Decoding Song Popularity: A Machine Learning Perspective

Authors: Siddharth Das, Summer Mohammad, Diego Martinez, Sakura Garcia

March 23, 2024

1 Introduction, Background, and Data Description

Spotify is a popular music streaming service that enables users to explore and enjoy a wide variety of music from different artists and genres. Considering the extensive variety of popular music genres and songs, we want to utilize statistical methods to increase our understanding regarding the foundational aspects of creating famous music.

The main purpose of this research is to identify any underlying patterns that exist between audio attributes and song popularity. Furthermore, we desire to comprehend the optimal method for predicting the total number of song streams on Spotify. We will achieve this goal by implementing three Supervised Machine Learning models. The Machine Learning models we will utilize include Ridge regression, LASSO regression, and Tree-based regression. Additionally, we will compare these prediction methods to further understand why a certain model may outperform alternative methods under this setting.

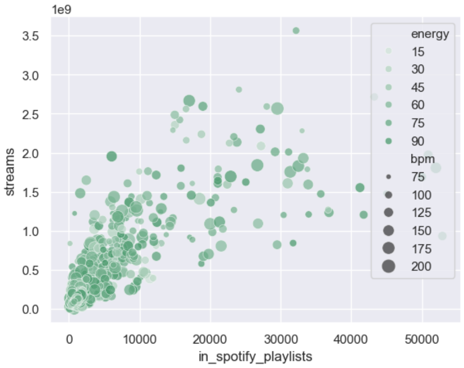
As previously mentioned, this scientific study will examine the relationship between the most significant audio attribute predictor variables and the response variable: the total number of song streams on Spotify. We intend for this investigation to provide meaningful insights regarding sustained song appeal, artist longevity, audience preference, and marketing effectiveness. Record labels, artists, and the music industry as a whole can utilize the knowledge discovered in this scientific analysis to optimize decision-making for maximum profit and popularity.

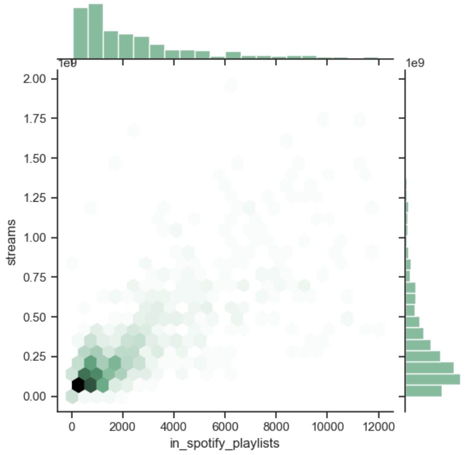
To support our scientific analysis, we will investigate the Most Streamed Spotify Songs 2023 dataset, found on Kaggle. This dataset comprises an expansive list of the most famous songs of 2023, specifically providing insights into each song's audio attributes, streaming statistics, playlist presence, and chart ranking on various music platforms. The exact audio attributes we will utilize for our Machine Learning model predictions include bpm, key, mode, danceability, valence, energy, acousticness, intrumentalness, liveness, and speechiness. The exact definitions of these variables can be found here.

Before proceeding, we must clean the data to ensure the accuracy and interpretability of our Exploratory Data Analysis and Machine Learning model predictions. To begin, we analyzed the unique values of all columns to understand the data preprocessing necessities. First, we adjusted the column names for simplicity. Next, we removed observations with invalid or empty values in the Streams and Key columns. Importantly, we noticed 3 quantitative variables were incorrectly interpreted by RStudio as character variables. Considering this error would negatively impact the accuracy and interpretability of our EDA and model predictions, we removed the commas and converted them into integer variables. Finally, we confirmed that the whole dataset was devoid of invalid values and duplicate observations. Before this data preprocessing, the dataset consisted of 953 observations. As a result of our data cleaning, we removed 154 observations, and proceeded with the dataset consisting of 799 observations. Next, we will proceed with this investigation by implementing an EDA.

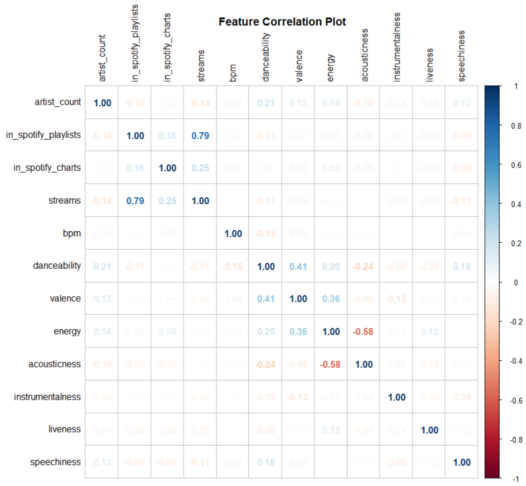
2 Exploratory Data Analysis (EDA)

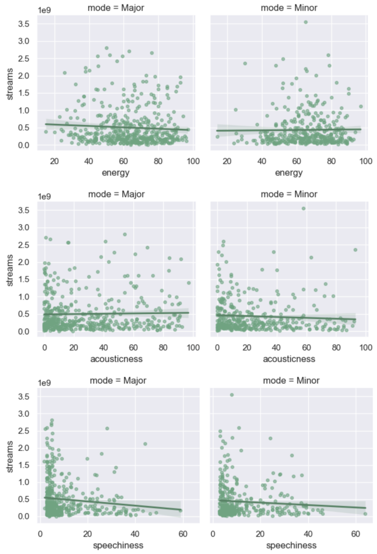
We began with a simple look into some summary and descriptive statistics, such as quartiles, mean, and standard deviation for some predictors of interest. Notable observations included a mean release year of 2018, a minimum and maximum of about 31 and 52,000 Spotify playlists that a song is in, respectively, and a minimum and maximum of about 0 and 530 Apple Music playlists that a song is in, a notably different range than its Spotify counterpart. A feature correlation matrix suggested that there was no meaningful relationship between the number of streams that a song received and any other predictor of interest besides the number of Spotify playlists that a song was in, a sign of limited predictive power of any potentially constructed models.





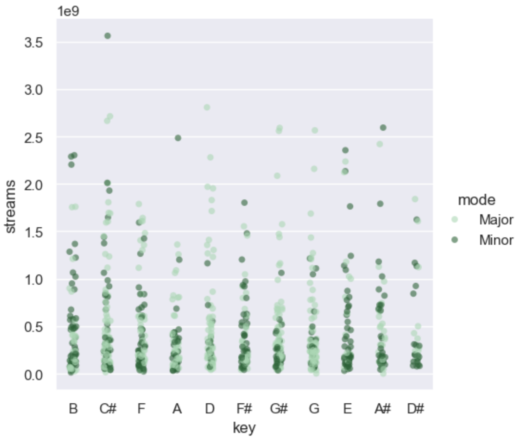
We used the *variance inflation factor (VIF)* as a measure of potential multicollinearity between any numerical predictors of interest, although there weren't any computed VIFs that were indicative of multicollinearity. Next, we considered the relationship between streams and a few variables of interest when conditioned on song mode with a facet plot. Plots were constructed with simple linear regression lines of best fit and 95% confidence bands.

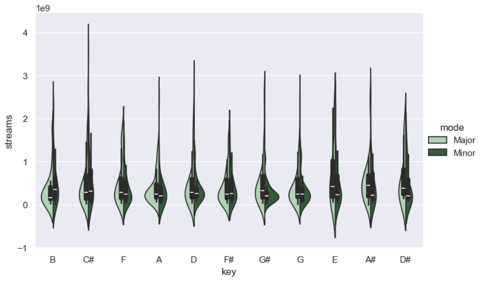




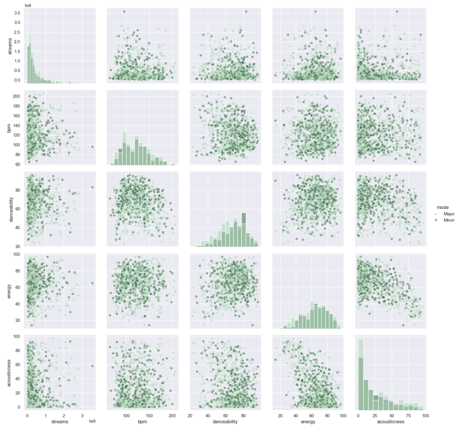
The confidence bands in the plots above suggest that there is no significant change in the slope coefficient when changing between song mode. Otherwise, we can infer that there isn't any linear relationship between streams and any of the above predictors, following suit with our VIF calculations. Next, we constructed a simple categorical scatterplot on song key and mode (insert plot later). To aid in visualization, we used *seaborn*'s kernel density estimation tool to recreate the plot with estimates for the stream distribution with respect to different song keys and modes. The plot also includes boxplots for each distribution estimate to compensate for not being able to see the actual observations.

As seen below, means in the *streams* variable appear to be uniform and pretty much equal across keys and modes except for a few song keys, such as B and E. It's worth noting that since analysis on predictors beforehand suggested that there isn't any meaningful relationship between *streams* and the majority of predictors, it's likely that the differences in sample means illustrated above can be attributed to noise or artifacts in the sampling process due to a rather small dataset. A general pairwise grid plot was constructed to get a glimpse into how the different predictors interact with each other. Histograms to visualize variable distributions are plotted along the diagonals and pairwise scatterplots





between difference variables (including *streams*) are on the off-diagonal entries. (insert plot later). As we can see, there appears to be some sort of relationship between *acousticness* and *energy*, aligning with our intuition. However, the VIF computation from earlier suggests that this relationship won't have any substantial impact on the predictive power of regression models. Distribution of *streams*, *bpm*, *danceability*, and *energy* seem relatively equal when conditioned on major vs minor modes.



3 Methodology

For the purpose of determining the optimal Machine Learning model to best predict the total number of song streams on Spotify, we considered numerous methods learned in STA 141C. Since we are interested in the prediction of streams, which is a quantitative variable, we were able to eliminate various Classification and Unsupervised learning techniques. We excluded Classification because those models are suited for qualitative response variables, whereas our response variable is quantitative. Additionally, we dismissed Unsupervised learning because these models are not intended for prediction, whereas we are certainly interested in predicting the total number of streams.

After considering the characteristics of our specific dataset and the assumptions for numerous models, we decided to proceed with [1]Ridge regression, LASSO regression, and [2]Tree-based regression. Ridge and LASSO regression are alternative forms of linear regression models that include a regularization parameter for shrinking, or reducing, the coefficient estimates. Given that our primary goal was to better understand how each song attribute or statistic affected the total number of streams, interpretability was paramount in choosing an appropriate regression method. LASSO and tree based regression methods are notorious for their interpretability given that they select the most important

[1] Lecture Notes - Linear Model Selection and Regularization: https://hastie.su.domains/ISLR2/Slides/Ch6_Model_Selection.pdf

[2] Lecture Notes - Tree-based methods: https://hastie.su.domains/ISLR2/Slides/Ch8_Tree_Based_Methods.pdf

predictors, but Ridge regression fails to maintain interpretability in high dimensional datasets. In our specific case, loss of interpretability through Ridge regression was not determined to be an issue since our dataset was relatively low dimensional. Shrinkage methods like LASSO and Ridge are also helpful in reducing the model's chance at overfitting the data, which is important given the limited observations we have to train our model after splitting 799 observations into a training and testing set.

The Ridge Regression coefficient estimates are calculated as the values that minimize

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\text{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2 \qquad (\text{L2 Regularization Term: } \lambda\textstyle\sum_{j=1}^{p}\beta_j^2)$$

In contrast, the LASSO regression coefficient estimates are calculated as the values that minimize

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\text{RSS}+\lambda\sum_{j=1}^{p}|\beta_j| \qquad (\text{L1 Regularization Term: } \lambda\textstyle\sum_{j=1}^{p}|\beta_j|)$$

The MSE (Mean Squared Error) from each of these methods will be calculated using R. Lasso Regression has a feature selection property which is characterized by the L1 regularization term that produces sparse solutions and shrinks coefficients to exactly zero. Ridge Regression has an L2 regularization term that keeps all coefficients in the model. However, the L2 regularization terms penalizes large coefficients and shrinks them towards zero, but not exactly to zero. Both shrinkage methods require an appropriate selection of the tuning parameter , which can be done using cross-validation. To do so, we choose a grid of λ values and compute the error for each value, and select the λ with the smallest cross-validation error. The model is refitted using all of the available observations and the optimal value of [3]λ. The regression tree was fit using both categorical and numerical predictors that were directly related to Spotify statistics for a given song. We also avoided using any predictors related to time as this would heavily bias predictions.

4 Main Results

Before implementing all three Machine Learning techniques, we obtained a numerical summary of our 10 audio attribute predictor variables and the response variable. By comparing the mean, median, and range of values for all 11 variables, we realized that the predictor and response variables were on significantly different scales. When the predictor and response variables are on different scales, the accuracy and interpretability of our results could be diminished. To find a solution, we considered excluding certain variables from the model. However, after further research into potential solutions, we decided the best solution was to scale the data. Initially, we only scaled the response variable, and quickly realized that this method was incorrect as the variables were still not on the same scale. After further testing, we discovered and implemented the correct solution by scaling all the quantitative predictor variables and the response variable. Now that the data was completely prepared for optimal accuracy and interpretability from our Machine Learning predictions, we were ready to proceed with training and testing our models.

After implementing all three Machine Learning techniques, we were absolutely stunned by the results. According to all three models, the 10 audio attributes were deemed insignificant for predicting the total number of streams on Spotify. Recall that LASSO shrinks insignificant coefficient estimates exactly to zero, whereas Ridge regression only shrinks insignificant coefficient estimates close to zero, but not exactly zero. By analyzing the coefficient estimates of the Ridge and LASSO models, we learned that LASSO shrunk every single audio attribute to zero! In other words, LASSO regression determined that all 10 audio attributes were absolutely insignificant in predicting the total number of streams on Spotify. As we expected, the Ridge regression model did not shrink any of the coefficient estimates to zero. However, most of the coefficient estimates were extremely close to zero. What does this mean? We interpreted these shocking findings as clear evidence that the number of song streams on Spotify cannot be accurately predicted by the song's audio attributes. As far as we could tell, there

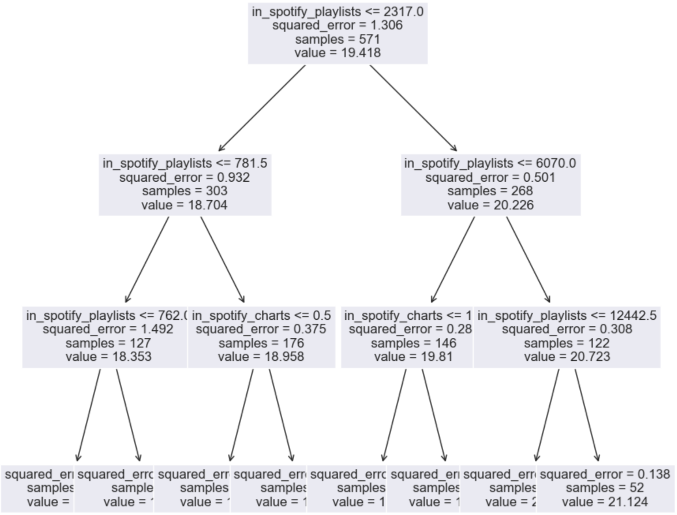
[3] Lecture Notes - Resampling Methods: Cross Validation and Bootstrap: https://hastie.su.domains/ISLR2/Slides/Ch5_Resampling_Methods.pdf

is little to no consistency between song attributes and song popularity. This highlights the amazing talent of musical artists, given the fact that it is extremely difficult to consistently create popular music.

Furthermore, we compared the prediction error of all 3 models to better understand which technique was most accurate. The MSE of Lasso was calculated to equal 0.4153, while Ridge was equal to 0.4268. The Regression tree produced a MSE equal to 0.44, which is estimated to be 1.9 million streams. The Regression Tree demonstrates that the number of Spotify playlists that the song appeared in is the most important predictor, followed by the song's ranking in Spotify charts. In comparison, the audio attributes were simply insignificant. Similarly, the Lasso model demonstrates that the number of playlist inclusions for a song is the strongest predictor for streams, while the rest of the coefficients related to song metrics and audio features are set to zero. Since a lower MSE is indicative of more accurate predictions, we observed that Lasso yields the model with the best accuracy when compared to ridge regression and the regression tree. Moreover, the Lasso regression model had the least amount of predictor variables, since LASSO shrinks coefficients to zero. The combination of relatively lower MSE and the least number of predictor variables lead us to conclude that LASSO is the optimal Machine Learning model to predict the total number of streams on Spotify.

Model	# of Predictors	MSE	R^2
LASSO	1	0.4153	0.5449
Ridge	9	0.4268	0.5332
Regression Tree	13	0.4400	0.62

In addition to having the highest accuracy, the Lasso Regression model has a higher R^2 value when compared to the two other models. Given that R^2 is a measure of how well the model fits the data, we observe that Lasso Regression provides the best fit for our data and successfully captures 54.49% of the variance within our dataset. Comparatively, the Ridge model captures a slightly lower amount of variance of 53.32%, while the regression tree captured about 62% of the variance. After fitting the tree, we observed that the tree partitioned the predictor space by using the number of Spotify playlists and then number of Spotify charts that a song is in, consistent with the suggested importance of features from earlier EDA.



To clarify on our predictor variable features and the relation to Spotify playlist inclusions, recall that the main purpose of this project is to understand the relationship between audio attributes and their impact on the total number of streams on Spotify. Our initial prediction models were prepared to be real-world applicable, only utilizing audio attributes to predict song streams. We hypothesized that we would find at least some moderate relationships between significant audio attribute predictor variables and total song streams. After implementing the three Machine Learning techniques described previously, we were stunned to learn that there are no significant relationships between audio attributes and song streams. Since the results we found did not align with what we hypothesized or desired, we decided to further explore prediction models utilizing other combinations of predictor variables that are not real-world applicable. This is where we added the variable for the number of Spotify playlists the song is included in. Utilizing the Ridge, LASSO, and Tree-based regression models, we confirmed that the number of Spotify playlists the song is included in is indeed statistically significant. In other words, when predicting with the 10 audio attributes and the Spotify playlist inclusions variable, LASSO again shrunk all 10 audio attributes to zero, and only included the Spotify playlist inclusions variable with a coefficient estimate approximately equal to 0.78. Similarly, the Ridge regression shrunk all 10 audio attributes close to 0. Only the Spotify playlist inclusions variable was not relatively close to 0, approximately equaling 0.77. Before modeling, we intuitively hypothesized that there would be a significant relationship between Spotify playlist inclusions and the total number of streams on Spotify. While our models did confirm this hypothesis to be true, the additional predictor variable of Spotify playlist inclusions did not necessarily provide meaningful insights in the real-world. Even though it was interesting to confirm the significant relationship between Spotify playlist inclusions and total streams on Spotify, this insight would not provide any significant knowledge to record labels, artists, or anyone else in the music industry that is interested in creating popular music.

5 Discussion and Outlook

The results from the data analysis imply that the Lasso Regression model is a better fit for the data than Ridge Regression. This is probably due to the fact that Lasso has a feature selection property while Ridge Regression does not. The feature selection property of Lasso Regression provides an advantage to the model by identifying which predictors are important to keep and eliminate any irrelevant ones in the model. Whereas, Ridge Regression keeps all of the predictors in the model regardless of importance.

Furthermore, our data analysis results show that the Lasso Regression model was superior to the model obtained via the Regression tree. While the Regression tree still implicitly performs feature selection by partitioning the feature space, this method may still fail to completely suppress the influence of irrelevant variables. As stated previously, Lasso is successful in producing a sparse model that forces irrelevant predictors to be zero.

While our EDA explicitly showed that only one predictor (number of playlist inclusions for Spotify) was strongly correlated with the number of streams, perhaps in the future it would be useful to use alternate forms of feature selection such as forward stepwise, backward stepwise, or best subset selection. Doing so on an expanded dataset with more than 799 valid observations would improve model performance and reduce variance. Since our best model only captured 54.49% of the variance within our dataset, it is likely that the number of streams is best predicted by variables outside of the standard Spotify metrics and audio features. Future analysis should then consider predictors related to artist popularity, such as time spent on the top 100 charts, global ranking, number of monthly listeners, etc.

6 Conclusion

The implementation of different Machine Learning Models helped us identify and understand any underlying patterns that exist between audio attributes and song popularity. We found that using Lasso regression produces the best model that predicts the total number of song streams on Spotify. From this model and EDA, we found that the number of playlist inclusions is the most correlated with streams and is thus the most important variable to consider when predicting the number of streams for a given song. This finding contradicts our initial hypothesis, which was that song attributes such as energy, bpm, acousticness, etc. would be somewhat correlated with number of streams and integral to accurately predicting this target variable. In reality, this can be interpreted as proof that there is extreme diversity in top performing songs and no intrinsic formula for engineering popular music. Some questions that can be considered in future research: Are streams predicted by the number of playlist inclusions, or are the number of playlist inclusions predicted by streams (popularity)?