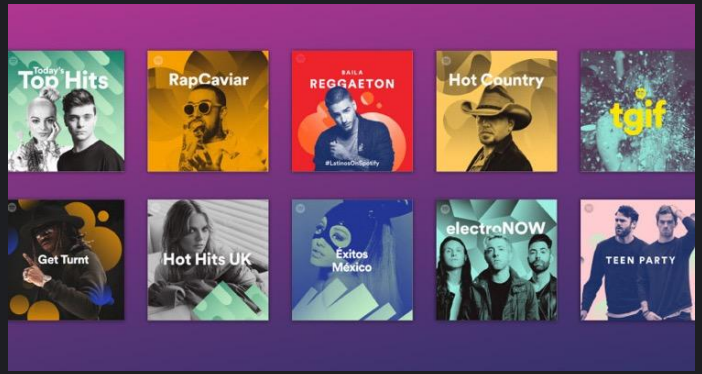
Spotify

Classification of Song Genres



Andrew Wong

Goals

- Create model to classify song genres
- Determine most important features
- Examine insights

Design

- Performance metric: Accuracy
- Business application
    - Improve Spotify recommended playlists
    - Enhance user experience
    - Enhance product

Data

- CSV download from /r/datasets subReddit
- 25809 songs
- 9 genres:
  - 'Emo', 'Pop', 'Underground Rap', 'DnB', 'Hardstyle', 'Psytrance', 'TechHouse', 'Techno', 'Trance'
- 12 features:
  - (incl. Danceability, Energy, Instrumentalness, Speechiness, Tempo)

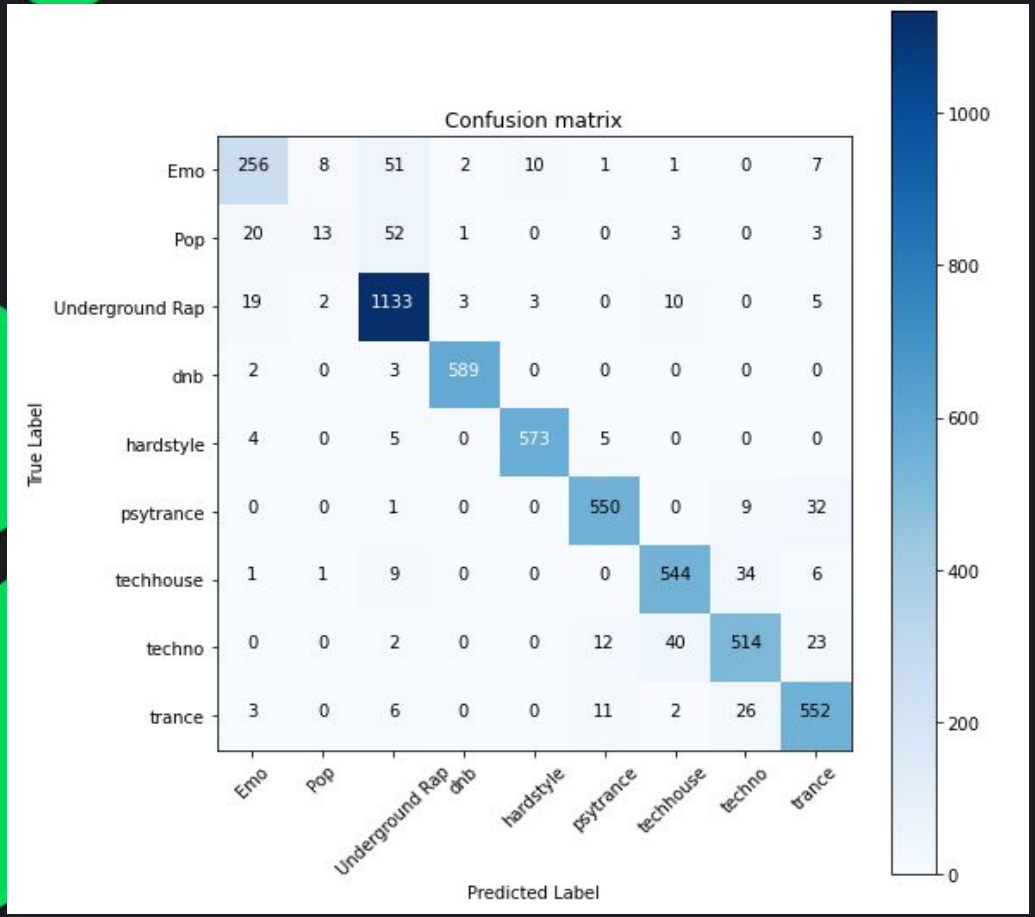
Data

- Class Imbalance
- Underground Rap (U_R) = Largest
    - U_R : Pop = 12:1
    - U_R : Emo = 3.5:1
    - U_R : Rest = 2:1

Models: KNN

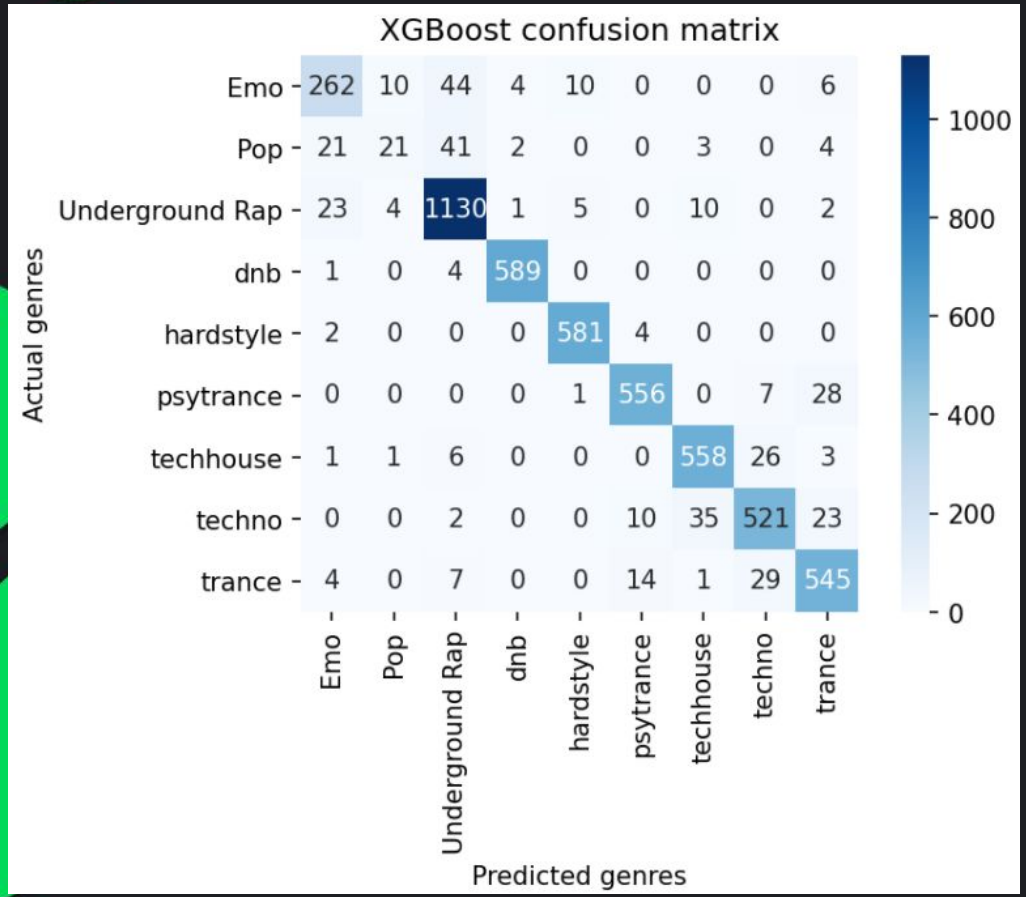
- Scaled feature values, N neighbors = 5
  - Val Accuracy = 0.815
  - Test Accuracy = 0.835
- KNN CV & KNN GridSearch CV
  - 0.829 & 0.832

Models: Random Forest



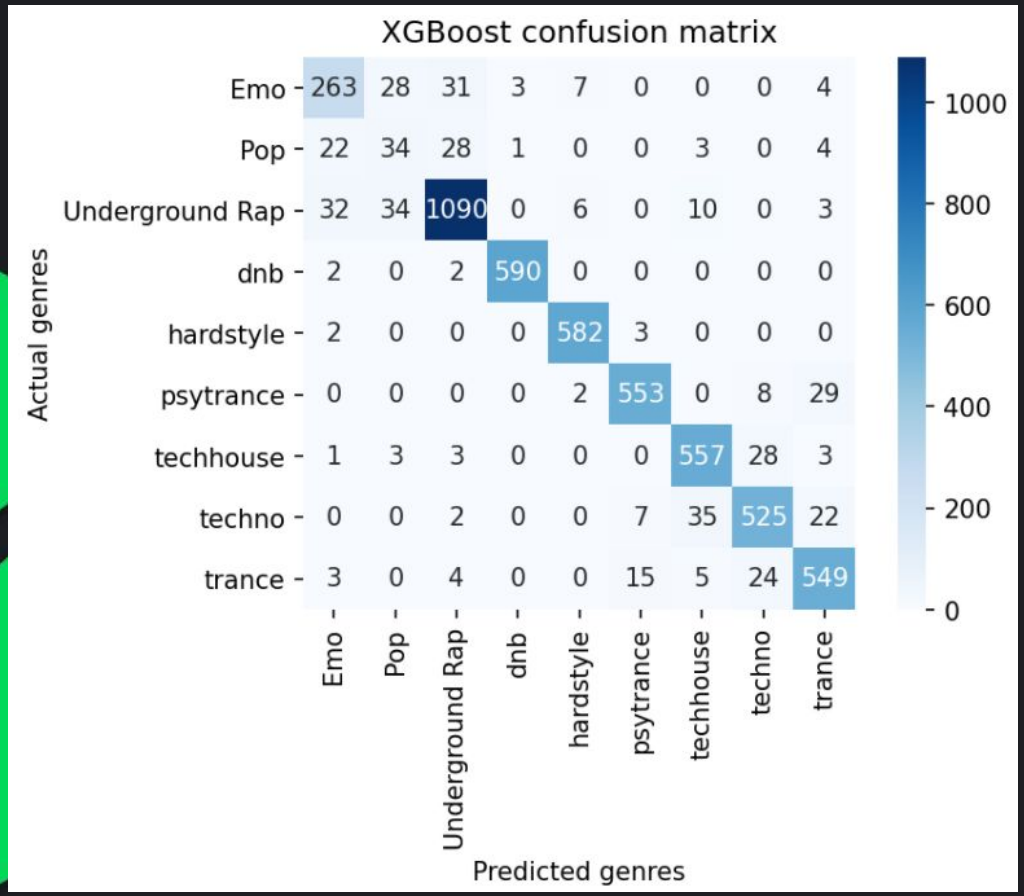
- N estimators = 400
  - Val Accuracy = 0.91
  - Test Accuracy = 0.92

Models: XGBoost



- Accuracy = 0.923
- Top 3 Most Accurate:
  - Drum and Bass = 0.99
  - Hardstyle = 0.99
  - Psytrance = 0.94

Models: XGBoost

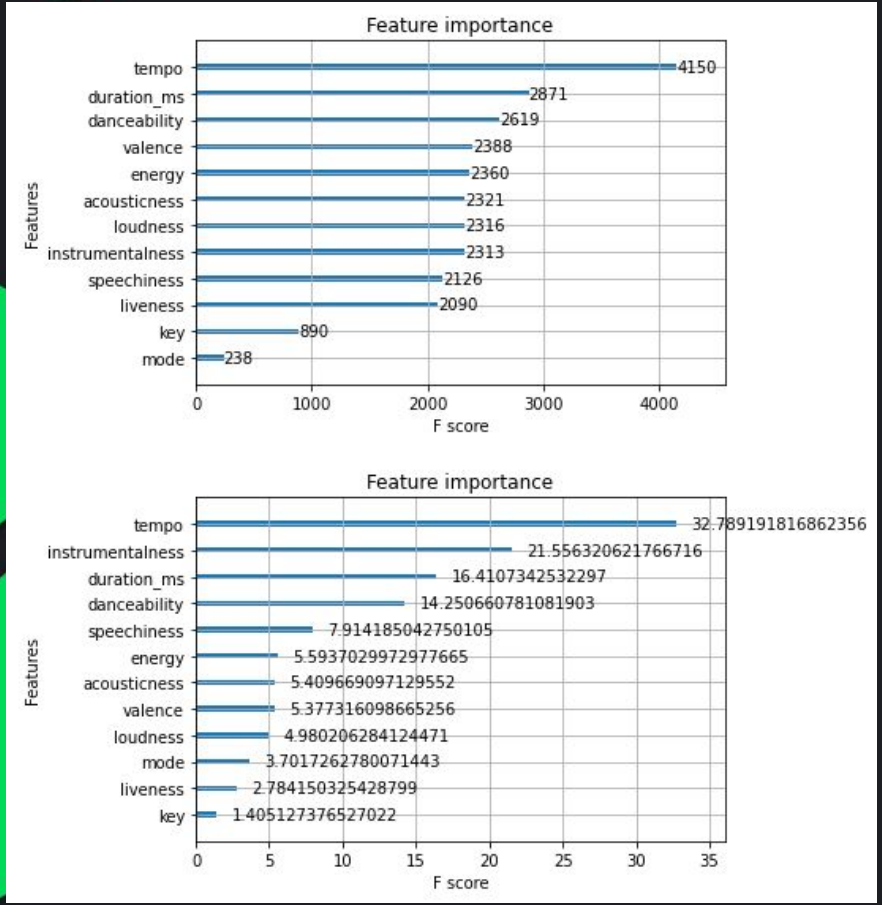


- Random Over Sampling
  - Upsampled to Underground Rap size
  - Accuracy = 0.919

Results/Insights

- XGBoost: Highest Accuracy
- Lowest predictive accuracy: Pop
    - Fewest data points
    - Loosely defined category, sub-genre
    - Mislabeled most on Underground Rap

Results/Insights: XGBoost



- Gain: Top 5 Features

1. Tempo
2. Instrumentalness
3. Duration
4. Danceability
5. Speechiness

Future Work

- More data, more genres
- Tuning
- Lyrics

Appendix

https://www.reddit.com/r/datasets/comments/k7apq3/i_created_a_dataset_of_mostly_edmtrap_songs_for_a/