Classification of Song Genre Based on Spotify Track Attributes

Mia Price, pricem4@rpi.edu

[1]Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States

Abstract

How we define a genre is always changing. Genre is constructed by patterns that we hear in music, but these patterns can change with time, with a movement, and sometimes instantly. The way that we can classify a song by genre is not always based on audio attributes.

The aim of this project is to investigate a potential model for determining the genre of a song track based on numerous audio attribute traits such as song duration, key, mode, time signature, acousticness, liveness, danceability, energy, instrumentalness, loudness, speechiness, valence, and tempo. The models that were constructed include classification trees and support vector machines.

This project was motivated by a simple interest in trying to understand and build connections genre with the song attributes provided by Spotify.

Workflow

- **Collect Song Data**
  - Songs that were released in 2018 and 2019 were collected by scraping SongList.com
- **Use Spotify API**
  - Spotify API was used to collect song attributes (danceability, accousticness, energy, etc.), popularity, and genre
- **Exploratory Data Analysis**
  - EDA was conducted in order to determine the relationships and patterns that existed in the data.
  - Here, data cleaning occurred and any incomplete entries were removed.
- **Analytics**
  - Determination of Genre using:
    - Classification Tree
    - Support Vector Machine
- **Validation and Conclusions**
  - The accuracy of the models were evaluated
  - Conclusions of about the validity of the models were made.

The Data

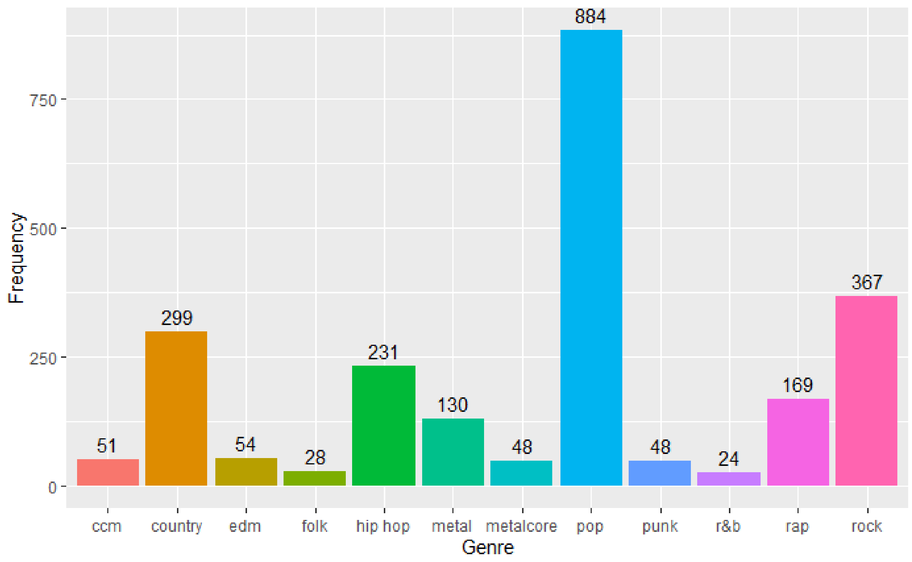

Figure 1: Frequency of Genre in Dataset

The data was obtained by scraping songlist.com for songs released in 2018 and 2019. The Spotify API was used to collect genre, popularity of the song, and song attributes. Song attributes include song duration,key, mode, time signature, acousticness, liveness, danceability, energy, instrumentalness, loudness, speechiness, valence, and tempo.

After collection, the data was cleaned and all incomplete entries were removed. This provided 3333 data points for analysis. The frequency of genre can be seen in Figure 1.

Note that the frequency of pop songs is much greater than the frequency of any other genre.

Classification Tree

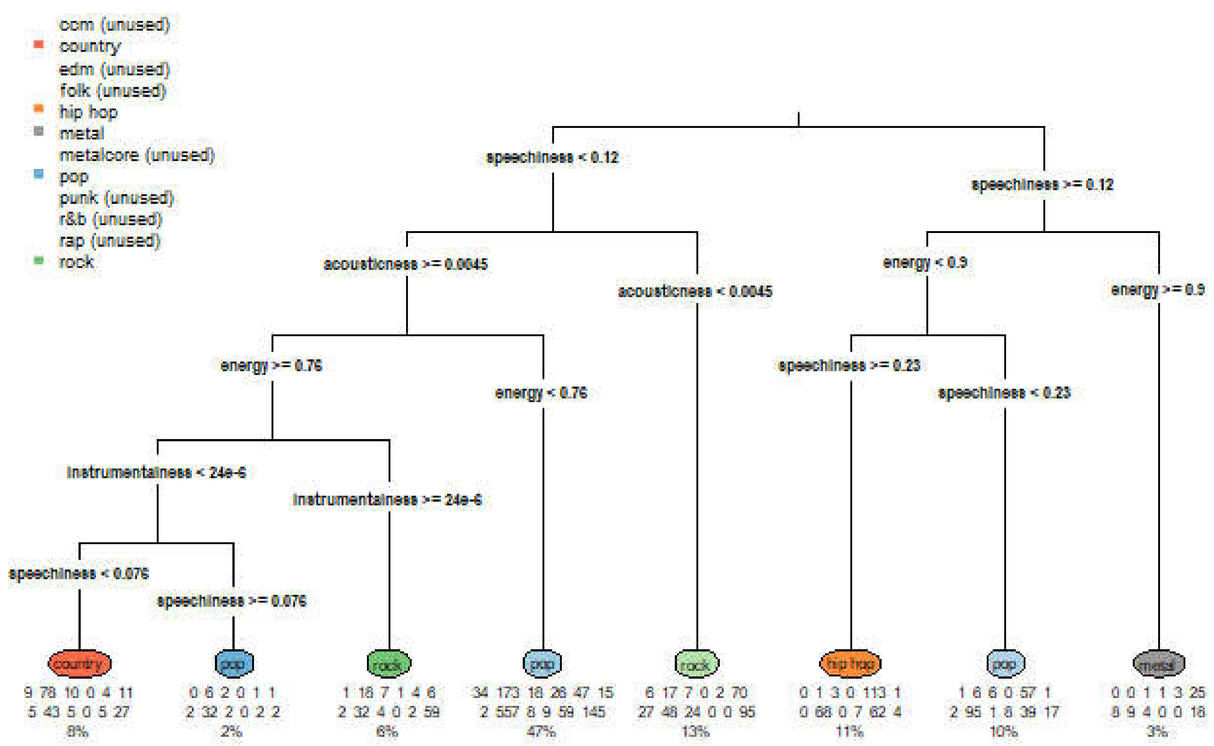

Figure 2: Classification Tree for the determination of Genre

Genre was determined by key, mode, time signature, acousticness, danceability, energy, instrumentalness, loudness, speechiness, valence, and tempo. The tree has a misclassification rate of about 54%.

Support Vector Machine

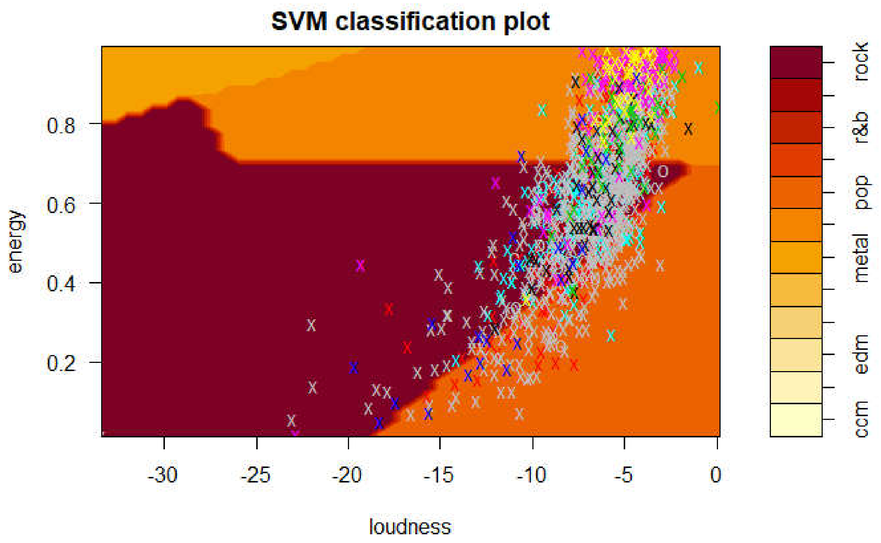

Figure 3: Linear SVM for Classifying Genre

A plot of the linear support vector machine showing energy versus loudness. This model has an accuracy of 43%.

Predicted \ Actual	ccm	country	edm	folk	hip hop	metal	metalcore	pop	punk	r&b	rap	rock
ccm	0	0	0	0	0	0	0	0	0	0	0	0
country	0	10	1	0	0	1	0	5	2	0	0	4
edm	0	0	0	0	0	0	0	0	0	0	0	0
folk	0	0	0	0	0	0	0	0	0	0	0	0
hip hop	0	0	0	0	24	0	0	16	0	1	21	3
metal	0	0	1	0	0	23	3	2	2	0	1	9
metalcore	1	0	0	0	0	0	0	0	0	0	0	0
pop	4	60	12	3	28	5	3	159	3	4	28	58
punk	0	0	0	0	0	0	0	0	0	0	0	0
r&b	0	0	0	0	0	0	0	0	0	0	0	0
rap	0	0	0	0	0	0	0	0	0	0	0	0
rock	3	9	1	2	1	9	3	17	8	0	0	33

Table 1: Confusion Matrix for SVM Model

Predicted versus actual for the test dataset

Conclusion

Overall, the support vector machine and classification tree are not perfect and do not produce the best results. In fact, both trees produce an accuracy less than 60% (SVM: 43%, Classification Tree: 46%). In addition, both of the proposed models for determining genre have trouble identifying certain genres (ccm, edm, folk, etc.) and the classification tree, leaves these genres out all together.

Future Direction

In the future, I would want to obtain a more representative dataset of the music that is released.
In addition, I would like to refine the models that I have created for this project and look into other models that would produce better results.

Resources:
Songlist.com – provided a list of songs released in 2018 and 2019
Spotify API – used to obtain song attributes

Glossary:
API – Application programming interface
SVM – Support Vector Machine