Using Classification Models to Predict Song Popularity

Objectives

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

HELLO!

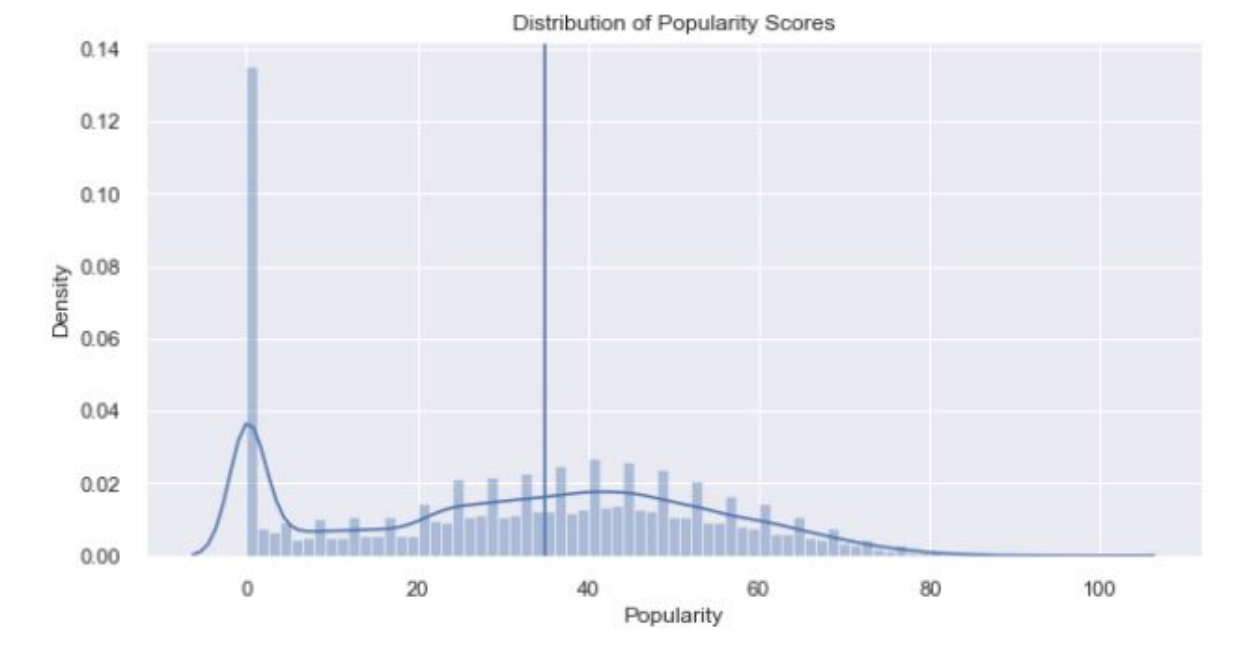
I am Gabby Amparo

Full-Time Data Science Student at Flatiron School

Overview of the Data

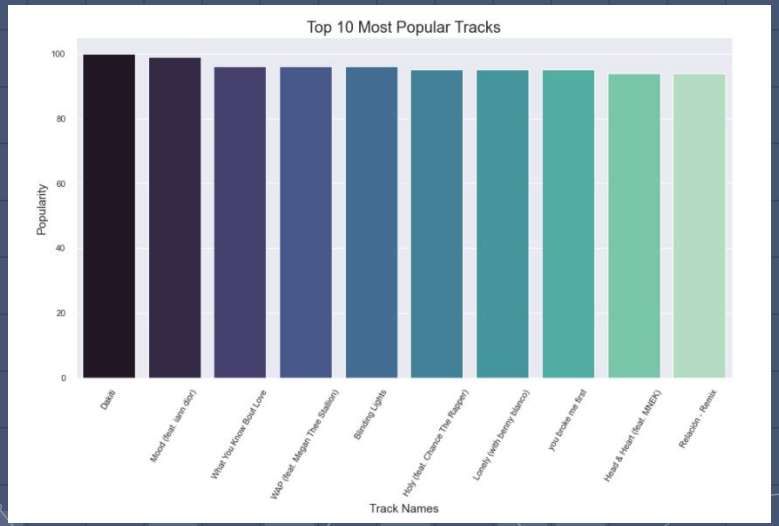
- Spotify is one of the most popular music streaming services around. They have an immense collection of songs dating back to 1921.
- I obtained data from the Spotify API and dataset from the kaggle website which contains over 175,000 songs between the years 1921-2020
- The dataset contained:
  - 170,000+ tracks
  - About 30,000+ artists
  - 16 track audio features
    - Not all audio features were used in this analysis

Exploratory Data Analysis

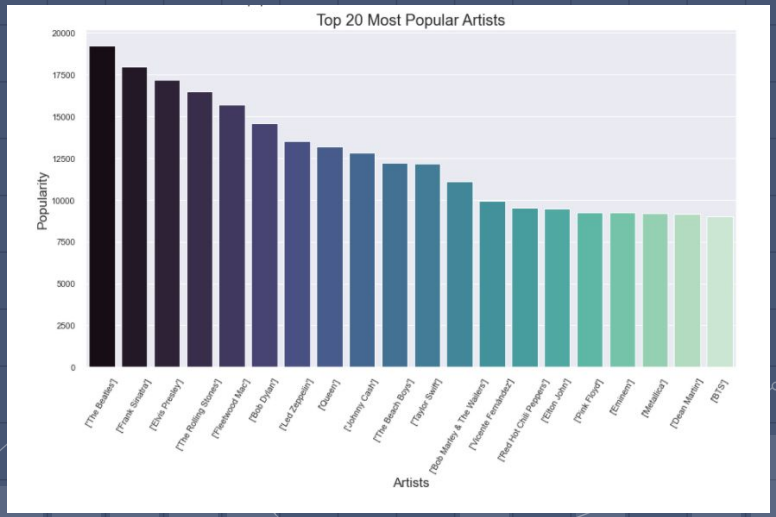


Exploratory Data Analysis

Top 10 Tracks

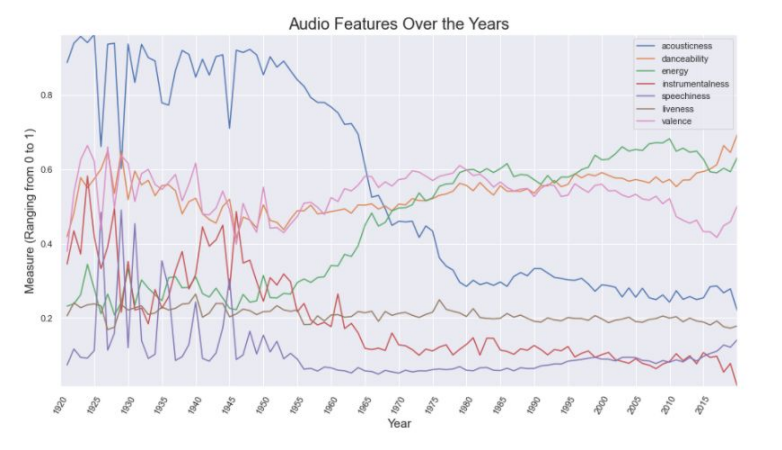


Top 20 Artists

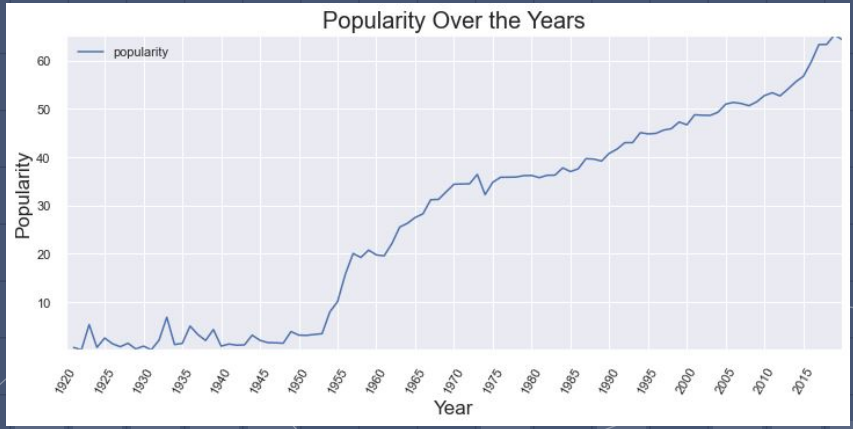


Exploratory Data Analysis

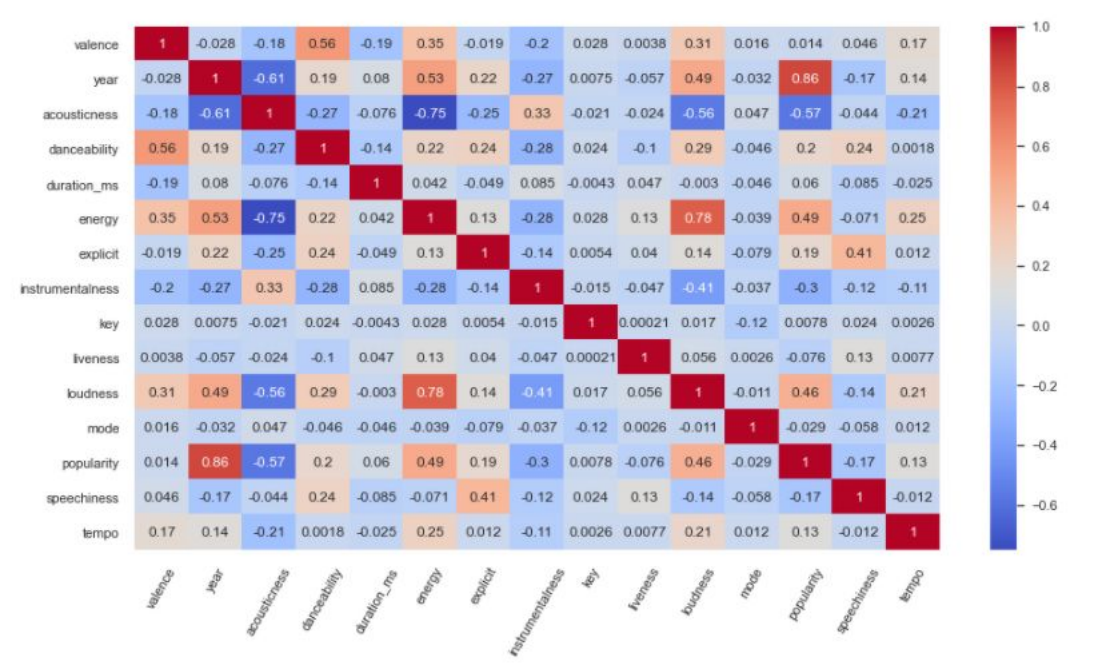
Time series analysis of audio features



Time series analysis of popularity



Exploratory Data Analysis



Classification Models

Built various classification models to compare their accuracies.

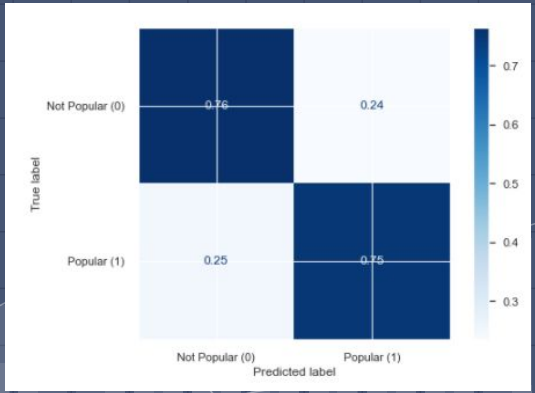
I also performed GridSearchCV on each model to get the best parameters and compared the accuracies.

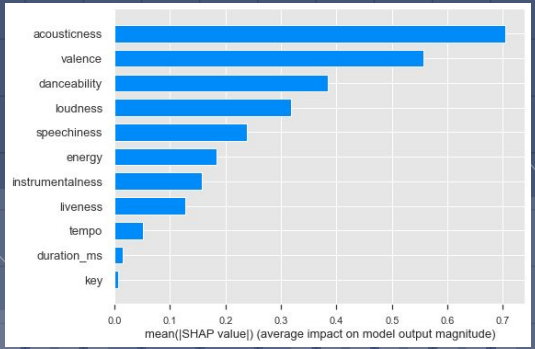
The models included were:

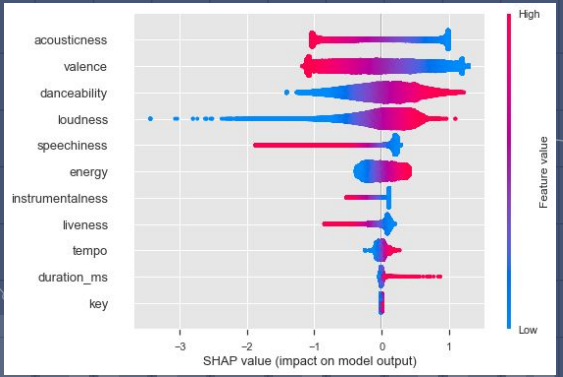
- Baseline LogisticRegression model
- LogisiticRegressionCV model
- Baseline DecisionTrees model
- Bagged Decision Tree
- Baseline RandomForests model

Best Logistic Regression Model

- The LogisticRegressionCV model performed the best
- The AUC value calculated from the ROC Curve for this model was 0.82
- The LogisticRegressionCV model had a performance accuracy of 75.7%
- Acousticness, valence, danceability, loudness, and speechiness to be the five most important features







Best Logistic Regression Model

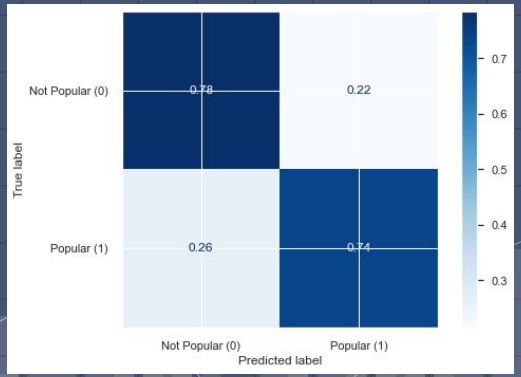
The summary plot summarizes the following:

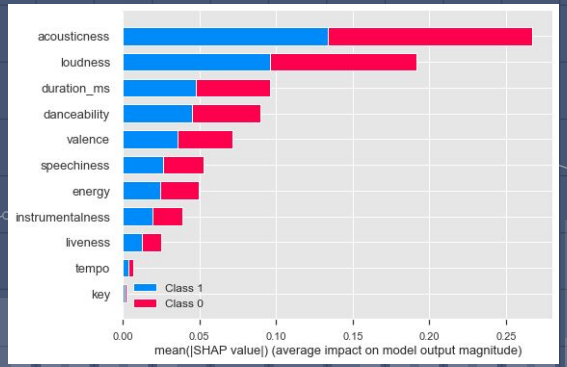
- When the level of acousticness of a track is low, it has a positive shap value and is more likely to be "popular"
- When the level of valence of a track is low, it has a positive shap value and is more likely to be "popular"
- When the level of danceability of a track is high, it has a positive shap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

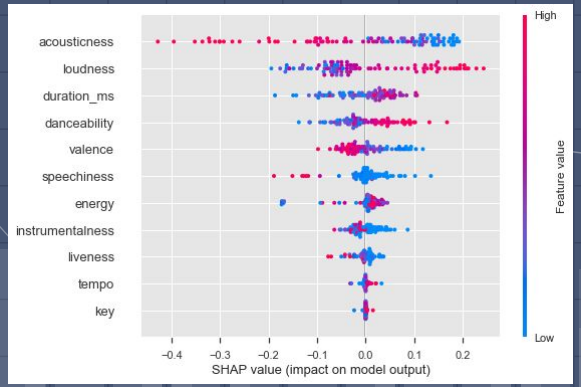
	coef
danceability	0.469579
loudness	0.404460
energy	0.214315
tempo	0.064652
duration_ms	0.026815
key	0.006954
liveness	-0.184105
instrumentalness	-0.197915
speechiness	-0.360529
valence	-0.651481
acousticness	-0.767210

Best Decision Trees Model

- The baseline Decision Trees model performed the best
- The AUC value calculated from the ROC Curve for this model was 0.76
- This model has a performance accuracy of 76.11%
- Acousticness, loudness, duration_ms, danceability, and valence to be the five most important features, meaning they have higher predictive power.

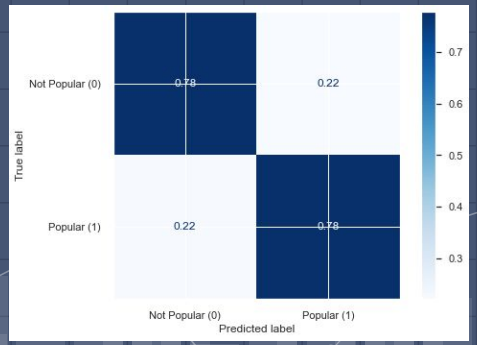


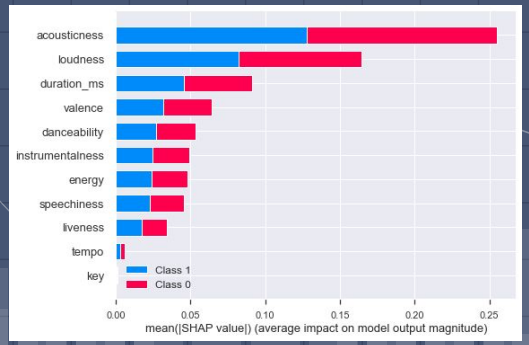


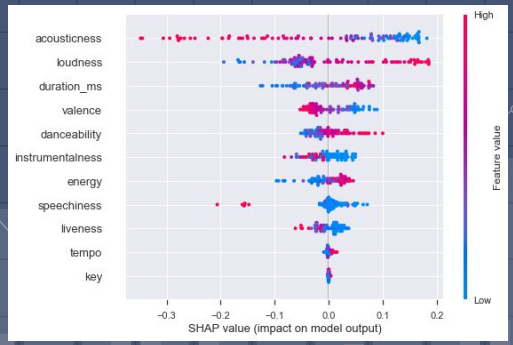


Best Random Forests Model

- The baseline Random Forests model performed the best
- The AUC value calculated from the ROC Curve for this model was 0.78
- This model has a performance accuracy of 77.68%
- Acousticness, loudness, duration_ms, valence, and danceability to be the five most important features, meaning they have higher predictive power.







Conclusions

- The Baseline Random Forests model performed the best of all models at an accuracy of 77.68%.
- For an artist that wants to create popular music I would recommend to create songs with low acoustics, a high loudness level, low valence, and high danceability.
- For the Logistic Regression models, acousticness, valence, danceability, loudness, and speechiness were ranked to be the five most important features.
- As for the Decision Trees and Random Forests models, acousticness, loudness, duration_ms, valence, and danceability were ranked to be the five most important features.

Recommendations to improve models / 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