Automating Genre Labelling With Classification Models

Matt Segall, Automation and Machine Learning

Why?

Hypotheses

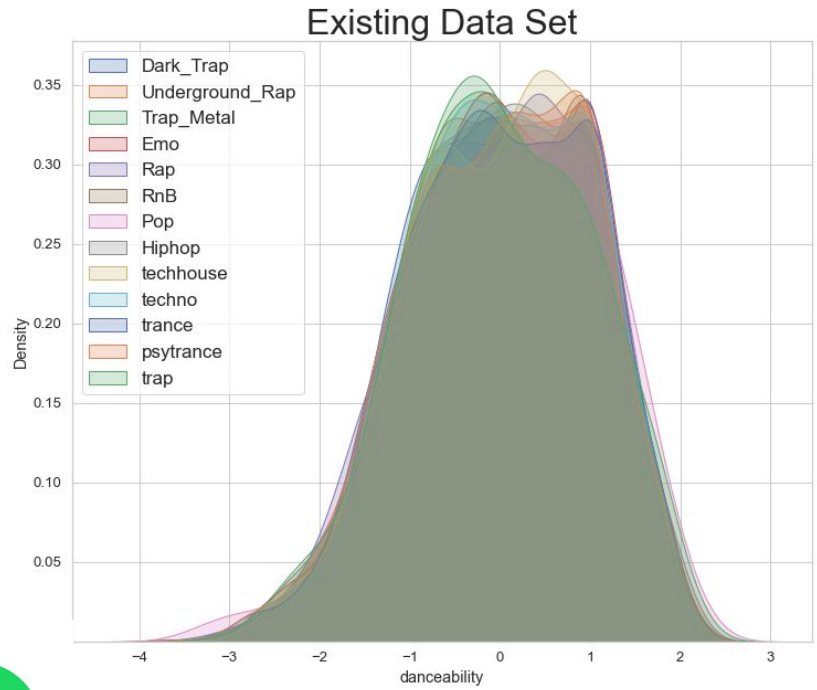
More Important

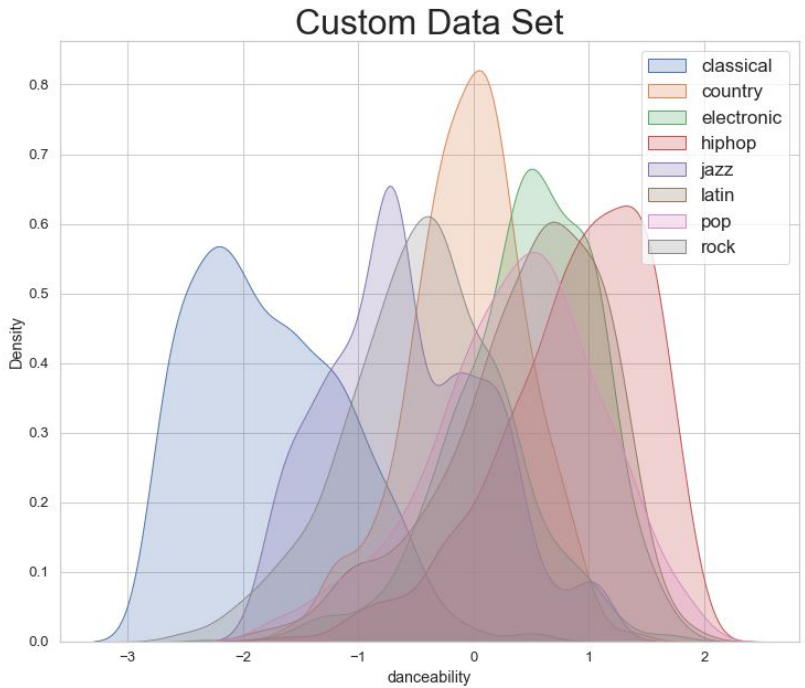
- Proprietary Attributes
  - Energy
  - Acousticness
  - etc
- Length
- Loudness
- Tempo

Less Important

- Name
- Artist
- Key
- Mode

Difference in Data Sets





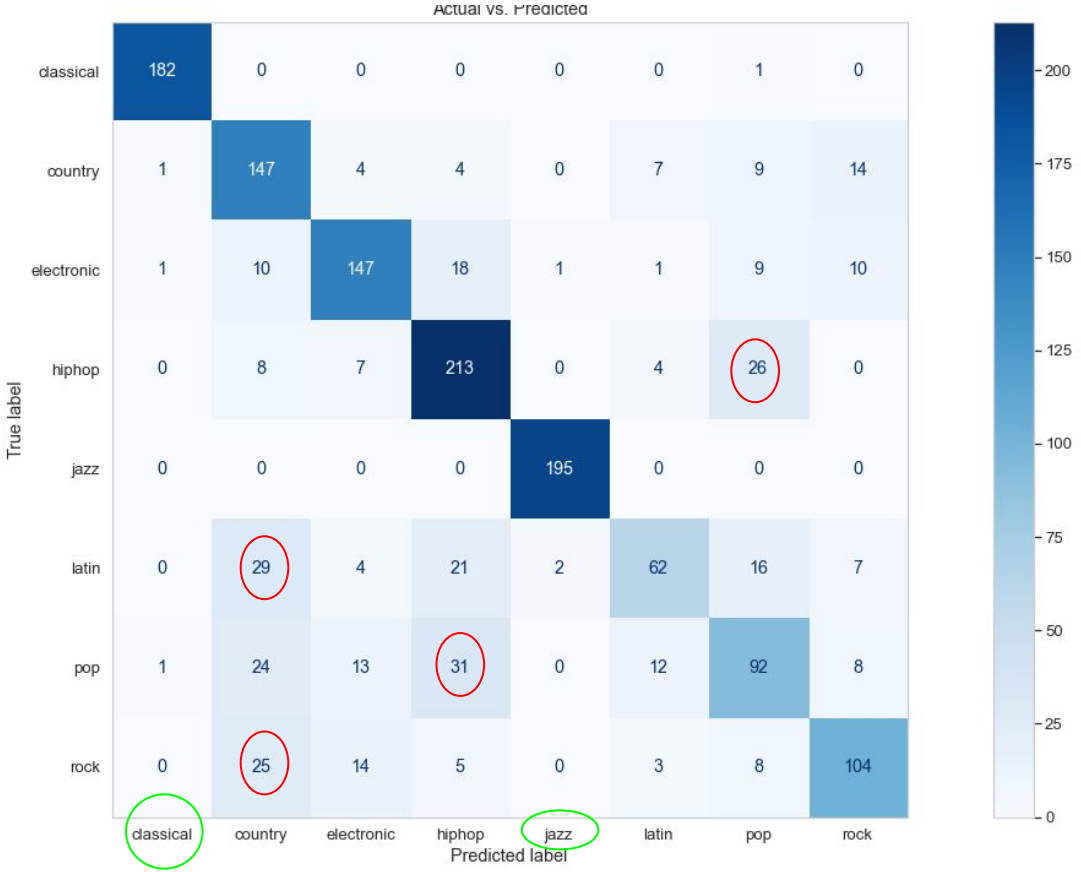
Model Selection

	K-Nearest Neighbors	Logistic Regression	Naive Bayes	**Random Forest**	XGBoost
Accuracy	65.39%	62.87%	54.20%	**77.93%**	76.26%
ROC AUC-score	0.91	0.92	0.89	**0.92**	0.92
Run Time (seconds)	.0101	0.2861	.0104	**4.8590**	2.1143

76.13%

Accuracy on new data (expected)

Model Trends



Looking forward

- How to balance run time and accuracy?
  - What levels would we like to see before implementation?
- Prioritize cleanliness of certain genres?
  - Different metrics
- MORE data
- Modeling raw audio