Predicting song popularity

Artists: Abeda Salsabil, Han Yu, Khevana Patel, Sumaiya Nathani

Overview

- Goal: To create a model that predicts the popularity of a song in June based on audio features
- Research Question
- Data collection and Cleaning
- Analysis and Visualizations
- Our models
- Model optimization
- Summary of findings
- Limitations and future directions

Research question

What audio features make a song popular within current times?

Audio features

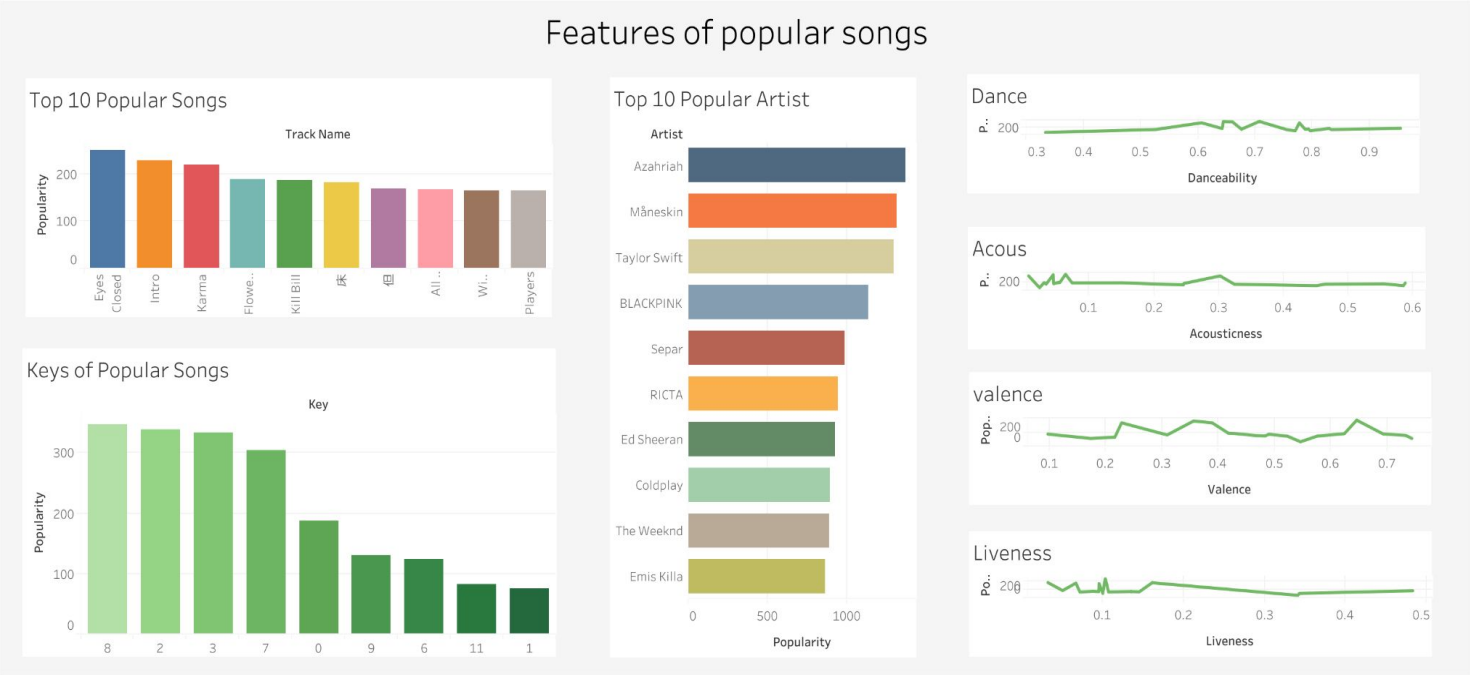
- **Acousticness**: a confidence measure of how acoustic a track is
- **Danceability**: how suitable a track is for dancing depending on tempo, beat strength, rhythm stability and overall regularity
- **Energy**: a perceptual measure of intensity and activity
- **Instrumentalness**: whether a track contains no vocals
- **Key**: values represent the key/pitch a track is in
- **Liveness**: probability of whether an audience is present in a track
- **Loudness**: overall loudness of a track in decibels
- **Speechiness**: presence of spoken words in a track
- **Tempo**: speed/pace of track in beats per minute
- **Valence**: the musical positiveness conveyed by a track

Data collection and cleaning

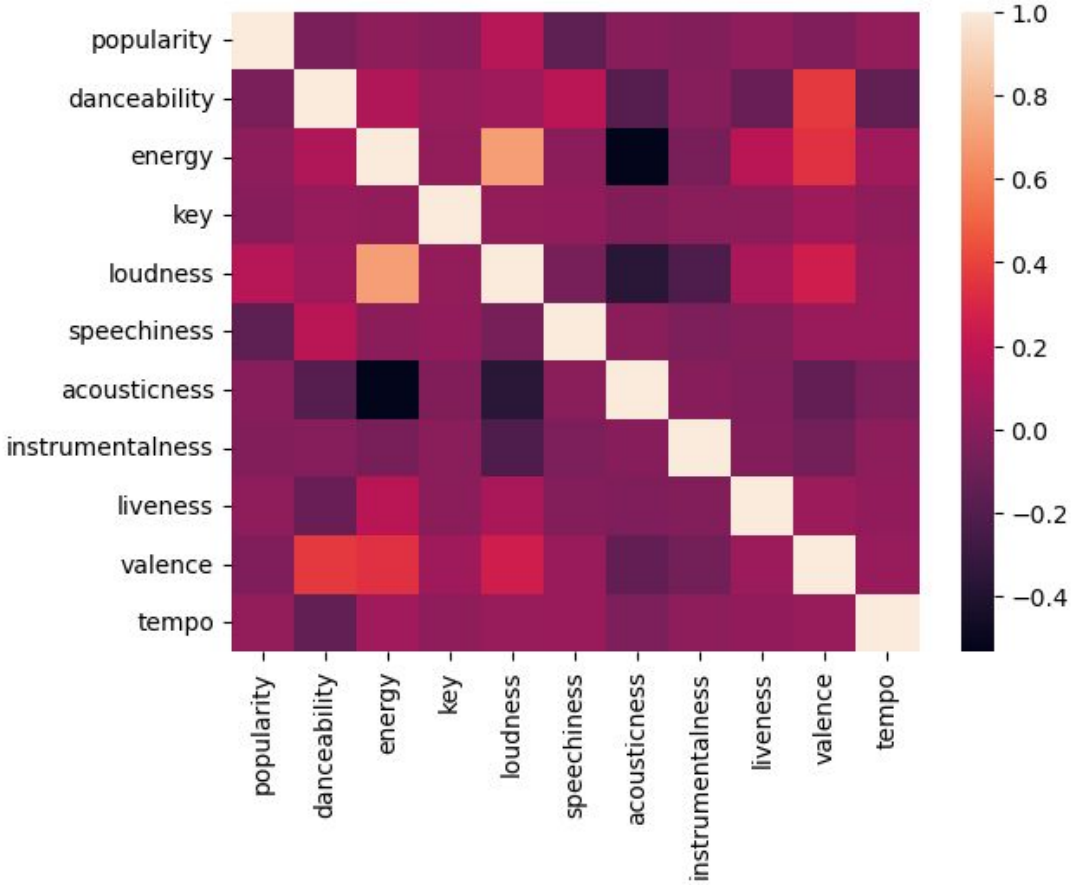
- Dataset obtained from Kaggle
- Small dataset: 3000+ songs from the official Spotify playlists
- Across **63 countries** for the month of **May, 2023**
- Once the data was cleaned, it was read in using Spark

Visualizations and exploratory data analysis

https://public.tableau.com/views/proj4_16866329241620/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link



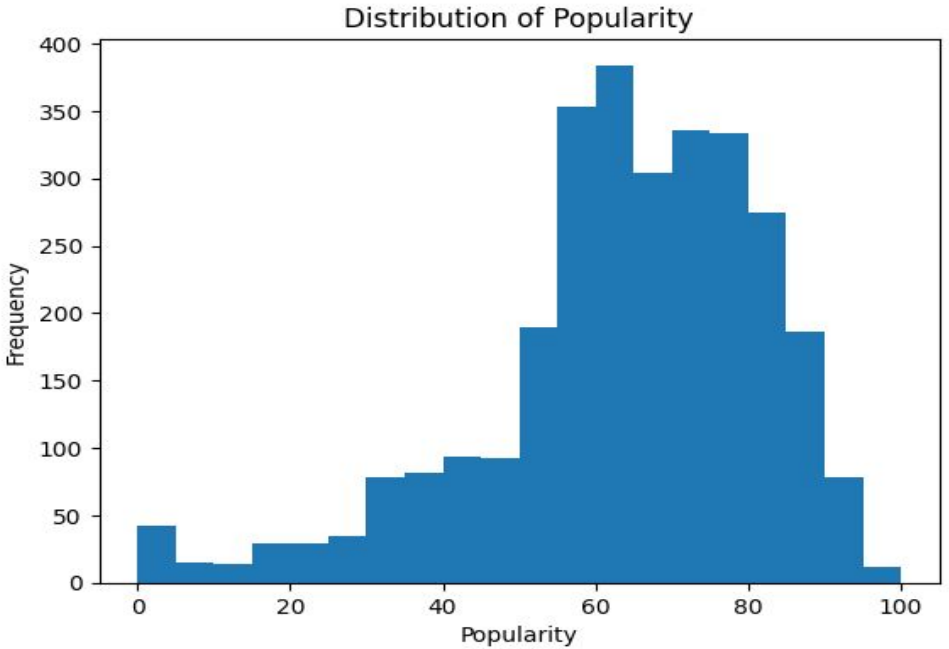
Determining correlation between variables



- Energy and loudness
- No significant correlation between any 2 variables

Data Overview

- Mean Popularity: 63.281999324552515
- Median Popularity: 65.0
- To binarizing the popularity column,considered 65.0 as cutoff value
- Popular_song :>=65
- Not Popular_song:<65



Model 1: Logistic Regression

With original data

- Training Data Score: 60.1%
- Testing Data Score: 61.7%
- Balanced Accuracy Score: 61.4%

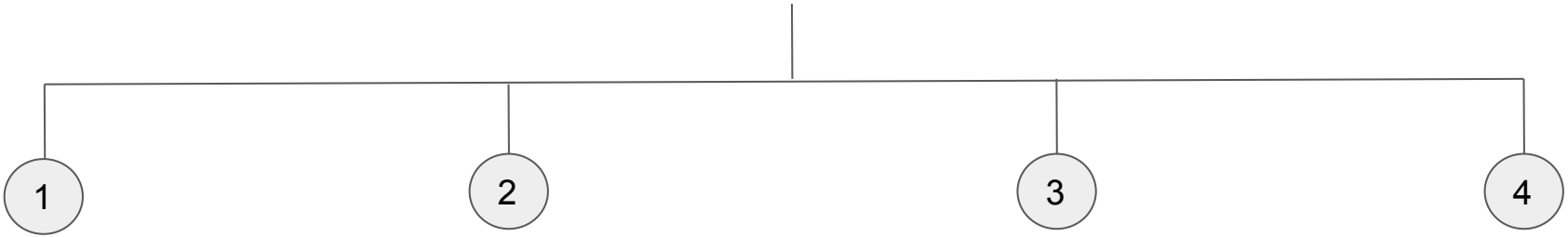
	Precision	Recall	F1-Score	Accuracy
Class 0	0.64	0.50	0.56	61.7%
Class 1	0.60	0.73	0.66	

With resampled data

- Training Data Score: 59.8%
- Testing Data Score: 61.9%
- Balanced Accuracy Score: 61.7%

	Precision	Recall	F1-Score	Accuracy
Class 0	0.63	0.54	0.58	61.9%
Class 1	0.61	0.69	0.65	

Model 2: Neural Networks



Layer 1 = 10 nodes
Activation function = relu

Layer 2 = 20 nodes
Activation function = relu

Loss: 68.1%
Accuracy: 58%

Changed activation function

Layer 1 = 10 nodes
Activation function = relu

Layer 2 = 20 nodes
Activation function = tanh

Loss: 66.3%
Accuracy: **60.6%**

Added hidden layer

Layer 1 = 10 nodes
Activation function = relu

Layer 2 = 20 nodes
Activation function = relu

Layer 3 = 30 nodes
Activation function = relu

Loss: 69.5%
Accuracy: 60%

Changed activation function

Layer 1 = 10 nodes
Activation function = relu

Layer 2 = 20 nodes
Activation function = tanh

Layer 3 = 30 nodes
Activation function = tanh

Loss: 67.1%
Accuracy: 57.9%

Model 3: Random Forest

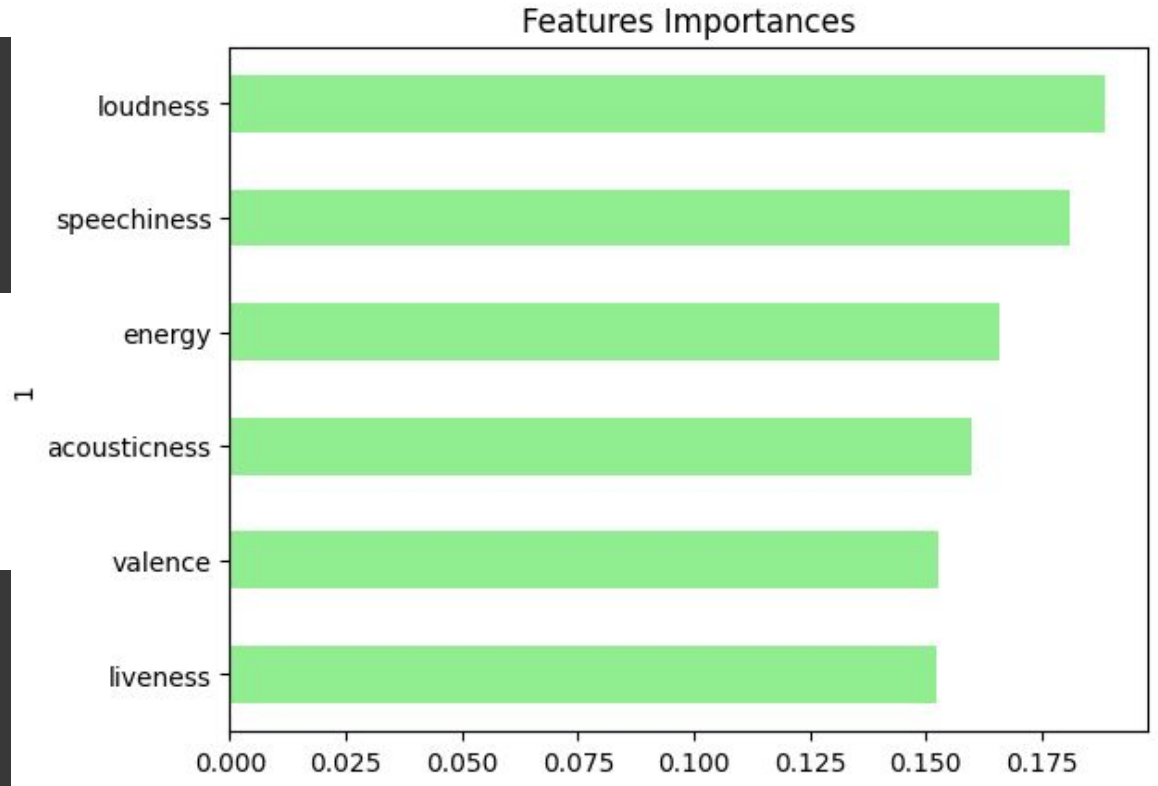
Accuracy: 57.7%

	precision	recall	f1-score	support
0	0.59	0.55	0.57	304
1	0.56	0.61	0.58	289
accuracy			0.58	593
macro avg	0.58	0.58	0.58	593
weighted avg	0.58	0.58	0.58	593

Optimized

Accuracy: 59%

	precision	recall	f1-score	support
0	0.61	0.55	0.58	304
1	0.57	0.63	0.60	289
accuracy			0.59	593
macro avg	0.59	0.59	0.59	593
weighted avg	0.59	0.59	0.59	593



Summary of findings

- Best model from logistic regression
  - Resampled data
  - **Test accuracy of 61.9%**
- Best neural network model
  - 2 hidden layers, relu and tanh activation functions
  - Test accuracy of 60.6%
- Best random forest model
  - Optimised model
  - Test accuracy of 59%

Limitations and future directions

- Use more data (Jan-April spotify data), to increase accuracy
- Considering further measures of popularity such as YouTube views/likes and Billboard chart rankings
- Considering further audio parameters like genre and artists to overcome underfitting

References

Dataset:

- https://www.kaggle.com/datasets/bwandowando/daily-spotify-top-50-of-60-countries?datasetId=2691244

- https://developer.spotify.com/documentation/web-api/reference/get-audio-features

Thank you for watching!