Now Playing

Popular Song Prediction

Syed Ishmum

Let machine learning do the hard work for you!

1:23 2:36

On repeat

CAPSTONE LIFECYCLE

Building a predictive model

1 - Data Collection and Description

After evaluating the capstone idea, we need to acquire a fitting dataset.

2 - Cleaning and EDA

This is where we focus on data cleaning, looking at correlations and doing visuals to understand our data.

3 - Feature Engineering

This is an important step, where we shape the data according to what types of modeling we will be doing.

4 - Modeling and Irritation

This is where we use the cleaned and curated data and fit them into various modeling systems to get the best potential outcome.

5 - Delivery

This is the stage we will work on our presentations and talk about how our model impacts the relevant fields going forward.

So why popularity prediction is important?

01

It can help new and upcoming artists understand if their songs would be popular

02

Very useful for record labels to estimate a song's potential

03

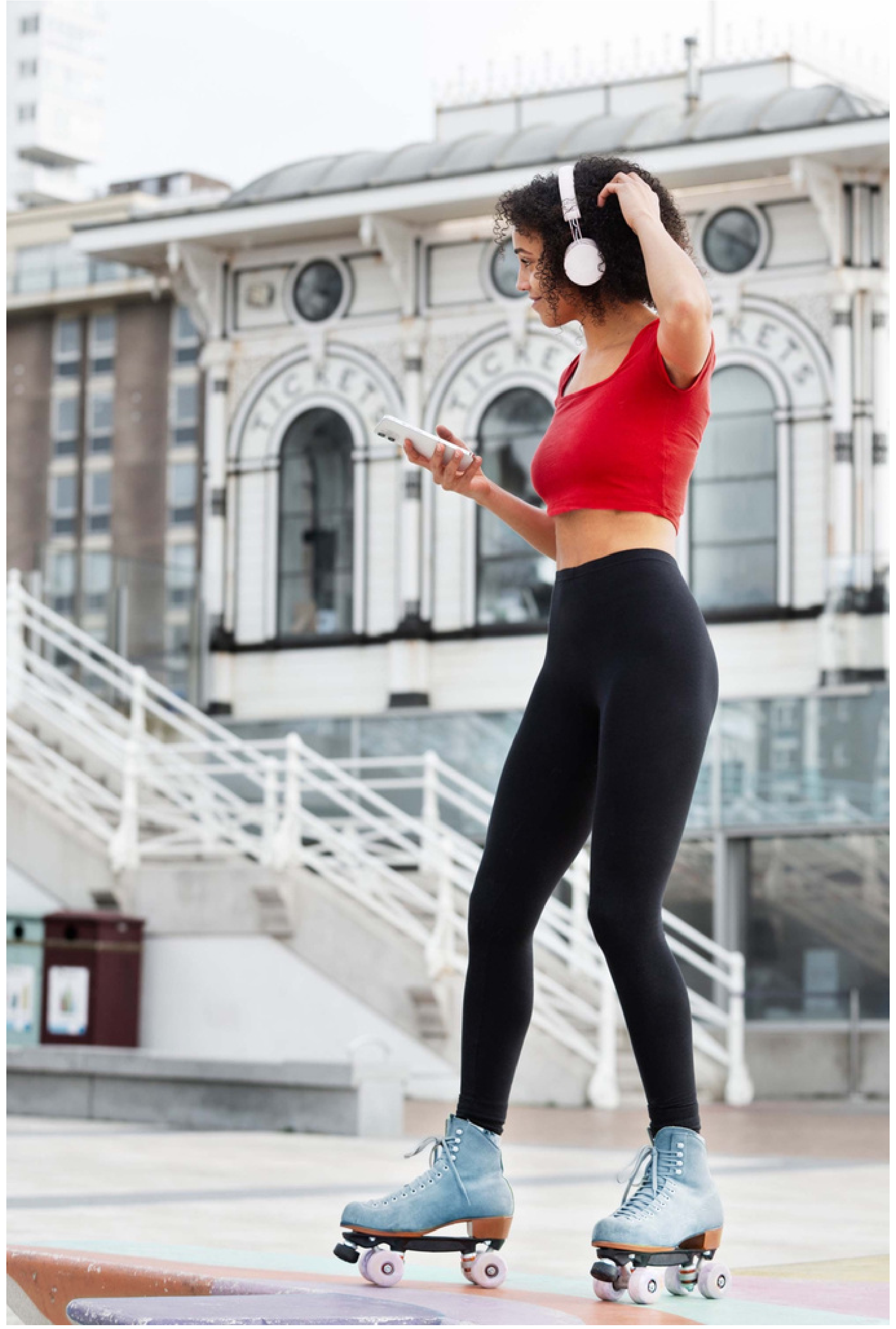
Help us better understand the causation of popularity

04

Can solve many cold-start problems in the future with access to user data

So why Spotify

Spotify has a vast library of music with over 70 million tracks from various genres, reported having over 365 million monthly active users worldwide.



Data Collection

Spotify Track Dataset

The dataset was sourced from Spotify's open API with an Open Database License (ODbL) allowing everyone to access their music collection database.

Data Collection

Spotify Track Dataset

The dataset has 114k rows and 21 columns, with 125 different genres, these aspects make this dataset amazing for training a robust model

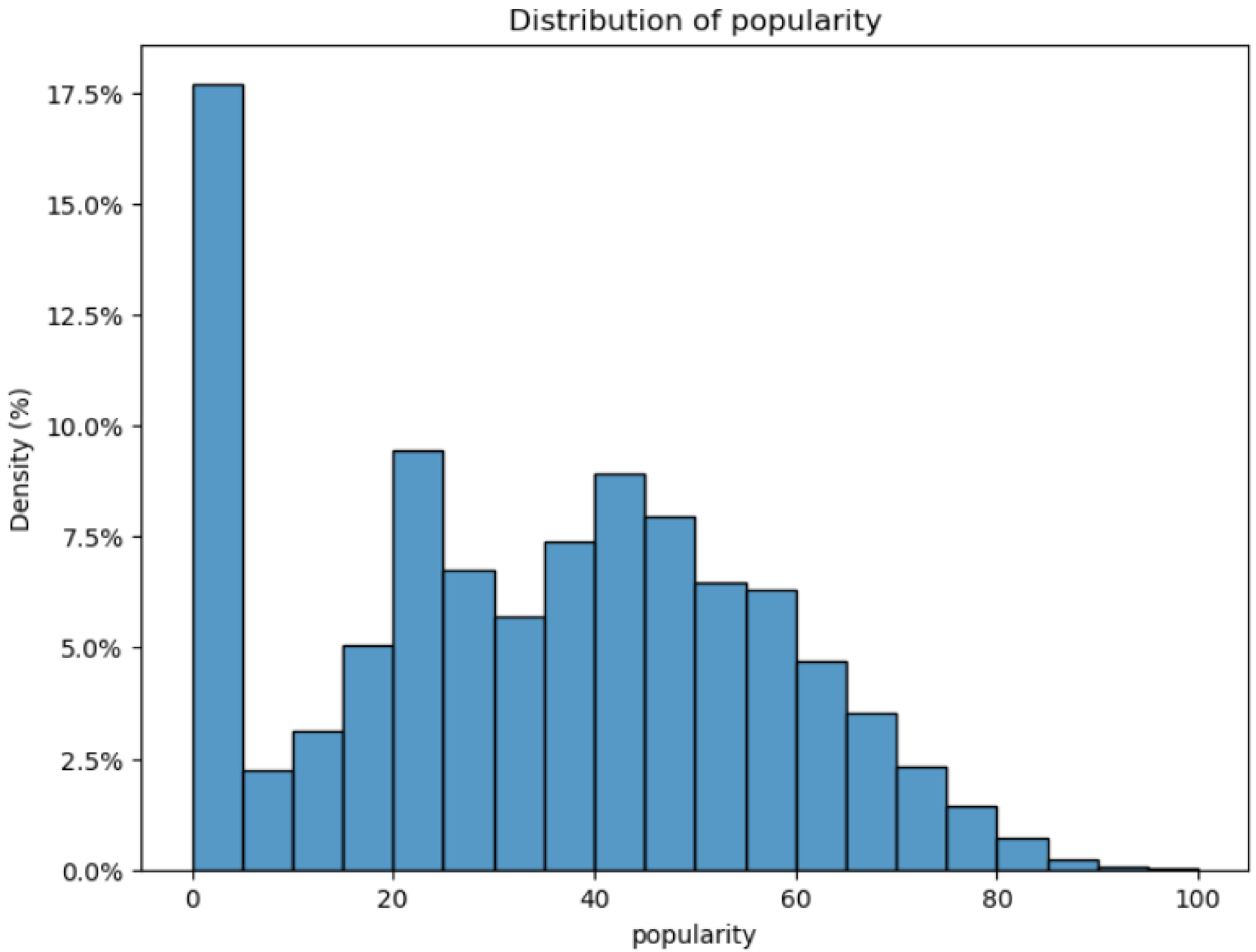
Cleaning & EDA

The dataset was very clean so there was not much to clean, but we did have some amazing visuals of really interesting features of what a song has to offer

EDA Visuals

Popularity

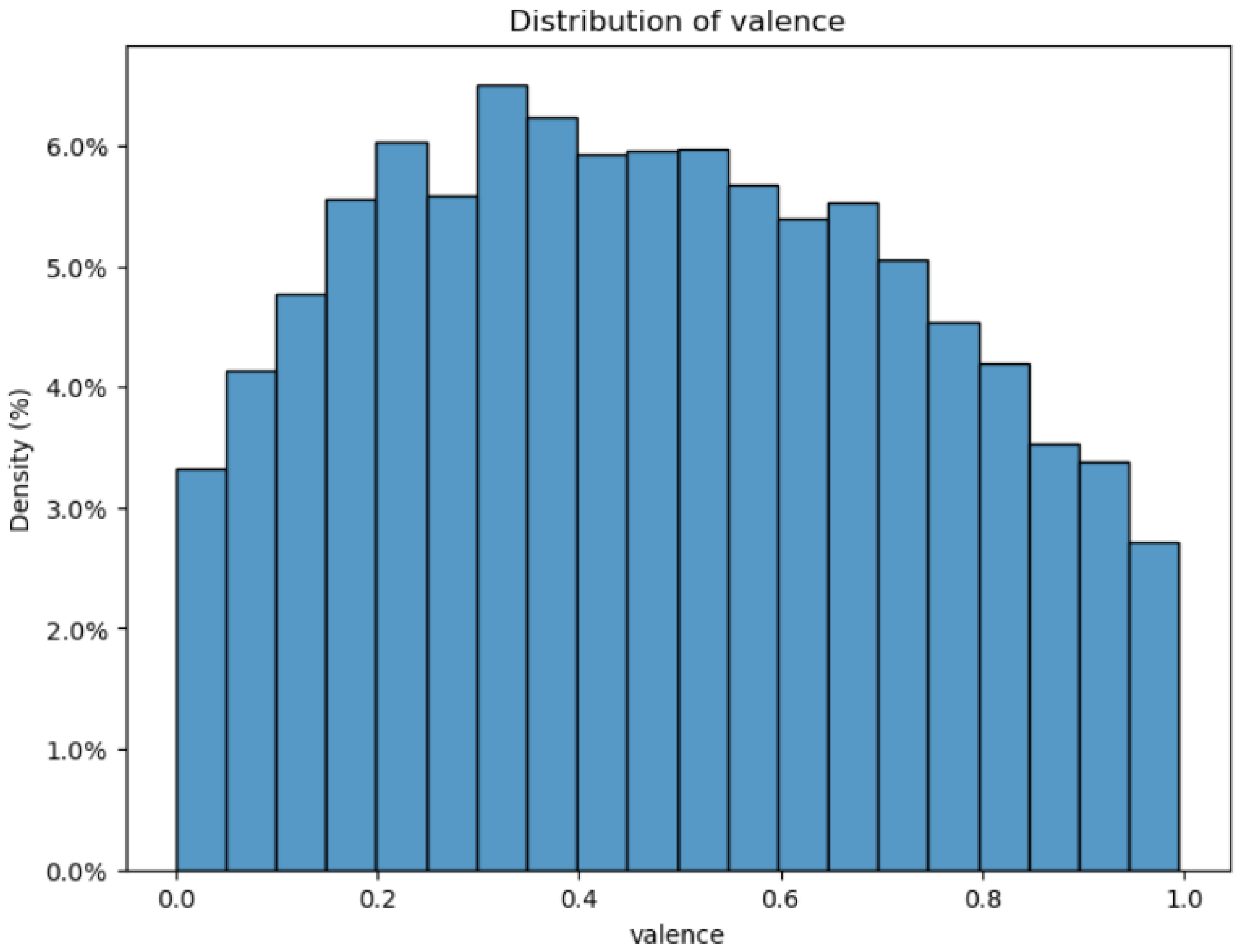
Popularity has a left-skewed distribution, showing there are more songs that are lower on the end of the spectrum



EDA Visuals

Valence

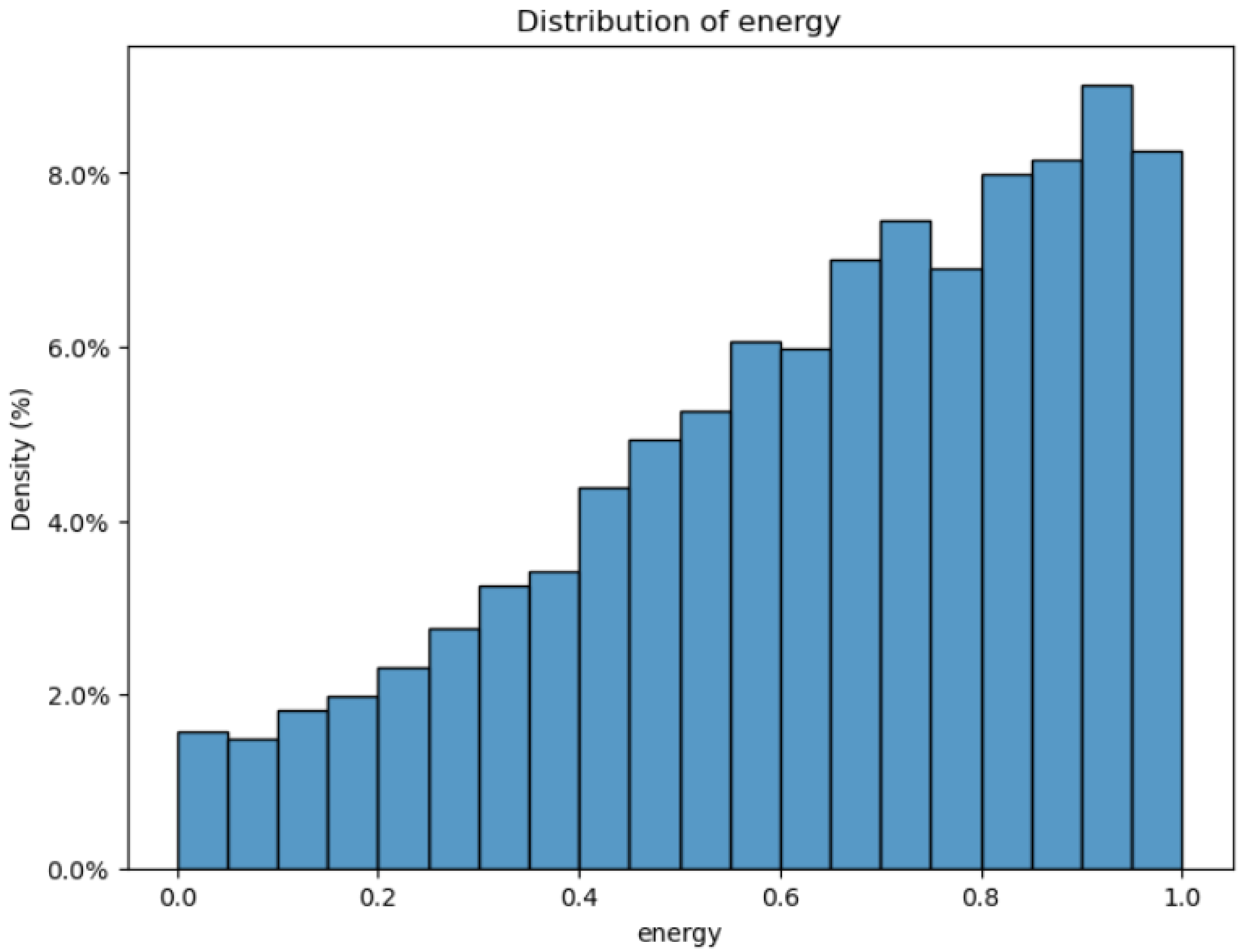
A measure of the musical positiveness conveyed by a track,
has the most normal distribution out of all the numerical columns



EDA Visuals

Energy

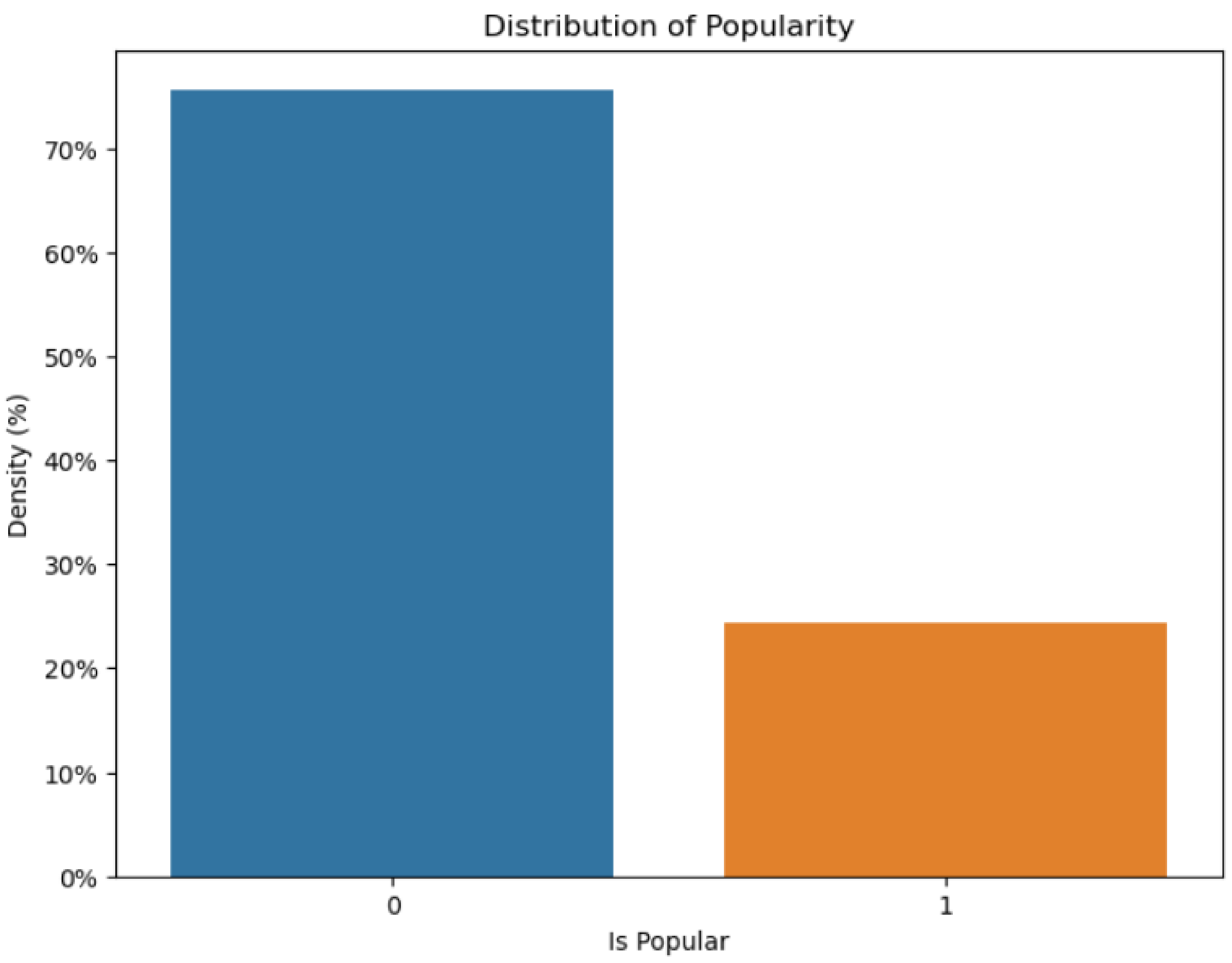
Represents a perceptual measure of intensity and activity, has a positive increasing right-skewed distribution



Feature Engineering

This is where we get the data to be ready so we can fit our model properly

Feature Engineering



Binarization

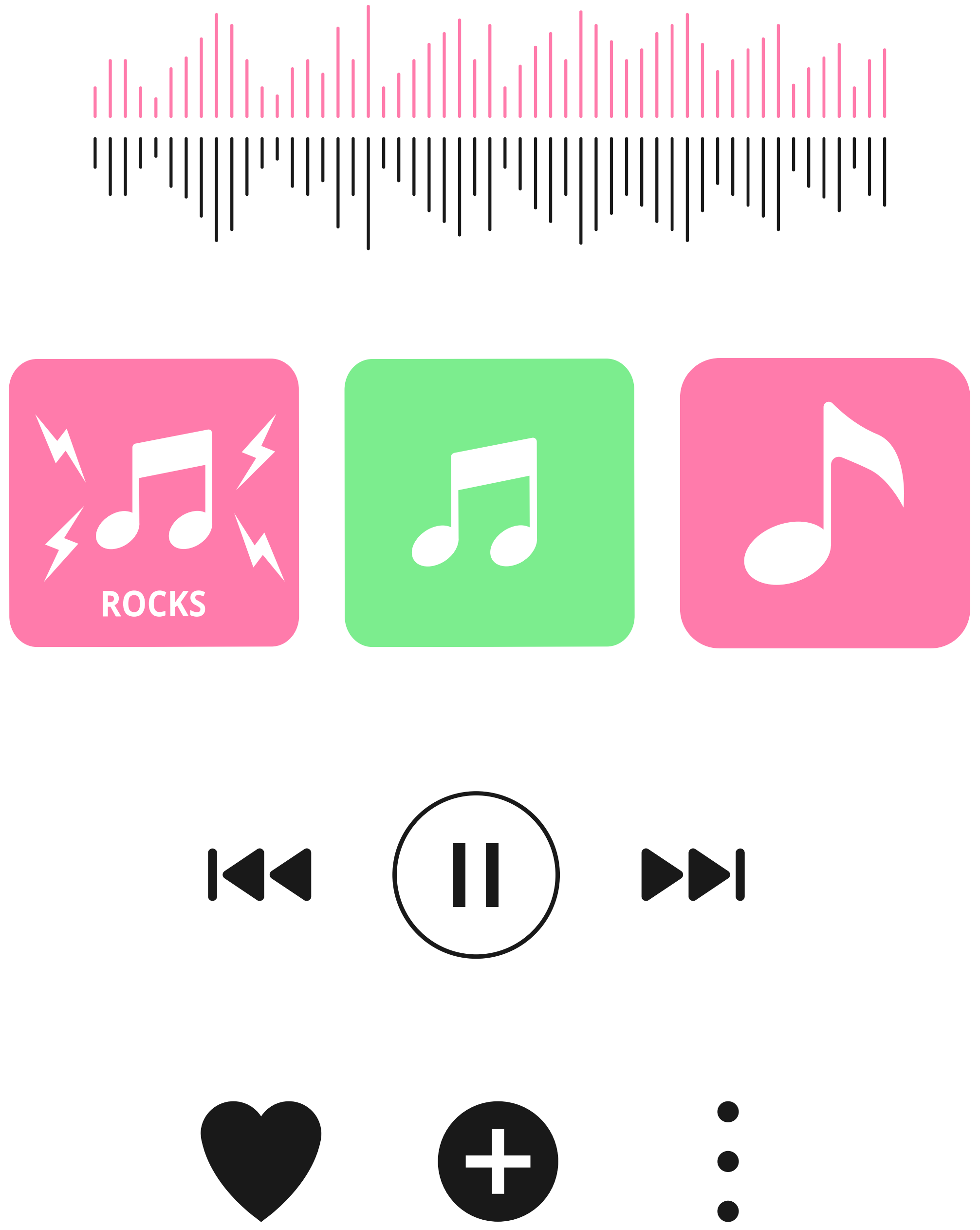
Simplifying our popularity data into 1's and 0's, representing popular or not popular.

Feature Engineering

One-Hot Encoding

This is another way of converting some non-numerical columns to numerical ones by turning them into dummy variables.

Modeling



This is the heart of our problem statement. We are going to model over Logistic Regression, Decision Tree, Random Foresting, and XGBoost!

Modeling

Baseline Modeling

As you can see, not all models performed well, from here on we want to focus on the top 2 accuracy scores

Base Model Comparision

Models	Accuracy	Recall (Is Popular)	Precision (Is Popular)	Recall (Not Popular)	Precision (Not Popular)	F1-Score (Is Popular)	F1-Score (Not Popular)
XGBoost	0.85	0.58	0.73	0.93	0.87	0.65	0.90
Random Forest	0.84	0.49	0.76	0.95	0.85	0.59	0.90
Decision Tree	0.77	0.51	0.52	0.85	0.84	0.51	0.85
Logistic Regression	0.76	0.00	0.00	1.00	0.76	0.00	0.86

Hyper-parameter Tuning

Now we want to focus on tuning our models so we can have better precision that way our models can be way more confident in their prediction of popular songs

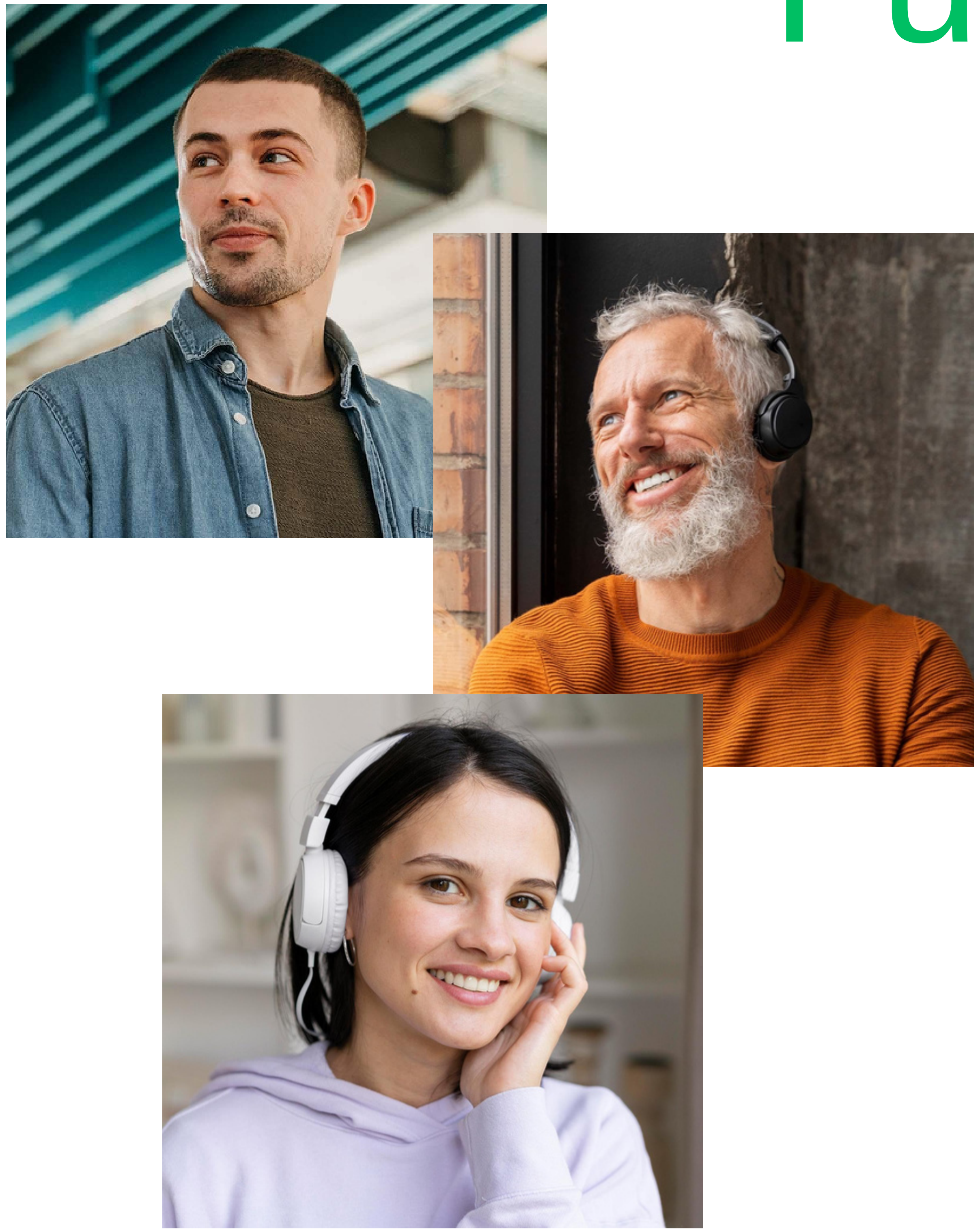
And with a precision-tuned XGBooot we got

86%
Accuracy

80%
Precision

We want high precision in this case because if a song is predicted to be popular by the model, then it is important that it holds up to that expectation

Future Planning



Eventually, use live user data from either new artists or record labels in order to help them identify songs that have a lot of potential and make data-driven decisions.

Feature Engineering

Feedback

We also want to take all the feedback from people using our product and apply better modeling in the future to make everything more robust!

Let's Connect

Syed Ishmum

Data Scientist | Python SQL Tableau | 5+ Years In Optometry Field | Fitness Enthusiast

THANK YOU

<u>Syed Ishmum</u>

- www.linkedin.com/in/syed-ishmum/
- sydishmum@gmail.com
- https://github.com/sai-gh