HOME

YOUR TOP SONGS

PLAYLIST

01 *Risk*

02 *Borderline*

03 *Save Your Tears*

04 *Bombay Rhapsody*

Spotify Track Popularity Predictor

By: Aanvi Goel

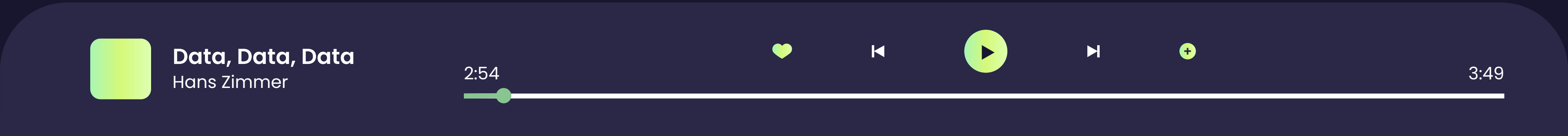
Data, Data, Data
Hans Zimmer

2:54 3:49

Overview

- **Spotify** is a Swedish audio streaming company that that has taken over globally, with 33 million monthly active users, including 188 million paying subscribers, as of June 2022

- **The Big Question:** Can we predict if a track will be popular or not before it’s launch on Spotify?



Objective

Build a Machine Learning model to classify if a track will be **Popular or Not** based on audio features such as danceability, acousticness, tempo etc

Mars Is a Cold Place
The 15th Planet

2:54

3:49

Methodology

Data Ingestion

Two Dataset Sources: Spotify Audio Features Dataset + Spotify Developers Web API

Data Analysis

Performing EDA and Converting to a categorical target

Baseline Model

Building a baseline Logistic Regression Model

Handling Class Imbalance

Applying oversampling, adjusting class weights and probability thresholds

Ensemble and Tree Based Models

Applying different tree based models and optimizing them for best performing metrics

Final Model

Comparing model performance and finalizing the best model for use-case

Data, Data, Data
Hans Zimmer

2:54 3:49

Data Ingestion

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 37934 entries, 0 to 38999
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                37934 non-null  object
 1   name              37934 non-null  object
 2   album             37934 non-null  object
 3   artists           37934 non-null  object
 4   artist_ids        37934 non-null  object
 5   explicit          37934 non-null  object
 6   danceability      37934 non-null  float64
 7   energy            37934 non-null  float64
 8   key               37934 non-null  int64
 9   loudness          37934 non-null  float64
 10  mode              37934 non-null  int64
 11  speechiness       37934 non-null  float64
 12  acousticness      37934 non-null  float64
 13  instrumentalness  37934 non-null  float64
 14  liveness          37934 non-null  float64
 15  valence           37934 non-null  float64
 16  tempo             37934 non-null  float64
 17  duration_ms       37934 non-null  int64
 18  year              37934 non-null  int64
 19  release_date      37934 non-null  object
 20  track_pop         35060 non-null  float64
 21  artist_pop        35060 non-null  object
 22  genres            35060 non-null  object
dtypes: float64(10), int64(4), object(9)
memory usage: 6.9+ MB
```

Spotify Audio Features Kaggle Dataset (Format: .csv)

Queried from Spotify Web API (JSON file)

Data, Data, Data

Hans Zimmer

2:54

3:49

Converting Target to Categorical

Track Popularity

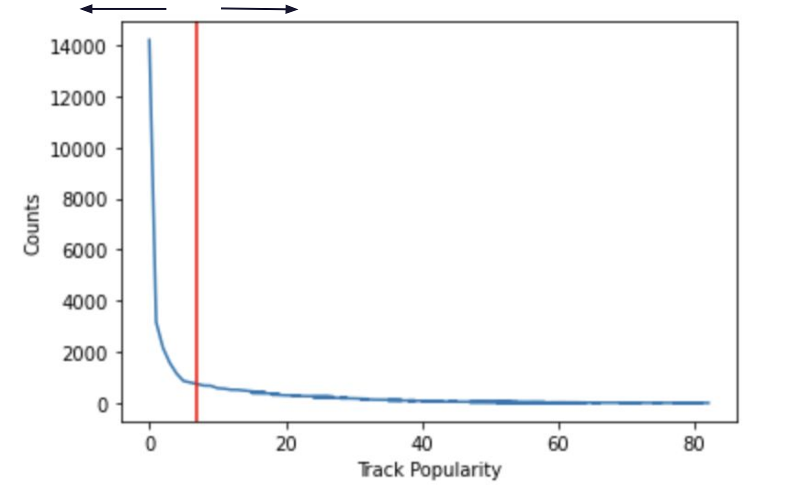
The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.

Track Popularity Stats:

count	35060.000000	25%	0.000000
mean	7.145693	50%	2.000000
std	11.330565	75%	10.000000
min	0.000000	max	82.000000

Class: Not Popular

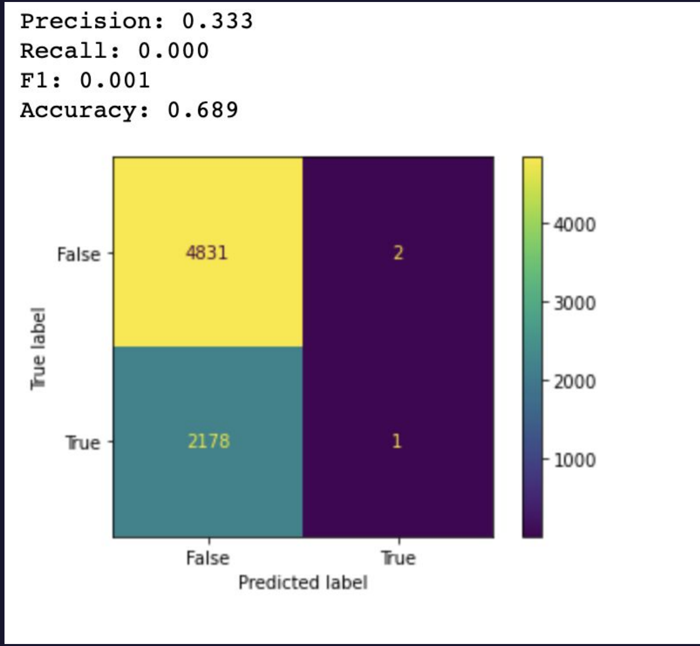
Class: Popular



Data, Data, Data
Hans Zimmer
2:54
3:49

Logistic Regression

- Using a simple Logistic Regression, the model is only predicting False
- This behavior can be attributed to **heavy class imbalance**



Data, Data, Data
Hans Zimmer
2:54
3:49

Handling Class Imbalance

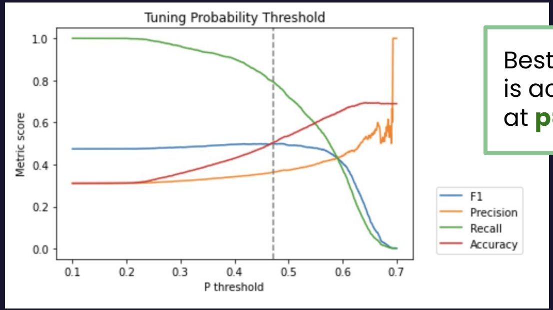
1. Over Sampling :

Training data is resampled with a 2:1 ratio using RandomOverSampler()

2. Adjusting Class Weights :

Class Weights are adjusted during model training 1.15:1 ratio to upweigh minority class

3. Tuning Probability Threshold :



Best F1 score is achieved at **p=0.472**

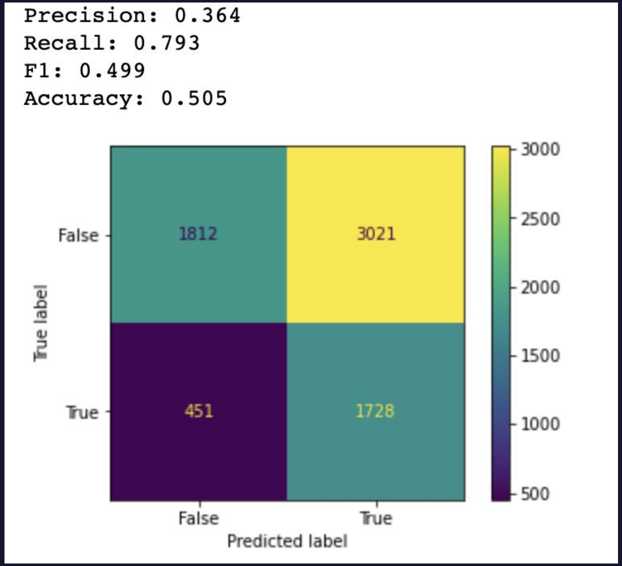
Data, Data, Data

Hans Zimmer

2:54

3:49

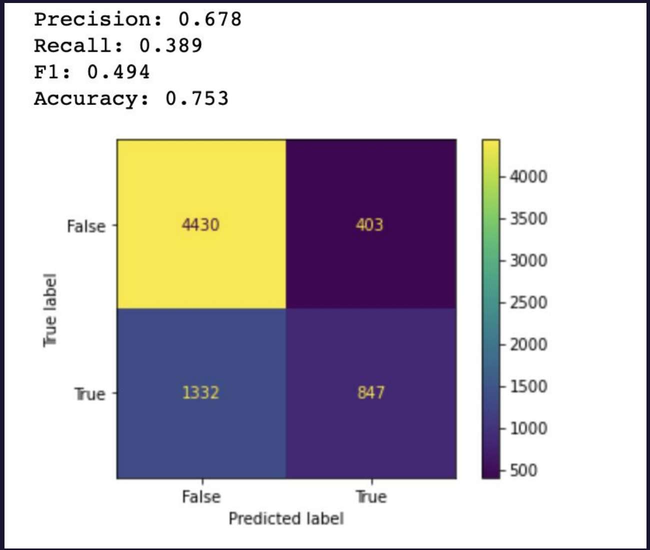
Baseline Model: Logistic Regression



Features	% Odds
danceability	66.543911
energy	-9.528595
key	0.3107
loudness	2.308803
mode	-8.048337
speechiness	-18.082337
acousticness	-47.848309
instrumentalness	-40.08758
liveness	-4.445637
valence	18.8713
tempo	0.020145
year	0.019897

Data, Data, Data
Hans Zimmer
2:54
3:49

Tree Based Ensemble: Random Forest Classifier



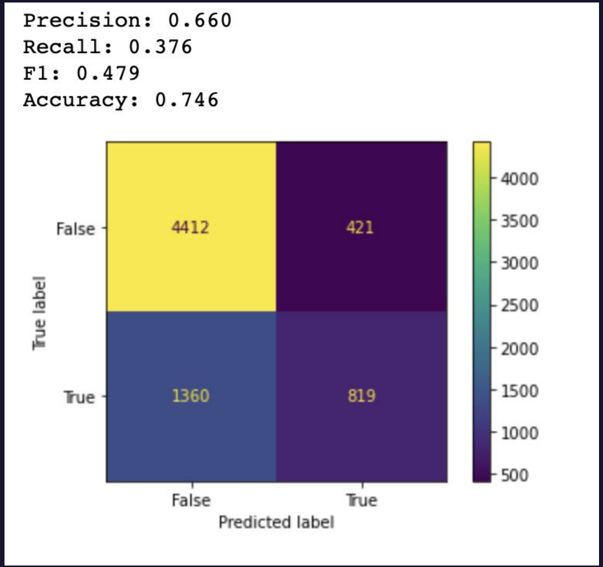
Hyperparameters tuned using **RandomizedSearchCV()**:

{'n_estimators': 1800,
'min_samples_split': 5,
'min_samples_leaf': 1,
'max_features': 'auto',
'max_depth': 70,
'bootstrap': False}

Data, Data, Data
Hans Zimmer

2:54 3:49

Tree Based Ensemble: XGBoost



Hyperparameters tuned using **GridSearchCV():**

{'colsample_bylevel': 0.4, 'learning_rate': 0.01, 'max_depth': 10, 'n_estimators': 1000}

Data, Data, Data
Hans Zimmer

2:54

3:49

Model Comparison

Model		Precision	Recall	F1	Accuracy
Logistic Regression	Simple Logistic Regression	0.333333	0.000459	0.000917	0.689104
	Logistic Regression w/ Balanced Weights	0.383166	0.681046	0.490416	0.560183
	Optimized Logistic Regression	0.383243	0.680128	0.490241	0.560468
	Logistic Regression w/ OverSampling	0.395402	0.639284	0.4886	0.584141
	Logistic Regression w/ SMOTE	0.391471	0.610831	0.477146	0.583999
	Logistic Regression w/ 2:1 Class Weights	0.395645	0.642038	0.489589	0.583999
	Logistic Regression w/ OS + Class Weights	0.373102	0.721891	0.491947	0.536651
	Baseline: Logistic Regression + Handling Class Imbalance	0.363866	0.793024	0.498845	0.504849
Tree-Based Models	Decision Tree	0.451066	0.475906	0.463153	0.657159
	Base Random Forest	0.614097	0.319872	0.42064	0.726184
	Optimized Random Forest	**0.6776**	**0.38871**	**0.494022**	**0.752567**
	XGBoost	0.660484	0.37586	0.479087	0.746007

Data, Data, Data
Hans Zimmer
2:54
3:49

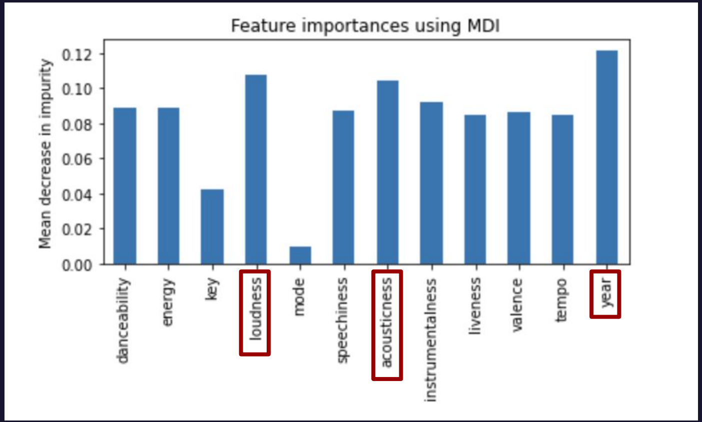
Final Model

Random Forest Classifier

Performance Metrics:

Precision: 0.678
Recall: 0.389
F1: 0.494
Accuracy: 0.753

Feature Importance:



Data, Data, Data
Hans Zimmer

2:54 3:49

Future Improvements

- Apply Feature Engineering on the features data
- Data acquisition: Increase the size of the initial dataset
- Tweak the Percentile Cut-off for the categorical conversion

Data, Data, Data
Hans Zimmer

2:54 3:49