

Hit or miss?

Oskar Diyali and Logan Smith

Motivation and Reasoning

Motivation:

Can song features and lyrics of a song predict hits?

Why is predicting hit songs valuable?

- Artists
- Listeners
- Labels
- Streaming Platforms
- Marketing

What “hit” means in your project?

- Popularity, from Spotify API ≥ 45 cutoff

Guess whether it's a hit or miss!



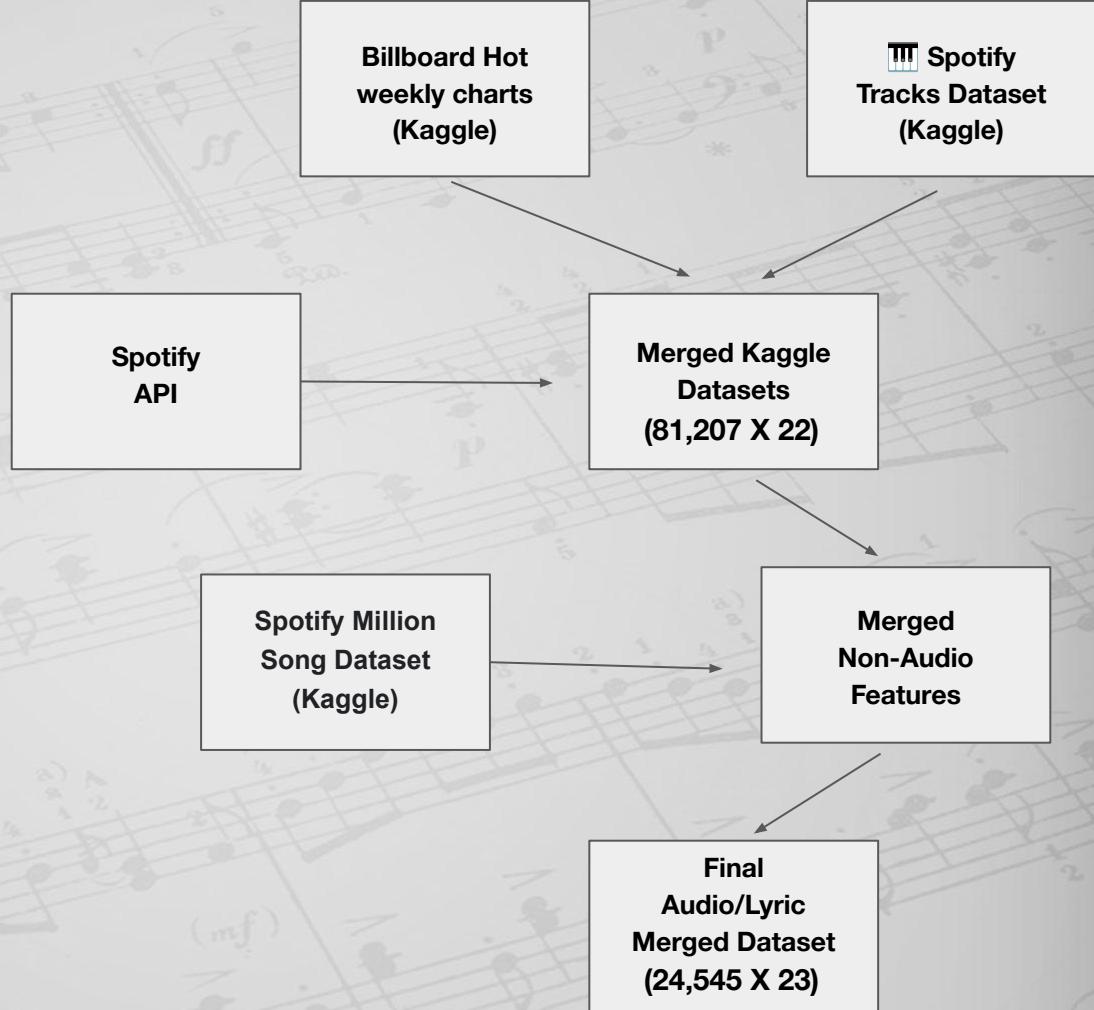
Dataset & Features

Numeric Features:

- **Audio:**
Danceability, Energy,
Speechiness, Tempo, Liveness,
Loudness, Instrumentalness,
Valence
- **Non-Audio:**
Tracks in Album, Artist
Popularity,
Artist Followers

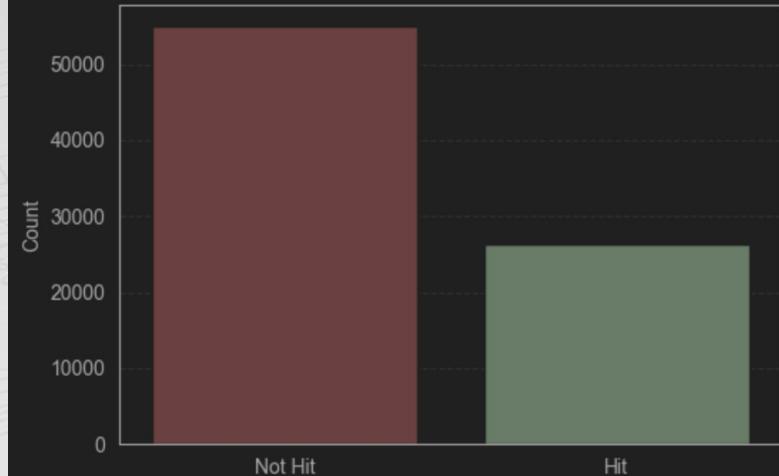
Categorical Features:

- Explicit, Mode, Album Type,
Label Group, Artist Genre,
Track Genre, Lyrics

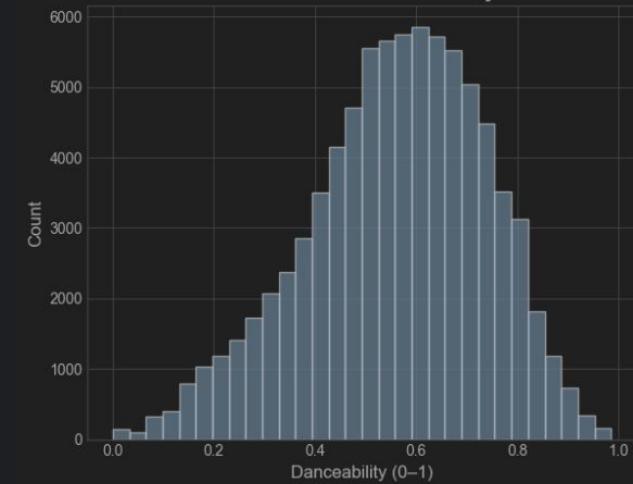


Exploratory Data Analysis

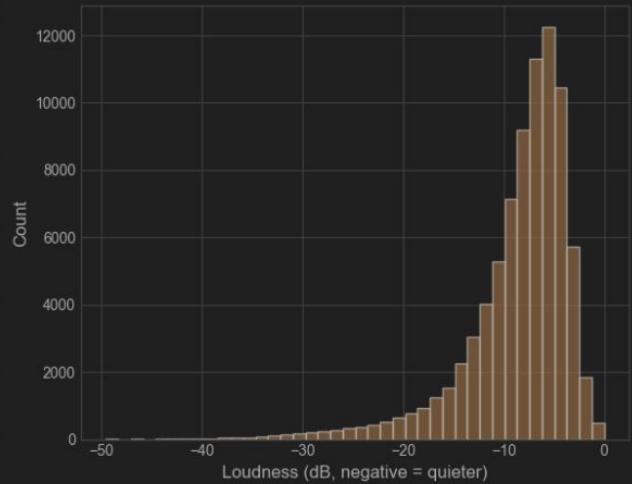
Class Balance: Hit vs Non-Hit



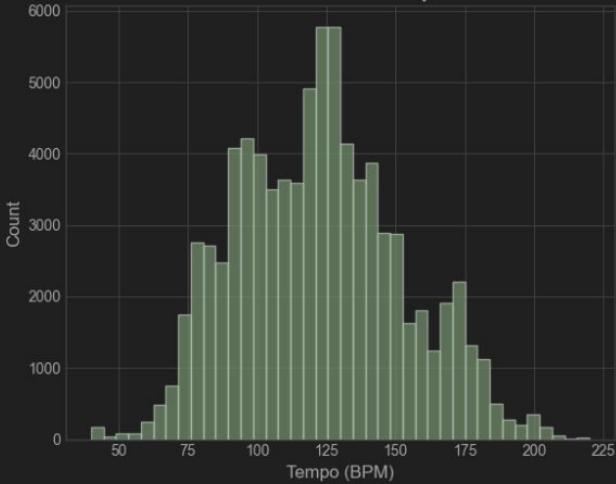
Distribution of Danceability



Distribution of Loudness



Distribution of Tempo



Challenges and Featured Engineering

Challenges in the Data

- Missing values
- Inconsistent formats in Genres and Labels
- Key stored as numbers, instead of musical notes
- Imbalanced target: only ~19% of songs are “hits”

Feature Engineering Solutions

- Extracted release_year, release_month, added flags
- Simplified genres and grouped labels into groups
- Log transforms
- Interaction of features

BEFORE (original features)

	duration_ms	tempo	loudness	energy	valence	danceability
0	230666.0	87.917	-6.746	0.461	0.715	0.676

AFTER (engineered features added)

	log_duration	log_tempo	log_loudness	energy_valence	dance_tempo
0	12.34873	4.487703	2.047177	0.329615	59.431892

Methods

Modeling Approach

- Train/test split (80/20, stratified).
- Models tested: Logistic Regression, Random Forest, LightGBM, CatBoost, **XGBoost (tuned)**.
- Evaluation metrics: Accuracy, Precision, Recall, F1, ROC-AUC.

Model Comparison

Model	Accuracy (%)
Tuned XGBoost	88.2%
LightGBM	88%
XGBoost (default)	87%
Random Forest	84%
Logistic Regression	82%

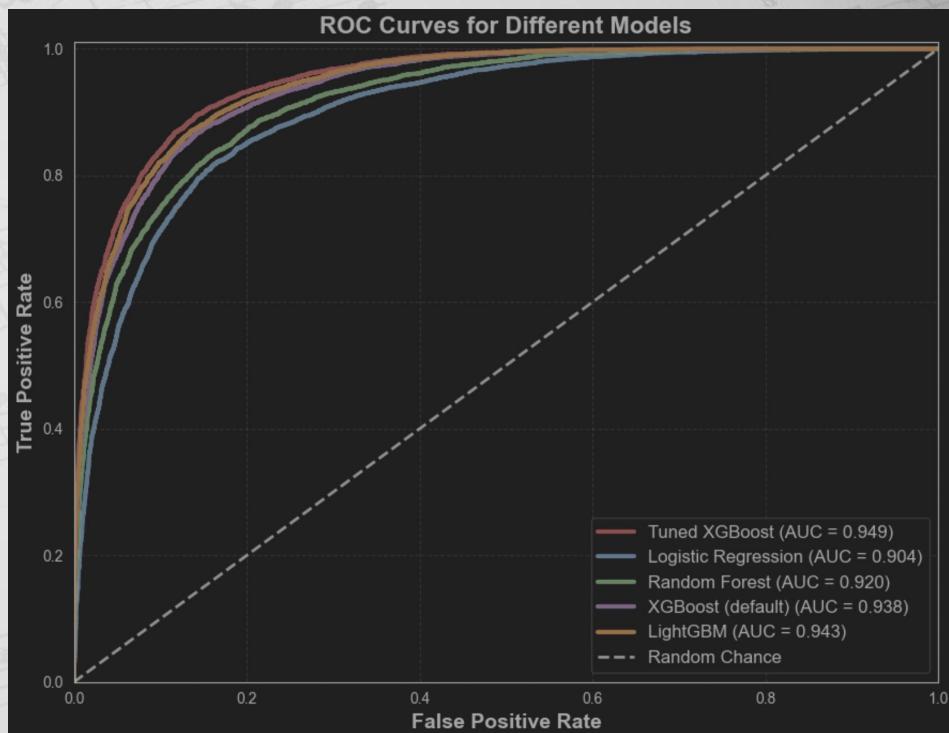
Tuned XGBoost (threshold=0.50)

	precision	recall	f1-score
0	0.909	0.918	0.913
1	0.824	0.807	0.815

accuracy

0.882

ROC-AUC Comparison



Methods pt. 2

Lyrical Analysis

- Standardized lyrics
- Converted lyrics into numeric features using term frequency-inverse document frequency (TF-IDF)
- Split and run logistic regression
- Extract top words
- WordCloud

The song was....

A HIT



Top 25 most common hit lyrics

A word cloud graphic featuring the top 25 most common hit lyrics. The words are arranged in a cluster, with larger words representing higher frequency. The colors of the words vary, creating a visual hierarchy. The words and their associated meanings are:

- send: broken
- send: nice
- send: orleans
- lean: hosanna
- lean: lay
- lean: hero
- blame: shaking
- blame: holiday
- blame: gloria
- blame: satellite
- blame: suffer
- blame: fair
- huh: blind
- imagination: seventeen
- imagination: human
- imagination: crush
- imagination: afraid
- imagination: moment

Reflection

Positive:

- Importance of non-audio features
- Learned TF-IDF

Negative:

- Remove duplicates
- Limitations of Spotify API

Any questions?

