EN.580.475 Biomedical Data Science
Final Project Report

Marion Pang
mpangwa1
GitHub ID: gnaprs

Predicting Song Popularity using Audio Features by Multivariable Linear Regression

Problem

In this project, we aim to predict the popularity of a song based on its audio features. Specifically, I used extract data from 19000 songs on Spotify using Spotify's in-built API[1], and analyzed each track's structure and musical content, including tempo, acousticness, duration etc. Specifically, we define our variables:

***Popularity**: Calculated by algorithm (total number of plays/how recent plays are) –between 0 and 100, with 100 being the most popular.*

***Tempo**: The overall estimated tempo of the section in beats per minute (BPM).*

***Key**: The estimated overall key of the section.*

***Mode**: Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.*

***Acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.*

***Danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.*

***Energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.*

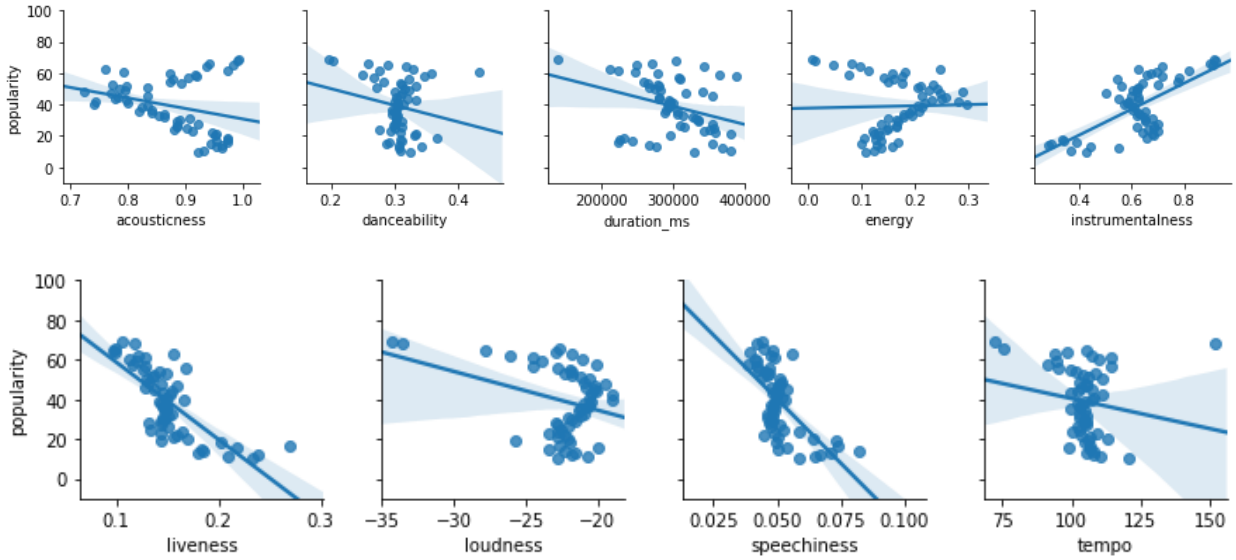
***Instrumentalness**: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".*

***Liveliness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.*

***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.*

Data

Data was extracted from Spotify and loaded into pandas using a dataframe, categorized by genre. Songs were grouped by their popularity and the mean for each (numerical) feature was calculated. Preliminary analysis revealed that Classical music held the strongest single-variable linear correlations with popularity, and thus further analysis was performed using the Classical music dataset of 7929 songs.



Model Fitting

To avoid overfitting, the data was split into training and test sets. A multivariate linear regression model was fitted using training data on a `statsmodels.formula.api` Ordinary Least Squares Regression model, with popularity as the predictor and variables with the strongest linear correlations from before.

```
results = smf.ols('popularity ~ acousticness + energy + instrumentalness + liveness', data =
classicalTrainMean).fit()
```

[1] https://developer.spotify.com/documentation/web-api/reference/tracks/get-several-audio-features/

Marion Pang
mpangwa1
GitHub ID: gnaprs

Coefficients for each variable were obtained from minimizing:

$$\sum\left(Popularity - \beta_0 - \beta_1 x_{acoustic} - \beta_2 x_{speech} - \beta_3 x_{instru} - \beta_4 x_{live}\right)^2$$

The interpretation of the regression coefficient is the estimated expected change in the outcome for a single unit change in the regressor with all other regressors held constant i.e. for β_1, it is the estimated change in popularity for a single unit change in acousticness. From our model's fit, we have:

$$\beta_0 = 104.0, \beta_1 = 59.02, \beta_2 = -143.6, \beta_3 = 33.39, \beta_4 = -185.6$$

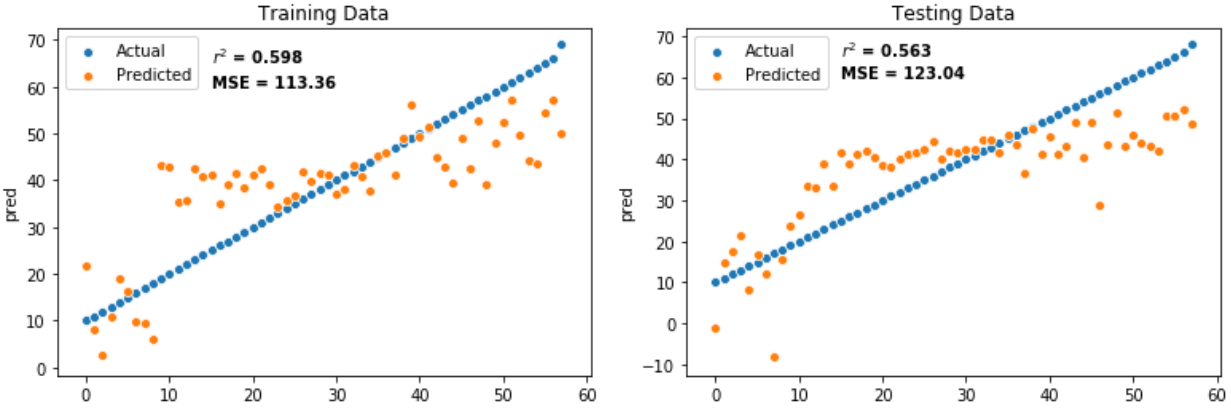
Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.568
Dependent Variable:	popularity	AIC:	448.9744
Date:	2019-10-06 21:40	BIC:	459.2766
No. Observations:	58	Log-Likelihood:	-219.49
Df Model:	4	F-statistic:	19.74
Df Residuals:	53	Prob (F-statistic):	5.34e-10
R-squared:	0.598	Scale:	124.06

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	104.0369	18.9055	5.5030	0.0000	66.1174	141.9565
acousticness	-59.0218	18.1056	-3.2599	0.0020	-95.3370	-22.7066
speechiness	-143.6381	131.5060	-1.0923	0.2797	-407.4058	120.1296
instrumentalness	33.3890	12.5054	2.6700	0.0100	8.3064	58.4716
liveness	-185.6042	47.3051	-3.9236	0.0003	-280.4861	-90.7222

Omnibus:	0.959	Durbin-Watson:	0.494
Prob(Omnibus):	0.619	Jarque-Bera (JB):	1.001
Skew:	-0.196	Prob(JB):	0.606
Kurtosis:	2.489	Condition No.:	134

Predicted popularities were evaluated using the model on both the training and test set, and the mean squared error was evaluated. Results predicted from the training set and the testing set achieved r^2 coefficients of 0.598 and 0.563, and Mean-squared errors of 113.36 and 123.04 respectively.



Interpretation

Thus, we have obtained a model that predicts a Classical song's popularity relatively accurately based on song metrics such as acousticness, speechiness, instrumentalness and liveness. It may thus be possible to predict the popularity of a new (Classical) song on spotify by analyzing the metrics and using the model. From the model, it appears that less speech and liveness (i.e. a better recording) and greater instrumentation tends to correlate with a higher popularity.

However, we do note that the model only applies to a subset of songs – specifically, I removed songs with less than 20 popularity from analysis, in other to obtain a less skewed distribution of song popularity. In addition, further analysis can be done to identify more features that affect song popularity, such as Key or Mode (in supplementary data, not in linear regression). Lastly, this form of analysis can also be applied to the other 26 song genres in the data set, to identify any interesting trends.