AI Academy Final Presentation

Thursday, February 2, 2023

Agenda

1 | Business Understanding

2 | Data Understanding

3 | Modeling

4 | Further Data Analysis

5 | Conclusion

Business Understanding

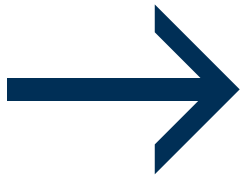
Business Understanding

Below is a quick overview on our business understanding, which answers what problem we are trying to solve for music industry related stakeholders.

ISSUE

Many music related businesses and services rely on music curation & playlist creation, to establish their presence in their respective music scene.

Currently, it is becoming increasingly difficult for emerging music businesses to establish their presence as “taste makers” due to the over saturation of the industry, and inability to put songs in their playlists that have the possibility to become popular.



GOAL

I am looking to see if it is possible to create a song popularity prediction system, to help these businesses choose songs that have the best chance of becoming popular, to apply to their playlists to achieve maximum visibility and exposure.

I will be basing song popularity on a few metrics : Genre, Energy, and Danceability

Hypothesis

Below is a quick overview on our business understanding, which answers what problem we are trying to solve for music industry related stakeholders.

Hypothesis :

I predict that song popularity will be most affected by 3 different measurements: Genre, Danceability, and Energy.
If a song belongs to a highly popular genre (pop), and has high danceability and energy, then the song will be considered popular. Further breakdown of my prediction is below:

Genre: Most songs that are considered popular will be from genres such as Pop, Rap, and Rock

Danceability: Most popular songs tend to be highly danceable

Energy: Most popular songs, tend to have higher energy

Data Preparation & Understanding

Data Preparation & Understanding

Below is a quick overview on the data that I have used for this project, along with some of the important features that come along with it, as well as steps I took for data preparation

1

What data will be collected?

Data for this project will include song data from a Spotify dataset that includes over 232,000 songs from 2019.

2

What is the plan for obtaining the data?

The song data I used for this project came in the form of a direct download, from Kaggle.

I am using two different datasets:

1. Overall Spotify song data containing 232,000 songs
2. Spotify Song data set containing the top 100 songs for the year 2019

3

What are the key features to be used in your model?

The following are the key data features that I will be using in my project and models:

1. Song Popularity
2. Song Genre
3. Danceability
4. Energy

Data Preparation & Understanding (Continued)

Below is a quick overview on the data that I have used for this project, along with some of the important features that come along with it, as well as steps I took for data preparation

1 Data Cleaning: Scrubbing

- Original song dataset contained duplicative information.
  - Example 1: The genre “Children’s Music” shows up twice in the data set. I had to merge the two “Children’s Music” values to achieve consistency
  - Example 2: There were a total of 55,951 duplicative song tracks in the dataset. I had to aggregate all of the duplicate songs into a single row

2 One Hot Encoding

- In order to pre-process most of the categorical features of my dataset for my machine learning models, I had to do some one hot encoding.

3 Outlier Removal

- Although this occurred during my model building, I thought it would be nice to include here
- When viewing song popularity from the top 100 songs on Spotify from 2019, there were a big number of outliers, which would skew my results if I kept in.
- I performed an Interquartile Range method to remove my dataset of any statistical outliers

Modeling

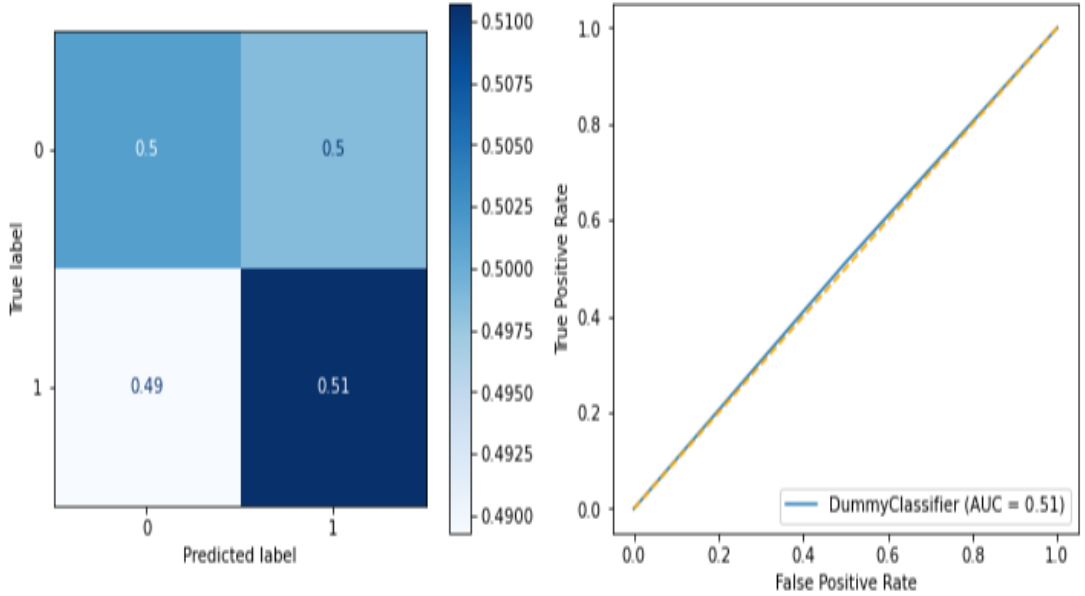
Modeling – Baseline Model (Dummy Classifier)

Below is a view of my baseline machine learning model:

```
CLASSIFICATION REPORT
---------------------------------------
              precision    recall  f1-score   support

           0       0.89      0.50      0.64     47002
           1       0.12      0.51      0.19      6031

    accuracy                           0.50     53033
   macro avg       0.50      0.51      0.42     53033
weighted avg       0.80      0.50      0.59     53033
```



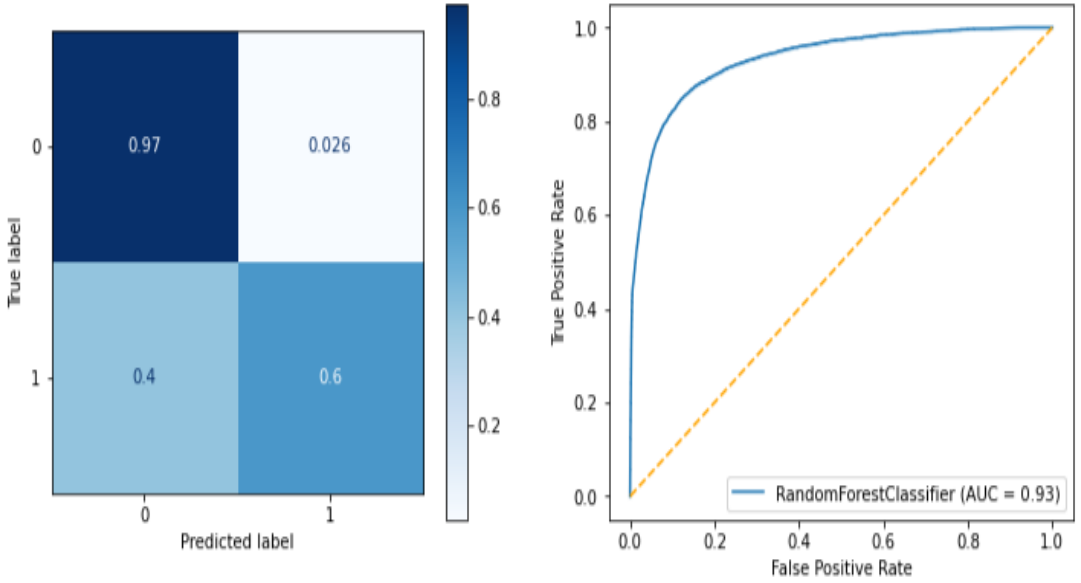
Accuracy	50%
Recall Score	50%

Modeling – Random Forests Classifier

Below is a view of my random forest's classifier model:

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.95	0.97	0.96	47002
1	0.75	0.60	0.66	6031
accuracy			0.93	53033
macro avg	0.85	0.79	0.81	53033
weighted avg	0.93	0.93	0.93	53033



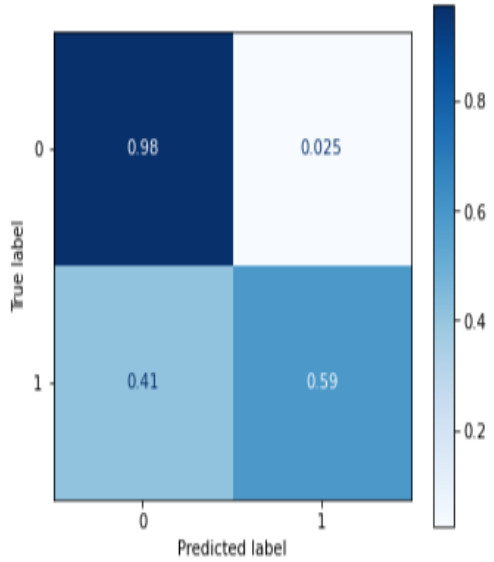
Accuracy	93%
Recall Score	97%

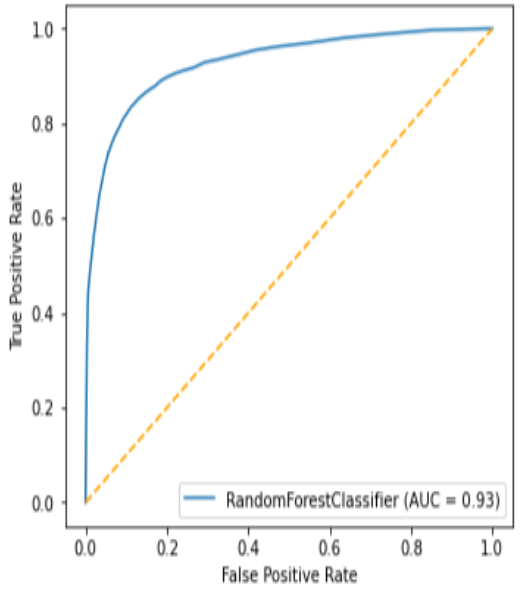
Modeling – Random Forests Classifier Tuned

Below is a view of my random forest's classifier model after hyperparameter tuning:

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.95	0.98	0.96	47002
1	0.75	0.59	0.66	6031
accuracy			0.93	53033
macro avg	0.85	0.78	0.81	53033
weighted avg	0.93	0.93	0.93	53033





Accuracy	93%
Recall Score	98%

Further Data Analysis

Further Data Analysis – Feature Engineering

To help me come to a better conclusion on what is the most popular style of song to pick, I performed some feature engineering on my random forests model

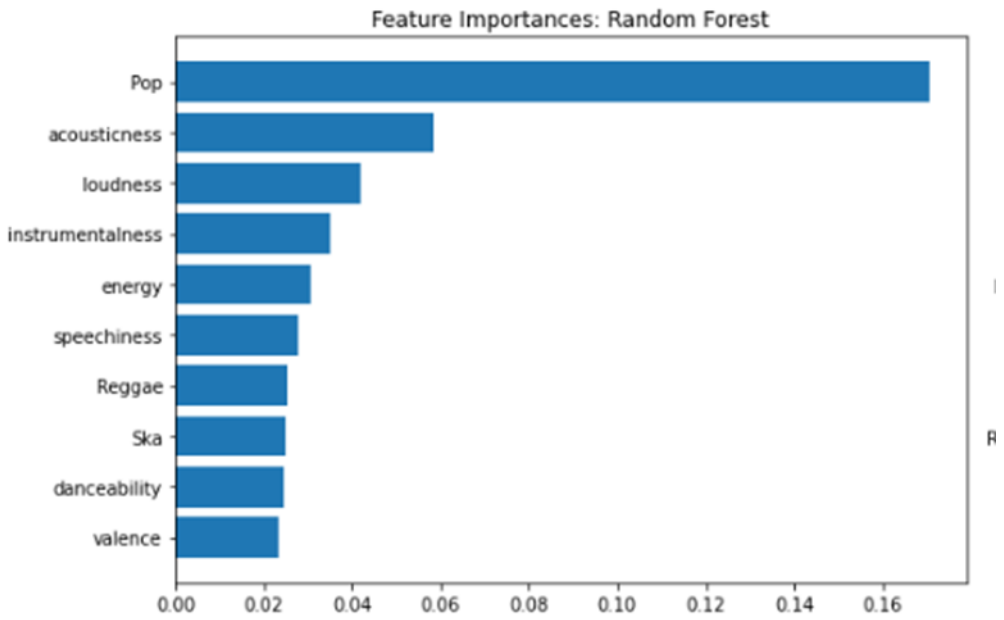
Feature Importance's

Below is a top 8 view of the most important attributes when it comes to song popularity

	RF-Attribute	RF-Importance
0	Pop	0.170450
1	acousticness	0.058465
2	loudness	0.042098
3	instrumentalness	0.035188
4	energy	0.030445
5	speechiness	0.027606
6	Reggae	0.025547
7	Ska	0.025070
8	danceability	0.024540

Feature Importance's: Visualized

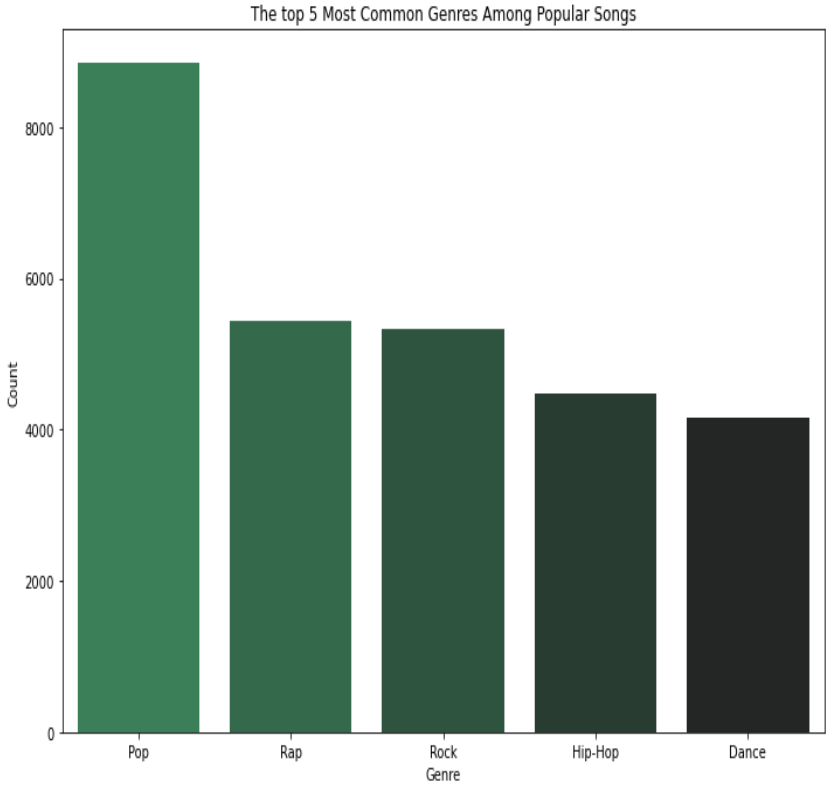
Below is a top 8 view of the most important attributes when it comes to song popularity, visualized



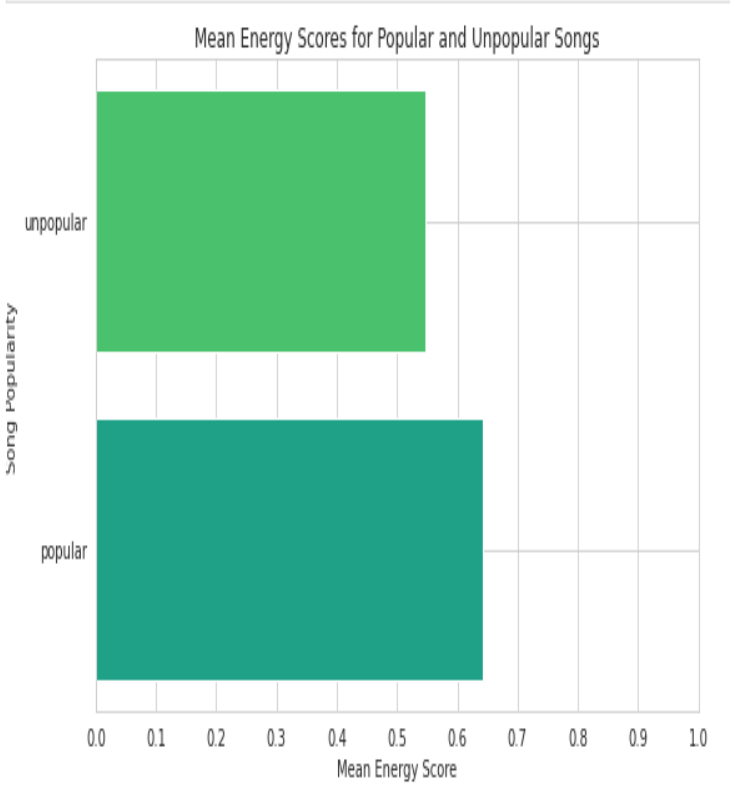
Further Data Analysis (Continued)

To help me come to a better conclusion on what is the most popular style of song to pick, here are some data visualizations

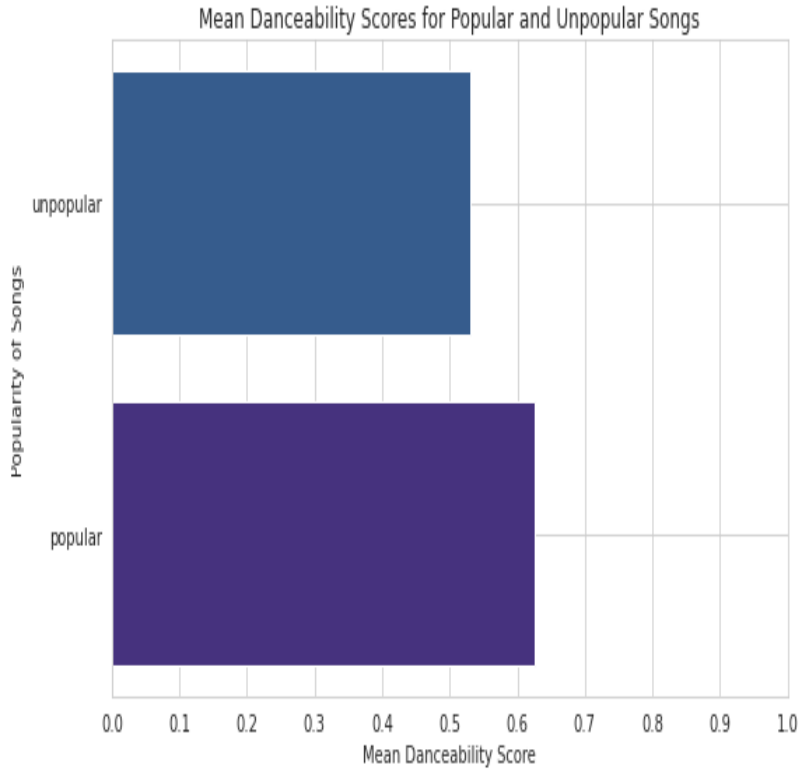
Top 5 Most Common Genres Among Popular Songs



Mean Song Energy Scores for Popular/Unpopular Songs



Mean Danceability Scores for Popular and Unpopular songs



Conclusion

Conclusion

Below are the conclusions I reached based off my modeling and data analysis, and how it will benefit businesses in the music industry

Final Model

- Random Forest Classifier with an accuracy of 93% and a recall score of 98%

Overall Conclusions Based off Model & Data Analysis

- These are the key things to consider when trying to predict a popular song to put into your playlist:
  - Most songs that are considered popular, are from Pop, Rap, Rock, Hip Hop, and Dance genres
  - Most songs that are considered unpopular, are from Children's Music, Comedy, Soundtrack, and Classical genres
  - Most popular songs tend to be more danceable, and have higher energy
- So, with this information in mind, music related businesses can thrive on creating the most popular/listenable playlists, to attract more customers to their businesses!

Limitations

- There may be more factors/data attributes that affect song popularity
  - Chord Progression, Obscurity, Reverb, etc…

Next Steps for Improvements

- The ideal next step for improvement, would be to gather more song data, from the latest year (2022), to have a more accurate assessment of what songs may be popular, since some genres/song types fade in popularity with time.