Spotify and Youtube Music Analysis

Kaviraj Gaire and Justin Won

Introduction

With more and more music being released and the rise of platforms that allow individuals to publish their own music independently, it has become ever increasingly difficult to find songs that you like. In this paper, we hope to help with finding what makes a song good by exploring the question: What are the features of a song that are important in determining its success? By trying to answer this question it will be easier to know what to search for when trying to find a new song for your playlist.

Previous research into this area has looked at factors such as tempo (LeBlanc et al., 1988), which looked at the effects of tempo on music preference. The study determined that tempo was an important factor so we will use that in our analysis. However, this study focuses on preferences among different age groups but we want to be able to determine overall popularity. Therefore our method is more similar to Bethapudi (2024), which looked at Spotify music data to determine that acousticness, valence, and loudness were high predictors of popularity and they were able to achieve high accuracy. We will be using both Spotify and Youtube music data, looking at similar predictors to this study. Where we differ from this study is that the study focuses on high prediction accuracy which determines important factors but isn't very easy to interpret and doesn't answer what levels of acousticness, valence, and loudness lead to popular songs. We will instead focus on creating a model that is more balanced so it is predictive and can still be somewhat interpreted.

Methods

In this study, we used the data from Kaggle which contains various attributes of songs and the corresponding view counts from YouTube. The attributes include views, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration_ms. The dataset had many values so we specifically chose these predictors following a similar study by Çimen, A., & Kayış, E. (2021, July 22) which used the same predictors to achieve a successful result. First, to clean the data, we removed rows with zero views; rows with missing values. Then, to validate our model later on, we split the cleaned data into training and testing sets using a 60/40 ratio since our dataset is relatively big. Also, to prevent the predictors from being highly collinear, we calculated the Variance Inflation Factor (VIF) for each predictor variable since multicollinearity affects the interpretation of the results. These data-cleaning processes were similar to both studies conducted by Çimen, A., & Kayış, E. (2021, July 22) and Bethapudi, P. (2024) on music popularity. Then, to get the basic idea of the dataset, we developed an initial model where we fit a multiple linear regression model with the attribute "Views" as the dependent variable and all other attributes as the independent variables. This initial model helped us understand the basic relationships and served as a good baseline model for later analysis. From this model, using Cook’s distance and standardized residuals, we identified outliers and observations with a Cook’s distance greater than a certain threshold were removed from the dataset to improve model accuracy. However, given the Q-Q plot not being linear and the residual vs fitted plot following a trend, we used the Box Cox method to determine that a log transformation was the best transformation. A linear regression model then was fitted using this new variable which helped stabilize the variance and achieve a better model fit. Then, as a comparison model, we fit a polynomial regression model to account for potential non-linear relationships between the predictors and the response variable. This model specifically included second-degree polynomial terms for each predictor. We then compared the log-transformed linear model to the polynomial regression model using partial F test and we saw that the polynomial regression model had a better fit. Based on the F test results, it came down to the model reduction where we removed the non-significant predictors and made a reduced polynomial model. To confirm that this reduced model is not performing worse, we conducted another partial F test on the full polynomial model and the reduced polynomial model where we ensured that the reduced polynomial model is not compromising the model’s explanatory power. Lastly, the final model selection was guided by AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to choose the best model to generalize new data. We concluded that the reduced polynomial model is the best-suited model and found that it performs around the same on the test set as the training set.

Results

Danceability	Energy	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Duration_ms
1.608108	3.439921	3.241914	1.083586	1.910722	1.572255	1.066807	1.517134	1.06301	1.013938

Table 1: VIF for all predictors

We used the standard “VIF > 5” threshold which can be found in table 1, and we saw that none of these VIF values corresponded to that. This meant that we didn’t have to investigate these variables for problematic multicollinearity.

The full model is:

$$\text{Views} = B_0 + B_1\text{Energy} + B_2\text{Loundness} + B_3\text{Speechiness} + B_4\text{Acousticness} + B_5\text{Instrumentalness} + B_6\text{Valence} + B_7\text{Tempo} + B_8\text{Duration\_ms} + \varepsilon$$

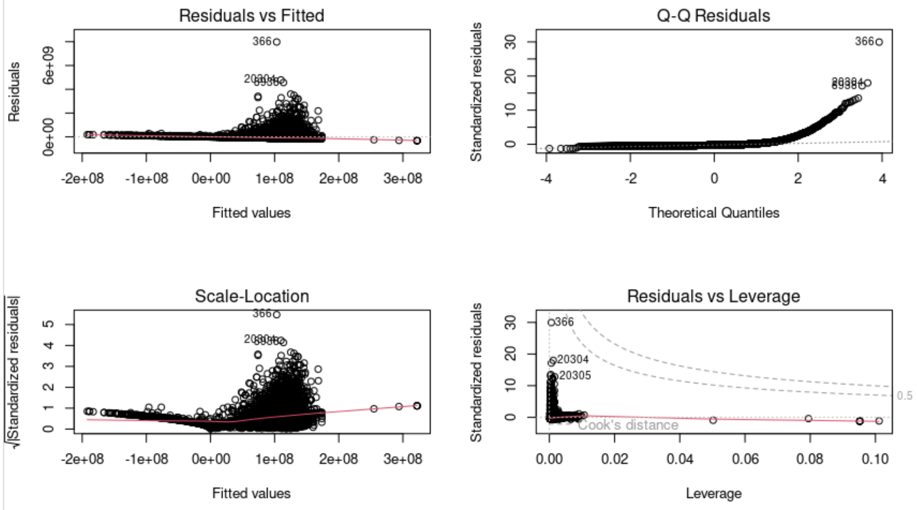


Figure 1: Full Model plots

Based on the residuals vs fitted graph for the full model(figure 1), there is a pattern which suggests that constant variance is violated. The QQ plot also suggests that normality is violated. We decided to use a transformation which was guided by the box cox method which gave an optimal lambda of 0.1010101. Since we wanted to keep the model simple and the optimal lambda was close to 0, we decided to use a log transformation. The residuals vs leverage plot also suggests that there could be some problematic points.

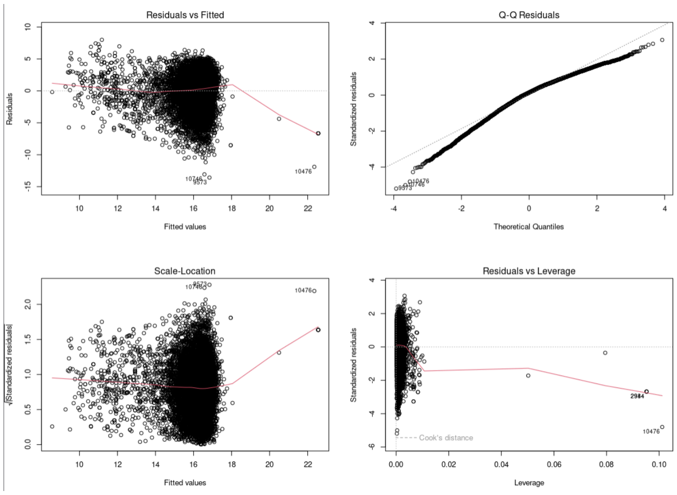


Figure 2: Log Model plots

The QQ plot for the log transformation is linear which suggests that the normality criteria is met. The Residuals vs fitted plot is mostly randomly distributed around zero except near the end which indicates that the model might not fully explain the data. In order to resolve this we used a polynomial model by adding quadratic terms in order to keep the model as simple as possible.

Using a partial f test comparing the linear model with the polynomial model where H_0: The linear model is sufficient. We obtained an F score of 42.501 and a p value < 2.2e-16. Since the p value for the F test is < 0.05, we can reject the null hypothesis, thus the polynomial model is sufficient. By using the polynomial model, the residuals vs fitted plot is randomly distributed around 0.

Based on the t test result of the polynomial model which is summarized in table 3 in the appendix, some of the predictors are not significant. This implies that the model can be reduced and so by removing those predictors and using a partial F test to compare the reduced vs full polynomial model, we obtained an F score of 1.399 and a p value of 0.2314 with H_0: The reduced model is sufficient. Since, the p value is > 0.05 so we cannot reject the null hypothesis indicating that the reduced model is sufficient.

Original Full Model		Log Transformed Full Model		Log transformed polynomial model		Log transformed reduced polynomial	
Summary Statistic	**Value**	**Summary Statistic**	**Value**	**Summary Statistic**	**Value**	**Summary Statistic**	**Value**
Residual standard error	266600000	Residual standard error	2.615	Residual standard error	2.569	Residual standard error	2.57
Multiple R-squared	0.02173	Multiple R-squared	0.1263	Multiple R-squared	0.1569	Multiple R-squared	0.1565
Adjusted R-squared	0.02089	Adjusted R-squared	0.1255	Adjusted R-squared	0.1555	Adjusted R-squared	0.1554
F-statistic (on 10 and 11703 DF)	25.99	F-statistic (on 10 and 11703 DF)	169.2	F-statistic (on 20 and 11693 DF)	108.8	F-statistic (on 16 and 11697 DF)	135.7
p-value (F-statistic)	< 2.2e-16	p-value (F-statistic)	< 2.2e-16	p-value (F-statistic)	< 2.2e-16	p-value (F-statistic)	< 2.2e-16

Table 2: Summary of Models

We can compare all the models by using the adjusted R^2 value to account for increasing model complexity and the F statistic. From the data in table 3, By using the log transformation, the adjusted R^2 value goes up significantly from the original model. Then, by using a polynomial model the F statistic goes down slightly while the R^2 goes up slightly. Finally the reduced polynomial model has a similar R^2 to the full polynomial model but a higher f-statistic. Based on the results, all models are sufficient but the reduced polynomial model is the best for prediction as it has a similar R^2 to the full polynomial model while being more simple for interpretation.

Coefficient	B_0	Danceability	Energy	Loudness	Loudness^2	Speechiness	Speechiness^2	Acousticness	Acousticness^2	Instrumentalness	Instrumentalness^2	Valence^2	Tempo	Duration_ms	Duration_ms^2
Estimate	15.95	1.283	-1.448	83.8	23.63	-25.3	-23.6	-11.56	9.436	-27.8	13.96	7.73	0.002	21.84	-39.4
Std. Error	0.204	0.167	0.199	4.699	2.897	2.688	2.654	3.612	2.948	3.345	2.65	2.727	0.001	2.594	2.612
Confidence Interval	[15.550, 16.350]	[0.957, 1.609]	[-1.840, -1.056]	[74.590, 93.010]	[17.953, 29.307]	[-30.569, -20.031]	[-28.812, -18.388]	[-18.661, -4.459]	[3.658, 15.214]	[-34.356, -21.244]	[8.766, 19.154]	[2.385, 13.075]	[0.001, 0.004]	[16.755, 26.925]	[-44.520, -34.280]

Table 3: Results for reduced polynomial model

Using the results from table 3, the final model is:

$$\text{Log(Views)} = 15.95 + 1.283 * \text{Danceability} - 1.448 * \text{Energy} + 83.80 * \text{Loudness} + 23.63 * \text{Loudness}^2 - 25.30 * \text{Speechiness} - 23.60 * \text{Speechiness}^2 - 11.56 * \text{Acousticness} - 9.436 * \text{Acousticness}^2 - 27.80 * \text{Instrumentalness} + 13.96 * \text{Instrumentalness}^2 - 7.730 * \text{Valence}^2 + 0.00195 * \text{Tempo} + 21.84 * \text{Duration\_ms} - 39.40 * \text{Duration\_ms}^2$$

Model	AIC	BIC
Full model	55772.31	55860.74
Polynomial model	55374.1	55536.21
Reduced polynomial model	55371.7	55504.34

Table 4: AIC and BIC values for all the models after log transform

In order to validate the final model, we compared the AIC and BIC values for all the models. From table 4, since the reduced polynomial model had the lowest AIC and BIC values, we can say that it is the best model.

We then tested the final model on the test dataset which resulted in an adjusted R^2 of 0.131 and a F statistic of 74.86 with a p value < 2.2e-16 which is < 0.05, therefore we can reject the null meaning that the model is statistically significant.

Discussion

The initial purpose of this research, we wanted to determine the popularity of a song with the characteristics of the song. Looking at the p-values in the chart, we can see that all the characteristics are significant since they are all less than 0.05. The intercept, which is the baseline level of popularity when all other characteristics are zero is 15.95. For danceability, there was a weak positive correlation to popularity which makes sense since music is always linked to dancing, but everyone has a different music taste. Similar to danceability, energy had a weak positive correlation to popularity. For loudness which had two terms, both terms indicate a relatively strong correlation to popularity. We interpret this as a “music mastering” issue. Mastering is a process in music creation where one polishes the track by increasing the volume and adjusting the overall sound of the track to match the average volume of the listeners. This means that relatively quiet tracks failed to go through this process leading to a decrease in the quality of listening. Also, famous music producers and artists never skip this process meaning that loudness can be a predictor of music popularity. Speechiness had two terms and both terms had a relatively high negative correlation to popularity. Similar to speechiness, acousticness had two terms where both terms had a relatively high negative correlation to popularity. Instrumentalness had two terms where the first term showed a negative correlation to popularity and the second term showed a positive correlation to popularity. For valence, there was a relatively weak negative correlation to popularity. For tempo, there was a weak positive correlation to popularity which is similar to danceability and energy. For duration_ms, there were two terms and the first term indicated a relatively strong positive correlation to popularity while the second term indicated a relatively strong negative correlation to popularity. We interpreted it as such: since it is difficult to put all the elements into a song in a short duration, increasing the duration helps the quality of the song until it reaches the point where the track feels dragged out and lengthy which drops the overall enjoyment. In summary, the two highest correlated predictors of popularity were loudness and duration.

One limitation that we ran into was the fact that our R^2 value was on the lower side meaning that the model is not as good at explaining the variability in the dependent variable. Also, it was difficult to interpret some characteristics such as speechiness since both terms had a relatively strong negative correlation to popularity which contradicted since the most popular genres at the moment are rap and pop. Acousticness and valence were also difficult to interpret since the negative correlation could not be explained well. Also, another limitation is in which predictors were used. Bethapudi, P. (2024) had other predictors such as music listened to at certain times of the month which could explain the results that we could not interpret. For example, they found that valence, which is how happy the song is, is more highly correlated in the winter season due to Christmas.

References

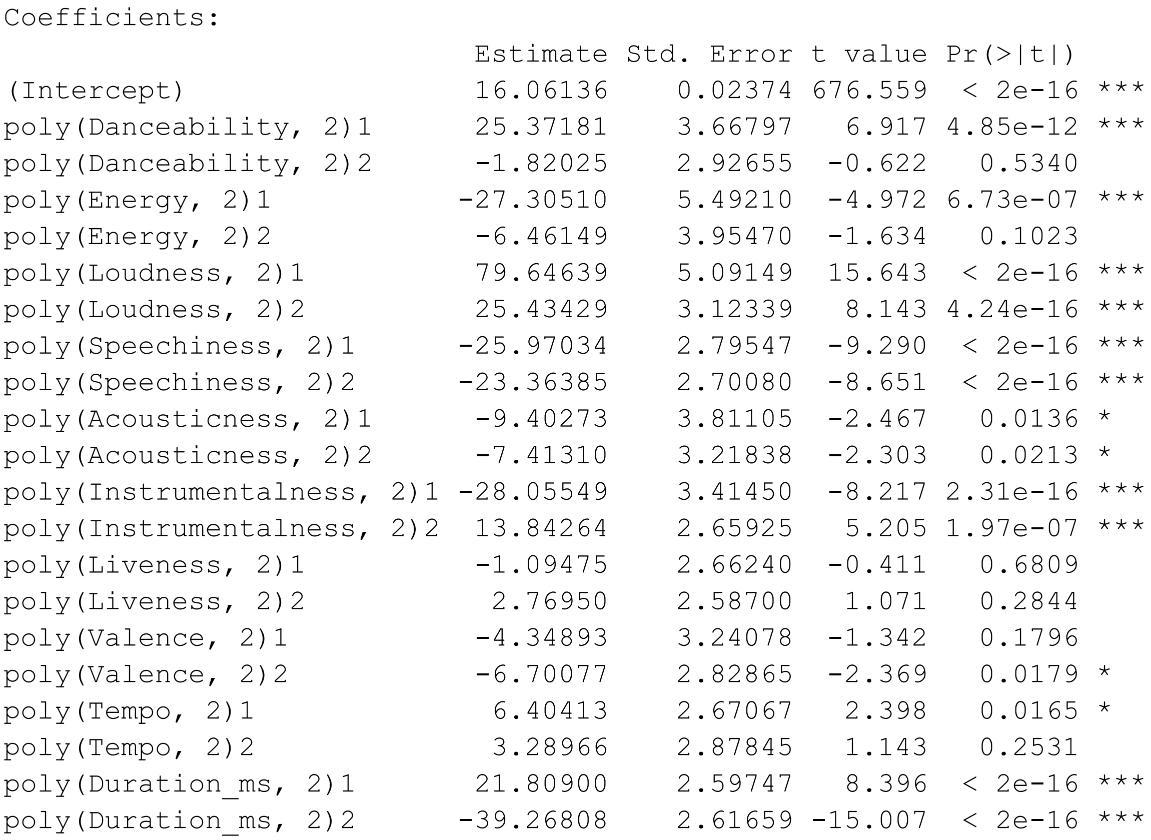
Bethapudi, P. (2024). Spotify data analysis and song popularity prediction. SSRN. http://dx.doi.org/10.2139/ssrn.4793176

Çimen, A., & Kayış, E. (2021, July 22). A longitudinal model for song popularity prediction. https://doi.org/10.5220/0010607700960104

LeBlanc, A., Colman, J., McCrary, J., Sherrill, C., & Malin, S. (1988). Tempo preferences of different age music listeners. Journal of Research in Music Education, 36(3), 156-168. https://doi.org/10.2307/3344637

Rastelli, S. (2023, March 20). Spotify and YouTube. Kaggle. https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Appendix



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 16.06136    0.02374 676.559  < 2e-16 ***
poly(Danceability, 2)1      25.37181    3.66797   6.917 4.85e-12 ***
poly(Danceability, 2)2      -1.82025    2.92655  -0.622   0.5340    
poly(Energy, 2)1           -27.30510    5.49210  -4.972 6.73e-07 ***
poly(Energy, 2)2            -6.46149    3.95470  -1.634   0.1023    
poly(Loudness, 2)1          79.64639    5.09149  15.643  < 2e-16 ***
poly(Loudness, 2)2          25.43429    3.12339   8.143 4.24e-16 ***
poly(Speechiness, 2)1      -25.97034    2.79547  -9.290  < 2e-16 ***
poly(Speechiness, 2)2      -23.36385    2.70080  -8.651  < 2e-16 ***
poly(Acousticness, 2)1      -9.40273    3.81105  -2.467   0.0136 *  
poly(Acousticness, 2)2      -7.41310    3.21838  -2.303   0.0213 *  
poly(Instrumentalness, 2)1 -28.05549    3.41450  -8.217 2.31e-16 ***
poly(Instrumentalness, 2)2  13.84264    2.65925   5.205 1.97e-07 ***
poly(Liveness, 2)1          -1.09475    2.66240  -0.411   0.6809    
poly(Liveness, 2)2           2.76950    2.58700   1.071   0.2844    
poly(Valence, 2)1           -4.34893    3.24078  -1.342   0.1796    
poly(Valence, 2)2           -6.70077    2.82865  -2.369   0.0179 *  
poly(Tempo, 2)1              6.40413    2.67067   2.398   0.0165 *  
poly(Tempo, 2)2              3.28966    2.87845   1.143   0.2531    
poly(Duration_ms, 2)1       21.80900    2.59747   8.396  < 2e-16 ***
poly(Duration_ms, 2)2      -39.26808    2.61659 -15.007  < 2e-16 ***
```

Figure 3: Predictors and t test for full polynomial model