Tanisha Batra
Data Scientist

CAPSTONE PROJECT

SONG POPULARITY PREDICTOR

Using statistics, machine learning and data science to make predictions

Start Slide

Tanisha Batra

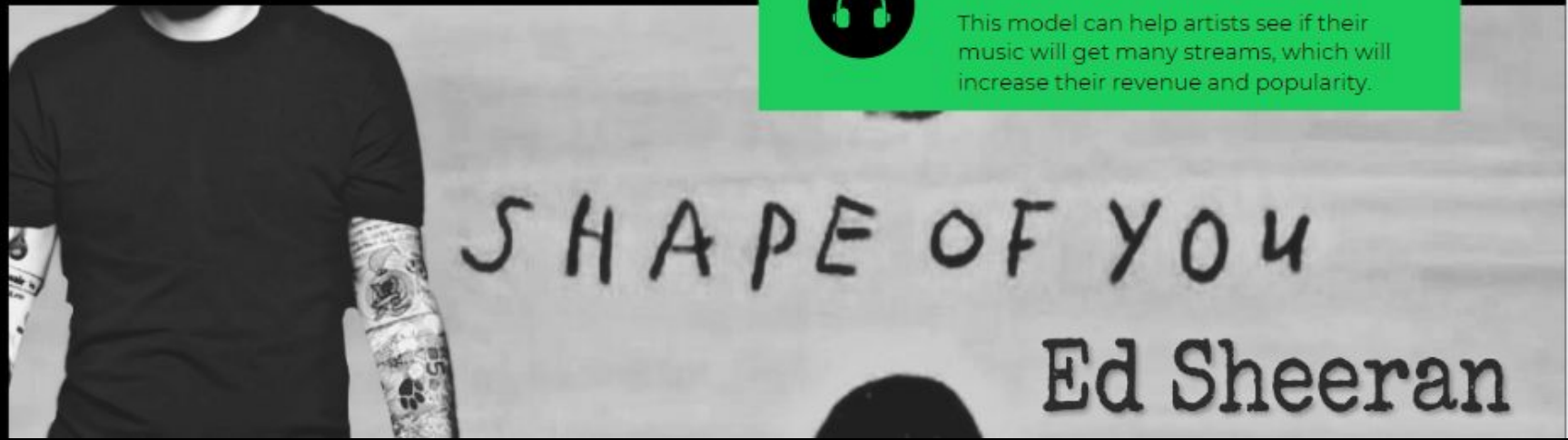
Part 1: Introduction

PROBLEM STATEMENT

Artists and music producers want to know what secret recipe they need in their song to get hits and make money. How popular will their song be?

Value-Add

This model can help artists see if their music will get many streams, which will increase their revenue and popularity.



USER JOURNEY

1

Make a song

2

Use Distrokid's AI bot Dave to extract the features (used by 30% of all musicians on streaming platforms)

3

Use the model to predict how popular the song will be

DATA COLLECTION

This dataset was created by extracting songs from the last decade from Spotify's API.

The dataset contains approximately 18000 rows of songs and their features on Kaggle

Downloaded as a csv
13070 unique values

M YASSER H · UPDATED 5 MONTHS AGO

Song Popularity Dataset

Song Popularity Prediction - Regression Problem

Activity Overview

ACTIVITY STATS

VIEWS
23854

DOWNLOADS
3388

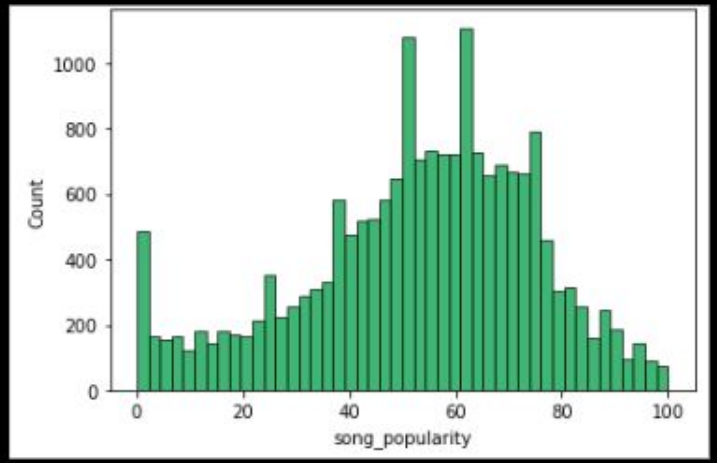
DOWNLOAD PER VIEW RATIO
0.14

TOTAL UNIQUE CONTRIBUTORS
14

TARGET VARIABLE: SONG POPULARITY

Song popularity is a score from 0-100 (least to most popular)
Spotify's internal value to evaluate music.
The higher the score, the more streams you are likely to get.

Frequency of Popularity Scores



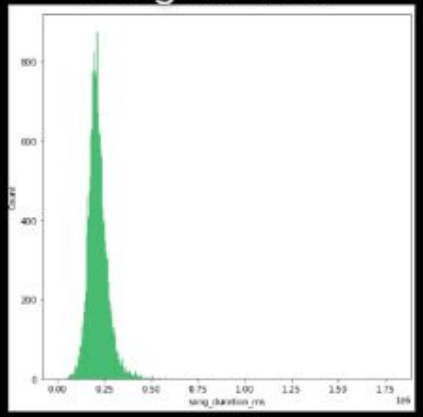
Most songs have a popularity score between 40 and 80

SONG FEATURES

Song duration is the only feature which seems most normally distributed.
The rest have a left or right skew.

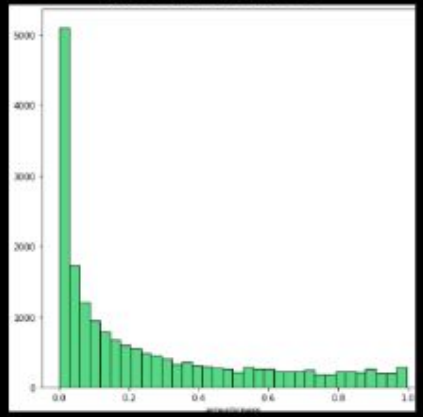
*Features with Gaussian distributions help the model perform better

Song duration



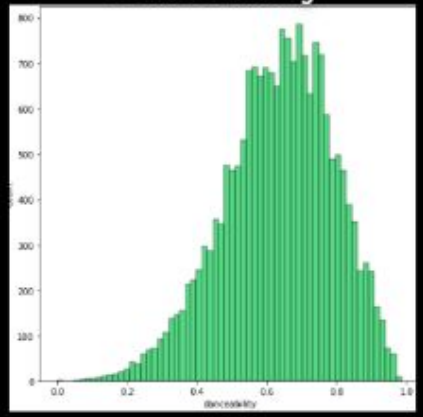
Song length in milliseconds

Acousticness



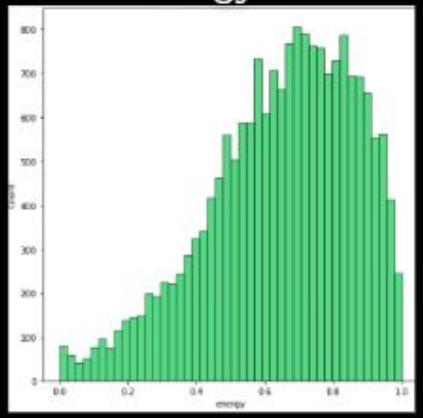
Confidence in acoustics (0-1)

Danceability



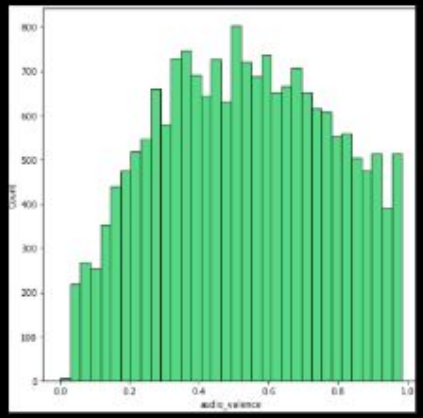
Suitability for dancing (0-1)

Energy



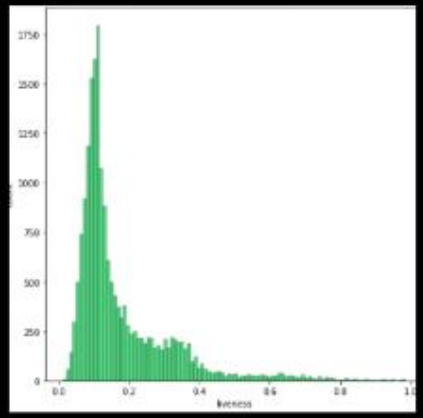
Intensity and activity (0-1)

Audio Valence



Musical positiveness (0-1)

Liveness



Audience present (0-1)

PROCESS

CLEANING, EDA AND SCALING

- Dropped many duplicates
- Feature Engineering
- Scaling the data

REGRESSION MODELING

- Realizing the scores are too low

MODELING FOR BINARY CLASSIFICATION

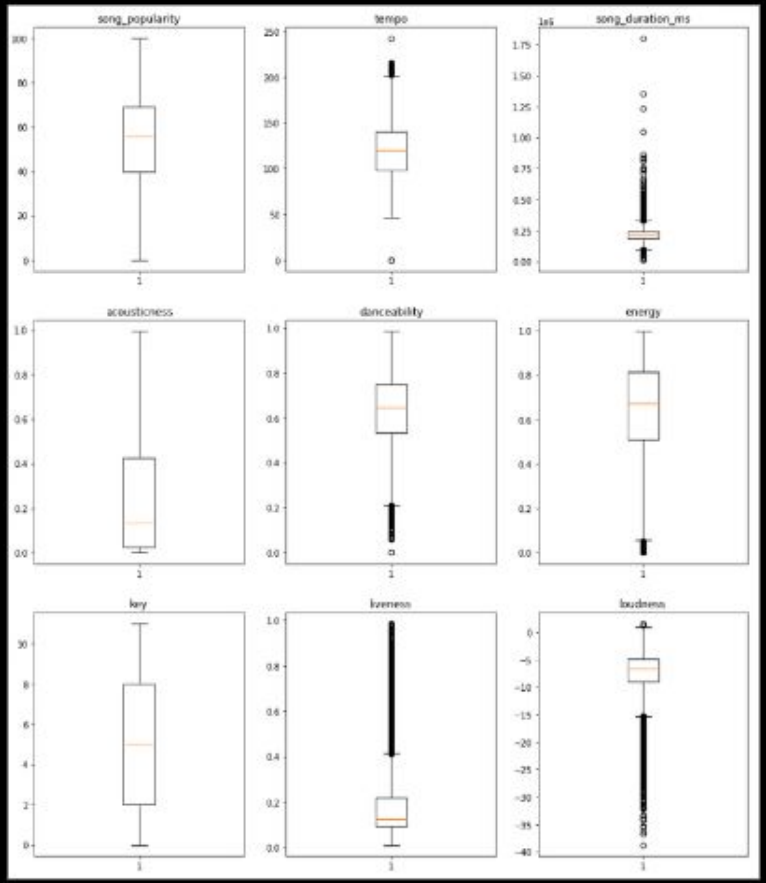
- A score below 52 is a "flop" vs above is a "bop"
- Used OOP to ultimately choose SVM

IMPROVING THE MODEL

- Removing outliter
- Feature engineering
- Hyperaparameter optimization

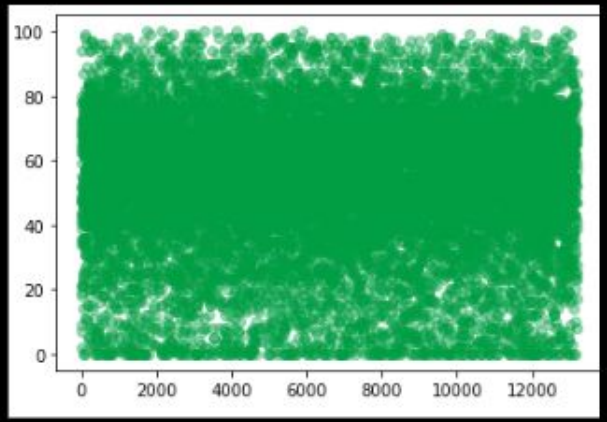
DISTRIBUTIONS

- The dataset contained many outliers
- These were removed to improve the performance of the model



Y VALUE FOR EVERY X

BEFORE SCALING



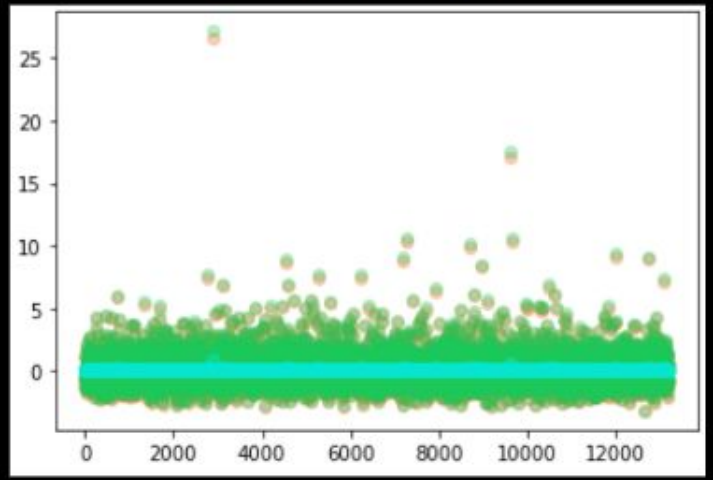
StandardScalar

Robust Scaling

MinMax Scalar

Many of the features are already measured between 0-1

AFTER SCALING



LOW SCORES ACHIEVED WITH REGRESSION MODELS

LINEAR REGRESSION

4.67%

POLYNOMIAL REGRESSION (DEGREE OF 2)

8.64%

K-NEAREST NEIGHBOURS (1 NEIGHBOUR)

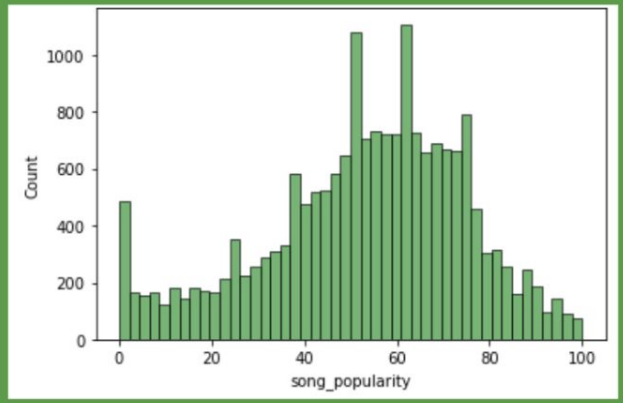
16.11%

RANDOM FOREST REGRESSOR

38.5%

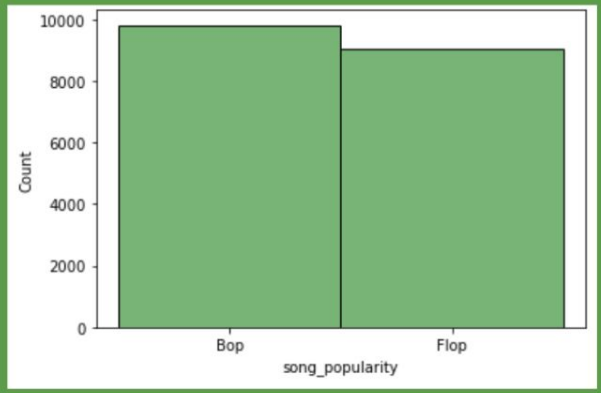
BELOW 52 "FLOP", ABOVE IS A "BOP"

1



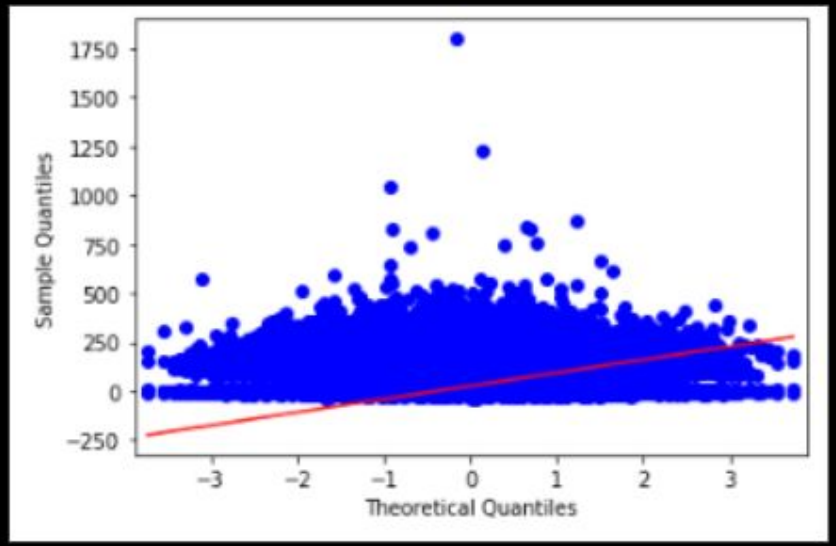
REGRESSION

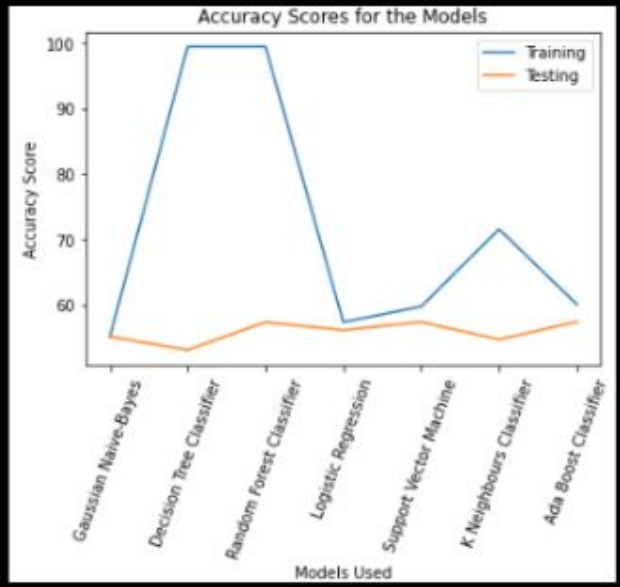
2



CLASSIFICATION

CHOOSING THE RIGHT CLASSIFIER





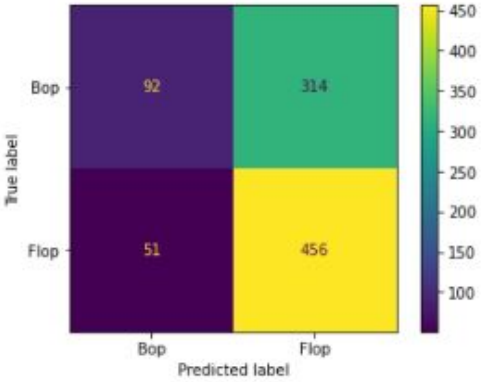
Used object-oriented programming to see which model performed the best.

SVC FINAL MODEL

Testing accuracy score

60.03

#	Score	Entries	Last	Code
1	0.57820	14	6mo	<>



	precision	recall	f1-score	support
Bop	0.64	0.23	0.34	406
Flop	0.59	0.90	0.71	507
accuracy			0.60	913
macro avg	0.62	0.56	0.52	913
weighted avg	0.61	0.60	0.55	913

Tanisha
Consulting

POPULAR

UNPOPULAR

HAPPIER

Marshmellow ft. Bastille

DETAILS

Oliver Heldens ft. Boy Matthew

Popularity Predictor

WHAT'S THE DIFFERENCE?

1. Keep songs around 3.5 minutes
2. Professional studio quality
3. Loud, major key, not too many words

Next Slide

1

Future Enhancements

Adding more dimensions or filtering by genre, artist, release period, country songs were popular in, language and song lyrics. Choosing a non-black-box like model for more interpretability.

2

Extensibility

The processes attempted in this report can be extended to other domains like likelihood of getting a disease, financial services, retail/marketing product success etc.

3

Key Learnings

Real-world data can be messy! It is important to be meticulous in your approach, so you can answer your hypothesis accurately. Sometimes a simpler model can lead to a better accuracy.

KEY TAKEAWAY

The final model understands how different features of a song can be used to predict its popularity.

The model is successful in helpings artist and producers learn whether their song will be a hit or a flop, so they know whether or not to go back to the drawing board.

Open to Data Scientist Positions

TANISHA BATRA

Previous experiences as a data researcher and programming instructor position me as a data scientist with strong communication and technical expertise.

Email
jsm.batra@outlook.com

LinkedIn
https://www.linkedin.com/in/tanisha-batra/

End Slide