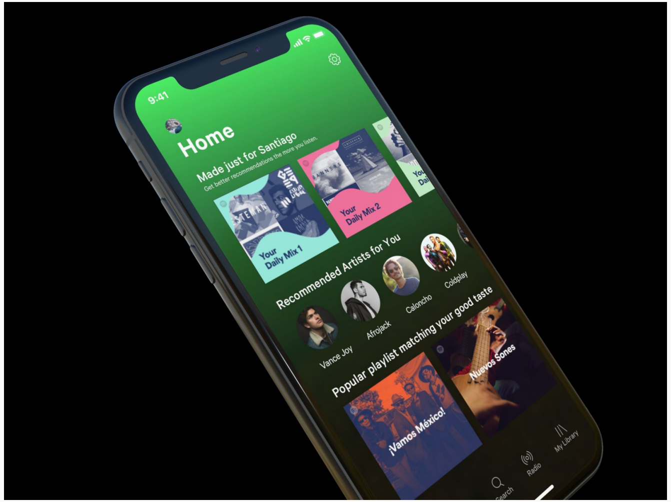
Group 5 - Project 4:
Spotify Song Popularity Predictor

By Jesse Heaton, Azkya Said, Charles Martin,
Rehman Butt, & Crystal Burkhardt

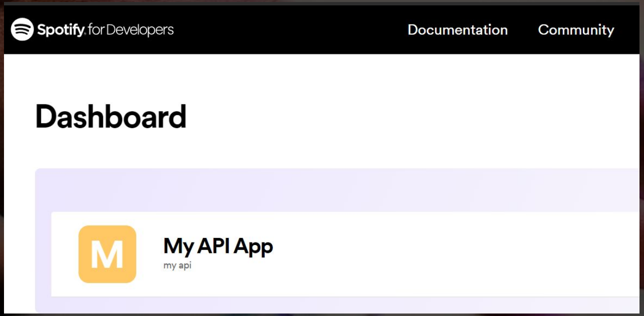
Machine Learning Model Overview

- We created a Random Forest model that will predict a new song's popularity based on the song features of 2000 of the top songs
- Spotify's API call data set comes with a 'popularity' feature based on how often a song is listened to currently
- We've defined a 'Popular' song as a song with a Spotify popularity rating of 70% or above



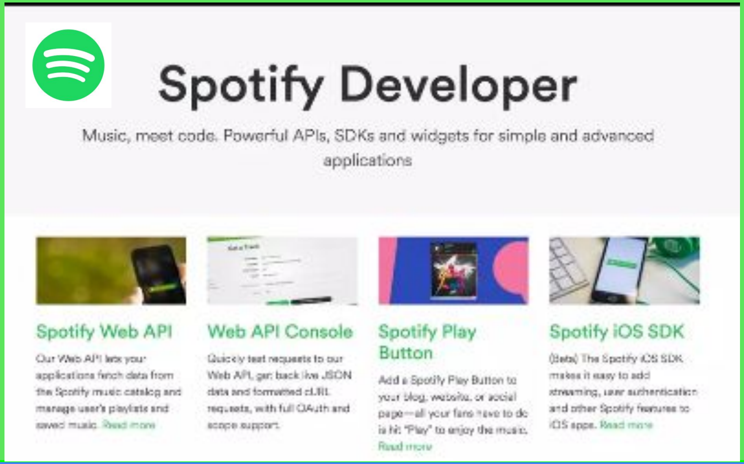
The API Call:

It's free!



Accessing the Spotify API:

- Comprehensive documentation
- We retrieved the client ID, client secret code, and setup the redirect URL - Server verifies that the redirect URL in this request matches the redirect URL
- Request to exchange an authorization code for an access token



The Machine Learning Process

Pre-Process the Data

Build and Train the Random Forest Model

Optimize the Model to fit our Specific Target Accuracy and Loss

Our Goal & Data Set

Goal : Create a model that has at least a 75% accuracy rate for classifying a song's likelihood of becoming popular or unpopular.

Data Set: Contains 2000 tracks; of which include the top 100 songs of each year since 2015, and the top 100 songs of multiple genres

What information can be retrieved with the Spotify API?

Albums

Millions of albums with

- Cover art
- Data on available markets, copyrights, genres, IDs, popularity

Artists

Millions of artists with

- Artists images
- Data on followers, genres, IDs, related artists, popularity, top tracks

Tracks

30+ Million tracks

- 30 second previews (most tracks)
- Data on artists, available markets, explicit lyrics, IDs, popularity

Playlists

2+ billion playlists with

- Full track listings
- Playlist images
- Data on followers

User Profiles

Spotify user profiles with

- Profile images
- Data on country, followers and following, subscriptions
- Top Tracks and Artists

And more...

- Search across albums, artists, tracks and playlists
- Top lists & new releases
- Browse by category
- Track recommendations
- Audio features

Data Set Overview

song_id	song_name	album	artist	duration(ms)	release_date	acousticness	danceability	energy	insturmentalness	key	liveness	loudness	mode	spechiness	tempo	time_signature	valence	popularity	binary_popularity
3k79jB4aG	vampire	vampire	Olivia Rodrigo	219724	6/30/2023	0.169	0.511	0.532	0	5	0.311	-5.745	1	0.056	137.827	4	0.322	90	1
2IGMVunI	Paint The Town Red	Paint The Town Red	Doja Cat	231750	8/4/2023	0.269	0.868	0.538	3.34E-06	5	0.0901	-8.603	1	0.174	99.968	4	0.732	100	1
4KULAymI	I Remember Everything (feat. Ka	Zach Bryan	Zach Bryan	227195	8/25/2023	0.554	0.429	0.453	2.00E-06	0	0.102	-7.746	1	0.0459	77.639	4	0.155	91	1
1Lo0QY9c	Fast Car	Gettin' Old	Luke Combs	265493	3/24/2023	0.186	0.712	0.603	0	8	0.115	-5.52	1	0.0262	97.994	4	0.67	89	1
1BxfuPKG	Cruel Summer	Lover	Taylor Swift	178426	8/23/2019	0.117	0.552	0.702	2.06E-05	9	0.105	-5.707	1	0.157	169.994	4	0.564	98	1
7K3BhSpA	Last Night	One Thing At A Time	Morgan Wallen	163854	3/3/2023	0.467	0.492	0.675	0	6	0.142	-5.456	1	0.0389	203.759	4	0.478	86	1
2i8f4VnnE	bad idea right?	bad idea right?	Olivia Rodrigo	184783	8/11/2023	0.00196	0.628	0.878	1.23E-05	9	0.0649	-3.468	1	0.0901	129.976	4	0.796	86	1
4iZ4pt7kv	Snooze	SOS	SZA	201800	12/9/2022	0.141	0.559	0.551	0	5	0.11	-7.231	1	0.132	143.008	4	0.392	91	1
4rXLjWdF	fukumean	a Gift & a Curse	Gunna	125040	6/16/2023	0.119	0.847	0.622	0	1	0.285	-6.747	0	0.0903	130.001	4	0.22	95	1
1vYXt7VSj	Dance The Night - From Barbie Th	Dance The Night (From Barbi	Dua Lipa	176579	5/25/2023	0.0207	0.671	0.845	0	11	0.329	-4.93	0	0.048	110.056	4	0.775	95	1
2gyxAWH	get him back!	GUTS	Olivia Rodrigo	211141	9/8/2023	0.0135	0.546	0.846	0	5	0.607	-5.719	1	0.181	162.043	4	0.74	92	1
0WtM2NB	Calm Down (with Selena Gomez)	Calm Down (with Selena Gon	Rema	239317	8/25/2022	0.382	0.801	0.806	0.000669	11	0.114	-5.206	1	0.0381	106.999	4	0.802	90	1
34sOdxW	all-american bitch	GUTS	Olivia Rodrigo	165833	9/8/2023	0.254	0.43	0.692	4.17E-06	6	0.148	-4.234	1	0.0429	156.261	4	0.401	90	1
4YQImHfl	Bongos (feat. Megan Thee Stallio	Bongos (feat. Megan Thee St	Cardi B	175099	9/8/2023	0.0791	0.726	0.738	0	0	0.0297	-4.788	1	0.355	121.953	4	0.896	83	1
741UUVE2	Barbie World (with Aqua) [From	Barbie World (with Aqua) [Fr	Nicki Minaj	109750	6/23/2023	0.519	0.77	0.58	0.000127	0	0.233	-8.393	1	0.247	144.072	4	0.753	92	1
3NI5Okkm	the grudge	GUTS	Olivia Rodrigo	189386	9/8/2023	0.916	0.548	0.307	0	10	0.0828	-9.214	1	0.0781	127.923	4	0.317	88	1
0yLdNVW	Flowers	Flowers	Miley Cyrus	200454	1/13/2023	0.0632	0.707	0.681	5.15E-06	0	0.0322	-4.325	1	0.0668	117.999	4	0.646	74	0
6HgWWaN	All My Life (feat. J. Cole)	All My Life (feat. J. Cole)	Lil Durk	223878	5/12/2023	0.15	0.829	0.436	0	3	0.0954	-8.205	1	0.327	143.031	4	0.693	84	1
5CscrLqFB	making the bed	GUTS	Olivia Rodrigo	198866	9/8/2023	0.346	0.537	0.413	0	7	0.12	-7.736	1	0.0319	133.21	4	0.226	88	1
53dtP2iUN	logical	GUTS	Olivia Rodrigo	231907	9/8/2023	0.853	0.499	0.246	0	7	0.1	-8.082	1	0.0345	81.051	4	0.153	87	1
0ec7EBr0n	Religiously	Religiously	Bailey Zimmerman	178723	3/17/2023	0.426	0.57	0.67	0	1	0.108	-6.519	1	0.0274	140.924	4	0.593	37	0
78Du4CMI	Rich Men North of Richmond	Rich Men North of Richmond	Oliver Anthony Music	187244	8/11/2023	0.753	0.678	0.264	0	10	0.302	-12.384	1	0.0622	121.722	4	0.528	88	1
6QT6j7rKt	lacy	GUTS	Olivia Rodrigo	177212	9/8/2023	0.803	0.455	0.379	0	3	0.109	-7.628	1	0.0341	76.967	4	0.413	87	1
5sp71CUt	ballad of a homeschooled girl	GUTS	Olivia Rodrigo	203369	9/8/2023	0.0557	0.385	0.873	0	1	0.353	-3.331	1	0.069	120.393	4	0.469	87	1
26QLJMK8	love is embarrassing	GUTS	Olivia Rodrigo	154516	9/8/2023	0.00261	0.52	0.831	0	9	0.156	-3.432	1	0.0968	160.035	4	0.677	87	1
0Nfh4u0J	Used To Be Young	Used To Be Young	Miley Cyrus	191185	8/25/2023	0.6	0.451	0.632	0	4	0.139	-3.843	1	0.0475	147.06	4	0.455	85	1
0PAcdVzh	Thinkinâ€™ Bout Me	One Thing At A Time	Morgan Wallen	177387	3/3/2023	0.492	0.656	0.757	0	3	0.117	-5.775	0	0.0308	139.971	4	0.429	84	1
5jUOG0ae	Need A Favor	Need A Favor	Jelly Roll	197400	12/9/2022	0.0123	0.504	0.771	1.89E-05	6	0.078	-4.231	1	0.0319	157.985	4	0.555	40	0
3OHfY25t	Kill Bill	SOS	SZA	153946	12/9/2022	0.0543	0.644	0.728	0.169	8	0.161	-5.75	1	0.0351	88.993	4	0.43	86	1
6W9l02gR	pretty isnâ€™t pretty	GUTS	Olivia Rodrigo	199422	9/8/2023	0.00394	0.554	0.854	1.78E-06	2	0.0798	-4.338	1	0.0561	123.926	4	0.594	86	1
02Zkkf2zN	Anti-Hero	Midnights (3am Edition)	Taylor Swift	200690	10/22/2022	0.133	0.638	0.634	1.23E-06	4	0.152	-6.582	1	0.0457	96.953	4	0.519	71	0
6wf7Yu7c	What Was I Made For? [From The	What Was I Made For? [From	Billie Eilish	222369	7/13/2023	0.959	0.444	0.0911	1.05E-06	0	0.098	-17.665	1	0.0307	78.403	4	0.142	96	1
1KUZ33cO	Peaches & Eggplants (feat. 21 Sav	Peaches & Eggplants (feat. 21	Young Nudy	203813	2/27/2023	0.00897	0.911	0.536	3.73E-06	4	0.111	-7.884	0	0.282	146.014	4	0.584	78	1
4pCbJC43j	Hey Driver (feat. The War and Tre	Zach Bryan	Zach Bryan	227466	8/25/2023	0.494	0.583	0.351	0	0	0.368	-5.836	1	0.0306	79.014	4	0.462	86	1

Audio Features

Audio features for a track are provided as numerical values (mostly normalized between 0 and 1) which are useful for analysis, but especially as features when fitting them to our random forest model.

	acousticness	danceability	energy	insturmentalness	key	liveness	loudness	mode	spechiness	tempo	valence
0	0.169	0.511	0.532	0.000000	5	0.3110	-5.745	1	0.0560	137.827	0.322
1	0.269	0.868	0.538	0.000003	5	0.0901	-8.603	1	0.1740	99.968	0.732
2	0.554	0.429	0.453	0.000002	0	0.1020	-7.746	1	0.0459	77.639	0.155
3	0.186	0.712	0.603	0.000000	8	0.1150	-5.520	1	0.0262	97.994	0.670
4	0.117	0.552	0.702	0.000021	9	0.1050	-5.707	1	0.1570	169.994	0.564

Other Audio Features

key integer

The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.

Example value: 9
Range: -1 - 11

time_signature integer

An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".

Example value: 4
Range: 3 - 7

Other Audio Features

mode integer

Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

Example value: `0`

valence number [float]

A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Example value: `0.428`

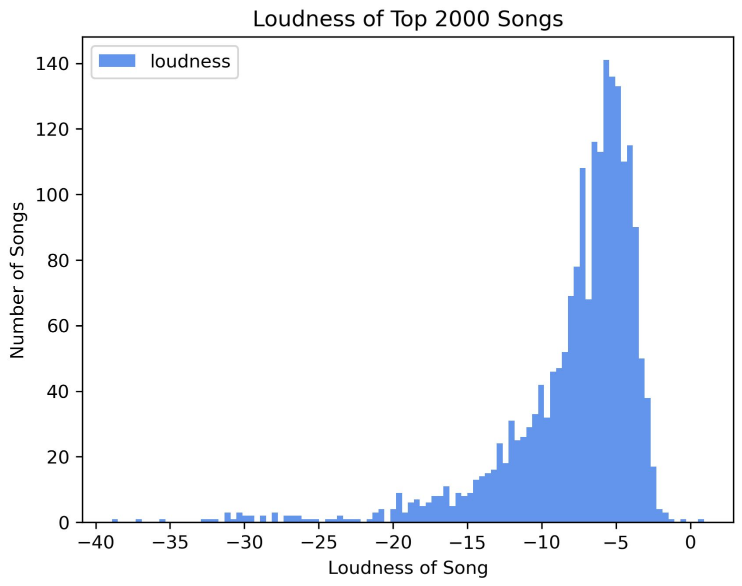
Range: `0` - `1`

Audio Feature Importance

Using Random Forest's feature importance function, we can see which of these features are most important to our model. The four most important features being loudness, instrumentalness, acousticness, and duration.

```
[0.10101179 0.11060855 0.07104946 0.09981881 0.23400324 0.00166388
 0.00825408 0.27048543 0.00041343 0.07717698 0.0085983  0.01691605]
Index(['duration(ms)', 'acousticness', 'danceability', 'energy',
       'insturmentalness', 'key', 'liveness', 'loudness', 'mode', 'spechiness',
       'tempo', 'valence'],
      dtype='object')
```

Audio Feature: Loudness



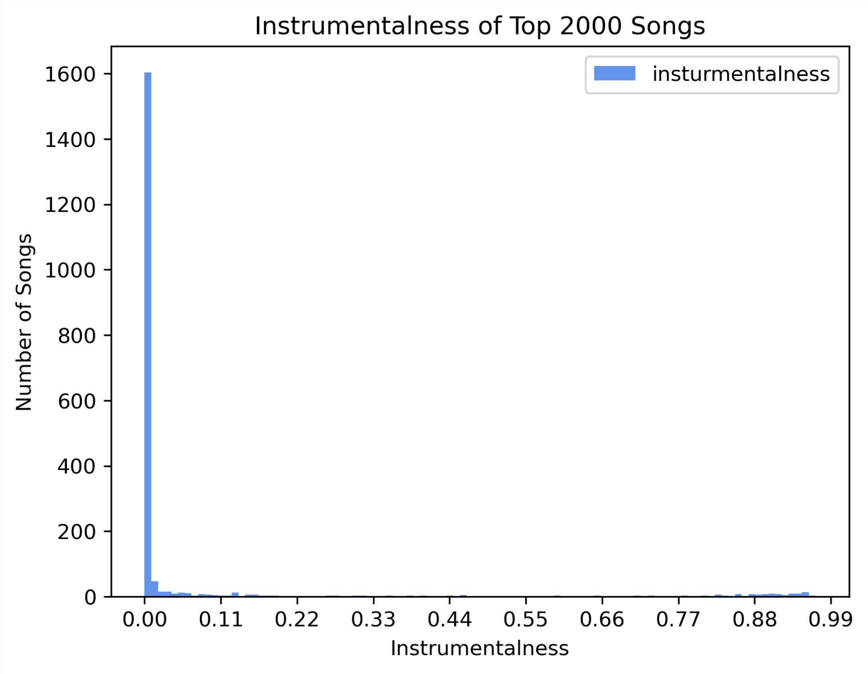
loudness number [float]

The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.

Example value: `-5.883`

- The loudness of the song is the most important feature for the model
- 80% of the songs are between -10 and 0
- Around -30db is a very soft piano melody

Audio Feature: Instrumentalness



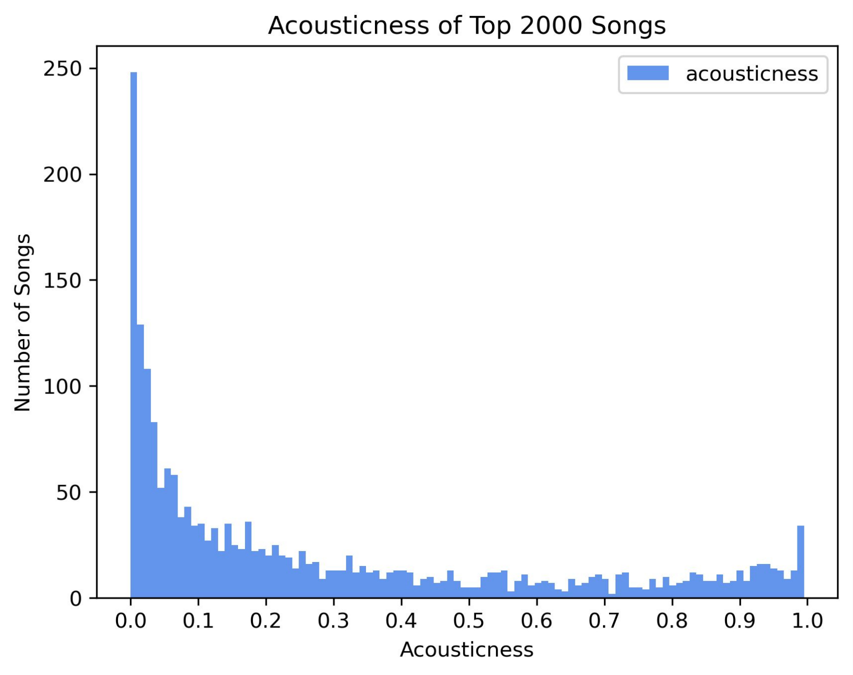
instrumentalness number [float]

Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

Example value: `0.00686`

- Instrumentalness is our 2nd most important feature for the model
- Over 90% of entries are between 0 and 0.1
- There are very few tracks between .2 and .75, so there are few outliers

Audio Feature: Acousticness



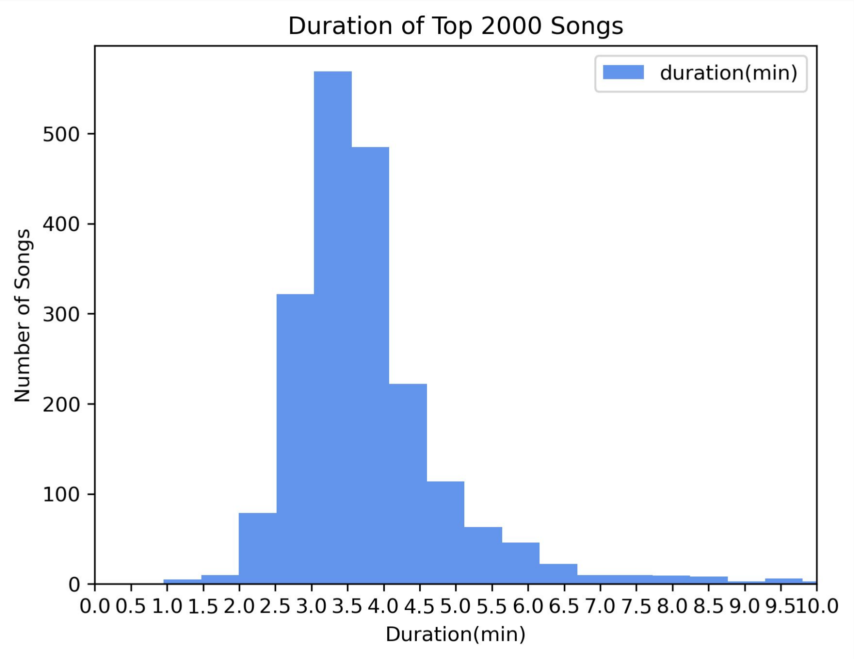
acousticness number [float]

A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

Example value: `0.00242`
Range: `0` - `1`

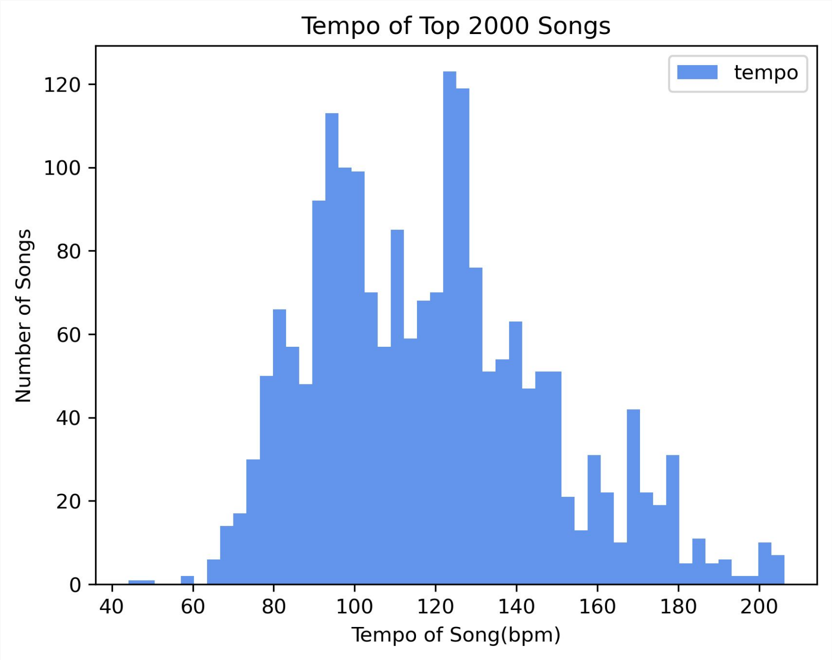
- Acousticness is our 3rd most important feature
- Acousticness represents whether and instrument was used or if the track was made electronically
- Around 40% of our data is between 0 and .1, meaning a lot of the songs are made without traditional instruments

Audio Feature: Duration



- Duration was our 4th most important feature for the model
- 70% of songs are between 2 and 4 minutes

Audio Feature: Tempo



tempo number [float]

The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

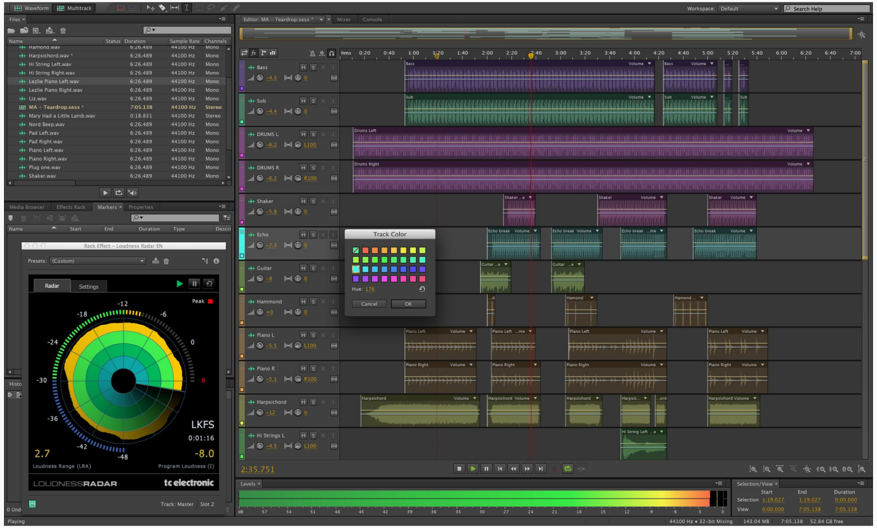
Example value: `118.211`

Rough BPM by Genre:
Downtempo - 70-100
R&B - 60-80
Reggae - 60-90
Chill-out - 90-120
Hip-Hop - 60-115
Jazz & Funk - 120-125
Pop - 100-160
Rock - 110-140
Metal - 100 -160
EDM - 140-180

Best Audio Features

Based on these features, would would the best song features look like?

Loudness: -7.75 db
Instrumentalness: .08 (Singing/Lyrical)
Acousticness: .28 (Electronically Produced)
Duration: 3.7 mins
Tempo: 118 bpm



Our Model

The results of our Random Forest hit prediction model:

Overall Accuracy: 71%

- The model can classify a hit song based on features 76% of the time
- It can only predict if a song will not be popular based on its features 65% of the time
- The biggest struggle for the model is that it classifies non-hits as hits more often that we would like based on the class 0 recall score and the class 1 precision scores
- Given more time, we would like to optimize the model to be more precise in classifying non-hits

--------Confusion Matrix---------

	Predicted 0	Predicted 1
Actual 0	133	109
Actual 1	35	223

Accuracy Score 0.71

Classfication Report

	precision	recall	f1-score	support
0	0.79	0.55	0.65	242
1	0.67	0.86	0.76	258
accuracy			0.71	500
macro avg	0.73	0.71	0.70	500
weighted avg	0.73	0.71	0.70	500

Technology Stack Utilized

```
# Dependencies
import os
import seaborn
import spotipy
import requests
import pandas as pd
import tensorflow as tf
from spotipy.oauth2 import SpotifyOAuth
from matplotlib.ticker import MultipleLocator
from config import SPOTIFY_CLIENT_ID, SPOTIFY_CLIENT_SECRET
from imblearn.over_sampling import RandomOverSampler, SMOTE
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
from sklearn.metrics import accuracy_score
```

Note: Spotipy is a lightweight Python library for the Spotify Web API. Once you've pip installed Spotipy, you must set the Spotipy environment variables.