Predicting the Next Dance Craze?

Modeling Music Danceability Based on Song Metrics

Sebastian Rivera-Chepetla, Ramiro Romero, Arnav Talukder, Jack Keeton, Gary Ramos, and Daniel Smith

Background

Newer social media platforms such as TikTok and Instagram reels have been centers for the promotion of new musical trends because of their ability to make a song go viral. Consequently, the quick and constant upload of musical media to these platforms has created a culture where trends are short-lived and difficult to define characteristics of songs such as their "catchiness" or "danceability" dictate whether they have the opportunity to go viral. This lucrative interplay between social media platforms and the music industry has made it more important than ever to be able to effectively and accurately predict which songs have the best chance at going viral.

Our dataset is sourced from Kaggle, and contains a list of 2,000 songs generated by Spotify, a leading music service. Each song was curated by Spotify's AI for a "Top Hits of the 2000's" or "Top Hits of the 2010's" playlist. It features a number of song metadata variables including duration (in milliseconds), popularity, energy, loudness, key, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and our variable of focus, danceability.

Hypothesis

H_0:
H_a:

Danceability is determined by Spotify as a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.

Though we were initially surprised to observe a lack of apparent strong relationships in our data, we were able to conduct hypothesis testing that gave us confidence to reject the null hypothesis and conclude that danceability is sigificantly related to at least some of our other variables.

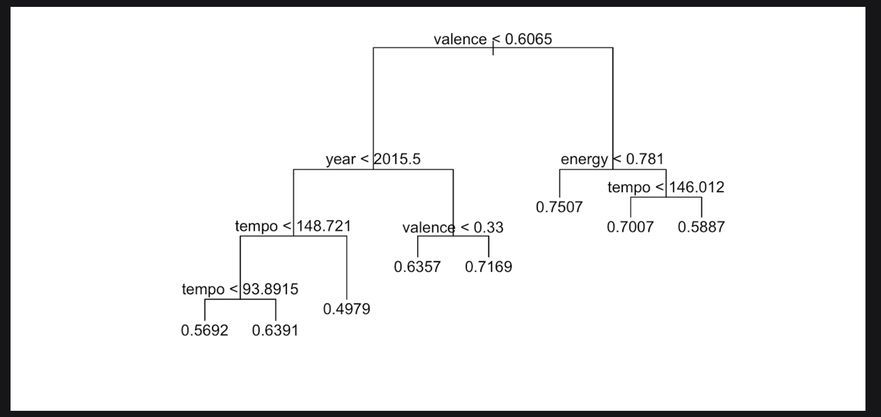
Methods

We ran a total of 8 different models for our predictions. Based on information gained from PCR testing, we switched our focus from regression modeling to developing more powerful methods, namely using a Random Forest. For our modeling we ran models with three variations of the data, the original data (data1), data with a boolean for each genre (data2), and one where a new genre variable was added to the original data (data3). The three different versions of the data were split into training and test data with a 70/30 ratio.

Model

Ultimately, in our decision to implement a more complicated and powerful model than simple linear regression could provide alone, we opted for a Random Forest method because we believed it offered the flexibility we needed for our large set of variables and, crucially, would provide the most insight into which variables were most useful for generating accurate predictions. Hence, from our Random Forest models we determined that the most important variables in predicting danceability were "energy", "valence" and "tempo".

Conclusion

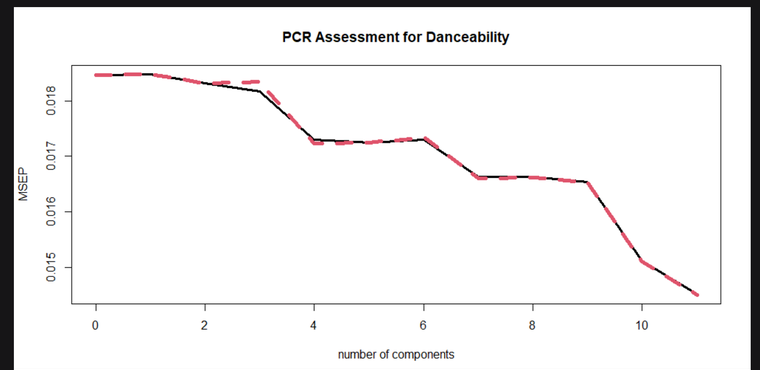


We were able to create a model that effectively predicts the danceability score of a song.

Fundamentally, we set out with the goal of analyzing a 2000 song long list of Spotify data with the intention that we could identify key relationships between certain known song metrics and a useful indicator variable that a service like Spotify may wish to market with or otherwise expand on. For these ends, our indicator was "danceability", a variable designed by Spotify based on other song metrics such as rhythm and, as the name would suggest, intending to assess how easy it is for people to dance along with a song. In the course of our experimentation, we confirmed that, using general Spotify metadata such as the information provided in this dataset, danceability could be modeled effectively, with three most important variables for prediction being, "energy", "valence" and "tempo". This finding is particularly insightful given that our early analysis found strong relationships between our variables to be generally lacking despite our initial preconceived notions. Hence, we are satisfied with our cutting through the noise of data and discovering a fruitful model that could assist Spotify in how it recommends music to its customers or perhaps even inform artists themselves of how to better tailor their songs for mass consumption.

our code

Analysis



Model	MSE
Simple linear regression	0.0146909
PCR Model	0.0189560
Tree with Data (1)	0.0153234
Tree with Data (2)	0.0153234
Random Forest with Data (1)	0.0112762
Random Forest with Data (2)	0.0110524
Tree with Data (3)	0.0153234
Random Forest with Data (3)	0.0099611

While the MSE's of all the models we ran were between 0.009 to 0.019 it was seen that the random forest model with the third variation of the data gave the lowest MSE. Given that the danceability values only ranged from 0.1290 to 0.9750 we concluded that our more complex random forest model would be better at predicting danceability due to it having the lowest MSE of approximately 0.0099. Ultimately, compared to simple linear regression, using a Random Forest method allowed us to increase our model R square value from 0.232 to 0.452, an increase that not only exceeded our expectations, but is well in line to be an excellent model by established sociological study standards (Ozili).

References

Ozili, Peterson. (2022). The Acceptable R-Square in Empirical Modelling for Social Science Research. SSRN Electronic Journal. 10.2139/ssrn.4128165.

UCLA