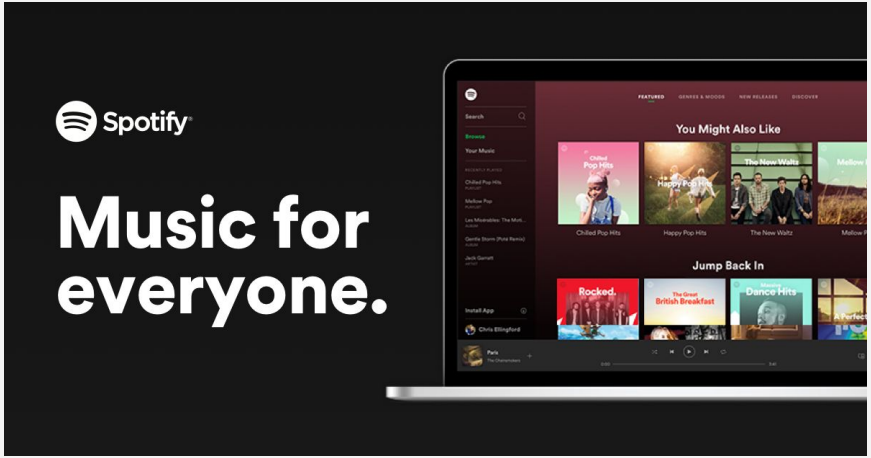
MSBA 212 - Group #9

Chengwu Weng
Mira Daya
Misha Khan
Yue Fang

Overview

Background

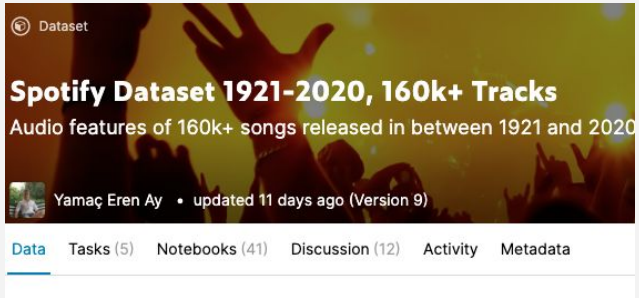
Spotify is the #1 competitor in the streaming service industry.



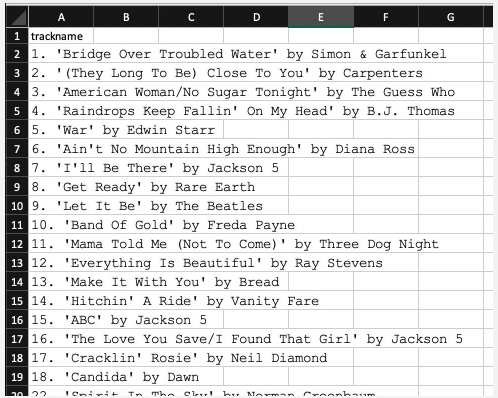
Objective

Our goal is to predict which 2020 songs will be popular based on past Billboard Hits and Spotify popularity ratings.

Data Acquisition



Kaggle.com



	A	B	C	D	E	F	G
1	trackname						
2	1. 'Bridge Over Troubled Water' by Simon & Garfunkel						
3	2. '(They Long To Be) Close To You' by Carpenters						
4	3. 'American Woman/No Sugar Tonight' by The Guess Who						
5	4. 'Raindrops Keep Fallin' On My Head' by B.J. Thomas						
6	5. 'War' by Edwin Starr						
7	6. 'Ain't No Mountain High Enough' by Diana Ross						
8	7. 'I'll Be There' by Jackson 5						
9	8. 'Get Ready' by Rare Earth						
10	9. 'Let It Be' by The Beatles						
11	10. 'Band Of Gold' by Freda Payne						
12	11. 'Mama Told Me (Not To Come)' by Three Dog Night						
13	12. 'Everything Is Beautiful' by Ray Stevens						
14	13. 'Make It With You' by Bread						
15	14. 'Hitchin' A Ride' by Vanity Fare						
16	15. 'ABC' by Jackson 5						
17	16. 'The Love You Save/I Found That Girl' by Jackson 5						
18	17. 'Cracklin' Rosie' by Neil Diamond						
19	18. 'Candida' by Dawn						

Billboard Hit Songs Dataset

Data Overview

Variable	Description
acousticness	Measures the acousticness of a track from 0.0 to 1.0 with 1.0 being the highest acoustic
artists	Artist of the song
danceability	Measures how suitable a track is for dancing based on tempo, rhythm stability, beat strength, & regularity from 0.0 to 1.0 with 1.0 being the most danceable
duration_ms	The duration of the track in milliseconds.
energy	Measures intensity and activity based on dynamic range, loudness, timbre, onset rate, & general entropy from 0.0 to 1.0 with 1.0 being the most energy
explicit	Measures how much profanity, inappropriate references, & other unsuitable content for children with 0.0 being non explicit and 1.0 being explicit
hitsong	Top 100 songs from Billboard dataset by year (2014-2019)
id	The Spotify ID for the track
instrumentalness	Measures the instruments in a track from 0.0 to 1.0 with 1.0 being purely instrumental (rap or spoken word tracks are very vocal so they would be considered more towards 0.0)
key	All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on

Variable	Description
liveness	Measures the presences of an audience in the track from 0.0 to 1.0 with 1.0 being track was performed live
loudness	Measures the overall loudness of a track in decibels from -60 db to 0 db
mode	Measures the modality (major = 1.0, minor = 0.0) of a track
name	Name of the song
popularity	Popularity of the artist from 0 to 100 with 100 being the most popular
release_date	Date of release in mostly YYYY-MM-DD format (precision may vary)
speechiness	Measures the presence of spoken word from 0.0 to 1.0 with 1.0 being mostly speech like podcasts, audio books, poetry, etc
tempo	Measures the beats per minute/ pacing of the track
year	Year of release

Preprocessing & Cleaning

Step 1: Clean 'artists' and 'track name'

Kaggle

Before ->

	artists	name
0	['Linkin Park']	Final Masquerade
1	['Hippie Sabotage']	Ridin Solo - Njomza Remix
2	['Bleachers']	Wild Heart
3	['together PANGEA']	Sick Shit
4	['David Guetta', 'Showtek', 'VASSY']	Bad (feat. Vassy) - Radio Edit

After ->

	artists	name
0	linkin park	final masquerade
1	hippie sabotage	ridin solo njomza remix
2	bleachers	wild heart
3	together pangea	sick shit
4	david guetta showtek vassy	bad feat vassy radio edit

Billboard

	trackname
0	1. 'Happy' by Pharrell Williams
1	2. 'Dark Horse' by Katy Perry Featuring Juicy J
2	3. 'All Of Me' by John Legend
3	4. 'Fancy' by Iggy Azalea Featuring Charli XCX
4	5. 'Counting Stars' by OneRepublic

	trackname	artist
0	happy	pharrell williams
1	dark horse	katy perry featuring juicy j
2	all of me	john legend
3	fancy	iggy azalea featuring charli xcx
4	counting stars	onerepublic

Preprocessing & Cleaning

Step 2: Join Kaggle data with Billboard Hot 100 Songs data

Join Kaggle dataset with Billboard dataset by ‘track name’.

	genre	track_id	acousticness	artists	danceability	duration_ms	energy	explicit	instrumentalness	name	popularity	hitsong
817	Dance	4T652DlATVHe0jdLKaN3Bw	0.17300	Ariana Grande	0.662	222947.0	0.600	1.0	0.000137	in my head	72.0	1
818	Dance	6QfS2wq5sSC1xAJCQsTSlj	0.41600	Lady Gaga Bradley Cooper	0.575	217213.0	0.330	0.0	0.000000	Shallow - Radio Edit	77.0	0
819	Dance	27356GVuMPFWiJSZCragoM	0.08440	Ariana Grande	0.671	140693.0	0.714	1.0	0.000001	make up	68.0	0
820	Dance	4VUwkH455At9kENOfzTqmF	0.34600	Bazzi Camila Cabello	0.638	180000.0	0.717	0.0	0.000000	Beautiful (feat. Camila Cabello)	79.0	0
821	Dance	2qT1uLXPVPzGgFOx4jtEuo	0.04000	Ariana Grande	0.699	205920.0	0.713	0.0	0.000003	no tears left to cry	82.0	1

Preprocessing & Cleaning

Step 3: Decide the threshold

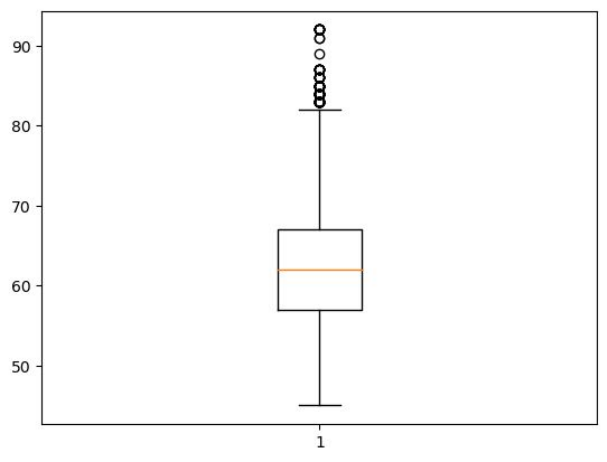
Set threshold using median 62.00 for popularity score

- Popular songs = 47.78%
- Not a popular song = 52.22%

Positive Instance Percentage	*0.47783489*
Negative Instance Percentage	*0.52216510*

$$highRate = \begin{cases} 1, & if\ popularity > 62 \\ 0, & otherwise \end{cases}$$

Average Popularity	*62.60*
Max Popularity	*92.00*
Min Popularity	*45.00*
Q1 Quantile	*57.00*
Q2 Quantile	*62.00*
Q3 Quantile	*67.00*
100th Quantile	*53.00*



Exploratory Data Analysis

Explore factors that affect song popularity

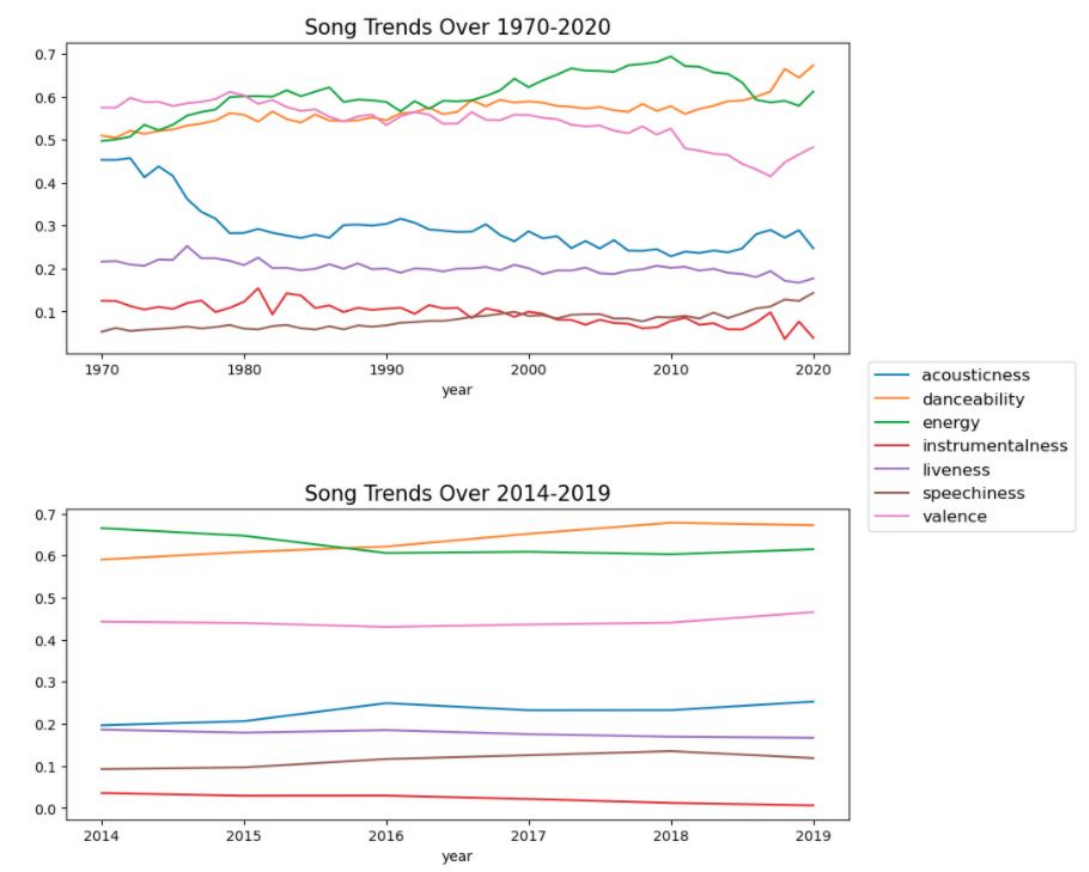
genre	Not Popular	Popular	Percentage Change %
acousticness	0.215912	0.234719	8.710782
danceability	0.614144	0.651891	6.146146
duration_ms	226500.257007	217887.326921	3.802614
energy	0.632306	0.617821	2.290733
explicit	0.405953	0.443108	9.152596
instrumentalness	0.032957	0.016335	50.435634
key	5.136553	5.252120	2.249891
liveness	0.182914	0.174426	4.640387
loudness	-6.931887	-6.651304	4.047705
mode	0.637711	0.611442	4.119385
speechiness	0.111864	0.116243	3.914114
tempo	120.794659	121.251517	0.378210
valence	0.430704	0.447575	3.916874
hitsong	0.056289	0.140615	149.808822
duration_min	3.775004	3.631455	3.802614

The percentage change shows how each variables affect on whether a song is popular.

Top five factors are:

1. Hit song
2. Instrumentalness
3. Explicit
4. Acousticness
5. Danceability

Exploratory Data Analysis

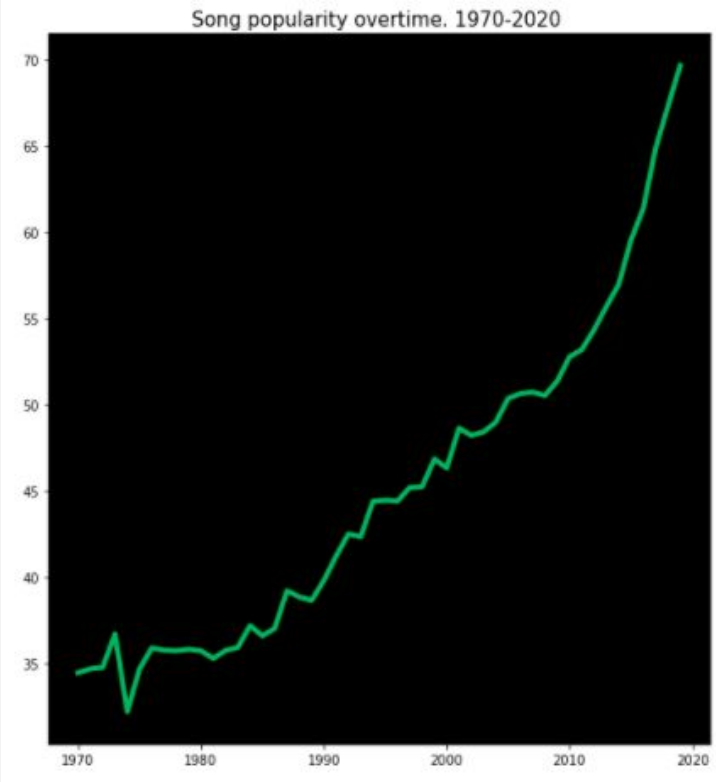


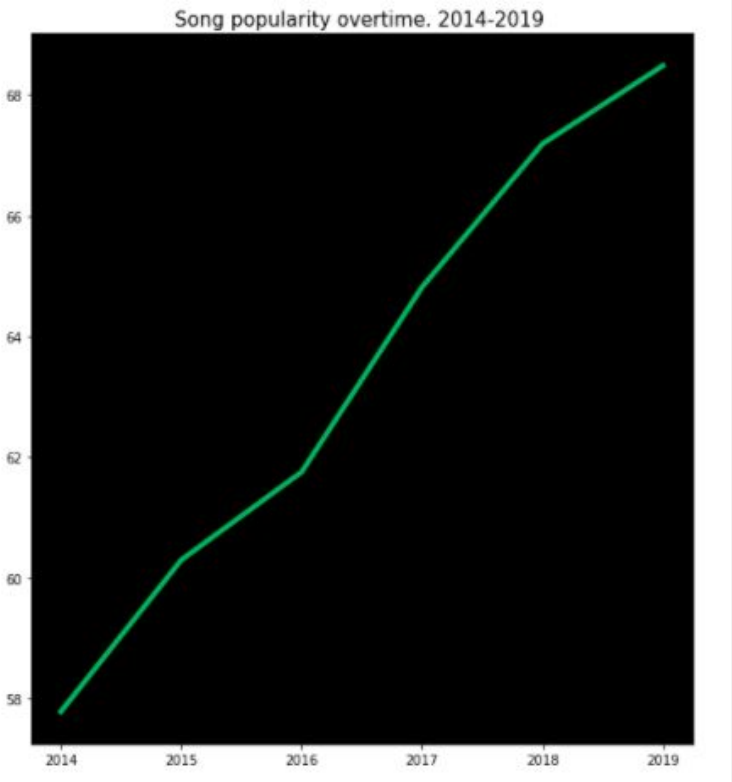
> ***"Popular music started out fairly acoustic in the '50s. After that, its "acousticness" declined steadily, decade after decade, mirroring technology's** integration into greater society at large," explains its blog post."***

- The Guardian

** Meaning less organic means more electric and more click-tracky (think relentlessly pounding techno).

Exploratory Data Analysis



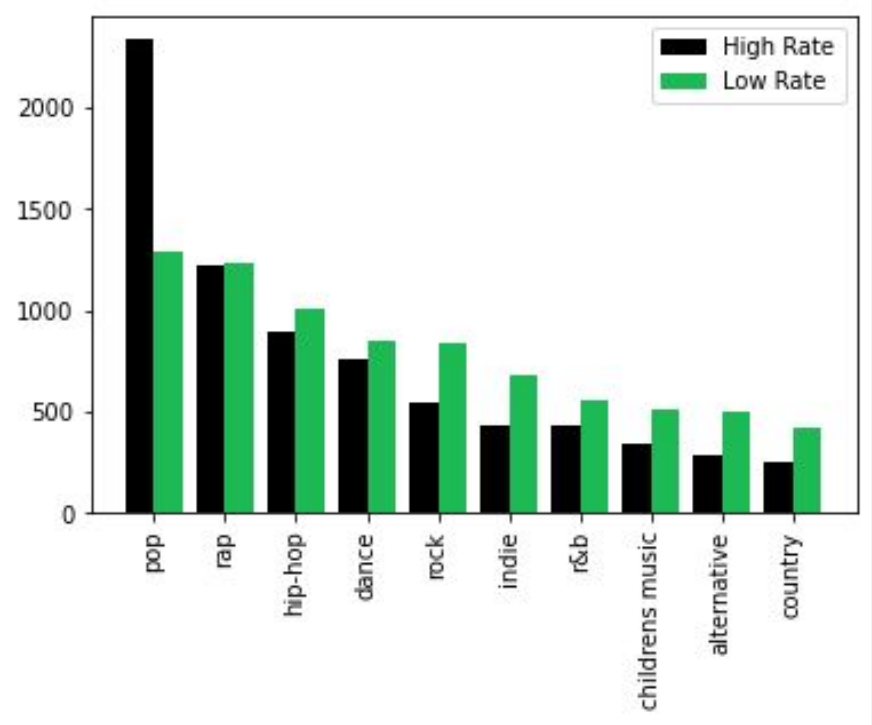


Exploratory Data Analysis

Exploring Genre 1970 - 2020

Exploratory Data Analysis

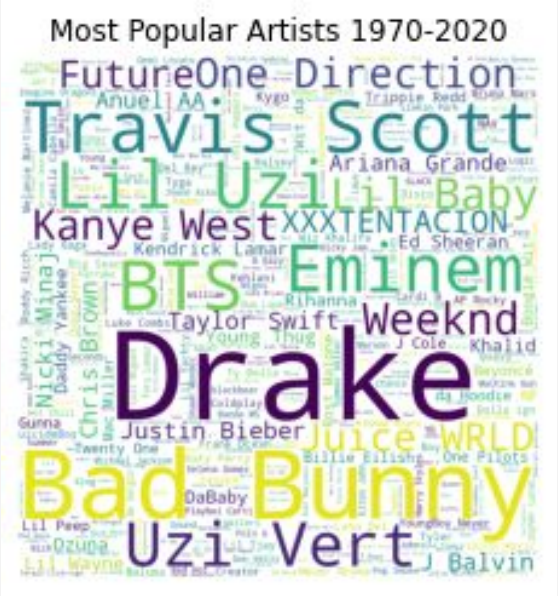
Explore how genre affects songs popularity

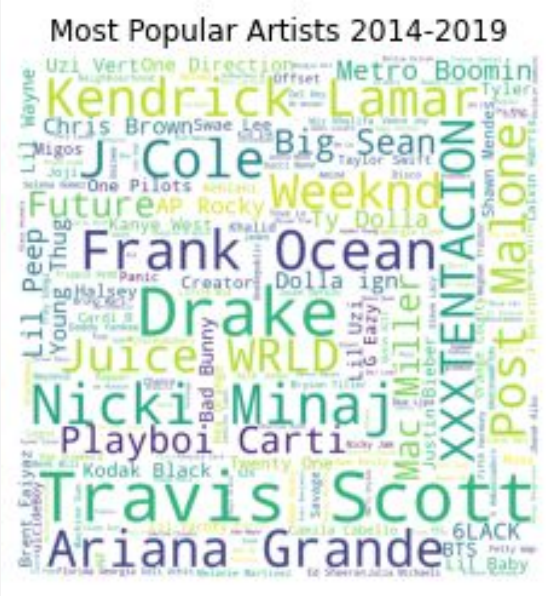


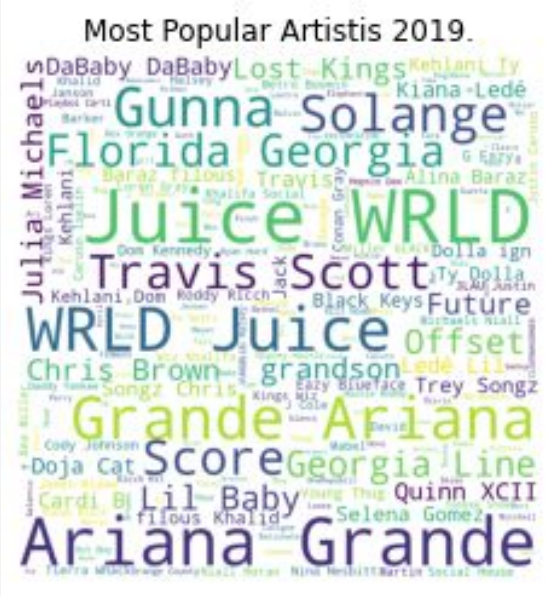
- Higher chance to be popular song: Pop, Rap, hip-hop, dance

- Lower chance to be popular song: rock, indie, r&b, children's music, alternative, country

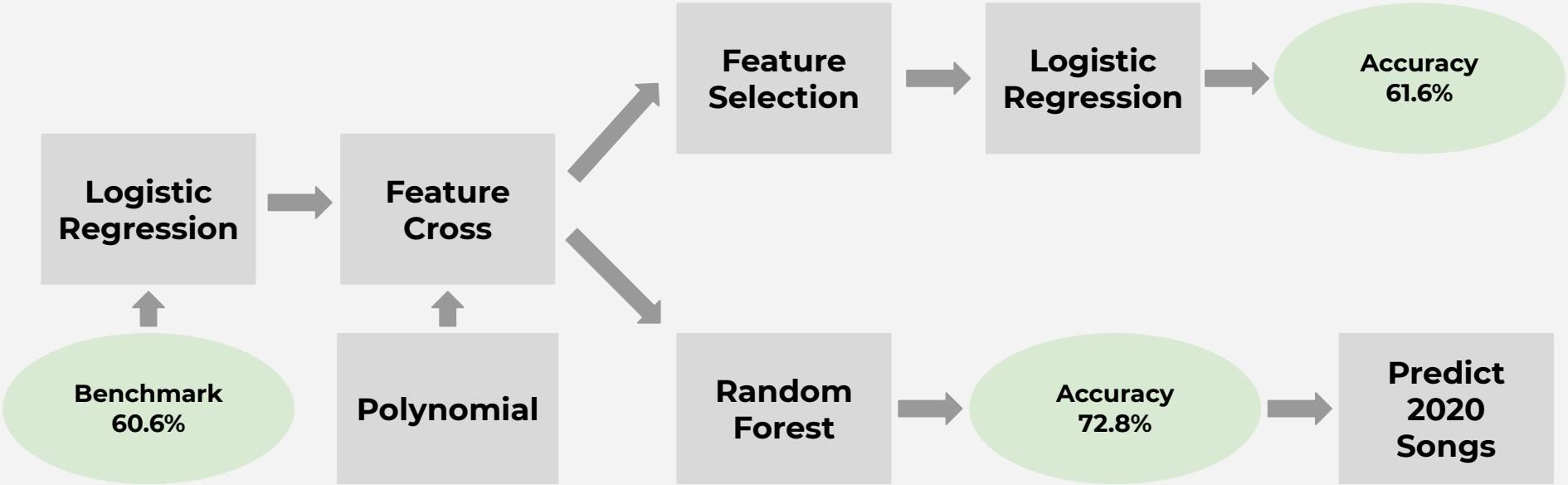
Exploratory Data Analysis







Models Overview

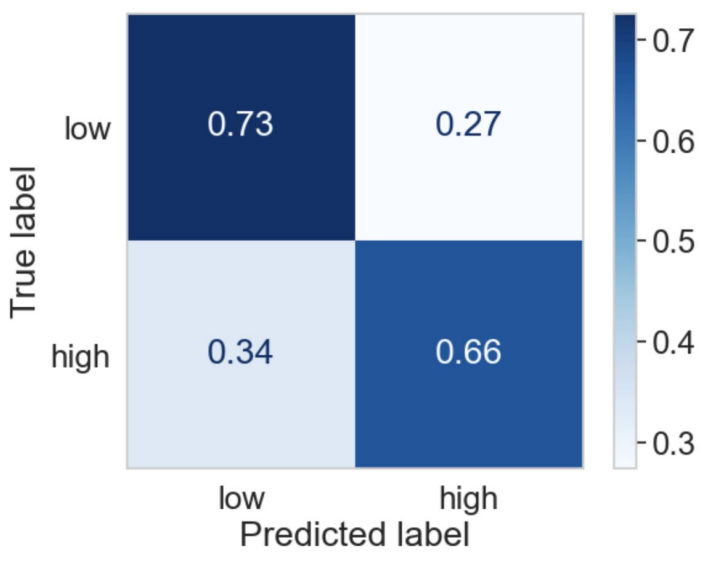


Logistic Regression

Check Accuracy Rate

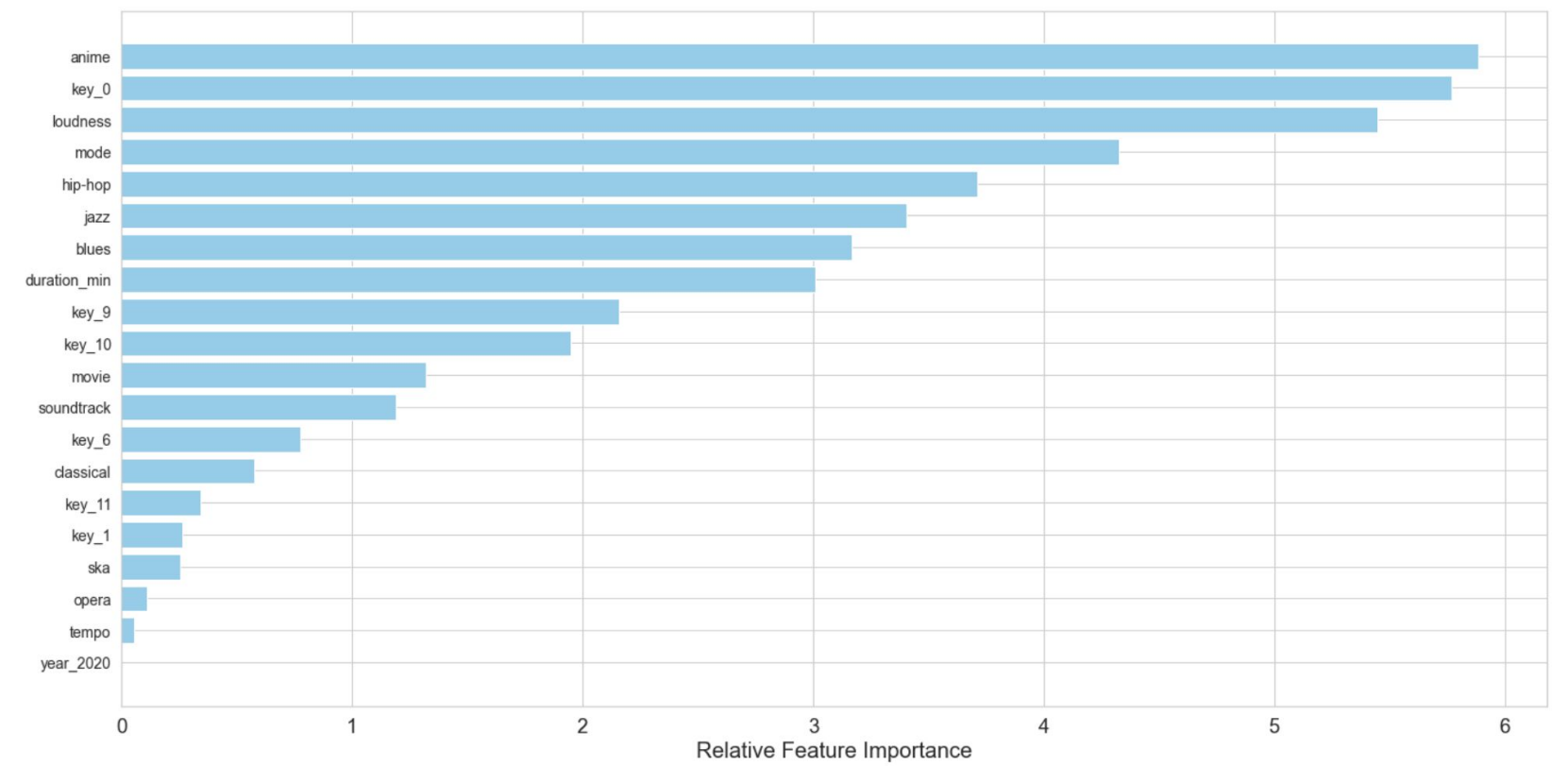
Benchmark Accuracy Rate = 60.58%

```
Average accuracy:  0.6058420035829473
Average precision:  0.6167415216885045
Average recall:  0.5799191857766968
```



Logistic Regression

Feature Importance



Polynomial Features

Top 7 linear correlated features top popularity

```
The most linear correlated features to POPULARITY are:
year_2018          -->  0.26 (abs)
year_2014          -->  0.23 (abs)
pop                -->  0.18 (abs)
hitsong            -->  0.14 (abs)
danceability       -->  0.13 (abs)
year_2015          -->  0.13 (abs)
year_2019          -->  0.12 (abs)
```

Degree = 2 for a better fit of our models

Examples:

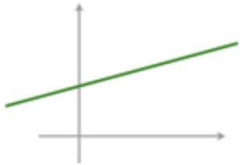
- Features * Features
- Genre * Features

1st degree polynomial

$y = a + bx^1$

straight line with no peaks and no valleys

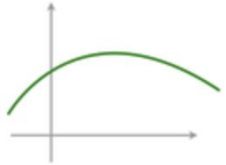
often written as $y = a + bx$



2nd degree polynomial

$y = a + bx^1 + cx^2$

curved line with only one peak or one valley.



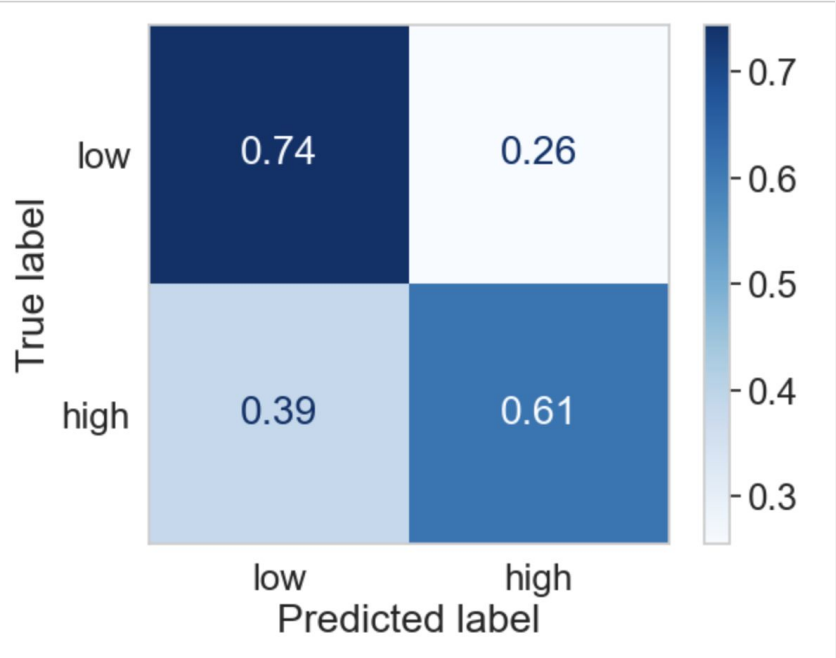
Feature Selection & Logistic Regression

Benchmark Accuracy Rate = 60.58%

Feature Cross & Logistic Regression = 61.14%

Slightly **higher** than the benchmark

```
Average accuracy:  0.6114071663334629
Average precision:  0.6366552051879262
Average recall:  0.558272015874794
```



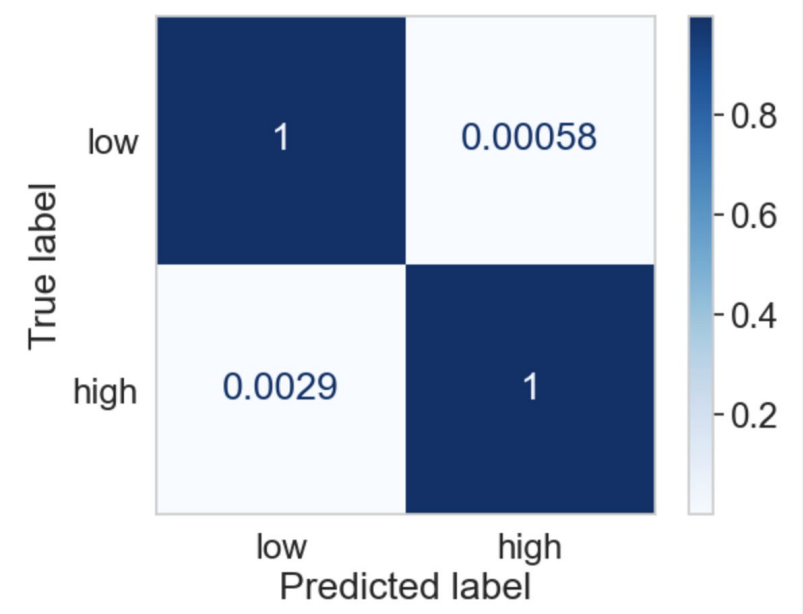
Random Forest

Benchmark Accuracy Rate = 60.58%

Feature Cross & Random Forest = 72.84%

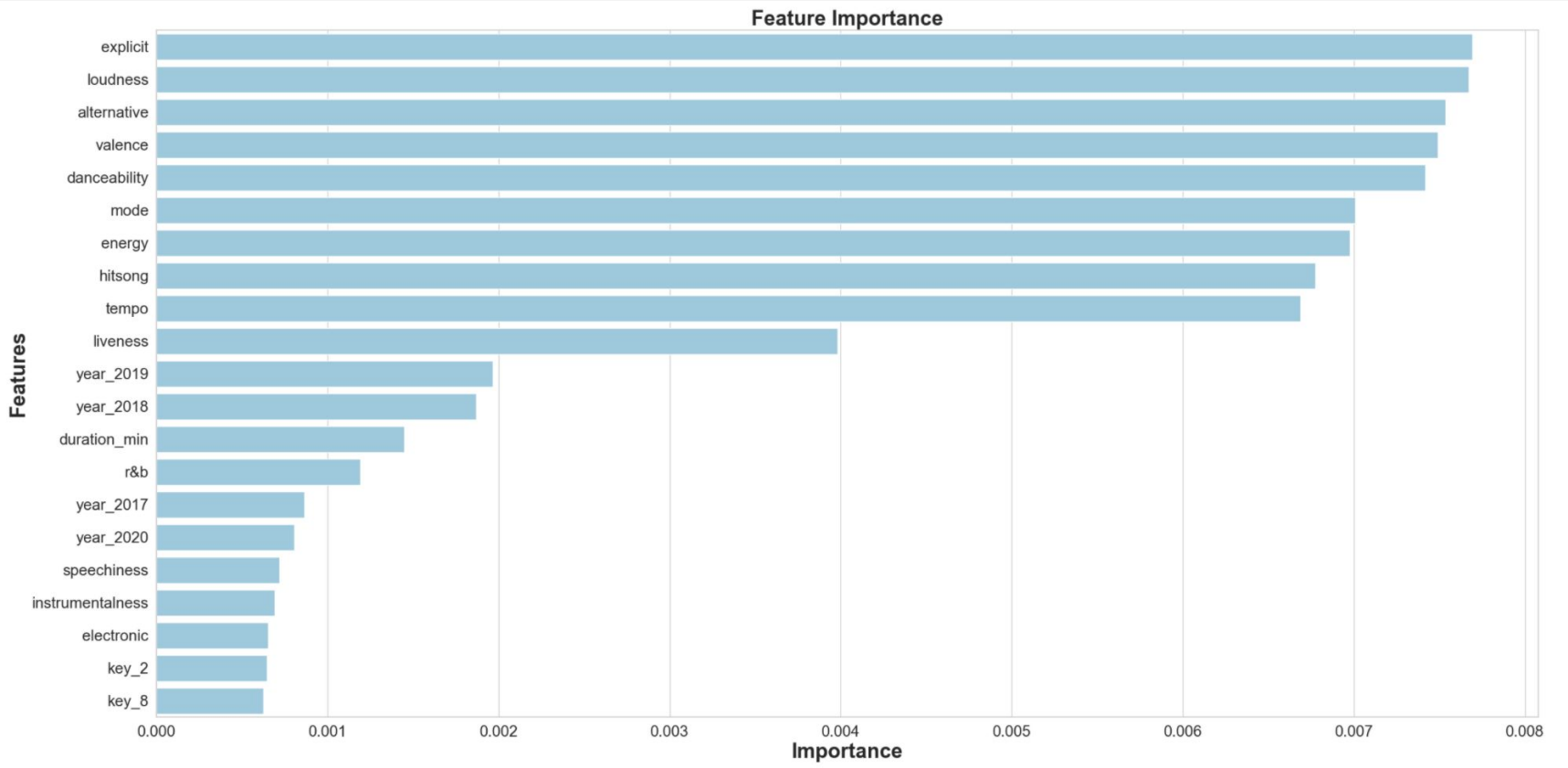
Increased by 12%

```
Average accuracy:  0.728484028526054
Average precision:  0.7620654606378716
Average recall:  0.7377658467890349
```



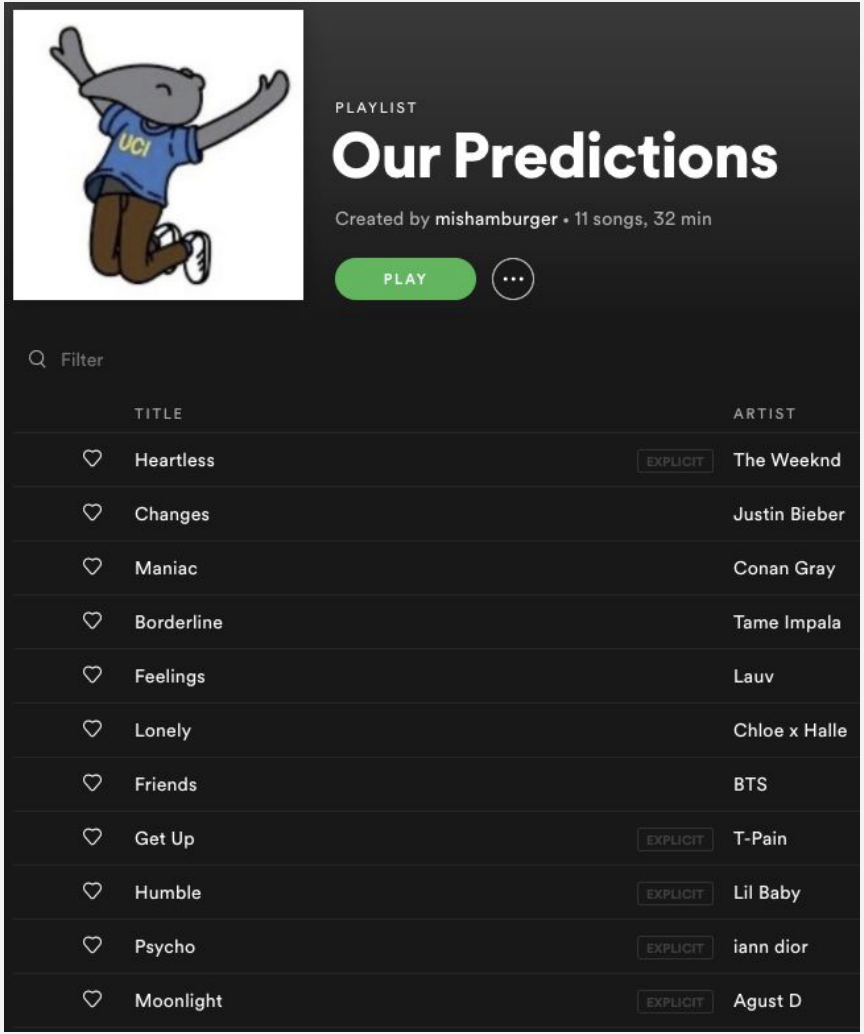
Feature Cross & Random Forest

Feature Importance



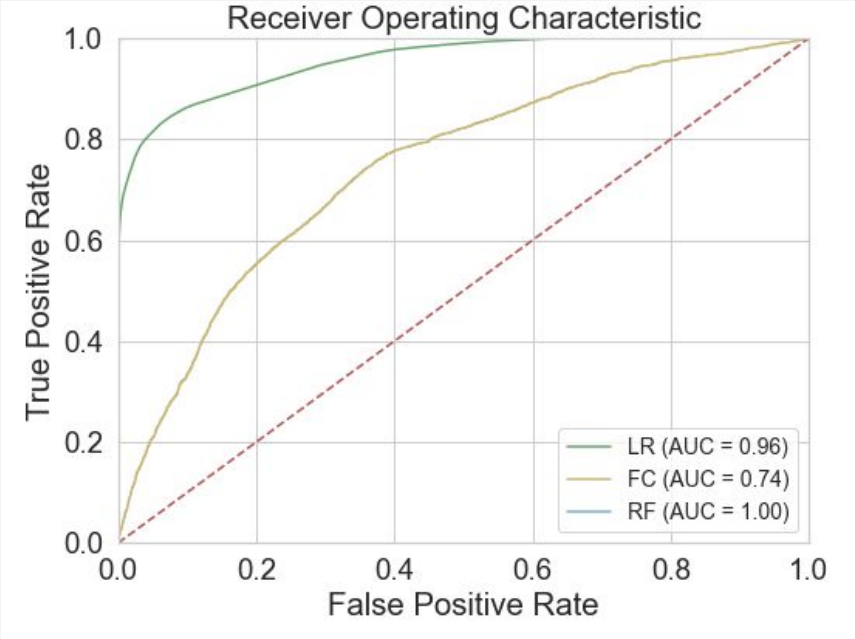
Predict 2020 Songs

Billboard Hot 100 Songs
18. Kings & Queens - Ava Max
20. More Than My Hometown - Morgan Wallen
24. Levitating - Dua Lipa
39. WHATS POPPIN - Jack Harlow
50. Love You Like I Used To Russell Dickerson
73. Took Her to the O - King Von
79. Good Time - Niko Moon
89. Martin & Gina - Polo G



Key Takeaways

Logistic Regression	Feature Cross	Random Forest
60.5842%	61.1407%	72.8484%

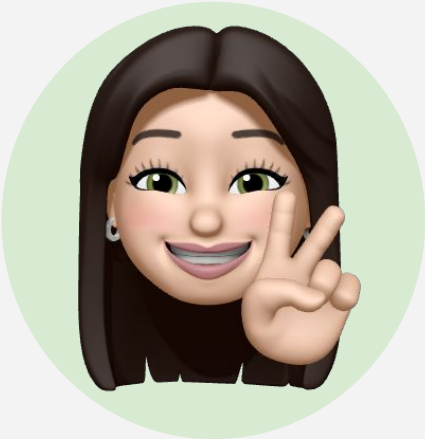


Additional Due Diligence

1. Fix models to accurately predict 2020 songs
2. Use models to define what is the most important attribute

Chengwu Weng

Mira Daya



Misha Khan

Yue Fang