Summary

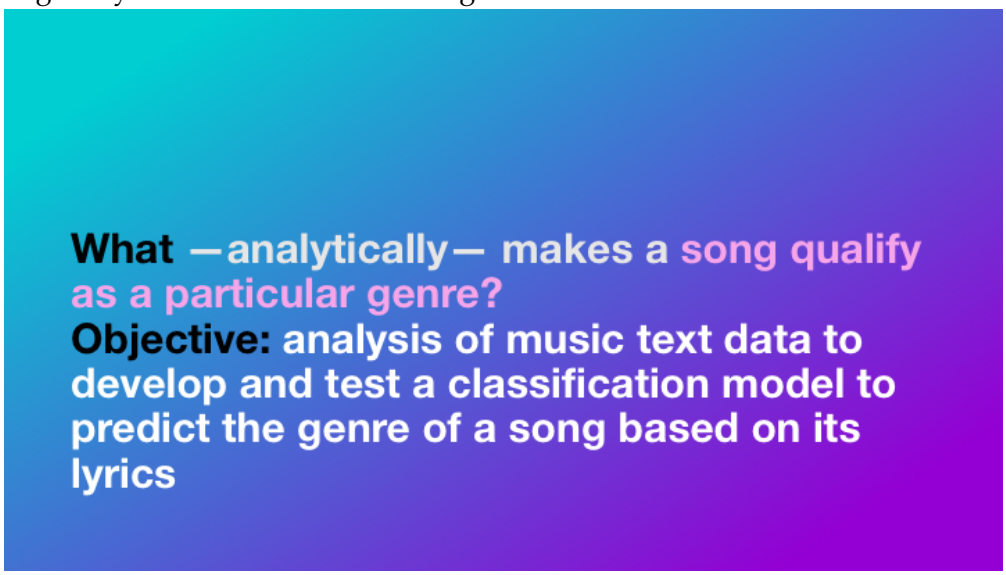
October 13, 2019

0.1 Classification of Music Lyrics for Artists and Songs (I've) played since 2016

0.2 Introduction

Music streaming has uppended the music industry in irreversable ways. Gone are the days of CD distribution and mp3 players with capacity constraints. Today, consumers can access nearly any song ever create - from top artists to recently-minted SoundCloud singer.

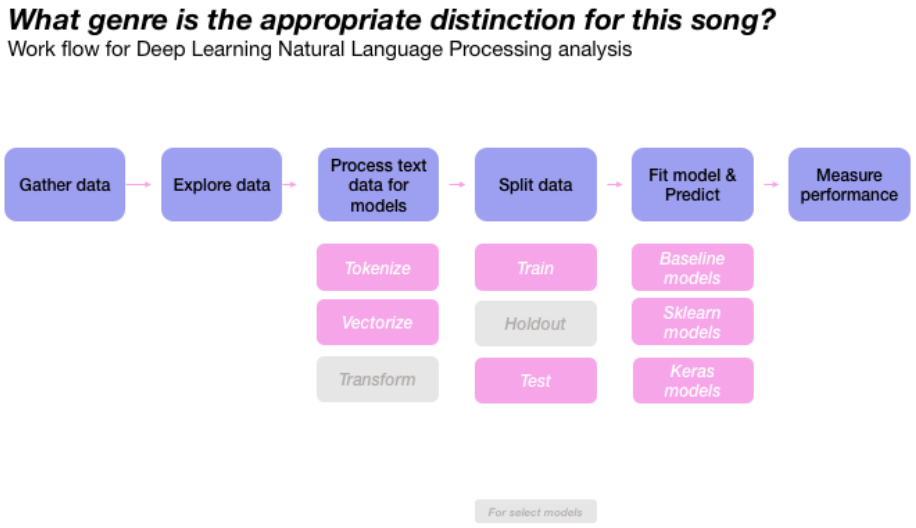
Instant access to music with significantly less financial barriers supports the experimentation and exploration of various music forms that conusmers would have passed on in previous phases. Given this, it is helpful to understand how listeners consume music: which genres, artists, and - because there is less emphasis on records - which songs and content. On content, the messages that are or aren't conveyed in music can speak to an artists strategic and creative approach. The following will explore the content in today's music through a process of Natural Language Processing analysis to answer the following:



0.3 Project motivation

For this project, I'm looking to build on the previous NLP and Machine Learning skills that I built on my Module 3 work while incorporating Deep Learning approaches and packages. I'd also like to focus on acquiring data through APIs and webscrapping. Most importantly is my interest in musicians and the art they produce, and possibly gaining better understanding in how they influence, with their words, me at the least.

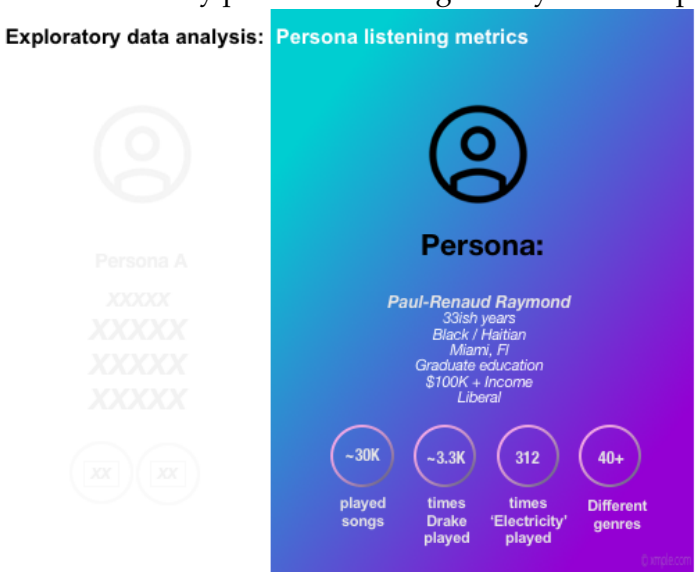
0.4 Project outline



1. Obtain my music listening history data from Apple Music through a database request. More seamless process than using the API.
2. Obtain music lyrics for each song and artist pair in my listening history data set by using the Genius API.
3. Tokenize and vectorize my text data for initial insights on most frequent ngram phrases. Additionally, I'd like to analyze trends in my listening activity over time.
4. Build and train a baseline, NB, and neural network models to classify Genres and Artist.
5. Compare the performance of all approaches.

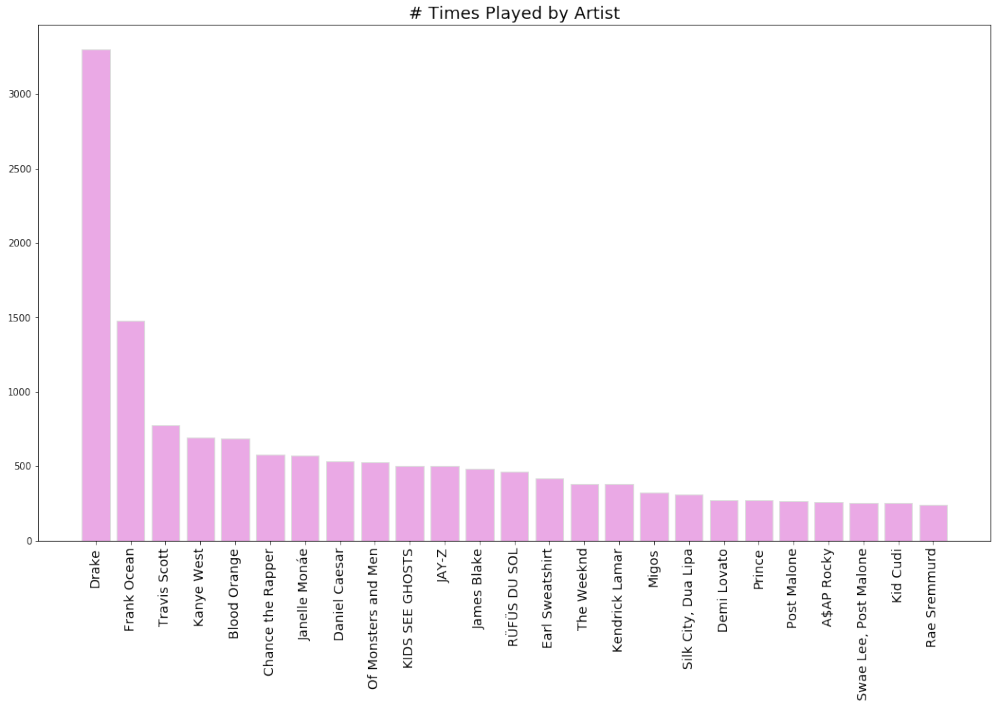
0.4.1 Exploratory data analysis

I reviewed data on my perosnal listening history to develop an initial persona.

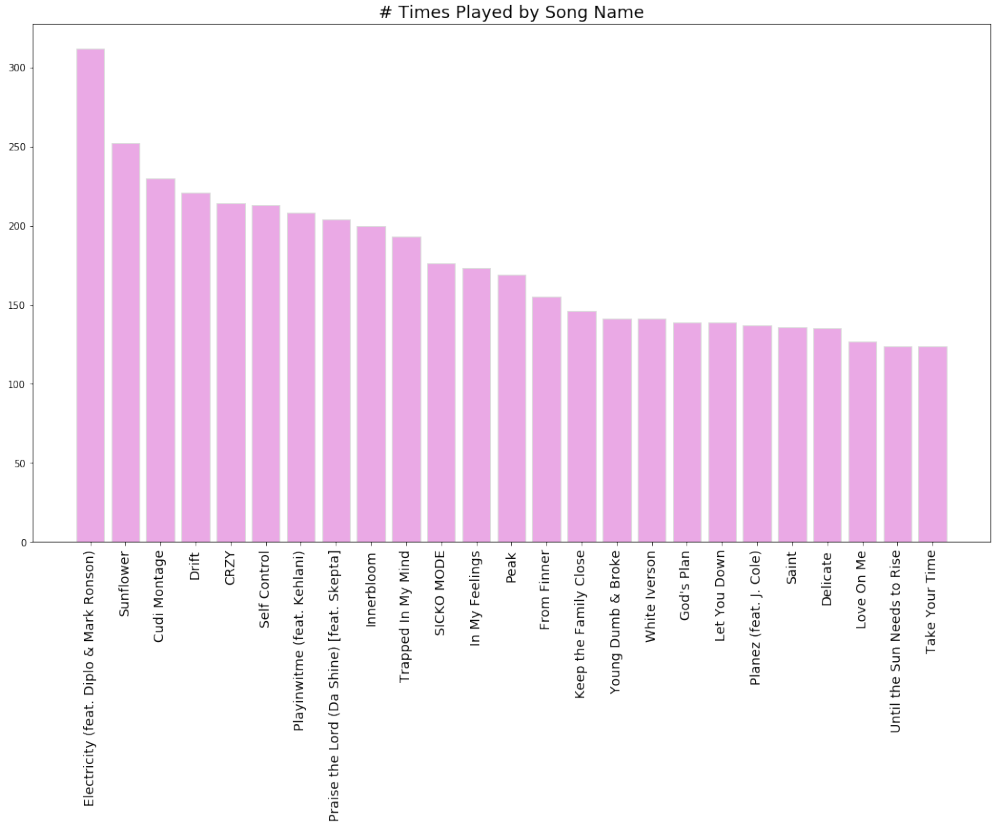


Below are additional snapshots of my listening activity during the ~3 year period.

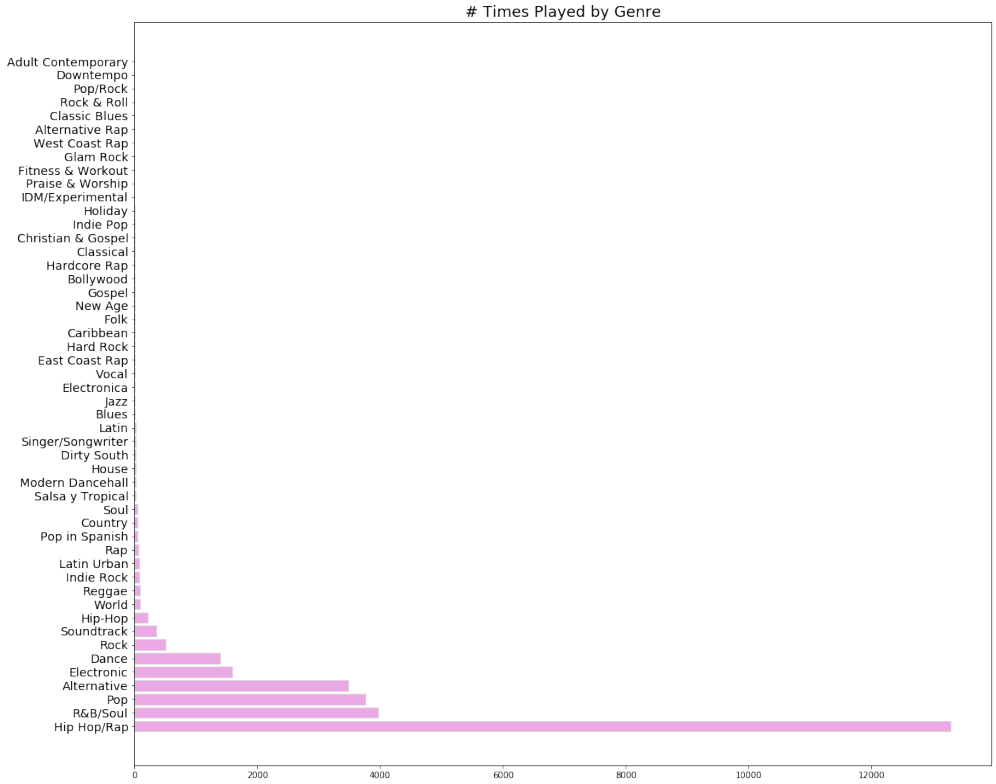
Artist plays



Song plays



Genre plays

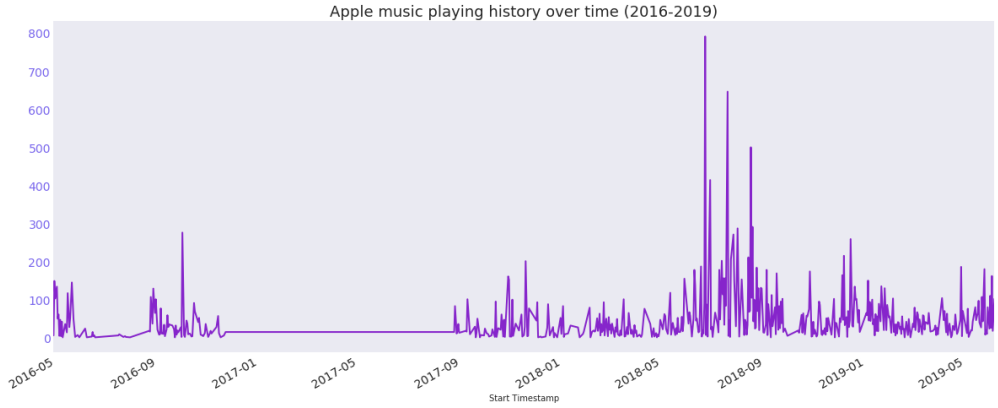


Nothing surprising and shameful here.

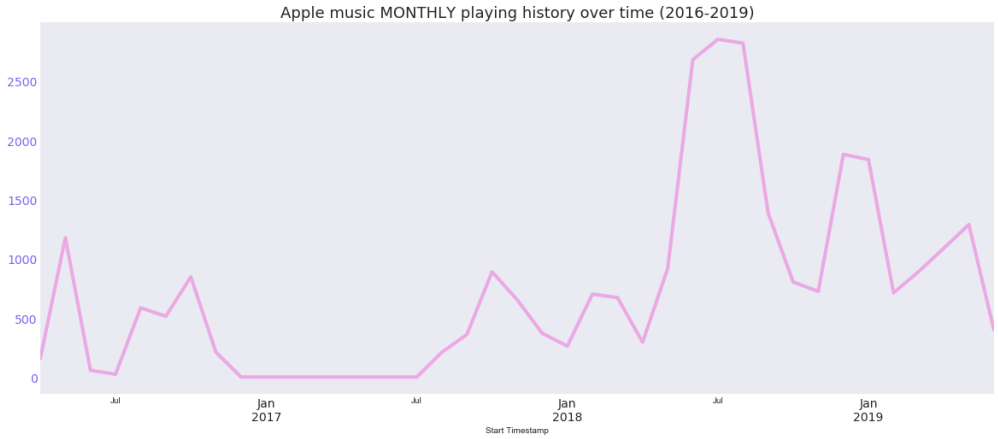
0.4.2 Time-series analysis

Additionally, to build on the initial exploratory analysis, I'll next dive into trends in listening habits over time.

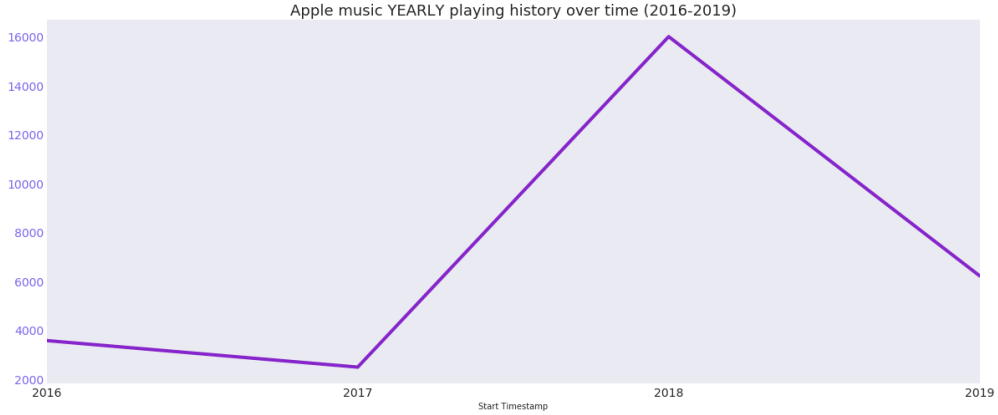
Overall listening activity



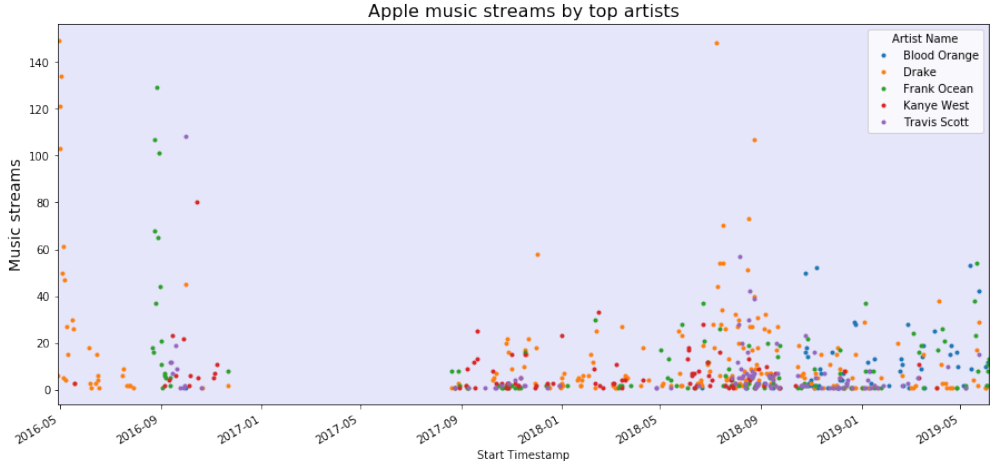
Monthly listening activity



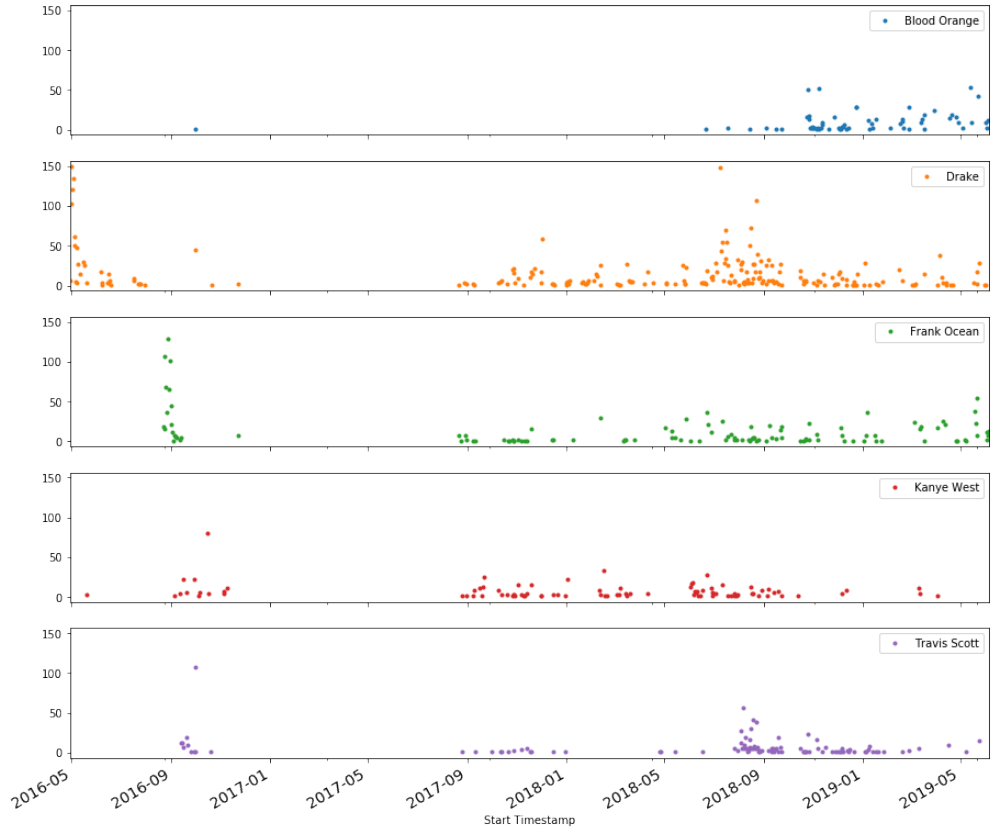
Annual listening activity



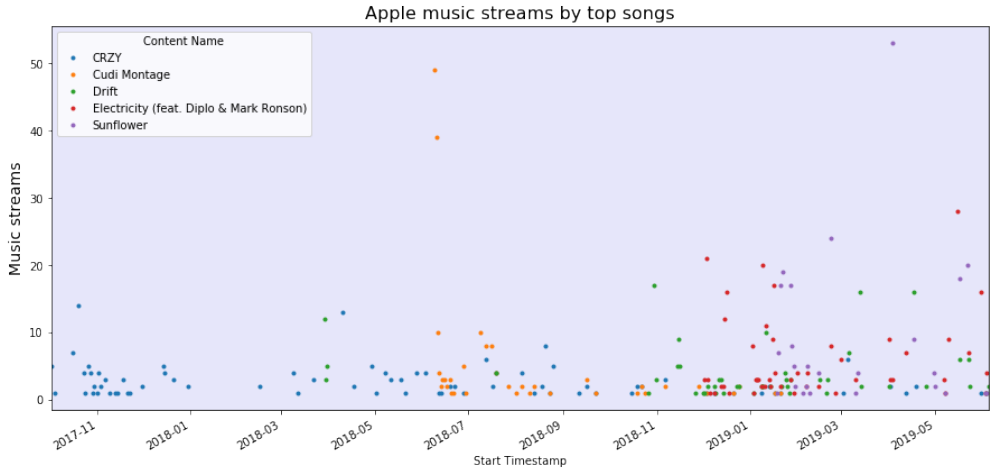
Trends by top artists



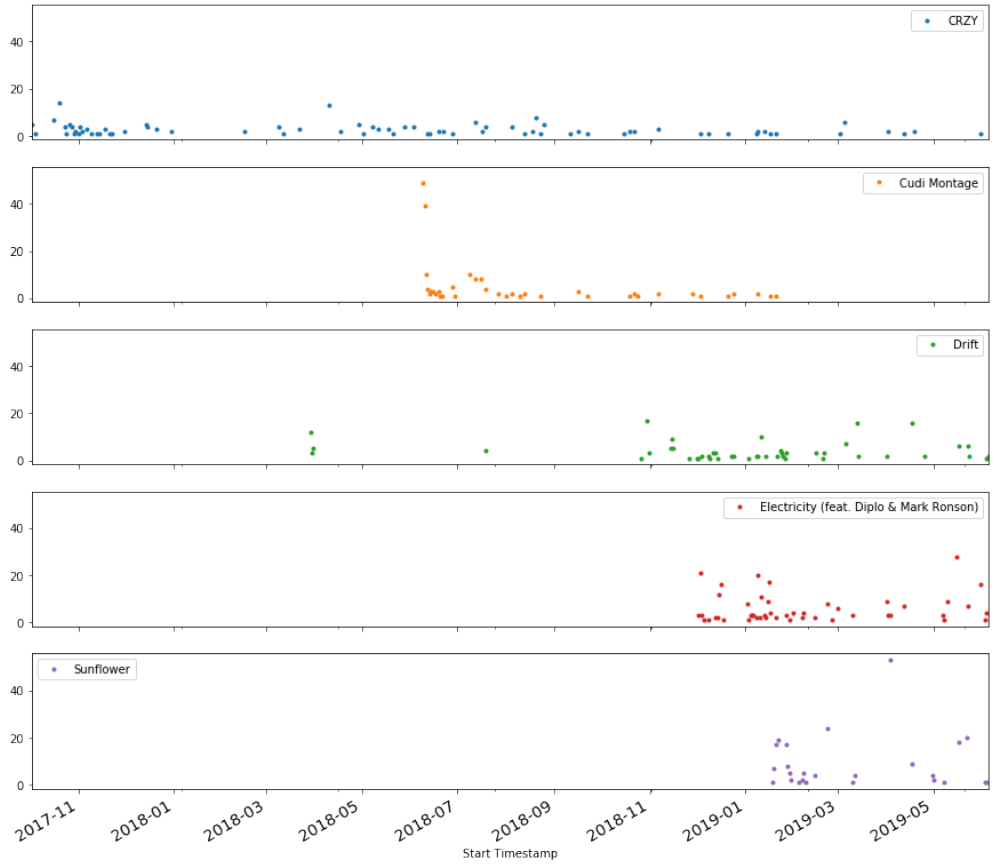
Subplots



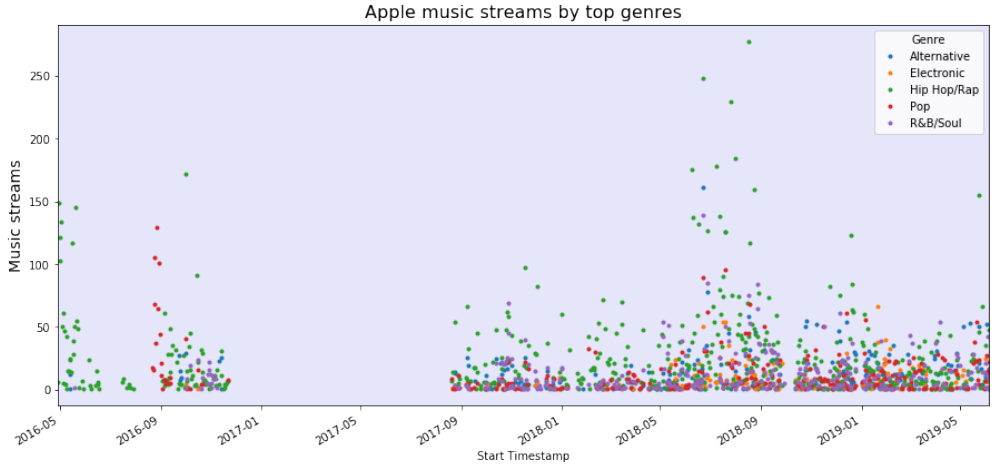
Trends by top songs



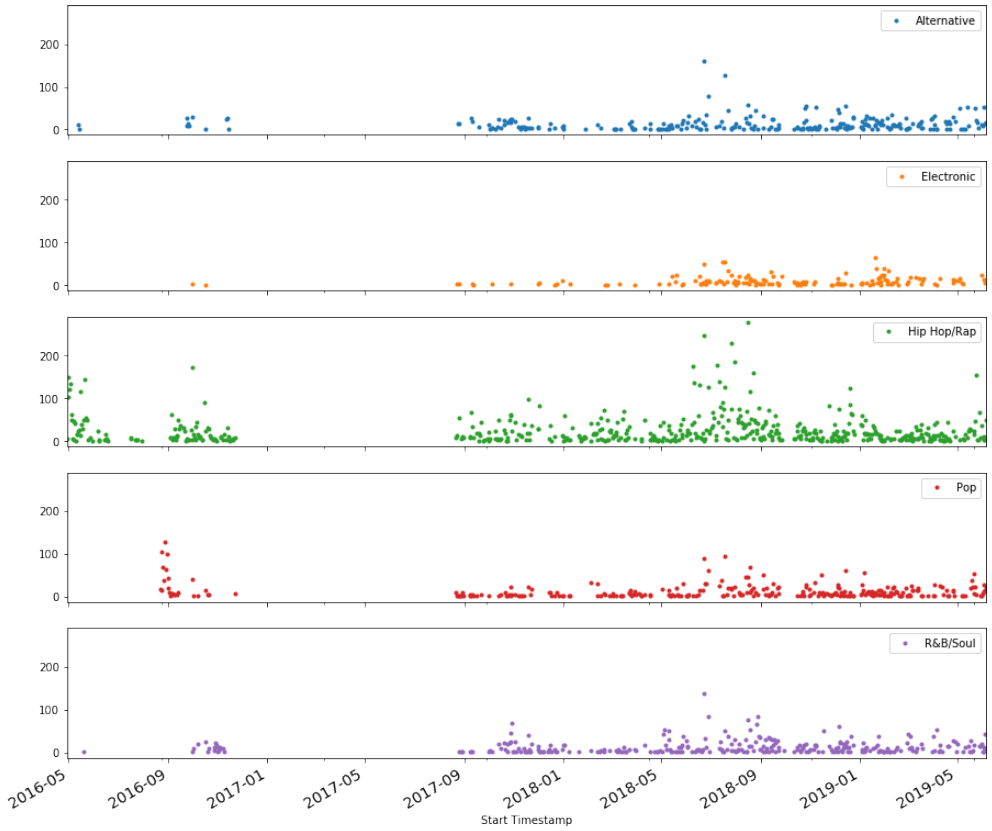
Subplots



Trends by top genres



Subplots



From these visualizations, I notice that I've played Hip Hop most consistently through the time period, with a considerable uptick in Summer 2018. I have also played more electronic music since Spring 2018.

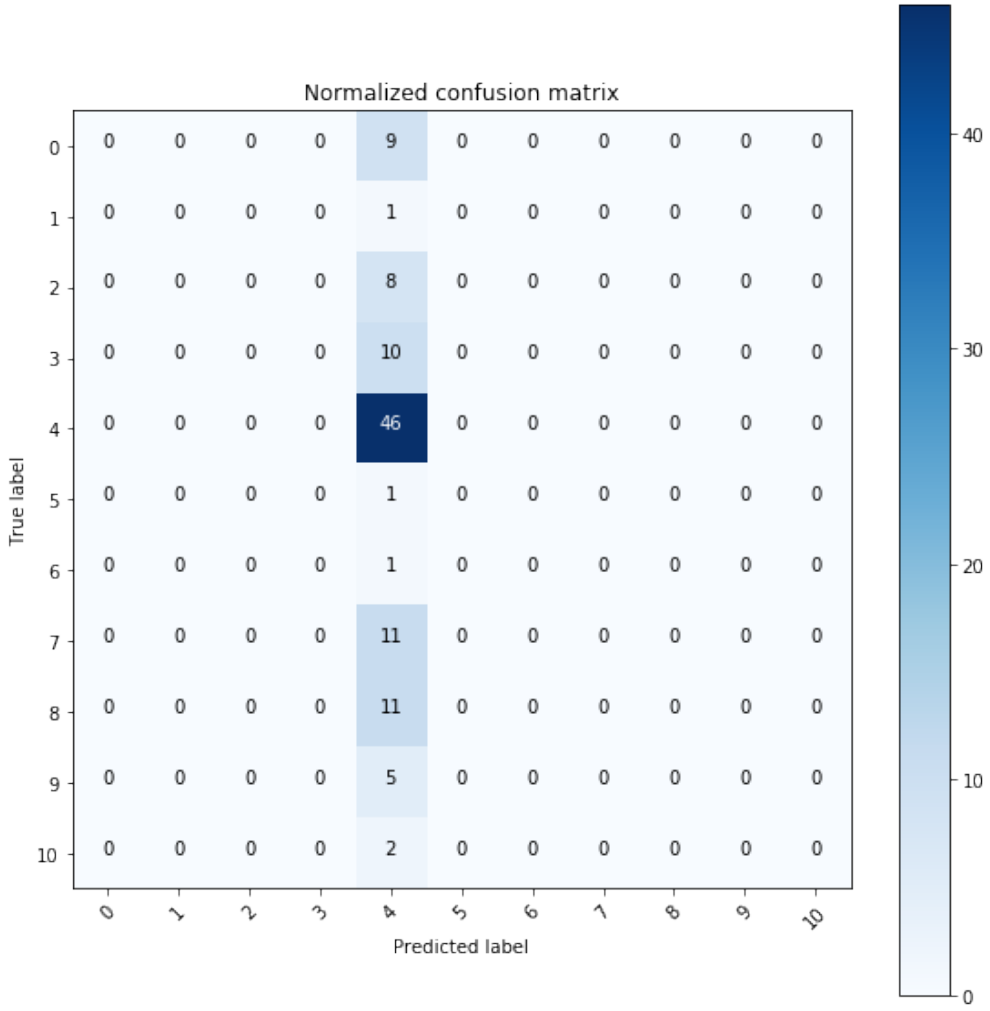
0.4.3 NLP classification

Now that the necessary data is sourced and organized, the next phase will focus on preprocessing the text data with the use of Natural Language Processing tools. For a series of analysis, the text will be tokenize, vectorized, and transformed in additional manners to create embeddings and to classify target features in the model.

I'll employ various approaches to build classification models for Genre feature of the data. I'll start with a Log Regression as baseline and after compare performance with Naive Bayes, Random Forest, GRU, LSTM, and lastly Bidirectional LSTM models. This classification task will entail double digit classes and likely some class imbalance challenges to work through.

Initial baseline

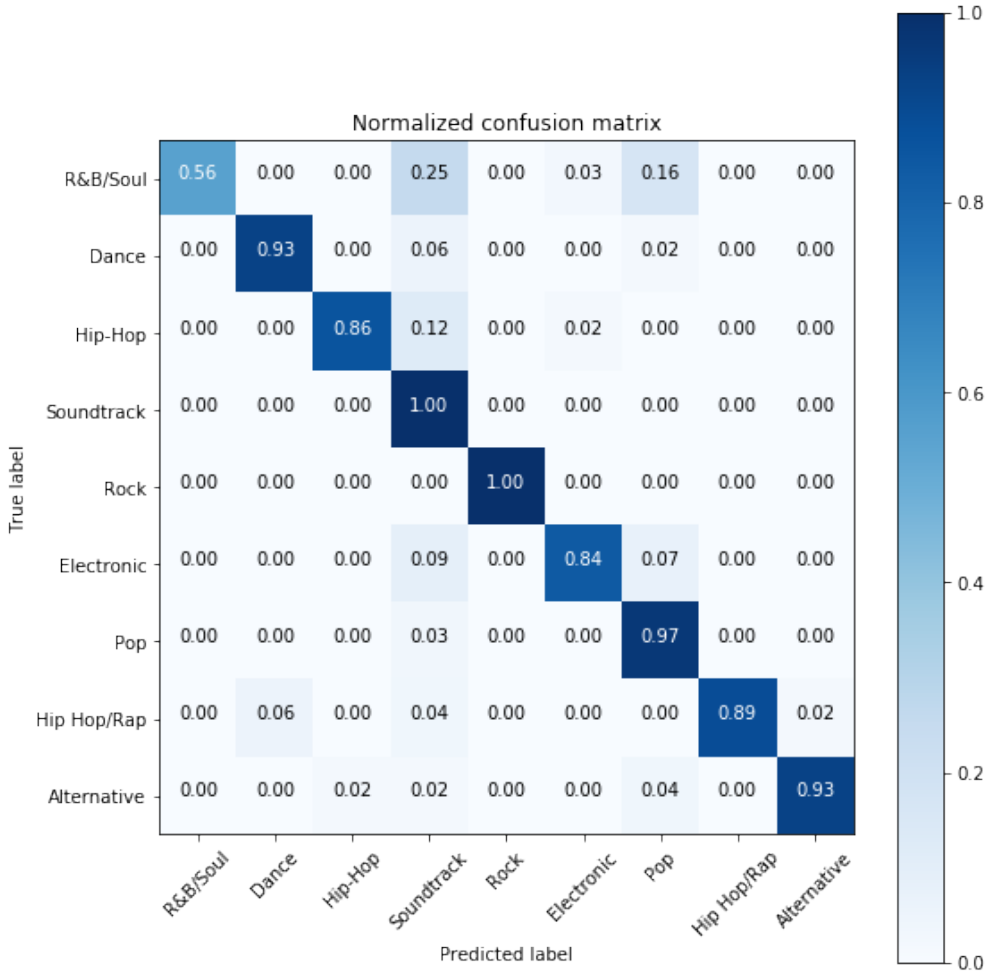
The baseline— Multinomial Naive Bayes— model has a 44% accuracy. It has particularly poor performance with the minority classes due to high class imbalance with Hip Hop/Rap genre.



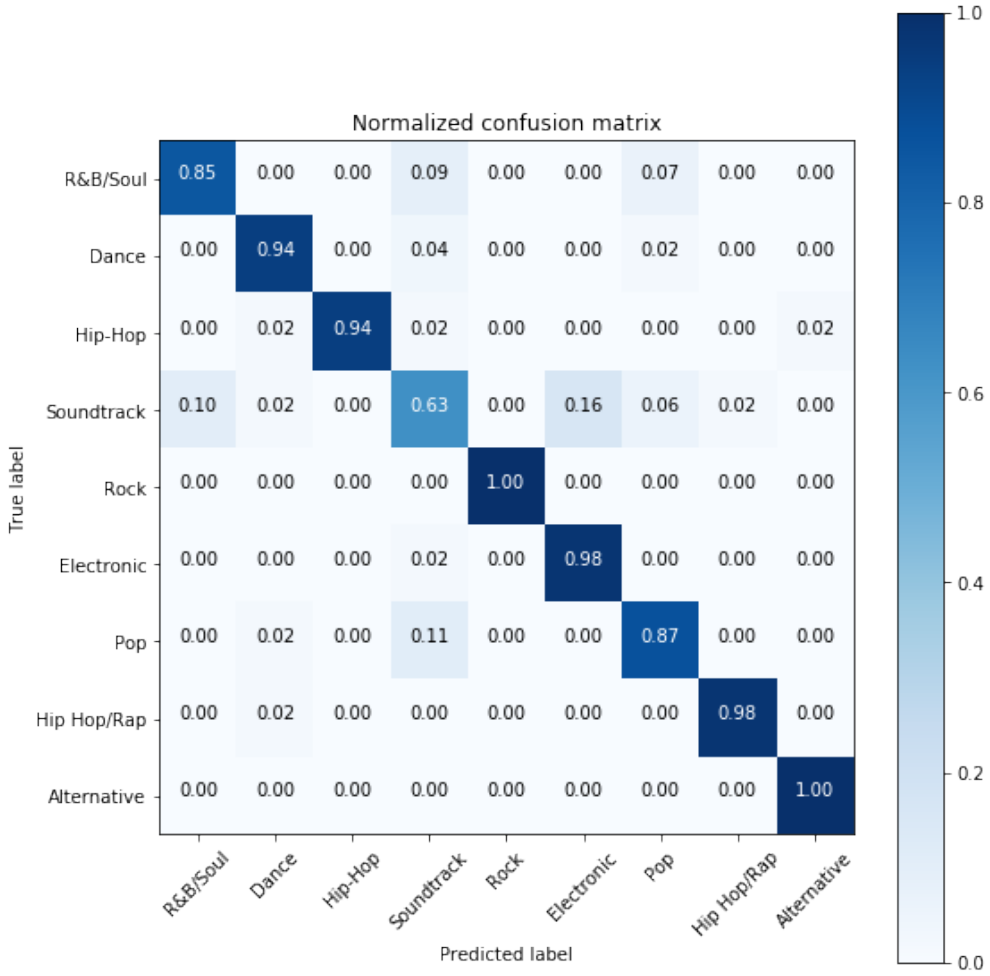
A seperate baseline Random Forest model performs slightly better yet is also affected by the class imbalance.

Completing these models with SMOTE improves performance significantly.

SMOTE Naive Bayes

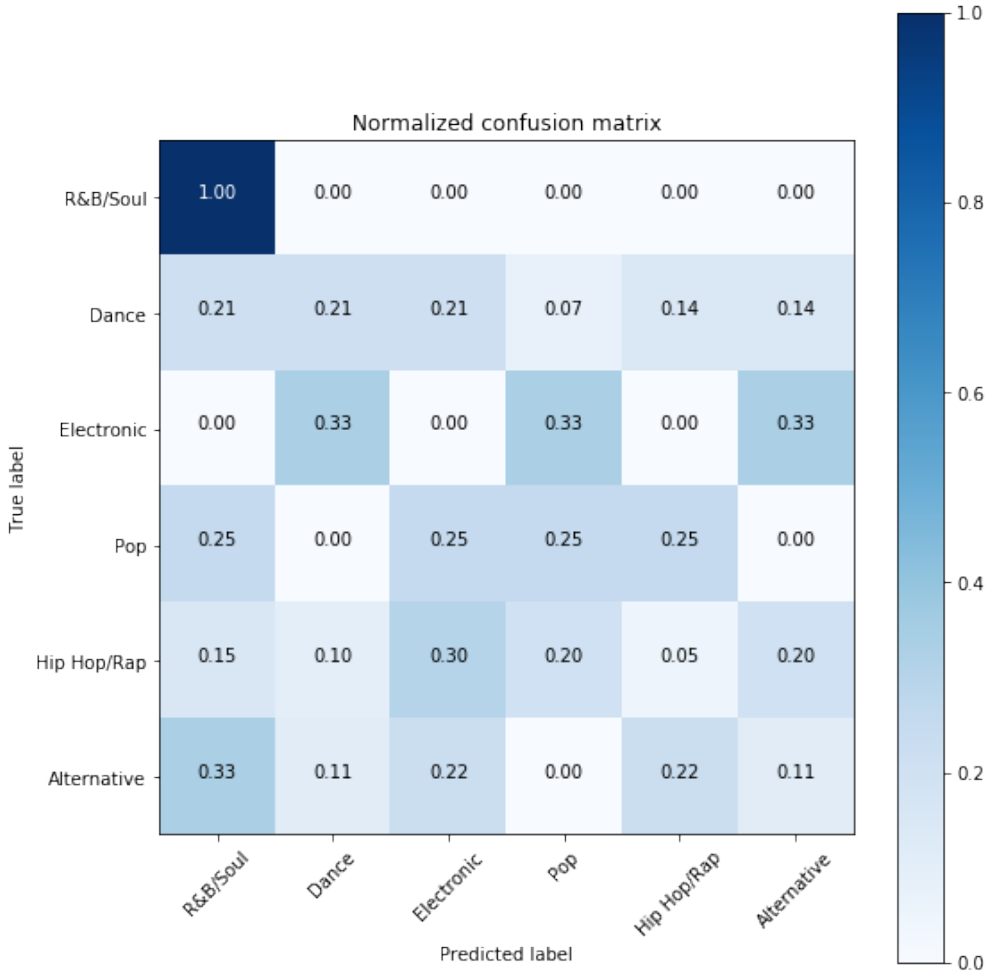


SMOTE Random Forest

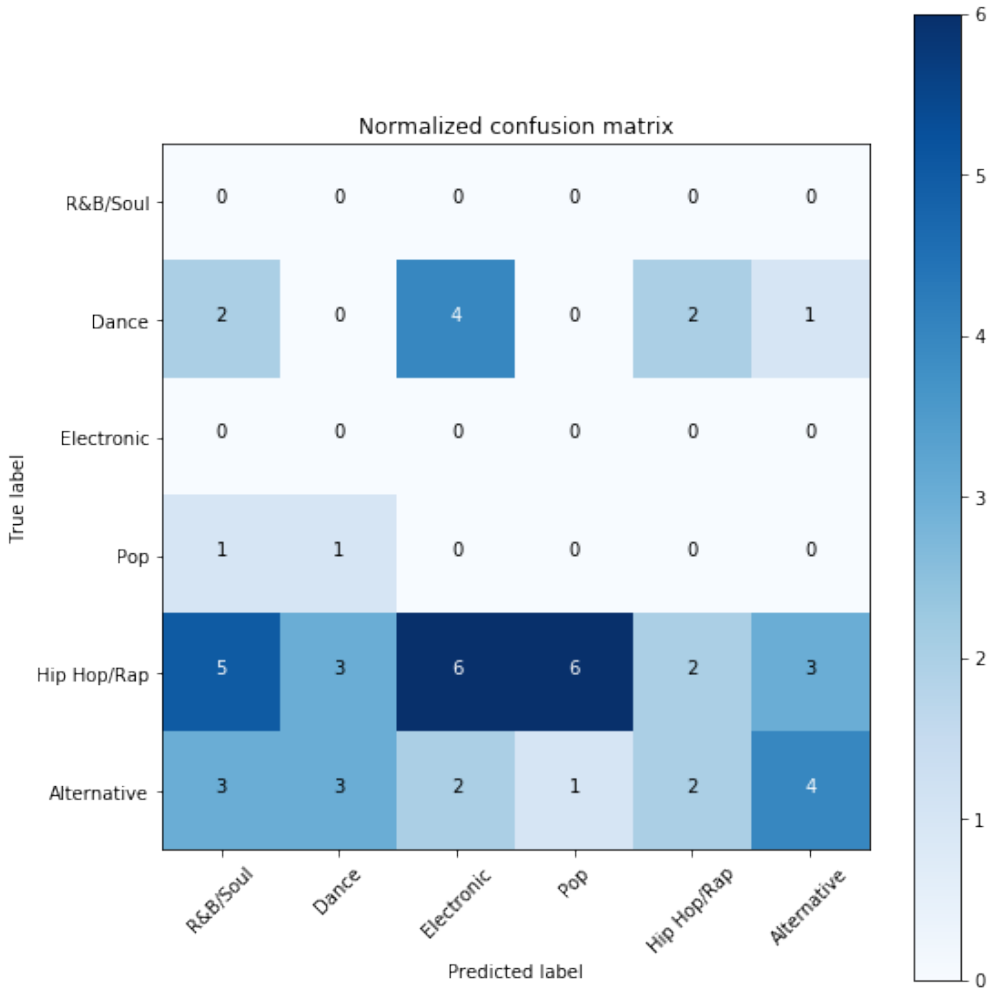


It is possible that the synthethic samples may overfit the model. I'd like to also complete these models with undersampled sets, for comparison.

Undersampled Naive Bayes with GridSearch



Undersampled Random Forest with GridSearch



Even with Gridsearch, the undersampled models perform poorly. In this case, the sample sizes are likely too small to provide both sufficient data to train and test on.

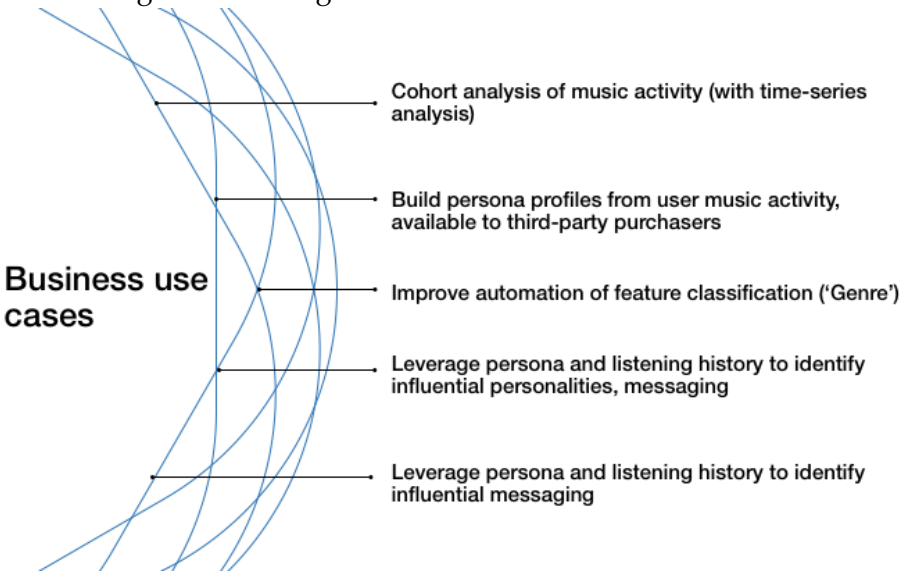
Given these limitation of combination of class imbalance and relatively small dataset, I'll explore how deep learning approaches may improve classification performance.

The deep learning models I used are: LSTM, GRE, and Bidirectional LSTM. Each had greater accuracy than the baseline models, with Bidirectional LSTM performing the best at 89% accuracy!

NLP Classification: Performance across models

	Accuracy	Precision	Recall	F-1	N-size
Baseline NB	44%	19%	44%	27%	105
Baseline RF	51%	35%	51%	40%	102
SMOTE NB	85%	89%	85%	82%	439
SMOTE RF	91%	91%	91%	91%	439
Undersampled NB GridSearch	14%	30%	14%	13%	51
Undersampled RF GridSearch	12%	5%	12%	7%	51
LSTM	49%				41
GRE	49%				41
Bidirectional LSTM	89%				41

This data and analysis could be helpful for various reasons. I'm interested in understanding my own behavior better but other organizations could find this information insightful for various reasons including the following:



0.5 Conclusion

I used various approaches and packages to build and test classification models for predicting song genres. Overall, the bidirectional LSTM model performed best, not a surprise given the ability to account not only for word frequency but also sequencig, and thus extracting greater meaning from the text data during the training stage.

In future iterations, I plan to refine my deep learning models further and to assess classification performance on other attritibutes such as artist name.