Using Classification Models to Predict Song Popularity

Presented by: Gabby Amparo

Objectives

- The objective of this analysis was to build classifying models that could predict a song's popularity given various audio features in hopes of helping artists gain popularity.

HELLO!

I am Gabby Amparo

Full-Time Data Science Student at Flatiron School

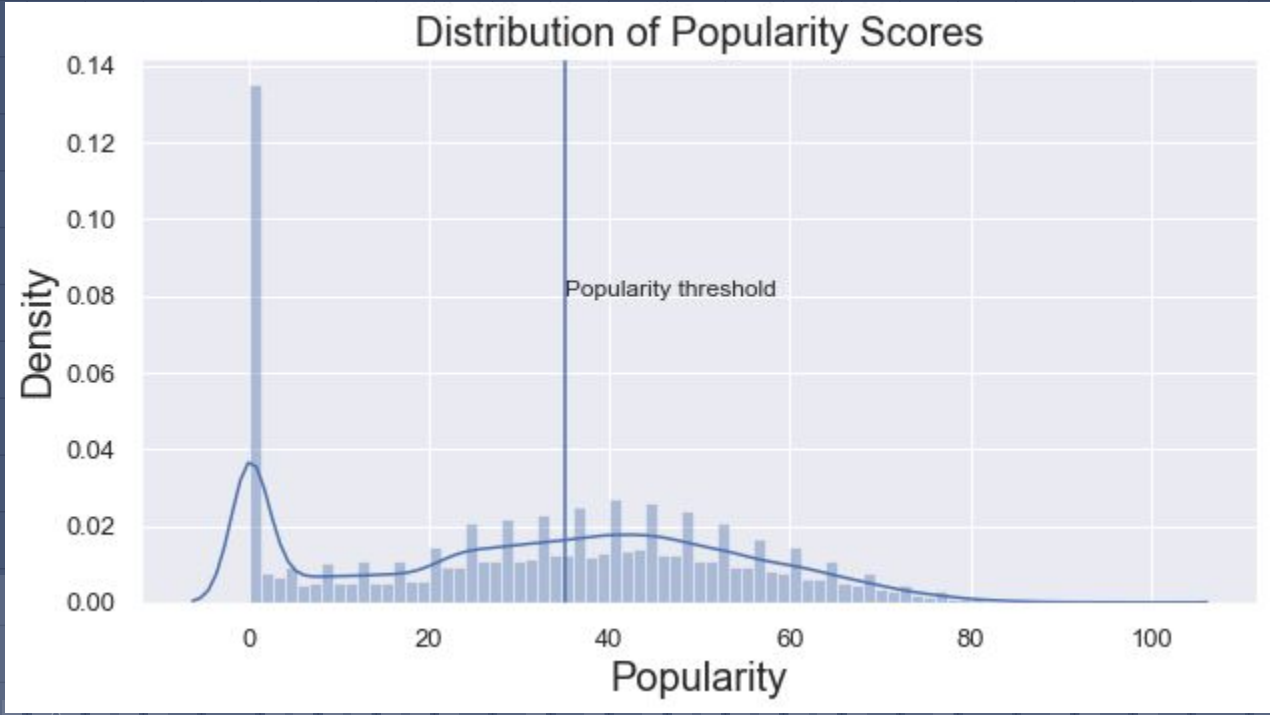
Overview of the Data

- Spotify is a popular music streaming service
- Data was obtained from the Spotify API and a dataset from the kaggle website
- The dataset contained:
  - 170,000+ tracks
  - About 30,000+ artists
  - Tracks dating back to 1921
  - Audio features

Overview of the Data: Summary of Audio Features

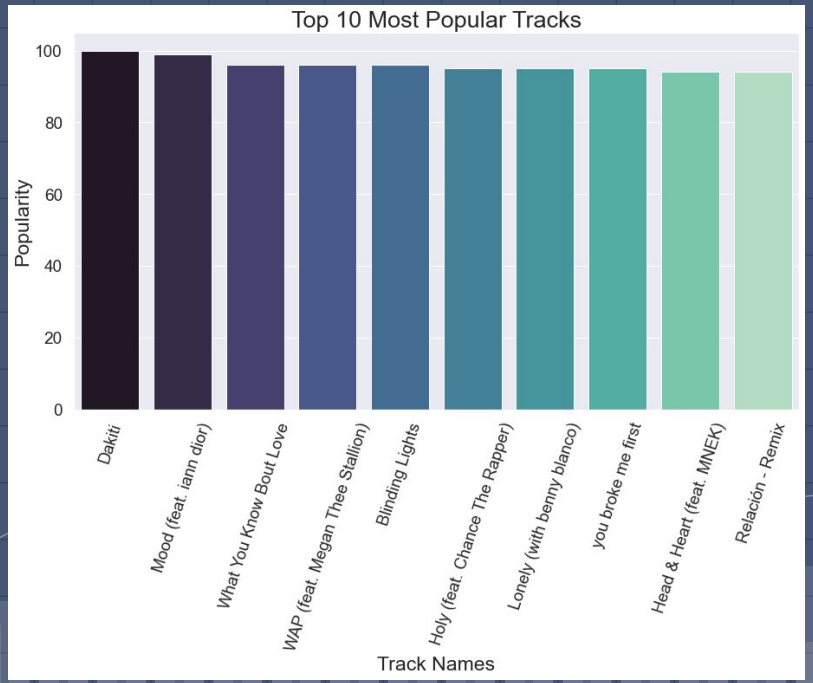
- Acousticness
  - Measure from 0.0 to 1.0 of whether the track is acoustic.
- Danceability
  - Measure from 0.0 to 1.0 how suitable a track is for dancing
- Duration_ms
  - Duration of the track in milliseconds.
- Energy
  - Measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- Valence
  - Measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
- Instrumentalness
  - Measure from 0.0 to 1.0 that predicts whether a track contains no vocals
- Liveness
  - Measure from 0.0 to 1.0 that detects the presence of an audience in the recording
- Loudness
  - The overall loudness of a track in decibels (dB) from -60 to 0
- Speechiness
  - Measure from 0.0 to 1.0 that detects the presence of spoken words in a track
- Key
  - The key the track is in
  - 0 = C, 1 = C♯/D ♭ , 2 = D, and so on.
- Tempo
  - The overall estimated tempo of a track in beats per minute (BPM).

Exploratory Data Analysis

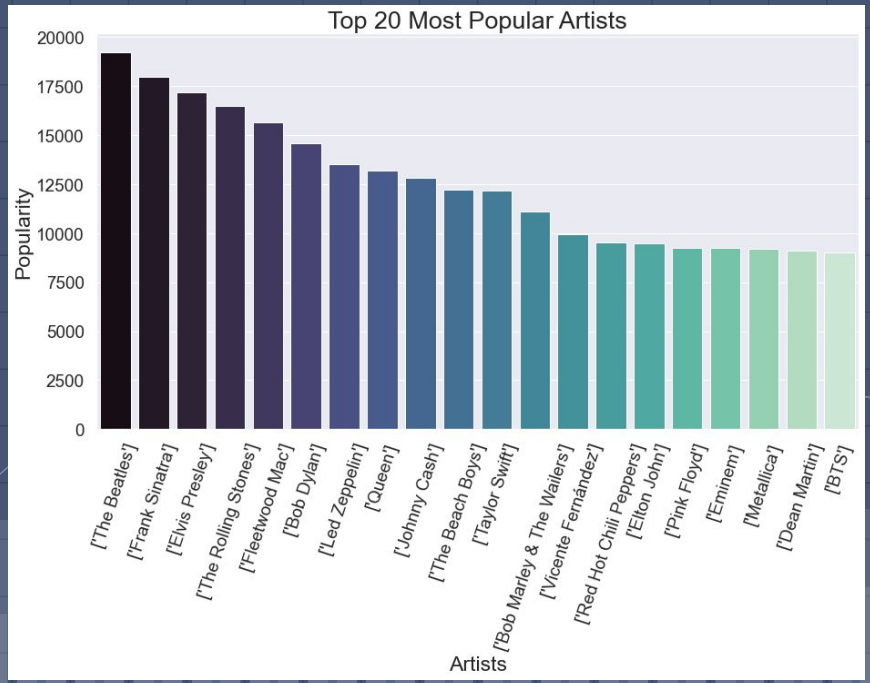


Exploratory Data Analysis

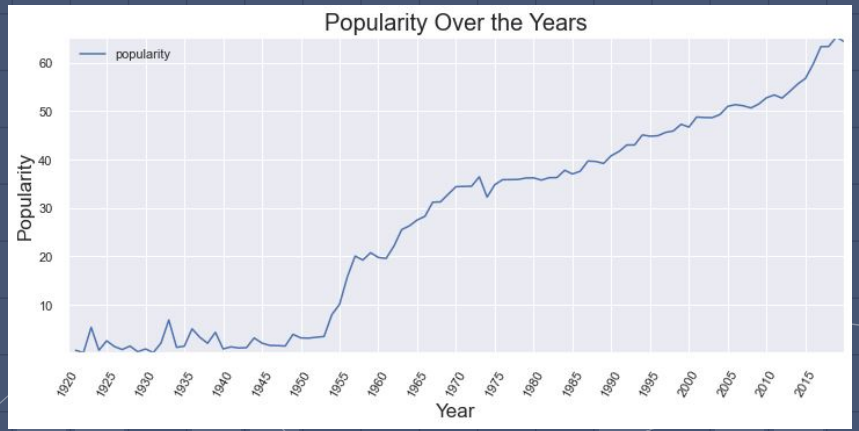
Top 10 Tracks



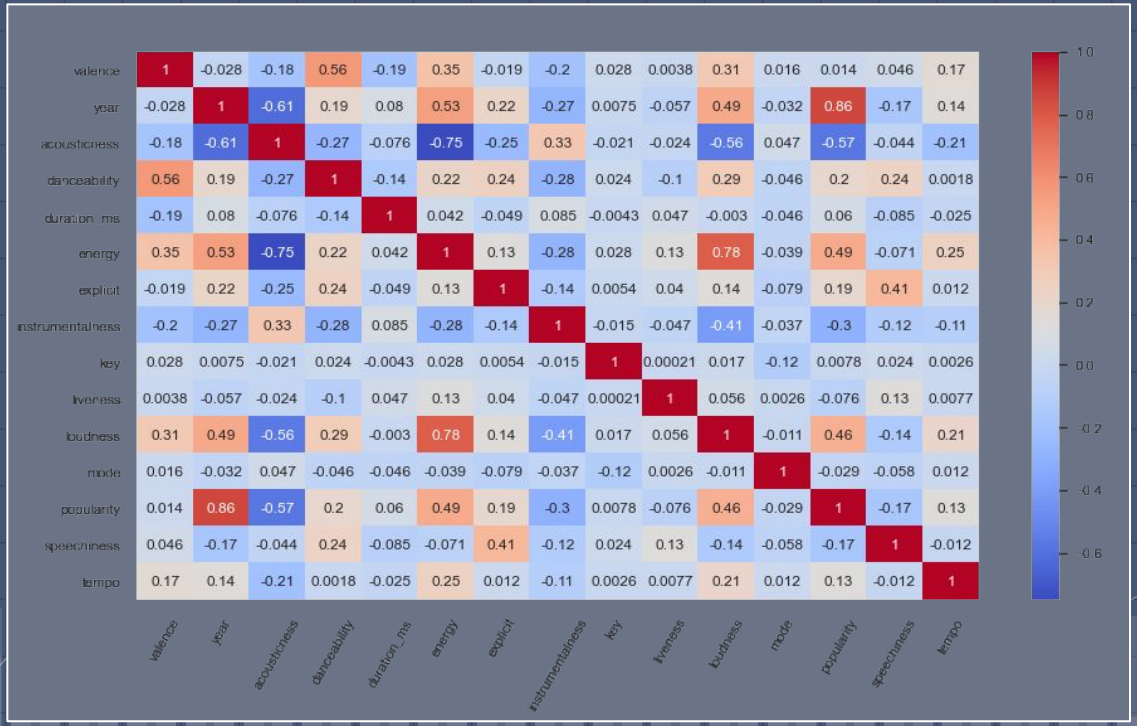
Top 20 Artists



Exploratory Data Analysis



Exploratory Data Analysis



Classification Models

Built Logistic Regression classification models to compare their accuracies

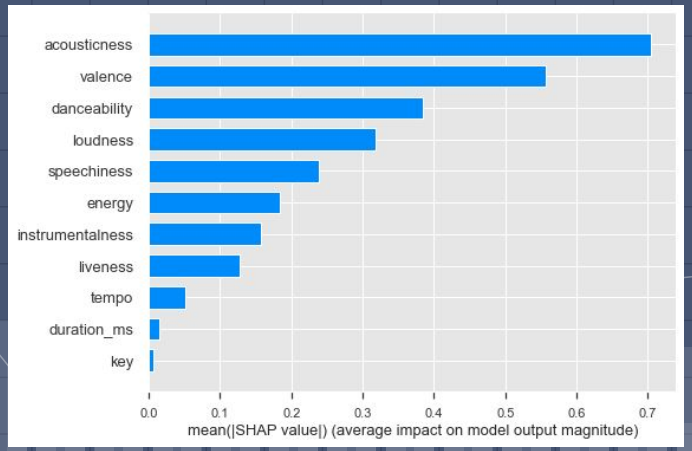
Performed GridSearchCV on each model to get the best parameters

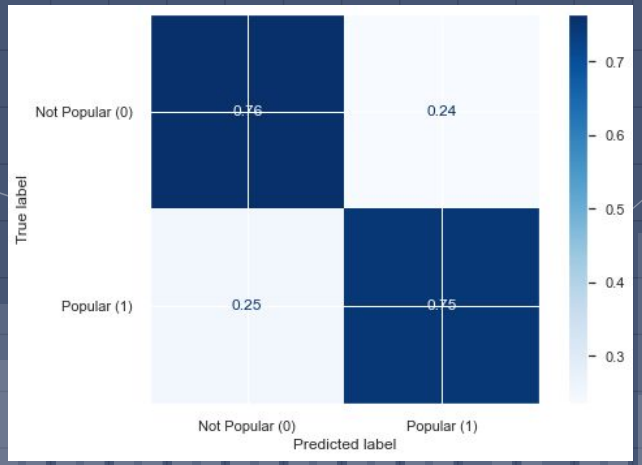
The models included were:

- Baseline LogisticRegression model
- LogisiticRegressionCV model

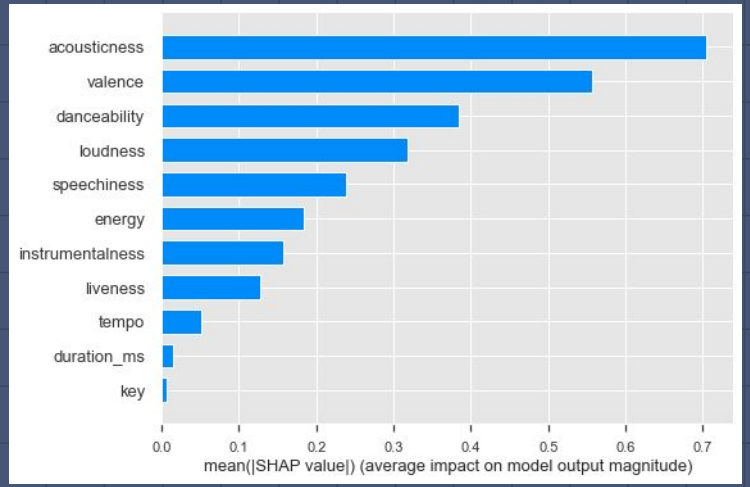
Logistic Regression Model

- The LogisticRegressionCV model performed the best
- The LogisticRegressionCV model had a performance accuracy of 75.7%
- Model has false positive and false negatives present
- Misclassification rate : 0.245





Logistic Regression Model



The Impact of Audio Features on the Model

	coef
acousticness	-0.767210
valence	-0.651481
danceability	0.469579
loudness	0.404460
speechiness	-0.360529
energy	0.214315
instrumentalness	-0.197915
liveness	-0.184105
tempo	0.064652
duration_ms	0.026815
key	0.006954

Direction of Impact

Impactful Audio Features:

- Acousticness
  - Negative correlation
- Valence
  - Negative correlation
- Danceability
  - Positive correlation
- Loudness
  - Positive correlation
- Speechiness
  - Negative correlation

Conclusions

Most Impactful Audio Features on Model

1. Acousticness
2. Valence
3. Danceability
4. Loudness
5. Speechiness

- Accuracy is 75.7%, could improve
- Misclassification rate = 0.245

Recommendations

For artists that want to create popular music I would recommend:

- Create songs with little to no acoustics, more electronic beats
  - Data shows that songs with a high acoustic levels are not popular
- Songs that are sad / gloomy tend to be more popular with listeners
  - Try creating songs about heartbreak or betrayal
- Listeners enjoy music that they can dance to
  - Create songs that have a fast tempo, strong beats, a stable rhythm
- Popular songs have lyrics, but not too many
  - Create songs with simple, repetitive lyrics
- Loud songs are more popular
  - Focus on electro-beats

Future Work

- Looking at the date and time when a song was uploaded to Spotify would improve the models.
- Use the same modeling techniques on a different popularity threshold
- Removing songs from 1920s - late 1940s

THANKS!

Any questions?

You can find me at:

- gabbyamparo97@gmail.com
- @gabbyamparo

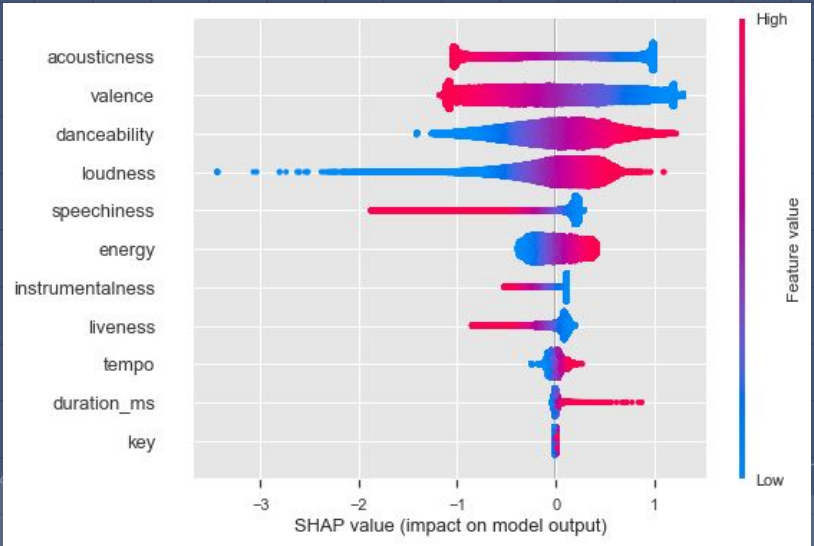
Appendix: Programs and Libraries

The following software libraries were used within Python to conduct data analysis:

- Numpy - for mathematical computation
- Pandas - allows for data organization & analysis
- Matplotlib - for data visualization
- Seaborn - works with Matplotlib to make clean graphics
- Sklearn - machine learning
- Shap - ML model interpretation

Appendix: Shap Summary Plot

Summary plot of Audio Features



The summary plot summarizes the following:

- When the level of acousticness of a track is low, it has a positive shap value and is more likely to be "popular"
- When the level of valence of a track is low, it has a positive shap value and is more likely to be "popular"
- When the level of danceability of a track is high, it has a positive shap value and is more likely to be "popular"
- When the level of loudness of a track is high, it has a positive shap value and is more likely to be "popular"
- When the level of speechiness of a track is high, it has a negative shap value and is less likely to be "popular"