Team 3

Spotify Song Popularity Prediction

Colab Notebook Click Link and QR code

Agnes Shih, Fu Yi Pao, Yu Ting Hung and Yuesen Zhang

HOME

Table of contents

THANKS!

Table of contents

Mars Is a Cold Place
The 15th Planet

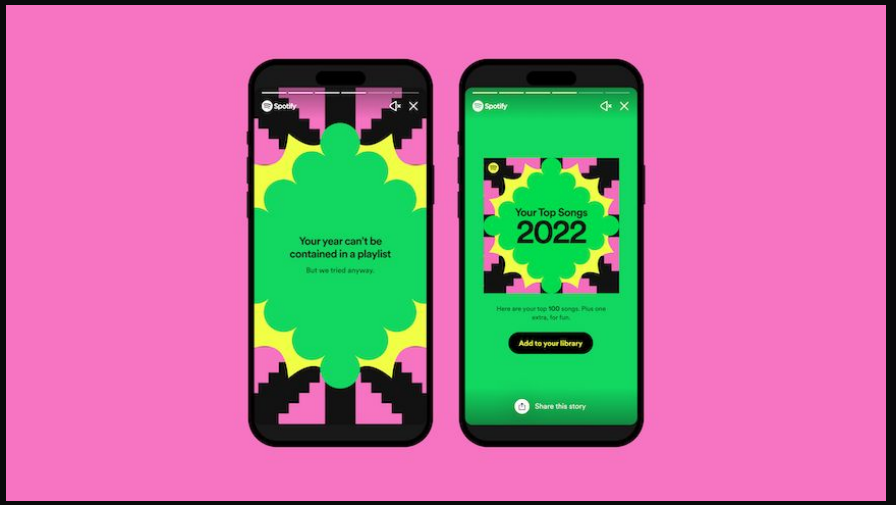
2:54

3:49

Spotify Introduction

Global leader in streaming music sector

- 456 million monthly active users
- 195 million paying subscribers



Evaluate machine learning models from Python to predict Spotify song popularity and identify what makes a song popular.

In short, the Spotify popularity index is calculated by:
Total streams of a song.
How recently a song has been played.
The frequency that a track has been played.

Project Problem Statement

Music has become an integral to our life.

For most people, it is part of their daily routine. However, this outbreak has increased our music listening activities.

We aim to understand the features that popular songs have in common.

Interested in knowing if a song's artist, the genre of the song and the musical features like energy, loudness and danceability among others can help distinguish hit and non-hit songs.

Each song values from 20 different attributes/features are mostly numerical values, but also include some categorical data

Assigns each song a popularity score, based on total number of listens/clicks

Benefits Artists and Record Labels

Streaming is s powerful way for artists to share their music and build a following

- artists get paid based on the number of times their song is played on a platform
- In 2021, spotify paid out over $1.6 b to the record labels, less than 5% of the streaming revenue they make

records labels benefit from streaming revolution

The music labels can get paid for this extra music in a shorter time period. With more artists now using streaming services, it seems likely that the overall revenue will continue to grow.

Preview of Dataset

track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	..
5SuOikwiRyPMVoIQDJUgSV	Gen Hoshino	Comedy	Comedy	73	230666	False	0.676	0.4610	
4qPNDBW1i3p13qLCt0Ki3A	Ben Woodward	Ghost (Acoustic)	Ghost - Acoustic	55	149610	False	0.420	0.1660	
1iJBSr7s7jYXzM8EGcbK5b	Ingrid Michaelson;ZAYN	To Begin Again	To Begin Again	57	210826	False	0.438	0.3590	
6lfxq3CG4xtTiEg7opyCyx	Kina Grannis	Crazy Rich Asians (Original Motion Picture Sou...	Can't Help Falling In Love	71	201933	False	0.266	0.0596	
5vjLSffimiIP26QG5WcN2K	Chord Overstreet	Hold On	Hold On	82	198853	False	0.618	0.4430	

113999 rows, 20 columns

Dataset from Kaggle: Spotify Track Dataset

Spotify tracks over a range of 114 different genres, 89k songs, 31k singers in total.

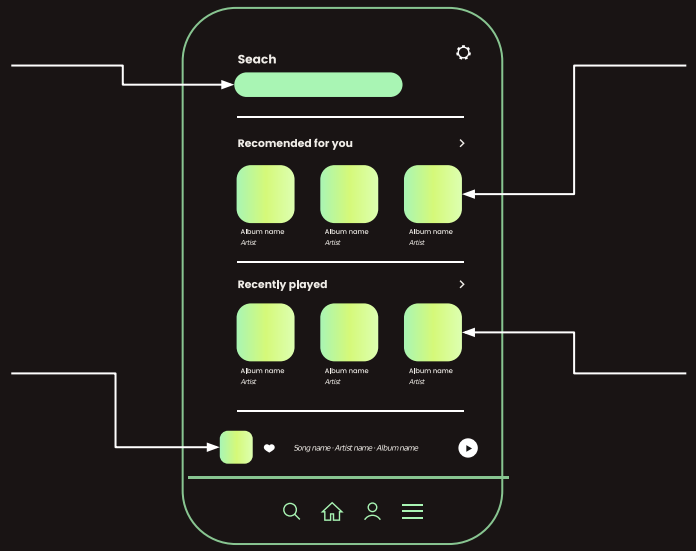
Dataset Description

Categorical

- track_id
- track_genre
- artists
- album_name
- track_name
- explicit
- key
- time signature
- mode

Numeric

- tempo
- duration_ms



Numeric

- popularity
- danceability
- energy
- valence

Numeric

- speechiness
- liveness
- loudness
- acousticness
- instrumentalness

- Perform data cleaning, exploratory the dataset, and lastly build ML models in Python.

Data Cleaning

Column Type

Column	Type
Track_id	Object
Popularity	Integer
Energy	Float

▶

Type
String
Integer
Float

Null Value

Artists
Album_name
Track_name

Duration_ms

millisecond -> minute

Zero values for Tempo

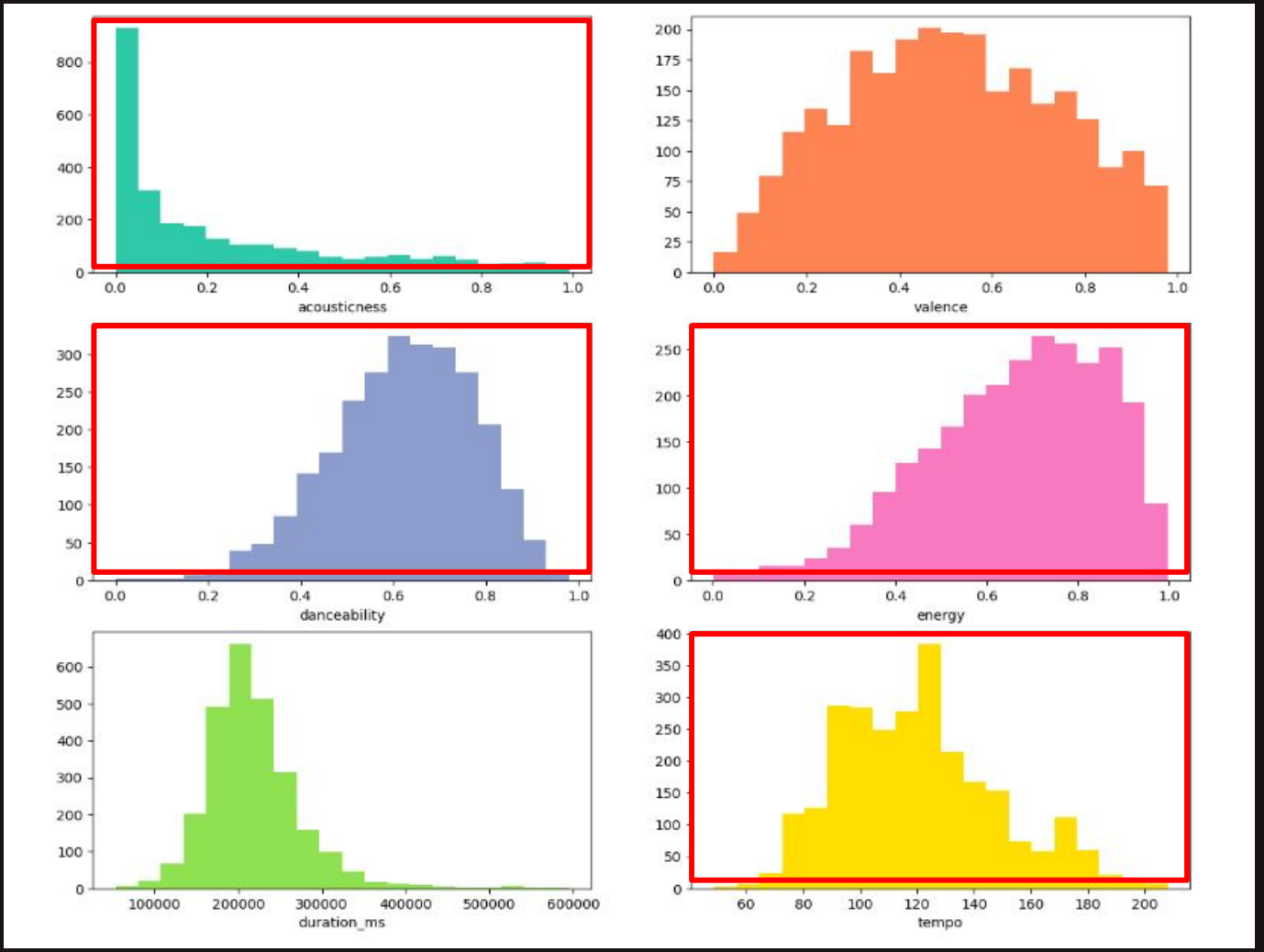
0 Tempo track = Avg Tempo of the genre

Duplicates

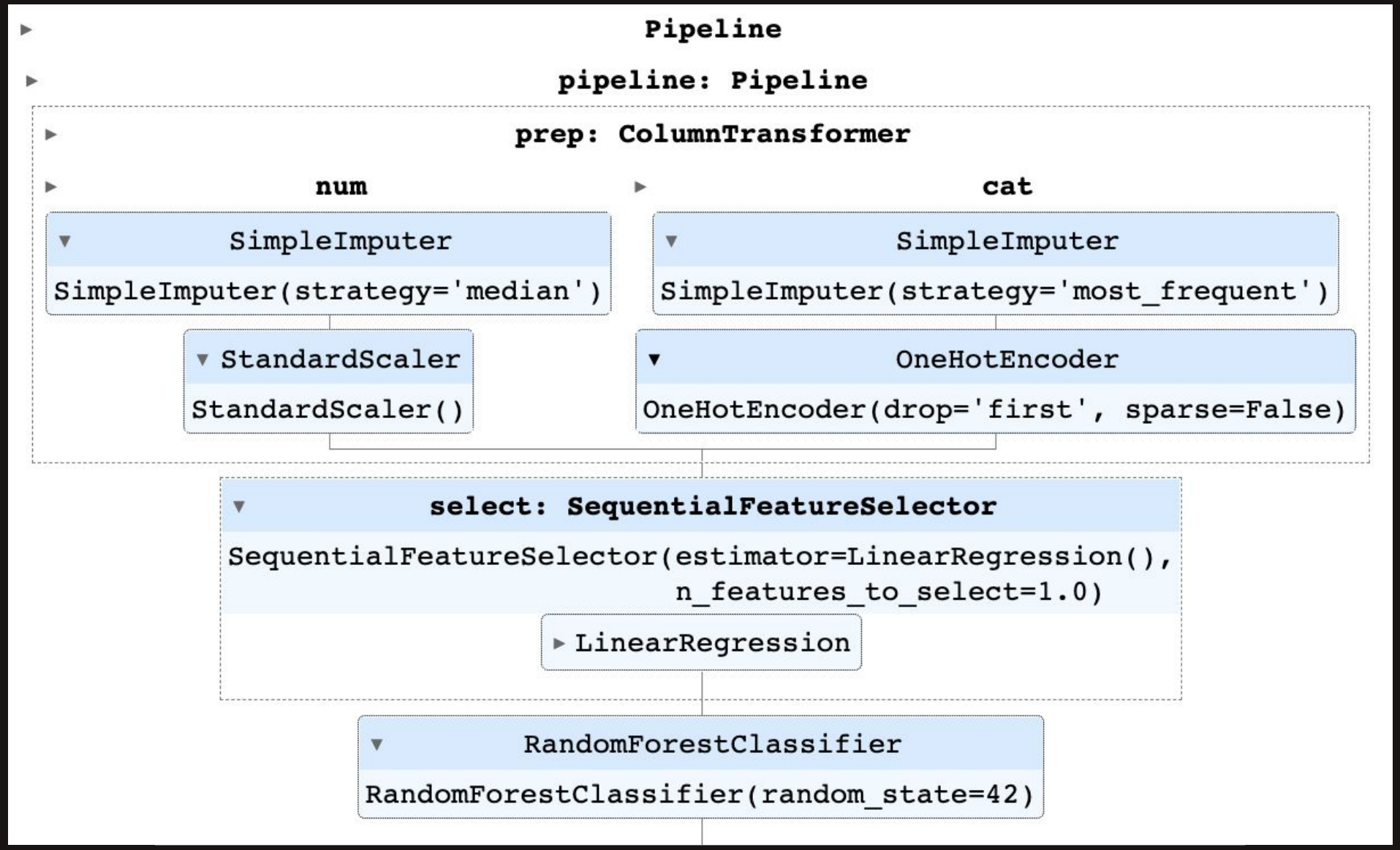
Total 894, Same track_id, track_name, artists, album

track_id	track_name	artists	album_name	popularity
3KKk48f33mlB56F5L5nbJk	"Don Carlos" Roderigo'S Death Aria	Nikolay Kopylov	Popular Opera Arias	0
3KKk48f33mlB56F5L5nbJk	"Don Carlos" Roderigo'S Death Aria	Nikolay Kopylov	Popular Opera Arias	0
4lvfOnCUxyT3aKKamZ3WXu	12 Variations in C Major on "Ah, vous dirai-je...	Wolfgang Amadeus Mozart;Danielle Laval	Mozart - Inspiring Classics	6
4lvfOnCUxyT3aKKamZ3WXu	12 Variations in C Major on "Ah, vous dirai-je...	Wolfgang Amadeus Mozart;Danielle Laval	Mozart - Inspiring Classics	6
1SZp7sIqzHHh1YMaMu8FL2	12 Variations on an Allegretto in B Flat, K.50...	Wolfgang Amadeus Mozart;Danielle Laval	Mozart - A Classical Dawn	9
1SZp7sIqzHHh1YMaMu8FL2	12 Variations on an Allegretto in B Flat, K.50...	Wolfgang Amadeus Mozart;Danielle Laval	Mozart - A Classical Dawn	9

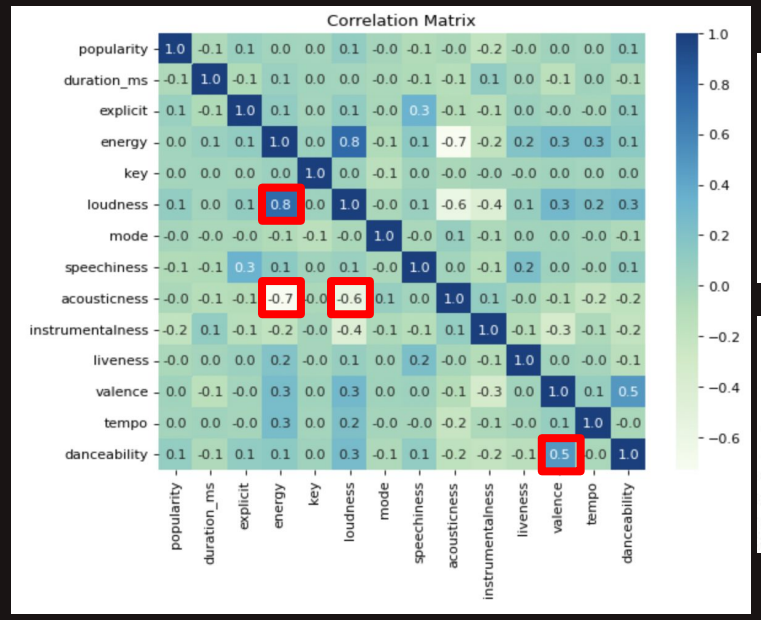
Popularity Analysis: Score>70



Pipeline building



Linear Regression Result



OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.062
Model:	OLS	Adj. R-squared:	0.062
Method:	Least Squares	F-statistic:	206.8

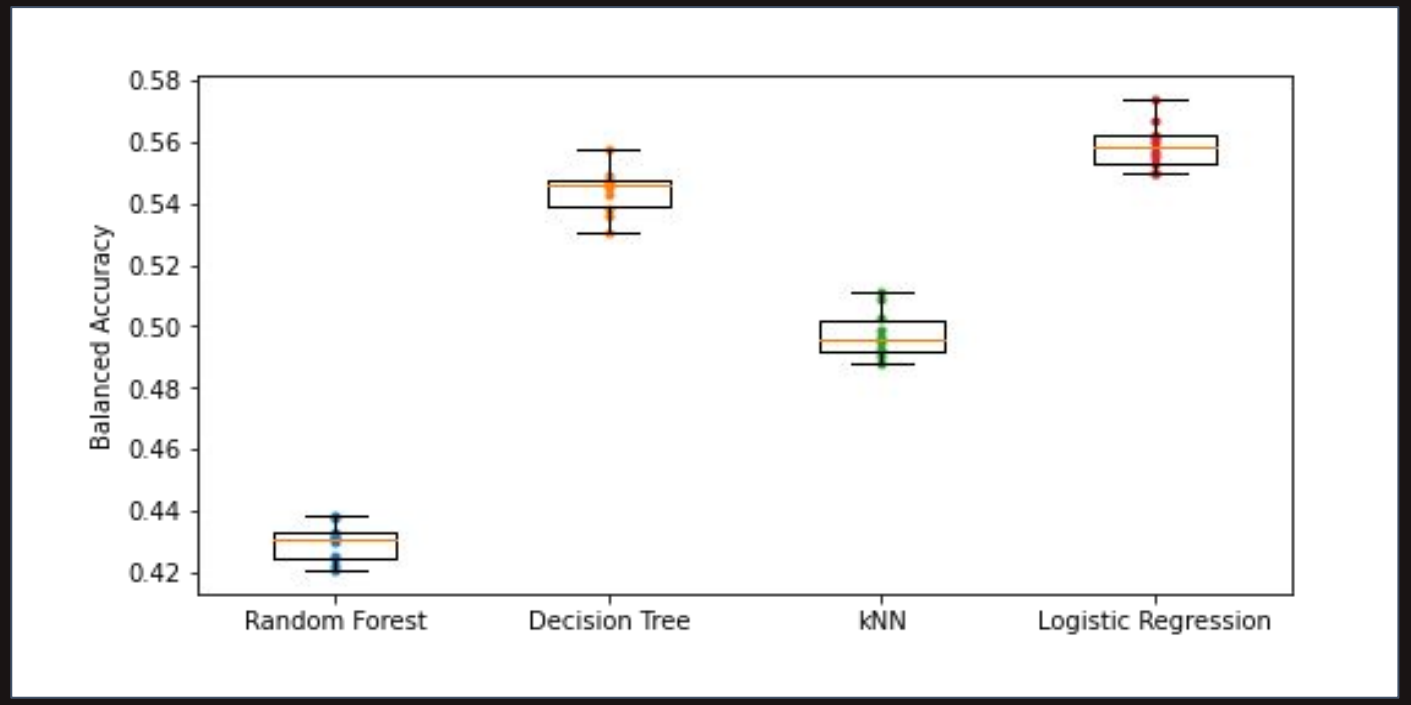
OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.084
Model:	OLS	Adj. R-squared:	0.084
Method:	Least Squares	F-statistic:	266.2

Popularity to Popularity Type

100–70	Popular
70–35	Somewhat Popular
0–35	Unpopular

Balanced Accuracy Scores



We compare four models with their cross validation scores

Decision Tree classifier and Logistic Regression perform better than Random Forest and kNN

Random Forest Result

All variable

0.47

Mean Test Score

6.44

Mean Fit time

Numerical variable

0.42

Mean Test Score

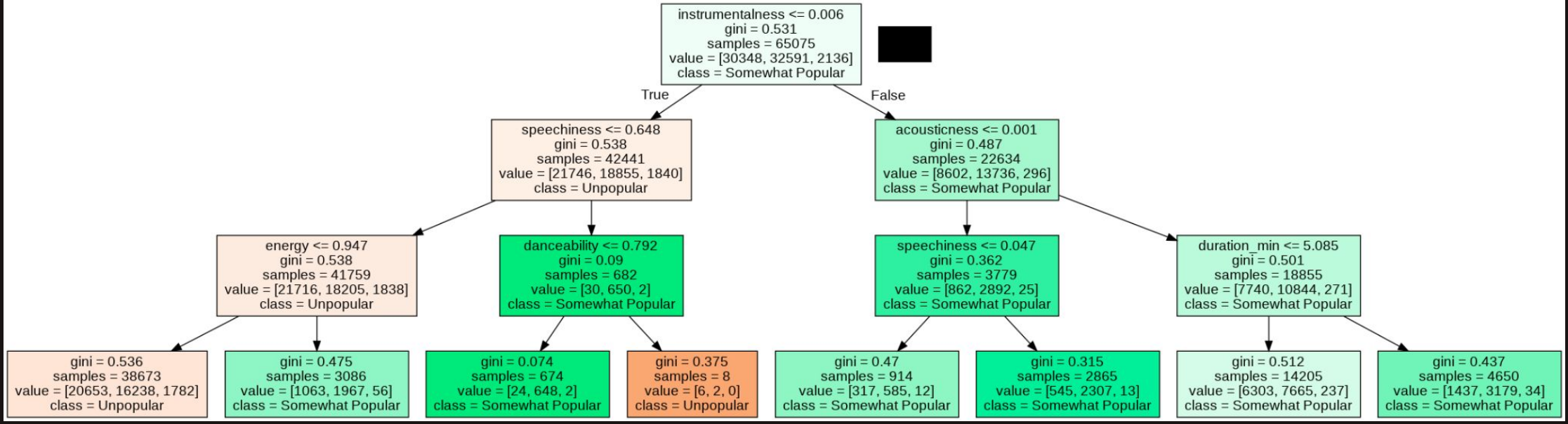
5.48

Mean Fit time

Genre and key still play important role in our dataset

Decision Tree:

from Unpopular to Somewhat Popular



0.55

Test Score

Finding during our process:

duration_ms vs duration_min

0.69
Test Score of duration_ms

0.55
Test Score of duration_min

The test score (of Decision Tree classifier) dropped after we alter the duration from ms to min

Single variable might influence a lot to our model

The Ensembles:
Bagging & Boosting

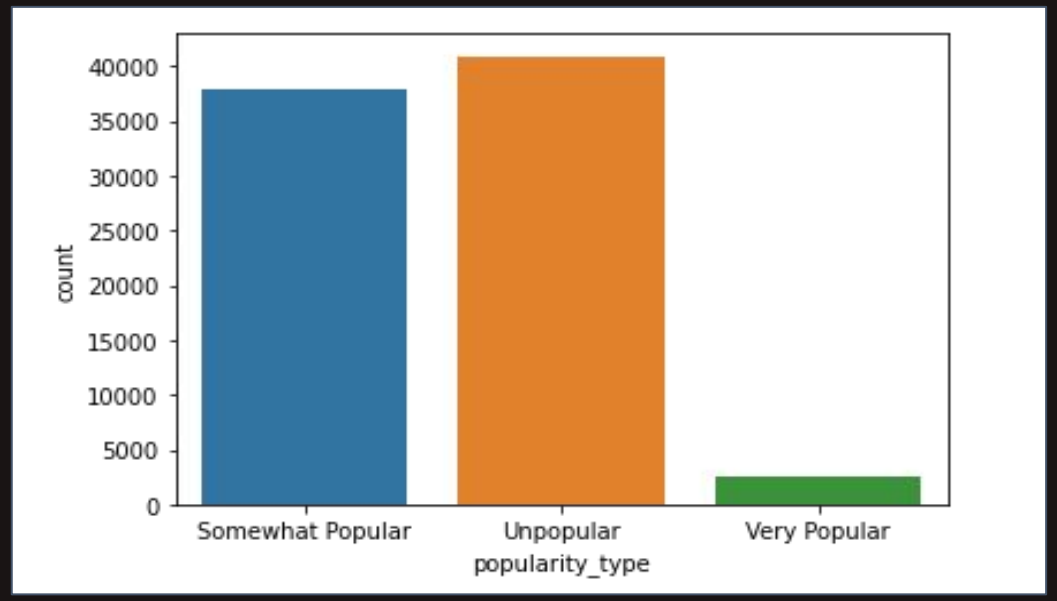
Bagging + Decision Tree Classifier:
0.643 (0.639 out-of-bag score)

Hist Gradient Boosting Classifier:
0.61 > 0.63(Random Search)

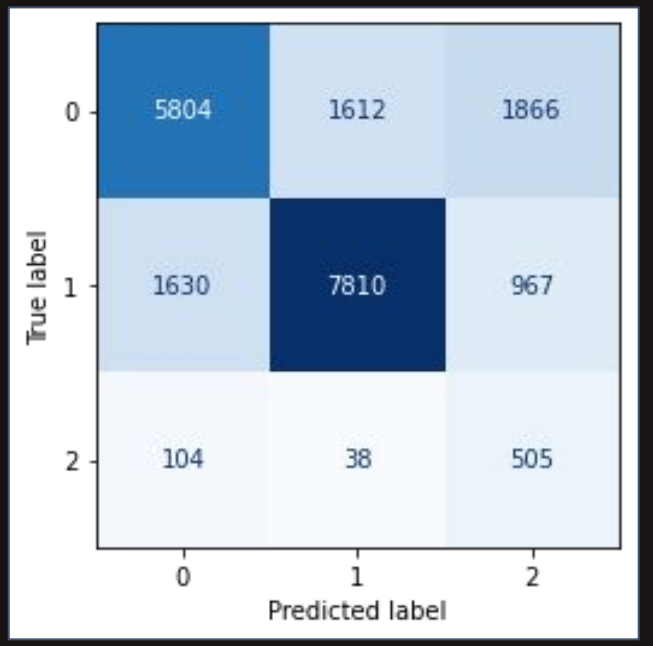
eXtreme Gradient Boosted Trees (XGBoost):
0.622

Class Imbalanced:

Over and Under Sampling



RandomOverSampler
SMOTE
BorderlineSMOTE
RandomUnderSampler



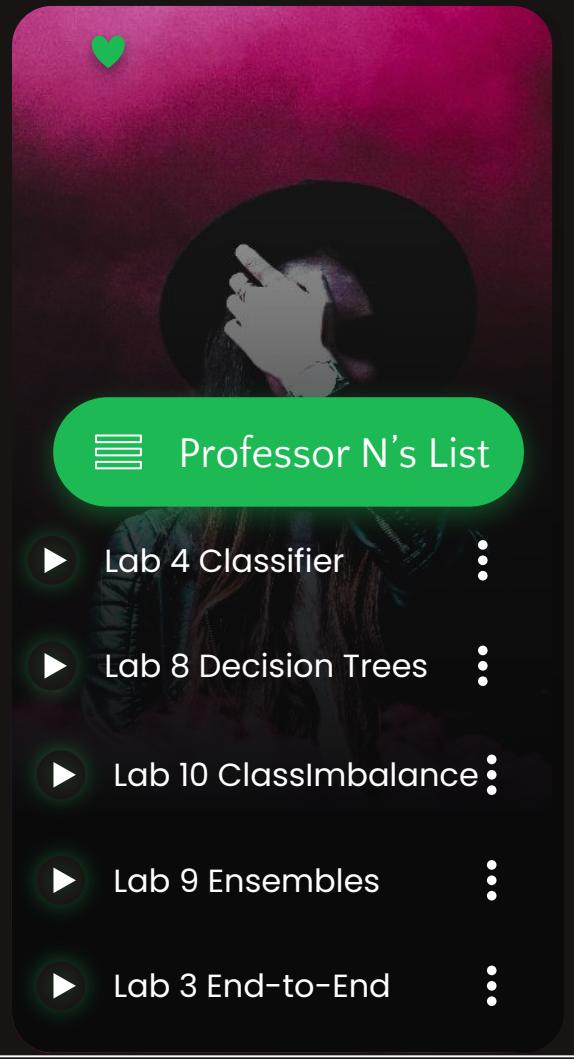
Accuracy: 0.694
Balanced Accuracy: 0.718
F1 Score: 0.725

Conclusion

Biggest Challenge

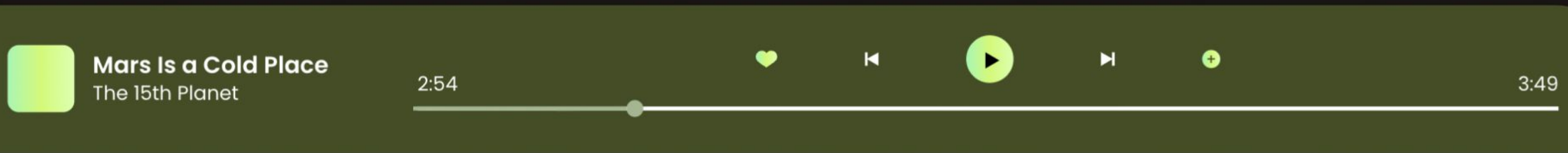
0.46 > 0.6 > 0.71

Score Improvement



Reference

- Professor N's Labs
- Sree, sunku sowmya. "Spotify-Song Prediction and Recommendation System." *Medium*, The Startup, 26 Mar. 2021, https://medium.com/swlh/spotify-song-prediction-and-recommendation-system-b3bbc71398ad.
- "Building a Music Recommendation Engine." *Section*, https://www.section.io/engineering-education/building-spotify-recommendation-engine/.
- "Exploring the Spotify API in Python." *Exploring the Spotify API in Python | Steven Morse*, https://stmorse.github.io/journal/spotify-api.html.
- Fadelli, Ingrid. "Using Spotify Data to Predict What Songs Will Be Hits." *Tech Xplore - Technology and Engineering News*, Tech Xplore, 9 Sept. 2019, https://techxplore.com/news/2019-09-spotify-songs.html.
- "How Spotify's Algorithm Works? A Complete Guide to Spotify Recommendation System [2022]: Music Tomorrow Blog." *How Spotify's Algorithm Works? A Complete Guide to Spotify Recommendation System [2022] | Music Tomorrow Blog*, https://www.music-tomorrow.com/blog/how-spotify-recommendation-system-works-a-complete-guide-2022.
- Jmcabreira. "A-Music-Taste-Analysis-Using-Spotify-API-and-Python./playlist_analysis_ .Ipynb at Master · JMCABREIRA/A-Music-Taste-Analysis-Using-Spotify-API-and-Python." *GitHub*, 13 Dec. 2019, https://github.com/jmcabreira/A-Music-Taste-Analysis-Using-Spotify-API-and-Python./blob/master/Playlist_analysis_%20.ipynb.
- Marin, Max. "AC209A Final Report: Predicting Playlist Success on the Spotify Platform." *Sitewide ATOM*, https://maxgmarin.github.io/AC209a_FinalProject_EEM/.
- Vhtrieu. "Spotify Track Popularity - Analysis and Prediction." *Kaggle*, Kaggle, 31 Oct. 2022,



Thank You

Any Questions?