Predicting Music Popularity Through Spotify Song Descriptions

Trevor Goodnight – Graduating May 2023

Faculty Advisors: Michael Frankel, Dr. Marla Bell

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

INTRODUCTION

Currently, Spotify is considered as the most popular mobile app when it comes to streaming music and podcasts in the United States and boasts about having over 100 million songs in its library for users to stream. In 2021, a sample of approximately 586,000 of these songs were selected from Spotify and each song's release date, popularity, duration, tempo, loudness level, and several other factors were recorded. Musical producers and artists in today's music scene are always looking to see if they can answer the question of what makes a song popular as trends always seem to change. This analysis will investigate and attempt to answer this question as well as see if it is, in fact, possible to create a popular song solely based off of the song's statistics.

METHODS

Data Wrangling/Manipulation: used to transform quantitative variables into categorical variables, as well as filter and clean variables in the data that have missing/coded values. These transformations were performed in order to do all three methods in this analysis.

Logistic Regression: used to create a potential predictive model of song popularity using a song's measurement of its danceability, energy level, loudness level, speech level, acoustic level, instrument level, liveliness level, positivity level, tempo, duration, modality, and explicitness.

One-Way ANOVA: used to see if there is a relationship between a song's popularity on Spotify and the year that the song was released in. The independent variable, which represents the song's release year, was discretized into categories or "Eras," as the dataset being used has songs that were released over a span of 100 years.

Stepwise Multiple Linear Regression: used to create a predictive linear model for a song's popularity based on its danceability, energy level, instrument level, liveliness level, tempo, duration, as well as whether the song was explicit or not

Variable Clustering: used to determine and remove all variables that may pose a cause of multicollinearity when looking at a predictive linear regression model of a song's popularity based on its observed specifications.

References

https://www.businessofapps.com/data/music-streaming-market/

https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks?select=tracks.csv

CODE

GitHub

LinkedIn

Table 1: Results of Variable Clustering to Remove Multicollinearity

8 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	energy	0.8811	0.1249	0.1359
	loudness	0.7503	0.1061	0.2794
	acousticness	0.6962	0.0612	0.3235
Cluster 2	danceability	0.7610	0.0718	0.2574
	valence	0.7610	0.1064	0.2674
Cluster 3	explicit	1.0000	0.0218	0.0000
Cluster 4	speechiness	0.5718	0.0496	0.4506
	duration_s	0.5718	0.0252	0.4393
Cluster 5	liveness	1.0000	0.0216	0.0000
Cluster 6	mode	1.0000	0.0033	0.0000
Cluster 7	tempo	1.0000	0.0483	0.0000
Cluster 8	instrumentalness	1.0000	0.0607	0.0000

Table 2: Final Model from Stepwise Multiple Linear Regression

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	16.33326	0.14168	115.28	<.0001	0
danceability	1	8.98795	0.14436	62.26	<.0001	1.17684
energy	1	13.81925	0.09645	143.28	<.0001	1.17800
instrumentalness	1	-7.45187	0.09277	-80.33	<.0001	1.07592
liveness	1	-6.70459	0.12054	-55.62	<.0001	1.04455
tempo	1	0.00470	0.00076706	6.13	<.0001	1.07045
duration_s	1	0.00707	0.00018764	37.69	<.0001	1.02049
explicit	1	14.23878	0.10552	134.93	<.0001	1.03209

Figure 1: Stratified Box-Plot of Song Popularity Stratified by Release Date Era

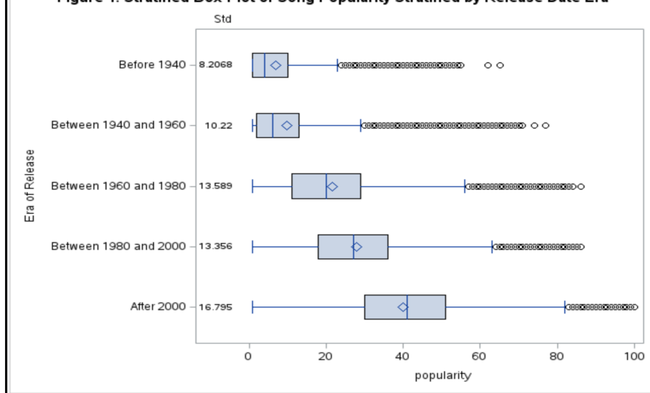

Figure 2: Results of Tukey Post-Hoc Test for One-Way ANOVA

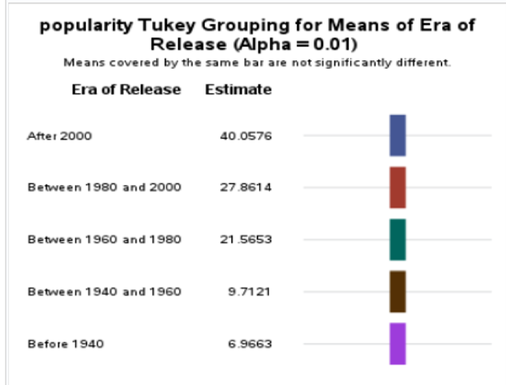

Figure 3: One-Way ANOVA Results of Popularity Based on Release Date Era

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	45848350.4	11462087.6	53845.1	<.0001
Error	541976	115371350.3	212.9		
Corrected Total	541980	161219700.7			

R-Square	Coeff Var	Root MSE	popularity Mean
0.284384	48.88893	14.59012	29.84341

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Era of Release	4	45848350.41	11462087.60	53845.1	<.0001

Table 3: Final Model of Stepwise Logistic Regression

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8536	0.0386	488.4371	<.0001
danceability	1	1.9447	0.0289	4522.0577	<.0001
energy	1	-0.1460	0.0308	22.5284	<.0001
loudness	1	0.0848	0.00135	3951.1515	<.0001
speechiness	1	0.1155	0.0230	25.3153	<.0001
acousticness	1	-1.0543	0.0156	4585.3297	<.0001
instrumentalness	1	-0.4408	0.0172	655.7248	<.0001
liveness	1	-0.3461	0.0207	279.6374	<.0001
valence	1	-1.4568	0.0189	5954.8020	<.0001
tempo	1	0.00119	0.000129	84.9322	<.0001
duration_s	1	0.000362	0.000032	126.3221	<.0001
explicit	1	1.1005	0.0225	2389.4357	<.0001
mode	1	0.0209	0.00766	7.4464	0.0064

Figure 4: ROC Curve of Logistic Model

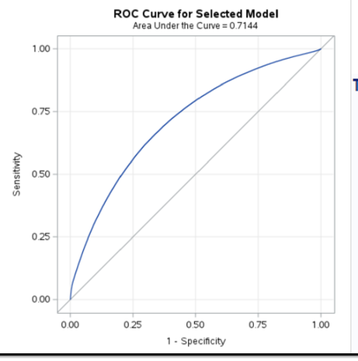

Table 4: Results of Logistic Regression Model Accuracy

The FREQ Procedure

Frequency / Percent	Table of Popular by preds		
	preds		
Popular	0	1	Total
0	63065 33.25	33328 17.57	96393 50.82
1	31341 16.52	61958 32.66	93299 49.18
Total	94406 49.77	95286 50.23	189692 100.00

Results

Variable Clustering: Table 1indicates that the variables loudness, acousticness, valence, and speechiness should be removed from the linear regression model as their respective 1 - R^2 ratio are the largest in their respective clusters.

Stepwise Multiple Linear Regression on Popularity: Table 2 indicates that, based on the stepwise linear model, songs with a higher energy level and songs that are marked as explicit will have a higher popularity rating.

Stratified Box Plots of Song Popularity by Era of Release: Figure 1 shows that the category "After 2000" has the largest amount variation and the category "Before 1940" has the least amount of variation when it comes to song popularity. The ratio of the standard deviations of these categories equates to approximately 2.04, which can be considered a homogeneous distribution.

One-Way ANOVA of Song Popularity by Era of Release: Figure 2 shows that the One-Way ANOVA that was conducted gives an F-Statistic of 53,845.1 and P-Value of less than 0.0001. This concludes that the Era of a song's release is a significant factor when predicting its popularity.

Tukey's Post-Hoc Test: Figure 3 shows that songs released after 2000 have a significantly higher mean popularity rating than any other era of release. It also indicates that as the song's era of release becomes more current, its popularity rating will be higher.

Logistic Regression: Table 3 shows that, based on the logistic model, explicit songs that have a high danceability level and have a louder sound are more likely to be received well and be considered as popular.

ROC Curve: Figure 4 indicates that the area under the curve of the logistic model created to predict song popularity is 0.7144. This indicates the relation of having false positive and false negative in the logistic model as the rates increase.

Model Prediction Table: Table 4 indicates that after scoring the logistic model, it leads to a 66% accurate "hit rate," in which the Type I and Type II error are 17.57 and 16.52, respectively.

Conclusions

- The following linear model can be considered a potential model for predicting song popularity where multicollinearity is minimized:
  Popularity = 16.333 + 8.068(**Dance**) + 13.819(**Energy**) – 7.452(**Instr.**) – 6.705(**Live**) + 0.005(**Tempo**) + 0.007(**Dur.**) + 14.239(**Exp.**)
- The One-Way ANOVA test showed that there is a significant relationship between the Year a song is released and a song's popularity.
- Using all predictor variables gives us a model that is approximately 66% accurate at predicting song popularity.