Spotify Sequential Skip Prediction Challenge

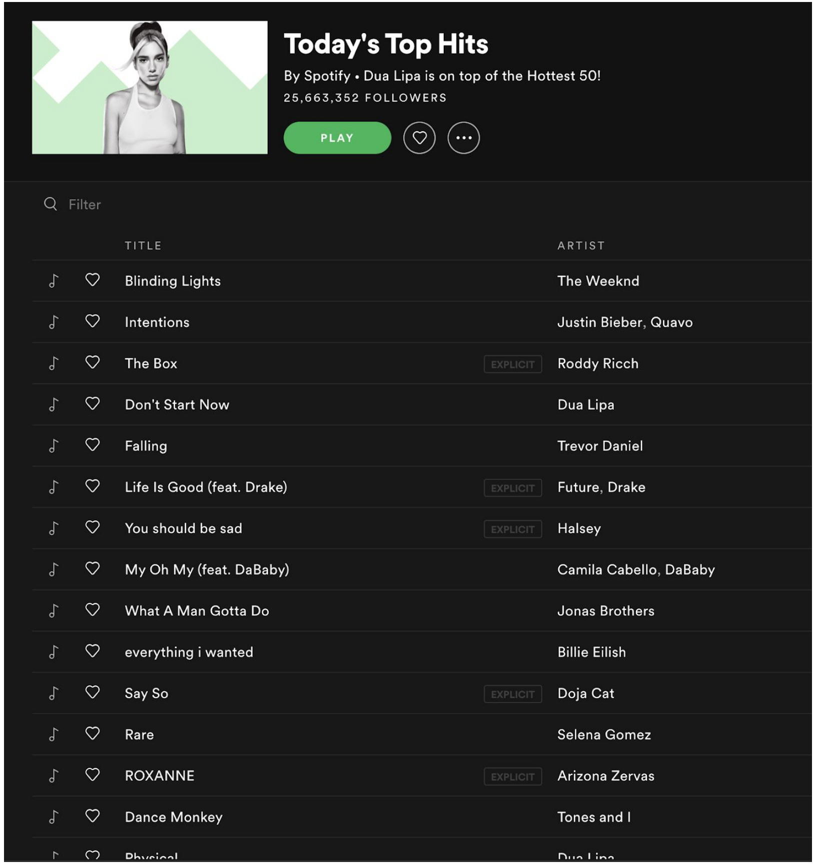
Austin Poor

Introduction

“Spotify has over 190 million active users interacting with over 40 million tracks”

The Challenge

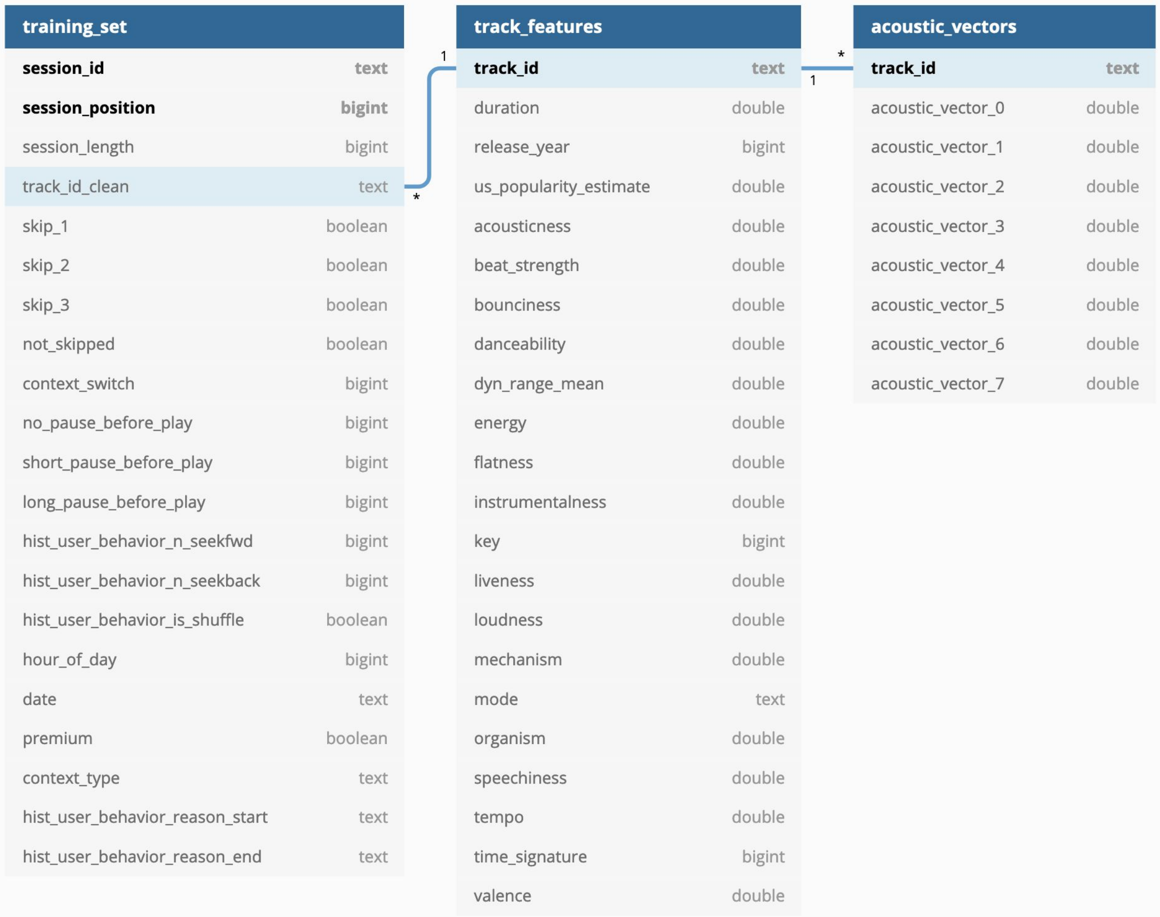
Goal: *Predict the likelihood of a user skipping any given song during a listening session*



Methodology

The Data

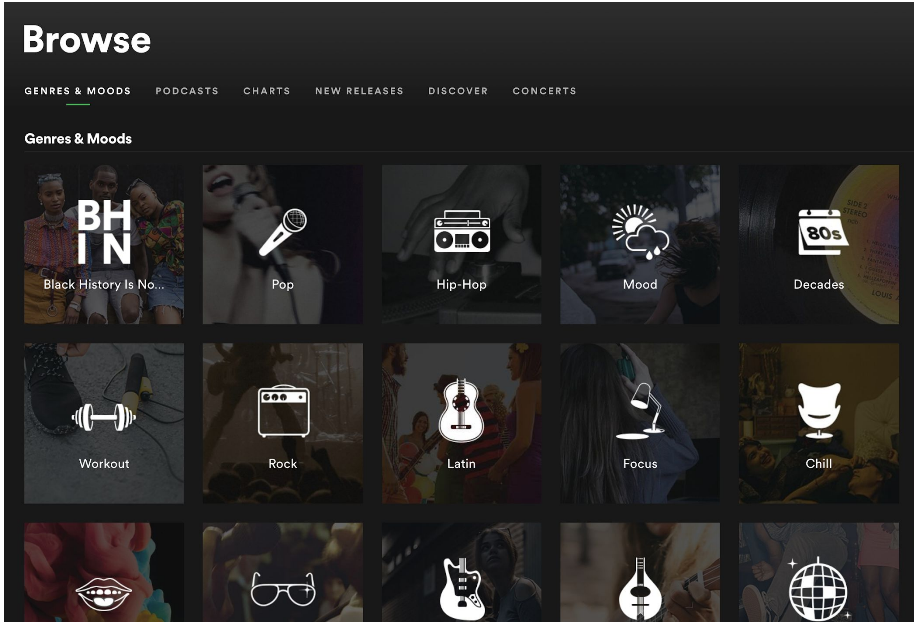
- 10-20 song sessions
- Song and user data was anonymized
- Balanced classes (0.517 skips)
- 350GB of CSV data
- Used subset of the data
  - 100k Session Rows
  - 50k Track Rows
- Stored data in PostgreSQL DB



Feature Engineering

Generate features to account for a user's listening history

Added previous track features, including if that track was skipped

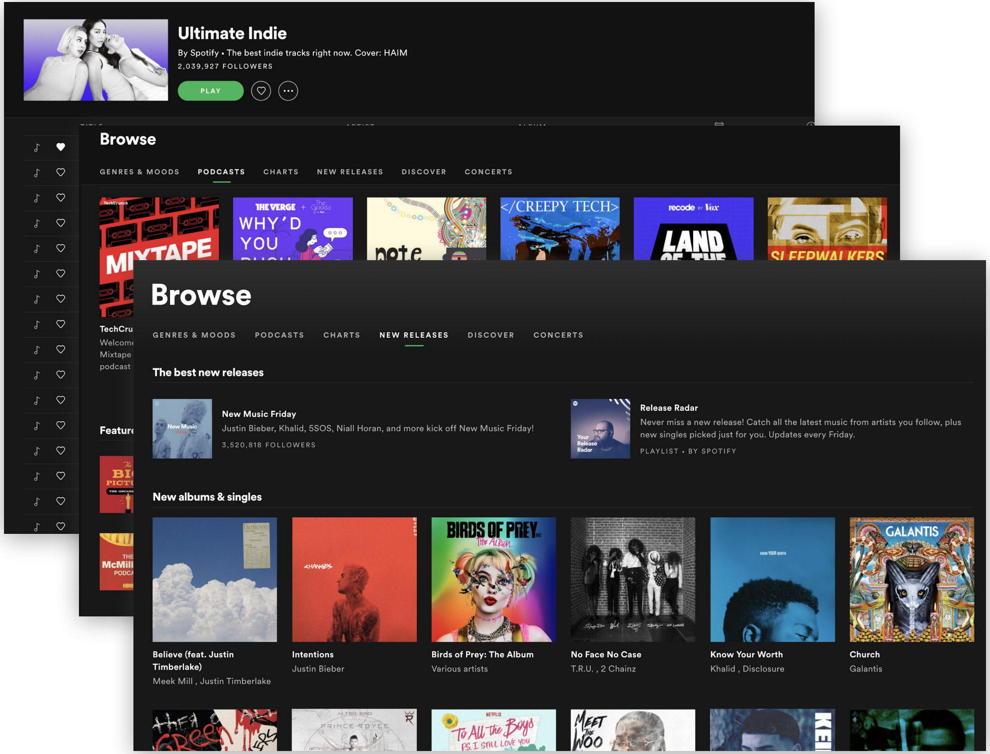


Model Selection

Target Metric **Accuracy** per competition guidelines

Baselined with Logistic Regression

Moved to tree-based models which would automatically handle feature interactions

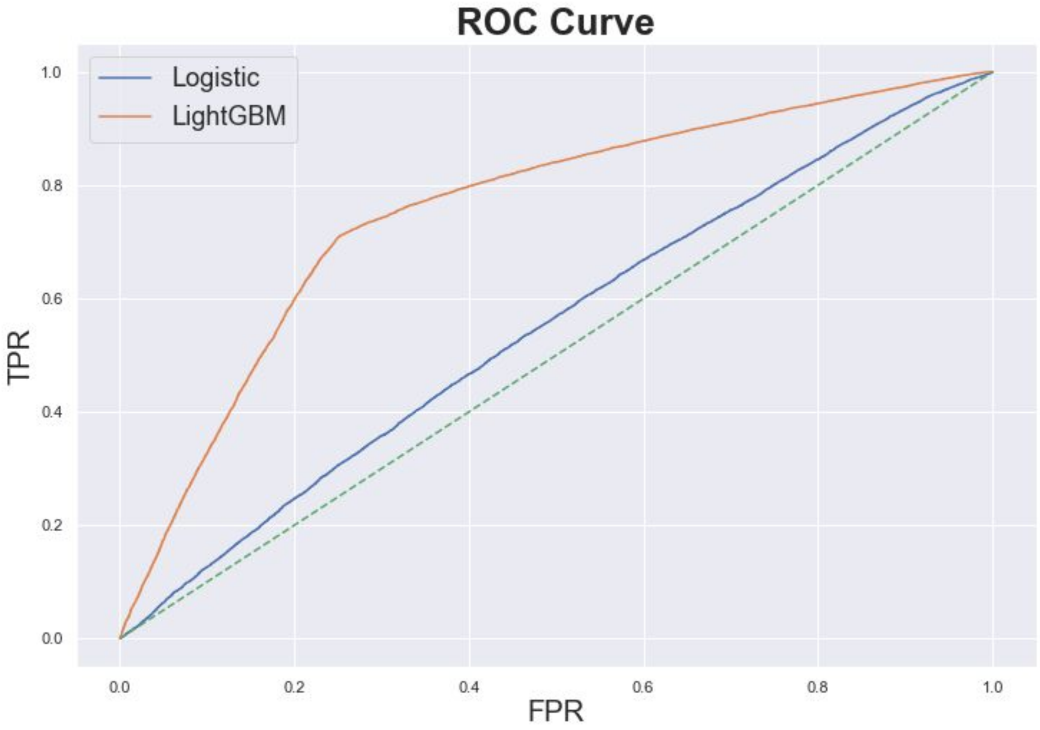


Results

Model Results

Best Test Accuracy: **0.73** (with LightGBM)

Error Analysis: No clear trend in the model residuals

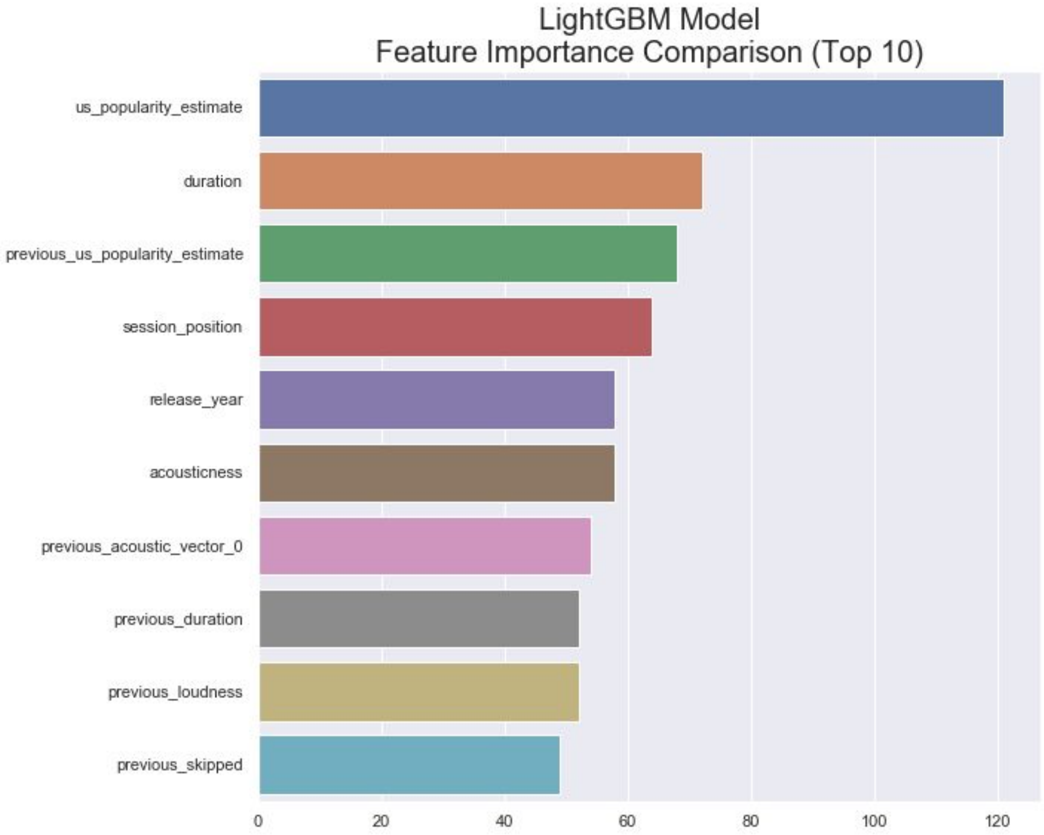


Conclusions

Conclusions

us_popularity_estimate had the highest feature importance followed by duration and then previous_us_popularity_estimate

Model results are pretty good but with room for improvement



Future Work

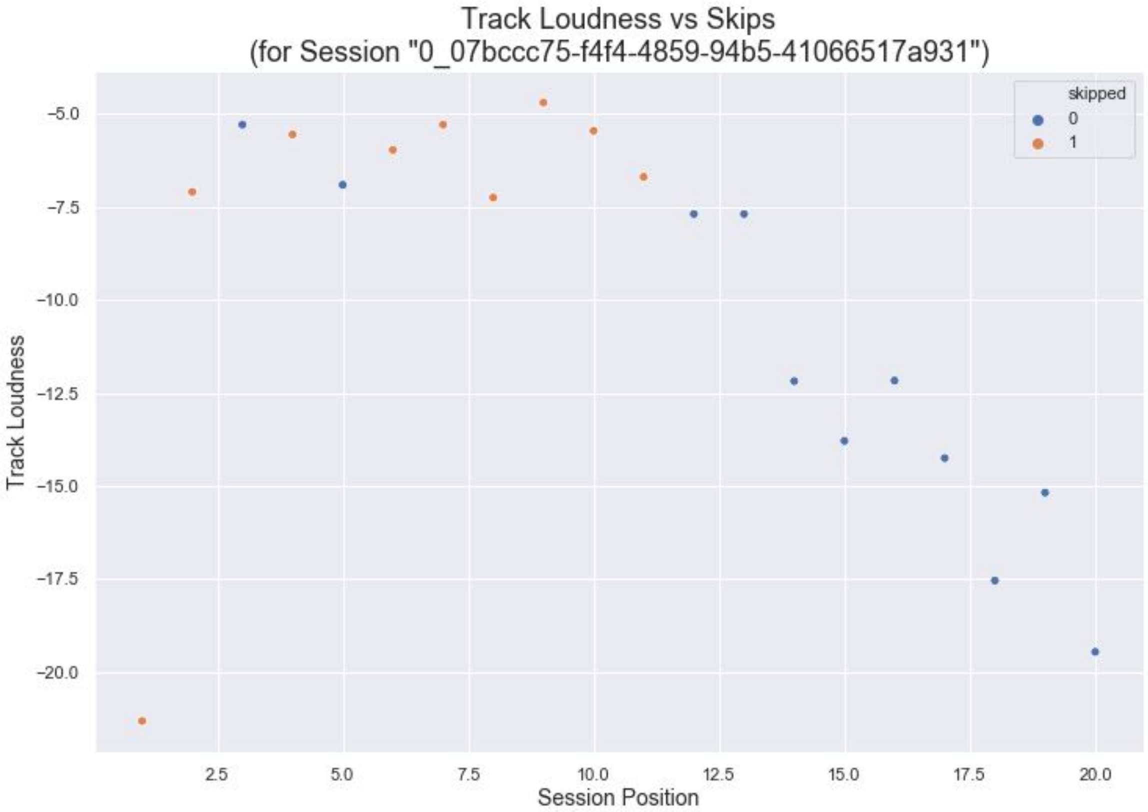
Future Work

- Test other types of algorithms:
  - Unsupervised Learning to cluster songs
  - RNN to predict based on the sequence of tracks
- Supplement the dataset with more data from the Spotify API
- Create a Flask app to visualize predictions using D3

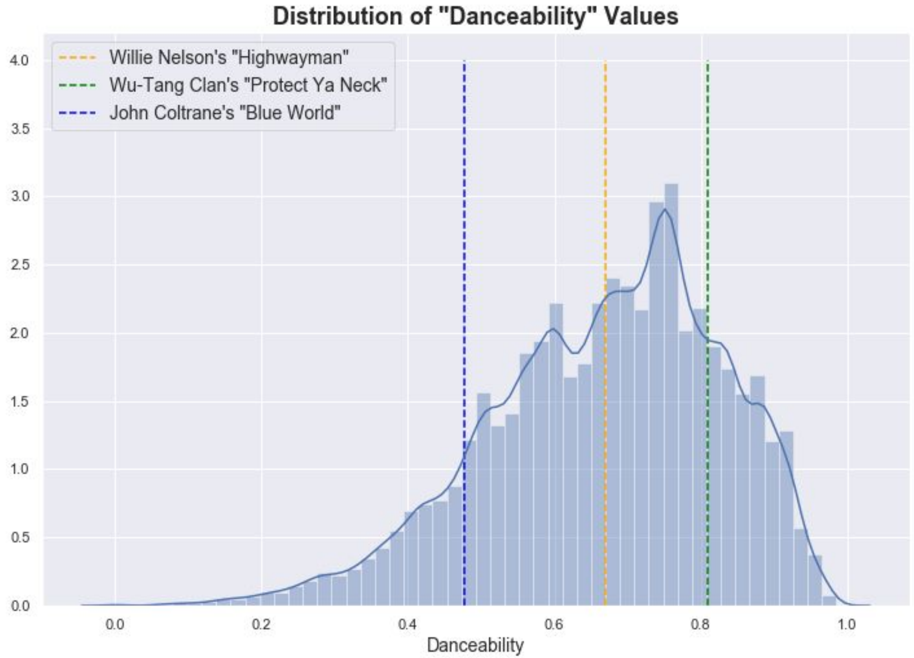
Thank you

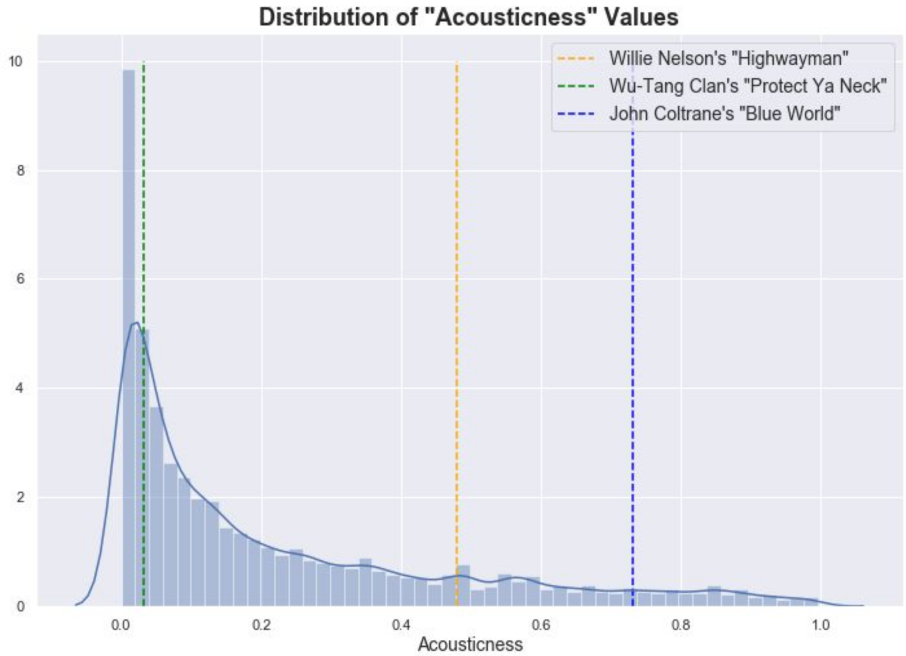
Appendix

Visualizing a Listening Session

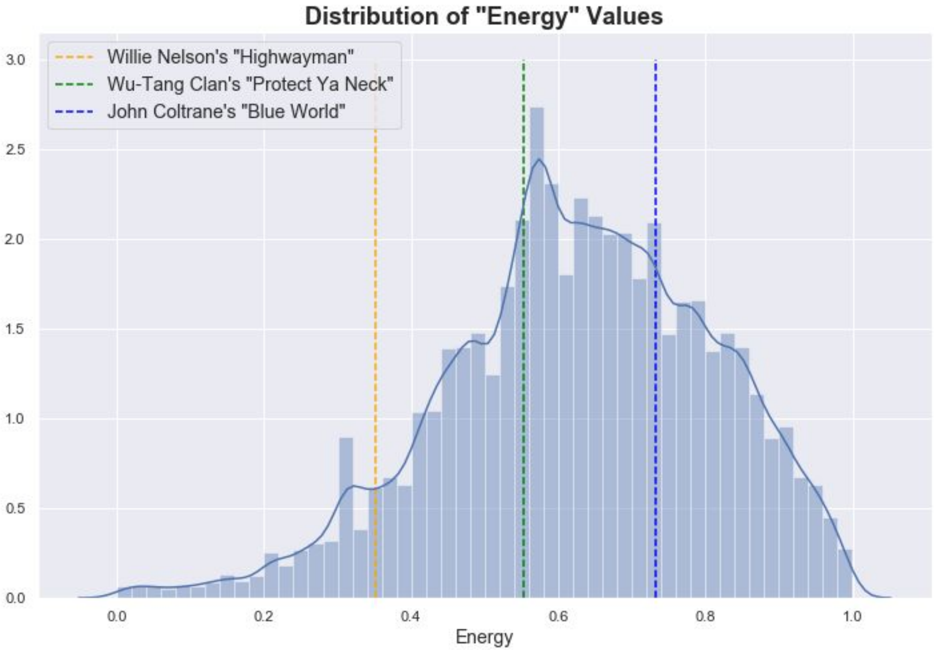


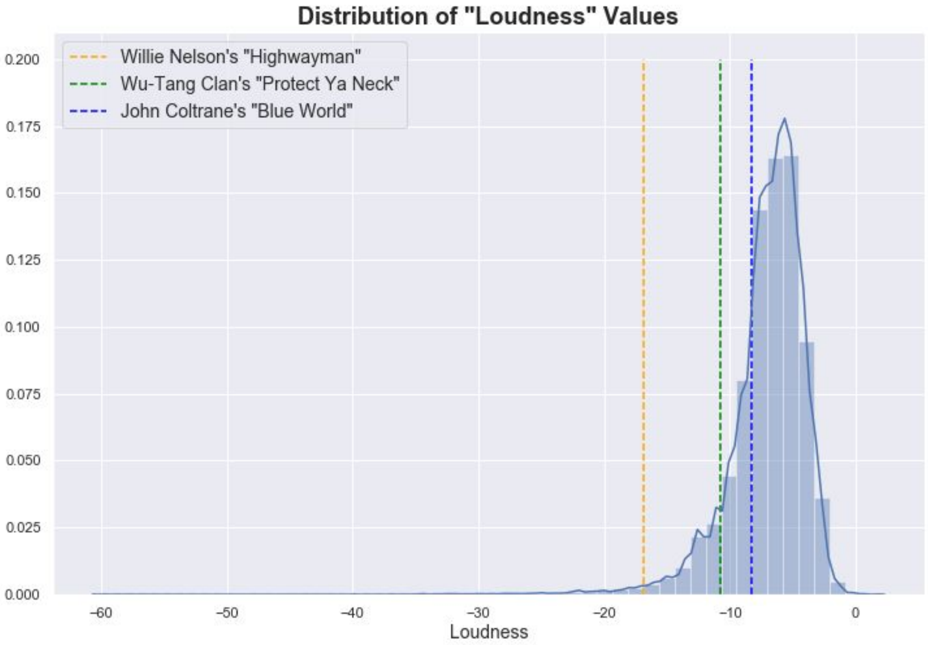
Exploring Features





Exploring Features

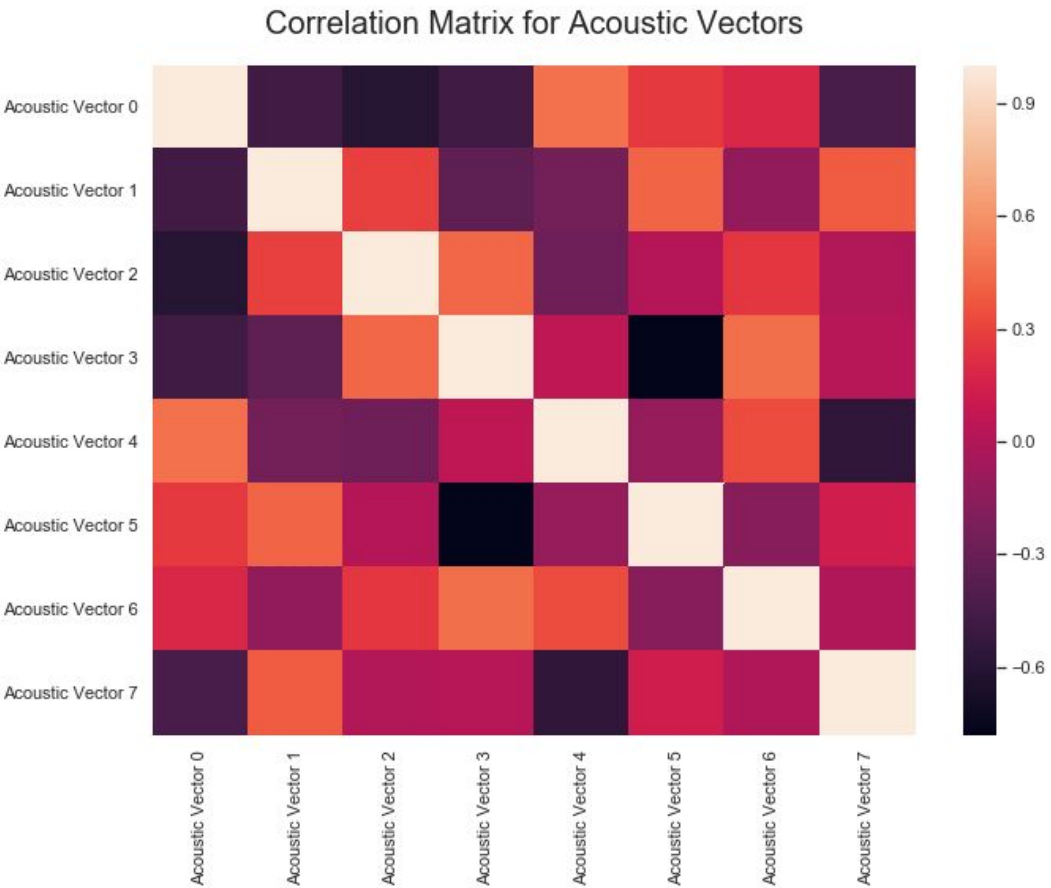


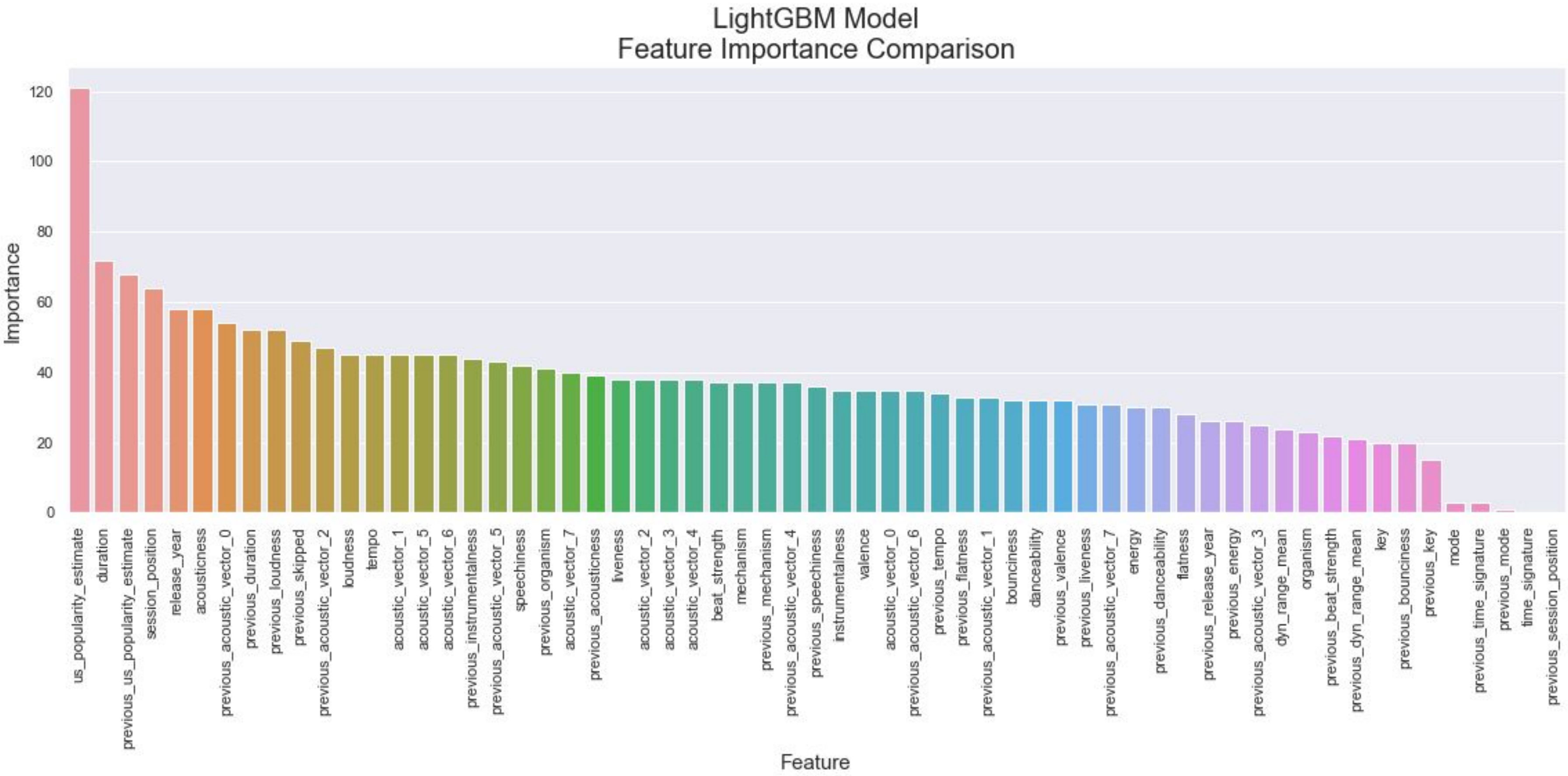


Exploring Features

Feature	Description
Acousticness	Likelihood a track is acoustic
Danceability	Describes how suitable a track is for dancing
Energy	Measure of a track’s intensity
Valence	Level of “positiveness” conveyed by a track
Speachiness	Detects the presence of spoken word

Acoustic Vector Correlations




LightGBM Model
Feature Importance Comparison
Importance
Feature
0
20
40
60
80
100
120
us_popularity_estimate
duration
previous_us_popularity_estimate
session_position
release_year
acousticness
previous_acoustic_vector_0
previous_duration
previous_loudness
previous_skipped
previous_acoustic_vector_2
loudness
tempo
acoustic_vector_1
acoustic_vector_5
acoustic_vector_6
previous_instrumentalness
previous_acoustic_vector_5
speechiness
previous_organism
acoustic_vector_7
previous_acousticness
liveness
acoustic_vector_2
acoustic_vector_3
acoustic_vector_4
beat_strength
mechanism
previous_mechanism
previous_acoustic_vector_4
previous_speechiness
instrumentalness
valence
acoustic_vector_0
previous_acoustic_vector_6
previous_tempo
previous_flatness
previous_acoustic_vector_1
bounciness
danceability
previous_valence
previous_liveness
previous_acoustic_vector_7
energy
previous_danceability
flatness
previous_release_year
previous_energy
previous_acoustic_vector_3
dyn_range_mean
organism
previous_beat_strength
previous_dyn_range_mean
key
previous_bounciness
previous_key
mode
previous_time_signature
previous_mode
time_signature
previous_session_position

Flask App with D3 Visualizations