Genre Predictions by Artists

Authors: Jake Ash & Donna Lee

Dataset description:

Labeled Genre Percentage
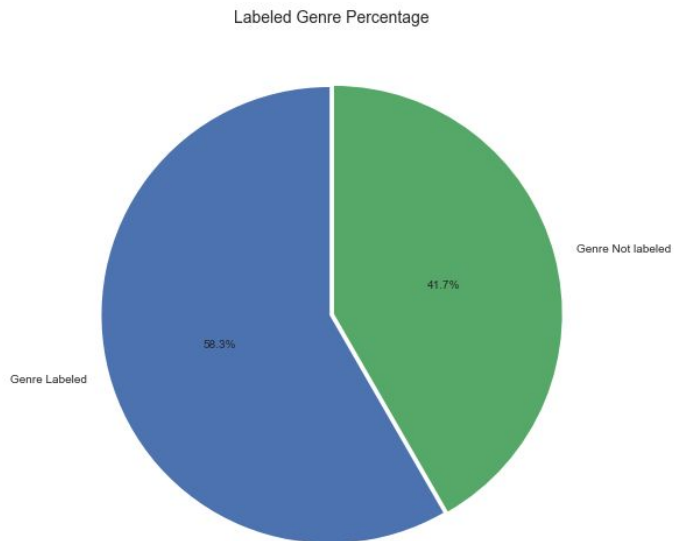
- Spotify dataset: Artists by genre and mean attributes of songs
- Attributes:
  - Acousticness
  - Danceability
  - Duration
  - Energy
  - Instrumentalness
  - Liveness
  - Loudness
  - Speechiness
  - Tempo
- Years: 1921 to 2021
- Genre Labels: 58%

Data Mapping Process:

	artists	genres	clean_genres
91	5th Ward Boyz	['dirty south rap', 'dirty texas rap', 'gangster rap', 'houston rap']	rap
194	Aaron Copland	['american modern classical', 'classical', 'early modern classical']	classical
197	Aaron Lewis	['contemporary country', 'country', 'country road', 'modern country rock', 'outlaw country', 're...	country
204	Aaron Watson	['contemporary country', 'country', 'country road', 'modern country rock', 'texas country']	country
236	Ace	['dirty south rap', 'gangster rap', 'hip hop', 'miami hip hop', 'pop rap', 'rap', 'southern hip ...	rap
...	...	...	...
18727	Zoot Sims	['bebop', 'contemporary post-bop', 'cool jazz', 'hard bop', 'jazz', 'jazz saxophone', 'soul jazz...	jazz
18771	convolk	['dark trap', 'emo rap', 'melodic rap', 'sad rap']	rap
18807	ilyTOMMY	['emo rap', 'sad rap', 'viral rap']	rap
18834	nothing,nowhere.	['emo rap', 'pop rap', 'sad rap']	rap
18883	Ólafur Arnalds	['bow pop', 'compositional ambient', 'icelandic classical', 'neo-classical', 'neoclassical darkw...	classical

We chose four distinct general genres to map each artist to based on their tags. To test this model, we started out with just four genres - rap, country, classical and jazz. Future steps would be to expand this list.
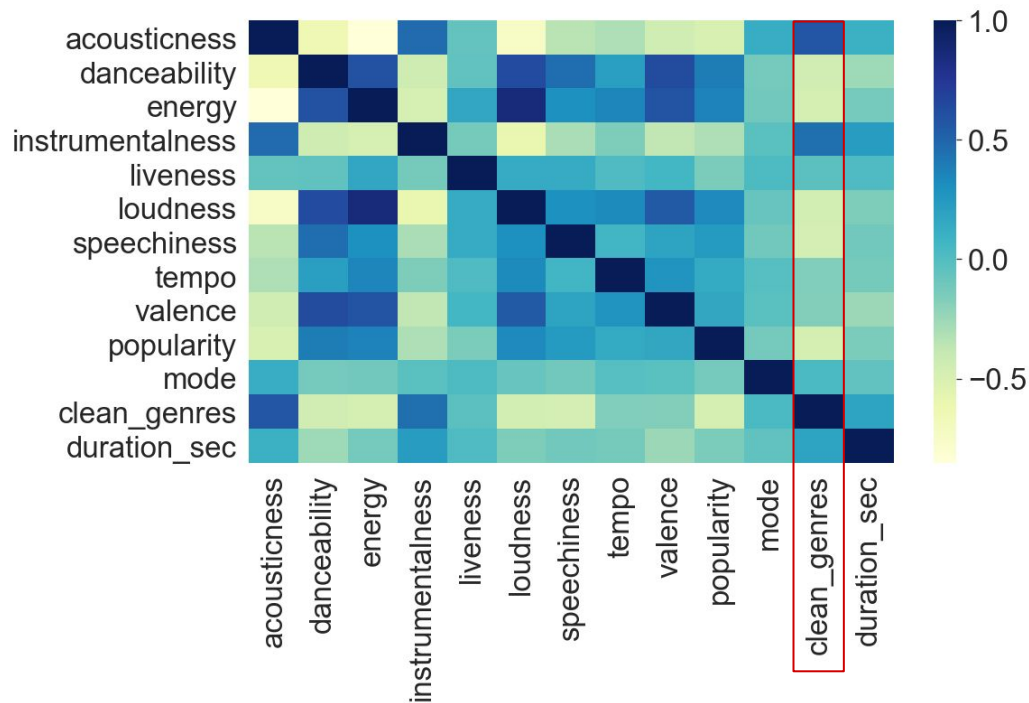
Spotify®

Genre Distribution:
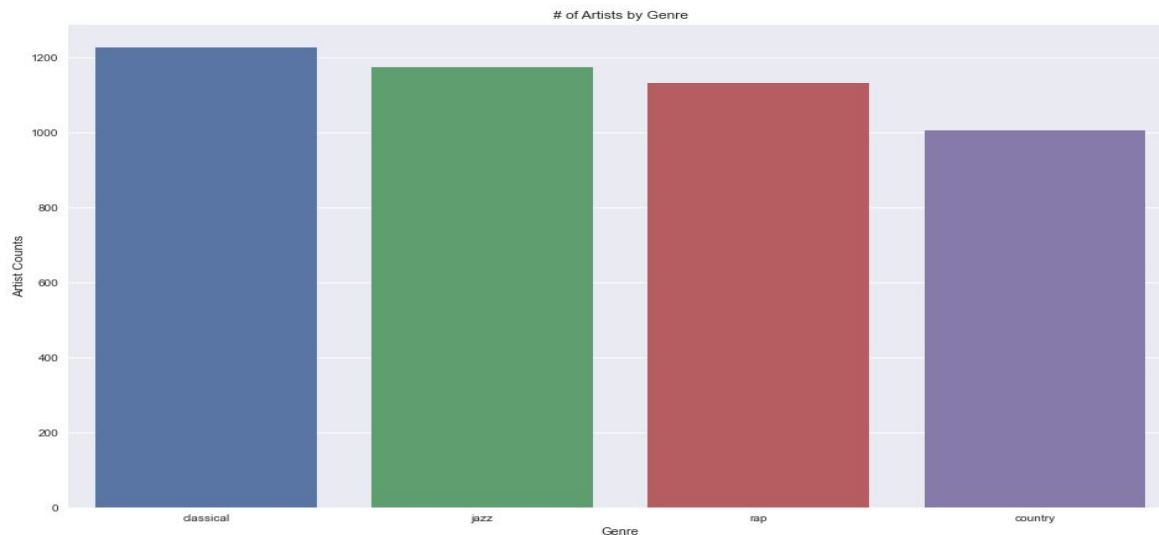
All four genres had similar levels of artist counts so class imbalance was not a big concern. The four genres all had a little over 1,000 artists.

Important Features:

Target Variable

Some important features in predicting genres:

- Acousticness
- Instrumentalness
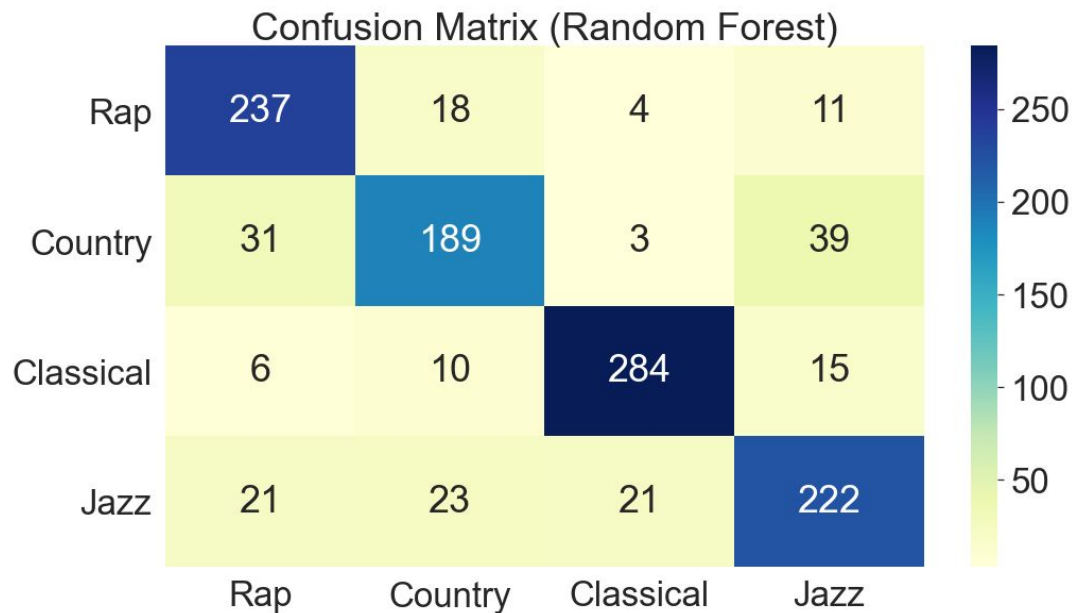- Duration (Seconds)

Spotify®

Feature Modeling Results:

Model	F1 Score
Random Forest	0.8226
KNN	0.6548
XGBoost	0.8182

We ran 3 models - Random Forest, KNN and XGBoost to predict the genres.

After predictions, we calculated the F1 score of each model and Random Forest came out on top.

Spotify®

Random Forest Predictions:

Confusion Matrix (Random Forest)

Based on the confusion matrix, our random forest model had limited number of wrong predictions.

This model was especially good at predicting classical music.

Predictions:

Here are a couple of our predictions. You can tell us how accurate you think it is!

	artists	clean_genres
4971	Hot Girl Meg	rap
12365	Trey Lewis	country
570	Andrew Andraos	classical
966	Avenue Blue (Featuring Jeff Golub)	jazz

Next Steps:

- Expanding super genre list to be more comprehensive

- Automate clustering of genres to remove manual mapping

- Artist radio based on genres

Spotify®

References:

- Kaggle Dataset - https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data_w_genres.csv
- Spotify Logo - https://storage.googleapis.com/pr-newsroom-wp/1/2018/11/Spotify_Logo_RGB_Green.png