06/12/2022

IBM Advanced Data Science Specialization

Capstone Project

Spotify Daily Global Top 200 Analysis

Krishan Deo

Agenda

1) Stakeholder Presentation

- The Dataset
- The Use Case
- The Solution
  - Notebooks
  - Model Performance
  - Feature Importances
  - Key Insights + Next Steps

2) Data Science Peers Presentation

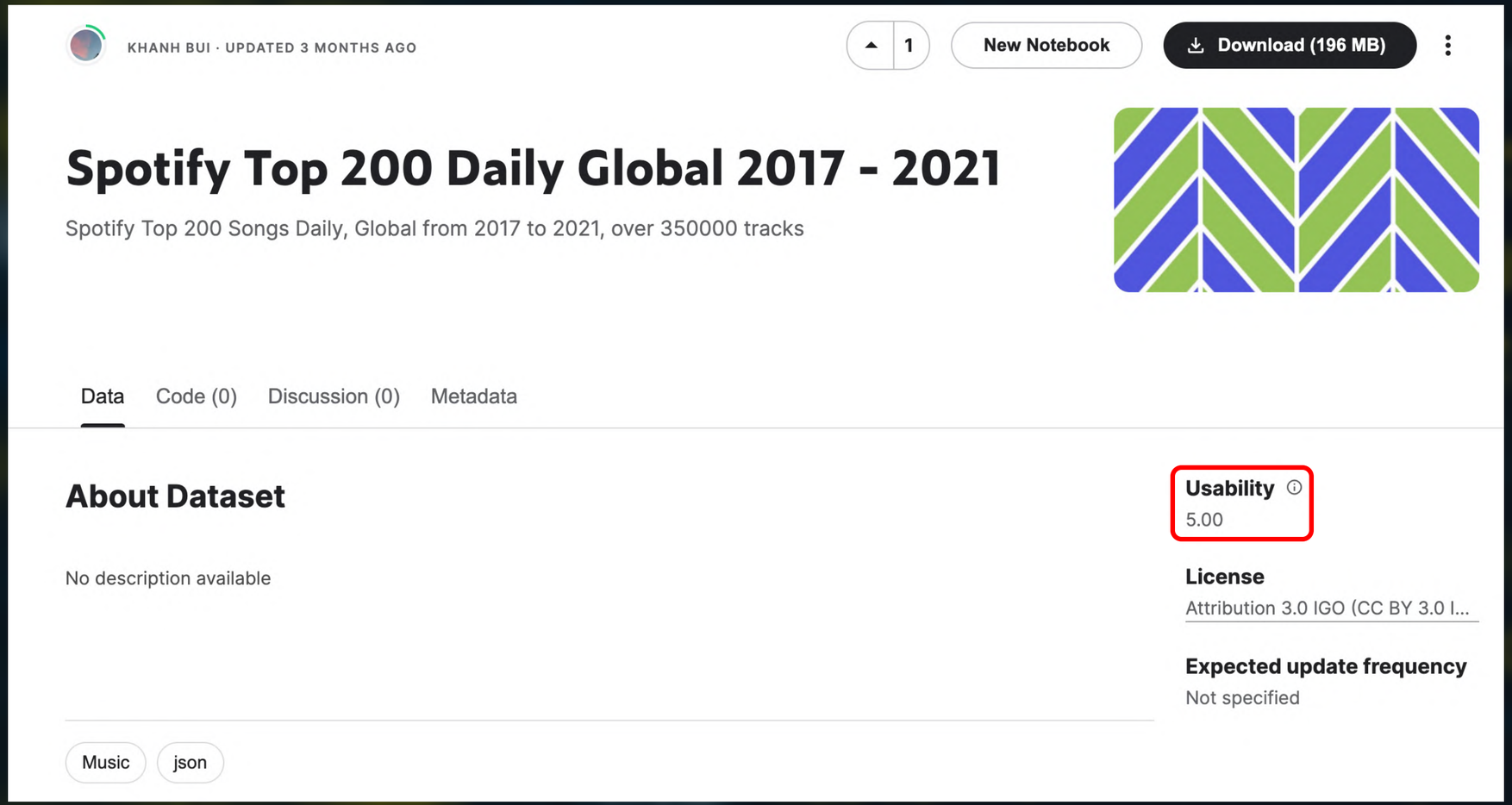
- Architectural Choices
- Data Quality Assessment
- Data Pre-Processing
- Feature Engineering
- Model Performance Indicators
- Chosen Algorithm(s)
- Thank You!

Spotify®

Stakeholder Presentation

Krishan Deo

The Dataset



https://www.kaggle.com/c0lydxmas/spotify-top-200-daily-global-2017-2021

The Dataset


```
"root" : [ 12 items
  0 : { 33 items
    "genres" : [...] 3 items
    "danceability" : float 0.681
    "energy" : float 0.594
    "key" : int 7
    "loudness" : float -7.028
    "mode" : int 1
    "speechiness" : float 0.282
    "acousticness" : float 0.165
    "instrumentalness" : float 0.00000349
    "liveness" : float 0.134
    "valence" : float 0.535
    "tempo" : float 186.054
    "type" : string "track"
    "id" : string "5aAx2yezTd8zXrkmtKl66Z"
    "uri" : string "spotify:track:5aAx2yezTd8zXrkmtKl66Z"
    "track_href" : string "https://api.spotify.com/v1/tracks/5aAx2yezTd8zXrkmtKl66Z"
    "analysis_url" : string "https://api.spotify.com/v1/audio-analysis/5aAx2yezTd8zXrkmtKl66Z"
    "duration_ms" : int 230453
    "time_signature" : int 4
    "album" : {...} 13 items
    "artists" : [ 2 items
      0 : { 6 items
        "external_urls" : {...} 1 item
        "href" : string "https://api.spotify.com/v1/artists/1Xyo4u8uXC1ZmMpatF05PJ"
        "name" : string "Daft Punk"
        "type" : string "artist"
        "uri" : string "spotify:artist:4tZwfgrHOc3mvqYlEYSvVi"
      }
    ]
    "available_markets" : [] 0 items
    "disc_number" : int 1
    "explicit" : bool true
    "external_ids" : {...} 1 item
    "external_urls" : {...} 1 item
    "href" : string "https://api.spotify.com/v1/tracks/5aAx2yezTd8zXrkmtKl66Z"
    "is_local" : bool false
    "name" : string "Starboy"
    "popularity" : int 1
    "preview_url" : NULL
    "track_number" : int 1
    "time" : string "2017-01-02"
}
```

ETL Script

In [ ]:

```python
# ETL Script - Spotify Top 200 Data

# Imports

import pandas as pd
import glob, os, json
import numpy as np

pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth',1000)

# Nested JSON Parsing and Concatenation

json_dir = 'C:/Users/krishand/Downloads/SPOTIFY/archive/'

json_pattern = os.path.join(json_dir, '*.json')
file_list = glob.glob(json_pattern)

dfs = []
for file in file_list:
    with open(file, 'r', encoding='utf-8') as f:
        json_data = pd.json_normalize(json.loads(f.read()))
        json_data['site'] = file.rsplit("/", 1)[-1]
    dfs.append(json_data)
df = pd.concat(dfs)

df = df.reset_index()
df['index'] = df['index']+1
df = df.rename(columns={'index': 'position'})

# Extracting Artist Names

df2 = pd.json_normalize(df['artists'])[0].to_frame(name="js")
df3 = pd.json_normalize(df2['js'])['name']

df_2 = pd.json_normalize(df['artists'])[1].to_frame(name="js2")
df_3 = pd.json_normalize(df_2['js2'])['name']

df_2a = pd.json_normalize(df['artists'])[2].to_frame(name="js3")
df_3a = pd.json_normalize(df_2a['js3'])['name']

df_2b = pd.json_normalize(df['artists'])[3].to_frame(name="js4")
df_3b = pd.json_normalize(df_2b['js4'])['name']

df['artistname']=df3
df['featuredartist']=df_3
df['featuredartist2']=df_3a
df['featuredartist3']=df_3b

# Filtering Out Songs Released in a Different Year (1 Year Window)

df['time'] = pd.to_datetime(df['time'])
df['album.release_date'] = pd.to_datetime(df['album.release_date'])
df['chart_year'] = df['time'].dt.year
df['release_year'] = df['album.release_date'].dt.year

df = df[df['chart_year']==df['release_year']]
```

```python
# Extract Initial Genre Information

df['numOfGenres'] = df['genres'].str.len()
df['primary_genre'] = df['genres'].str[0]
df['secondary_genre'] = df['genres'].str[1]
df['genre3'] = df['genres'].str[2]
df['genre4'] = df['genres'].str[3]
df['genre5'] = df['genres'].str[4]
df['genre6'] = df['genres'].str[5]
df['genre7'] = df['genres'].str[6]
df['genre8'] = df['genres'].str[7]
df['genre9'] = df['genres'].str[8]
df['genre10'] = df['genres'].str[9]
df['genre11'] = df['genres'].str[10]
df['genre12'] = df['genres'].str[11]
df['has_feature'] = pd.notnull(df['featuredartist'])
df['multigenre'] = pd.notnull(df['secondary_genre'])

# Feature Engineering - Sorting Genres By Popularity Instead of Alphabetic

genrecombine = pd.concat([df['primary_genre'], df['secondary_genre'],df['genre3'],df['genre4'],df['ger

sorter = genrecombine.to_frame().rename(columns={0: "Genre"}).groupby(by='Genre').size().sort_values(a
sorter['row_num'] = np.arange(len(sorter))+1
genlist = sorter['Genre'].tolist()
ranklist = sorter['row_num'].tolist()
sort_fin = dict(zip(genlist, ranklist))

df['sorted_genres'] = df['genres'].apply(lambda x: sorted(x,key=sort_fin.get))

# Extracting Sorted Genre Information

df['primary_genre'] = df['sorted_genres'].str[0]
df['secondary_genre'] = df['sorted_genres'].str[1]
df['genre3'] = df['sorted_genres'].str[2]
df['genre4'] = df['sorted_genres'].str[3]
df['genre5'] = df['sorted_genres'].str[4]
df['genre6'] = df['sorted_genres'].str[5]
df['genre7'] = df['sorted_genres'].str[6]
df['genre8'] = df['sorted_genres'].str[7]
df['genre9'] = df['sorted_genres'].str[8]
df['genre10'] = df['sorted_genres'].str[9]
df['genre11'] = df['sorted_genres'].str[10]
df['genre12'] = df['sorted_genres'].str[11]

# Filter Out Null Data

df = df[df['album.release_date_precision']=='day']
df = df[df['primary_genre'].notnull()]

# Create Label - Top 100 vs Bottom 100 Placement

df['binary_performance_bin'] = df['position'].apply(lambda x: 'Bottom 100' if x > 100 else 'Top 100')
```

```python
# Replace Null Values to Preserve Categorical Feature Utility in Future Steps

df.fillna('NABlankValue', inplace=True)

# Drop Unnessecary and No Longer Needed Columns to Shrink Dataset

df = df.drop(columns=['track_number','album.total_tracks','is_local','disc_number','artists','available_m

# Export

df.to_csv('C:/Users/krishan/Downloads/SPOTIFY/archive/2017.csv')
```

The Dataset

position	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	explicit	name
8	0.928	0.481	9	-9.35	0	0.287	0.105	0	0.176	0.613	134.007	210937	4	TRUE	Fake Love
13	0.852	0.773	8	-2.921	0	0.0776	0.187	3.05E-05	0.159	0.907	102.034	195840	4	FALSE	Chantaje (feat. Maluma)
16	0.49	0.485	4	-6.237	0	0.0406	0.0592	0	0.337	0.196	133.889	195840	4	FALSE	In the Name of Love
17	0.927	0.665	11	-5.313	1	0.244	0.061	0	0.123	0.175	127.076	343150	4	TRUE	Bad and Boujee (feat. Lil Uzi Vert)
21	0.469	0.68	11	-4.921	0	0.117	0.137	0	0.11	0.374	147.734	208733	4	FALSE	Mercy
22	0.697	0.691	2	-4.757	1	0.146	0.214	0	0.185	0.305	137.853	239293	4	FALSE	Bad Things (with Camila Cabello)
25	0.624	0.803	10	-4.105	0	0.246	0.123	0	0.114	0.821	166.018	187973	4	FALSE	Treat You Better
30	0.544	0.809	8	-5.098	1	0.0363	0.0038	0	0.323	0.448	145.017	197640	4	FALSE	All Night
32	0.781	0.57	11	-5.874	0	0.188	0.273	0	0.196	0.858	107.059	193181	4	FALSE	Now and Later
35	0.486	0.713	2	-3.949	0	0.0524	0.0853	0	0.0839	0.297	121.028	226720	4	FALSE	I Would Like
47	0.952	0.318	10	-10.357	1	0.467	0.174	0	0.205	0.665	120.077	209640	4	TRUE	Caroline
52	0.739	0.833	1	-5.012	1	0.0463	0.35	0	0.266	0.699	117.99	195313	4	FALSE	Perfect Strangers
57	0.78	0.575	1	-5.628	0	0.139	0.106	0	0.129	0.273	81.502	222360	4	TRUE	Bounce Back
59	0.602	0.707	9	-4.097	1	0.302	0.393	0	0.165	0.554	75.087	200186	4	FALSE	Human
70	0.422	0.852	6	-3.546	1	0.208	0.0845	0	0.0687	0.666	82.914	248386	4	TRUE	Lot to Learn

numOfGenres	primary_genre	secondary_genre	genre3	genre4	genre5	genre6
5	rap	hip hop	canadian pop	toronto rap	canadian hip hop	NABlankValue
5	pop	dance pop	latin	latin pop	colombian pop	NABlankValue
7	pop	dance pop	edm	pop dance	tropical house	progressive house
5	rap	pop rap	hip hop	trap	atl hip hop	NABlankValue
4	pop	dance pop	canadian pop	viral pop	NABlankValue	NABlankValue
2	pop rap	ohio hip hop	NABlankValue	NABlankValue	NABlankValue	NABlankValue
4	pop	dance pop	canadian pop	viral pop	NABlankValue	NABlankValue
4	pop	dance pop	post-teen pop	boy band	NABlankValue	NABlankValue
6	dance pop	rap	pop rap	trap	southern hip hop	hyphy
10	pop	dance pop	edm	pop dance	tropical house	post-teen pop
5	pop	rap	hip hop	underground hip hop	portland hip hop	NABlankValue
6	pop	dance pop	edm	pop dance	tropical house	uk dance
7	pop	rap	pop rap	hip hop	trap	southern hip hop
1	neo soul	NABlankValue	NABlankValue	NABlankValue	NABlankValue	NABlankValue
2	pop rap	indie pop rap	NABlankValue	NABlankValue	NABlankValue	NABlankValue

artistname	featuredartist
Drake	NABlankValue
Shakira	Maluma
Martin Garrix	Bebe Rexha
Migos	Lil Uzi Vert
Shawn Mendes	NABlankValue
Machine Gun Kelly	Camila Cabello
Shawn Mendes	NABlankValue
The Vamps	Matoma
Sage The Gemini	NABlankValue
Zara Larsson	NABlankValue
AminÃ©	NABlankValue
Jonas Blue	JP Cooper
Big Sean	NABlankValue
Rag'n'Bone Man	NABlankValue
Luke Christopher	NABlankValue

genre11	genre12	has_feature	multigenre	binary_performance_bin
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	FALSE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100

The Dataset - Features

Danceability	Energy	Loudness	Mode	Speechiness	Acousticness	Instrumentalness
Liveness	Valence	Tempo	Duration	Explicit	Has Feature	Multi-Genre
Key	Time Signature	Artist Name	Album Type	Featured Artist	Featured Artist 2	Featured Artist 3
Primary Genre	Secondary Genre	Genre 3	Genre 4	Genre 5	Genre 6	Genre 7
Genre 8	Genre 9	Genre 10	Genre 11	Genre 12		

Numeric

Categorical

The Use Case

Goals:

1) Achieve 80%+ AUROC in Binary Classification between Top 100 and Bottom 100 Songs for the Datasets 2017, 2018, 2019, 2020, and 2021 (Considered Excellent)
2) Determine the Primary Drivers (Feature Importances) behind the above mentioned Classifications
3) Utilize these Drivers to Generate Actionable Insights and Recommendations for Business Stakeholders
4) Determine Next Steps to Further Improve Models in Future Iterations

Methodology: Utilize a Combination of Categorical and Numerical Features within the 2017, 2018, 2019, 2020, and 2021 Datasets to Train 5 Different Binary Classifiers (one for each year) and Extract their Associated Feature Importances

The Solution - Notebooks

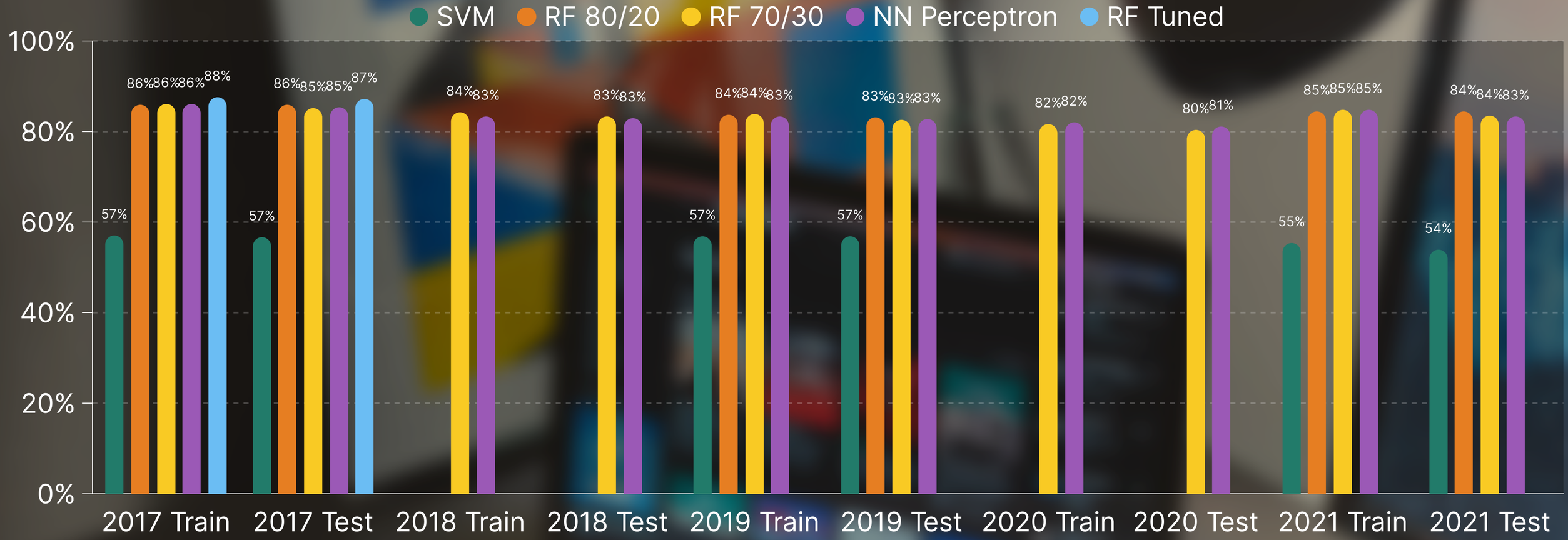
https://github.com/k-deo/Spotify-Daily-Global-Classifier-KD

The Solution - Model Performance

	SVM Baseline	RF 70/30	NN Perceptron	RF Tuned
2017 Train	57.0%	86.1%	86.2%	87.6%
2017 Test	56.7%	85.2%	85.4%	87.2%
2018 Train		84.3%	83.3%	
2018 Test		83.4%	83.0%	
2019 Train	56.9%	83.9%	83.3%	
2019 Test	56.8%	82.5%	82.7%	
2020 Train		81.7%	82.0%	
2020 Test		80.4%	81.1%	
2021 Train	55.4%	84.8%	84.8%	
2021 Test	53.8%	83.5%	83.4%	

* No Overfitting *

The Solution - Model Performance



* No Overfitting *

Spotify

The Solution - Feature Importances

	Feature	2021	2020	2019	2018	2017	SUM	AVERAGE
0	artistname_indexed	0.123817	0.075257	0.107869	0.091307	0.118017	0.516267	0.103253
1	featuredartist_indexed	0.059671	0.111273	0.072024	0.096596	0.120896	0.460461	0.092092
2	primary_genre_indexed	0.093213	0.069956	0.080219	0.081358	0.067475	0.372221	0.074444
3	scaled_tempo	0.045616	0.065078	0.050962	0.065487	0.054648	0.281791	0.056358
4	loudness	0.047695	0.080769	0.056587	0.042986	0.048175	0.276212	0.055242
5	liveness	0.050288	0.053624	0.071057	0.043948	0.041774	0.260691	0.052138
6	scaled_duration	0.048711	0.043295	0.041363	0.073865	0.050354	0.257587	0.051517
7	valence	0.048183	0.036589	0.046523	0.050744	0.074429	0.256469	0.051294
8	energy	0.053490	0.056037	0.040528	0.047522	0.049571	0.247148	0.049430
9	acousticness	0.072903	0.048430	0.037916	0.041026	0.043189	0.243463	0.048693
10	danceability	0.049102	0.052344	0.055830	0.043539	0.042011	0.242826	0.048565
11	speechiness	0.053945	0.051496	0.045922	0.041756	0.044902	0.238021	0.047604
12	secondary_genre_indexed	0.044220	0.023136	0.053115	0.059886	0.020453	0.200811	0.040162
13	key_indexed	0.030143	0.027351	0.046301	0.033584	0.025448	0.162827	0.032565
14	genre3_indexed	0.025539	0.022880	0.026544	0.039013	0.028611	0.142587	0.028517
15	genre4_indexed	0.027885	0.041189	0.023399	0.021424	0.017283	0.131180	0.026236

15	genre4_indexed	0.027885	0.041189	0.023399	0.021424	0.017283	0.131180	0.026236
16	featuredartist2_indexed	0.020499	0.031430	0.018991	0.026663	0.031119	0.128703	0.025741
17	instrumentalness	0.021463	0.023071	0.015027	0.040586	0.021797	0.121942	0.024388
18	genre5_indexed	0.011635	0.020489	0.018614	0.020604	0.021145	0.092488	0.018498
19	genre6_indexed	0.005597	0.014967	0.030040	0.011845	0.015825	0.078274	0.015655
20	explicit	0.022303	0.005914	0.020234	0.006554	0.003462	0.058468	0.011694
21	albumtype_indexed	0.004896	0.006503	0.012298	0.008748	0.013093	0.045539	0.009108
22	genre7_indexed	0.009225	0.002869	0.005202	0.007024	0.012304	0.036625	0.007325
23	featuredartist3_indexed	0.003279	0.016846	0.006238	0.002096	0.005356	0.033814	0.006763
24	genre8_indexed	0.004627	0.002218	0.003002	0.009360	0.004880	0.024088	0.004818
25	multigenre	0.009968	0.004491	0.003322	0.002411	0.003132	0.023323	0.004665
26	mode	0.004323	0.004095	0.004686	0.004291	0.003045	0.020439	0.004088
27	has_feature	0.004710	0.003378	0.002935	0.002884	0.006138	0.020045	0.004009
28	timesig_indexed	0.001998	0.004784	0.001997	0.000882	0.001389	0.011030	0.002206
29	genre9_indexed	0.000341	0.000119	0.000985	0.000582	0.008992	0.011020	0.002204
30	genre10_indexed	0.000271	0.000085	0.000230	0.001347	0.001069	0.003002	0.000600
31	genre11_indexed	0.000444	0.000037	0.000038	0.000101	0.000018	0.000638	0.000128
32	genre12_indexed	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Primary (Top 3) Feature Importances:

- **Artist Name (Average: 10.3%)**
- **Featured Artist (Average: 9.2%)**
- **Primary Genre (Average: 7.4%)**

Spotify®

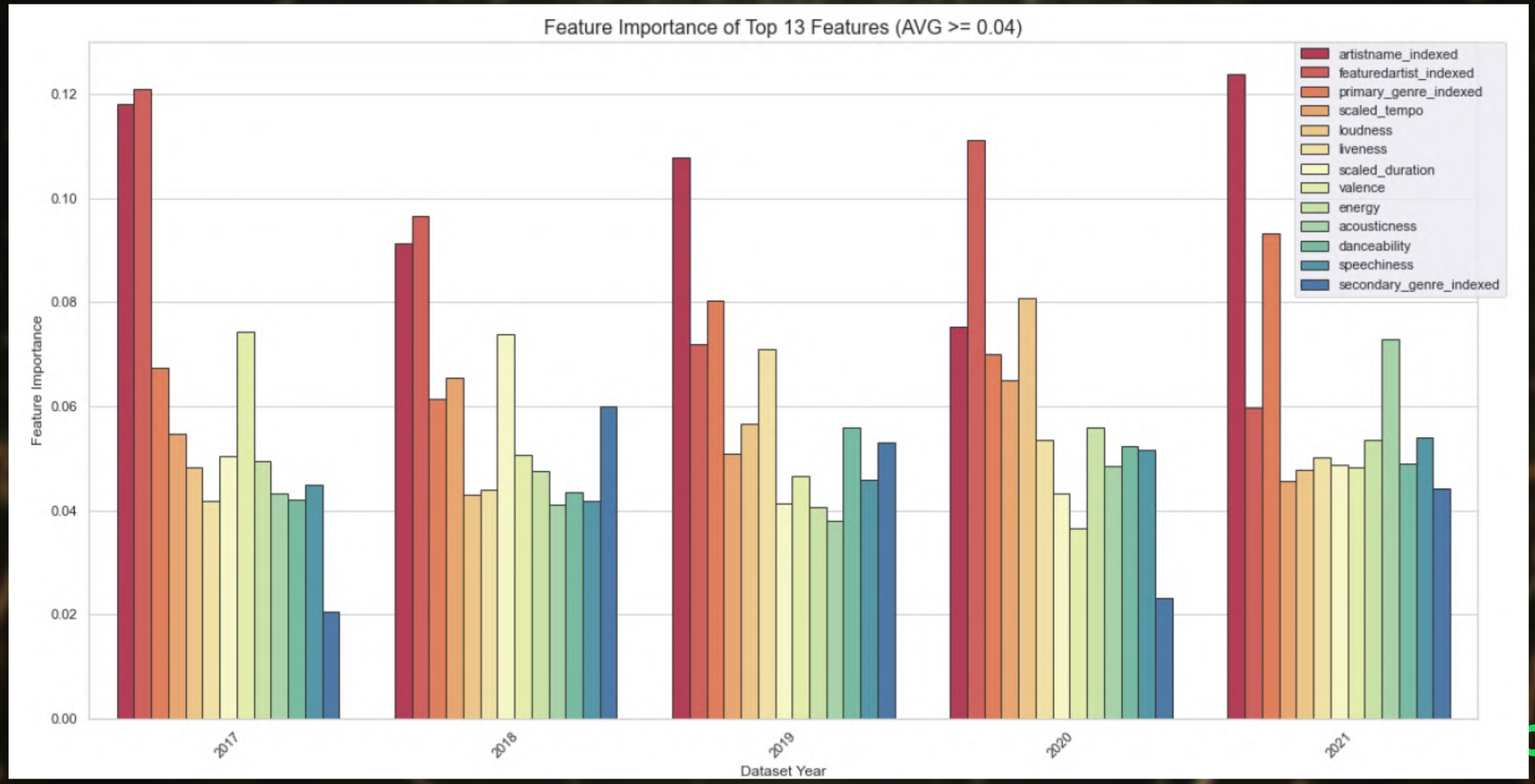
The Dataset - Feature Selection

Danceability	Energy	Loudness	Mode	Speechiness	Acousticness	Instrumentalness
Liveness	Valence	Tempo	Duration	Explicit	Has Feature	Multi-Genre
Key	Time Signature	Artist Name	Album Type	Featured Artist	Featured Artist 2	Featured Artist 3
Primary Genre	Secondary Genre	Genre 3	Genre 4	Genre 5	Genre 6	Genre 7
Genre 8	Genre 9	Genre 10	Genre 11	Genre 12		

Keep

Eliminate

The Solution - Top 13 Feature Importances Visualized



The Solution - Primary Feature Importances - 2017

artistname

VALUES:	25,686 (100%)		18,078 (100%)			
MISSING:	---		---			
DISTINCT:	164	(<1%)	272	(2%)		
1,525	6%	856	5%	Ed Sheeran		
862	3%	537	3%	Kendrick Lamar		
835	3%	627	3%	Drake		
613	2%	34	<1%	Luis Fonsi		
610	2%	149	<1%	Imagine Dragons		
575	2%	241	1%	The Chainsmokers		
501	2%	518	3%	Migos		
20,165	79%	15,116	84%	(Other)		

1,525	6%	856	5%	Ed Sheeran
862	3%	537	3%	Kendrick Lamar
835	3%	627	3%	Drake
613	2%	34	<1%	Luis Fonsi
610	2%	149	<1%	Imagine Dragons
575	2%	241	1%	The Chainsmokers
501	2%	518	3%	Migos
494	2%	146	<1%	Calvin Harris
474	2%	162	<1%	Kygo
466	2%	349	2%	Martin Garrix
457	2%	61	<1%	Khalid
444	2%	59	<1%	J Balvin

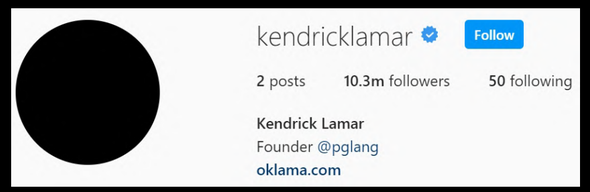
184	<1%	50	<1%	ZAYN
183	<1%	143	<1%	Cheat Codes
182	<1%	32	<1%	Enrique Iglesias
178	<1%	32	<1%	Demi Lovato
177	<1%	34	<1%	Post Malone
5,812	23%	8,702	48%	(Other)

featuredartist

VALUES:	25,686 (100%)		18,078 (100%)			
MISSING:	---		---			
DISTINCT:	135	(<1%)	220	(1%)		
12,288	48%	9,164	51%	NABlankValue		
644	3%	129	<1%	Ozuna		
614	2%	215	1%	Daddy Yankee		
459	2%	91	<1%	Alessia Cara		
422	2%	289	2%	Nicki Minaj		
395	2%	153	<1%	Rihanna		
367	1%	78	<1%	Justin Bieber		
10,497	41%	7,959	44%	(Other)		

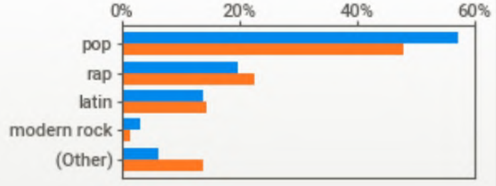
12,288	48%	9,164	51%	NABlankValue
644	3%	129	<1%	Ozuna
614	2%	215	1%	Daddy Yankee
459	2%	91	<1%	Alessia Cara
422	2%	289	2%	Nicki Minaj
395	2%	153	<1%	Rihanna
367	1%	78	<1%	Justin Bieber
366	1%	135	<1%	J Balvin
344	1%	115	<1%	Quavo
315	1%	304	2%	Gucci Mane
306	1%	2	<1%	Coldplay
268	1%	77	<1%	Selena Gomez

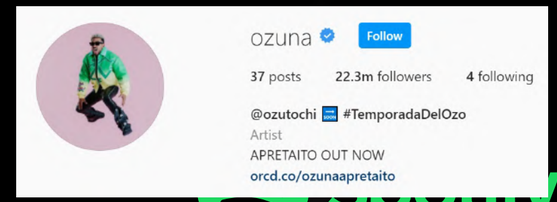
80	<1%	7	<1%	Nicky Jam
80	<1%	56	<1%	XXXTENTACION
80	<1%	58	<1%	Camila Cabello
79	<1%	45	<1%	Seeb
77	<1%	47	<1%	Oh Wonder
72	<1%	21	<1%	Julia Michaels
1,519	6%	3,997	22%	(Other)



primary_genre

VALUES:	25,686 (100%)	18,078 (100%)	
MISSING:	---	---	
DISTINCT:	31 (<1%)	48 (<1%)	





The Solution - Primary Feature Importances - 2018

artistname

VALUES:	24,916 (100%)		17,613 (100%)			
MISSING:	---		---			
DISTINCT:	195	(<1%)	321	(2%)		
	1,464	6%	415	2%	Post Malone	
	1,312	5%	382	2%	Drake	
	1,092	4%	528	3%	XXXTENTACION	
	701	3%	102	<1%	Ariana Grande	
	520	2%	235	1%	Camila Cabello	
	508	2%	255	1%	Khalid	
	468	2%	262	1%	Juice WRLD	
	18,851	76%	15,434	88%	(Other)	

1,464	6%	415	2%	Post Malone
1,312	5%	382	2%	Drake
1,092	4%	528	3%	XXXTENTACION
701	3%	102	<1%	Ariana Grande
520	2%	235	1%	Camila Cabello
508	2%	255	1%	Khalid
468	2%	262	1%	Juice WRLD
444	2%	237	1%	Cardi B
433	2%	270	2%	6ix9ine
411	2%	353	2%	Travis Scott
398	2%	104	<1%	Marshmello
387	2%	28	<1%	Calvin Harris

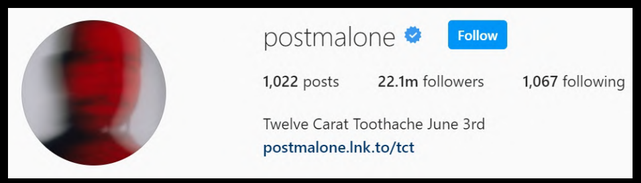
150	<1%	149	<1%	Logic
150	<1%	114	<1%	Martin Garrix
140	<1%	68	<1%	Lady Gaga
139	<1%	102	<1%	George Ezra
5,767	23%	8,769	50%	(Other)

featuredartist

VALUES:	24,916 (100%)		17,613 (100%)			
MISSING:	---		---			
DISTINCT:	174	(<1%)	276	(2%)		
	10,836	43%	8,362	47%	NABlankValue	
	717	3%	330	2%	Drake	
	448	2%	183	1%	Bad Bunny	
	393	2%	184	1%	Nicki Minaj	
	385	2%	81	<1%	21 Savage	
	371	1%	98	<1%	Khalid	
	355	1%	204	1%	J Balvin	
	11,411	46%	8,171	46%	(Other)	

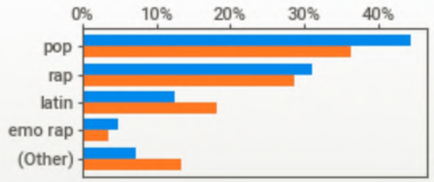
10,836	43%	8,362	47%	NABlankValue
717	3%	330	2%	Drake
448	2%	183	1%	Bad Bunny
393	2%	184	1%	Nicki Minaj
385	2%	81	<1%	21 Savage
371	1%	98	<1%	Khalid
355	1%	204	1%	J Balvin
348	1%	187	1%	Natti Natasha
333	1%	137	<1%	Anne-Marie
333	1%	51	<1%	Dua Lipa
300	1%	54	<1%	Young Thug

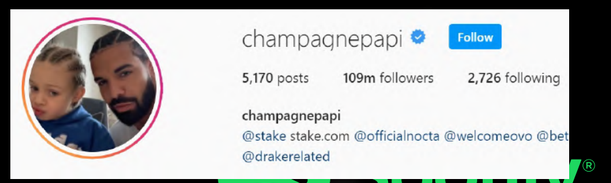
93	<1%	41	<1%	Justin Bieber
90	<1%	69	<1%	Luis Figueroa
79	<1%	127	<1%	Lil Uzi Vert
78	<1%	42	<1%	Joyner Lucas
77	<1%	19	<1%	Jeon
2,073	8%	4,541	26%	(Other)



primary_genre

VALUES:	24,916 (100%)	17,613 (100%)	
MISSING:	---	---	
DISTINCT:	30 (<1%)	47 (<1%)	





The Solution - Primary Feature Importances - 2019

artistname

VALUES:	23,555 (100%)		15,713 (100%)	
MISSING:	---		---	
DISTINCT:	235 (<1%)		353 (2%)	

1,285	5%	781	5%	Billie Eilish
1,038	4%	354	2%	Ariana Grande
957	4%	272	2%	Post Malone
800	3%	157	<1%	Ed Sheeran
675	3%	171	1%	Lil Nas X
542	2%	288	2%	Taylor Swift
443	2%	74	<1%	Sam Smith
17,815	76%	13,616	87%	(Other)

1,285	5%	781	5%	Billie Eilish
1,038	4%	354	2%	Ariana Grande
957	4%	272	2%	Post Malone
800	3%	157	<1%	Ed Sheeran
675	3%	171	1%	Lil Nas X
542	2%	288	2%	Taylor Swift
443	2%	74	<1%	Sam Smith

425	2%	206	1%	Sech
387	2%	226	1%	Khalid
385	2%	45	<1%	Shawn Mendes
372	2%	136	<1%	Maluma
352	1%	222	1%	Jonas Brothers

141	<1%	67	<1%	Calvin Harris
128	<1%	2	<1%	Tainy
126	<1%	101	<1%	Tyler, The Creator
123	<1%	21	<1%	Camilo
122	<1%	8	<1%	blackbear
5,320	23%	7,582	48%	(Other)

featuredartist

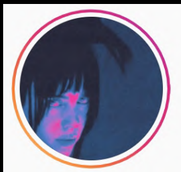
VALUES:	23,555 (100%)		15,713 (100%)	
MISSING:	---		---	
DISTINCT:	192 (<1%)		311 (2%)	

10,761	46%	7,378	47%	NABlankValue
951	4%	352	2%	J Balvin
622	3%	228	1%	Ozuna
493	2%	126	<1%	Daddy Yankee
416	2%	46	<1%	Justin Bieber
406	2%	30	<1%	Camila Cabello
345	1%	117	<1%	Bad Bunny
9,561	41%	7,436	47%	(Other)

10,761	46%	7,378	47%	NABlankValue
951	4%	352	2%	J Balvin
622	3%	228	1%	Ozuna
493	2%	126	<1%	Daddy Yankee
416	2%	46	<1%	Justin Bieber
406	2%	30	<1%	Camila Cabello
345	1%	117	<1%	Bad Bunny

293	1%	77	<1%	Normani
279	1%	5	<1%	Billy Ray Cyrus
260	1%	61	<1%	Disclosure
259	1%	76	<1%	Snow
240	1%	19	<1%	Goodboys
214	<1%	22	<1%	Khalid

83	<1%	40	<1%	Social House
82	<1%	99	<1%	Bebe Rexha
79	<1%	9	<1%	ZAYN
78	<1%	33	<1%	9lokknine
78	<1%	63	<1%	Rita Ora
1,899	8%	4,271	27%	(Other)



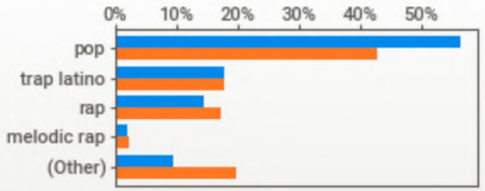
billieeilish Follow

659 posts 102m followers 0 following

BILLIE EILISH
billieeilish.lnk.to/HappierThanEver

primary_genre

VALUES:	23,555 (100%)	15,713 (100%)
MISSING:	---	---
DISTINCT:	40 (<1%)	60 (<1%)



jbalvin Follow

13,001 posts 52m followers 2,035 following

J Balvin
hoo.be/jbalvin

The Solution - Primary Feature Importances - 2020

artistname

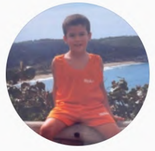
VALUES:	21,169 (100%)		15,144 (100%)		
MISSING:	---		---		
DISTINCT:	217	(1%)	355	(2%)	
1,213	6%	763	5%	Bad Bunny	
796	4%	236	2%	Dua Lipa	
714	3%	379	3%	Juice WRLD	
698	3%	161	1%	Justin Bieber	
666	3%	438	3%	J Balvin	
661	3%	224	1%	The Weeknd	
634	3%	365	2%	Pop Smoke	
483	2%	101	<1%	Ariana Grande	
483	2%	160	1%	Drake	
470	2%	257	2%	Rauw Alejandro	
429	2%	270	2%	Taylor Swift	
15,787	75%	12,578	83%	(Other)	

featuredartist

VALUES:	21,169 (100%)		15,144 (100%)		
MISSING:	---		---		
DISTINCT:	188	(<1%)	300	(2%)	
10,941	52%	7,808	52%	NABlankValue	
498	2%	199	1%	Drake	
372	2%	51	<1%	Roddy Ricch	
350	2%	282	2%	Anuel AA	
312	1%	9	<1%	beabadoobee	
311	1%	261	2%	J Balvin	
285	1%	405	3%	Juice WRLD	
275	1%	77	<1%	Quavo	
261	1%	51	<1%	Justin Bieber	
254	1%	290	2%	Daddy Yankee	
226	1%	66	<1%	Camilo	
216	1%	31	<1%	Jason Derulo	
8,100	38%	6,129	40%	(Other)	

127	<1%	51	<1%	Khalid
125	<1%	96	<1%	Myke Towers
113	<1%	36	<1%	ROSALÍA
111	<1%	57	<1%	NLE Choppa
3,948	19%	6,737	44%	(Other)

63	<1%	-	-	Jhay Cortez
54	<1%	27	<1%	Rihanna
54	<1%	44	<1%	Felix Jaehn
49	<1%	38	<1%	Zara Larsson
49	<1%	56	<1%	Demi Lovato
46	<1%	40	<1%	Lenny Santos
1,253	6%	3,013	20%	(Other)

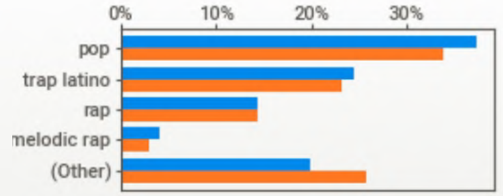


badbunnypr Follow

6 posts 38.8m followers 0 following

primary_genre

VALUES:	21,169 (100%)	15,144 (100%)	
MISSING:	---	---	
DISTINCT:	51 (<1%)	77 (<1%)	



champagnepapi Follow

5,170 posts 109m followers 2,726 following

champagnepapi

@stake stake.com @officialnocta @welcomeovo @bett

@drakerelated

The Solution - Primary Feature Importances - 2021

artistname

VALUES:	19,443 (100%)		11,578 (100%)	
MISSING:	---		---	
DISTINCT:	228	(1%)	353	(3%)

1,742	9%	311	3%	Olivia Rodrigo
972	5%	187	2%	Doja Cat
836	4%	232	2%	Justin Bieber
728	4%	277	2%	Drake
605	3%	54	<1%	Lil Nas X
451	2%	156	1%	Rauw Alejandro
412	2%	187	2%	J Balvin
13,697	70%	10,174	88%	(Other)

1,742	9%	311	3%	Olivia Rodrigo
972	5%	187	2%	Doja Cat
836	4%	232	2%	Justin Bieber
728	4%	277	2%	Drake
605	3%	54	<1%	Lil Nas X
451	2%	156	1%	Rauw Alejandro
412	2%	187	2%	J Balvin
410	2%	187	2%	Billie Eilish
402	2%	64	<1%	Ed Sheeran
383	2%	39	<1%	Bruno Mars
364	2%	109	<1%	BTS
349	2%	42	<1%	The Weeknd

featuredartist

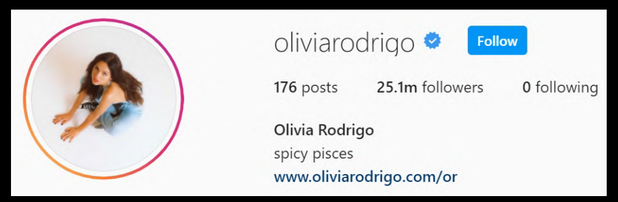
VALUES:	19,443 (100%)		11,578 (100%)	
MISSING:	---		---	
DISTINCT:	197	(1%)	270	(2%)

9,538	49%	5,580	48%	NABlankValue
393	2%	22	<1%	Bad Bunny
383	2%	39	<1%	Anderson .Paak
351	2%	175	2%	Rauw Alejandro
339	2%	95	<1%	Maria Becerra
290	1%	89	<1%	The Weeknd
286	1%	2	<1%	Daniel Caesar
7,863	40%	5,576	48%	(Other)

9,538	49%	5,580	48%	NABlankValue
393	2%	22	<1%	Bad Bunny
383	2%	39	<1%	Anderson .Paak
351	2%	175	2%	Rauw Alejandro
339	2%	95	<1%	Maria Becerra
290	1%	89	<1%	The Weeknd
286	1%	2	<1%	Daniel Caesar
284	1%	27	<1%	Ariana Grande
272	1%	5	<1%	SZA
271	1%	136	1%	Myke Towers
240	1%	141	1%	Lil Baby
237	1%	43	<1%	iann dior

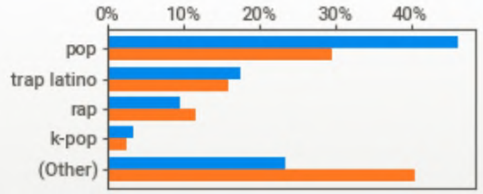
110	<1%	30	<1%	Post Malone
109	<1%	68	<1%	Gera MX
106	<1%	54	<1%	OneRepublic
104	<1%	62	<1%	Marc Seguí
103	<1%	74	<1%	Juice WRLD
3,476	18%	6,410	55%	(Other)

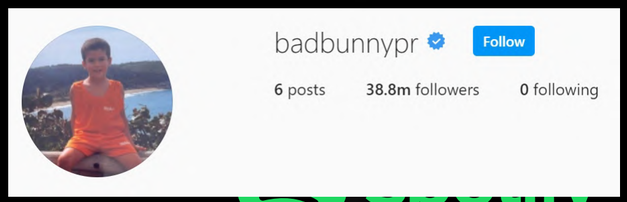
55	<1%	21	<1%	Khalid
55	<1%	5	<1%	Kali Uchis
49	<1%	24	<1%	Chance the Rapper
49	<1%	24	<1%	Travis Barker
49	<1%	17	<1%	Trueno
45	<1%	55	<1%	Ozuna
1,178	6%	3,052	26%	(Other)



primary_genre

VALUES:	19,443 (100%)	11,578 (100%)	
MISSING:	---	---	
DISTINCT:	64 (<1%)	88 (<1%)	





The Solution - Actionable Insights

Artist Name + Featured Artist

- Importance of Branding (I.e. "Who They Are")
- Year-Over-Year Turnover
- Smaller, Lesser-known Artists + Featured Artists are Less Likely to be Top 100

Primary Genre

- Pop Songs Dominate Top 100
- Tastes Change Year-Over-Year (Rap, Trap Latino, etc.)
- Smaller, Niche Genres are Less Likely to be in the Top 100

Changes In Taste

- Relative Feature Importances Rankings Fluctuate Year-Over-Year
- Reflective of Changing Tastes

The Solution - Recommendations for Success

Independent Artists

- Spend Time Defining Your Brand and Growing Your Social Media Presence / Fanbase
- Avoid Overly Niche Genres
- Aim to Secure Features with Well-Known Artists

Artist Managers

- Focus on Building your Artist's Brand and Social Media Presence
- Incorporate "Pop" Into Metadata during Submission
- Secure Features with Well-Known Artists for Your Artist

Spotify

- Mine Social Media Statistics to Identify Strong Upcoming Potential Artists
- Promote Artists with More Popular Genres in Metadata
- Promote Artists with Strong Feature Collaborations

Record Labels

- Mine Social Media Statistics to Identify Strong Upcoming Potential Artists
- Incorporate "Pop" Into Metadata during Submissions
- Organize Features Between Well-Known Artists for Cross-Pollination and Growth

Thank You!

LinkedIn

- https://ca.linkedin.com/in/krishandeo
- Happy to Connect :)

GitHub

- https://github.com/k-deo
- Jupyter Notebooks (.ipynb)
- Cleansed Datasets (.csv)
- Architectural Design Document (.pdf)
- Data Product (.pdf)
  - PDF Reports from Jupyter Notebooks
  - Final Presentation PDF

Data Science Peers Presentation

Krishan Deo

Architectural Choices

Model deployment comes in many shapes. The key to everything is that the business insights that result from the model are made available to stakeholders. This can happen in various ways. At the simplest level a PDF report is generated (e.g. using a jupyter notebook in Watson Studio) and handed over to business stakeholders. Alternatively, the model is encapsulated behind a REST API and made either available to be consumed by a data product or sold internally or externally as a API (e.g. by using IBM Watson Machine Learning or Fabric for DeepLearning).

Depending on your use case, please choose and implement an appropriate model deployment option and justify your decisions in the ADD.

Programming Language: Python 3.7
Engine: PySpark 2.4.5 (SparkML), Tensorflow 2.4.4 (Keras)
Development Environment: Jupyter Notebooks via IBM Watson Studio
Storage: Cloud Object Store via IBM Cloud

Reasoning: Putting into Practice the Learnings from the IBM Advanced Data Science Specialization via Coursera.Note: Python 3.7

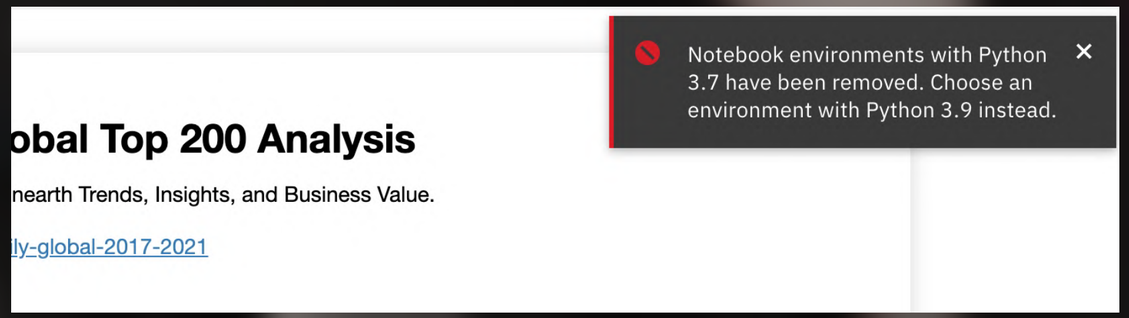
Exploratory Data Analysis Library: SweetViz
Reasoning: SweetViz Generates Descriptive Statistics and Visuals for Labeled Subsets of the Dataset with Very Few Lines of Code

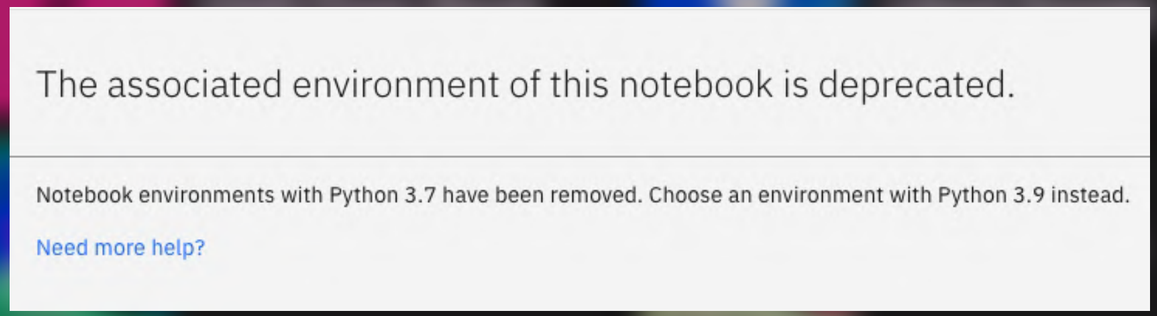
Data Product Choice: PDF Report (Jupyter Notebooks + Slide Deck)
Reasoning: PDF Report is Better Suited to Conveying Actionable Insights and Recommendations for Business Users

Architectural Choices - 2

Note: Python 3.7 was Deprecated on Watson Studio Midway through Project Development

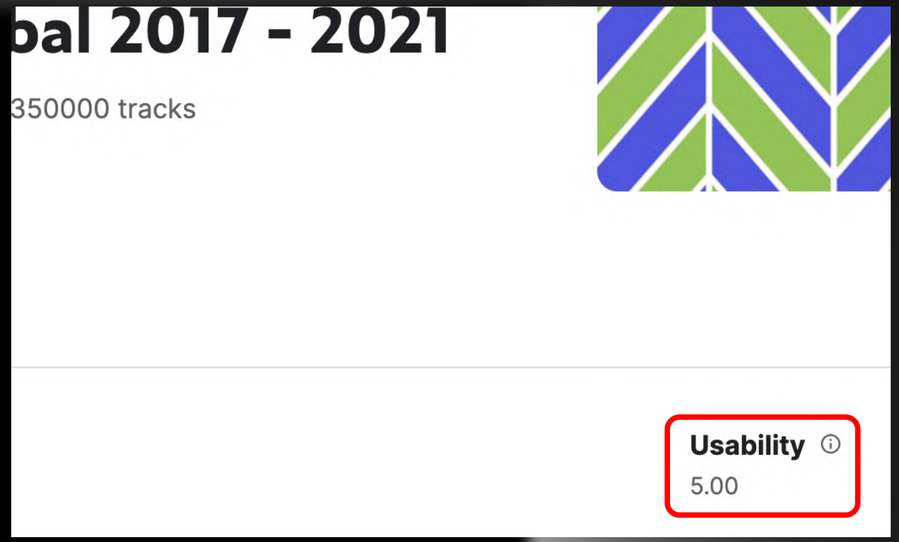
- Broken Compatibility with Python 3.9
- Prevented Planned Hyperparameter Tuning
  - Unable to Spin Up Stronger, Parallelized Environments for Python 3.7

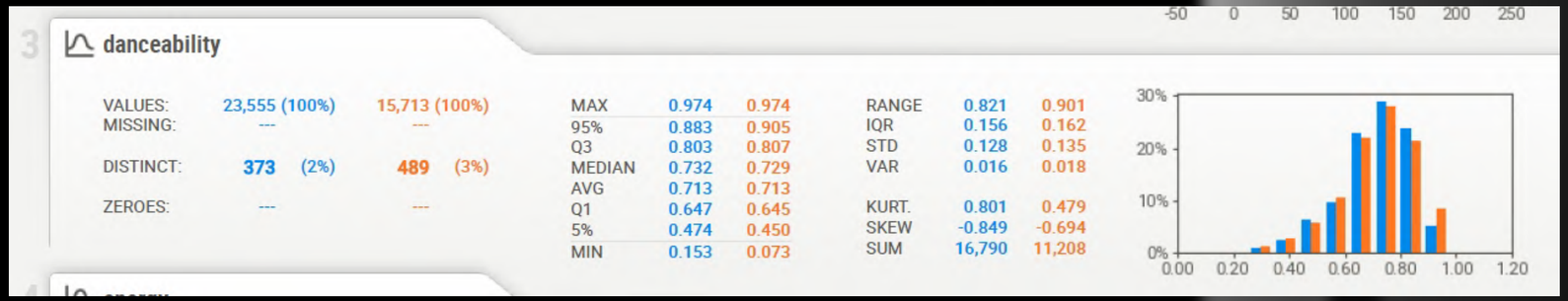




Data Quality Assessment

- **Alphabetically Ordered Genres:**
  - Not Easily Comparable
- **Nested JSON with Non-Uniform 1 to Many Relationships:**
  - Varying # of Genres
  - Varying # of Artists + Featured Artists
- **Unscaled Numerical Data:**
  - Tempo
  - Duration
- **Missing Values:**
  - Retroactively Removed Songs
- **Outliers:**
  - Songs with 4+ Years on the Bottom Half of the Charts
- **Slightly Imbalanced Label Distribution**





Data Quality Assessment

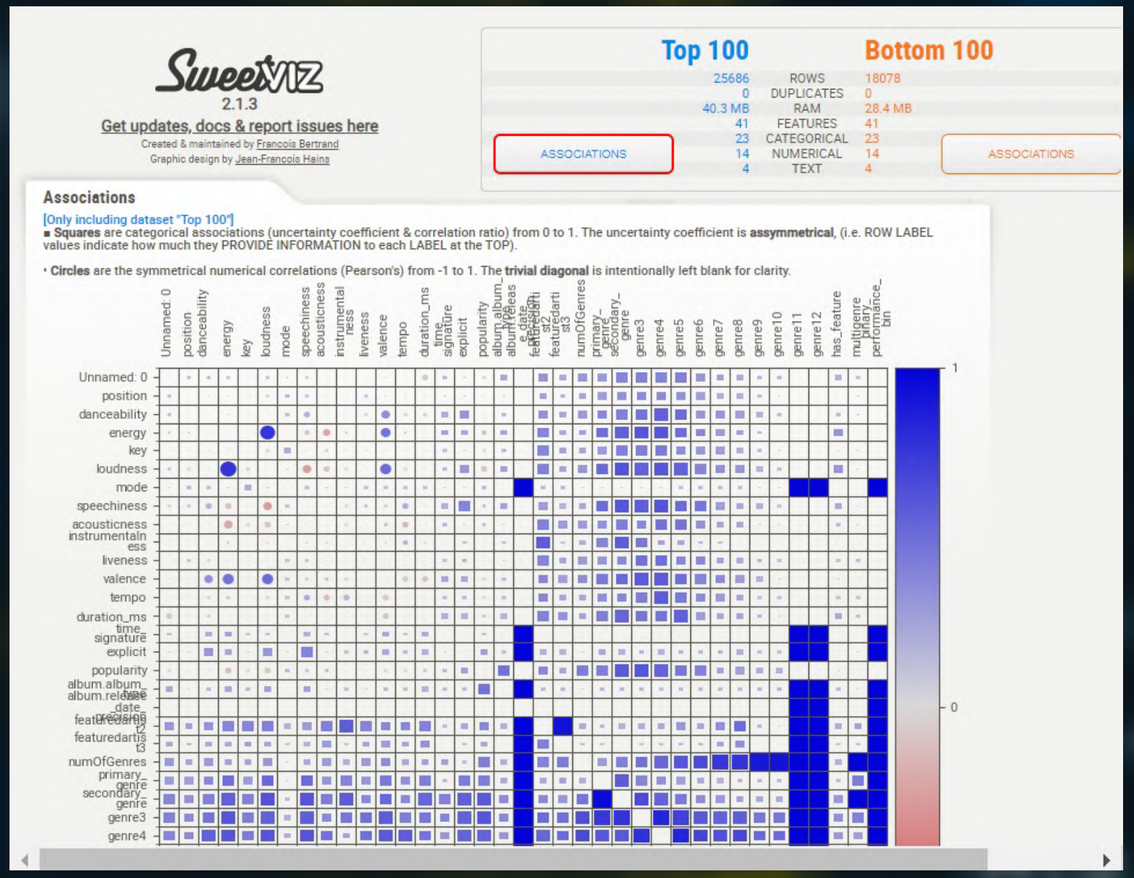
position	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	explicit	name
8	0.928	0.481	9	-9.35	0	0.287	0.105	0	0.176	0.613	134.007	210937	4	TRUE	Fake Love
13	0.852	0.773	8	-2.921	0	0.0776	0.187	3.05E-05	0.159	0.907	102.034	195840	4	FALSE	Chantaje (feat. Maluma)
16	0.49	0.485	4	-6.237	0	0.0406	0.0592	0	0.337	0.196	133.889	195840	4	FALSE	In the Name of Love
17	0.927	0.665	11	-5.313	1	0.244	0.061	0	0.123	0.175	127.076	343150	4	TRUE	Bad and Boujee (feat. Lil Uzi Vert)
21	0.469	0.68	11	-4.921	0	0.117	0.137	0	0.11	0.374	147.734	208733	4	FALSE	Mercy
22	0.697	0.691	2	-4.757	1	0.146	0.214	0	0.185	0.305	137.853	239293	4	FALSE	Bad Things (with Camila Cabello)
25	0.624	0.803	10	-4.105	0	0.246	0.123	0	0.114	0.821	166.018	187973	4	FALSE	Treat You Better
30	0.544	0.809	8	-5.098	1	0.0363	0.0038	0	0.323	0.448	145.017	197640	4	FALSE	All Night
32	0.781	0.57	11	-5.874	0	0.188	0.273	0	0.196	0.858	107.059	193181	4	FALSE	Now and Later
35	0.486	0.713	2	-3.949	0	0.0524	0.0853	0	0.0839	0.297	121.028	226720	4	FALSE	I Would Like
47	0.952	0.318	10	-10.357	1	0.467	0.174	0	0.205	0.665	120.077	209640	4	TRUE	Caroline
52	0.739	0.833	1	-5.012	1	0.0463	0.35	0	0.266	0.699	117.99	195313	4	FALSE	Perfect Strangers
57	0.78	0.575	1	-5.628	0	0.139	0.106	0	0.129	0.273	81.502	222360	4	TRUE	Bounce Back
59	0.602	0.707	9	-4.097	1	0.302	0.393	0	0.165	0.554	75.087	200186	4	FALSE	Human
70	0.422	0.852	6	-3.546	1	0.208	0.0845	0	0.0687	0.666	82.914	248386	4	TRUE	Lot to Learn

numOfGenres	primary_genre	secondary_genre	genre3	genre4	genre5	genre6
5	rap	hip hop	canadian pop	toronto rap	canadian hip hop	NABlankValue
5	pop	dance pop	latin	latin pop	colombian pop	NABlankValue
7	pop	dance pop	edm	pop dance	tropical house	progressive house
5	rap	pop rap	hip hop	trap	atl hip hop	NABlankValue
4	pop	dance pop	canadian pop	viral pop	NABlankValue	NABlankValue
2	pop rap	ohio hip hop	NABlankValue	NABlankValue	NABlankValue	NABlankValue
4	pop	dance pop	canadian pop	viral pop	NABlankValue	NABlankValue
4	pop	dance pop	post-teen pop	boy band	NABlankValue	NABlankValue
6	dance pop	rap	pop rap	trap	southern hip hop	hyphy
10	pop	dance pop	edm	pop dance	tropical house	post-teen pop
5	pop	rap	hip hop	underground hip hop	portland hip hop	NABlankValue
6	pop	dance pop	edm	pop dance	tropical house	uk dance
7	pop	rap	pop rap	hip hop	trap	southern hip hop
1	neo soul	NABlankValue	NABlankValue	NABlankValue	NABlankValue	NABlankValue
2	pop rap	indie pop rap	NABlankValue	NABlankValue	NABlankValue	NABlankValue

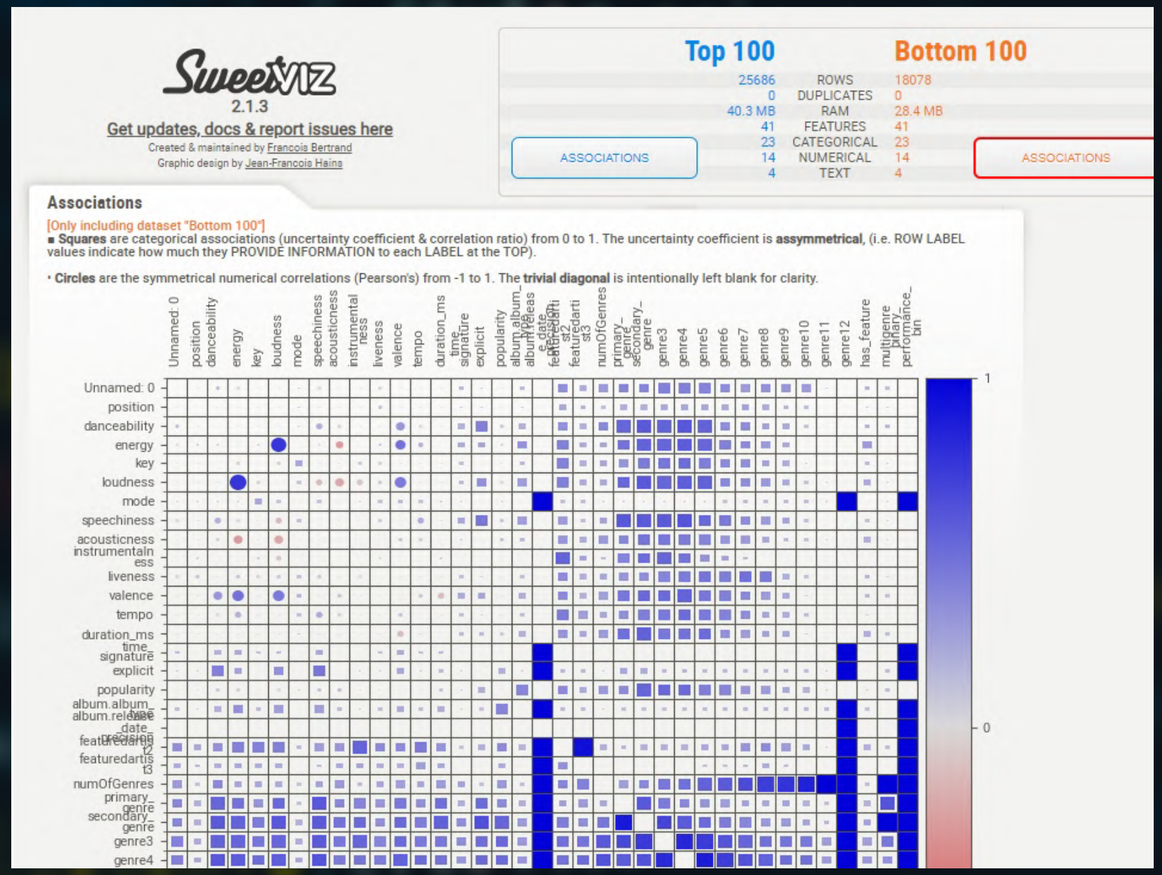
artistname	featuredartist
Drake	NABlankValue
Shakira	Maluma
Martin Garrix	Bebe Rexha
Migos	Lil Uzi Vert
Shawn Mendes	NABlankValue
Machine Gun Kelly	Camila Cabello
Shawn Mendes	NABlankValue
The Vamps	Matoma
Sage The Gemini	NABlankValue
Zara Larsson	NABlankValue
AminÃ©	NABlankValue
Jonas Blue	JP Cooper
Big Sean	NABlankValue
Rag'n'Bone Man	NABlankValue
Luke Christopher	NABlankValue

genre11	genre12	has_feature	multigenre	binary_performance_bin
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	TRUE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100
NABlankValue	NABlankValue	FALSE	FALSE	Top 100
NABlankValue	NABlankValue	FALSE	TRUE	Top 100

Data Quality Assessment

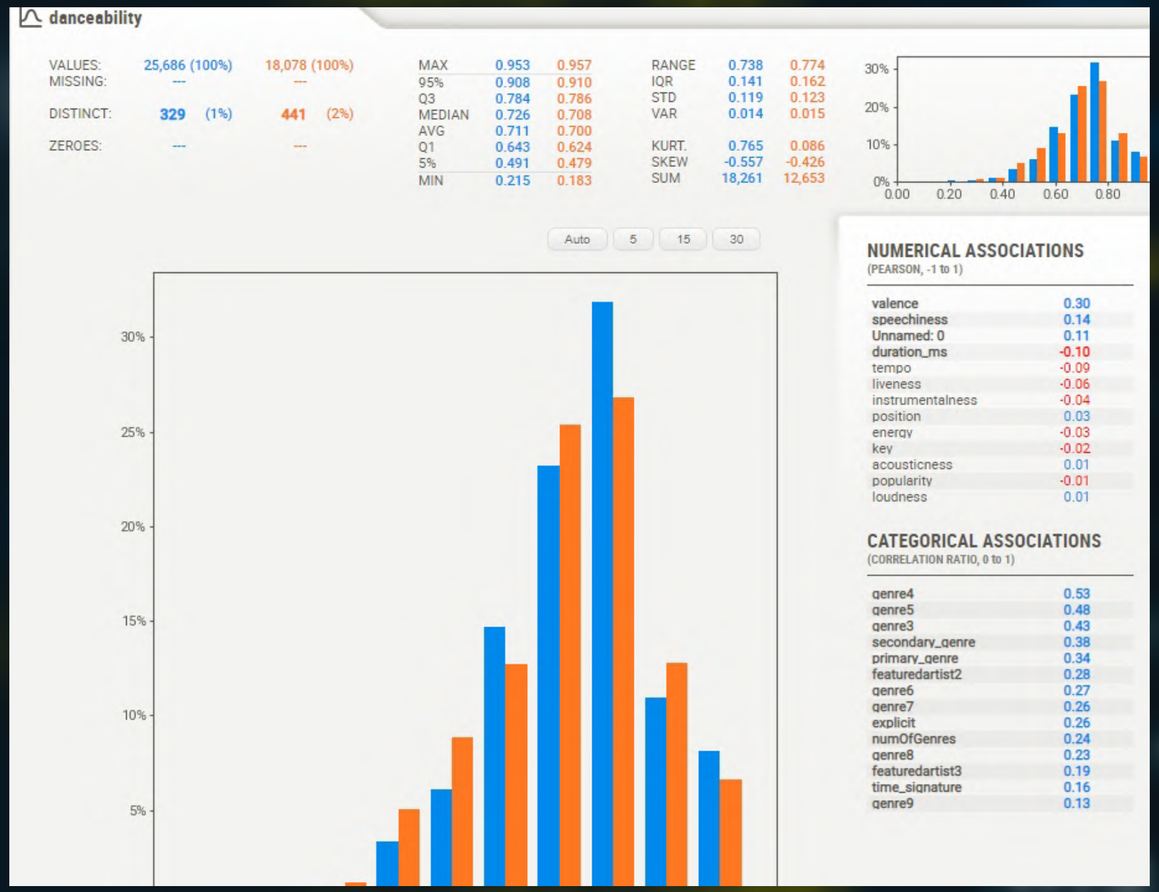


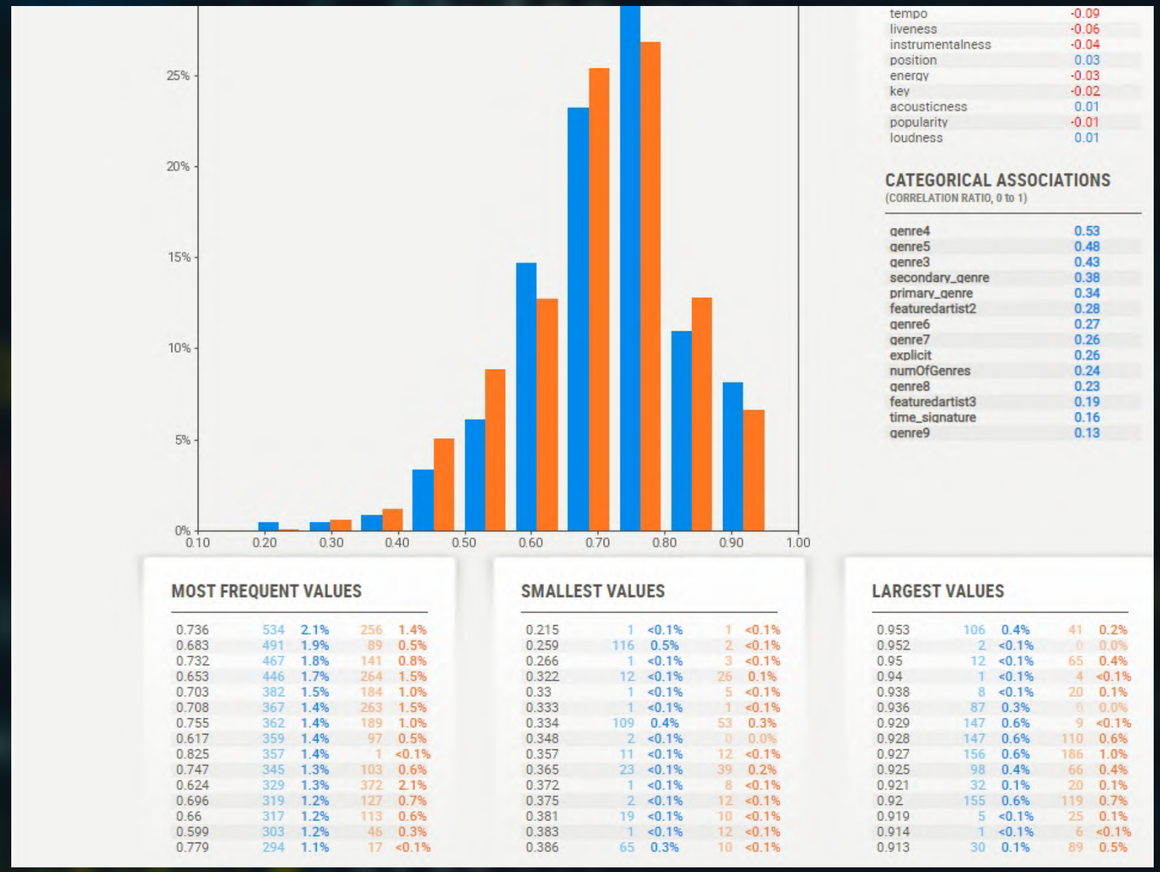
Top 100



Bottom 100

Data Quality Assessment





Data Pre-Processing and Feature Engineering

- **Custom ETL Script Built to Extract and Normalize Nested JSON Data**
- **Top 100 / Bottom 100 Label Created**
- **Datasets were Limited to Songs Released Within Calendar Year**
  - Avoided Penalization of Songs with Notable Longevity
  - Allowed Year-to-Year Idiosyncratic Taste Changes to be Captured
- **Genres were Extracted and Sorted based on Popularity within the Calendar Year**
  - Genre Information was Previously Useless due to Alphabetical Sorting
  - Improved Performance from 70s to 80s
- **Records with Key Missing Values were Removed**
  - Ex: Artist Name, Primary Genre, etc.
- **Optional Missing Values were Imputed as "NABlankValue" to Avoid Errors in Model Training and Evaluation**
  - Ex: Featured Artist, Secondary Genre, etc.
- **"Multi-Genre" and "Number of Genres" Features were Engineered**

Step 1: Custom ETL Script to Create 2017, 2018, 2019, 2020, and 2021 Datasets

In [ ]:

```
# Imports

import pandas as pd
import glob, os, json
import numpy as np

pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth',1000)

# Nested JSON Parsing and Concatenation

json_dir = 'C:/Users/krishand/Downloads/SPOTIFY/archive/'

json_pattern = os.path.join(json_dir, '*.json')
file_list = glob.glob(json_pattern)

dfs = []
for file in file_list:
    with open(file, 'r', encoding='utf-8') as f:
        json_data = pd.json_normalize(json.loads(f.read()))
        json_data['site'] = file.rsplit("/", 1)[-1]
    dfs.append(json_data)
df = pd.concat(dfs)

df = df.reset_index()
df['index'] = df['index']+1
df = df.rename(columns={'index': 'position'})

# Extracting Artist Names

df2 = pd.json_normalize(df['artists'])[0].to_frame(name="js")
df3 = pd.json_normalize(df2['js'])['name']

df_2 = pd.json_normalize(df['artists'])[1].to_frame(name="js2")
df_3 = pd.json_normalize(df_2['js2'])['name']

df_2a = pd.json_normalize(df['artists'])[2].to_frame(name="js3")
df_3a = pd.json_normalize(df_2a['js3'])['name']
```

ETL + EDA + Data Quality + Feature Engineering Jupyter Notebook

Model Performance Indicators

Label Type: Binary (Top 100 / Bottom 100)

Performance Indicator Choice: AUROC

Reasons for Choice:

- Measures Ability to Discriminate Classes
- Handles Imbalanced Datasets

Key Insight Driver Choice: Feature Importance

Reasons for Choice:

- Uncovers Relative Feature Hierarchy + Idiosyncracies within Each Year
- Informs Feature Selection for Future Project Iterations

Dataset Imbalance

	Top 100	Top 100 %	Bottom 100	Bottom 100 %	Total
2017	25,686	58.7%	18,078	41.3%	43,764
2018	24,916	58.6%	17,613	41.4%	42,529
2019	23,555	60%	15,713	40%	39,268
2020	21,169	58.2%	15,144	41.7%	36,313
2021	19,443	62.7%	11,578	37.3%	31,021

Model Algorithms

Initial Baseline - SparkML Based Support Vector Machine

- Strictly Numerical Features
- Categorical Features Excluded to Prevent Worsened Performance and Inefficient Training due to One Hot Encoding (High-Cardinality)

Final Model - SparkML Based Random Forest

- SparkML's RF Implementation is able to handle High-Cardinality, Categorical Features without One Hot Encoding via Spark's Metadata Management

Deep Learning Model - Neural Network (Keras) Based Binary Perceptron

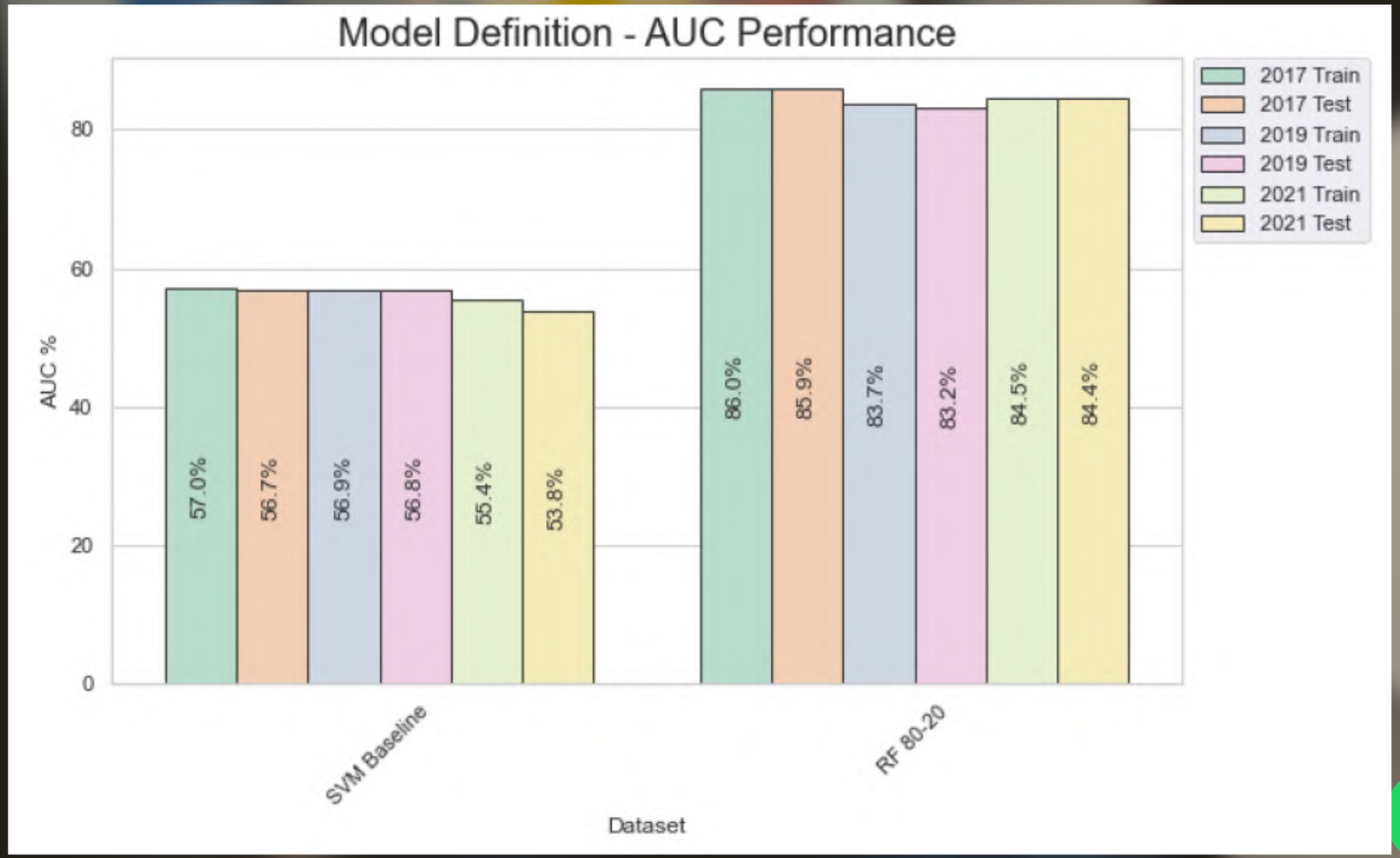
- 2 Dropouts for Regularization
- "ReLU" Activation Function for Hidden Layers
- "Sigmoid" Activation Function for Output Layer
- "Adam" as Optimizer
- "Categorical Crossentropy" as Loss Function

```
# Neural Network Structures

model1 = Sequential()  # Instantiate Sequential Model
model1.add(Dense(28, input_dim=14, activation='relu'))
model1.add(Dense(56, activation='relu'))
model1.add(Dropout(0.1))
model1.add(Dense(112, activation='relu'))
model1.add(Dropout(0.2))
model1.add(Dense(num_classes1, activation='sigmoid'))
model1.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Model Definition Jupyter Notebook

Model Definition Stage - Performance Visualization



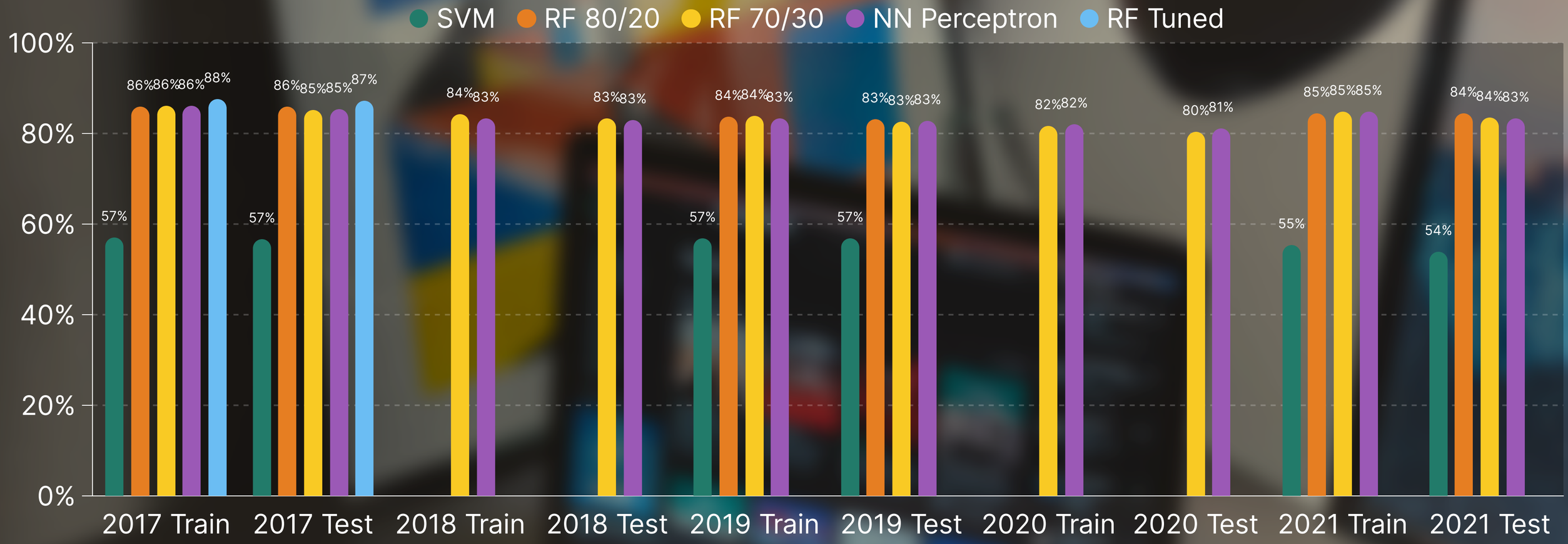
Model Training + Evaluation Jupyter Notebook

Model Performance

	SVM Baseline	RF 80/20	RF 70/30	NN Perceptron	RF Tuned
2017 Train	57.0%	86.0%	86.1%	86.2%	87.6%
2017 Test	56.7%	85.9%	85.2%	85.4%	87.2%
2018 Train			84.3%	83.3%	
2018 Test			83.4%	83.0%	
2019 Train	56.9%	83.7%	83.9%	83.3%	
2019 Test	56.8%	83.2%	82.5%	82.7%	
2020 Train			81.7%	82.0%	
2020 Test			80.4%	81.1%	
2021 Train	55.4%	84.5%	84.8%	84.8%	
2021 Test	53.8%	84.4%	83.5%	83.4%	

* No Overfitting *

Overall Model Performance - Visualization



* No Overfitting *

Feature Importances

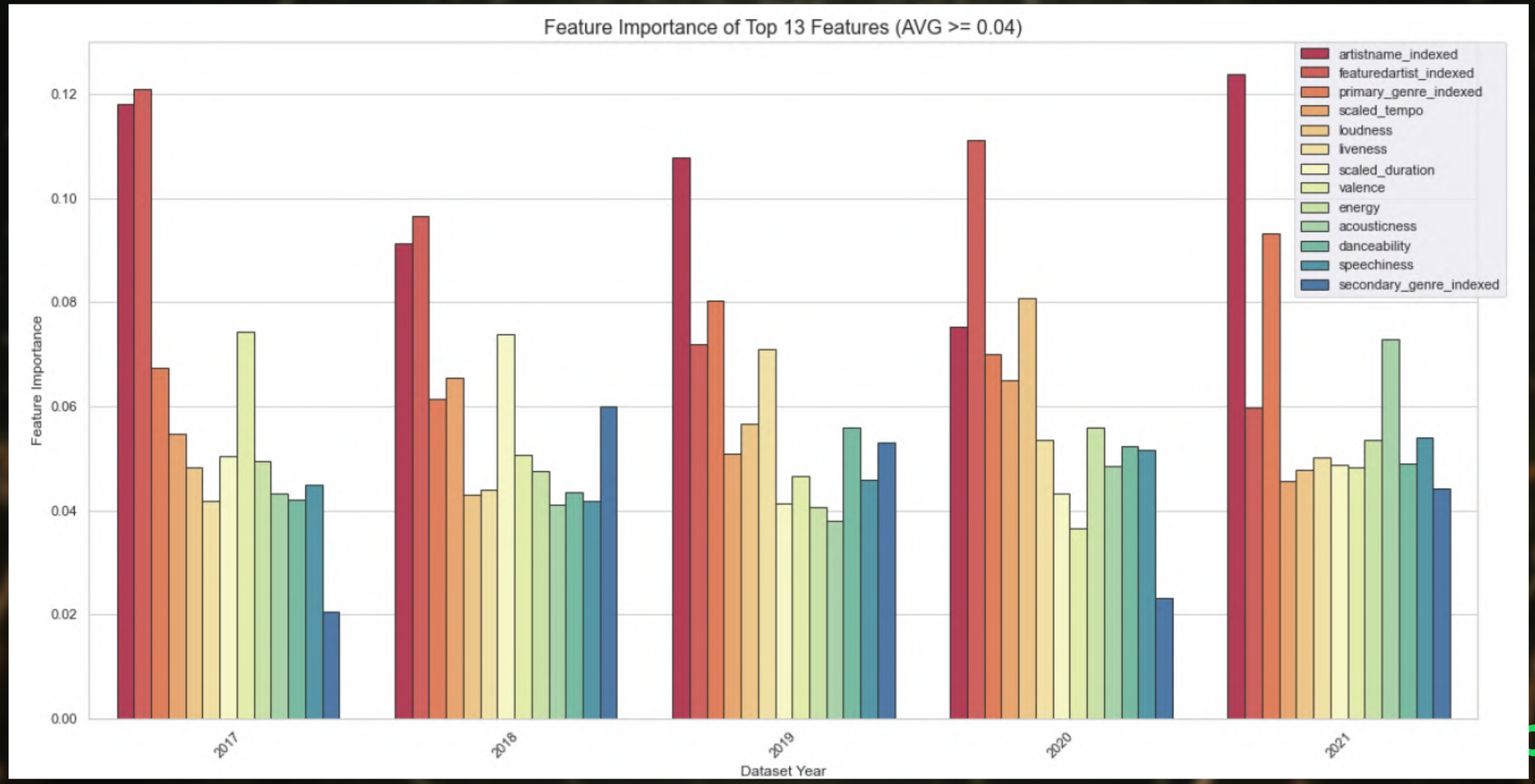
	Feature	2021	2020	2019	2018	2017	SUM	AVERAGE
0	artistname_indexed	0.123817	0.075257	0.107869	0.091307	0.118017	0.516267	0.103253
1	featuredartist_indexed	0.059671	0.111273	0.072024	0.096596	0.120896	0.460461	0.092092
2	primary_genre_indexed	0.093213	0.069956	0.080219	0.081358	0.067475	0.372221	0.074444
3	scaled_tempo	0.045616	0.065078	0.050962	0.065487	0.054648	0.281791	0.056358
4	loudness	0.047695	0.080769	0.056587	0.042986	0.048175	0.276212	0.055242
5	liveness	0.050288	0.053624	0.071057	0.043948	0.041774	0.260691	0.052138
6	scaled_duration	0.048711	0.043295	0.041363	0.073865	0.050354	0.257587	0.051517
7	valence	0.048183	0.036589	0.046523	0.050744	0.074429	0.256469	0.051294
8	energy	0.053490	0.056037	0.040528	0.047522	0.049571	0.247148	0.049430
9	acousticness	0.072903	0.048430	0.037916	0.041026	0.043189	0.243463	0.048693
10	danceability	0.049102	0.052344	0.055830	0.043539	0.042011	0.242826	0.048565
11	speechiness	0.053945	0.051496	0.045922	0.041756	0.044902	0.238021	0.047604
12	secondary_genre_indexed	0.044220	0.023136	0.053115	0.059886	0.020453	0.200811	0.040162
13	key_indexed	0.030143	0.027351	0.046301	0.033584	0.025448	0.162827	0.032565
14	genre3_indexed	0.025539	0.022880	0.026544	0.039013	0.028611	0.142587	0.028517
15	genre4_indexed	0.027885	0.041189	0.023399	0.021424	0.017283	0.131180	0.026236
16	featuredartist2_indexed	0.020499	0.031430	0.018991	0.026663	0.031119	0.128703	0.025741
17	instrumentalness	0.021463	0.023071	0.015027	0.040586	0.021797	0.121942	0.024388
18	genre5_indexed	0.011635	0.020489	0.018614	0.020604	0.021145	0.092488	0.018498
19	genre6_indexed	0.005597	0.014967	0.030040	0.011845	0.015825	0.078274	0.015655
20	explicit	0.022303	0.005914	0.020234	0.006554	0.003462	0.058468	0.011694
21	albumtype_indexed	0.004896	0.006503	0.012298	0.008748	0.013093	0.045539	0.009108
22	genre7_indexed	0.009225	0.002869	0.005202	0.007024	0.012304	0.036625	0.007325
23	featuredartist3_indexed	0.003279	0.016846	0.006238	0.002096	0.005356	0.033814	0.006763
24	genre8_indexed	0.004627	0.002218	0.003002	0.009360	0.004880	0.024088	0.004818
25	multigenre	0.009968	0.004491	0.003322	0.002411	0.003132	0.023323	0.004665
26	mode	0.004323	0.004095	0.004686	0.004291	0.003045	0.020439	0.004088
27	has_feature	0.004710	0.003378	0.002935	0.002884	0.006138	0.020045	0.004009
28	timesig_indexed	0.001998	0.004784	0.001997	0.000882	0.001389	0.011030	0.002206
29	genre9_indexed	0.000341	0.000119	0.000985	0.000582	0.008992	0.011020	0.002204
30	genre10_indexed	0.000271	0.000085	0.000230	0.001347	0.001069	0.003002	0.000600
31	genre11_indexed	0.000444	0.000037	0.000038	0.000101	0.000018	0.000638	0.000128
32	genre12_indexed	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Primary (Top 3) Feature Importances:

- **Artist Name (Average: 10.3%)**
- **Featured Artist (Average: 9.2%)**
- **Primary Genre (Average: 7.4%)**

Spotify®

Feature Importances



Next Steps / Future Improvements

- Refactor Code for Python 3.9+
- Re-Train, Re-Tune, and Re-Evaluate Random Forest Models via Stronger, Parallelized Environment on IBM Watson Studio
  - Final 13 Features
- Using Embedding to Incorporate High-Cardinality Features into NN Perceptron Models
  - Final 13 Features

Thank You!

LinkedIn

- https://ca.linkedin.com/in/krishandeo
- Happy to Connect :)

GitHub

- https://github.com/k-deo
- Jupyter Notebooks (.ipynb)
- Cleansed Datasets (.csv)
- Architectural Decisions Document (.pdf)
- Data Product (.pdf)
  - PDF Reports from Jupyter Notebooks
  - Final Presentation PDF

Spotify®