**How To Predict A Hit Song**

Daniel Huang

Carlo Lopez Hernandez

Vibha Naiknavare

George Cole

University of Notre Dame - Mendoza School of Business

MSBA - Sports Analytics

**MSSA-60510 Machine Learning**

**2023 Fall**

**Introduction**

Music is a powerful universal language that profoundly influences our emotions and culture. Thanks to music streaming platforms, we can now easily access a wide range of songs, making music more inclusive, supporting new artists, and encouraging innovation. These platforms offer us personalized playlists and song recommendations based on what we like, making our music experience richer and ensuring that music continues to play a vital role in our lives.

However, this digital era has brought a new challenge for musicians: getting lots of people to listen to their music online. Having a large number of listeners can lead to success and help artists connect with a broader audience. To tackle this challenge, we've launched a project to explore what makes a song a big hit. This information is incredibly valuable in the music industry, where artists aim to attract many listeners. By analyzing data from top-streamed songs in 2023 (over 1000 songs), we'll uncover the key qualities and factors that contribute to a song's popularity using visualization, and correlation heatmap. We'll use various tests like Linear Regression, Random Forest, Decision Trees, and XGBoost to achieve this. Our goal is to gain a better understanding of what made the most-streamed songs of the year successful, so we can apply that knowledge to create future songs and hits.

**Related Research**

In a 2018 paper published in IEEE Transactions on Multimedia, researchers Lee and Lee made significant strides in the realm of music analysis and its implications for artists and the music industry (Lee & Lee, 2018). Their work was centered on the study of music popularity. The study looked into the aspects of music that contribute to its popularity. By using advanced data analysis techniques, they explored the various factors that play a role in determining which songs become hits. This research shed light on what resonates with listeners and what drives high stream counts. Their findings were significant in the context of today's music landscape, which is heavily influenced by digital platforms and streaming services. Lee and Lee's offers practical implications for artists and the music industry, aiding in the creation and promotion of music that resonates with audiences in an increasingly digital and data-centric world.

Our research sets itself apart by focusing on music streaming platforms and the strategies to boost song streams. In today's music landscape, marked by platforms such as Spotify and Apple Music, understanding what propels a song to success is pivotal for artists and industry experts. Our approach in 2023 relies on current data and predictive tests to forecast song streams. This approach distinguishes our work from other authors, who concentrated solely on identifying music resonance with audiences. We take a broader perspective, exploring factors that influence stream counts, and recognizing that music streaming encompasses more than audience resonance. Thus, our research extends previous study by aiming to create songs that not only resonate but also garner streams and success.

**Data Description**

Our dataset, sourced from Kaggle (Appendix 1), encompasses various attributes (Appendix 2) and offers insights into the most popular songs of 2023. With 946 unique values, it provides ample data for our project. Notably, it includes variables like "in\_spotify\_charts," "in\_apple\_charts," "in\_deezer\_charts," and "in\_shazam\_charts," indicating song rankings throughout the year. To focus on Spotify charts, we excluded these columns. Additionally, it contains variables such as "bpm" (beats per minute), "danceability\_%" (reflecting a song's danceability), and "valence\_%" (reflecting a song's positivity), which will aid our exploration of what defines a hit song and its impact on stream counts. We transformed specific variables into factors, including "key" and "mode," and converted "streams" to a numeric format. A logarithmic transformation was applied to normalize data and enhance robustness. Data was divided into training and test sets, with 700 observations allocated to training and 252 to testing. We also excluded categorical data for XGBoost analysis. Recognizing the significance of data cleaning, we acknowledge that removing specific variables can improve the quality of our findings. Summary and density plots of the data can be seen in (Appendix 3 and 4), respectively.

**Methods**

Linear regression is used to model the relationship between a response and one or more explanatory variables. For this topic, we wanted to see if any of the variables had a relationship with the increasing number of streams. From this, we can visualize relationships and further interpret them through correlation. Though we believe it is a great way to explain the relationship, one of its limitations is that it only focuses on linear relationships. Then, our group decided to do a decision tree which is used to categorize or make predictions on how a previous set of questions were answered. This will help us find the factors that affect the number of streams the most and group them together with other significant factors. We have to be careful of overfitting the data using decision trees. So, we will use random forest analysis to account for the overfitting decision trees might do. This analysis groups multitudes of decision trees together to better measure the prediction accuracy. Finally, we will XG Boost which is a combination of gradient-boosted decision trees. This helps us find the best decision tree. XG Boost can emphasize outliers, so we will need to be careful of that. All of these methods will help us find which variables are major predictors of a hit song.

**Results**

Visualization: Initially, our group used the methods from ggplot to visualize the relationship between streams and different variables. In the first place, we created a scatter plot between log\_streams and danceability with a linear regression line. Looking at the regression line and the distribution of points, we noticed that the prediction line has a slightly negative slope meaning that the danceability is inversely related to streams. Besides analyzing continuous variables, our group also explored whether the key of songs (a categorical variable) can influence their streams. Therefore, we created a box plot to visualize the distribution of streams for the songs based on 12 different keys. The resulting graph reveals 2 interesting features: first, the median stream of songs with “A” key is lower than the median streams of songs with other keys, while the songs with “A#” keys have the highest median stream value. This difference explains why songs written in certain keys may be more popular among listeners.

**Regression analysis: Factors Influencing Streams (Appendix**

The regression analysis aimed to uncover the factors influencing the number of streams on Spotify. Among the variables considered, speechiness and acousticness emerged as statistically significant predictors. Specifically, higher speechiness, indicating the presence of spoken words, was associated with a decrease in expected stream counts. Similarly, increased acousticness, signifying a more acoustic sound, was linked to a reduction in song streams. While danceability and liveness exhibited some trends, they did not reach statistical significance in predicting stream counts. It is important to mention that the adjusted R-squared value is relatively low, suggesting that the included variables explain only a limited portion of the variance in stream counts. These findings highlight the importance of speechiness and acousticness in shaping song popularity on Spotify, but additional factors may also contribute to listeners' streaming preferences.

**Correlation and Heatmap (appendix 5):**

The correlation analysis of musical attributes and song popularity reveals several key relationships. Popularity, measured by the logarithmic of song streams, shows a very weak positive correlation with beats per minute (bpm). “Danceability” songs tend to have lower acousticness and higher valence, suggesting that upbeat and positive songs are more popular. Additionally, high-energy songs are less likely to be acoustic, and instrumental songs generally have lower “speechiness”. Liveness has a minimal influence on “speechiness”. These findings provide valuable insights into the factors influencing song popularity, with "danceability," "valence," and "energy" appearing to play significant roles.

To further explore the impact of "valence" and "danceability," we conducted an improved regression analysis by adding interaction terms. The results indicated a positive significance (valence\_.:danceability\_.: 0.0002363, p-value: 0.04649\*), highlighting the importance of these attributes in predicting song popularity.

**Decision Tree:**

Decision tree analysis was conducted on a dataset of 952 songs to uncover the factors affecting song popularity, measured by song streams. The tree identified “speechiness” categories as the initial splitting criterion, segregating songs into high and low “speechiness” categories. Within the low “speechiness” category, further splits occurred based on beats per minute (bpm), classifying songs with lower bpm as more popular. For songs with higher bpm, the tree considered musical key and acousticness as distinguishing factors, revealing specific combinations of these attributes that influence popularity. The decision tree provides insights into how different musical features interact to impact song success on the streaming platform.

**Random Forest:**

Our random forest shows signs of the variables being good predictors of the number of streams. We used all of the variables as predictors to see if these variables actually had any predictive power on streams. Since we didn’t have a proper classification of streams, we decided to look at the Root Mean Squared Error (RMSE) value of the prediction model instead of a correlation matrix. The RMSE we received was 1.19 which is very close to zero. Meaning that on average our model is off by 1.19 units compared to the actual data. This proves that our model is relatively accurate. Then, we applied a bagging model to our random forest which gave us the results of 1.38 for Mean of Squared Residuals (MSR). Also, we reduced our RMSE to 1.15 which is an even better RMSE than before. Based upon that, we still want to improve the model and we would like to try XGBoost on our model One challenge we faced was creating the training and test sets. We ran into the problem of using more observations than were in the data. So, we created a code to subtract the 700 from the training set to then create the test set. Overall, our random forest model provides excellent predictive power with the predicted stream values being very close to the actual stream values.

**XGBoost:**

To further prove our findings, we conducted an XGBoost on our data. A problem our group ran into was not all of our data was a numeric type, we decided to remove those indicators before running the model. After the initial XGBoost model, we were able to find a RMSE value of 1.29. Since our RMSE increased eventually, we need to tune our XGBoost model. After a set of tuning methods, we found the optimal parameters for our XGBoost model should be eta = 0.3, max. depth = 15, min\_child\_weight =3, gamma = 0.1, subsample = 0.8, colsample\_bytree = 0.6. Eventually, the accuracy of our improved XGBoost model slightly improved which has a RMSE = 1.26. Compared with our random forest model, our XGBoost seems to have a higher error. However, as XGBoost is a comprehensive model, one of our future works is to decide whether we should pick the XGBoost model vs. the Random Forest model. In the end, we build an importance variable graph based on our final XGBoost model (appendix 6). The top 3 important variables are valence, bpm, and energy.

**Discussion**

The development of a program that understands the key factors behind hit songs is highly valuable in the music industry. This technology has the potential to benefit the music industry significantly, offering artists and labels a means to streamline songwriting, ultimately saving them valuable time and effort. While such a program may not entirely replace human creativity and emotion, it serves as a valuable tool that complements the creative process and expands the possibilities within music production, as well as helps artists reach a higher level of success. Thus, utilizing analytics could be the next step in the music industry’s growth.

**Conclusion and Future Work:**

Our work has shed light on the various factors that influence a song's streaming success. Through machine learning tests such as Linear Regression, Correlation, Random Forest, and XGBoost, we were able to achieve our original goal of uncovering factors that can help predict a song's success. A potential extension of our work could be utilized by Spotify or other streaming platforms. By leveraging our data and expanding upon it, platforms could employ our findings to make better recommendations for users. For instance, Spotify offers personalized playlists for each user based on their most-streamed songs. By applying our methods, we could likely identify key factors within a user's most-streamed songs and discover songs that align with them, resulting in a more personalized and enjoyable listening experience.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Contribution:**

Report: Daniel, Vibha, Carlo, and George

Presentation: Carlo, Vibha, George, and Daniel

R code: Daniel, Vibha, Carlo, and George

**Bibliography:**

Lee, J., & Lee, J.-S. (2018, March 29). *Music Popularity: Metrics, Characteristics, and Audio-Based Prediction*. IEEE Xplore. https://ieeexplore.ieee.org/document/8327835

**Appendix:**

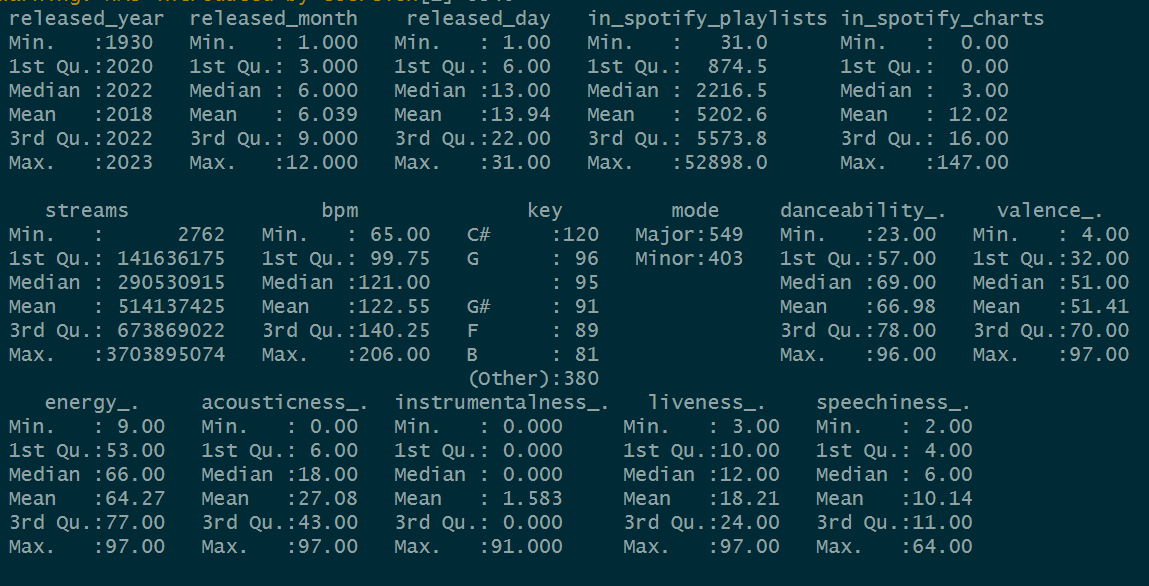
Appendix 1:

Spotify: <https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>

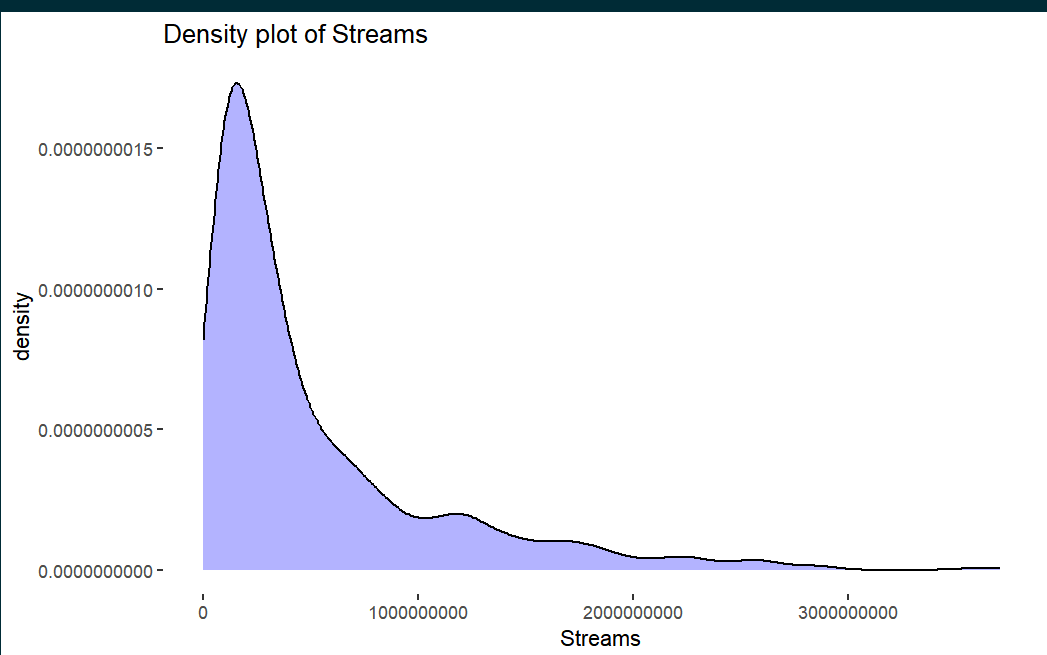
Appendix 2: Variables used in Dataset

* **track\_name:** *Name of the song*
* **artist(s)\_name**: *Name of the artist(s) of the song*
* **artist\_count**: *Number of artists contributing to the song*
* **released\_year**: *Year when the song was released*
* **in\_spotify\_playlists**: *Number of Spotify playlists the song is included in*
* **in\_spotify\_charts**: *Presence and rank of the song on Spotify charts*
* **streams**: *Total number of streams on Spotify*
* **bpm**: *Beats per minute, a measure of song tempo*
* **key**: *Key of the song*
* **mode**: *Mode of the song (major or minor)*
* **danceability\_%**: *Percentage indicating how suitable the song is for dancing*
* **valence\_%**: *Positivity of the song's musical content*
* **energy\_%**: *Perceived energy level of the song*
* **acousticness\_%**: *Amount of acoustic sound in the song*
* **instrumentalness\_%**: *Amount of instrumental content in the song*
* **liveness\_%**: *Presence of live performance elements*
* **speechiness\_%***: Amount of spoken words in the song*

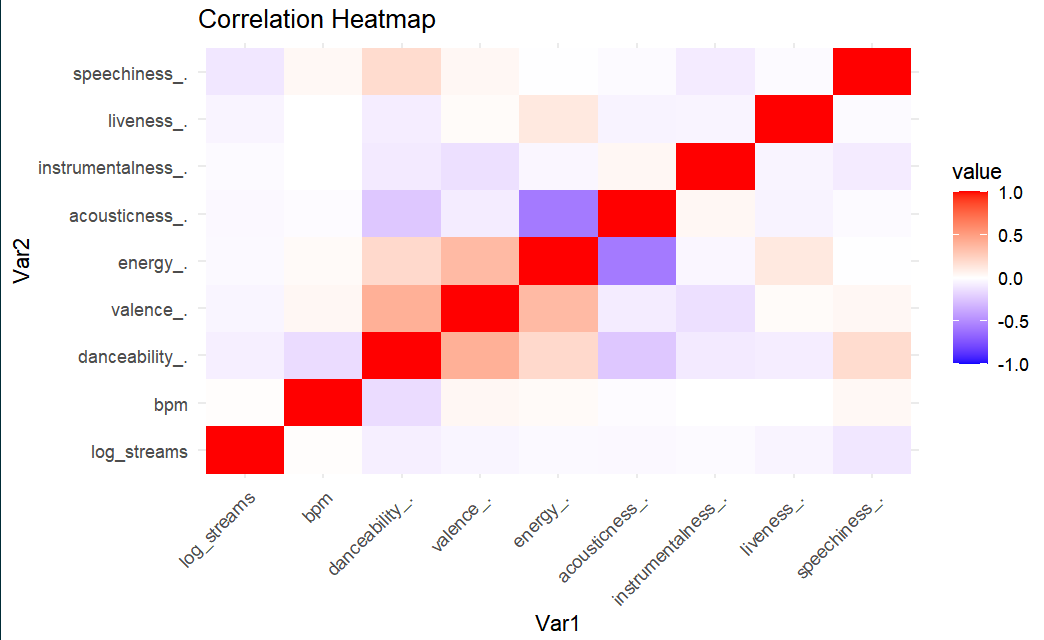
Appendix 3: Summary of Spotify data



Appendix 4: Density Plot



Appendix 5:



Appendix 6 (Importance Variable):

