Predicting Music Genre

Sean Simon

Music is Important

- Warm home environment
- Lively shopping experience
- Comfortable hotel atmosphere

Curated music selections with a wide breadth of sound space improve the settings in which they're played.

Genre is key.

Music Data is Accessible

3rd party music service Spotify serves open access data through an API

Subscribers:

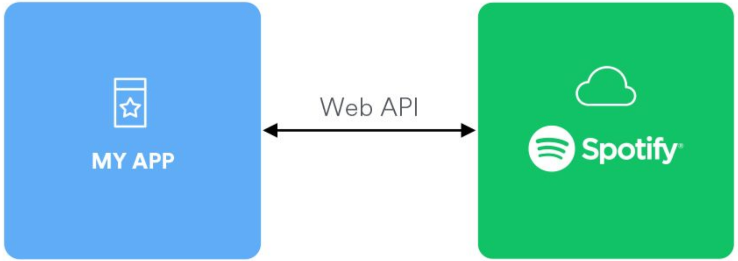
172 million

Tracks:

70 million

Availability:

184 markets




```
curl --request GET \
  --url https://api.spotify.com/v1/audio-features \
  --header 'Authorization: ' \
  --header 'Content-Type: application/json'
```

The Data

{“id”, “artist”, “track”, “popularity”, “acousticness”, “danceability”, “duration”, “energy”, “instrumentalness”, “key”, “liveness”, “loudness”, “mode”, “speechiness”, “tempo”, “obtained_date”, “valence”, “genre”}

Fields:

18 features

Representation:

10 genres

Size:

50,000 tracks

Source:

https://www.kaggle.com/vicsuperman/prediction-of-music-genre

Format:

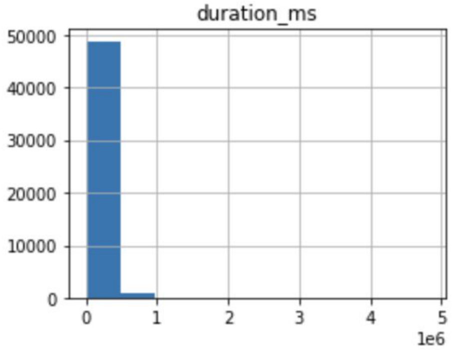
CSV

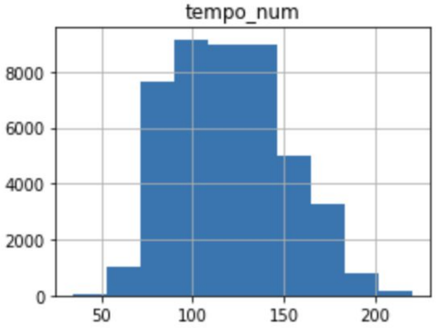
Engineering the Data: Cleaning Missing Values

Data is never perfect.

- **Median** imputation for skewed features

- **Mean** imputation for normal features





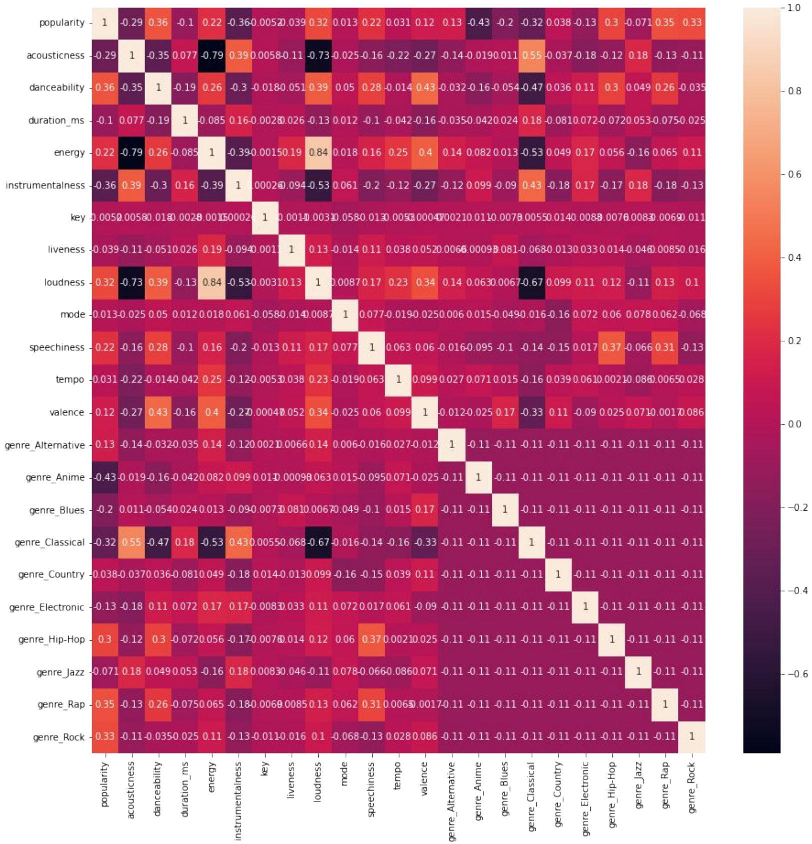
Engineering the Data: Cleaning High Entropy Features

	instance_id	artist_name	track_name	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key
0	32894.0	Röyksopp	Röyksopp's Night Out	27.0	0.00468	0.652	-1.0	0.941	0.79200	A#
1	46652.0	Thievery Corporation	The Shining Path	31.0	0.01270	0.622	218293.0	0.890	0.95000	D
2	30097.0	Dillon Francis	Hurricane	28.0	0.00306	0.620	215613.0	0.755	0.01180	G#
3	62177.0	Dubloadz	Nitro	34.0	0.02540	0.774	166875.0	0.700	0.00253	C#
4	24907.0	What So Not	Divide & Conquer	32.0	0.00465	0.638	222369.0	0.587	0.90900	F#

- Minimize low predictive power
- Maximize generability

Engineering the Data: Confidence in the Features

- Correlation = Redundancy
- Correlation shown by intensity of *light* or *dark*



Each feature demonstrates unique information.

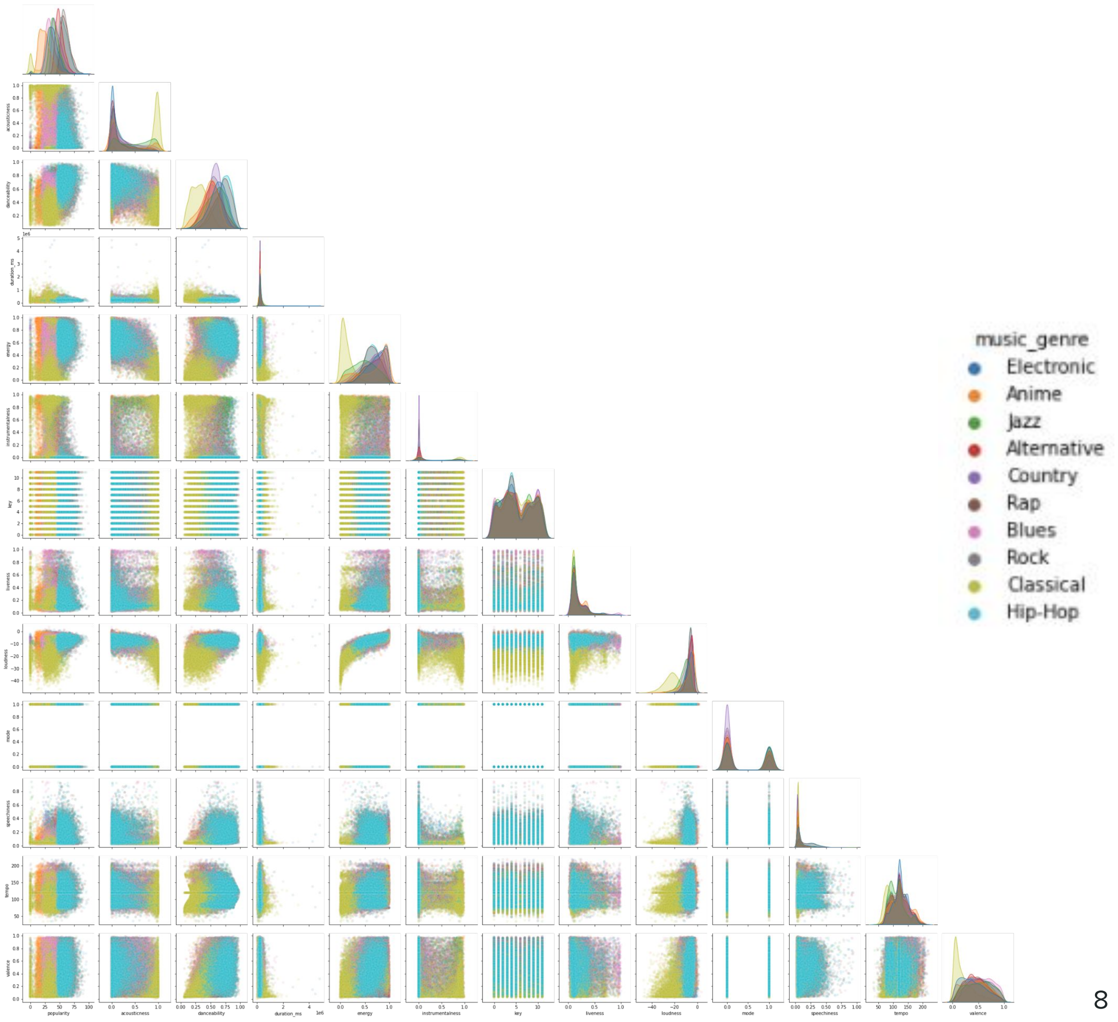
Visualizing Genre

Separation:

- Anime, Blues, Classical, Hip-hop

Overlap:

- Hip-Hop & Rap
- Rock & Country



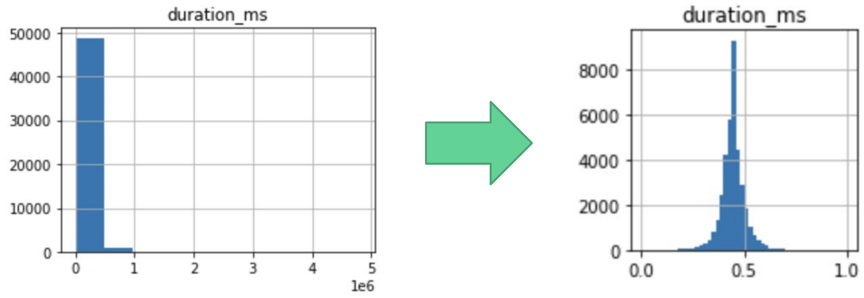
Engineering the Data: Categories, Distributions, and Scale

1

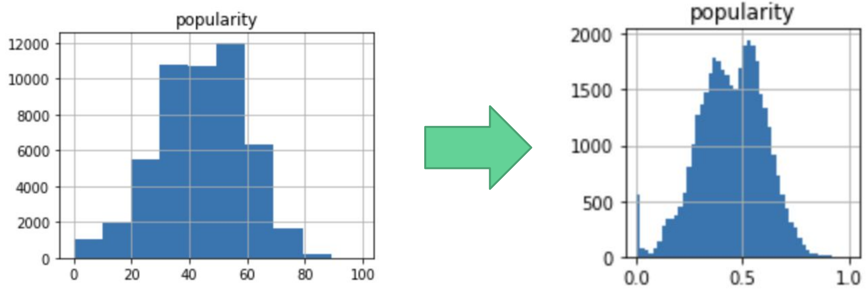
...	C	C#	D	D#	E	F	F#	G	G#	Minor

1. One Hot Encoding
2. Power Transformation
3. Min-Max Scaling

2



3



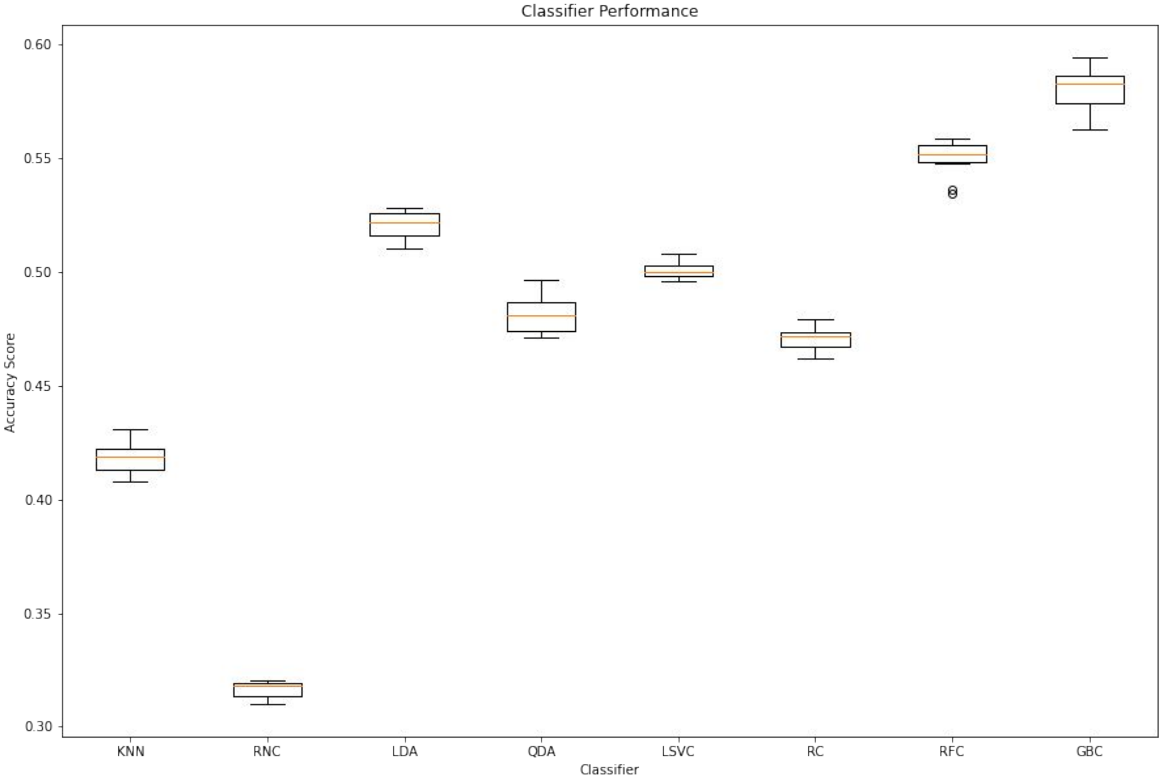
Model Selection

- Default scikit-learn classifiers
- 10-fold CV scoring
- Hyperparameter tuning

Gradient Boost Classification outperforms 7 other classifiers by roughly 8%

Training accuracy:

58%



Model Performance

Test Accuracy:

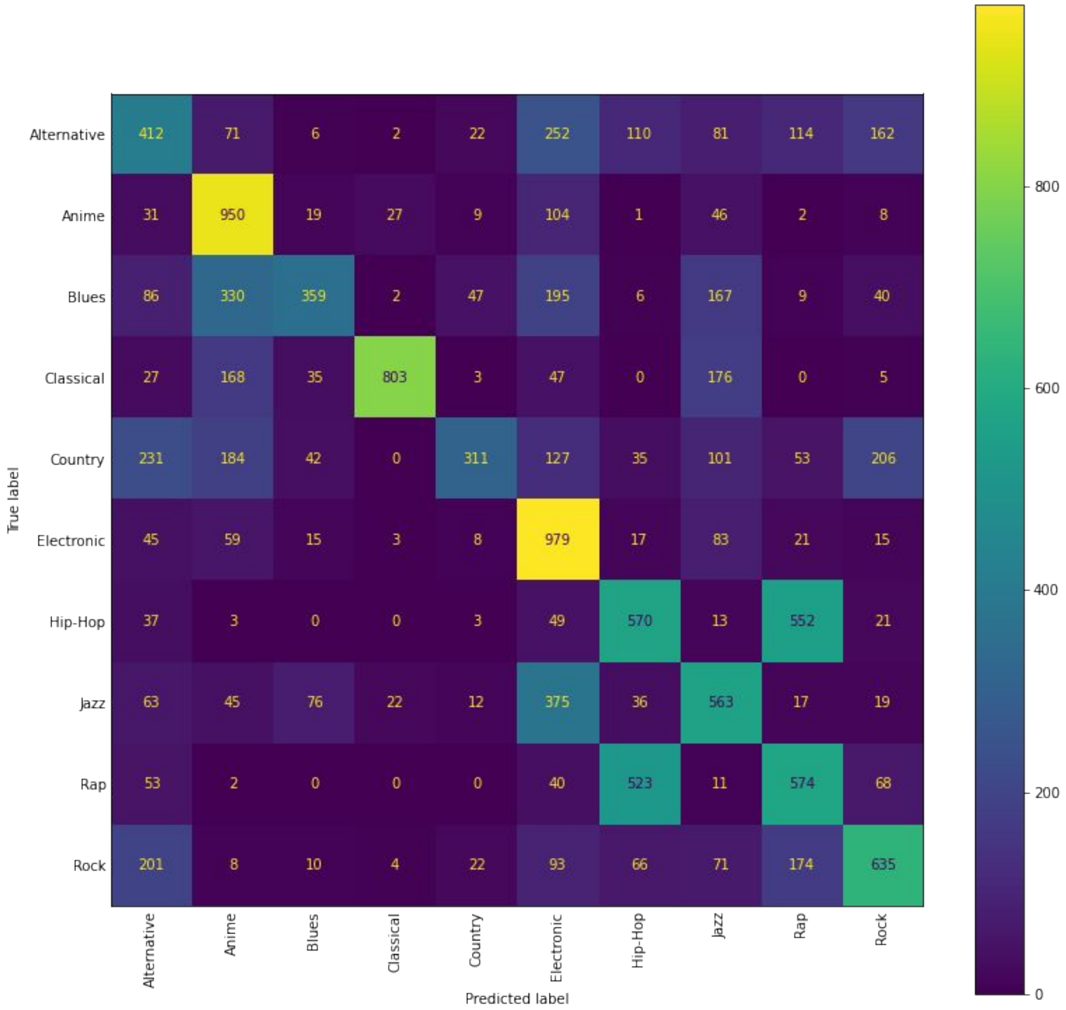
- 49%

Strengths:

- Anime, Classical, Electronic

Weaknesses:

- Alternative, Blues, Country



Summary

- 40% more accuracy than randomly picking a new song
- Robust playlist building engine
  - Add tracks from within genre
  - Decent generalization accuracy for newly released music
  - Build playlist depth by adding songs from adjacent genres
- Music for Mood exploration tool

Future Work

- Better accuracy through PCA
- DSP characteristic approach to generalization
- Music for Mood song suggestion web app