Predicting Music Popularity Using Spotify Song Metadata

Introduction

The digital revolution led by streaming giants like Spotify, Apple Music, and YouTube has not only relegated CDs to history but also democratized music distribution, granting global access to vast music catalogs. This transformation has ushered in a new era of music discovery, where algorithms cater to individual tastes, enabling listeners to explore music uniquely tailored to them. For emerging artists, this means unprecedented opportunities to reach audiences worldwide.

In this landscape, Hit Song Science (HSS) emerges as a crucial endeavor within Music Information Retrieval, seeking to decode the essence of a song's appeal. This research holds personal significance as it bridges my fascination with music's universal language and its analytical dissection. For the industry, understanding a song's potential hit factors pre-release could pivot the artistic direction and mitigate financial uncertainties, marking a strategic evolution in music production and marketing.

Existing literature offers diverse methodologies for forecasting song success, underscoring the complexity and multifaceted nature of hit song prediction.

Paper 1[1] in the cited works investigates the symbiosis of Spotify audio features and YouTube social metrics, revealing that the incorporation of social media data notably enhances predictive accuracy. However, this approach limits itself to using a dataset of songs that were only popular in 2021.

[1] Yee, Y. K., & Raheem, M. (2022, September 21). Predicting music popularity using Spotify and YouTube features. SRS Journal. https://indjst.org/articles/predicting-music-popularity-using-spotify-and-youtube-features

Paper 2[2] in the cited works delves into the realm of song lyrics, proposing that lyrical complexity, as manifested through rhyme and meter, holds predictive power for song popularity. This study broadens the predictive landscape by introducing textual analysis into hit song science, an area less explored in previous research, and is something that can be incorporated once the initial model has been built.

Paper 3[3] in the cited works introduces the concept of 'early adopters' behavior from social media platforms, like Last.FM, as a significant predictor of a song's popularity. This novel approach underscores the potential of leveraging early listener data to forecast song success, a method not fully explored in prior studies. Additionally, it incorporates Spotify song metadata in determining a song's popularity but again limits itself to using Top 20 Dance songs between 2011 to 2013.

Building upon these foundational insights, this paper endeavours to use a dataset encompassing hit songs over a wide range of genres and over a much larger period, 2000 to 2020. Similar to previous research, Spotify's metadata of popular songs will be leveraged to investigate what makes a song popular and a model of popularity shall be built around this concept.

Methodologies

Dataset

Data was sourced from a Kaggle repository containing the audio statistics of the top 2000 tracks on Spotify from 2000-2019. The data included 18 columns listing the tracks and their audio qualities. This dataset was chosen because it contained popular songs

[2] Can song lyrics predict hits? (n.d.). https://cs.uwaterloo.ca/~browndg/CMMR15data/CMMR2015paper.pdf

[3] Herremans, D., & Bergmans, T. (2020, October 16). Hit song prediction based on early adopter data and audio features. arXiv.org. https://arxiv.org/abs/2010.09489

across various genres, allowing for the building of a more precise and accurate predictive model.

The dataset contained the following variables:

- **artists**: The artists' names who performed the track.
- **track_genre**: The genre in which the track belongs
- **song**
- **popularity**: Value between 0 and 100, with 100 being the most popular.
- **explicit**: Whether or not the track has explicit lyrics (true = yes it does; false = no)
- **duration_ms**: The track length in milliseconds
- **danceability**: Measures how suitable a track is for dancing, ranging from 0 to 1.
- **energy**: Represents intensity and activity on a scale from 0 to 1
- **loudness**: The overall loudness of a track in decibels (dB)
- **mode**: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is 1 and minor is 0
- **speechiness**: Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
- **Acousticness**: Music that solely or primarily uses instruments that produce sound through acoustic means
- **instrumentalness**: Predicts whether a track contains no vocals Scale is 0 to 1 where 1 is high instrumentalness
- **liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live
- **valence**: Valence measures the musical positiveness conveyed, from 0 to 1
- **tempo**: Beats per Minute

A. EDA

In the exploratory data analysis, I plotted histograms and boxplots to understand the spread of the predictor & response variables and identify outliers that could affect the coefficients of my regression model. I applied Log transformations and a cube-root transformation to correct for the right skew of several predictor variables. I also plotted a correlation matrix to check for multicollinearity between predictors and calculated their Variance Inflation Factors (VIFs). High multicollinearity can lead to large changes in the estimated regression coefficients for small changes in the model or the data, reducing its reliability.

B. Checking GLM Assumptions & Fitting of the Model

Given that I was predicting popularity I opted to fit the data to a Poisson regression model. Such models are appropriate when modelling count data that is non-negative integers (0, 1, 2).

However, the Poisson regression model makes several assumptions that must be satisfied.

Assumption 1: Checking for whether Popularity (Y) follows the correct distribution. To check for this, I performed a Pearson's chi-squared test which compared the expected values of the Popularity data vs the actual values to see if there were any significant differences.

Influential observations are detrimental to a model as they skew the results of a regression analysis. The tests I did to check for influential observations were computing Cook's distance and calculating DFBETA values. These are used to identify and assess the influence of individual data points on the overall fit of a regression model and measure the influence of a single observation on each regression coefficient respectively.

Once I identified the influential values, I removed them and re-fit the Poisson regression model.

Assumption 2: Correct Link Function (Log Link) and Assumption 3: Linear Relationship in Link Function and Assumption 4: Error Independence. These assumptions were all verified via deviance residuals plotted against predictors. Ideally the residuals should should be randomly distributed around zero without clear patterns or trends.

Additionally, Pearson's residuals (deviance residuals) are used to check the goodness-of-fit of models, helping determine if the model adequately captures the relationship between the Popularity and the predictors. The results from this step caused me to go back and transform some of my predictor variables as well as remove outliers from certain boxplots.

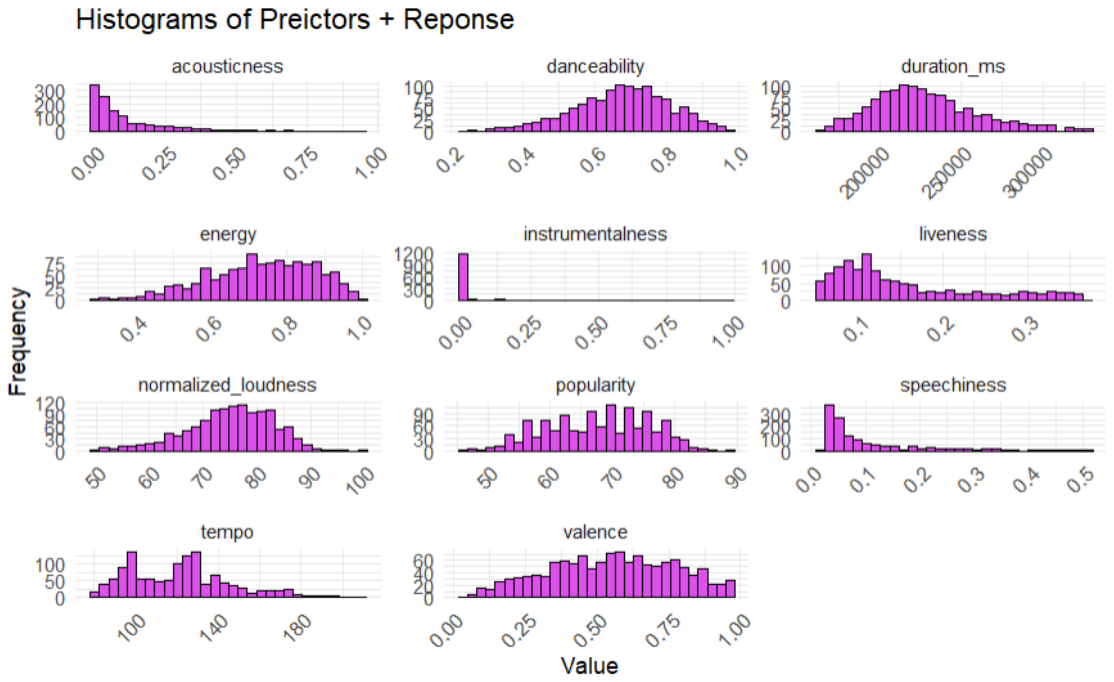
The next step in my model-building process was stepwise variable selection using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to check the goodness of fit of the reduced models. This process led me to drop certain variables (danceability, cube_root_acousticness, log_liveness and log-tempo (BIC)) and proceed with re-fitting the data to a Poisson regression model.

To test whether the reduced models were better than the original, I implemented the Likelihood ratio test which in addition to being a goodness of fit test, is also used to compare 2 models with different covariates.

The final step in my analysis was model validation - testing whether it would be able to accurately correctly predict a song's popularity on a test dataset - via K-Fold Cross Validation.

Results

The histograms of the predictors + response variable (popularity) revealed that certain variables were right and left skewed. However, whether these skews would affect the Poisson model would be revealed by the deviance residuals plots. Given that the maximum value of loudness was 0 and the minimum value was -20, I went ahead and normalized it to a [0,1] scale to make it comparable to other attributes. It is worth noting that the following variables had mean values greater than 50 or 0.5: danceability (0.6), popularity (60), energy (0.72), valence (0.57), tempo (120 beats per minute) suggesting that most Popular songs are are high energy and positive (valence = positivity)

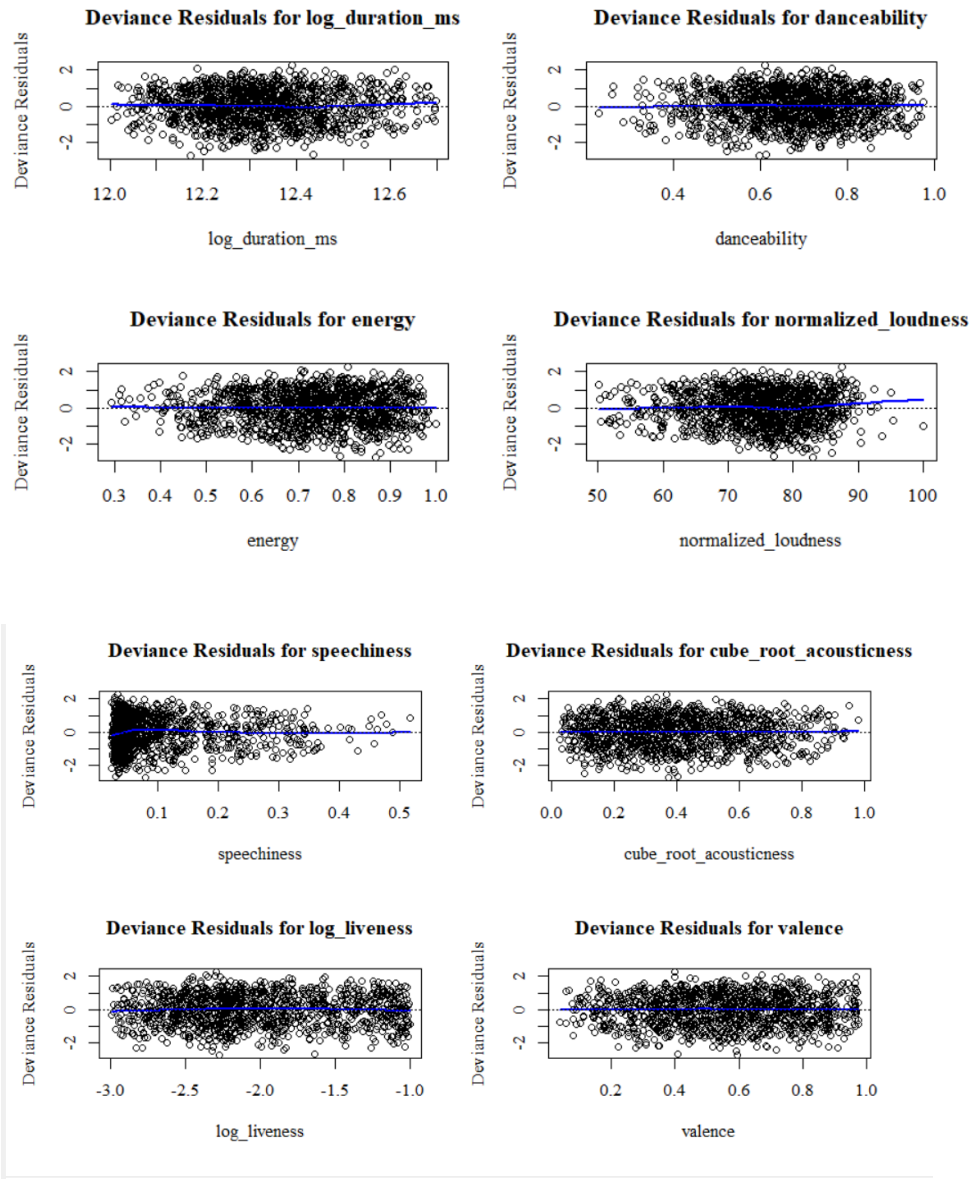


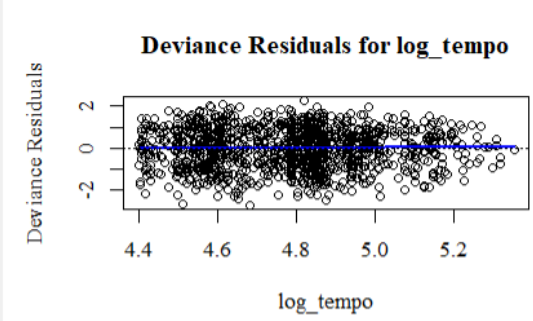
Histograms of Spotify audio features in dataset.

The histograms revealed that some variables were right-skewed and a series of deviance residual plots revealed the right-skewness of some variables was negatively affecting the Poisson model. Clear, discernible patterns were observed in the residual plots and there were many values outside the [-2,2] threshold, suggesting the presence of outlier values for these variables: acousticness, duration_ms, liveness and tempo

for the following variables. A mixture of log transformations and manual outlier removal (removing values above and below the 1.5 IQR) were implemented to correct for this issue

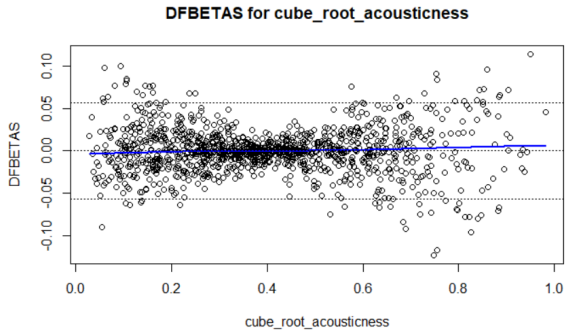
The other 2 skewed variables, speechiness and instrumentalness were removed from the model because even with transformations, discernible patterns were observed in their residual plots

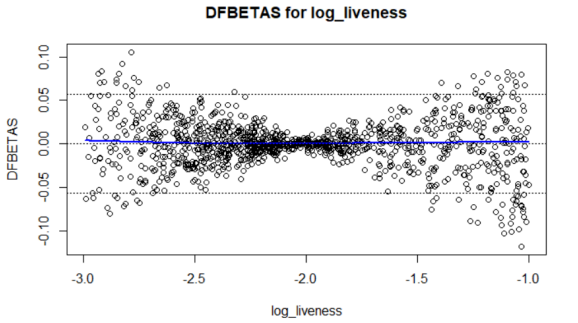


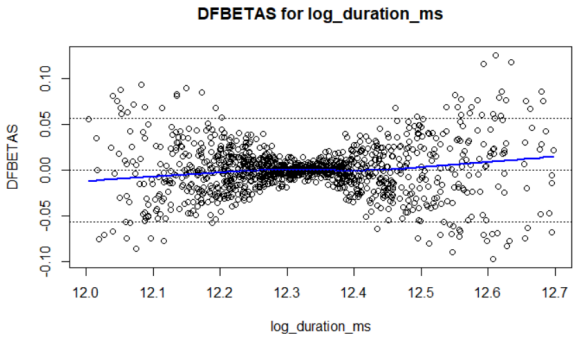


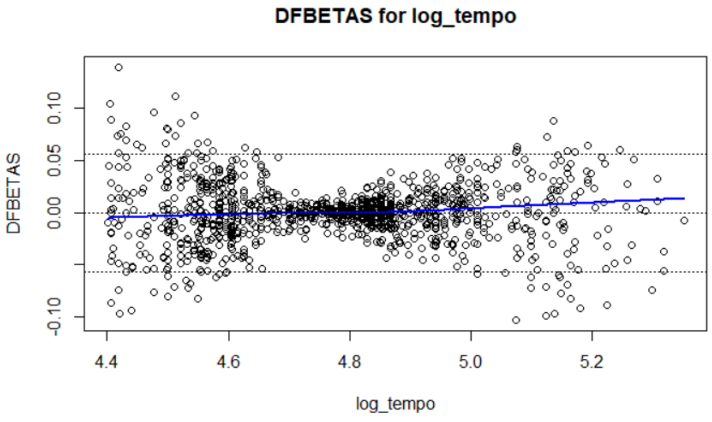
Deviance residuals from the Preliminary model fit

After removing influential observations via Cook's distance calculations and DFFITs, the DFBETA values for certain variables was still not satisfactory and nonlinear patterns were observed as shown below

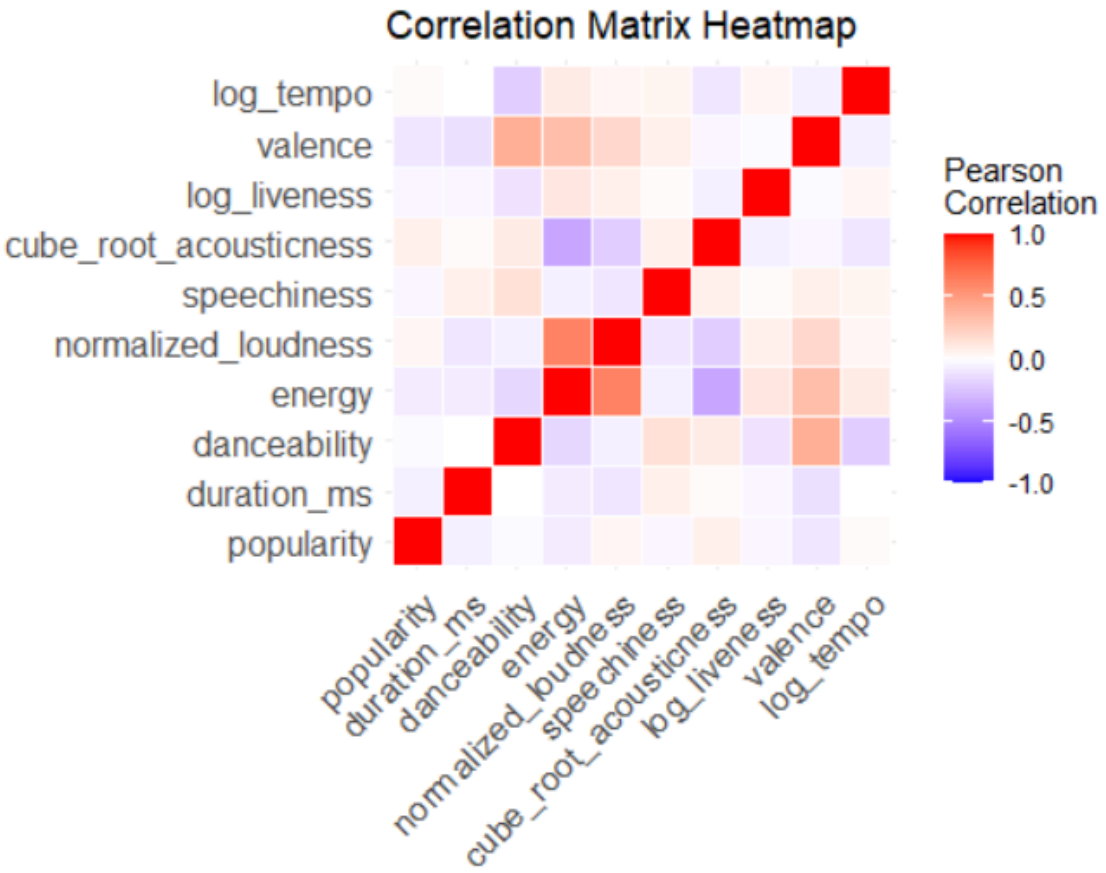








DFBETA values for the Poissson model with transformed predictors and removed influential values.



Correlation heatmap of audio features

The correlation heat map above revealed that energy and normalized loudness were highly correlated while Valence and Danceability were mildly correlated. This intuitively made sense since energetic hyped-up songs would be loud and danceable songs have positive sounding musical elements. The VIFs for each predictor, calculated after fitting the Poisson model were all below 5 (VIFs above 5 mean multicollinearity is affecting the predictor).

With this original complete model, AIC and BIC values were as follows: 8517.345 and 8563.371 respectively.

Once stepwise selection was used to figure out which values to remove, the following variables remained in the model: "log_duration_ms", "energy",

"normalized_loudness", "valence" and "log_tempo". Except for log_tempo, the other predictors with problematic DFBETAs we removed.

The reduced model was fit again and this time, the AIC value was slightly lower, `8513.983` suggesting that the reduced model was not much different from the original. This was conifmred by a Likelihood ratio test where the chi-squared value was 2.63, with df =3 and p = 0.45 (above 0.05 sig level)

Finally, the model was validated and returned a prediction error of 65%, suggesitng that the model was not very good at predicting song popularity.

References

Yee, Y. K., & Raheem, M. (2022, September 21). Predicting music popularity using Spotify and YouTube features. SRS Journal.

https://indjst.org/articles/predicting-music-popularity-using-spotify-and-youtube-features
Can song lyrics predict hits? (n.d.).

https://cs.uwaterloo.ca/~browndg/CMMR15data/CMMR2015paper.pdf
Herremans, D., & Bergmans, T. (2020, October 16). Hit song prediction based on early adopter data and audio features. arXiv.org. https://arxiv.org/abs/2010.09489

Dataset source: https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019