What Makes Songs Popular? Using Multilevel Modeling

Colin Pi & Matt Zacharski

6/08/2020

Introduction

Music has become increasingly more accessible over the past several decades with the rise of smartphones and various smartphone apps that allow users to stream music. This, paired with companies in the music industry's desire to gain an edge over competitors and a general curiosity from others has led several people to ponder what makes popular songs popular. In 2012, several researchers working under a grant from the National Science Foundation published a paper titled: *The Million Song Dataset Challenge*. The Million Song Dataset Challenge challenged researchers to predict songs someone would listen to given their listening history and various song level characteristics. Consequently, other papers have been published in response to this challenge. In the paper *Predicting Song Popularity*, several Stanford researchers found that measures such as artist familiarity, loudness, and select genres were associated with a song scoring highly on their chosen popularity metric.

The exploration of what makes a song popular, sometimes called Hit Song Science, assumes that popular songs possess some similar characteristics that make them appealing to a majority of people. While its possible this is not the case, in this report we sought to understand which variables from our dataset were associated with a high popularity score using multilevel modeling.

Methods

The data we used for this report comes from a dataset found on Kaggle.com. This data was taken from Spotify, a large music streaming service located in Sweden and founded in 2006. The data was uploaded to Kaggle by a Kaggle user, Yamaç Eren Ay, who describes themself as a "self-taught data scientist and music enthusiast". Yamaç used Spotify's Web API to develop a dataset of over 160,000 songs with various song level characteristics. To be able to run our models, we took a random sample of 10,000 songs from the original dataset to create the dataset we used for modeling. The characteristics included in the dataset are scores of acousticness, danceability, duration in milliseconds, instrumentalness, popularity, and more. The focus of this report and our response variable was the popularity score. For a complete list of all the variables and their respective descriptions please refer to the table below.

Since popularity is our response variable, it is worth going into greater detail about its calculation. According to Spotify's Web API guides, the popularity score of a song is calculated by an algorithm that strongly weights the total number of plays a song has and how recent those plays are. Consequently, songs that are currently being played frequently tend to have higher popularity than songs that were played a lot in the past.

Variable Name	Description	Unit (range)
popularity	Index denoting popularity of a song	Numeric (0 100)
acousticness	Index denoting acousticness of a song	Numeric (0 1)
danceability	Index denoting danceability of a song	Numeric (0 1)
energy	Index denoting energy of a song	Numeric (0 1)
duration	Duration of a song	Milliseconds (200k 300k)
instrumentalness	Index denoting instrumentalness of a song	Numeric (0 1)
valence	Index denoting valence of a song	Numeric (0 1)
tempo	Tempo of a song	BPM (50 150)
liveness	Index denoting liveness of a song	Numeric (0 1)
loudness	Index denoting loudness of a song	dB (-60 0)
speechiness	Index denoting speechiness of a song	Numeric (0 1)
year	Year a song was published	Numeric (1920 2020)
note	Note of a song	Factor (0 = Minor, 1 = Major)
explicit	A song containing explicit contents or not	Factor (0 = No, 1 = Yes)
key	Key of a song	Factor (0 (C) 11 (B))
artist	Artist of a song	Character

Table 1: Table of the variables with description

Results

Exploratory Data Analysis

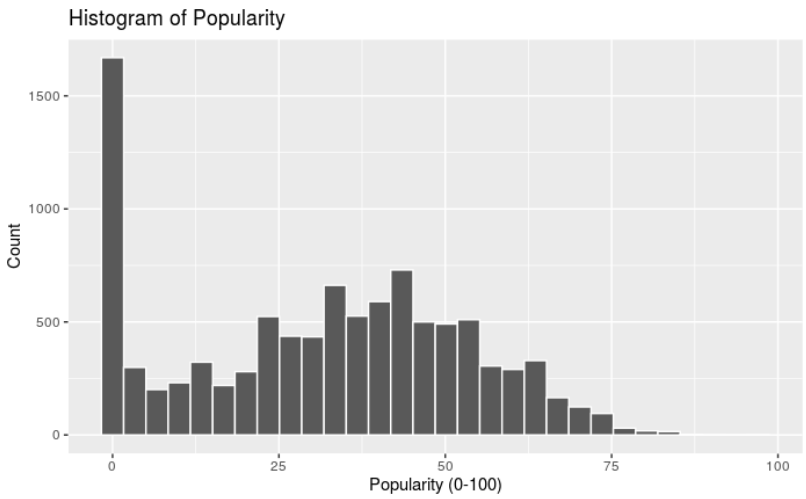

Figure 1: Distribution of Popularity

Performing our EDA, we discovered that the popularity distribution was strongly right-skewed with a high frequency of zero scores. To account for how this large body of zero scored songs influenced our modeling, we decided to construct several models to compare and contrast. Our first model, referred to as Model 1, was a multilevel model grouped on artists using the complete 10,000 observation subset from the dataset discussed above. Our second model, Model 2, was also a multilevel model grouped on artists. However, the data used for this model did not include songs with popularity scores of zero. Lastly, we considered a model, Model 3, a Generalized Linear Multilevel Model that used data where the popularity scores were adjusted to be 0 for 0 scores and 1 for any score greater than 0.

We also noticed several particularly interesting relationships between variables and popularity scores, depicted in the figure below.

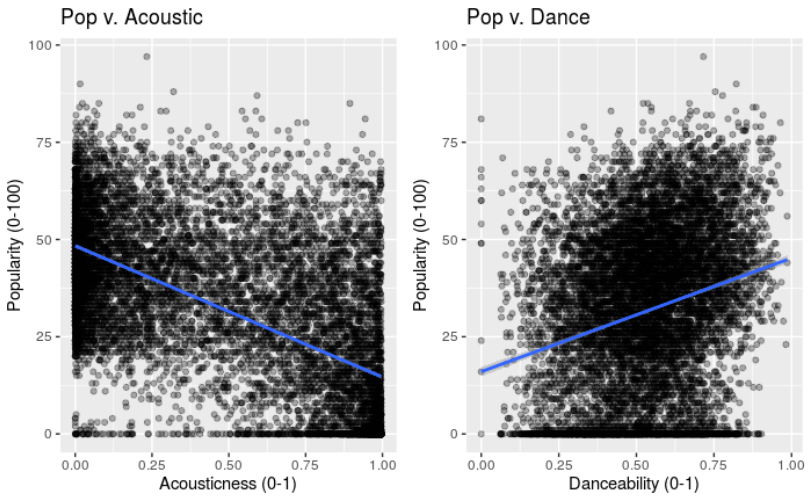

Figure 2: Plots of Response vs. Several Variables.

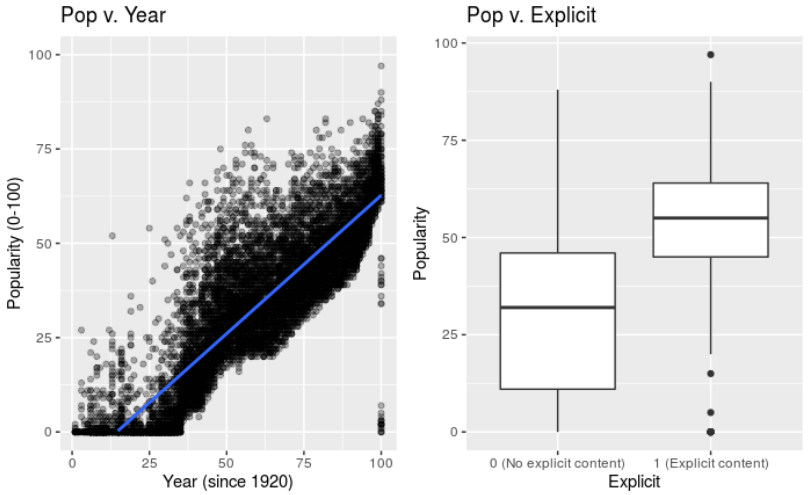

Figure 3: Plots of Response vs. Several Variables.

Model 1

The model specification for Model 1 is the following:

Level 1 (song):

$$Y_{i,j} = a_i + b_i acousticness + c_i danceability + d_i energy + e_i instrumentalness + f_i liveness + g_i speechiness + h_i year1920 + i_i explicit + j_i mode + \epsilon_{ij}$$

Level 2 (artist):

$$a_i = \alpha + u_i$$
$$b_i = \beta$$
$$c_i = \gamma$$
$$d_i = \nu$$

$$e_i = \delta$$
$$f_i = \eta$$
$$g_i = \kappa$$
$$h_i = \lambda$$
$$i_i = \xi$$
$$j_i = \iota$$

The tables below (Table 2 and 3) display the fixed and random effect estimates of Model 1.

Groups	Name	Variance	Std.Dev.
artists	(Intercept) ($\hat{u_i}$)	25.51	5.051
Residual		74.1	8.608

Table 2: Table of Random Effects for Model 1

Variables	Estimate	Std. Error	t value
(Intercept) ($\hat{\alpha}$)	-7.311234	0.752373	-9.718
acousticness ($\hat{\beta}$)	-3.025816	0.473722	-6.387
danceability ($\hat{\gamma}$)	4.556476	0.649911	7.011
energy ($\hat{\nu}$)	-1.504595	0.605707	-2.484
instrumentalness ($\hat{\delta}$)	-2.324892	0.360807	-6.444
liveness ($\hat{\eta}$)	3.225131	0.567216	-5.686
speechiness ($\hat{\kappa}$)	-7.427429	1.012761	-7.334
year1920 ($\hat{\lambda}$)	0.697012	0.005792	120.33
explicit1 ($\hat{\xi}$)	1.204245	0.449813	2.677
mode1 ($\hat{\iota}$)	-0.573283	0.213354	-2.687

Table 3: Table of Fixed Effects for Model 1

Model 2

For the Model 2, we included loudness (ρ) in addition to all the the explanatory varibles used in Model 1. The estimation of the model follows:

Groups	Name	Variance	Std.Dev.
artists	(Intercept) ($\hat{u_i}$)	19.75	4.444
Residual		83.77	9.152

Table 4: Table of Random Effects for Model 2

Variables	Estimate	Std. Error	t value
(Intercept) ($\hat{\alpha}$)	-1.57228	1.149787	-1.367
acousticness ($\hat{\beta}$)	-1.564609	0.520396	-3.007
danceability ($\hat{\gamma}$)	3.70677	0.725659	5.108
energy ($\hat{\nu}$)	-3.189854	0.872567	-3.656
instrumentalness ($\hat{\delta}$)	-2.334359	0.463339	-5.038
liveness ($\hat{\eta}$)	-3.411257	0.635855	-5.365
speechiness ($\hat{\kappa}$)	-6.018143	1.286606	-4.678
year1920 ($\hat{\lambda}$)	0.651488	0.006744	96.606
explicit1 ($\hat{\xi}$)	1.545377	0.473058	3.267
mode1 ($\hat{\iota}$)	-0.695478	0.240709	-2.889
loudness ($\hat{\rho}$)	0.097384	0.037079	2.626

Table 5: Table of Fixed Effects for Model 2

Model 3

The explanatory variable (Y) for Model 3 is coded as $Y = 0$ if the popularity is 0 and $Y = 1$ if the popularity is nonzero. We scaled the year variable, which is relatively wider in range ($0 \sim 100$) than the rest of the variables (mostly $0 \sim 1$) to address the convergence issue of the estimated model. The model specification follows:

$$Y_{i,j} = a_i + b_i acousticness + c_i liveness + d_i speechiness + e_i year + f_i liveness + g_i speechiness + h_i year1920$$

$$a_i = \alpha + u_i$$
$$b_i = \beta$$
$$c_i = \gamma$$
$$d_i = \nu$$
$$e_i = \delta$$

Table 6 and 7 show the estimates of the model.

Groups	Name	Variance	Std.Dev.
artists	(Intercept) ($\hat{u_i}$)	94.92	9.743

Table 6: Table of Random Effects for Model 3

Variables	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept) ($\hat{\alpha}$)	22.8825	1.6833	13.594	<2e-16
acousticness ($\hat{\beta}$)	-7.3313	1.1279	-6.500	8.03e-11
liveness ($\hat{\gamma}$)	-1.7054	0.8426	-2.024	0.042969
speechiness ($\hat{\nu}$)	-5.1673	1.4803	-3.49	0.000482
scale(year1920) ($\hat{\delta}$)	13.5321	0.9993	13.541	<2e-16

Table 7: Table of Fixed Effects for Model 3

Performing our modeling, we see substantial differences between the two models. Notably, the variable that appeared to have the most significant effect across all models was the time variable, which had relatively large t-scores in the first two models and a large z-score in the last model. Additionally, the models seemed to break down with random effects. The third model, especially, tended to break down once it became too complex. Our final models for the first two models ended up sharing almost every single explanatory variable

except Model 1 did not include the loudness variable. Additionally, their fixed effect sizes did not change substantially, typically changing less than a unit per variable. Most variables decreased in Model 2 compared to Model 1, except a variable indicating the explicit status of a song, which increased slightly. For Model 1, a danceability score of .5 was associated with a popularity score increase of 2.28, holding all other variables constant. For Model 2 on the other hand, a danceability score of .5 was associated with a popularity score increase of 1.85, holding all other variables constant.

Since Model 3 is a generalized linear mixed model, our interpretation of its effects are more nuanced and harder to compare to the previous models. Since this model predicts the odds of a song's popularity score being greater than 0, we should interpret its effects by considering odds and odds ratios. However, the estimates for the effects of this model are so large causing the associated odds to change to be so large it may be more clear to just consider the size of the estimate. The largest positive effect for Model 3 was the year variable, indicating that the more recent a song was released, the more likely it was to have a non-zero popularity score. The model estimate suggests that one standard deviation increase in year variable is associated with 753210 times increase in the odds of having nonzero popularity.

Additionally, after performing diagnostics for each model, we observed several outliers and high cooks distance values among some songs released in 2020. Yet running models with and without the outliers did not produce substantial differences, so the outliers were kept in all the models.

Discussion

Across all three of our models, we saw that the most significant variable was the variable that indicated how recently a song had been released. While other variables had greater effect estimates, it is worth noting that these variables typically only had ranges of between zero and one, while the variable indicating how recently a song had been released was measured in years since 1920. This means that this variable had the greatest potential to increase the popularity score. Intuitively, this makes sense that since Spotify's popularity score strongly weights a song's total number of plays and how recent those plays are. However, in hindsight, it may have been beneficial to scale the year variable to an even closer scale as the other variables to gain a better understanding of their relative influences on popularity.

The next most influential variable to a high popularity score appeared to be danceability. Looking closer at how Spotify calculates danceability scores for their tracks, we found that Spotify used a combination of song characteristics, namely tempo, rhythm stability, beat strength, and overall regularity, to develop a danceability score. The complexity of the danceability score may indicate why it is such a good predictor of a high popularity score. For a high score, a song must have "subscores" that fit whatever Spotify defines as favorable tempos, rhythms, etc for dancing. These "subscores" may also play a role in influencing a song's popularity score not captured in these models.

Returning to the study discussed in the introduction, *Predicting Song Popularity*, we see that the researchers found that artist familiarity, loudness, and several select genres were most indicative of a high popularity score. Unfortunately, due to different systems of variable calculation and several different explanatory variables, it is most likely impossible to rigorously compare our findings to theirs. However, it is worth noting that, similar to our study, they found a loudness score to be slightly indicative of a high popularity score, although like in our case it was overshadowed by more substantial predictors.

Future modelers or music producers interested in releasing the next big hit may find it beneficial to exclude the variable indicating how recent a song was released to better understand which variables besides time affect popularity scores.