Predicting Song Popularity on Spotify: An OLS Regression Analysis in SPSS

Introduction

The music industry has always been a lucrative industry for fame, fortune, and artistic commentary. The success of artists, producers and all those associated with a piece of musical work, heavily relies on how popular their songs are, or how widely their songs are heard. Earlier songs were only sold together in albums which were quite heavily priced and hence sales were largely based on artist popularity but today, with the development of modern streaming services like Spotify and Youtube, the ability of songs to go viral has changed. How many plays or streams have taken over how many sales. Gone are the days of "platinum albums". This has led to a democratization of the music industry by lowering the barriers to access for both creators and consumers of music. Consequently, more people want to break into the industry and achieve success.

But what makes a song popular? Does the success of a song depend on the lyrics? The artist? Or does the musicality of the song plays the biggest? Or is it marketing? In this exploratory analysis I limit the scope of factors affecting popularity to the musicality of the song. I try to discover which audio features are the most important in determining a song's popularity. If we can determine the most important audio features, this could potentially allow streaming services like Spotify to increase customer personalization and better predict which songs will become popular and use this information for promotional and other purposes. More importantly, it could also help artists understand their audiences music taste better and produce more tailored content.

Using the song attributes data from Spotify's Web API, I will use OLS regression analysis using SPSS to answer the research question: ***To what extent does song popularity on Spotify depend on certain audio features?***

Since this is an exploratory research paper, I do not have a hypothesis. I will analyse a numbero f Spotify's audio features to determine which were significant predictors for song popularity on their platform.

Data

I conducted primary analysis of a Spotify audio feature data set. The data set is from Kaggle, an online community that allows users to find and publish a variety of data sets. It is a crowd sourced collection of data sets. The data set used in this analysis was published on Kaggle by Zaheen Hamidani in July 2019. He used Spotify's web API to scrape audio features from their database. The data was collected on April 3, 2019. This is a purposive sample to get an equal sample of songs from a large variety of genres.

The data set is a collection of 232,725 unique songs across 26 genres of music. There are approximately 10,000 per genre. Each song has been analysed by an algorithm to produce song attributes like: duration of song, key, acousticness, danceability, energy, instrumentalness, liveliness, loudness, speechiness, tempo, valence (song positivity) as well as a popularity index.

Measures

All the variables used in this analysis represent interval level data, ideal for regression analysis. They are all calculated using Spotify's proprietary soundwave analysis algorithms and are defined on Spotify's web developer platform.

The dependent variable in the model is **popularity**

The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by a Spotify algorithm and is largely based on the total number of plays the track has had and how recent those plays are. Songs that are being played a lot recently will have a higher popularity than songs that were played a lot in the past. It is important to note that the popularity value may lag actual popularity by a few days: the value is not updated in real time. The popularity rankings for the data set analysed are unique to April 3 2019.

The possible independent variables or predictors in the model are:

- "**Danceability**" (scale of 0 – 1, where 0 = is least suitable for dancing and 1 = most suitable)
- "**Energy**" (scale of 0 – 1, where 0 = low measure of intensity and activity and 1 = most intensity and activity)
- "**Loudness**" (scale of -60 – 0 decibels [db], where -60 = quiet and 0 = loud)
- "**Speechiness**" (scale of 0 – 1, where 0 = more music/non-speech tracks and 1 = more speech-like tone such as poetry, etc. Rap songs have high speechiness)
- "**Acousticness**" (scale of 0 – 1, where 0 = song is not acoustic and 1 = high confidence that song is acoustic)

- "**Instrumentalness**" (scale of 0 – 1, where 0 = song contains a lot of vocal content and 1 = high confidence the song has no vocal content)
- "**Liveness**" (scale of 0 – 1, where 0 = song is not live and 1 = very strong likelihood the song is recorded live)
- "**Valence**" (scale of 0 – 1, where 0 = songs that are more negative, angry, depressed, etc. and 1 = tracks that are positive and cheery)
- "**Tempo**" (the estimated tempo of a song, measured in Beats per Minute)
- "**Duration**" (the length of the song in milliseconds)

Method

Since the aim to the research is to predict song popularity based on certain attributes, OLS regression analysis was the appropriate choice to create a model to understand the explanatory power each of the determining attributes.

However, before proceeding with a regression model I looked at *descriptives* like frequencies and ranges to eliminate variables that did not vary much as well modify the data set to improve analysis.

Examining the *correlation* matrix also pinpointed the variables most strongly correlated to popularity, helped eliminate uncorrelated variables and highlighted concerns for multicollinearity amongst several dependent variables.

The subsequent analysis consisted of making decisions on which variables to include in the *OLS regression* model based on explanatory power they provided as well as accounting for

multicollinearity by eliminating variables and possibly creating an index of conceptually similar independent variables. SPSS was used for all statistical analysis.

Analysis

Descriptive Analysis

The first step was performing descriptive statistics on the various audio features of the songs in the data set as well as the dependent variable, popularity.

It was observed that about 3% of the cases under popularity are 0 and hence 0 values were coded as missing and excluded from the regression analysis. The distribution looks quite close to normal.

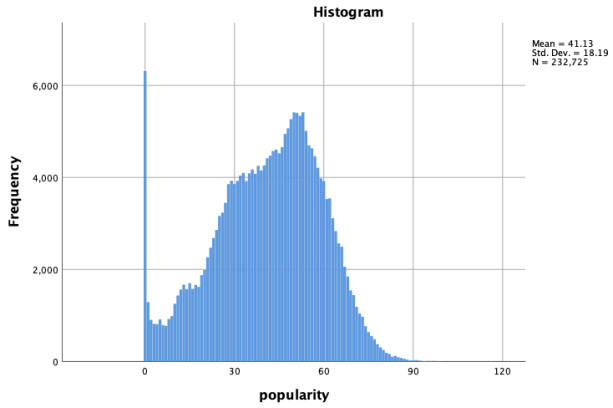

Figure 1: Distribution of the dependent variable popularity

Further, we can see in Figure 2 below that there is very little variation in the durations of songs with most songs being between 3:30 and 4:00 minutes in length, therefore duration is unlikely to have much explanatory power when explaining variation in popularity of songs. The

correlation analysis performed subsequently supported the elimination of the variable duration from the model.

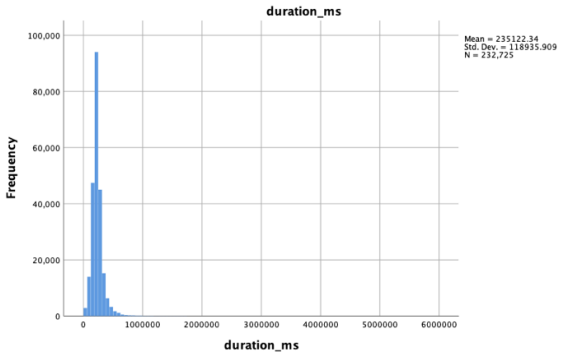

Figure 2: Distribution of the predictor duration

The variables liveness and instrumentalness were also excluded from further analysis. The measures for each represent a level of confidence whether is a track is instrumental or live. It is not a measure of degree on spectrum such as danceability or valence (positivity). In reality whether is song is live or not is a binary concept, i.e., the song is either live or it is not. As is the case with instrumentalness. Hence, even though figure 3 shows a spectrum of values, only the values very close to 1 are live songs and those are extremely few. The values between 0 and 1 represent error in Spotify's algorithm. Similarly instrumentalness (figure 4) shows almost no variation due to the very small number of songs that are instrumental in nature.

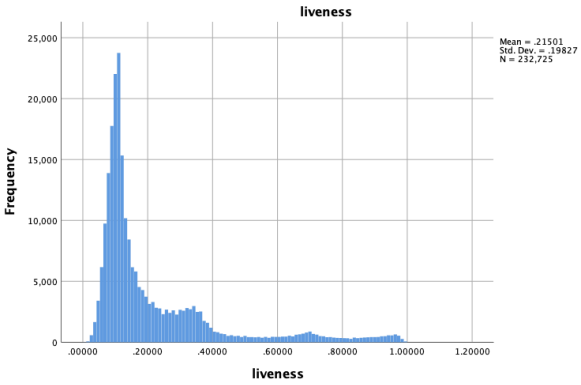

Figure 3: Distribution of liveness

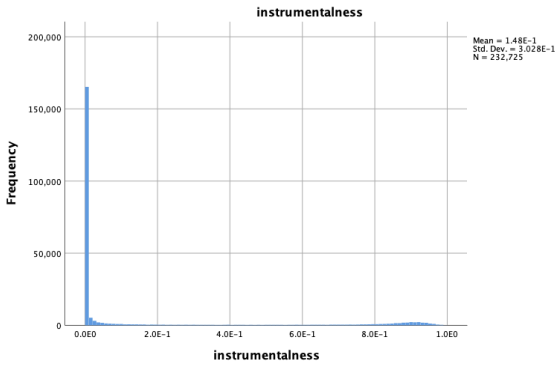

Figure 4: Distribution of instrumentalness

All other variables showed enough variation in values to be considered further as possible predictors in this exploratory analysis.

Correlation Analysis

Using the remaining variables I used a correlation matrix to select independent variables showing sufficient correlation with popularity. The table below shows the correlations of the independent variables with popularity. It should be noted that the extremely large sample size causes all correlations including negligible ones are significant.

	popularity
acousticness	-.362**
danceability	.288**
energy	.217**
loudness	-.350**
speechiness	-.166**
tempo	.085**
valence	.087**

Figure 5: Correlation with popularity

Other than tempo and valence all other variables show promise however, the correlation matrix also shows presence of multicollinearity between many variables. Loudness is highly correlated with acousticness (-.690) and energy (.816). Further loudness is possibly moderately collinear with danceability (.439) and valence (.400).

	loudness
acousticness	-.690**
danceability	.439**
energy	.816**
valence	.400**

Figure 6: Multicollinearity predictors

This makes sense because highly energetic or danceable songs will be louder. While acoustic songs will have lower decibel levels. On running another correlation matrix while controlling for loudness, the predictors danceability, energy and speechiness still showed a good correlation with popularity. Hence, they can be considered while building the predicting model for popularity.

OLS Regression Analysis

While loudness is the most powerful variable that seems to correlate with popularity it is not the most useful as merely the decibel level of the song cannot be helpful to identify a popular song. Other factors need to be considered but the other variables exhibit multicollinearity. A model containing loudness and with all the other variables in the dataset has the highest R^2 value of about 20% This is because when there are so many variables, that each variable adds a little bit to the explanatory power. The model is also significant due to the large sample size.

However, the collinearity statistics show high multicollinearity that needs to be dealt with. Collinearity tolerance values for variables were around .5

Creating an additive index such as DanceValence or LoudEnergy reduces the R^2 value substantially. Since energy was highly correlated with loudness and not collinear with danceability, it is ideal to replace loudness in the model. Multicollinearity was never an issue with speechiness which finds place in the model below along with danceability the remaining variable that showed correlation with popularity.

Model Summary

The R^2 value is 0.149 which means almost 15% of the variance in song popularity can be explained by the regression model containing the predictors – energy, danceability and speechiness.

The ANOVA F statistic value is 13193.361 and is statistically significant at the 0.00 level, hence it can be inferred that the overall regression model is statistically significant.

The MS regression is 3275071.58 while the MS residual is much smaller at 248.236. This shows that the explained variance to be greater than the unexplained variance (error) and so the overall model fit is acceptable.

The standardized regression equation for the model is:

popularity = .264danceability + .164energy -.226speechiness

It is inferred that danceability is the strongest predictor of song popularity is followed by speechiness and energy.

It is interesting that speechiness has a strong negative coefficient given the burgeoning popularity of rap songs and rap additions to pop and R&B music.

The t test tells us whether or not a regression coefficient is statistically significant. The significance level should be p≤.05. In this model, the t tests of all 3 coefficients are significant at the 0.00 level and hence statistically significant. The tables below outline further model details.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	.386[a]	.149	.149	15.756	.149	13193.361	3	226409	.000

a. Predictors: (Constant), danceability, speechiness, energy

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9825214.75	3	3275071.58	13193.361	.000[b]
	Residual	56202938.9	226409	248.236		
	Total	66028153.7	226412			

a. Dependent Variable: popularity

b. Predictors: (Constant), danceability, speechiness, energy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	25.136	.113		222.568	.000		
	speechiness	-20.750	.181	-.226	-114.553	.000	.970	1.031
	energy	10.700	.135	.164	79.309	.000	.880	1.137
	danceability	24.319	.190	.264	127.855	.000	.884	1.131

a. Dependent Variable: popularity

Figure 7: Model for predicting song popularity

Discussion

This research intended to answer the research question '*To what extent does song popularity on Spotify depend on certain audio features?*'. The question was analyzed by using Spotify's audio features taking an attribute based approach to a success prediction model for popularity of songs on Spotify. The results of the correlations showed that there were significant relationships between audio features and song popularity. The found relationships however were generally weak. And many audio features were highly correlated. Next, a regression

model was built from a selection of attributes after eliminated non-significant and multicollinear ones. Its explanatory power (R^2) is 15%, meaning that the model based on the predictors energy, danceability and speechiness explains 15% of the variation in the song popularity. Hence, I can conclude that my model is not as effective in explaining popularity on its own. There are some reasons that likely limit the explanatory power of the model. It could be mainly due to the question we are asking. Since 26 genres were included for measuring streaming popularity, it is likely that, because different genres do not share the same popular attributes, there will be error in the hit prediction model making for a lower R^2 value and lower correlations. However, I did try to use the only the pop songs from this dataset to come up with a popularity prediction model but it had an explanatory power of only 3%. Maybe another variation in data set is required.

This research can better be seen as a starting point for prediction since the analysis shows promising results for prediction with audio features. Future research can develop the work by creating prediction models from these features such as higher order regression models and decision trees for predicting song popularity. This research is limited to audio features extracted by Spotify from its own data. As we saw with the variables liveness and instrumentalness, the algorithms that calculate these values have high error rates. Other algorithms could be better or be used to extract different features that could provide more explanation in the variance of popularity. Additionally we again used popularity as defined my Spotify's index as opposed to something more representative such as the Billboard top 200. That maybe a better operational definition of the concept of popularity. Further, factors such as

song lyrics and artist popularity would contribute highly to popularity of songs. Machine learning would be a very useful tool for popularity prediction as it is adaptive and keeps improving with additional data. With more the 30 million songs on Spotify, machine learning algorithms would probably be the best bet to create a predictor model.

Success prediction in cultural markets and especially for products remains a very complex subject. The novel possibilities of the use of such data will generate substantial value for artists and music producers alike. It could however lead to reduced variation in musical art in order to chase mass popularity and hence reduce choices for consumers.