Music Genre Classification System in Python

Music enthusiasts can generally determine the genre of a song naturally. This analysis will attempt to train a computer to predict the genre of a song using audio feature characterizations from Spotify, a large music streaming platform, using various multi-classification models. A successful model could feed music recommendation systems or other similar uses.

1. Introduction

Music recommendation systems play are a big part of how music is consumed in the market place today. Recommender systems use two basic principles – collaborative based or content based. Collaborative based systems rely on massive user data and preferences to provide recommendations based upon the individual user and other users within the system. Collaborative systems struggle providing recommendations for new users or new content as the models rely on historical usage data which is not available. Content based models, alternatively, use inherent features of the content, songs in this case, to help derive similarities and create recommended predictions. If successful, content-based systems could bridge the gap for new content and provide better recommendations for new users with little historical data.

This analysis will attempt to create a content-based model to predict music genre, which may only be one aspect of music recommender system, using audio feature data available through the Spotify API. Additionally, an unsupervised clustering model will be used to group songs together based purely audio features. The clustered model will be compared to the defined genre model.

2. Gathering Data/Feature Selection

Dataset

The dataset is derived from genre data created and maintained by Every Noise at Once (Everynoise.com), which maintains lists of genres and tracks from users on Spotify. The dataset includes popular genres shown below:

- Alternative Rock
- Country
- Dance Pop
- Hip Hop
- Pop
- R&B
- Rock

The data set includes 5141 songs across the seven genres. Everynoise has created playlists on Spotify for each genre and each playlist track information was captured through the Spotify API.

Preprocessing & Data Cleaning:

The first step in capturing the data was to search through the Spotify playlists to identify the appropriate playlist ID. Spotify provides playlist owner information, so playlist id information was captured by searching for genre and owner. The information returned contained more playlists than specified. For example, the search returned any genre created which contain the key search terms 'pop', 'rock', 'country' etc. The playlists were manually filtered to include only the genres listed above.

In order to capture the required audio feature data, the key features of the prediction models, we need the Spotify ID of each song by genre. To capture the Spotify ID, the program captured song meta data from each playlist. The API uses a paginated json reply, so functions were created to handle multiple page returns as well as converting json formatted data to usable data. While the ID was the only data required, the functions captured other song data such as song name, artist and album data as well.

Once the dataset included ID info per genre, another function was created to iterate through each track to retrieve audio feature information. All the data was stored in panda data frame format. The meta data and audio feature data was joined into a single data frame and saved in csv format to provide clean data for the analysis.

The final song count by genre is shown below:

Row Labels	Songs
Alternative Rock	693
Country	569
Dance Pop	1145
Hip Hop	891
Pop	556
R&B	360
Rock	927
Grand Total	**5141**

Data Visualization & Investigation:

To obtain a sense of the data, plots were generated showing distribution of each genre over each feature. Upon review of the data, it became apparent that only 4 features provided enough potential variation to create useful prediction models. The charts are shown below.

These metrics are defined as

- danceability - suitability for dancing based upon tempo, rhythm, beat and other factors. Scale: 0.0 - 1.0
- energy - perceptual measure of intensity and activity. Scale: 0.0 - 1.0
- valence - describing the musical positiveness conveyed by a track. Higher values are 'happier'. Scale: 0.0-1.0
- tempo - overall estimated tempo of a track in beats per minute (BPM). This value will be converted to a scalar 0.0-1.0

Careful review of these charts pointed to potential problems when using the various models as the majority of the songs appear to have the same basic set of characteristics.

Feature Selection

The total feature set used to create the predictive models comes directly from Spotify audio features. All the features are described below.

- **danceability** - suitability for dancing based upon tempo, rhythm, beat and other factors. Scale: 0.0 - 1.0
- **energy** - perceptual measure of intensity and activity. Scale: 0.0 - 1.0
- **key** - estimated overall key of track. Scale: integer

- **loudness** - the overall loudness of a track in decibels (dB). This value will be converted to scalar 0.0-1.0
- **speechiness** - detects the presence of spoken words in a track. . Scale: 0.0-1.0
- **acousticness** - overall acousticness of track. Most tracks are have low acousticness. Scale: 0.0-1.0
- **instrumentalness** - predicts whether a track contains no vocals. Scale: 0.0-1.0
- **liveness** - detects the presence of an audience in the recording. Scale: 0.0-1.0
- **valence** - describing the musical positiveness conveyed by a track. Higher values are 'happier'. Scale: 0.0-1.0
- **tempo** - overall estimated tempo of a track in beats per minute (BPM). This value will be converted to a scalar 0.0-1.0

3. Methods

Genre Classification of Lyrics

We implemented three different classifiers in hopes of achieving high-fidelity genre classifications:

1. K Nearest Neighbor (KNN)
2. Support Vector Machine (SVM)
3. Logistic Regression
4. Decision Tree

All models were implemented in python. Please follow the link to the github site with all the coding (https://github.com/hackrdog/IBM-Data-Science-Capston).

During model creation and tuning, the accuracy of the predictive models appeared to change each time the model was executed. The data set was not big enough to segment into individual groups to lead to an aggregated accuracy metric, so I ran each model 100 times with newly selected random train/test data sets each time. Since the KNN model requires a variable be determined for each model run, this model was dropped from the final analysis.

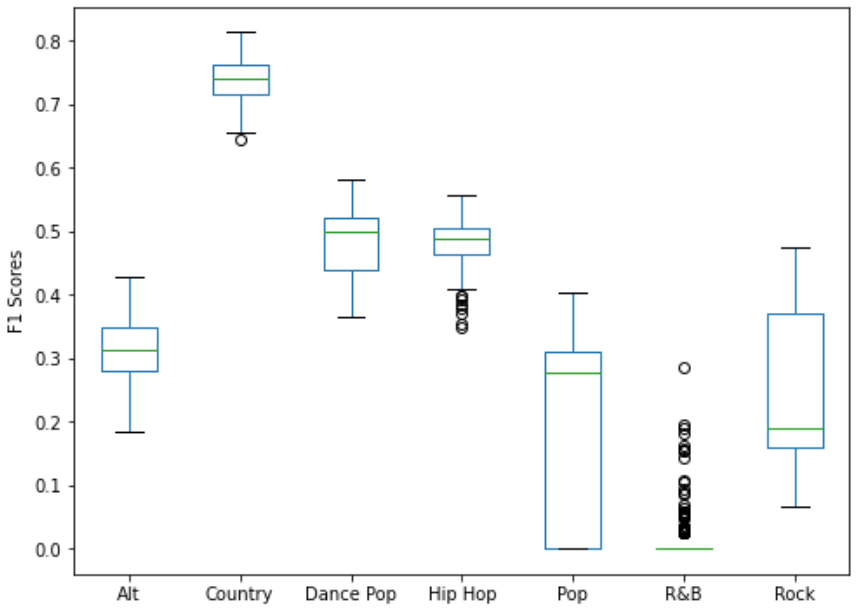
I also used a unsupervised multi-class model, K Means Cluster (KMC), against the data set to see how what type of overlap might occur from an unsupervised segmentation to a defined segmentation.

The F1 score and confusion matrixes were stored for each test and used exclusively to evaluate the success of each model. Box diagrams were generated for each module over the 100 executed samples. Aggregated confusion matrixes were also created for each model which used the mean value for each position in the array across all 100 samples.

4. Results and Discussion

Genre Classification – F1 Scores

Below are our F1 accuracy results by classifier:



The combined models do introduce some variability, but the trends are reasonably consistent across all models. Clearly, this data shows all three models have significant issues predicting genre based upon just the audio features provided.

Genre review:

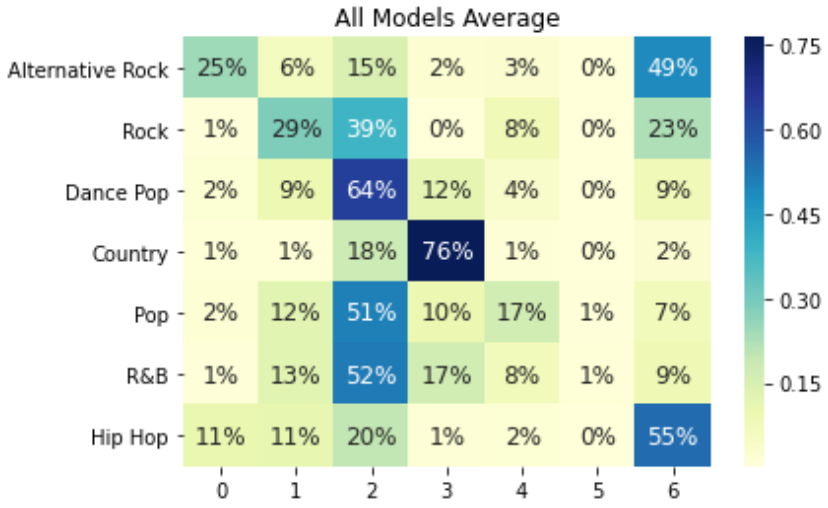
- Country genre shows the highest predictability across all the models with an mean F1 score = 0.73 across the models. Reasonably accurate results with few outliers.
- Dance and Hip-Hop score the next highest and provide relatively consistent results across all three models with Dance mean = 0.47 and Hip-Hop mean = 0.49. Hip-Hop does have a fair number of outliers below the line which requires further analysis.
- Alternative rock: The F1 mean = 0.33, which is pretty poor although all three models are fairly consistent.
- Rock: Results of each model vary significantly in this genre, although all provided poor predictive results. Decision tree mean = 0.40 with a span of 0.24 - 0.47. Both the other models resulted in mean scores of 0.17 with similar spans of 0.07 – 0.27. Highly unreliable.
- Pop: All models performed poorly with some substantial differences between the models. Essentially, the decision tree almost never predicted any song to belong to the Pop genre whereas the SVM and Log models obtained F1 scores of 0.30 with a span of 0.07 – 0.40.
- R&B: All models performed quite poorly and almost never predicted a song to belong to this genre with mean F1 ~= 0.00.

Genre Classification – Confusion Matrix

Looking at the aggregated confusion matrix provides some insight into the data and where most mis-classification errors occurred.

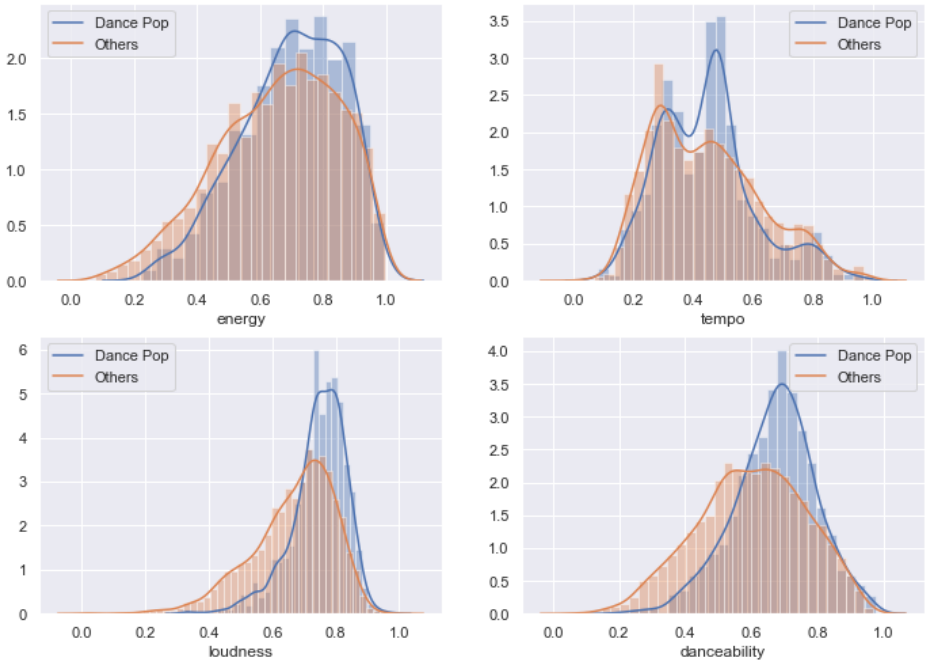
Observations

Looking at the *All Models Average* confusion matrix above, several observations about the models can be made.



- The F1 scores and confusion matrix for all three models are consistent with each other with some minor differences across the genres.
- All models tended to mis-classify Alternative music as Hip-Hop, with an average of 49% of all Alternative songs being classified as Hip-Hop
- Country music has the most definitive feature variation with an overall F1 mean average of 76% across all three models. The models were very accurate with low false positive rates in all other genres except Dance Pop (18%).
- All models generated significant False Positive predictions in the Dance Pop category with high False Positive rates for Rock (39%), Pop (51%) and R&B (52%). False Positives were significant in all remaining genres as well Alternative (15%), Country (18%) and Hip-Hop (20%).
- All models under predicted Pop genre with 17% Accuracy and low F1 False Positive scores in all other genres.
- All models rarely predicted R&B with R&B songs falsely attributed to Dance Pop (52%), Country (18%), Rock (12%) and the remaining 18% distributed across remaining genres.

The Dance Pop genre seems to overwhelm the models and review of the plots of this genre over the four key features provide some insight.



Tracks in the Dance genre almost completely encapsulate the data from all the other genres as shown here. Only a few songs with low energy, loudness or danceability can differentiate themselves.

The table shows the min-max-mean for each model/genre pair.

	Decision Tree				Logistic				Support Machine Vector			
Genre	Min	Max	Mean	Span	Min	Max	Mean	Span	Min	Max	Mean	Span
Alt	0.20	0.46	0.39	0.26	0.26	0.43	0.34	0.17	0.18	0.37	0.29	0.19
Rock	0.24	0.47	0.40	0.24	0.07	0.28	0.17	0.21	0.09	0.27	0.17	0.18
Dance Pop	0.36	0.48	0.42	0.12	0.47	0.57	0.51	0.10	0.48	0.58	0.51	0.10
Country	0.64	0.79	0.71	0.14	0.67	0.81	0.75	0.14	0.70	0.81	0.75	0.12
Pop	0.00	0.15	0.01	0.15	0.21	0.38	0.30	0.17	0.18	0.40	0.30	0.22
R&B	0.00	0.29	0.02	0.29	0.00	0.10	0.02	0.10	0.00	0.00	0.00	0.00
Hip Hop	0.35	0.53	0.45	0.18	0.43	0.54	0.49	0.12	0.42	0.56	0.50	0.13
Accuracy	0.41	0.48	0.44	0.07	0.44	0.52	0.47	0.07	0.44	0.51	0.47	0.07
Macro Avg	0.31	0.39	0.34	0.08	0.34	0.40	0.37	0.06	0.33	0.40	0.36	0.07
Weighted Avg	0.36	0.44	0.40	0.08	0.40	0.49	0.43	0.09	0.39	0.48	0.43	0.09

K Means Cluster Analysis

The K Means Cluster algorithm was applied to the same data in an unsupervised fashion with the following results as mapped against the original genre tag for each song.

	Clusters						
genre	0	1	2	3	4	5	6
Alternative Rock	6%	6%	26%	19%	16%	18%	8%
Rock	10%	13%	27%	18%	18%	6%	8%
Dance Pop	10%	10%	12%	26%	36%	2%	5%
Country	6%	20%	16%	18%	21%	0%	19%
Pop	5%	13%	11%	27%	16%	1%	28%
R&B	6%	18%	4%	30%	29%	0%	14%
Hip Hop	16%	13%	8%	20%	41%	1%	2%

Total	9%	12%	16%	22%	27%	4%	10%

5. Conclusion and Future Work

The results of the model conclusively show that using audio features available through the Spotify API are insufficient to building a successful music genre classification system. Perusing the topic on several data science sites, reveals that developing successful music classification algorithms is very difficult. The problem lies potentially in a lack of clearly defined limits for each genre, popular tracks are increasingly blending genres and the similarity of topics of song lyrics across all genres.

No clear standards exist for classifying genres. This study used a few genres selected from Everynoise.com, but this web site tracks over 4500 individual genres, with many songs mapping to multiple genres. Additionally, songs are increasing crossing over traditional genre classifications. The #1 Hot Billboard Song for 2019 was Old Town Road by Lil Nas X which could easily fit into either the Hip-Hop or Country genres. Songs from all genres cover topics about love, relationships, trials and tribulations rendering subtle, hard to discriminate differences in lyric features.

In my research, I have not come across models which combine lyric, Spotify audio features and spectrum analysis into a single model to see if a combined data set would provide higher accuracy in predictions.