Project Report | Spotify Hit Predictor

Project Objective

The objective of this project is to predict which songs are classified as a "Hit" and which songs are classified as a "Flop'. If the model is successful, playlist curators and radio program directors can use it as a tool to identify what music has the potential to appeal to a mainstream audience.

Data Collections

The dataset consisted of 6400 songs that were collected using Spotify's. I analyzed each class of song (Hit or Flop) using the following 15 audio features:

i. Danceability

ii. Energy

iii. Loudness

iv. Speechiness

v. Acousticess

vi. Instrumentalness

vii. Liveness

viii. Valence

ix. Tempo

x. Duration

xi. Chorus Hit

xii. Sections

xiii. Key

xiv. Mode

xv. Time Signature

Target Variable consists of two evenly balanced classes:

Hit (has a value of 1)

I. A track is defined as a Hit if it was listed on the Billboard Hot 100 chart at least once between 2010 and 2019

Flop (has value of 0)

A track is defined as a flop if any one of the following criteria is true:

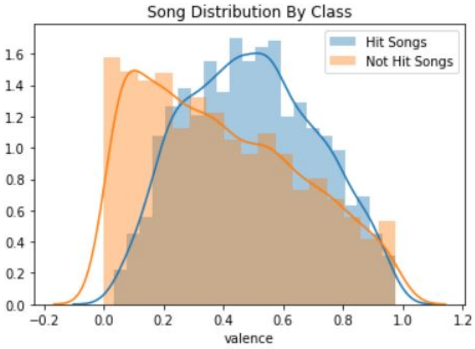
I. The track must not appear in the 'hit' list of that decade.

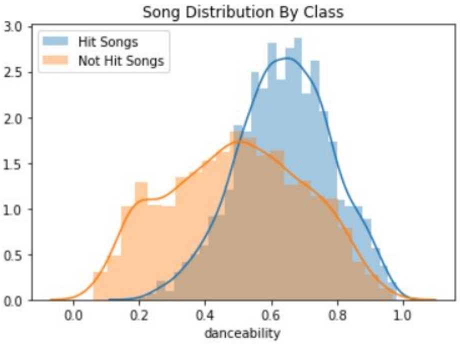
II. The track's artist must not appear in the 'hit' list of that decade.

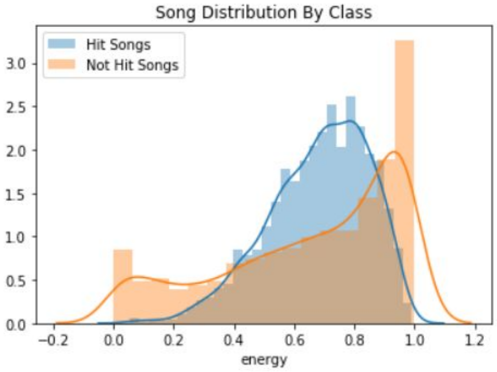
III. The track must belong to a genre that could be considered non-mainstream and/or avant-garde.

IV. The track's genre must not have a song in the 'hit' list.

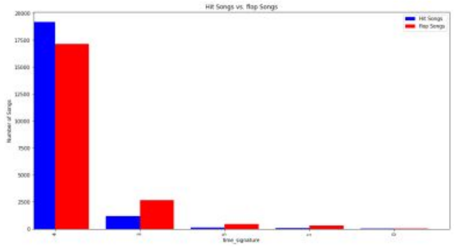
Exploratory Data Analysis

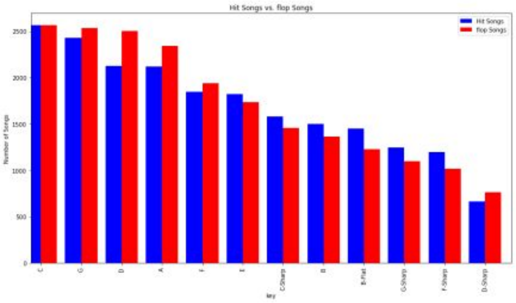


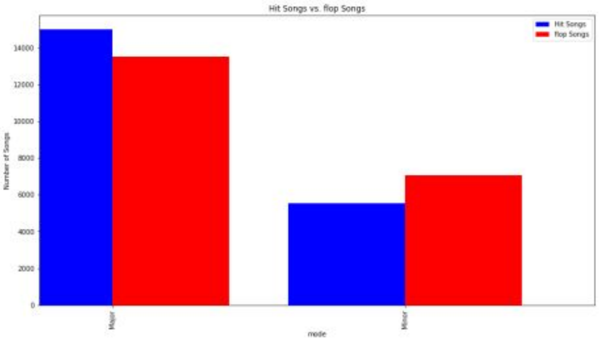




Categorical variable distribution analysis







Data Cleaning

The dataset was already clean when I received it. There were no missing values or outliers that I had to handle. I did look at all the variables to identify the continuous and categorical variables.

i. Categorical features are defined as any column that has less than 12 unique values (Key, Mode, Time Signature).

ii. Continuous features are all the others that have more than 12 unique values

iii. I converted the values for Key from an integer-based system to an music key notation (C, C-sharp, F minor, etc).

Data Exploration

After analyzing the distribution of my continuous variables, I noticed a difference in the variance between the two groups of songs. Just looking at Energy, Danceability, and Valence, the range of Hit songs appear to be much slimmer than non-hit songs. These insights are not groundbreaking, but nonetheless confirm the idea that pop songs are formulaic and have little variance.

Another useful insight came from looking at the categorical variables. For instance, Hit songs tend to favor the key of C and being in a major scale, confirming that pop songs are generally more upbeat and positive.

Modelling

I have trained and tested with 6 different models to evaluate the dataset, models I used are

1. Random forest

2. Decision tree

3. Support vector machine

4. Logistic regression

5. Xgboost

6. Neural network model