Predicting Music Genre with Lyrics and Machine Learning Algorithms

26 June, 2021

GEORGETOWN UNIVERSITY

School of Continuing Studies
Certificate in Data Science

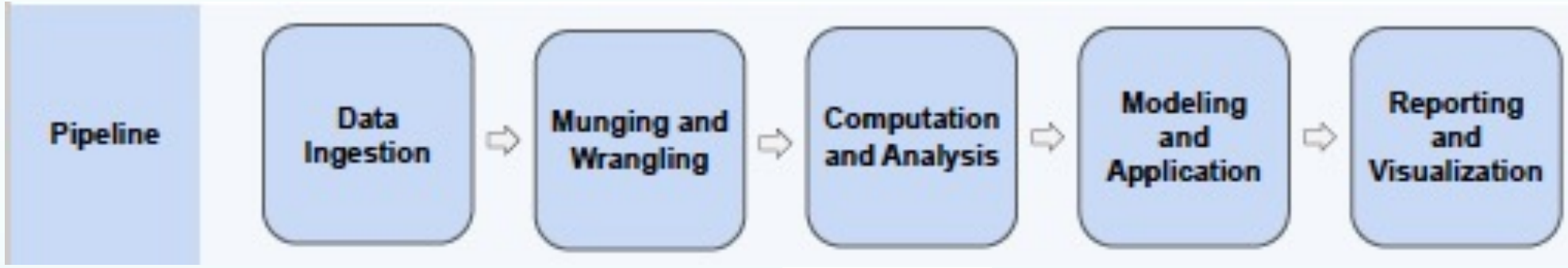
AGENDA

1. *Hypothesis and Framing*
2. *Planned vs. Actual*
3. *Data, Wrangling*
4. *Natural Language Processing (NLP) PreProcessing (PP), Munging*
5. *Exploratory Data Analysis (EDA)*
6. *Feature Engineering & Feature Selection*
7. *Algorithm Selection*
8. *Hyperparameter Tuning*
9. *Results*
10. *Conclusion*

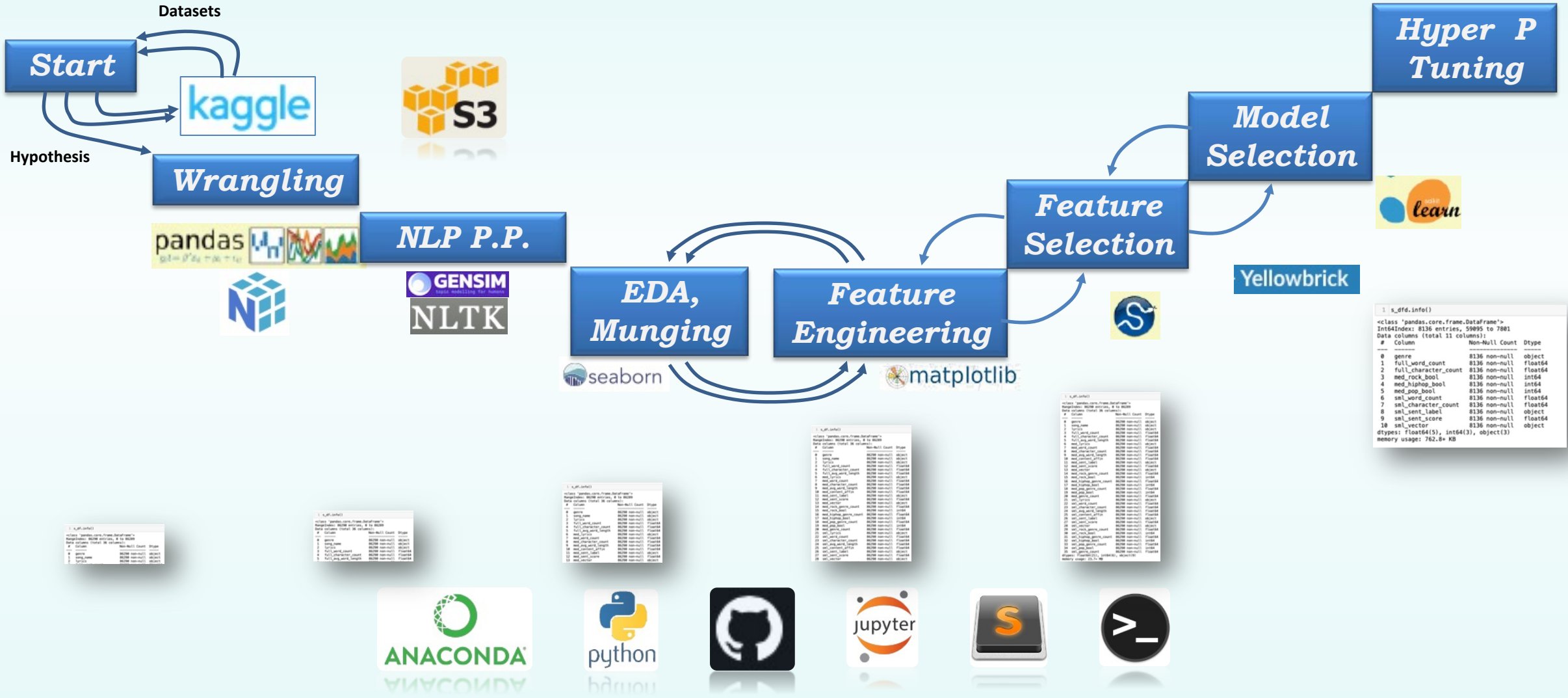
Hypothesis

- *Given only a song's lyrics, it is possible to classify the genre for that song using machine learning, with Hip Hop most readily identifiable.*
- *Blockers:*
  - *Natural Language Processing (NLP) Tools historically have been optimized for literature and tweets, and not focused on the core attributes of music.*
  - *Genre classification is subjective*
    - *Dataset covers decades of songs*
- *Mitigation:*
  - *Domain-specific stopwords list*
  - *Genre-specific corpus*

Planned Project Pipeline



Actual Project Pipeline



Data

Final Candidates

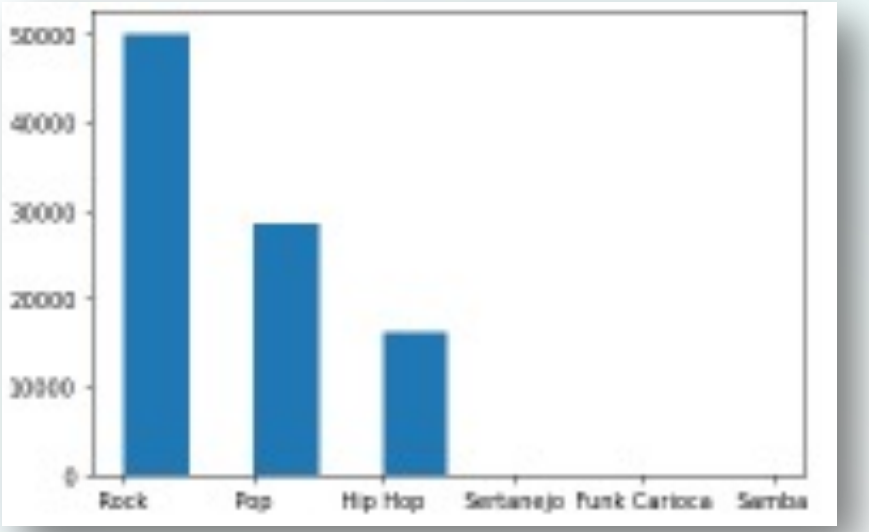
Song Lyrics from 6 Musical Genres		340kb	genres_artists_data.csv	Artist	# Songs	# Popularity	Link	Genre	Genres
167499 tracks	Rock, Pop, Sertanejo, Hip Hop, Funk Carioca	263.4MB	genres_lyrics_data.csv	Alink	Sname	Slink	Lyric	Idiom	
				/band name/	song name clean	web link	all the words	language in all caps	
Link									
Music Dataset: 1950 - 2019	pop, country, blues, rock, jazz	26.4mb	decades_tcc_ceds_music.csv	#	artist_name	track_name	release_date	genre	lyrics
23689 tracks							Plus 24 other other classifications ->		
Link									
Song Lyrics	No Genre Information	103.1kb	album_details_25k.csv	#	id	singer_name	album_name	type	year
25000 tracks		38.9MB	Lyrics_25k.csv	#	link	artist	song_name	lyrics	
Link		2.2MB	songs_details_25k.csv	#	song_id	singer_name	song_name	song_href	
???	No Genre Info	178.7MB	labled_lyrics_cleaned.csv	#	artist	seq	song	label	

Artist | Song | Lyric | Genre | Year

- *A hypothesis aligned with resources helped focus the data search*
- *'Musical Genres' Selected*
  - *Larger dataset*
  - *Multi-class (Rock, Pop, Hip Hop)*

"I used to think one day we'd tell the story of us. How we met and sparks flew instantly. People would say 'They're the lucky ones'. I used to know my place was a spot next to you. Now I'm searching the room for an empty seat. 'Cause lately I don't even know what page you're on. Oh. A simple complication, miscommunication. Has lead to fallout. Too many things that I wish you knew. So many walls up I can't break through. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. And I don't know what to say. Since the twist of fate. When it all broke down. And the story of us looks a lot like a tragedy now. Next chapter. How'd we end up this way?. Se me nervously pulling at my clothes. And trying to look busy. And you're doing your best to avoid me. I'm starting to think one day I'll tell the story of us. How I was losing my mind when I saw you here. But you held your pride like you should've held me. Oh. I'm scared to see the ending. Why are we pretending this is nothing?. I'd tell you I miss you but I don't know how. I've never heard silence quite this loud. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. And I don't know what to say. Since the twist of fate. When it all broke down. And the story of us looks a lot like a tragedy now. This is looking like a contest. Of who can act like they care less. But I liked it better when you were on my side. The battle's in your hands now. But I will lay my armor down. If you say you'd rather love than fight. So many things that you wish I knew. But the story of us might be ending soon. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. But I don't know what to say. Since the twist of fate. And it all broke down. And the story of us looks a lot like a tragedy now. Now, now. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. I don't know what to say. Since the twist of fate. 'Cause we're going down. And the story of us looks a lot like a tragedy now. The End"

Full Lyrics, with Punctuation



Number of English Lyrics, by Genre

NLP PreProcessing, Munging

Reduce corpus to words useful to machine learning

- *All lower case*
- *Remove digits, accented characters, whitespace, etc.*
- *Expand contractions, align to present tense*
- *Remove 2-digit words & genism stop_words list*

```
genres_df['lyrics']=genres_df['lyrics'].apply(lambda x: x.lower())
```

```
genres_df['lyrics']=genres_df['lyrics'].apply(remove_urls)
genres_df['lyrics']=genres_df['lyrics'].apply(remove_www)
genres_df['lyrics']=genres_df['lyrics'].apply(remove_special_characters)
genres_df['lyrics'] =genres_df['lyrics'].apply(remove_extra_whitespace_tabs)
genres_df['lyrics'] =genres_df['lyrics'].apply(remove_digits)
genres_df['lyrics'] =genres_df['lyrics'].apply(remove_accented_chars)
genres_df['lyrics'] =genres_df['lyrics'].apply(expand_contractions)

genres_df['lyrics'] =genres_df['lyrics'].apply(replace_punctuation)
genres_df['lyrics'] =genres_df['lyrics'].apply(stops_letters)
```

```
def lemmatized_word(text):

    word_tokens = nltk.word_tokenize(text)
    lemmatized_word = [wordnet_lemmatizer.lemmatize(word) for word in word_tokens]
    return  " ".join(lemmatized_word) #combine the words into a giant string that v
```

```
genres_df['vector'] = genres_df['lyrics'].apply(lemmatized_word)
```

Before

```
1 genres_df.iloc[84304]['lyrics']#genres Taylor Swift The Story of Us
```

"I used to think one day we'd tell the story of us. How we met and sparks flew instantly. People would say 'They're the lucky ones'. I used to know my place was a spot next to you. Now I'm searching the room for an empty seat. 'Cause lately I don't even know what page you're on. Oh. A simple complication, miscommunication. Has lead to fallout. Too many things that I wish you knew. So many walls up I can't break through. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. And I don't know what to say. Since the twist of fate. When it all broke down. And the story of us looks a lot like a tragedy now. Next chapter. How'd we end up this way?. Se me nervously pulling at my clothes. And trying to look busy. And you're doing your best to avoid me. I'm starting to think one day I'll tell the story of us. How I was losing my mind when I saw you here. But you held your pride like you should've held me. Oh. I'm scared to see the ending. Why are we pretending this is nothing?. I'd tell you I miss you but I don't know how. I've never heard silence quite this loud. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. And I don't know what to say. Since the twist of fate. When it all broke down. And the story of us looks a lot like a tragedy now. This is looking like a contest. Of who can act like they care less. But I liked it better when you were on my side. The battle's in your hands now. But I will lay my armor down. If you say you'd rather love than fight. So many things that you wish I knew. But the story of us might be ending soon. Now I'm standing alone in a crowded room. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. But I don't know what to say. Since the twist of fate. And it all broke down. And the story of us looks a lot like a tragedy now. Now, now. And we're not speaking. And I'm dying to know. Is it killing you like it's killing me?. I don't know what to say. Since the twist of fate. 'Cause we're going down. And the story of us looks a lot like a tragedy now. The End"

Average of 274 words per song.

```
1 genres_df.iloc[84304]['vector']#genres Taylor Swift The Story of Us
```

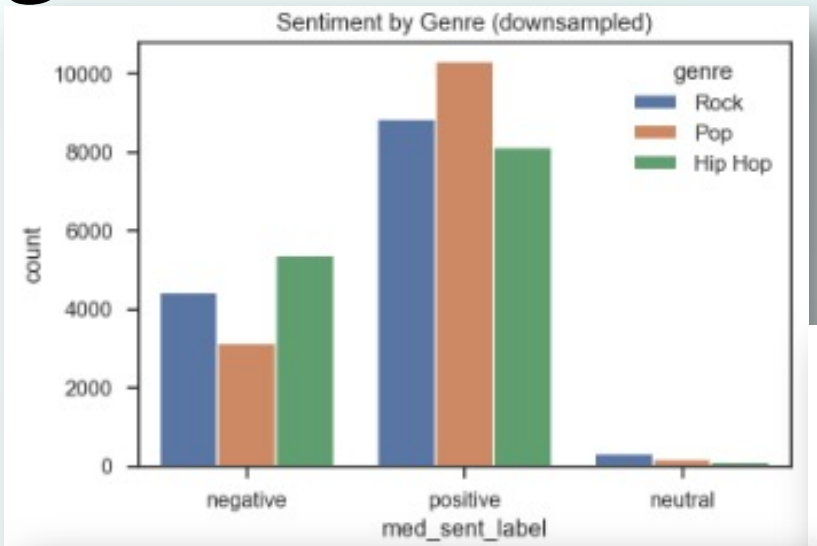
'think day tell story met spark flew instantly people lucky one know place spot searching room seat lately know page simple complication lead fallout thing wish knew wall break standing crowded room speaking dying know killing like killing know twist fate broke story look lot like tragedy chapter end way nervously pulling clothes trying look busy best avoid starting think day tell story losing mind saw held pride like held scared ending pretending tell miss know heard silence loud standing crowded room speaking dying know killing like killing know twist fate broke story look lot like tragedy looking like contest act like care liked better battle hand lay armor love fight thing wish knew story ending soon standing crowded room speaking dying know killing like killing know twist fate broke story look lot like tragedy speaking dying know killing like killing know twist fate going story look lot like tragedy end'

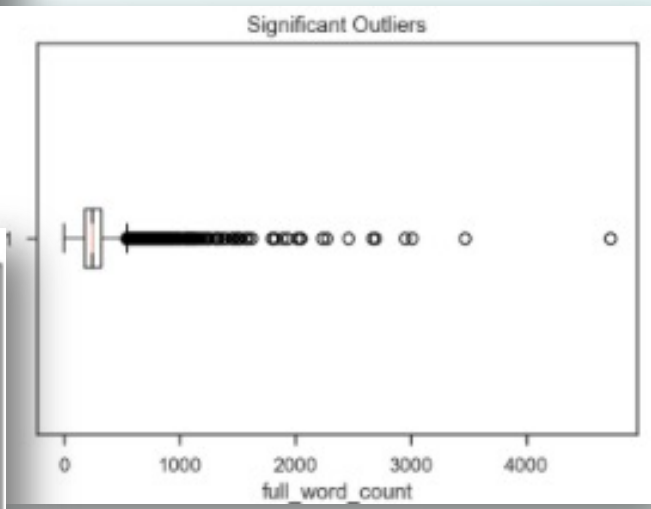
After NLP PP: Average of 108

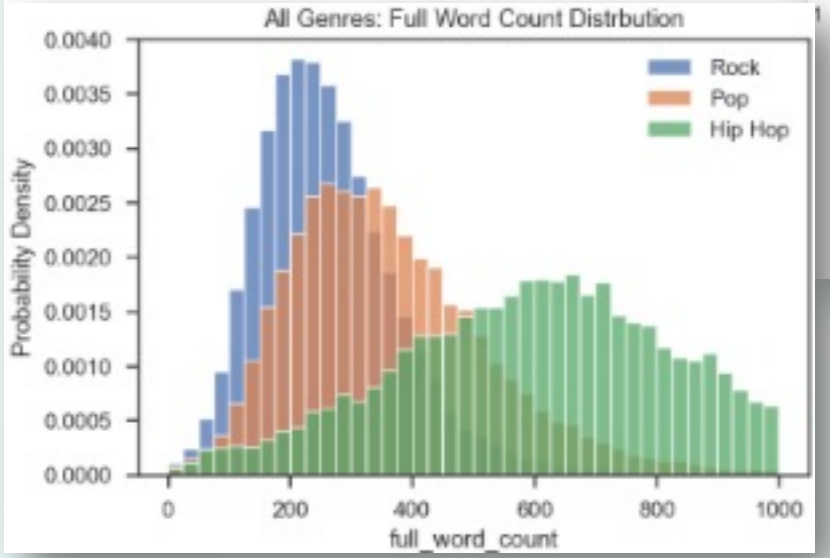
NLP PP, Munging

Generate Fundamental, Simple Information

- *Created Features*
  - *Full Lyrics*
    - *Word / Character Counts*
    - *Average character per word*
  - *Medium Lyrics*
    - *Counts*
    - *AFINN Lexicon Scores*
      - *[-1.0, 1.0]*
    - *TextBlob Sentiment Analysis:*
      - *Label & Score*
      - *[-0.5, 0.5]*







EDA – Words, Words, Words

- *Unigram, by genres, total frequency*
- *Entire corpus, words and frequency*
- *Tri-grams show impact of chorus*

	Hip Hop	0
0	like	46642
1	got	35368
2	know	33244
3	don	32809
4	just	25822
5	ain	21722
6	[illegible]	20765
7	love	19968
8	yeah	18635
9	let	17348
10	[illegible]	16286

	Rock	0
0	don	52511
1	love	43442
2	know	43362
3	just	40292
4	ll	37227
5	like	35838
6	oh	35330
7	got	29389
8	ve	28977
9	time	28532
10	let	23199

	Pop	0
0	love	49089
1	don	45462
2	oh	43967
3	know	41600
4	like	37362
5	just	34913
6	ll	27505
7	baby	26319
8	got	25833
9	let	24729
10	yeah	22165

```
4 all_the_words_df
```

	All	All_Count
0	like	149428
1	know	147687
2	got	140398
3	love	138435
4	want	124151
...	...	...
135660	hhhands	1
135661	dumdiddydum	1
135662	dunitdunitdunit	1
135663	crackalackalin	1
135664	whodundunit	1

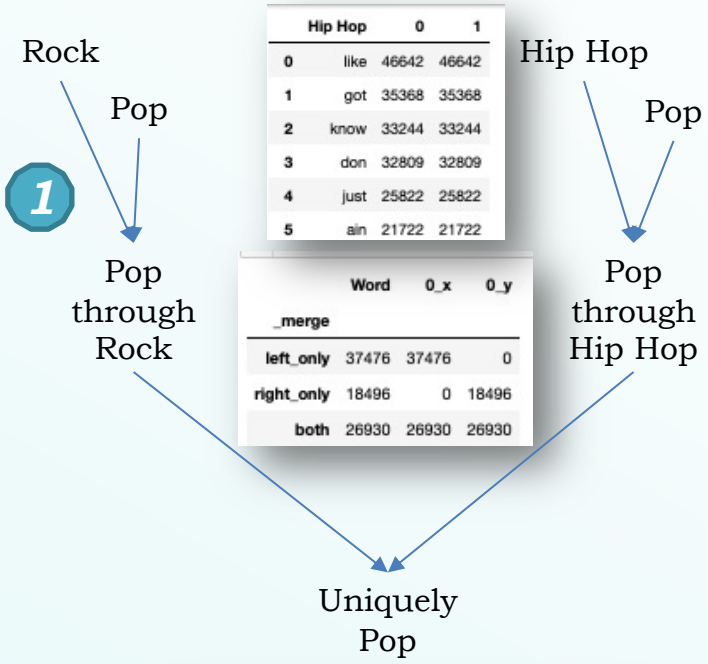
Expanded Stop Words List

```
1 n_gram_df.head(20)
```

	Rock	0	Pop	1	Hip Hop	2	All	3
0	yeah yeah yeah	3027	yeah yeah yeah	3658	yeah yeah yeah	1961	yeah yeah yeah	8646
1	love love love	2060	love love love	2997	love love love	976	love love love	6033
2	hey hey hey	1906	know know know	1442	hey hey hey	744	hey hey hey	3853
3	come come come	1177	want want want	1301	know know know	724	know know know	3296
4	know know know	1130	hey hey hey	1203	like like like	507	want want want	2796
5	want want want	1045	ooh ooh ooh	1202	baby baby baby	507	baby baby baby	2548
6	baby baby baby	995	baby baby baby	1046	got got got	452	come come come	2294
7	let let let	720	let let let	782	want want want	450	ooh ooh ooh	2037
8	run run run	539	work work work	774	come come come	412	let let let	1856
9	ooh ooh ooh	537	nah nah nah	725	work work work	369	got got got	1491
10	whoa whoa whoa	500	come come come	705	girl girl girl	358	like like like	1374
11	got got got	476	like like like	584	let let let	354	work work work	1204
12	time time time	443	got got got	563	stop stop stop	318	nah nah nah	1191
13	going going going	415	way way way	530	ooh ooh ooh	298	whoa whoa whoa	1039
14	doo doo doo	403	know want know	426	whoa whoa whoa	292	way way way	1037
15	alright alright alright	399	night night night	411	nah nah nah	290	run run run	965
16	away away away	359	shake shake shake	400	white white white	282	know want know	961
17	way way way	354	dance dance dance	364	money money money	276	want know want	886
18	round round round	352	want know want	357	low low low	257	time time time	875
19	want know want	349	run run run	349	right right right	248	round round round	865

Creation of Genre-Specific Corpus

35 Features



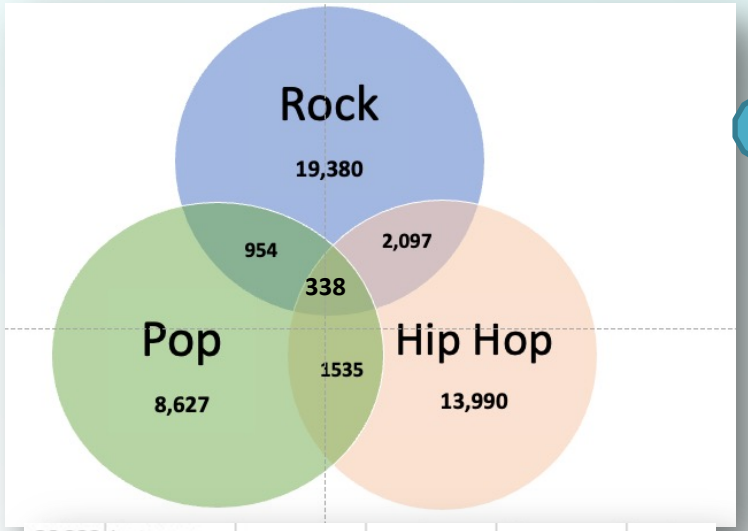
	Hip Hop	0	1
0	like	46642	46642
1	got	35368	35368
2	know	33244	33244
3	don	32809	32809
4	just	25822	25822
5	ain	21722	21722

_merge	Word	0_x	0_y
left_only	37476	37476	0
right_only	18496	0	18496
both	26930	26930	26930

Applied to 80% of the dataset

Expansion of CountVectorizer

Modification of nltk.corpus.stopwords.



86,290	Instances
41,997	Instances with > 1 genre specific word
5,205	Instances with >1 genre specifc word in more than one genre

_merge	Word	Count_x	Count_y
left_only	0	0	0
right_only	0	0	0
both	23091	23091	23091

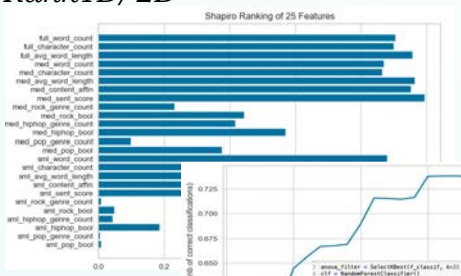
2

Total Words		Words Used	Unique Words (100%)	Unique Words (80%)	Canonical Words
135,665	Hip Hop	80,152	40,652	29,843	2,054
	Pop	55,477	23,091	19,010	1,073
	Rock	69,860	30,702	13,757	1,871

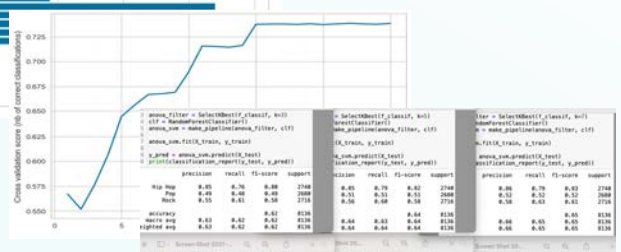
3

Final Feature Selection

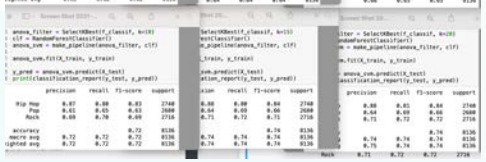
Rank1D/2D

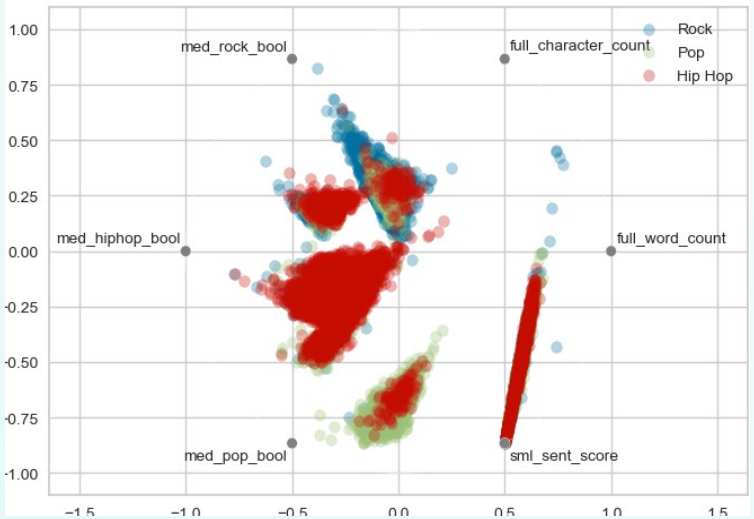


RFECV



SelectKBest ANOVA





Criteria

- ***Impactful***
- ***Avoid Co-Linearity***
- ***Most descriptive***

ColumnTransformer:

- *Target*
- *RobustScaler*
- *MinMaxScaler*
- *OneHotEncoder*
- *TfidfVectorizer*

#	Column	Non-Null	dType	Notes:	
0	genre	86290	*Object*	Target	
1	song_name	86290	*Object*	dataset	***Full Lyrics***
2	lyrics	86290	*Object*	dataset	
3	full_word_count	86290	**Int64**	lambda	
4	full_character_count	86290	**Int64**	lambda	
5	full_avg_word_length	86290	**float64**	lambda	
6	med_lyrics	86290	*Object*	cleaned lemmatized, gensim stopwords, 39% of full_lyrics	
7	med_word_count	86290	**Int64**	lambda	
8	med_character_count	86290	**Int64**	lambda	***Medium***
9	med_avg_word_length	86290	**float64**	lambda	
10	med_content_affin	86290	**float64**	AFFIN lexicon [-1.0, 1.0]	
11	med_sent_label	86290	*Object*	TextBlob - positive, negative, neutral	
12	med_sent_score	86290	**float64**	TextBlob [-0.5, 0.5]	
13	med_vector	86290	*Object*	NLTK punkt and wordnet, bag of words, in order.	
14	med_rock_genre_count	86290	**float64**	# of unique rock words in the lyrics X .01	
15	med_rock_bool	86290	**Int64**	0 or 1, rock word present?	
16	med_hiphop_genre_count	86290	**float64**	# of unique hiphop words in the lyrics X 100	
17	med_hiphop_bool	86290	**Int64**	0 or 1, hiphop word present?	
18	med_pop_genre_count	86290	**float64**	# of unique pop words in the lyrics X 1.0	
19	med_pop_bool	86290	**Int64**	0 or 1, pop word present?	
20	med_genre_count	86290	**Int64**	Sum of rock, pop, and hiphop genre counts.	
21	sml_lyrics	86290	*Object*	med_lyrics run through NLTK stop_words and genre_stopwords	
22	sml_word_count	86290	**Int64**	lambda	
23	sml_character_count	86290	**Int64**	lambda	***Small***
24	sml_avg_word_length	86290	**float64**	lambda	
25	sml_content_affin	86290	**float64**	AFFIN lexicon [-1.0, 1.0]	
26	sml_sent_label	86290	*Object*	TextBlob - positive, negative, neutral	
27	sml_sent_score	86290	**float64**	TextBlob [-0.5, 0.5]	
28	sml_vector	86290	*Object*	NLTK punkt and wordnet, bag of words, in order.	
29	sml_rock_genre_count	86290	**float64**	# of unique rock words in the lyrics X .01	
30	sml_rock_bool	86290	**Int64**	0 or 1, rock word present?	
31	sml_hiphop_genre_count	86290	**float64**	# of unique hiphop words in the lyrics X 100	
32	sml_hiphop_bool	86290	**Int64**	0 or 1, hiphop word present?	
33	sml_pop_genre_count	86290	**float64**	# of unique pop words in the lyrics X 1.0	
34	sml_pop_bool	86290	**Int64**	0 or 1, pop word present?	
35	sml_genre_count	86290	**Int64**	Sum of rock, pop, and hiphop genre counts.	

Model Selection

Supervised, Classification, Multi-Class, Downsampled

Ensemble, Tree, Effective

Effective

Baseline Model for Text Analysis

Neural Net

	Rock	Pop	Hip Hop	F1
LinearSVC,	0.746	0.646	0.858	0.750
SVC	0.674	0.672	0.822	0.723
BaggingClassifier,	0.708	0.669	0.856	0.744
ExtraTreesClassifier,	0.75	0.694	0.867	0.770
RandomForestClassifier,	0.744	0.666	0.859	0.756
DecisionTreeClassifier	0.669	0.615	0.82	0.701
AdaBoostClassifier	0.738	0.648	0.863	0.750
KNeighborsClassifier	0.659	0.626	0.827	0.704
LogisticRegressionCV,	0.747	0.694	0.87	0.770
LogisticRegression,	0.745	0.692	0.874	0.770
SDGClassifier	0.749	0.688	0.866	0.768
MultinomialNB	0.735	0.689	0.853	0.759
GaussianNB				
BernoulliNB	0.733	0.61	0.807	0.717
MLPClassifier	0.7	0.646	0.851	0.732
GradientBoostingClassifier	0.745	0.685	0.869	0.766
Avg all models	*0.723*	*0.663*	*0.851*	

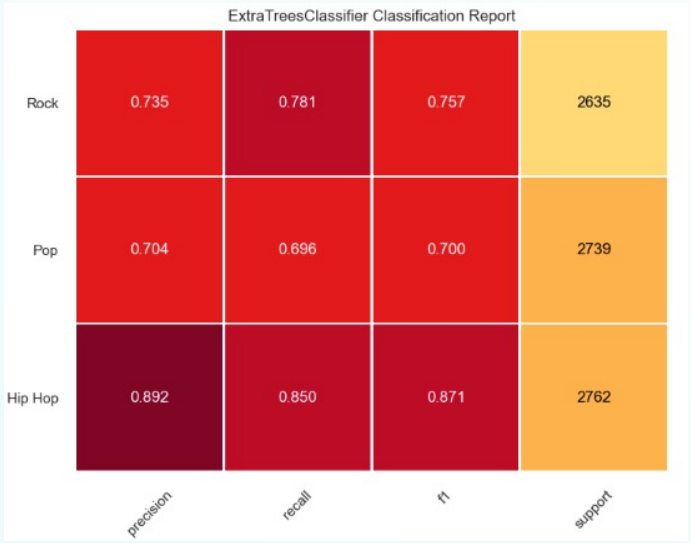
All have predict_proba(X) method

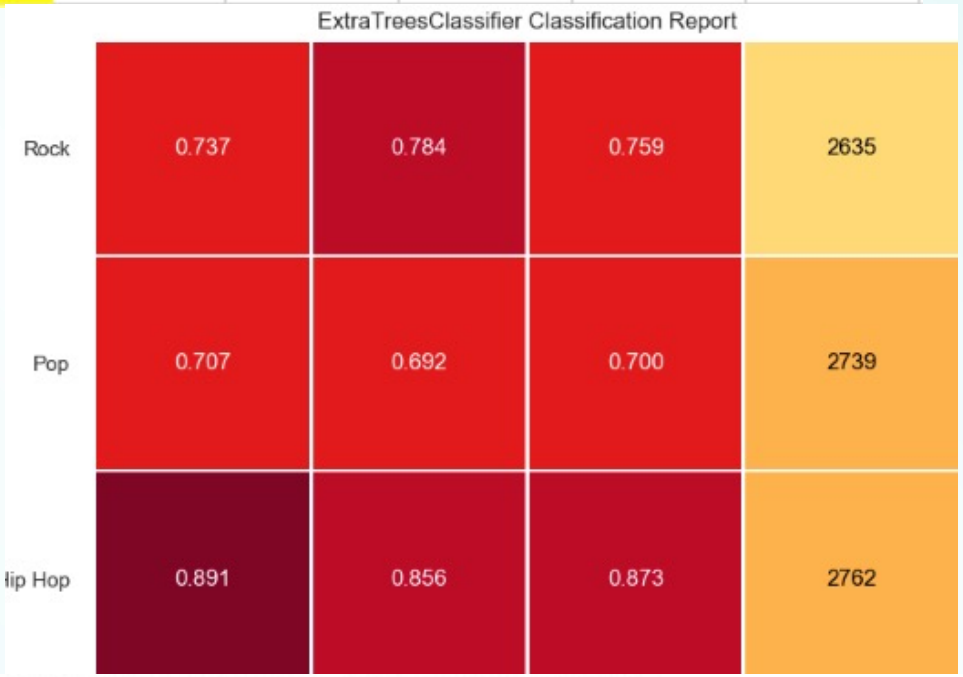
ExtraTreesClassifier

Well-tuned out of the box

Parameter1	Grid Serached	Default	Selected

n_estimators	10	50	100	125	150	175	200
criterion	gini	entropy					
max_depth	100	250	500	750	1000	none	
bootstrap	TRUE	FALSE					
oob_score	TRUE	FALSE					
warm_start	TRUE	FALSE					





	Rock	Pop	Hip Hop	F1
First Pass	0.75	0.694	0.867	0.770
Second Pass	0.759	0.7	0.873	0.777
Delta	0.009	0.006	0.006	0.007

```
Accuracy on x_train is 0.9999078171091446
Accuracy on x_test is 0.7764257620452311
CPU times: user 2min 43s, sys: 967 ms, total: 2min 44s
Wall time: 22.7 s
```

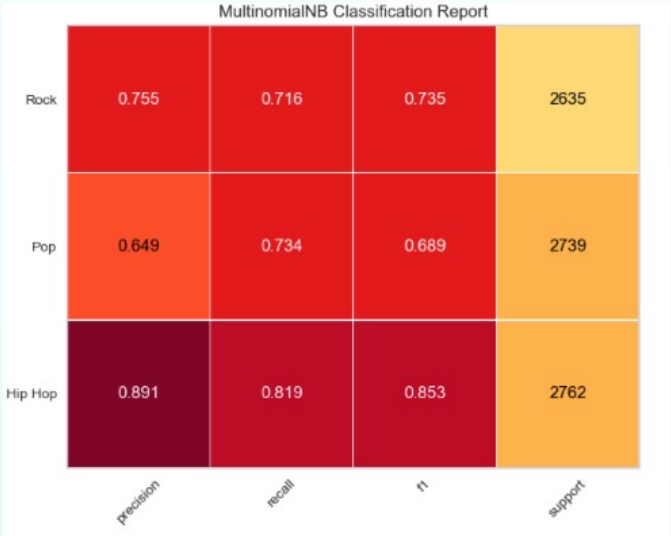
MultinomialNBClassifier

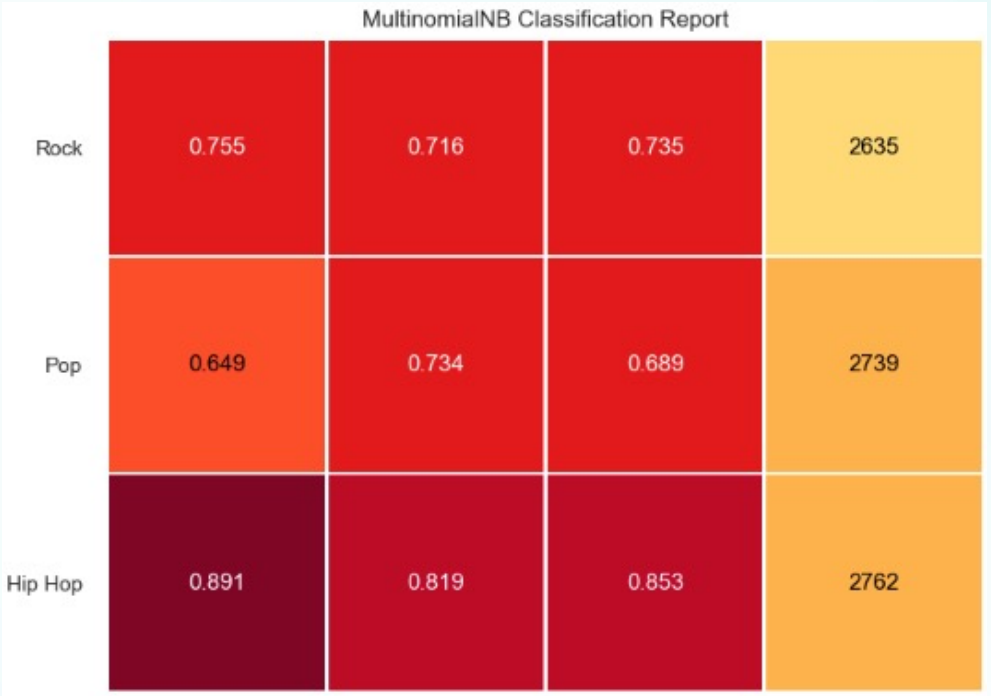
Not that many knobs to turn. Wicked Fast.

Parameter1	Grid Serached	Default	Selected

Alpha	0	0.5	1	1.5	2	3	
fit_prior	TRUE	FALSE					

Didn't matter





	Rock	Pop	Hip Hop	F1
First Pass	0.735	0.689	0.853	0.759
Second Pass	0.735	0.689	0.853	0.759
Delta	0	0	0	0

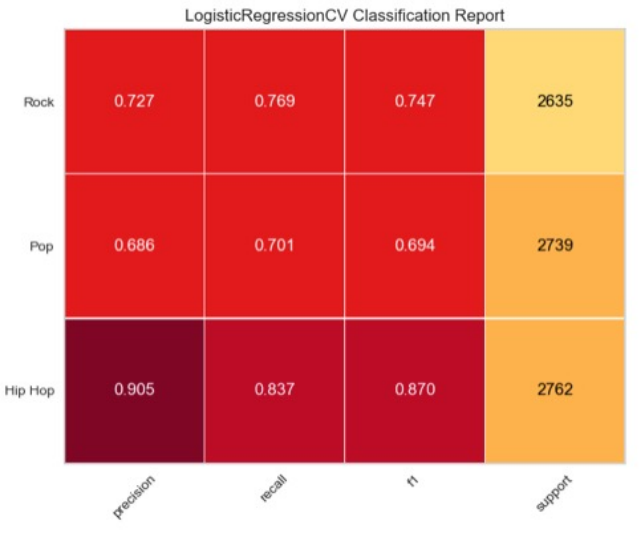
```
Accuracy on x_train is 0.7745206489675516
Accuracy on x_test is 0.7568829891838741
CPU times: user 33 ms, sys: 2.66 ms, total: 35.7 ms
Wall time: 33.7 ms
```

LogisticRegressionCV

Well-tuned out of the box.

Parameter1	Grid Serached	Default	Selected

Cs	1	5	10	25		
CV	1	5	12			
Solver	lbfgs	newton-cg	liblinear	sag	saga	saga
Penalty	l2	l2	l1, l2	l2	elasticnet	l1
multi_class	auto	ovr	multinomial		Wouldn't converge	



	Rock	Pop	Hip Hop	F1
First Pass	0.747	0.694	0.87	0.770
Second Pass	0.747	0.694	0.87	0.770
Delta	0	0	0	0

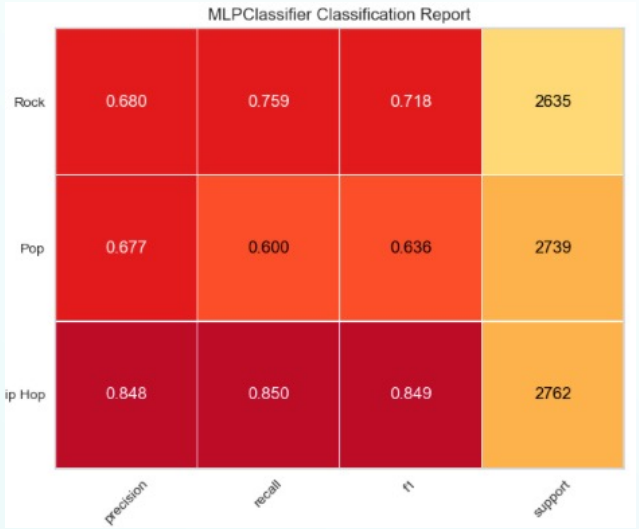

```
Accuracy on x_train is 0.8054633726647001
Accuracy on x_test is 0.7689282202556539
CPU times: user 1min 25s, sys: 2min, total: 3min 25s
Wall time: 1min 6s
```

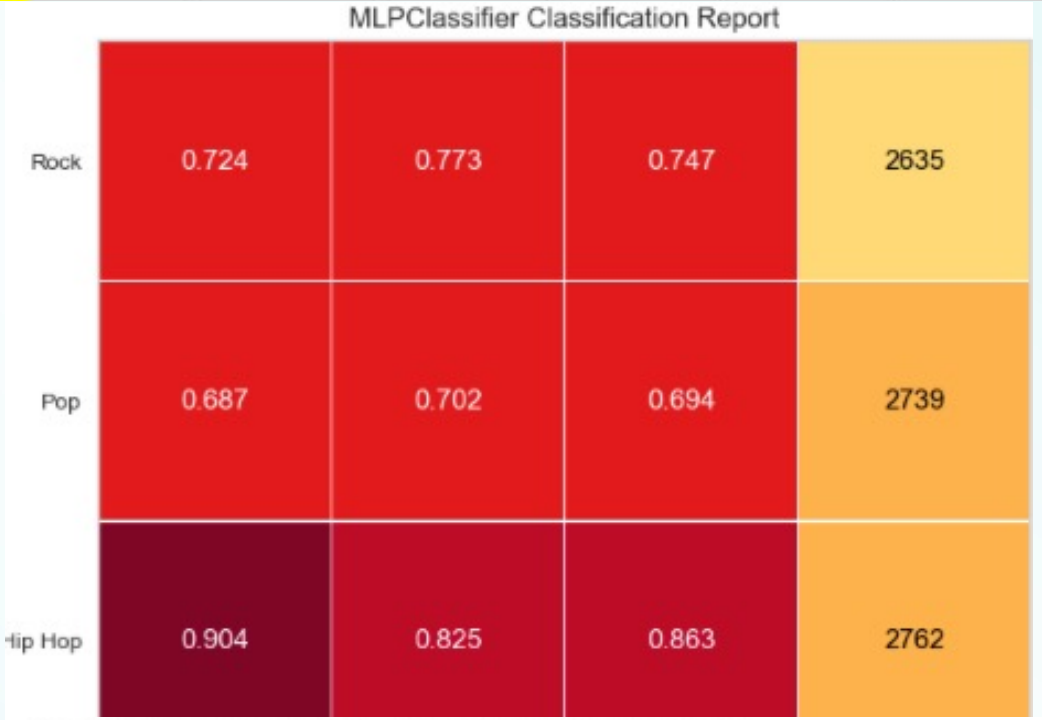
MLPClassifier

Has a max_fun parameter.

Parameter1	Grid Serached	Default	Selected

hidden_layer_sizes	1	3	5	50	100	150	200	300
solver	lbfgs	sgd	adam					
activation	relu	logistic	tanh					
max_fun	15000	17000						
alpha	0.00001	0.0001	0.001	0.1	1	5		





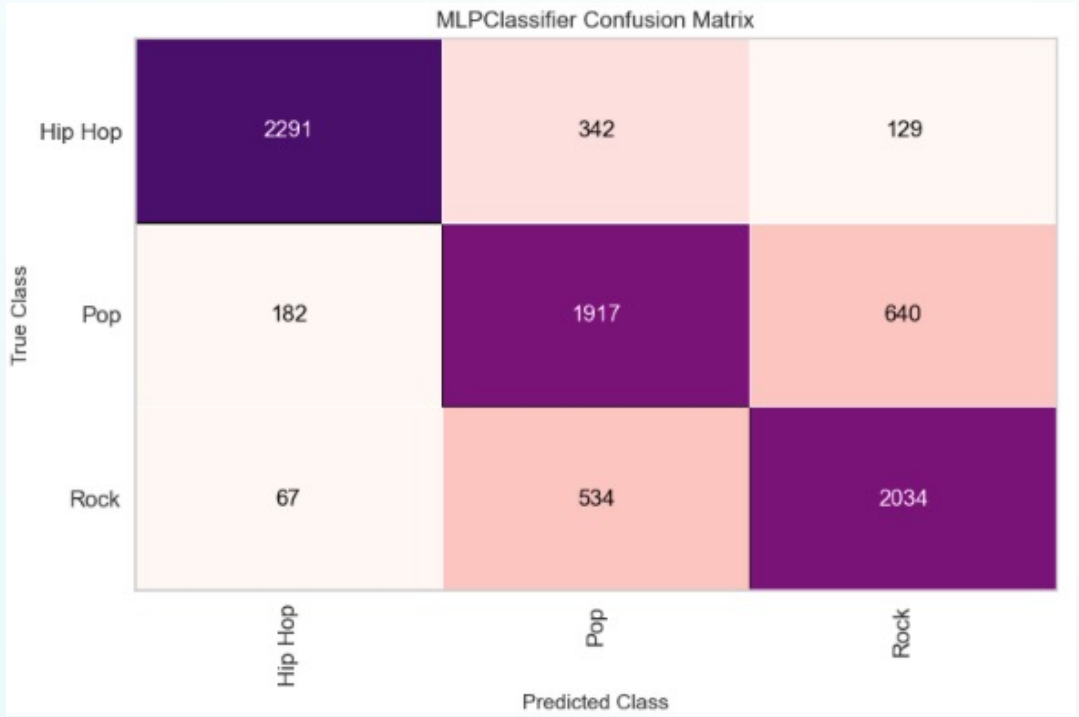
	Rock	Pop	Hip Hop	F1
First Pass	0.7	0.646	0.851	0.732
Second Pass	0.746	0.696	0.863	0.768
Delta	0.046	0.05	0.012	0.036

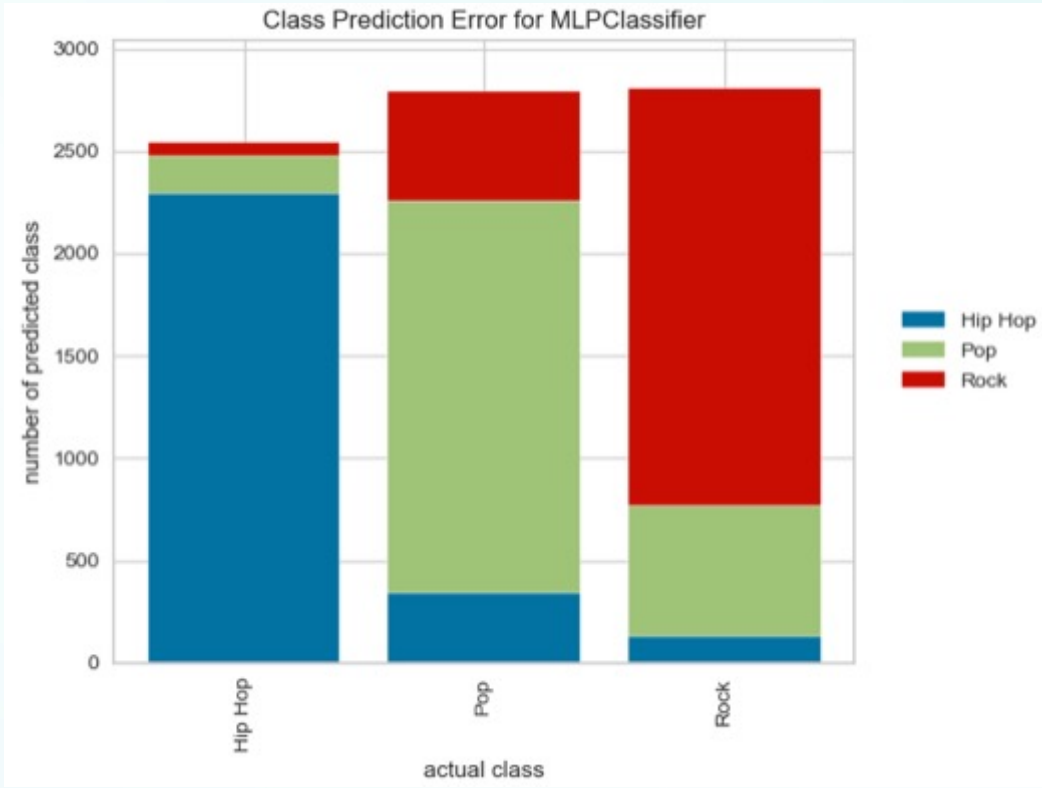
```
Accuracy on x_train is 0.7820796460176991
Accuracy on x_test is 0.7672074729596854
CPU times: user 12min 4s, sys: 21min 47s, total: 33min 51s
Wall time: 7min 13s
```

Parameter1	Grid Serached	Default	Selected

MLPClassifier

The line between Rock and Pop is more ambiguous.





Next Steps

- ***More Data***
  - *More Classes for More Relevance*
  - *Sound*
- ***More Wrangling***
  - *Identify Repeated Sets of Words (Choruses)*
    - *Evaluate Separately*
- ***More Classes, Identify Cross-Overs***
  - *Softmax Layer, predict_proba(X)*
- ***Better Model***
  - *Feature Union & Transformer Weights*
  - *Fine Tune Pruning and Weights*
- ***Data Product***
  - *Automated Scraping, Genre Lists, stop_words List -> "Musical Genre Lexicon"*

Hypothesis

- *Given only a song's lyrics, it is possible to classify the genre for that song using machine learning.*
- *Blockers:*
  - *Natural Language Processing (NLP) Tools historically have been optimized for literature and tweets, and not core attributes of music.*
  - *Genre classification is subjective*
    - *Dataset covers decades of songs*
- *Mitigation:*
  - *Domain-specific stopwords list*
  - *Genre-specific corpus*

GEORGETOWN UNIVERSITY