Benjamin Hanim 12/15/2023

Professor Schweikert Data Mining & Predictive Modeling

**Spotify Data Mining Research**

The goals of this research are to determine the most popular genres/subgenres of each year/decade and to see whether they are consistent throughout a decade. If they are, then I want to determine the greatest advanced attribute indicators of a popular song in each decade(including danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration). Using these results, I will then compare what is a better indicator of what period in time a song comes from, the genre it belongs to, or the advanced statistics.

First, I want to determine which playlist genres and subgenres are the most popular across all years of release dates to see what kind of music from the 60s, 70s, 80s, etc. is still most listened to now. I hypothesize that the 60s will feature rock as the Beatles ruled this time, the 70s will feature disco and hard rock, the 80s may feature techno-pop music, the 90s will likely include R&B and maybe an introduction to pop, the 2000s will be ruled by pop, and the 2010s will be a mix of pop and rap. Since my hypothesis alone may not have any legitimacy as I do not have an amazing knowledge of music dating back to the '60s, I used an article called *The Evolution of Popular Music.* Using this, we should have rock and roll featured in the 60s, punk in the 70s, hip-hop in the 80s, a mix of hip-hop, rock, and pop in the 90s, pop/hip-hop in the 2000s, and the article doesn’t cover the 2010s, but another article *Top 2010s Music Trends: Exploring the Evolution and Influences*, says that pop continues to dominate music, along with some more genre diversity than the past.

Since this dataset has many advanced attributes such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration, another goal of mine was to see what type of stats in music determined a hit throughout the decades. Of the attributes, I expect duration to have a bigger impact on music in recent decades as, throughout my lifetime, 3 minutes has been the sweet spot for many musicians to make a hit song. I think it likely varied more in the past since I have heard old songs that can be 7 minutes and some that are 3 minutes. Since we are not finding the actual duration, but the information value, I used an article called *Infographic: The Shorter the Song, the Sweeter the Stream?*, where a larger Spotify data set was analyzed, and the average song duration was calculated throughout the decades. Surprisingly, they found that from the 1930s through the 1990s, the song duration increased, and from 1990 to the 2010s, the song duration decreased back to the same duration as the 30s. We will see if the information value is linked in any way to the actual duration, I predict that the 90s will have the highest information value for the duration. Danceability is likely a big indicator of a hit song in every decade (I expect a peak in the 70s-80s for disco music). Additionally, I think energy will peak in the 70s and 80s(because of the Hard Rock) and tempo will likely have a peak in the 80s and 90s(because of the rise in hip-hop). One attribute that is interesting to take a look at is liveness. This determines whether there is an audience or not. I used an article called *Top 20 Greatest Live Musical Performances Ever* to determine when live performances were at their peak. I noticed that the 60s, 90s, and 2000s had the greatest performances. I assume that these years will have a higher liveness rating.

**1960s Genre Analysis:**

 

A close-up of a computer screen

Description automatically generated

Prediction: Rock

Result: Rock, some R&B, and some Pop (Very accurate 5/5)

**1970s Genre Analysis:**

A black text on a white background

Description automatically generated

Prediction: Disco, Hard Rock, Punk

Result: Rock, some R&B, Electropop (Somewhat accurate 3/5)

**1980s Genre Analysis:**

A white background with black text

Description automatically generated

Prediction: Techno-pop and Hip Hop

Result: Pop, Electro House, R&B (aka Hip Hop), and some Latin (Pretty Accurate 4/5)

**1990s Genre Analysis:**

A white background with black text

Description automatically generated

Prediction: Hip Hop, Rock, Pop  
Result: Pop, Rap (aka Hip Hop), Rock (Very accurate 5/5)

**2000s Genre Analysis:**

A computer screen shot of words

Description automatically generated

Prediction: Pop, Hip Hop

Result: Pop, EDM, Rock, Latin (Not very accurate 2/5)

**2010s Genre Analysis:**

A computer screen shot of words

Description automatically generated

Prediction: Pop, Rap, and more musical diversity than ever

Result: Pop, Rap, Latin (Pretty accurate 4/5)

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated A computer screen shot of a number

Description automatically generated

Predictions:

* Recent decades will have the highest information value for the duration.
* Danceability is likely a big indicator of a hit song in every decade (I expect a peak in the 70s-80s for disco music).
* Tempo will peak in 80s/90s, and energy will peak in the 70s/80s.
* The 60s, 90s, and 2000s will have a higher liveness rating.

Results:

* Duration had the greatest information value in the 1970s and the 2010s (Somewhat accurate 3/5)
* Danceability was by far the most indicative factor of a hit song in the 60s, it was also significant in the 70s and 2000s, but not as significant as expected across other decades. (Somewhat accurate 3/5)
* Surprisingly, tempo is a much greater indicator in the 2000s and 2010s than in years prior, however, energy did have a peak indication in the 70s and 80s as predicted. (Somewhat accurate 3/5)
* The 90s, 2000s, and 70s had the highest liveness ratings as compared to our prediction of the 90s, 2000s, and 60s. (Pretty accurate 4/5)

**Prediction:**

Having analyzed these two sets of data (genres/subgenres), and the advanced data, I wanted to see which metric would be better at predicting which decade a random song belongs to, its advanced stats, or its genre/subgenre. Now, having analyzed both options, I think neither one would have greater than a 60-80% success rate. This is because all decades of music (except maybe the 60s) have a great variety in the types of music. Because of this, along with the fact that the decision trees involving advanced attributes will likely be a lot larger due to less possible variation with categorical attributes as compared to continuous attributes, I think the advanced attributes will have a greater accuracy in determining the decade that a song belongs to.

To see how well a decision tree predicts the decade a song is from is likely to be less accurate if the song is not as popular, because popular songs must have some overlap as we saw in the previous data. For this reason, I wanted to pinpoint a few popularities and see how the two groups compare in each. We are going to start with fairly popular songs, with a popularity>50, then see all songs, with a popularity>=0, and finally very popular songs with a popularity>65.

Popularities>50:

**Popularity > 50:**

Genres

A screenshot of a computer program

Description automatically generated

Advanced Attributes

A screen shot of a computer program

Description automatically generated

**Popularity>=0:**

Genres

A screenshot of a computer program

Description automatically generated

Advanced Attributes

A screenshot of a computer program

Description automatically generated

**Popularity>65:**

Genres

A screenshot of a computer program

Description automatically generated

Advanced Attributes

A screen shot of a computer code

Description automatically generated

**Results:**

When analyzing decision trees with somewhat popular songs (popularity>50), it turns out the genres/subgenres can predict a song’s decade with 79.5% accuracy, while the advanced attributes can predict songs slightly better, with an 80.5% accuracy. This is not a very significant difference. Diving deeper into the data, when the popularity measure is greater than or equal to 0, the advanced attributes’ decision tree has an accuracy of 75.5% as compared to the genre/subgenre’s decision tree accuracy of 77.3%. This shows that genres have a significantly bigger impact on predicting the decade a song was created in when a song is not popular. When making the popularity very high (above 65), the advanced attributes have a decision tree with 87.3% accuracy as compared to the genre/subgenre’s decision tree of 81.5% accuracy. This shows that advanced attributes predict popular songs much better than genres do. We do see, however, both sets of data can predict the decade a song was made decently well. While genres are a better indicator of the decade a song was made in at a popularity score of 0 or greater, it is still not as accurate at assessing all songs as compared to assessing popular songs only. This makes sense as a genre will be fairly consistent throughout a decade (77.3%-81.5%), but advanced attributes will become increasingly accurate as popularity improves (75.5%-87.3%) because popular songs have more overlapping attributes with each other than they do with other songs.

**Bibliography**

***Dataset***

Arvidsson, J. (2023, November 1). *30000 Spotify Songs*. Kaggle. https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs

***Algorithms***

@article{scikit-learn,

title={Scikit-learn: Machine Learning in Python},

author={Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.},

journal={Journal of Machine Learning Research},

volume={12},

pages={2825--2830},

year={2011},

note={Mutual Information computation for scikit-learn: `mutual\_info\_regression` function}

}

@article{breiman1984classification,

title={Classification and Regression Trees},

author={Breiman, L. and Friedman, J.H. and Olshen, R.A. and Stone, C.J.},

journal={CRC Press},

year={1984}

}

***Articles***

Zandt, F., & Richter, F. (2022, January 7). *Infographic: The shorter the song, the sweeter the stream?*. Statista Daily Data. https://www.statista.com/chart/26546/mean-song-duration-of-currently-streamable-songs-by-year-of-release/#:~:text=Music%20through%20the%20decades&text=The%20average%20song%20released%20on,song%20lengths%20starting%20in%201990.

Kenny, M. (n.d.). *Top 20 greatest live musical performances ever: Articles on watchmojo.com*. WatchMojo. https://www.watchmojo.com/articles/top-20-greatest-live-musical-performances-ever

[Author removed at request of original publisher]. (2016, March 22). *6.2 the evolution of popular music*. Understanding Media and Culture. https://open.lib.umn.edu/mediaandculture/chapter/6-2-the-evolution-of-popular-music/

By:          Yellowbrick. (2023, August 24). *Top 2010s music trends: Exploring the evolution and influences*. Yellowbrick. https://www.yellowbrick.co/blog/music/top-2010s-music-trends-exploring-the-evolution-and-influences#:~:text=Pop%20music%20continued%20to%20dominate,EDM%2C%20and%20hip%2Dhop.