INFO 6105 Final Project, Fall 2025

Music Hit Prediction

Decoding Popularity with Machine Learning

By Da Lei, Wei Li

GitHub Repository: https://github.com/radar-iscs/music-hit-prediction

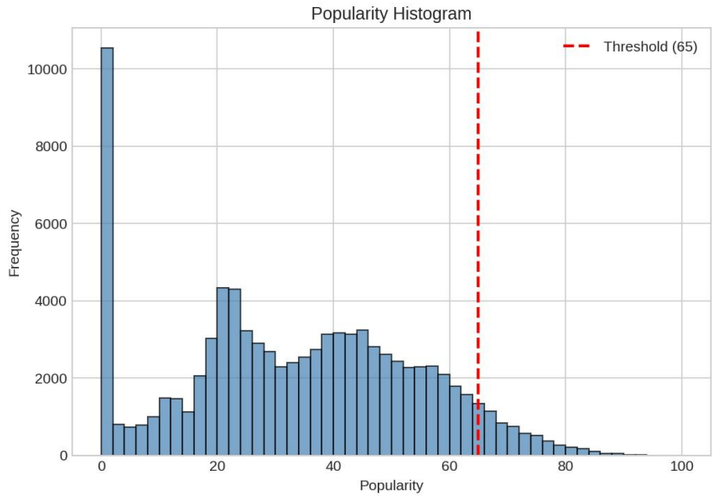
Problem

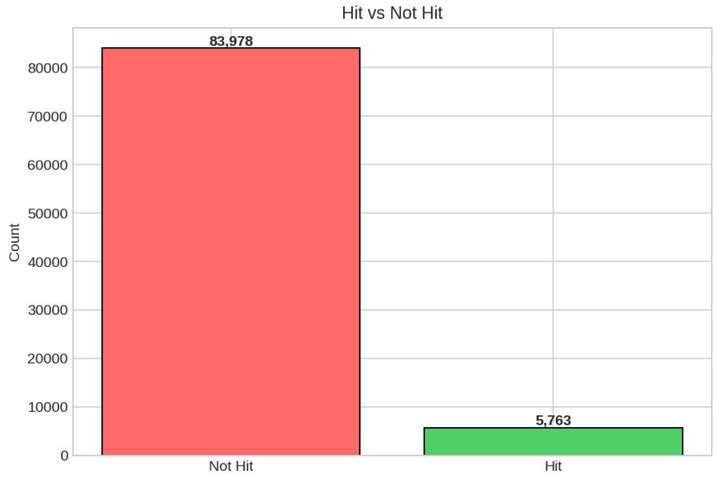
The Scale of the Problem

Over 100,000 songs are uploaded to streaming platforms every single day, creating an overwhelming digital tsunami where exceptional music drowns in the noise.

Threshold: We defined a "Hit" as having a Spotify popularity score ≥≥ 65 (red dashed line).

Severe Imbalance: Only 6.4% (5,763 tracks) are Hits, while 93.6% (83,978) are Not Hits





Solution: An End-to-End ML Pipeline

Massive Data Integration

89,000+ tracks from Spotify API and Last.fm, capturing comprehensive audio features and listener behavior patterns

Advanced Feature Engineering

Sophisticated analysis of duration, loudness, tempo, energy, and 12+ audio characteristics

Smart Targeting

Predicting tracks with popularity scores ≥65, identifying commercial hits before they break

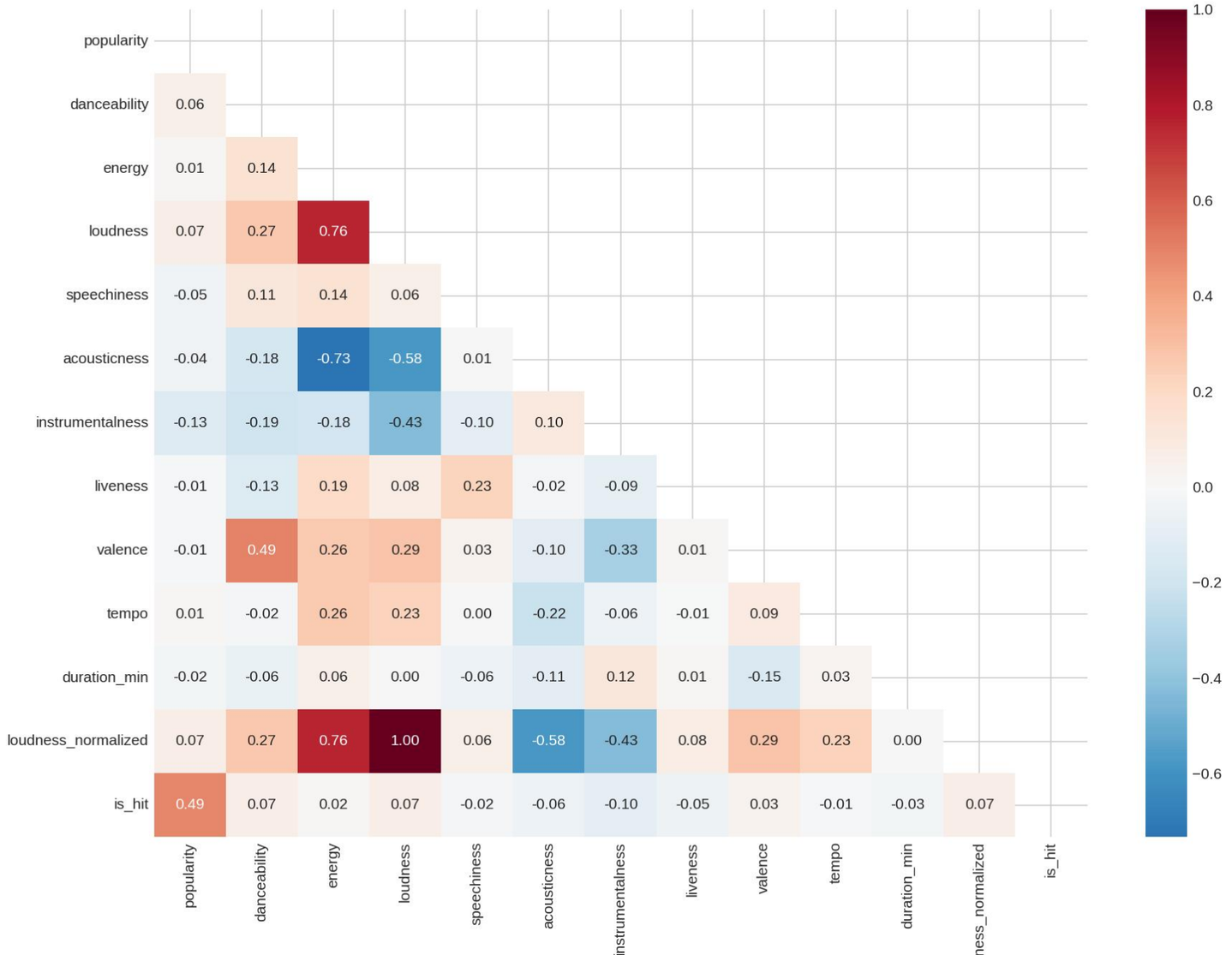
LVX VERITAS VIRTVS
Northeastern University

Data Analysis

Key Correlations

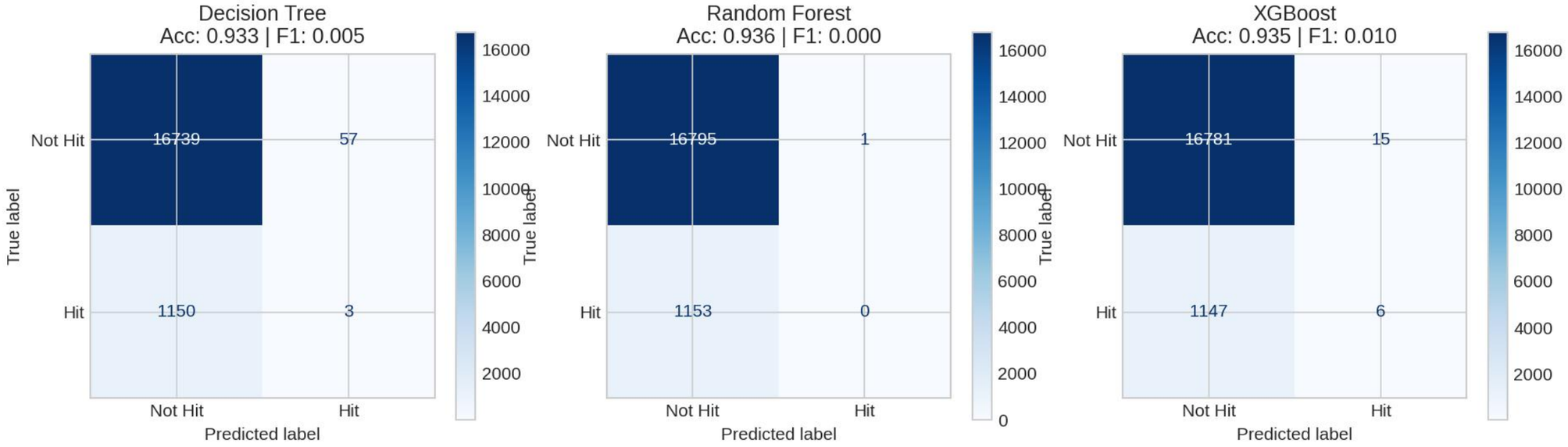
· Loudness and Energy are highly correlated (0.76), suggesting we might not need both

· Acousticness has a strong negative correlation with Energy (-0.73), confirming that energetic hits are rarely acoustic



Model Performance Metrics

- The Bias Problem: All three models (Decision Tree, Random Forest, XGBoost) excel at predicting "Not Hits" (High True Negatives)
- The Struggle: They struggle significantly to identify the minority "Hits" (Low True Positives)
- Random Forest: Predicted almost zero hits, defaulting to the majority class.
- XGBoost & Decision Tree: Showed slightly better ability to detect hits, but false negatives remain high due to the 93:7 class imbalance

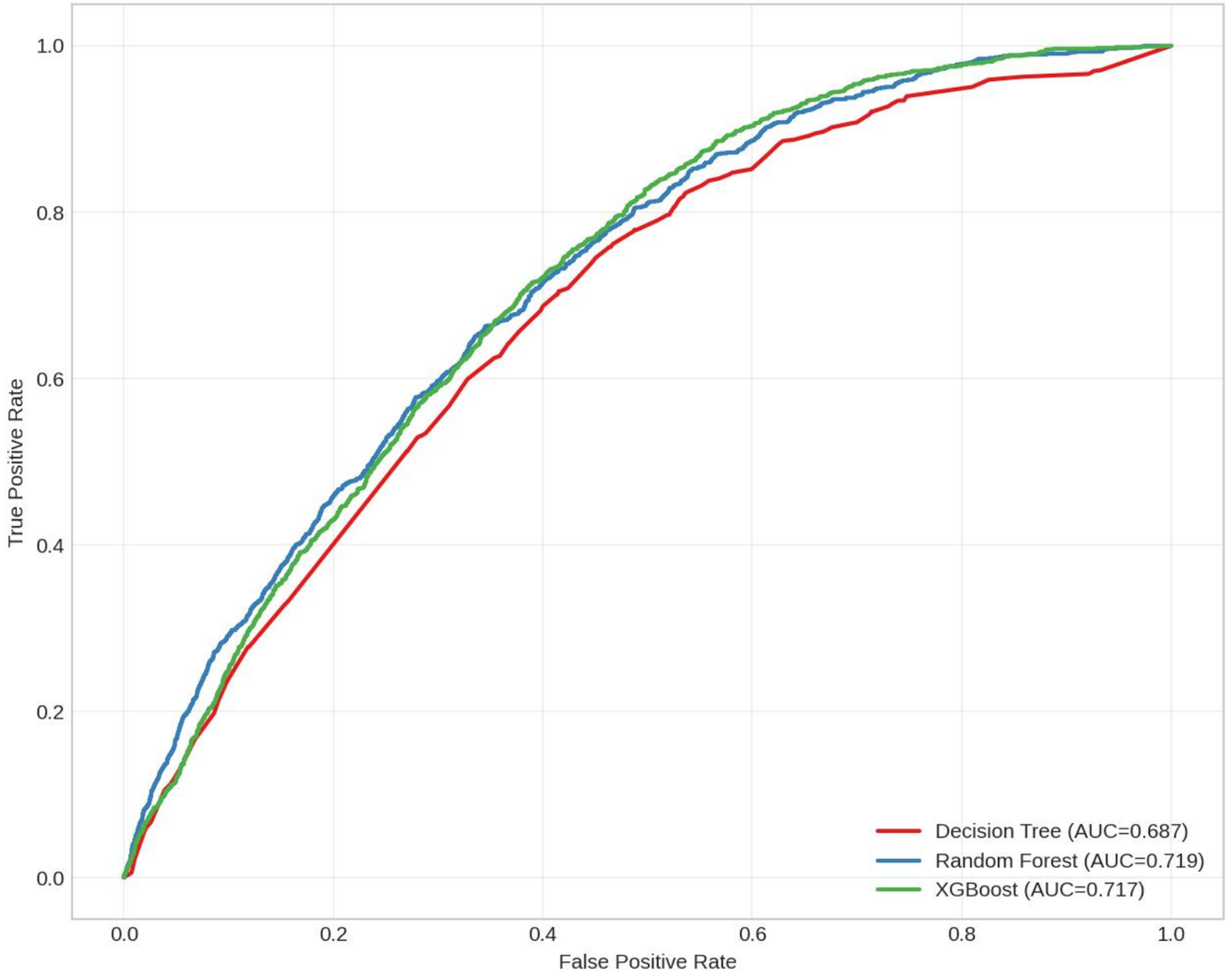


Model Evaluation

• Separation Ability: The AUC scores (~0.72) indicate the models are learning and performing better than random guessing (0.50)

• Model Comparison: Random Forest (0.719) and XGBoost (0.717) slightly outperform the Decision Tree (0.687)

• Trade-off: While Random Forest has a slightly higher AUC, the Decision Tree was chosen for its interpretability and balance in precision/recall trade-offs



What Makes a Hit?

Key Predictive Features

Our model identified three critical audio characteristics that distinguish commercial hits from the rest:

Tempo

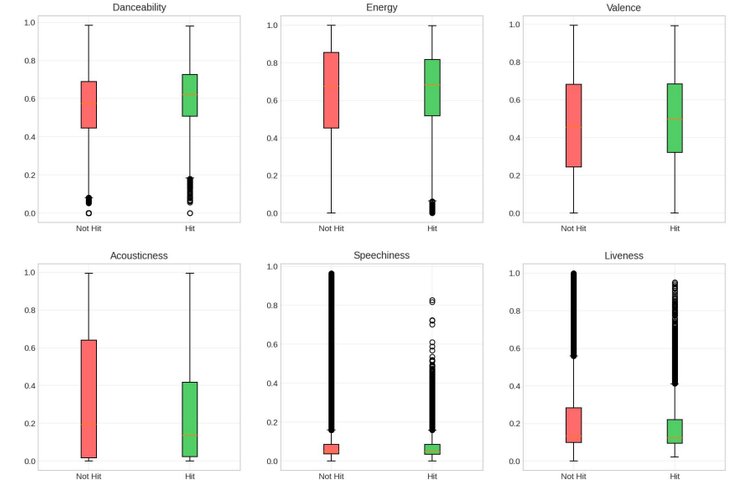
is the #1 most important predictor for a hit song

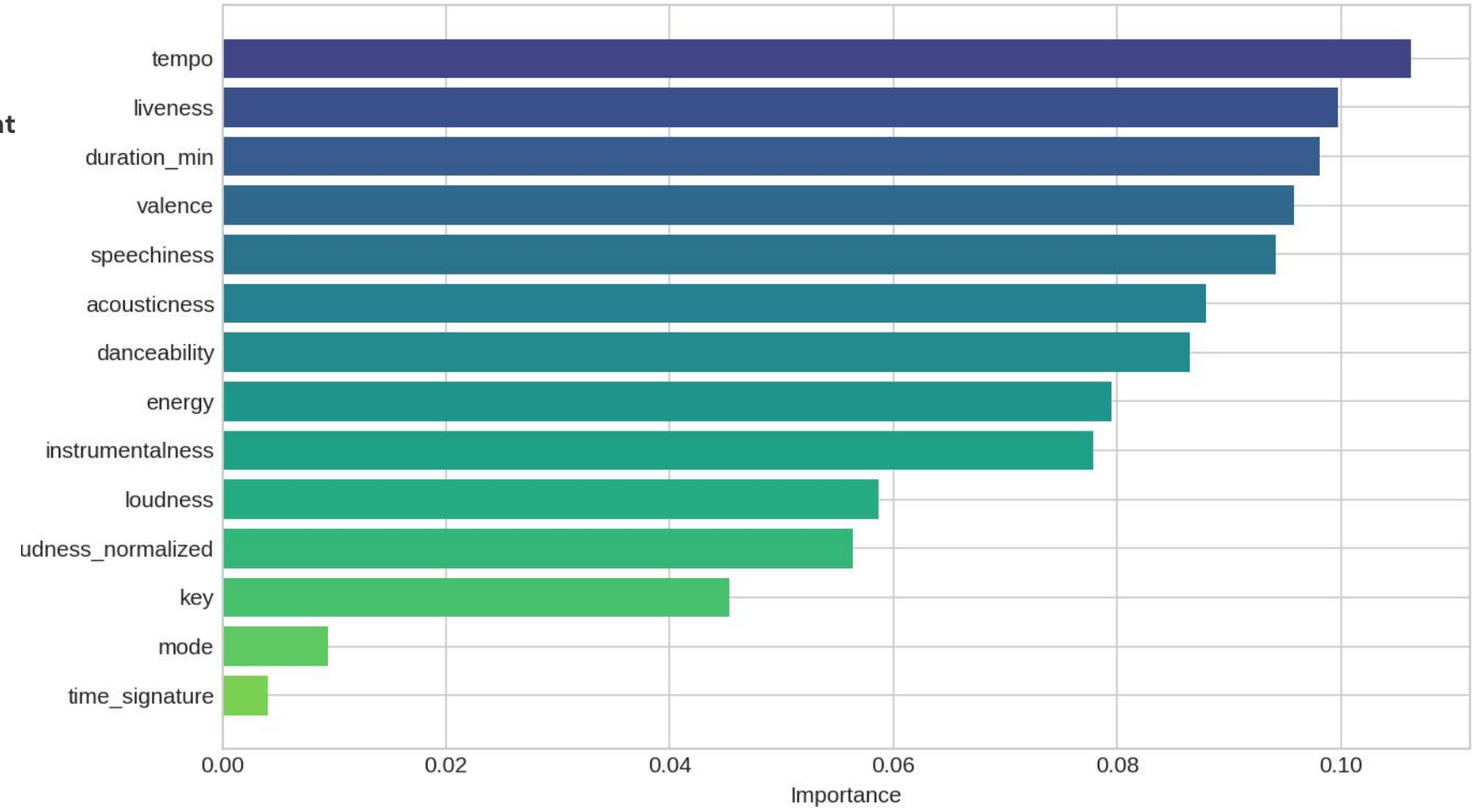
Liveness

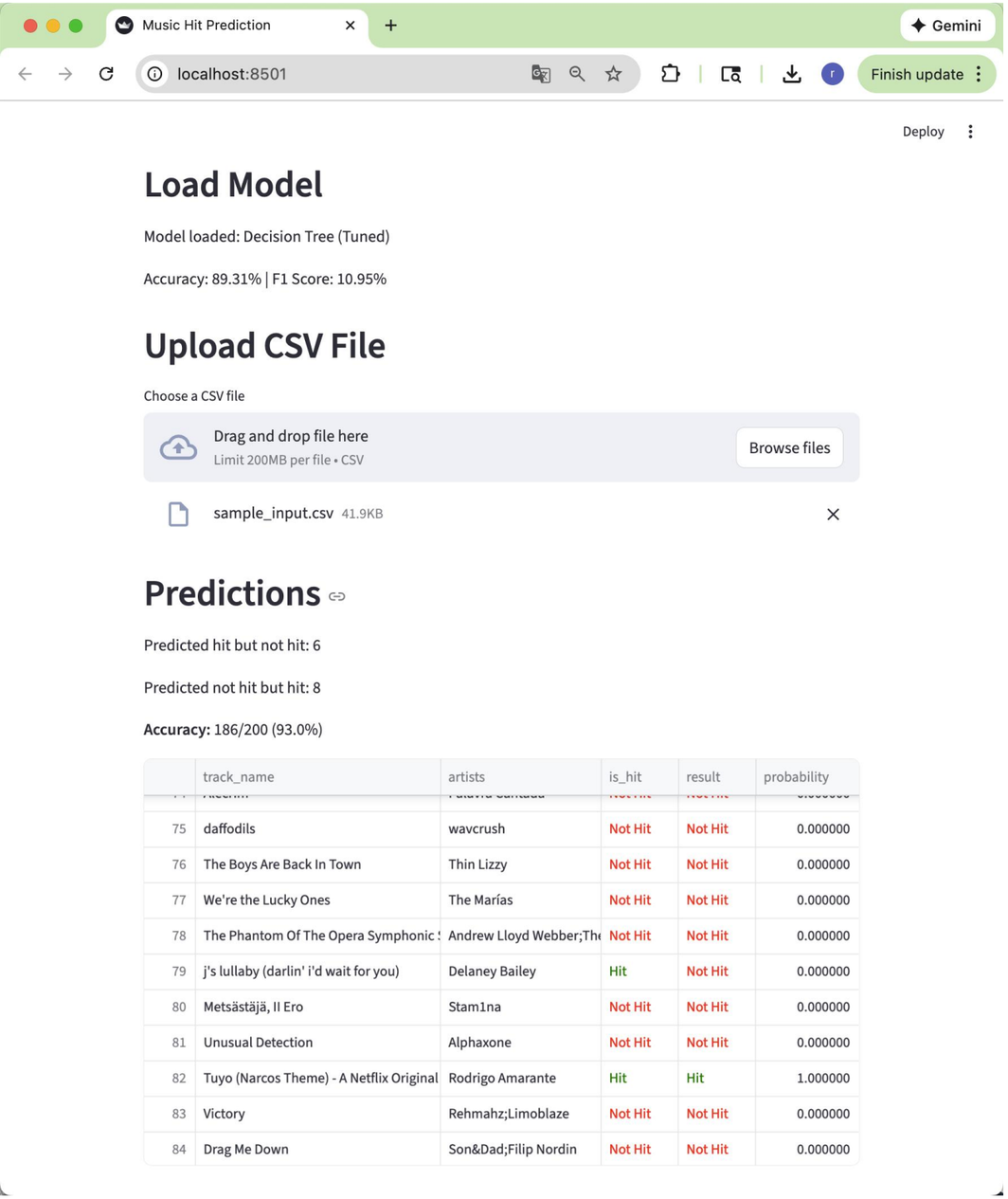
Authentic performance quality drives listener engagement

Duration

Optimal track length maximizes replay value and playlist inclusion







	track_name	artists	is_hit	result	probability
75	daffodils	wavcrush	Not Hit	Not Hit	0.000000
76	The Boys Are Back In Town	Thin Lizzy	Not Hit	Not Hit	0.000000
77	We're the Lucky Ones	The Marías	Not Hit	Not Hit	0.000000
78	The Phantom Of The Opera Symphonic	Andrew Lloyd Webber;The	Not Hit	Not Hit	0.000000
79	j's lullaby (darlin' i'd wait for you)	Delaney Bailey	Hit	Not Hit	0.000000
80	Metsästäjä, II Ero	Stam1na	Not Hit	Not Hit	0.000000
81	Unusual Detection	Alphaxone	Not Hit	Not Hit	0.000000
82	Tuyo (Narcos Theme) - A Netflix Original	Rodrigo Amarante	Hit	Hit	1.000000
83	Victory	Rehmahz;Limoblaze	Not Hit	Not Hit	0.000000
84	Drag Me Down	Son&Dad;Filip Nordin	Not Hit	Not Hit	0.000000

Streamlit Web Application

We built an intuitive Streamlit web app specifically designed for A&R executives and label decision-makers. The platform enables instant hit probability assessment through simple CSV upload, delivering actionable insights in seconds.

Upload a batch of tracks, receive probability scores, and make data-driven signing decisions with confidence. No technical expertise required—just upload and predict.

LVX VERITAS VIRTVS Northeastern University

Real-World Impact

Reduce Financial Risk

A&R teams can evaluate hundreds of demos instantly, focusing investment on tracks with proven commercial potential and avoiding costly mistakes on low-probability artists.

Empower Independent Labels

Indie labels gain enterprise-level analytics capabilities, leveling the playing field against major label resources and discovering hidden gems in their catalogs.

Combat Algorithmic Bias

Surface quality music that platform algorithms might overlook, ensuring talented artists receive the attention they deserve regardless of existing follower counts.

LVX VERITAS VIRTVS Northeastern University

Future Enhancements

Data

Fetch data from newly released songs in other ways

Features

Add more features like artist follower count, languages, professional ratings

Model

Combine multiple models to improve the accuracy of predictions

LVX VERITAS VIRTVS
Northeastern University