



Spotify Sequential Skip Prediction Challenge

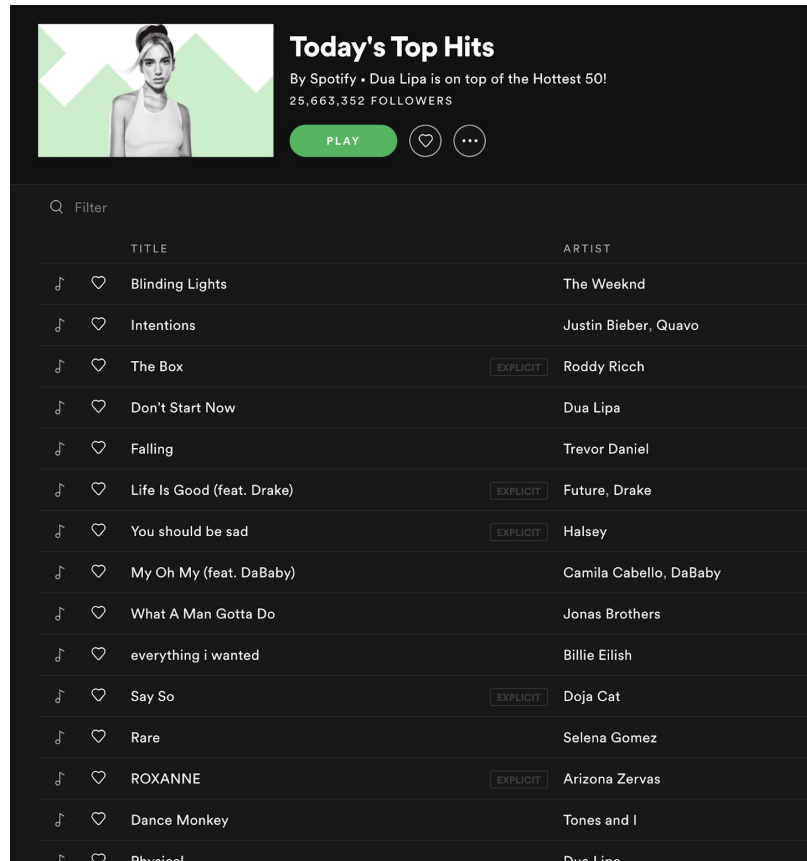
Austin Poor

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*



The screenshot shows the Spotify 'Today's Top Hits' playlist. At the top, there's a featured image of Dua Lipa and the text 'Today's Top Hits' followed by 'By Spotify • Dua Lipa is on top of the Hottest 50!' and '25,663,352 FOLLOWERS'. Below this is a 'PLAY' button and icons for a heart and a menu. A search bar with the text 'Filter' is present. The main content is a list of songs with columns for 'TITLE' and 'ARTIST'. Each song entry includes a play button icon, a heart icon, the song title, an 'EXPLICIT' label (if applicable), and the artist name.

	TITLE	ARTIST
🎵 ❤️	Blinding Lights	The Weeknd
🎵 ❤️	Intentions	Justin Bieber, Quavo
🎵 ❤️	The Box	EXPLICIT Roddy Ricch
🎵 ❤️	Don't Start Now	Dua Lipa
🎵 ❤️	Falling	Trevor Daniel
🎵 ❤️	Life Is Good (feat. Drake)	EXPLICIT Future, Drake
🎵 ❤️	You should be sad	EXPLICIT Halsey
🎵 ❤️	My Oh My (feat. DaBaby)	Camila Cabello, DaBaby
🎵 ❤️	What A Man Gotta Do	Jonas Brothers
🎵 ❤️	everything i wanted	Billie Eilish
🎵 ❤️	Say So	EXPLICIT Doja Cat
🎵 ❤️	Rare	Selena Gomez
🎵 ❤️	ROXANNE	EXPLICIT Arizona Zervas
🎵 ❤️	Dance Monkey	Tones and I
🎵 ❤️	Physical	Dua Lipa

Methodology

The Data

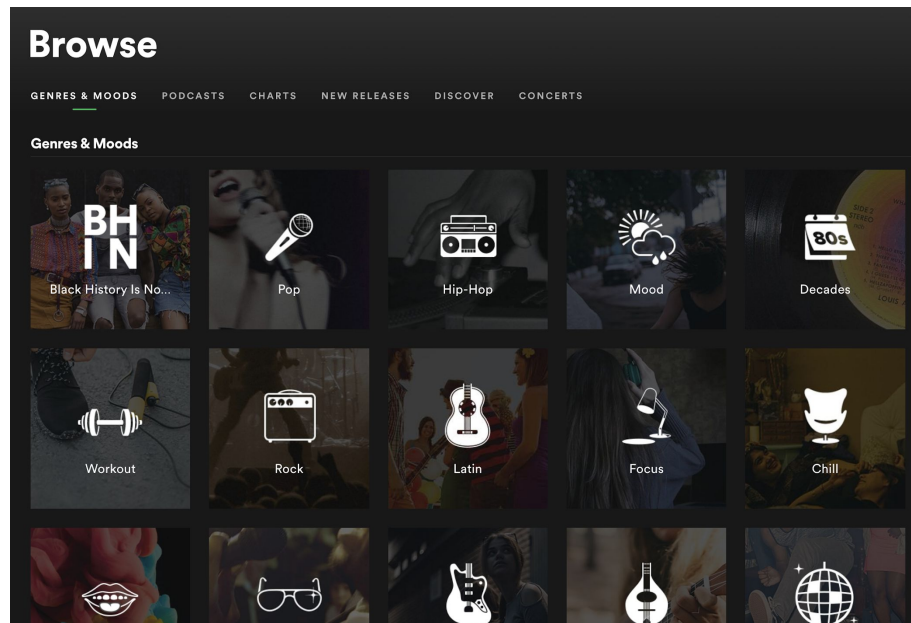
- 10-20 song sessions
- Song and user data was anonymized
- Balanced classes (0.517 skips)
- 350GB of CSV data
- Used subset of the data
 - 100k Session Rows
 - 50k Track Rows
- Stored data in PostgreSQL DB

training_set			track_features			acoustic_vectors		
session_id	text		track_id	text		track_id	text	
session_position	bigint		duration	double		acoustic_vector_0	double	
session_length	bigint		release_year	bigint		acoustic_vector_1	double	
track_id_clean	text	*	us_popularity_estimate	double		acoustic_vector_2	double	
skip_1	boolean		acousticness	double		acoustic_vector_3	double	
skip_2	boolean		beat_strength	double		acoustic_vector_4	double	
skip_3	boolean		bounciness	double		acoustic_vector_5	double	
not_skipped	boolean		danceability	double		acoustic_vector_6	double	
context_switch	bigint		dyn_range_mean	double		acoustic_vector_7	double	
no_pause_before_play	bigint		energy	double				
short_pause_before_play	bigint		flatness	double				
long_pause_before_play	bigint		instrumentalness	double				
hist_user_behavior_n_seekfwd	bigint		key	bigint				
hist_user_behavior_n_seekback	bigint		liveness	double				
hist_user_behavior_is_shuffle	boolean		loudness	double				
hour_of_day	bigint		mechanism	double				
date	text		mode	text				
premium	boolean		organism	double				
context_type	text		speechiness	double				
hist_user_behavior_reason_start	text		tempo	double				
hist_user_behavior_reason_end	text		time_signature	bigint				
			valence	double				

Feature Engineering

Generate features to account for a user's listening history

Added previous track features, including if that track was skipped

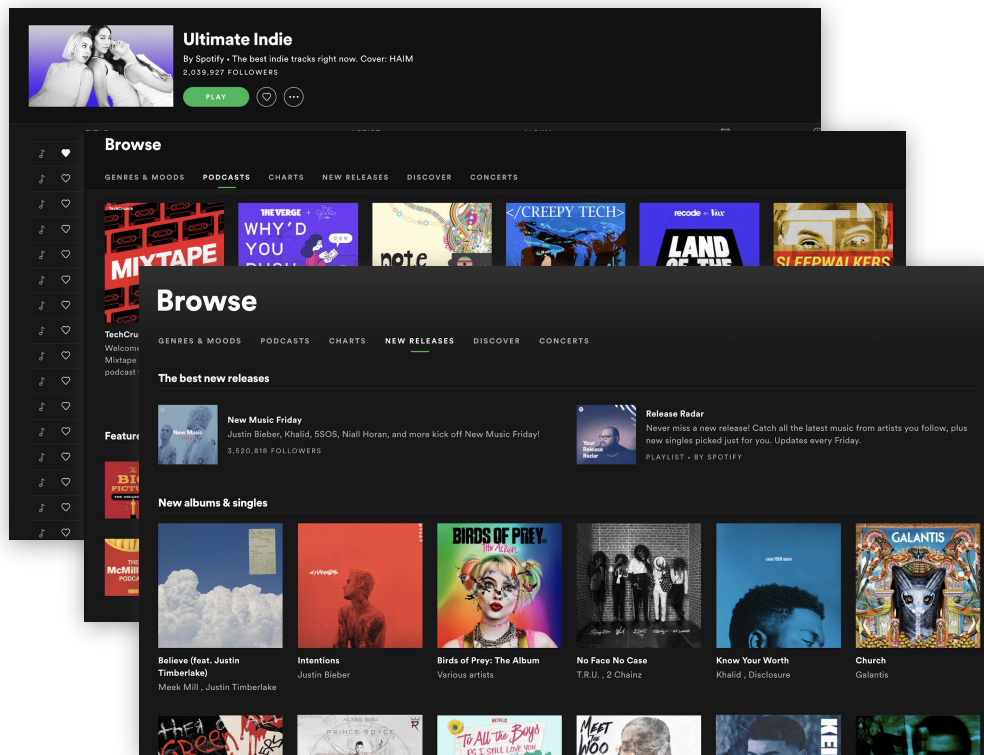


Model Selection

Target Metric **Accuracy** per competition guidelines

Baselined with Logistic Regression

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

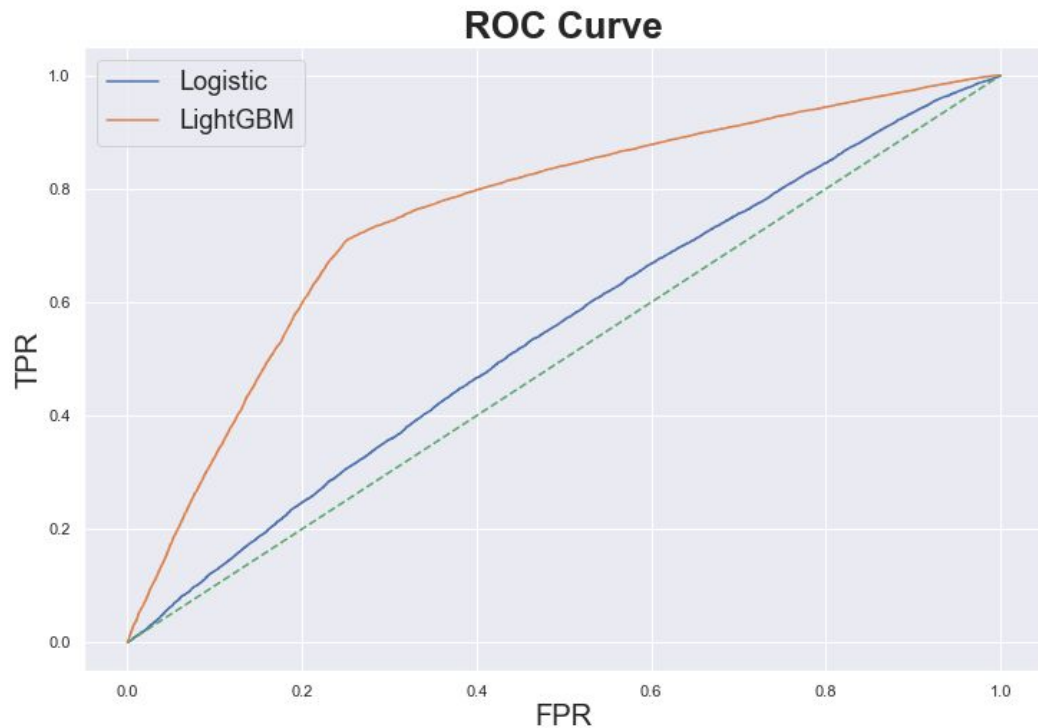


Results

Model Results

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

Error Analysis: No clear trend in the model residuals

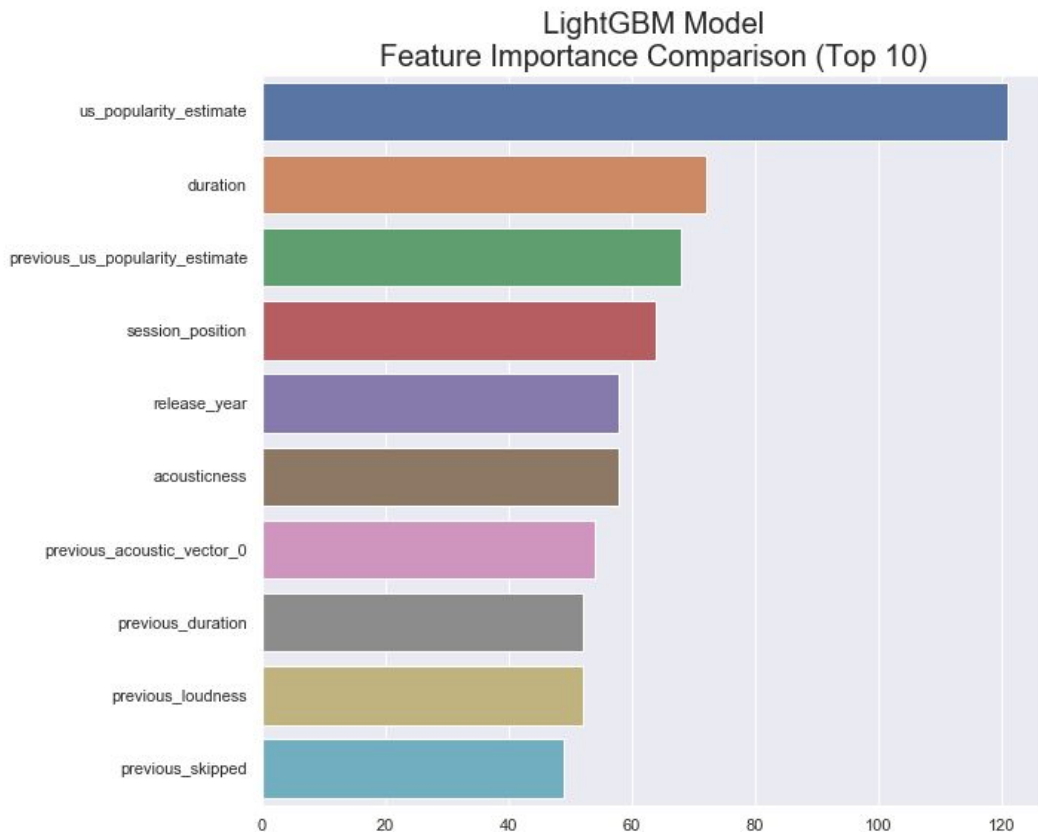


Conclusions

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

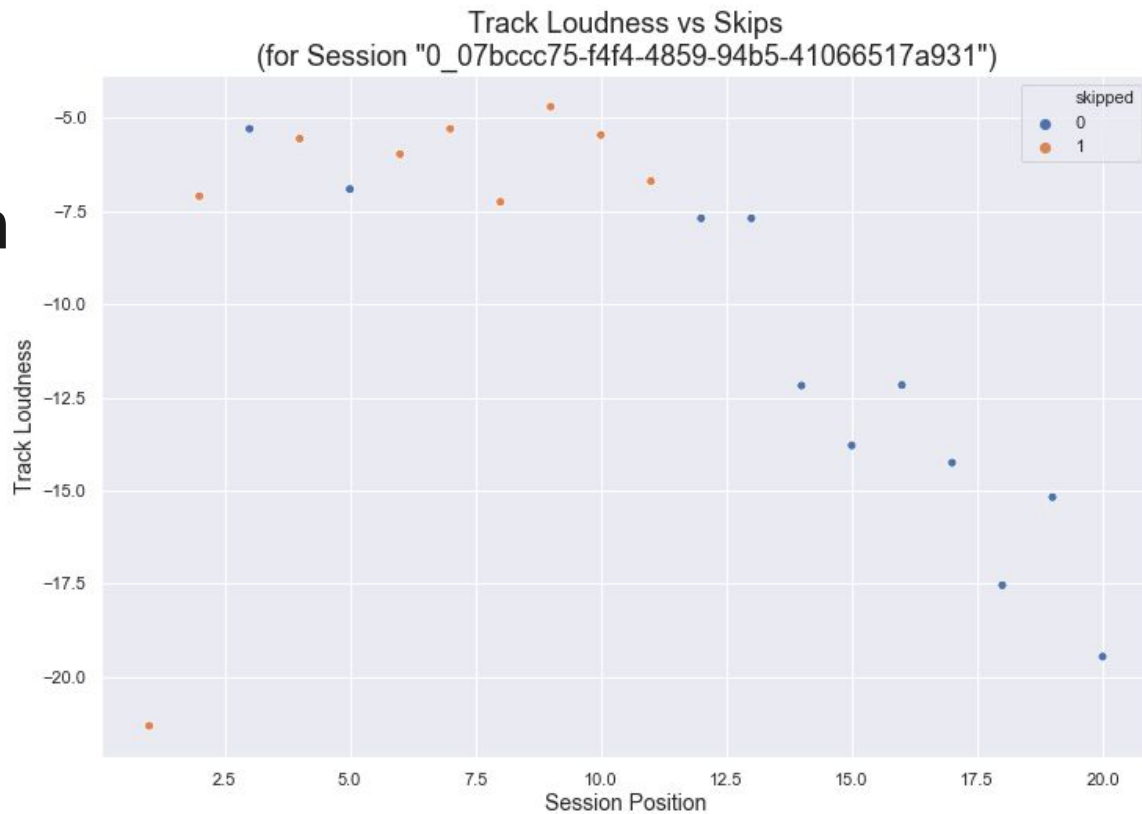
- Test other types of algorithms:
 - Unsupervised Learning to cluster songs
 - RNN to predict based on the sequence of tracks
- Supplement the dataset with more data from the Spotify API
- Create a Flask app to visualize predictions using D3

Thank you



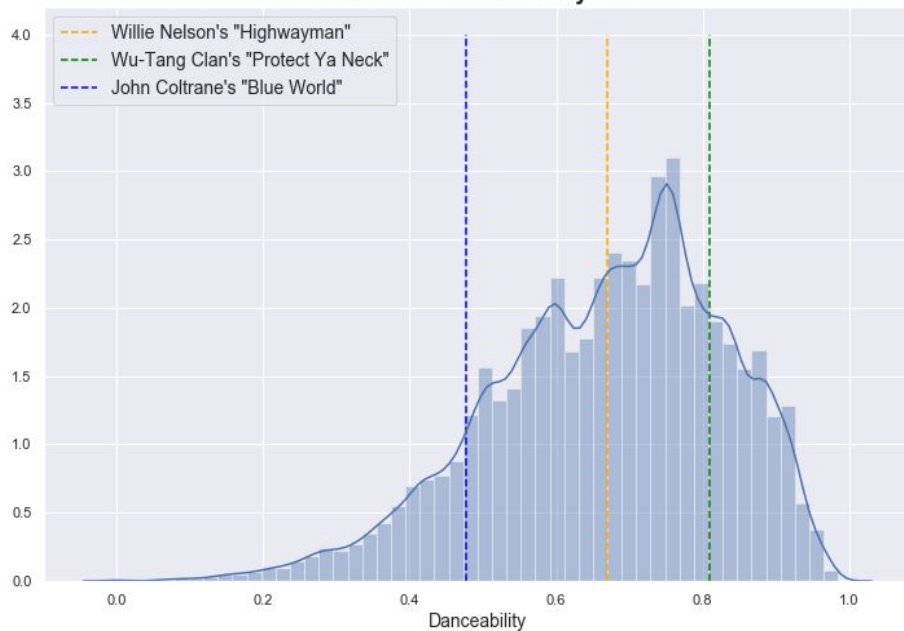
Appendix

Visualizing a Listening Session

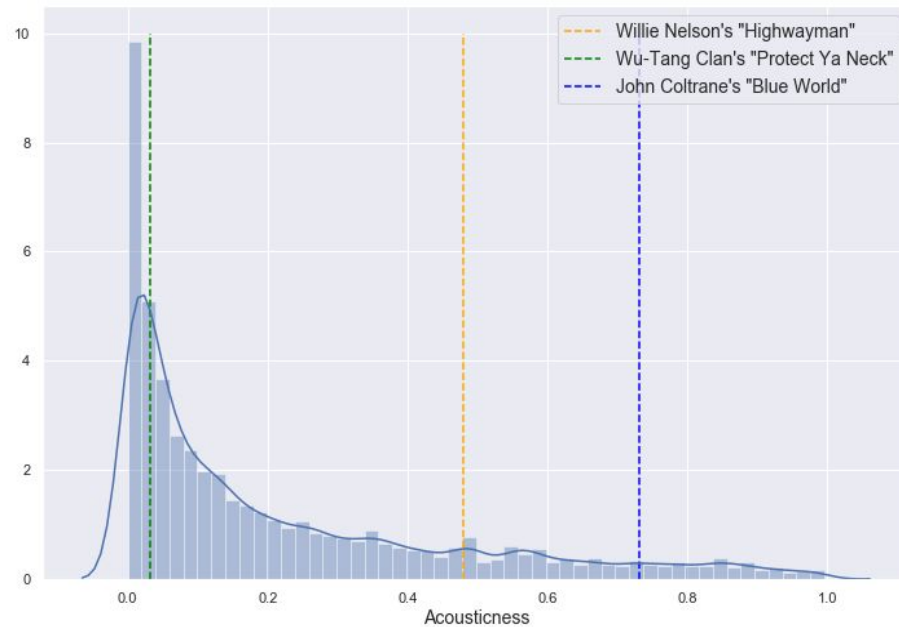


Exploring Features

Distribution of "Danceability" Values

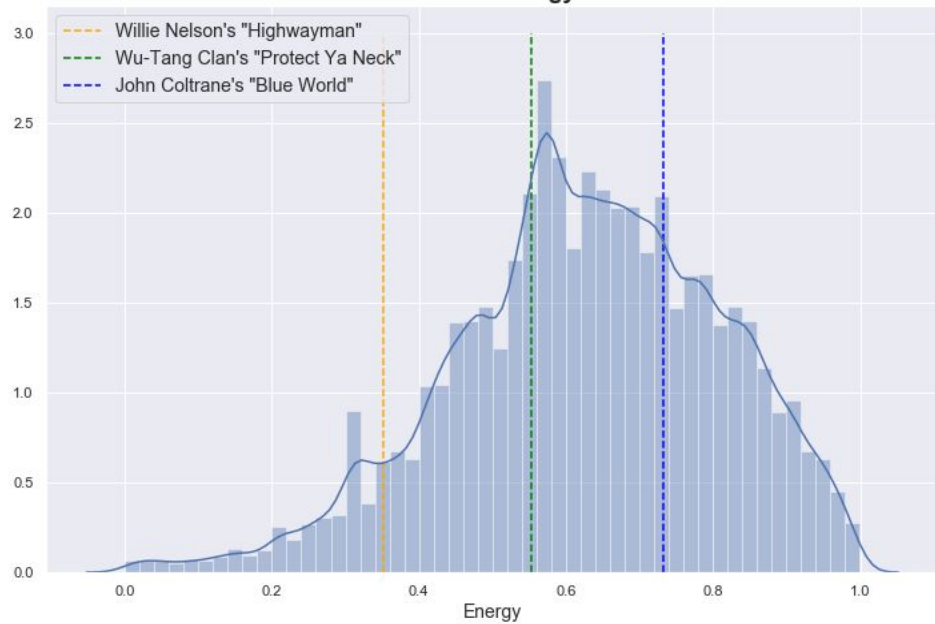


Distribution of "Acousticness" Values

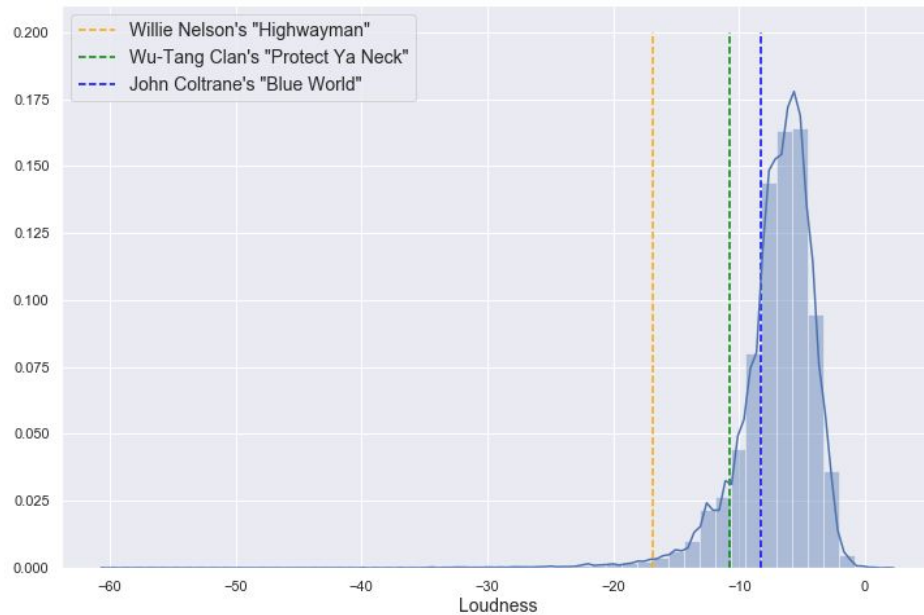


Exploring Features

Distribution of "Energy" Values



Distribution of "Loudness" Values



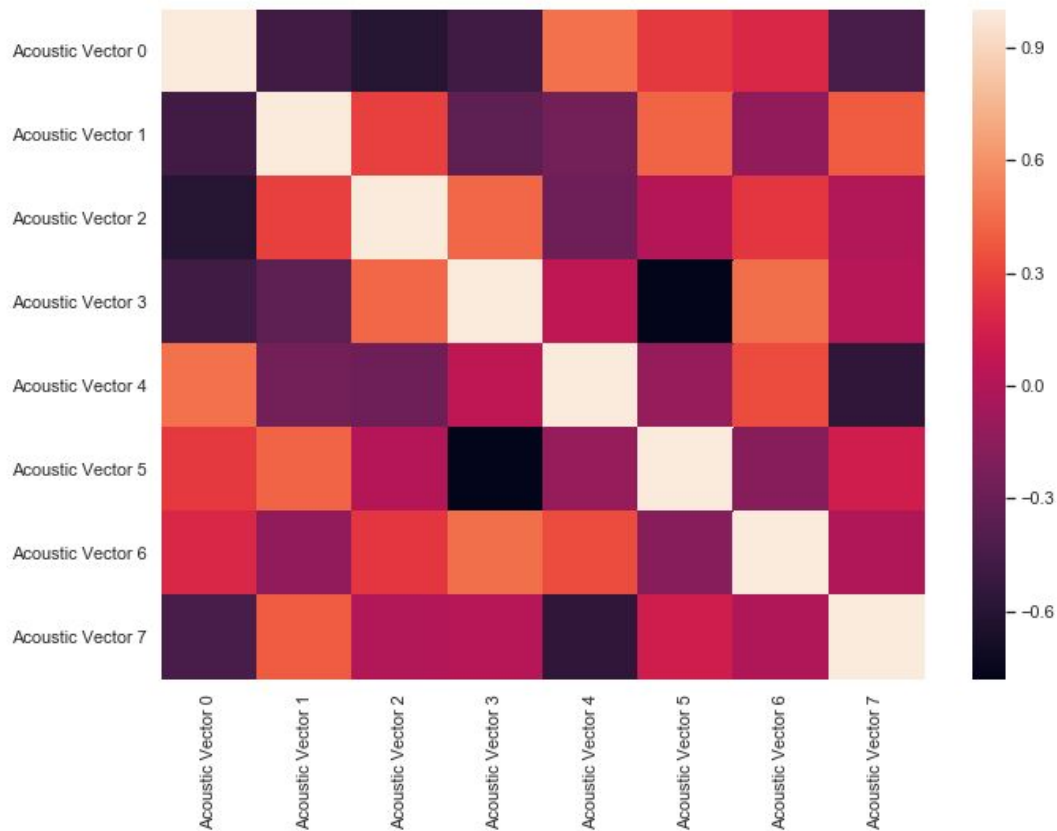


Exploring Features

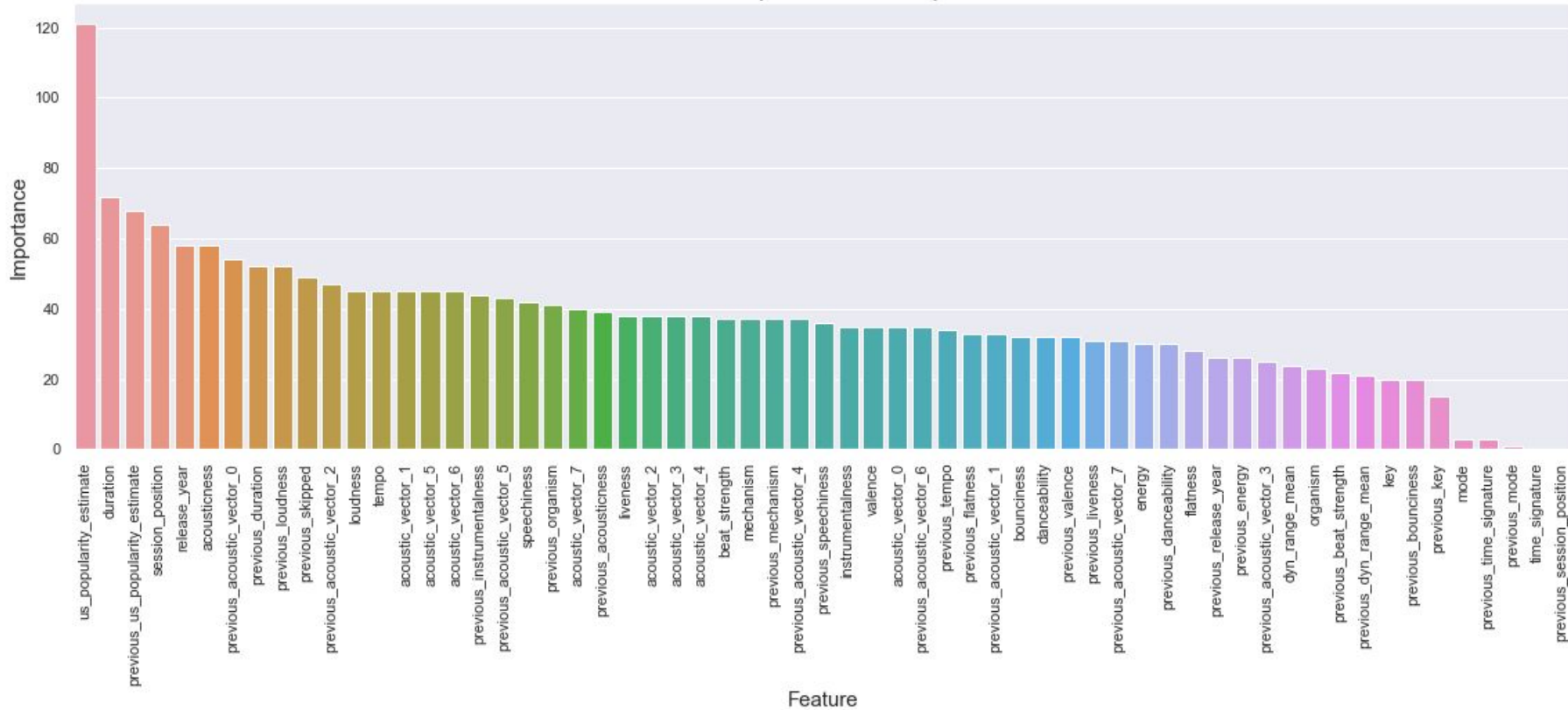
Feature	Description
<i>Acousticness</i>	Likelihood a track is acoustic
<i>Danceability</i>	Describes how suitable a track is for dancing
<i>Energy</i>	Measure of a track's intensity
<i>Valence</i>	Level of “positiveness” conveyed by a track
<i>Speechiness</i>	Detects the presence of spoken word


Acoustic Vector Correlations

Correlation Matrix for Acoustic Vectors



LightGBM Model
Feature Importance Comparison





Flask App with D3 Visualizations

