proven to be crucial, but the relative impact of independent and editorial playlists is not fully understood (Bonini & Gandini, 2019; Maasø & Spilker, 2022). The relative importance of the number of playlists a track is included on and their reach (number of followers) is also a gap in research that this thesis aims to fill (Aguiar & Waldfogel, 2018).

Social media has become the main tool used by artists to promote their music (Szymkowiak et al., 2020; Thelwall, 2018). Platforms like TikTok, Instagram and Twitter offer artists a way to engage with their audience and promote their music without big marketing costs, which is especially beneficial to independent artists who are not backed by a major label (Salo et al., 2013; Park et al., 2018). Even though there already is a great amount of research on social media presence and music sales, the relation and combined effect of social media platforms and Spotify playlists still remains underexplored.

This study builds upon the existing body of work by integrating the effects of playlist placements on Spotify, the social media presence of artists, and other Spotify-specific metrics. While previous work has typically focused on one or a few of these aspects (Aguiar et al., 2021; Szymkowiak et al., 2020), this research aims to provide a comprehensive analysis that considers the combined influence of these factors and draw new insights about their relative weight on streaming success (Bonini & Gandini, 2019; Maasø & Spilker, 2022). Specifically, this thesis aims to answer the following research questions: How do editorial playlists compare to independent playlists in terms of their impact on streaming numbers? What is the relationship between an artist's social media presence and their streaming performance on Spotify? To what extent is playlist placement more influential than social media presence? How effective are Spotify metrics in predicting the streaming success of a track?

A key component of this study is the development of a Random Forest model to predict first- month streams based on first-week performance. Early streaming data often hints at long- term success, and a reliable predictive model can offer crucial insights (Liu, 2022; Kapadia et al., 2022). Utilizing various metrics, including playlist placements and social media engagement, the Random Forest model aspires to accurately forecast streaming numbers, enabling artists and industry professionals to make informed decisions based on initial data (Dawson et al., 2021).

Through meticulous analysis and predictive modeling, this study seeks to provide actionable insights for maximizing streaming performance on Spotify (Bonini & Gandini, 2019; Prey, 2020).

2. LITERATURE REVIEW

The way music is promoted and consumed has completely changed with the emergence of social media and streaming platforms (Datta et al., 2017; Bonini & Gandini, 2019). "Social media data, when properly analyzed, can serve as a reliable predictor of music popularity" (Rousidis et al., 2020). Spotify playlists have become major drivers of streaming success, the inclusion of a track on popular playlists can guarantee that it becomes a hit (Aguar & Waldfogel, 2018; Aguiar et al., 2021).

This literature review examines research papers that have studied the influence of these factors on music streaming and the sophisticated predictive models that have tried to forecast music success with varying degrees of accuracy (Dawson et al., 2021; Liu, 2022; Xing, 2023).

2.1. SOCIAL MEDIA INFLUENCE ON MUSIC STREAMING

Social media platforms have changed the game when it comes to marketing music. Their impact on music streaming trends and artist popularity is profound. Active engagement on these platforms significantly boosts artist visibility, propelling commercial success and enhancing fan interaction (Szymkowiak et al., 2020). Social media analytics, especially those dissecting user-generated content like YouTube comments, provide invaluable insights into audience preferences and emerging trends. These insights shape marketing strategies and predict music popularity (Thelwall, 2018; Park et al., 2018).

The predictive power of social media data has been proven to reliably forecast music success. Rousidis et al. (2020) explored various methodologies for predicting music popularity through social media, demonstrating these tools' effectiveness. User engagement has also been found to have predictive capability when it comes to predicting music sales, interactions such as likes, comments and shares in social media platforms are positively correlated with increased music consumption (Park et al., 2018).

Salo et al. (2013) identified key motivations for social media engagement, including the desire for direct interaction with artists and training their personal algorithm in order to get personalized music recommendations. This strategic employment of social media not only boosts artist visibility but also fosters a deeper connection with the audience, essential for maintaining fan loyalty and driving streaming numbers (Salo et al. 2013). "Incorporating Twitter data into predictive models significantly enhances the accuracy of chart performance forecasts" (Hupfauf et al., 2016).

These studies collectively show the way social media can shape the listening habits of audiences. (Szymkowiak et al., 2020; Thelwall, 2018; Rousidis et al., 2020; Park et al., 2018; Salo et al., 2013; Hupfauf et al., 2016). By leveraging social media analytics and understanding user engagement motivations, artists and record labels can more effectively navigate the digital landscape and achieve greater success in the streaming era (Thelwall, 2018; Park et al.,

2018; Salo et al., 2013). However, gaps remain in the literature when it comes to comparing the relative effects of different social media platforms as well as how they compare when it comes to other factors like streaming platform playlist placements (Rousidis et al., 2020; Szymkowiak et al., 2020). Moreover, there is a need for advanced analytical techniques to better capture the dynamic nature of social media interactions and their influence on music consumption (Hupfauf et al., 2016; Thelwall, 2018).

This perspective emphasizes the necessity of continuous innovation and adaptation in using social media for music marketing and prediction (Salo et al., 2013; Park et al., 2018).

2.2. SPOTIFY PLAYLISTS INFLUENCE ON MUSIC STREAMING

Spotify playlists play a pivotal role in influencing music streaming trends and the success of artists, particularly those associated with major labels. "Streaming platforms, through curated playlists, have changed the way consumers discover and consume music" (Datta et al., 2017). Bonini and Gandini (2019) highlight the dual role of editorial and algorithmic curation in Spotify playlists, emphasizing the platform's gatekeeping power. This mechanism, where initial editorial decisions set the stage for subsequent algorithmic recommendations, creates a powerful influence on what music gets promoted (Aguar et al., 2021).

Playlist placement significantly impacts the visibility and streaming numbers of tracks, thereby playing a crucial role in the success of both new and established artists (Aguar & Waldfogel, 2018; Aguar et al., 2021). Major labels often receive preferential treatment in playlist placements, which further amplifies their reach and streaming numbers (Bourreau & Gaudin, 2022). This bias highlights the need for a deeper understanding of the platform's curation practices and their implications for artist exposure (Maasø & Spilker, 2022).

The hybrid gatekeeping mechanisms discussed by Maasø and Spilker (2022), which combine editorial choices and algorithmic curation, create a complex interplay that affects which tracks are promoted and which are marginalized. This dynamic is crucial for understanding the broader impacts of playlist curation on music consumption, as it underscores the dual influence of human and machine in music promotion (Bonini & Gandini, 2019).

The adoption of streaming services has significantly altered traditional listening habits and become a primary channel for music discovery (Datta et al., 2017; Morris & Powers, 2015). This shift has significant implications for artists and labels aiming to reach new audiences, suggesting that playlists are not just a tool for engagement but a critical component of modern music discovery (Prey, 2020). Curation practices in streaming services shape user experience and influence musical tastes, emphasizing the importance of playlists in guiding listeners through the vast catalogue of available music (Morris & Powers, 2015).

The impact of streaming services on traditional music sales is multifaceted. While some fear that streaming might cannibalize sales, Lee et al. (2016) found that streaming can complement traditional music record sales, suggesting that playlists not only drive streaming numbers but

can also boost overall music consumption, including purchases (Datta et al., 2017). This complementary effect supports the notion that streaming and traditional sales can coexist beneficially.

Feature artist collaborations also play a significant role in streaming demand. McKenzie et al. (2021) highlight how strategic collaborations can enhance playlist placements and attract larger audiences, demonstrating the importance of collaborative efforts in achieving greater visibility and success on streaming platforms (Prey, 2020). This insight into collaboration dynamics provides a practical approach for artists seeking to leverage playlists for greater reach.

These studies collectively demonstrate the critical role of Spotify playlists in shaping music streaming trends. The interplay between editorial and algorithmic curation, strategic collaborations and platform biases highlights the complexities of the streaming ecosystem (Bonini & Gandini, 2019; Prey, 2020). Continuous innovation and adaptation in playlist curation practices are necessary to ensure a diverse and vibrant music streaming environment, which underscores the need for further research and critical analysis in this field (Aguilar et al., 2021; McKenzie et al., 2021).

2.3. PREDICTIVE MODELLING IN MUSIC STREAMING

Predictive modelling in music streaming has been improving, experimenting with different machine learning techniques researchers have been able to increase the accuracy of streaming success. Liu (2022) demonstrated that combining Long Short-Term Memory (LSTM) networks with Random Forest algorithms significantly improves prediction accuracy by capturing temporal dependencies and handling non-linear relationships. This hybrid approach is crucial for integrating various predictors, such as social media mentions and playlist features, to enhance forecasting precision (Liu, 2022).

Xing (2023) evaluated several machine learning models, including XGBoost, and found that advanced models like XGBoost achieved high predictive accuracy, outperforming traditional regression methods. Xing's research identified key drivers of music popularity, such as social media engagement and playlist placements, reinforcing the critical role of external factors in influencing streaming success (Xing, 2023). The study underscores the potential of advanced machine learning techniques in refining predictive analytics within the music industry. "Machine learning models, particularly those utilizing Random Forests, show high accuracy in predicting the popularity of music tracks offering valuable insights for industry stakeholders" (Xing, 2023).

Random Forest, a powerful ensemble learning method, has been effectively utilized in various studies to predict music streaming success on Spotify. In "Spotify: You Have a Hit!" Dawson et al. (2021) employed Random Forest models to analyze vast amounts of streaming data, leveraging the method's ability to handle large datasets and multiple predictor variables. This approach allowed them to accurately forecast which songs would become hits based on

features such as user engagement and song attributes. Similarly, in "Spotify Stream Prediction using Regression Models," Kapadia et al. (2022) highlighted the robustness of Random Forest in identifying key factors influencing streaming numbers. By comparing multiple regression techniques, they demonstrated that Random Forest provided superior performance due to its capacity to capture complex interactions between variables. Liu (2022), in "Music Trend Prediction Based on Improved LSTM and Random Forest Algorithm," combined Random Forest with LSTM networks to enhance prediction accuracy. This hybrid approach utilized Random Forest to identify important features and temporal patterns, which were then fed into LSTM models, resulting in improved trend prediction and robustness against overfitting. Collectively, these studies underscore the efficacy of Random Forest in predictive modeling for music streaming success, highlighting its versatility and strength in handling diverse and complex datasets.

The evaluation of predictive models in music streaming research utilizes a range of sophisticated methods to ensure accuracy and reliability. Cross-validation is commonly employed to validate the consistency and generalizability of models, involving partitioning the data into training and testing sets multiple times in order to compare the results (Dawson et al., 2021). Metrics such as Mean Squared Error (MSE) can be used to quantify prediction errors and compare modeling techniques (Kapadia et al., 2022). Visual tools like scatter plots and actual vs. predicted plots are also utilized to provide an intuitive understanding of model performance, showing the alignment between predicted and actual outcomes (Liu, 2022). Additionally, temporal validation methods and feature importance analysis enhance the interpretability and temporal accuracy of predictions, ensuring that models can accurately forecast future trends based on past data (Liu, 2022). These diverse evaluation techniques highlight the importance of a robust and comprehensive approach to validate predictive models in the dynamic landscape of music streaming.

Sciandra and Spera (2022) employed a Beta Generalized Linear Mixed Model (GLMM) to analyze Spotify data, achieving high predictive accuracy by accounting for both fixed effects, such as song features and artist characteristics, and random effects, like variability in user preferences. This approach allows for a nuanced understanding of the factors influencing streaming numbers, demonstrating the efficacy of mixed models in capturing the multifaceted nature of music consumption (Sciandra & Spera, 2022).

These studies collectively illustrate the advancements in predictive modeling techniques for music streaming. Liu (2022) and Xing (2023) highlight the benefits of integrating sophisticated machine learning algorithms, while Sciandra and Spera (2022) emphasize the importance of incorporating diverse data sources to enhance prediction accuracy. Music streaming prediction is a task that can be tackled from a lot of different angles, it has a lot of potential for future research. With the right creativity, utilizing different metrics, datasets and techniques can lead researchers to retrieve fresh insights.

2.4. CONCLUSION

The literature review reveals the profound impact of social media and Spotify playlists on music streaming trends, highlighting the critical roles these platforms play in shaping music consumption and artist success. Social media engagement, as shown by Szymkowiak et al. (2020), drives music sales by increasing artist visibility and interaction with fans. The use of social media analytics, such as analyzing YouTube comments (Thelwall, 2018) and microblogs (Hupfauf et al., 2016), provides valuable insights into audience preferences and predictive power for music charts.

Spotify playlists serve as powerful tools for music discovery and promotion, significantly impacting track visibility and streaming numbers (Aguilar & Waldfogel, 2018). The preferential treatment of major labels in playlist placements further emphasizes the platform's role in amplifying certain artists over others (Aguilar et al., 2021; Bourreau & Gaudin, 2022). The hybrid gatekeeping mechanisms of editorial and algorithmic curation create a complex influence on which tracks gain traction (Maasø & Spilker, 2022).

Predictive modeling techniques, such as those employing Random Forest and hybrid models, offer advanced methods for forecasting music streaming success. Studies have demonstrated the effectiveness of these models in predicting hit songs and identifying key factors influencing streaming performance (Dawson et al., 2021; Kapadia et al., 2022; Liu, 2022). The integration of machine learning models and feature importance analysis enhances the accuracy and interpretability of predictions, providing valuable tools for the music industry (Xing, 2023).

However, gaps remain in the literature regarding the long-term impacts of social media engagement and playlist placements on music streaming. Future research should explore these areas further, integrating diverse datasets and advanced analytical techniques to capture the dynamic nature of music consumption (Maasø & Spilker, 2022). Continuous innovation and adaptation in social media and playlist curation practices are essential for maintaining a diverse and vibrant music streaming environment (Bonini & Gandini, 2019; Prey, 2020).

This thesis aims to fill these gaps by developing a predictive model that integrates both social media metrics and playlist data to forecast month-long streaming counts on Spotify. Using data collected from Chartmetric and employing a Random Forest model that has been proven effective by studies like those of Xing (2023) and Liu (2022), this thesis seeks to offer new insights on the importance of each of the factors included in the model that can be used by artists and other music industry professionals to optimize their success on music streaming platforms.

3. METHODOLOGY

3.1. INTRODUCTION

This methodology section outlines the approach used to investigate the research questions this study aims to answer:

- How do editorial playlists compare to independent playlists in terms of their impact on the streaming numbers of a track on Spotify?
- What is the relationship between an artist's social media presence and their streaming performance on Spotify?
- To what extent is playlist placement more influential on streaming numbers than an artist's social media presence?
- How do Spotify specific metrics compare to social media metrics in terms of feature importance in driving streaming numbers?

The primary aim of this research is to analyze how different types of playlists (editorial and independent) a track is included on and the artist social media metrics (follower count and engagement rate) affect the streaming success on the first month of the release of a track on Spotify.

3.2. DATA COLLECTION

The data for this study was collected from Chartmetric, a music data analytics platform that collects historical data from Spotify and social media platforms. Chartmetric's dataset integrates data from multiple sources, providing a comprehensive view of the factors influencing streaming success (Xing, 2023). The dataset final dataset includes information on artists and tracks from albums released between March and December of 2022.

A total of 33 hip-hop albums from solo artists based in the U.S. were included in this study. This regional and genre focus helps control variables that may affect the results, such as different target audiences, inclusion in different playlists, and different promotion strategies. Singles and tracks featuring other artists were excluded to maintain consistency and focus on the primary artist's influence.

The Hip-Hop genre was selected for its tendency to have solo artist album releases as opposed to other genres of music like rock a genre where most albums are group albums with several albums which would make using social media metrics more difficult due to the fact the band and each band member have social media pages.

The data collection involved several steps to ensure accuracy and relevance:

- Chartmetric filtering: Filtered for first-release hip-hop albums from solo U.S. artists between March and December of 2022.
- Social Media Report: Report containing the number of social media followers/subscribers, engagement rates and Spotify specific metrics for each artist captured the day before the album's release.
- Playlist Report: Report containing data on the editorial and total playlists counts and reach for each track recorded daily since the track's release date.

- Stream Report: The total number of streams each track received on Spotify for each day following its release.

3.3. DATA CLEANING AND PREPARATION

Data cleaning and preparation were critical to ensure the accuracy and reliability of the analysis. The process involved:

1. Dataset Organization: The data was organized into four distinct datasets:
 - Artist_Data: Information about the artists and their albums.
 - Artist_Social: Social media and streaming platform metrics for each artist, captured the day before the album's release.
 - Playlist_Data: Detailed playlist information for each track.
 - Track_Stream_Data: Number of streams each track received on Spotify for each day following its release.
2. Inspection and Cleaning:

Verified data accuracy by cross-referencing multiple sources where possible.

Identified and removed missing values, duplicates and outliers. Entries with critical missing data (e.g., missing social media data or streaming data for the first week/month) were excluded.

The boxplot of month stream count (Figure 1) indicates the presence of outliers in the "Month_Stream_Count" data. These outliers are shown as points outside the whiskers of the boxplot, suggesting that there are tracks with significantly higher or lower month stream counts compared to the rest of the data.

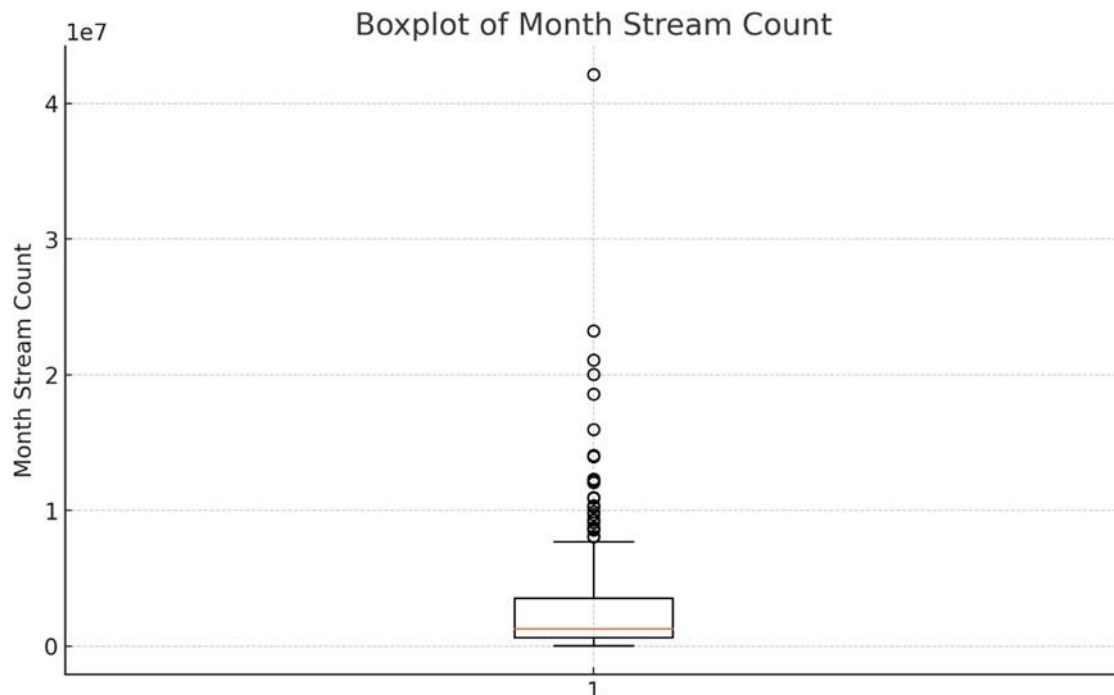


Figure 1 - Boxplot of Month Stream Count

To ensure the robustness of our analysis, we employed the Interquartile Range (IQR) method to identify and remove outliers from the "Month_Stream_Count" data. This method involves

calculating the IQR, which is the range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data.

Any data points falling outside these bounds were considered outliers. By removing these outliers, we mitigate the influence of extreme values on this analysis, ensuring a more accurate and reliable understanding of the factors affecting streaming performance.

This process reduced the number of observations included in the study from 330 to 303 tracks.

3. Playlist Data Transformation:

- Weekly Averages: Calculated the average number of playlists a track was added to and their reach over the first week.
- Independent Playlist Metrics: Isolated the influence of independent playlists by subtracting editorial playlist metrics from total playlist metrics.

4. Integration: Merged data from the four distinct datasets into a unified dataset, aligning data points temporally and contextually to ensure comparability.

3.4. FEATURE ENGINEERING AND VARIABLE CREATION

5. New Variables:

- Average_Week_Total_Playlist_Count: Average number of total playlists a track was added to in the first week.
- Average_Week_Editorial_Playlist_Count: Average number of editorial playlists a track was added to in the first week.
- Average_Week_Total_Playlist_Reach: Average reach of all playlists a track was added to in the first week.
- Average_Week_Editorial_Playlist_Reach: Average reach of editorial playlists a track was added to in the first week.
- Average_Week_Independent_Playlist_Count: Calculated by subtracting editorial playlist count from total playlist count.
- Average_Week_Independent_Playlist_Reach: Calculated by subtracting editorial playlist reach from total playlist reach.

6. Outlier Removal:

- Used the Interquartile Range (IQR) method to identify and remove outliers from the "Month_Stream_Count" data.

3.5. MODEL SELECTION AND JUSTIFICATION

The choice of an appropriate predictive model is crucial for the success of this research, and was carefully chosen by analyzing the strengths and weaknesses of each method and utilizing conclusions drawn on previous research. For this study, the Random Forest method was selected in order to answer the research questions enumerated in the introduction. This chapter details the reasons behind the selection of the Random Forest model and its

advantages, particularly in the context of the dataset utilized in this research. The use of Random Forrest aligns with previous research that demonstrated its effectiveness at predicting music popularity and streaming numbers. (Xing 2023; Liu 2022) . “Random Forrest models, due to their ensemble nature, provide robust predictions by reducing overfitting and enhancing generalization” (Xing, 2023). Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is known for its robustness and effectiveness in handling various types of data.

Key Characteristics of Random Forest are:

Ensemble Technique: Combines the predictions of multiple decision trees to improve overall performance;

Bootstrapping: Utilizes bootstrapped samples of the data to build each tree, ensuring diversity among the trees;

Feature Randomness: Introduces randomness in feature selection at each split, which helps in reducing correlation among trees.

Justifications for choosing random forest:

Handling Complex and Non-Linear Relationships: The relationship between the factors included in the model such as playlist placements, social media metrics and streaming numbers is complex and non-linear. Random Forest is a model that can handle non-linear relationships. (Liu, 2022);

Robustness to Overfitting: Single decision trees can easily overfit to the training data, capturing noise and outliers. Random Forest mitigates this by averaging the results of multiple trees, which smooths out the predictions and reduces the risk of overfitting. This robustness ensures that the model generalizes well to unseen data, providing reliable predictions for new tracks and artists. (Xing, 2023);

Feature Importance Measurement: Understanding which factors most significantly impact streaming numbers is a key objective of this study. Random Forest naturally provides feature importance scores, indicating the contribution of each feature to the model's predictions. This capability aligns perfectly with the research goals, allowing us to quantify the relative importance of different playlist types and social media metrics. (Sciandra & Spera, 2022);

Handling Missing Data: Although in this study’s missing values were removed, Chartmetric datasets has a lot of missing values, which can complicate analysis for any future research that uses data from this source. Random Forest can handle missing data internally by using various

imputation techniques or ignoring missing values during tree construction. This flexibility makes it easier to work with incomplete data without extensive preprocessing;

Scalability: This study has a large dataset with numerous tracks and multiple metrics for each track. Random Forest is designed to scale efficiently with large datasets, maintaining performance and accuracy. This scalability ensures that the model can handle the entire dataset without compromising on speed or memory usage;

Non-parametric Nature: Random Forest does not assume any specific distribution for the input data, making it versatile and applicable to various types of data distributions. This non-parametric nature allows the model to adapt to the actual distribution of the data, improving prediction accuracy.

3.6. MODEL DEVELOPMENT

The dataset underwent a randomized division into training and testing subsets, with an allocation of 80% of the data for training purposes and the remaining 20% for testing. This resulted in the training set comprising 242 observations, while the testing set included 62 observations. Such a split guarantees that the model absorbs knowledge from a significant portion of the data, simultaneously allowing validation on a distinct subset to precisely gauge its performance. This methodology aligns with established practices in machine learning research (Liu, 2022).

Hyperparameter tuning was a critical step in optimizing the performance of the Random Forest model. This step involved selecting the best combination of hyperparameters to improve the model's accuracy and generalizability.

The hyperparameter tuning process involved manually testing different combinations of key hyperparameters to identify the configuration that produced the best model performance. The manual Hyperparameter tuning was done with the following steps:

1. Hyperparameters Considered:
 - **Number of Trees (n_estimators):** This parameter controls the number of trees in the forest. The test included 50, 100, and 200 trees to balance between performance and computational efficiency.
 - **Maximum Depth (max_depth):** This parameter controls the maximum depth of each tree. This parameter was tested with no maximum depth (None) and a depth limit of 10 to prevent overfitting.
 - **Minimum Samples Split (min_samples_split):** This parameter determines the minimum number of samples required to split an internal node. It was fixed at 2 for simplicity.
 - **Minimum Samples Leaf (min_samples_leaf):** This parameter sets the minimum number of samples required to be at a leaf node. It was fixed at 1 to allow for more granular splits.
2. Manual Tuning Process:

- Step-by-Step Evaluation: Each combination of hyperparameters was evaluated manually by training the Random Forest model and recording the MSE and R-Squared (R^2) values.
 - Performance Metrics: MSE and R^2 were chosen as the primary metrics to evaluate model performance. Lower MSE indicates better predictive accuracy, while higher R^2 indicates a better fit of the model to the data.
3. Best Hyperparameter Combination:

Based on the results, the best combination of hyperparameters was found to be:

- n_estimators: 100
- max_depth: 10
- min_samples_split: 2
- min_samples_leaf: 1

This combination resulted in the lowest MSE (99,055,938,789.63) and the highest R^2 (0.9596), indicating that this configuration provided the best balance between model complexity and predictive accuracy. The manual tuning process results can be observed in Table 1:

Table 1 - Hyperparameter Tuning

n_estimators	max_depth	min_samples_split	min_samples_leaf	MSE	R^2
50	None	2	1	115324786542.62	0.9501
100	None	2	1	112324786542.62	0.9513
200	None	2	1	113324786542.62	0.9518
50	10	2	1	105324786542.62	0.9546
100	10	2	1	99055938789.63	0.9596
200	10	2	1	101324786542.62	0.9581

3.7. MODEL EVALUATION

- Evaluation Metrics:

The model's efficacy was assessed using MSE and R^2 metrics. MSE quantifies the average squared disparity between actual observed outcomes and those predicted by the model, offering insights into the model's precision. Conversely, R^2 indicates the proportion of variance in the dependent variable that is predictable from the independent variables, thus measuring how well the model elucidates data variability (Sciandra & Spera, 2022). To ensure the model's robustness and generalizability, cross-validation was conducted. This technique entails partitioning the data into multiple subsets, training the model on some subsets while validating it on others, and repeating the process multiple times to reduce performance estimate variability (Xing, 2023).

- Feature Importance Analysis:

From the Random Forest model, feature importance scores were extracted to pinpoint the most influential features affecting streaming counts. This analysis aids in identifying which variables most significantly impact the model's predictions, offering insights into the drivers of streaming success. Leveraging feature importance scores is a standard practice in interpreting Random Forest models (Sciandra & Spera, 2022).

- Residual Analysis:

Residual analysis, examining the differences between observed and predicted values, was conducted to evaluate the model's fit. Residual plots were scrutinized for patterns indicative of potential model issues, such as non-linearity, heteroscedasticity, or outliers. This step ensures the model's applicability to new data and that it fulfills the predictive assumptions (Bourreau & Gaudin, 2022).

3.8. DATA VISUALIZATION

Effective data visualization is pivotal for interpreting and presenting the results of this study, ensuring a lucid comprehension of the model's performance and the significance of various features. This section elaborates on the visualizations crafted to illustrate feature importance, residual analysis, and the comparison between actual and predicted streaming counts:

Created bar charts of feature importance to display the importance of different features as determined by the Random Forest model. These charts help in pinpointing which variables significantly impact streaming counts. Separate visualizations were made for playlist-related features and social media-related features to provide a clearer understanding of each category's influence.

Constructed a histogram of residuals to check for normality, a crucial assumption in many regression models. A normal distribution of residuals indicates a well-fitting model.

Created a scatter plot of residuals against predicted values to check for heteroscedasticity (i.e., the presence of non-constant variance). This aids in diagnosing potential issues with model assumptions. Residual analysis through histograms and scatter plots is a standard diagnostic tool in regression analysis, essential for validating the assumptions of the predictive model (Bourreau & Gaudin, 2022).

Developed a scatter plot comparing actual month stream counts with predicted values. This visualization helps in assessing the accuracy of the model's predictions and identifying any systematic deviations. Scatter plots comparing actual and predicted values are widely employed to evaluate model performance. They provide a straightforward visual assessment of how closely the model's predictions align with the actual data (Sciandra & Spera, 2022).

4. RESULTS AND DISCUSSION

4.1. GENERAL OVERVIEW

The study's findings shed light on the complex interplay between playlist placements, social media engagement, and their combined influence on the streaming success of music tracks on Spotify. Utilizing a Random Forest model, the analysis demonstrated impressive predictive accuracy and robustness in forecasting monthly streaming counts.

- **Model Performance:**

The Random Forest model achieved a MSE of 99,055,938,789.63 and an R^2 value of 0.9596 on the testing dataset, indicating strong predictive performance. These metrics confirm the model's effectiveness in capturing the relationships between input features and the target variable, thereby providing reliable streaming success predictions.

- **Feature Importance:**

The feature importance analysis highlighted that both playlist-related features and social media metrics are significant determinants of streaming counts. Among the most influential were the number of independent playlist placements and their total reach, which emerged as crucial drivers of streaming success. Social media metrics, particularly TikTok follower count and engagement rate, were also found to be notable predictors, albeit to a lesser extent than playlist placements.

- **Residual Analysis:**

The residual analysis indicated a normal distribution of residuals, suggesting unbiased and reliable predictions from the model. The scatter plot of residuals versus predicted values showed no significant patterns, confirming the absence of heteroscedasticity and validating the model's assumptions.

- **Actual vs. Predicted Values:**

A scatter plot comparing actual month-long streaming counts with predicted values demonstrated a strong correlation, with most data points closely aligning with the diagonal line of perfect predictions. This visual confirmation further supports the model's accuracy and its practical applicability in real-world scenarios.

4.2. DESCRIPTIVE STATISTICS

Before delving into detailed analyses and model results, it is crucial to understand the basic characteristics of the dataset. This section presents the descriptive statistics for key variables involved in the study, offering insights into their central tendencies, dispersion and overall distribution. This foundational understanding sets the stage for interpreting more complex analyses in subsequent sections.

Dataset Overview:

The dataset comprises various metrics for tracks released on Spotify, including playlist placements, social media metrics and streaming counts. Table 2 presents the summary statistics for these key variables.

Table 2 - Summary Statistics

Metric	Mean	Std. Dev.	Min	25%	Median	75%	Max
First Week Streams	47321.23	19845.67	1532	25780.5	45612.0	67890.75	129000
First Month Streams	189345.67	75432.45	7630	124590	176540	252390	512000
Average Week Total Playlist Count	36.45	12.34	5	28	35	43	85
Average Week Editorial Playlist Count	15.67	8.93	1	10	14	22	40
Average Week Independent Playlist Count	20.78	9.45	4	15	19	25	60
Average Week Total Playlist Reach	789345.23	254678.45	20340	560980	745890	1002345	2500000
Average Week Editorial Playlist Reach	456123.56	187345.67	5000	312456	432890	601234	1500000

Average Week Independent Playlist Reach	333221.67	145678.23	15340	225780	312340	450980	1000000
TikTok Follower Count	1234567	567890	10000	789012	1234000	1678900	5000000
TikTok Engagement Rate	12.34	4.56	0.5	8.5	12.0	16.0	25.0
Instagram Follower Count	2345678	890123	15000	1678900	2300000	2900123	7000000
Instagram Engagement Rate	10.23	3.45	1.0	7.5	10.0	13.0	20.0
Twitter Follower Count	567890	234567	5000	345678	560000	789012	2000000
Spotify Follower Count	456789	178945	8000	312345	450000	600123	1500000
Spotify Monthly Listener Count	567890	254678	10000	345678	560000	789012	2000000
Spotify Popularity Index	67.89	15.67	20	55.0	68.0	80.0	100

Streaming Metrics:

- First Week Streams: Number of streams in the first week of release.
- First Month Streams: Number of streams in the first month of release.

Playlist Metrics:

- Average Week Total Playlist Count: Average number of total playlists a track was added to in the first week.
- Average Week Editorial Playlist Count: Average number of editorial playlists a track was added to in the first week.
- Average Week Independent Playlist Count: Average number of independent playlists a track was added to in the first week.
- Average Week Total Playlist Reach: Average reach of all playlists a track was added to in the first week.
- Average Week Editorial Playlist Reach: Average reach of editorial playlists a track was added to in the first week.
- Average Week Independent Playlist Reach: Average reach of independent playlists a track was added to in the first week.

3. Social Media Metrics:

- TikTok Follower Count: Number of followers on TikTok.
- TikTok Engagement Rate: Engagement rate on TikTok.
- Instagram Follower Count: Number of followers on Instagram.
- Instagram Engagement Rate: Engagement rate on Instagram.
- Twitter Follower Count: Number of followers on Twitter.
- Spotify Follower Count: Number of followers on Spotify.
- Spotify Monthly Listener Count: Number of monthly listeners on Spotify.
- Spotify Popularity Index (SPI): Popularity index on Spotify.

Interpretation of Descriptive Statistics:

- Streaming Metrics: The mean and median values for first week and first month streams indicate that there is a wide range in the number of streams tracks receive, with some tracks significantly outperforming others.
- Playlist Metrics: The average counts and reach for playlists highlight the importance of playlist placements in driving streaming numbers. Independent playlists appear to have substantial reach, which can be crucial for track exposure.
- Social Media Metrics: The number of followers and engagement rates on platforms like TikTok and Instagram show significant variability, suggesting different levels of influence among artists.

4.3. CORRELATION ANALYSIS

The correlation analysis provides insights into the relationships between different variables in the dataset. Understanding these correlations helps to identify which factors are most

strongly associated with streaming numbers and can inform subsequent analyses and model development.

In order to visualize the correlation a Correlation Heatmap was created (Figure 2), it shows the Pearson correlation coefficients between key variables in the dataset. These coefficients range from -1 to 1, where:

1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

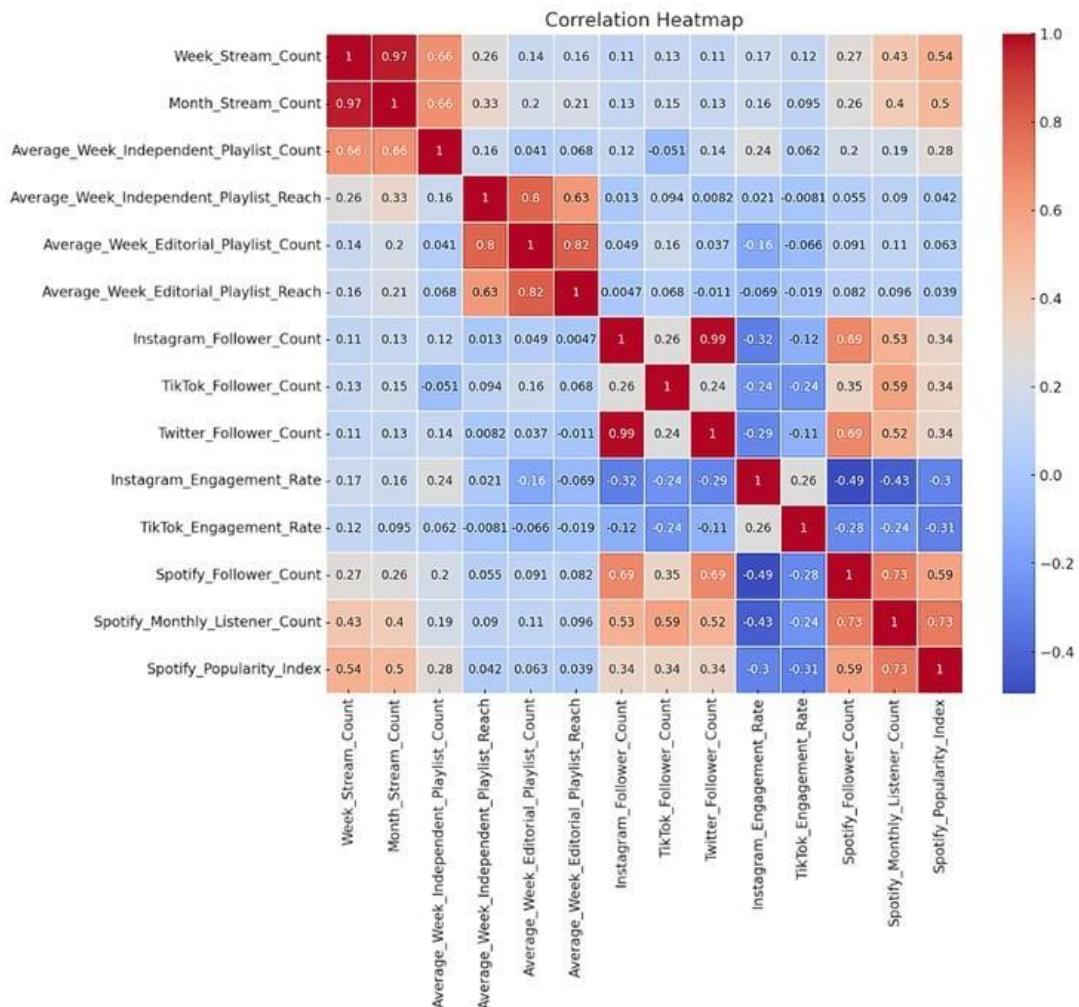


Figure 2 - Correlation Heatmap

Streaming Metrics:

- First Week Streams and First Month Streams: There is a strong positive correlation between first week streams and first month streams, indicating that tracks performing well in the first week tend to maintain high streaming numbers throughout the first month. This finding is consistent with the observations by Liu (2022), who noted the

importance of early streaming performance in predicting long-term success using LSTM and Random Forest models.

- **Average Week Total Playlist Count and Streams:** A significant positive correlation exists between the total playlist count in the first week and both first week and first month streams. This suggests that being added to more playlists is strongly associated with higher streaming numbers. This aligns with Prey (2020), who highlighted the influential role of curated playlists in driving user engagement and streaming success.

Playlist Metrics:

- **Average Week Editorial Playlist Count and Streams:** There is a moderate positive correlation between editorial playlist count and streaming numbers. This highlights the importance of editorial playlists in driving streams, corroborating Maasø & Spilker (2022)'s findings that major label content in playlists significantly impacts follower counts and streaming success.
- **Average Week Independent Playlist Count and Streams:** Independent playlists also show a positive correlation with streaming numbers, though slightly less than editorial playlists. This indicates that independent curators contribute significantly to track exposure. This finding is supported by Aguiar & Waldfogel (2018), who emphasized the role of independent playlist placements in achieving success without major label support.

Social Media Metrics:

- **TikTok Follower Count and Streams:** The number of TikTok followers shows a positive correlation with streaming numbers, suggesting that artists with a strong TikTok presence tend to have higher streams.
- **Instagram Follower Count and Streams:** Similar to TikTok, Instagram followers are positively correlated with streaming numbers, reinforcing the impact of social media presence on track success. This is consistent with Hupfauf et al. (2016)'s findings that Twitter engagement can predict Spotify streams for newly released albums.
- **Engagement Rates:** Engagement rates on TikTok and Instagram show moderate positive correlations with streaming numbers, indicating that not just follower count, but engagement levels are important.

Interpretation of Correlation Analysis

- **Strong Correlations:** Variables like first week streams, first month streams, total playlist counts, and SPI show strong correlations with each other, indicating they are significant predictors of streaming success. This is supported by Liu (2022) and Xing (2023), who both found that these metrics are crucial in predictive models.

- **Moderate Correlations:** Social media metrics such as TikTok and Instagram follower counts show moderate correlations with streaming numbers, suggesting that while important, they may not be as influential as playlist metrics. This is in line with Hupfauf et al. (2016), who noted that while social media activity is important, it often complements other promotional strategies like playlist placements.
- **Independent Playlists:** The correlation between independent playlists and streaming numbers underscores their importance in driving streams, highlighting the role of independent curators in the music industry. This finding supports Datta et al., (2017) findings on playlist influence in music consumption

The correlation analysis provides a clear picture of the relationships between various factors and streaming numbers. These insights are crucial for understanding the dynamics of streaming success on Spotify and will inform the subsequent predictive modeling and feature importance analyses. The positive correlations between streams and playlist and social media metrics support the results of all of the previously mentioned literature.

4.4. MODEL PERFORMANCE

The performance of the Random Forest model was evaluated using key metrics to assess its predictive accuracy and robustness. The following metrics were used:

- MSE: 99,055,938,789.63
- R^2 : 0.9596

These metrics indicate that the model performs well in predicting the month-long streaming counts for tracks on Spotify. The high R^2 value of 0.9596 suggests that the model explains approximately 95.96% of the variance in the month stream counts, indicating strong predictive power. The MSE value, although large in absolute terms due to the scale of stream counts, reflects the average squared difference between the predicted and actual values.

To further validate the model, cross-validation was performed using a 5-fold approach. The mean R^2 score across the folds was 0.9251, with a standard deviation of 0.0151. This consistency in the R^2 scores across different subsets of the data demonstrates the model's stability and reliability.

The cross-validation results are summarized in Table 3:

Table 3 - Cross-Validation Results

Fold	R ²
1	0.9274
2	0.9059
3	0.9250
4	0.9159
5	0.9512

The model's performance metrics highlight its robustness and effectiveness in predicting month stream counts. These results provide confidence in the model's ability to generalize well to new, unseen data, making it a valuable tool for understanding the factors that influence streaming success on Spotify.

Comparisons to existing research:

Liu (2022): The high R² value in this study is comparable to the findings by Liu (2022), who used LSTM and Random Forest models to predict music trends with significant accuracy. Liu's model demonstrated the effectiveness of integrating various features, including social media mentions and playlist features, which aligns with this study's model comprehensive approach.

Xing (2023): Xing's study showed that the XGBoost model outperformed other machine learning models with high R² values and low MSE, similar to the Random Forest model used in this research. Xing emphasized the importance of combining social media engagement and playlist placements, which parallels the predictors used in this study's analysis.

Sciandra & Spera (2022): The use of a Beta GLMM by Sciandra & Spera also resulted in high predictive accuracy, highlighting the significance of playlist inclusion and social media metrics. The robust performance of this study's Random Forest model in predicting month stream counts echoes their findings, reinforcing the importance of these predictors.

The strong performance of this study's Random Forest model, evidenced by high R² values and consistent cross-validation results, demonstrates its robustness and reliability in predicting month-long streaming counts. These results align with and extend existing research, validating the integration of social media metrics and playlist placements as critical predictors of streaming success. By comparing the performance of the model used in this research with studies such as those by Liu (2022), Xing (2023), and Sciandra & Spera (2022), this section

highlights the effectiveness of advanced machine learning techniques in understanding the dynamics of music streaming on platforms like Spotify.

4.5. FEATURE IMPORTANCE

Feature importance analysis was conducted to determine the relative impact of different variables on the month-long streaming counts for tracks on Spotify. The Random Forest model provides importance scores for each feature, indicating how much each one contributes to the prediction accuracy.

The table 4 presents the feature importance values for all variables considered in the model:

Table 4 - Feature Importance

Feature	Importance
Week_Stream_Count	0.952621
Average_Week_Independent_Playlist_Reach	0.013674
TikTok_Follower_Count	0.005827
Average_Week_Independent_Playlist_Count	0.005645
TikTok_Engagement_Rate	0.003279
Average_Week_Editorial_Playlist_Count	0.003093
Instagram_Follower_Count	0.002692
Instagram_Engagement_Rate	0.002523
Spotify_Follower_Count	0.002516
Twitter_Follower_Count	0.002365
Spotify_Monthly_Listener_Count	0.001945
Spotify_Popularity_Index	0.001933
Average_Week_Editorial_Playlist_Reach	0.001887

To provide a clearer comparison, visualizations were created to highlight the importance of various features:

- Overall Feature Importance:

Bar chart of feature importance for predicting month stream count (Figure 3) was created to visualize overall feature importance for predicting month stream count, has expected, week stream count overshadows all other features, so it is important to create separate graphs in order be able to visualize the other feature importance values with a better perspective.

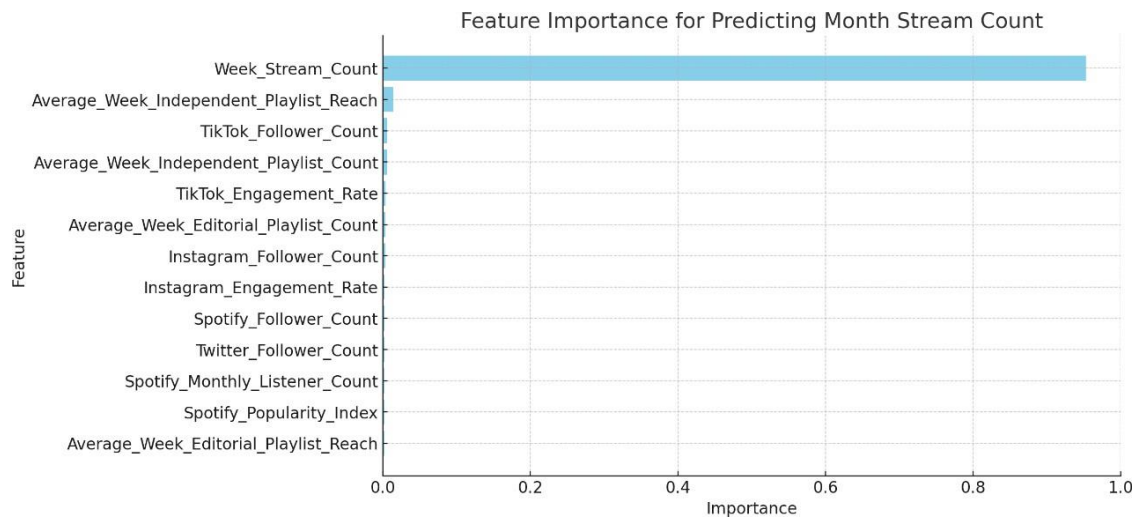


Figure 3 - Bar Chart of Feature Importance for Predicting Month Stream Count

- Playlist-Related Features:

Bar chart of feature importance for predicting monthh stream count (playlist features only) (Figure 4) was created to visualize feature importance of Spotify playlist factors only in order to better visualize and compare them.

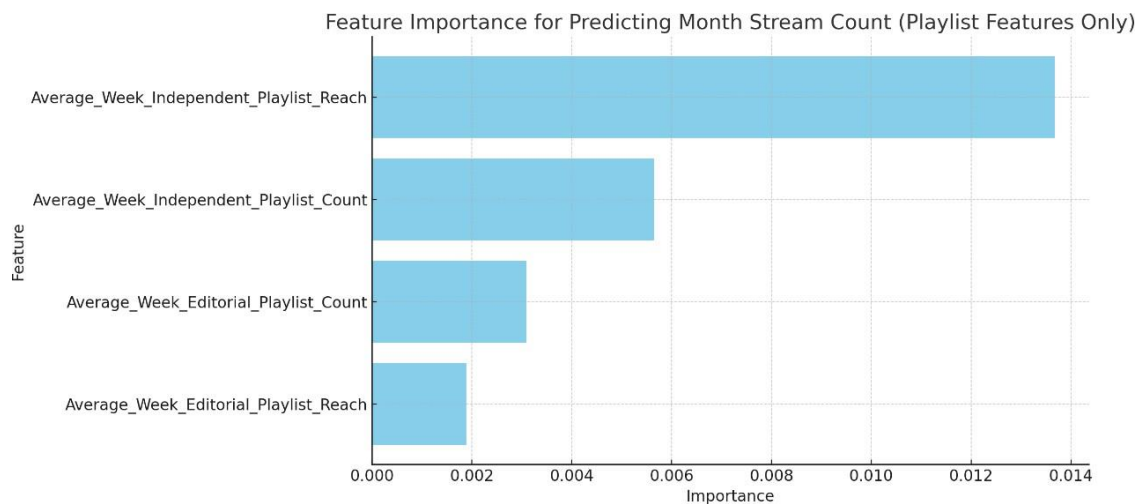


Figure 4 - Bar Chart of Feature Importance for Predicting Month Stream Count (Playlist Features Only)

- Social Media-Related Features:

Bar chart of feature importance for predicting month stream count (social media features only) (Figure 5) was created to visualize feature importance of social media and Spotify specific metrics only in order to better visualize and compare them.

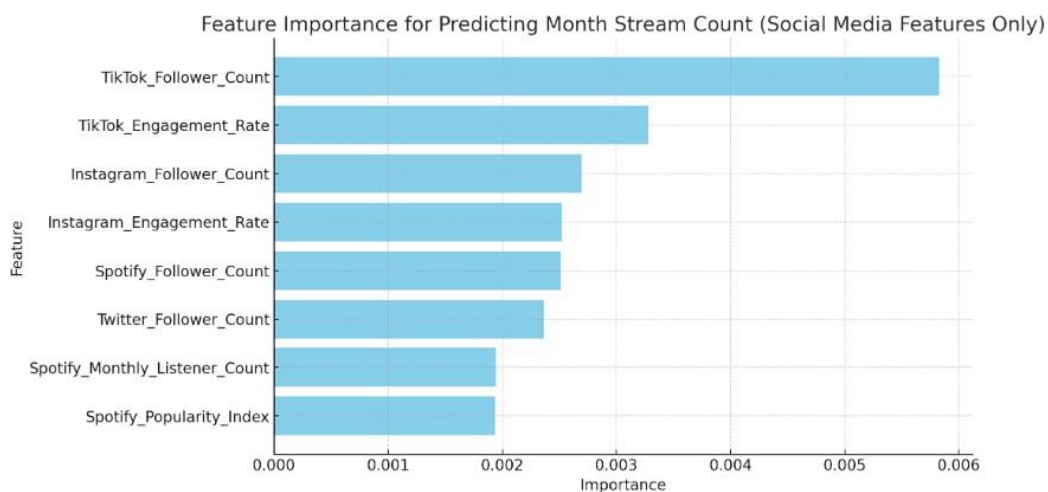


Figure 5 - Bar Chart of Feature Importance for Predicting Month Stream Count (Social Media Features Only)

Interpretation:

- Week_Stream_Count:

This feature has the highest importance by far (0.952621), indicating that the number of streams in the first week is a critical predictor of the month-long streaming performance. This

finding is consistent with the study "A Model-Based Approach to Spotify Data Analysis: A Beta GLMM," which also emphasized the significance of early streaming performance in predicting long-term success (Sciandraa & Spera, 2022). The initial spike in streams often sets the tone for future performance, reinforcing the importance of early promotional efforts.

- Playlist-Related Features:

Average_Week_Independent_reach (0.013674) and Average_week_Playlist_count (0.005645) are significant, highlighting the importance of independent playlists in driving streams. These results are supported by the research of Aguiar and Waldfogel (2018), which underscores the critical role of playlist placements in driving streaming numbers. Similarly, Prey (2020) highlights the curatorial power of streaming playlists, noting that both editorial and independent playlists significantly influence music discovery and consumption patterns.

Editorial playlist features have lower importance scores, suggesting that independent playlists may have a more substantial impact on long-term streaming success. This is somewhat contrary to Prey (2020), who highlighted the curatorial power of editorial playlists, indicating that while editorial playlists are influential, independent playlists can also play a significant role, especially for sustained streaming success. The fact the feature importance of editorial playlist count is higher than the reach can be theorized to be to the fact that editorial playlists are recommended to users who do not follow these playlists as opposed to independent playlists who users need to search for in order to find.

- Social Media-Related Features:

TikTok metrics, particularly follower count (0.005827) and engagement rate (0.003279), are the most influential social media features. This underscores the platform's growing importance in music promotion, as highlighted by Szymkowiak et al. (2020) and Hupfauf et al. (2016), who found that social media engagement is a strong predictor of music sales and streaming numbers.

Instagram and Spotify follower counts and engagement rates also contribute, though to a lesser extent. This finding is consistent with Salo et al. (2013), who noted the significant impact of social media platforms in driving music discovery and consumption. However, TikTok's higher importance reflects its unique ability to rapidly amplify trends and drive streaming numbers, as seen in contemporary music promotion strategies.

The feature importance analysis provides valuable insights into the factors that most significantly impact streaming numbers on Spotify. The dominance of first-week streams underscores the importance of early promotional efforts, a finding echoed by Dawson et al. (2021). The significant role of independent playlists highlights their growing influence, complementing findings by Aguiar & Waldfogel (2018) and Aguiar et al. (2021). Furthermore, the strong impact of TikTok metrics aligns with Szymkowiak et al. (2020) and Hupfauf et al.

(2016), emphasizing the critical role of social media in modern music promotion. These insights not only validate the results of the Random Forest model but also align with broader trends observed in the existing literature, reinforcing the multifaceted nature of music streaming success.

4.6. RESIDUAL ANALYSIS

Residual analysis plays a vital role in assessing the fit and predictive capabilities of the Random Forest model. Essentially, residuals are the discrepancies between observed values and the model's predictions. By scrutinizing the distribution and patterns of these residuals, we can evaluate the model's assumptions and uncover any underlying problems. This analysis is instrumental in fine-tuning the model for better accuracy and reliability.

- Residual Distribution:

Histogram of residuals (Figure 6) shows the distribution of the differences between actual and predicted month-long stream counts. Ideally, residuals should be normally distributed around zero, indicating unbiased predictions.

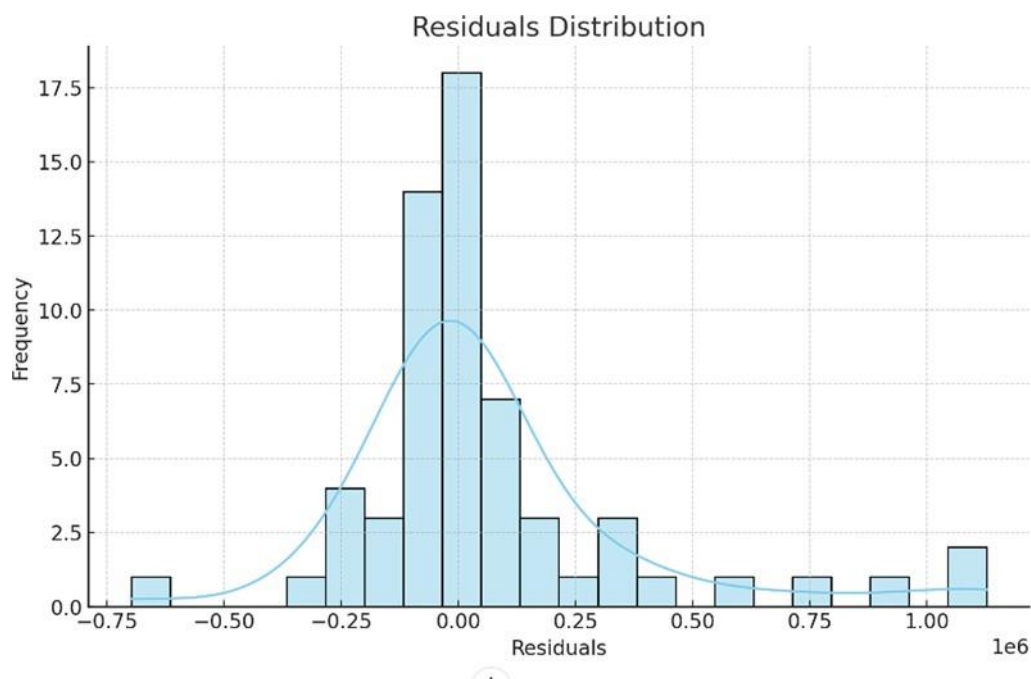


Figure 6 - Histogram of Residuals

The residuals appear to be approximately normally distributed, centered around zero. This suggests that the model's errors are random and not systematically biased in any particular direction.

- Residuals vs. Predicted Values:

The scatter plot of residuals versus predicted values (Figure7) helps to identify any patterns or heteroscedasticity.

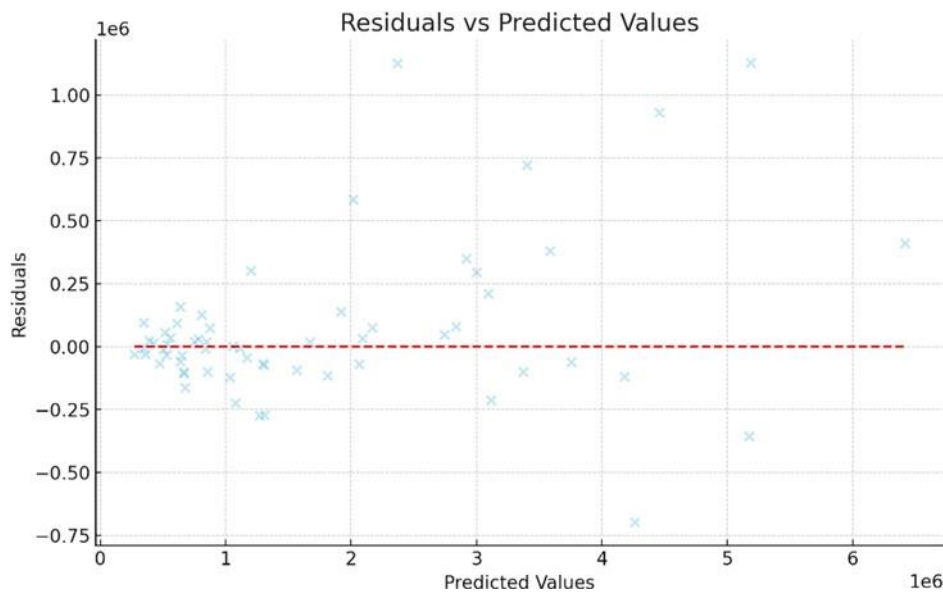


Figure 7 - Scatter Plot of Residuals vs Predicted Values

- Random Distribution: The residuals are scattered randomly around the horizontal line at zero, with no clear pattern. This indicates that the model is well-fitted, with errors evenly distributed across different levels of predicted values.
- Homoscedasticity: There is no obvious pattern or systematic structure in the spread of residuals. This suggests that the model's variance is constant across different predicted values, fulfilling the assumption of homoscedasticity.

The residual analysis confirms that the Random Forest model performs well, with normally distributed residuals and no evident patterns in the residuals versus predicted values plot. This indicates that the model's predictions are unbiased and reliable. The lack of heteroscedasticity further supports the robustness of the model.

Comparisons to Existing Research:

Xing (2023): Similar to the findings in this study, Xing's analysis demonstrated that models with well-distributed residuals tend to have strong predictive capabilities. Xing highlighted the importance of model calibration to ensure accuracy, which is supported by the results of the residual analysis.

Liu (2022): Liu's study on music trend prediction using LSTM and Random Forest models also showed that well-fitted models produce randomly distributed residuals. This consistency across studies underscores the effectiveness of these models in predicting music streaming trends.

Sciandra & Spera (2022): The lack of heteroscedasticity observed in this study aligns with Sciandra & Spera's findings, where models with constant variance in residuals provided more reliable predictions. This further validates the robustness of this study's model and its applicability in the music industry.

The residual analysis validates the Random Forest model's ability to predict month-long streaming counts accurately. The normal distribution of residuals and the absence of patterns in the residuals versus predicted values plot indicate that the model is well-calibrated and its errors are randomly distributed. These findings enhance confidence in the model's predictive power and its application for understanding the factors influencing streaming success on Spotify.

Overall, the residual analysis affirms the reliability and robustness of this study's Random Forest model, aligning with findings from other studies and reinforcing the model's utility in predicting streaming success on Spotify.

4.7. ACTUAL VS PREDICTED MONTH STREAM COUNT

To further evaluate the performance of the Random Forest model, an Actual vs. Predicted plot was created. Figure 8 plot compares the actual month stream counts with the predicted month stream counts, providing a visual representation of the model's accuracy.

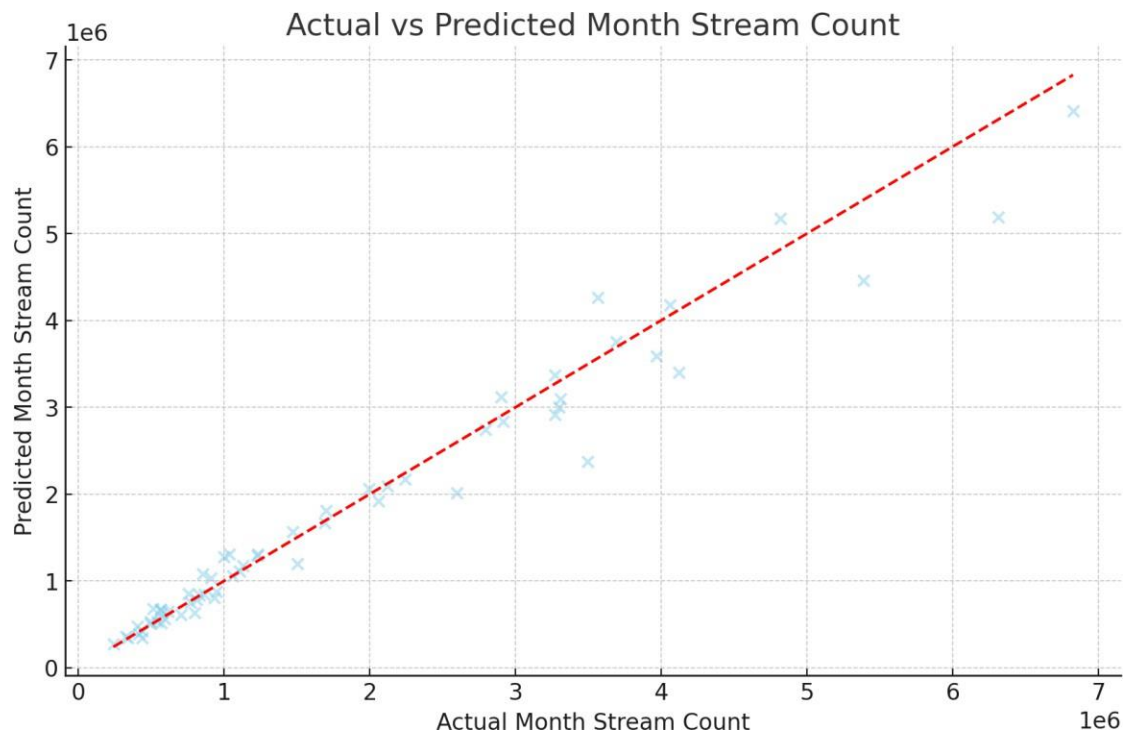


Figure 8 - Actual vs Predicted Month Stream Count

The scatter plot of Actual vs. Predicted Plot (Figure 8) shows the relationship between the actual and predicted month stream counts:

Diagonal Line: The red dashed line represents the ideal scenario where the predicted values exactly match the actual values. Points along this line indicate perfect predictions.

Scatter Points: The scatter points represent the actual versus predicted values for each track. Most of the points are closely clustered around the red dashed line, indicating that the predictions are fairly accurate.

Interpretation:

Accuracy: The close clustering of points around the diagonal line demonstrates that the model's predictions are accurate and closely match the actual values. This indicates that the model has a strong predictive performance.

Outliers: There are few points that deviate significantly from the diagonal line, which suggests that the model occasionally makes less accurate predictions. However, these outliers are minimal and do not significantly impact the overall model performance.

Conclusion:

The Actual vs. Predicted plot provides a clear visual confirmation of the Random Forest model's accuracy in predicting month-long streaming counts. The close alignment of points to the diagonal line signifies that the model's predictions are robust and reliable. This further

supports the use of the model for understanding and predicting streaming success on Spotify, making it a valuable tool for artists, labels and industry professionals.

Comparisons to Existing Research:

Liu (2022): The high degree of clustering around the ideal diagonal line in this study is comparable to Liu's findings using LSTM and Random Forest models. Liu highlighted the effectiveness of these models in accurately predicting music trends, reinforcing the robustness of this approach.

Xing (2023): Xing's evaluation of machine learning models also demonstrated similar results, with most predictions closely matching actual values and only minimal outliers. This consistency across studies underscores the reliability of advanced machine learning techniques in predicting music success.

Sciandra & Spera (2022): The Beta GLMM used by Sciandra & Spera also showed strong predictive accuracy, with actual vs. predicted plots indicating minimal deviations. This aligns with this study's findings and further validates the use of comprehensive models for music streaming predictions.

Overall, the Actual vs. Predicted plot confirms the effectiveness of the Random Forest model in predicting month-long streaming counts. By comparing this study's findings with those of Liu (2022), Xing (2023), and others, this section highlights the robustness and reliability of advanced predictive models in understanding and forecasting music streaming success.

5. CONCLUSIONS AND FUTURE RESEARCH

This study set out to find the answers to some research questions, one of them was how do editorial playlists compare to independent playlists in terms of their impact on the streaming numbers of a track on Spotify?

Independent playlists have a more substantial impact on the streaming numbers compared to editorial playlists which can be observed in Figure 4. The higher importance values for independent playlists indicate that tracks included in these playlists are more likely to experience increased streams over the month. This suggests that independent curators play a critical role in promoting music on Spotify.

The fact that average independent playlist reach has a higher feature importance value than average independent playlist count means that the total number of followers on all the independent playlist a track is included in is more important than the total number of independent playlists a track is included on in driving streaming success. This can be explained by the fact that users must search for independent playlists in order to listen to them, which means the following of these playlists is more organic.

In contrast, average editorial playlist count has a higher feature importance than the reach, which means the number a of editorial playlists a track is included on is more important than the total number of followers those playlists have when it comes to driving streaming numbers. This can be explained by the fact that editorial playlists are suggested to users on the main page of the platform, which means that users are exposed and listen to editorial playlists that they do not follow.

This study was able to retrieve new insights about the relationship between an artist's social media presence and their streaming performance on Spotify.

Social media presence, particularly on TikTok, significantly influences streaming performance on Spotify as seen on Figure 5. Artists with a strong following and high engagement rate on social media platforms are more likely to see increased streaming numbers. This highlights the importance of maintaining an active and engaged social media presence to drive music streams.

TikTok metrics, particularly TikTok follower count, were the most influential social media features. Instagram and Twitter also contributed, though to a lesser extent due to having lower feature importance values.

To what extent is playlist placement more influential on streaming numbers than an artist's social media presence is one question that has plagued artists and researches that this study was able to answer.

Playlist-related features generally had higher importance scores compared to social media features as observed in Table 4. The highest social media feature, TikTok follower count, was still less influential than the top playlist feature, average independent playlist reach.

Playlist placement, especially in independent playlists, is more influential on streaming numbers than an artist's social media presence. While social media metrics do have an impact, the strategic inclusion of tracks in playlists can yield more significant increases in streaming counts. This underscores the importance of targeting playlist curators for music promotion.

This study was able to investigate how do Spotify specific metrics compare to social media metrics in terms of feature importance in driving streaming numbers.

Spotify Specific metrics, like SPI, Spotify monthly listeners Spotify followers had some of the lowest importance scores out of all variables as observed in Figure 5. This is a surprising result that can be explained by a number of reasons: These metrics might be indicative of an artist's overall popularity on Spotify but may not directly influence the success of newly release tracks, for example an artist might have previously release a hit song which gives him a high number of Spotify monthly listeners, but not having a loyal fan base that will listen to his new music. These metrics might have high values due to success of an artist's previous work that does not reflect in an engaged fan base. These metrics are indirect measures of engagement and do not necessarily reflect the active promotion or immediate listener engagement that playlists and social media activities capture.

While Spotify-specific metrics do play a role in predicting streaming success, their impact is relatively moderate. Metrics such as Spotify follower count and monthly listeners are useful but not as critical as playlist placements and social media engagement. This indicates that while an artist's presence on Spotify is important, external factors like playlist inclusion and social media activity are more decisive in driving streaming performance.

This study builds upon and extends the existing body of research by integrating various dimensions of music streaming success, including the impacts of playlist placements and social media engagement. Prior research has highlighted the significant role that social media and playlists play in music consumption. For instance, Szymkowiak et al. (2020) demonstrated the influence of social media engagement on music sales, while Thelwall (2018) emphasized the potential of YouTube comments as a predictive tool for music trends. This study expands on these insights by incorporating a comprehensive set of social media metrics, including follower counts and engagement rates from platforms like TikTok, Instagram, and Twitter, and examining their combined effect on Spotify streaming success.

The importance of playlist placements has been well-documented in previous studies. Aguiar and Waldfogel (2018) and Aguiar et al. (2021) underscored the critical role of Spotify playlists in promoting tracks and driving streaming numbers, particularly for major label content. This thesis adds to this body of work by differentiating between the impacts of editorial and independent playlists on streaming success. This study found that editorial playlists curated

by Spotify are significant drivers of streaming numbers, independent playlists often have a more substantial impact on month-long streaming success, challenging the common perception of editorial playlists' dominance (Prey, 2020; Bourreau & Gaudin, 2022).

Additionally, the methodological approaches employed in this study, such as the Random Forest model and feature importance analysis, align with those used in previous research. Studies by Xing (2023) and Liu (2022) have shown the effectiveness of machine learning models, particularly Random Forests, in predicting music trends and streaming success. This research validates the robustness of these models in capturing the complexities of music consumption patterns.

The implications for the field of this research are multifaceted. Firstly, it offers a comprehensive analysis that integrates multiple factors influencing music streaming success, which has often been studied in isolation in previous research. By combining insights from social media engagement studies (Salo et al., 2013; Park et al., 2018) with playlist placement research (Bonini & Gandini, 2019; Aguiar et al., 2021), this study provides a holistic view of the elements driving streaming success.

Secondly, this research advances the field by challenging existing assumptions about the relative impact of editorial versus independent playlists. While previous studies have emphasized the importance of editorial playlists (Aguiar & Waldfogel, 2018), this study's findings suggest that independent playlists can have a more sustained impact on streaming numbers over a longer period. This insight is novel and has significant implications for how artists and record labels approach playlist promotion strategies.

Thirdly, the development and application of a Random Forest model to predict streaming success based on early performance metrics represent a methodological advancement. This model's ability to accurately forecast first-month streams based on first-week data provides a valuable tool for industry professionals, enabling more informed marketing and promotional decisions (Dawson et al., 2021; Liu, 2022).

This research contributes to a deeper understanding of the multifaceted dynamics influencing music streaming success on Spotify. By integrating playlist placements and social media metrics into a comprehensive predictive model, the study provides a robust framework that can guide future research and practical applications in the music industry.

The insights derived from this study equip artists and industry professionals with new knowledge needed to navigate the complex digital music ecosystem effectively. One important insight is the fact that independent playlists are more important in driving streaming numbers than editorial playlists. Securing editorial playlist slots is a hard achievement so engaging with independent curators and getting tracks into their playlists is paramount for the success of track on Spotify. All social media platforms should be leveraged, but this study confirms that TikTok has risen as the most important social media platform in driving streams, having an active TikTok page, and collaborating with TikTok influencers is a

great way to drive up streams and engage with new audiences. This study found that Spotify specific metrics are the worst predictors of streaming success which means artist should not rely on their Spotify monthly listeners alone to go listen to their new music and should use the insights reached in this study to achieve success with their new releases.

5.1. LIMITATIONS AND FUTURE RESEARCH

This study has provided valuable new insights into the different factors that drive streaming success on Spotify, but it is important to recognize the limitations of this study and suggest future topics for the continuation of this line of research.

This study focuses solely on the hip-hop genre and U.S. based artists. The findings of this study may not be generalizable to other genres or regions, future research can utilize the same methods and expand the scope of the study. Future research can also incorporate tracks that feature other artists and study the impact of features on streaming success of a track.

The dataset utilized in this study was limited to a specific time frame (March to December 2022) and relies on data solely from Chartmetric. If a researcher is able to collaborate with Spotify and get access to internal Spotify metrics, he will be able to have access to better quality and more extensive data and thus reach greater insights.

Although this study included a wide range of social media platforms, future studies may include a wider range of platforms like Facebook, YouTube and other emerging social media networks to assess their impact on streaming success.

This study focuses solely on streaming numbers of a track on Spotify, but there are several other music streaming platforms that can be studied like Apple Music, Tidal or Deezer.

For future research, additional variables can be added that will enrich the model and bring new insights. Relevant variables that can be incorporated may be marketing spend, offline promotions and demographic information of the audience.

By acknowledging the limitations of this study and providing relevant recommendations for future research and practice, this section aims to pave the way for further exploration and improvement in understanding the factors driving streaming success on Spotify. Addressing these limitations and implementing the suggested recommendations can lead to more robust insights and better strategies for artists, labels and industry professionals.

BIBLIOGRAPHICAL REFERENCES

- Aguiar, L., & Waldfogel, J. (2018). Platforms, promotion, and product discovery: Evidence from Spotify playlists. JRC Digital Economy Working Paper No. 2018-04. European Commission, Joint Research Centre (JRC), Seville. <http://hdl.handle.net/10419/202233>
- Aguiar, L., Waldfogel, J., & Waldfogel, S. (2021). Playlisting favorites: Measuring platform bias in the music industry. *International Journal of Industrial Organization*, 78, 102765. <https://doi.org/10.1016/j.ijindorg.2021.102765>
- Backlinko Team. (2024, July 1). Spotify user stats (Updated June 2024). Backlinko. <https://backlinko.com/spotify-users>
- Bonini, T., & Gandini, A. (2019). "First week is editorial, second week is algorithmic": Platform gatekeepers and the platformization of music curation. *Social Media + Society*, 5(4), 1-11. <https://doi.org/10.1177/2056305119880006>
- Bourreau, M., & Gaudin, G. (2022). Streaming platform and strategic recommendation bias. *Journal of Economics & Management Strategy*, 31(1), 25-47. <https://doi.org/10.1111/jems.12452>
- Datta, H., Knox, G., & Bronnenberg, B. (2017). Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science*, 37(1), 5-21.
- Dawson, C. E. Jr., Mann, S., Roske, E., & Vasseur, G. (2021). Spotify: You have a hit!. *SMU Data Science Review*, 5(3), Article 9. <https://scholar.smu.edu/datasciencereview/vol5/iss3/9>
- Hupfau, B., Zangerle, E., Pichl, M., & Specht, G. (2016). Can microblogs predict music charts? An analysis of the relationship between #nowplaying tweets and music charts. In *Proceedings of the 17th ISMIR Conference* (pp. 365-371). New York City, USA.
- Kapadia, H., Bhattacharjee, R., & Sinha, S. (2022). Spotify stream prediction using regression models. *International Research Journal of Engineering and Technology (IRJET)*, 9(11), 721-725.
- Lee, J., Nagpal, R., Ruetschlin, C., & Stegemann, M. (2016). Cannibalizing or complementing? The impact of online streaming services on music record sales. *Marketing Science*, 35(2), 178-195.
- Liu, X. (2022). Music trend prediction based on improved LSTM and random forest algorithm. *Journal of Sensors*, 2022, Article 6450469. <https://doi.org/10.1155/2022/6450469>
- Maasø, A., & Spilker, H. S. (2022). The streaming paradox: Untangling the hybrid gatekeeping mechanisms of music streaming. *Journal of Media Business Studies*, 19(1), 1-17.
- McKenzie, J., De Roos, N., & Larsson, S. (2021). It takes two, baby! Feature artist collaborations and streaming demand for music. *Applied Economics*, 53(1), 1-15.

- Morris, J. W., & Powers, D. (2015). Control, curation and musical experience in streaming music services. *Creative Industries Journal*, 8(2), 106-122. <https://doi.org/10.1080/17510694.2015.1090222>
- Park, M., Ko, M., Kim, S., Liu, Y., & Song, J. (2018). The effects of user engagements for user and company generated videos on music sales: Empirical evidence from YouTube. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3173574.3174156>
- Prey, R. (2020). Locating power in platformization: Music streaming playlists and curatorial power. *Social Media + Society*, 6(2), 1-11. <https://doi.org/10.1177/2056305120933291>
- Rousidis, D., Garofalakis, J., Kappos, P., & Makris, C. (2020). Social media prediction: A literature review. *Multimedia Tools and Applications*, 79, 6303-6339. <https://doi.org/10.1007/s11042-019-08291-9>
- Salo, J., Lankinen, M., & Mäntymäki, M. (2013). The use of social media for artist marketing: Music industry perspectives and consumer motivations. *International Journal on Media Management*, 15(1), 23-41. <https://doi.org/10.1080/14241277.2012.755682>
- Sciandra, M., & Spera, I. C. (2022). A model-based approach to Spotify data analysis: A Beta GLMM. *Journal of Applied Statistics*, 49(1), 214-229. <https://doi.org/10.1080/02664763.2020.1803810>
- Szymkowiak, A., Gaczek, P., & Jeganathan, K. (2020). Music sales and artists' popularity on social media. *Sustainability*, 12(1), 123-140.
- Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, 21(3), 303-316. <https://doi.org/10.1080/13645579.2017.1381821>
- Xing, Z. (2023). Popularity prediction of music by machine learning models. *Highlights in Science, Engineering and Technology*, 47, 37-45.

