Explainable AI (XAI) for Music Recommendation Systems: A Model Transparency Study

AASHIQ RAHAMAN

UJJWAL KAMILA

IMTIYAZ AHMAD

BTech Computer Science and Engineering, 7th Semester
Roll: 10000122020, Reg: 221000110175

Roll: 10000122033, Reg: 221000110162

Roll: 10000122054, Reg: 221000110181

Under the guidance of

Prof. Debasis De

Maulana Abul Kalam Azad University of Technology, W.B

November 2025

Abstract

Music recommendation algorithms drive user engagement on streaming platforms, yet their internal decision processes often remain opaque. Users typically receive suggestions without any explanation, contributing to the growing issue of mistrust toward algorithmic systems. In this study, we train a Random Forest model on a real-world song dataset and — importantly — apply Explainable AI (XAI) frameworks (LIME and SHAP) to interpret its functionality. According to our IPython notebook execution, the model achieves a perfect Accuracy of 1.0 and an F1-score of 1.0 on the test set. While these metrics are undeniably strong, they must be interpreted in the context of dataset characteristics, particularly slight class imbalance toward non-popular tracks.

We analyze not only what the model predicts, but why it predicts so. LIME explanations reveal local reasoning for individual predictions, while SHAP analysis uncovers global feature importance trends. We also discuss limitations of popularity-based metrics and ethical implications, including the risks of algorithmic homogenization and genre suppression. Ultimately, this work advocates for transparent recommendation systems and demonstrates how XAI methods create interpretable, accountable, and user-trustworthy AI.

1. Introduction

1.1 Background and Motivation

Music streaming services like Spotify, Apple Music, and YouTube Music host massive song catalogs — often exceeding 100 million tracks. Since users cannot reasonably explore such volume manually, recommendation systems have become central to music discovery. However, these systems often operate invisibly: recommendations are experienced, but the reasoning behind them remains hidden.

Many users express sentiments such as:

> "I like the recommendations, but I don't know why they're recommended to me."

This is the essence of the "black-box" problem in recommendation AI.

1.2 The Imperative of Explainability

Pure predictive accuracy is insufficient when recommendations influence culture, listening habits, artist visibility, and even revenue streams. If a model silently suppresses entire genres due to learned biases, it introduces hidden discrimination.

Thus, we employ Explainable AI (XAI) — a methodology for translating algorithmic behavior into human-interpretable reasoning.

2. Dataset and Preprocessing

We used a Kaggle-sourced dataset consisting of approximately 170,000 songs. Each track includes numeric features such as energy, tempo, acousticness, instrumentalness, danceability, etc.

2.1 Target Classification

Popularity was originally a numeric score between 0 and 100. We binarized it into:

- Popular (1)
- Not Popular (0)

2.2 Data Splitting and Scaling

- Train-test split: 80/20
- Standardization: mean = 0, std = 1

These steps were implemented exactly as shown in the IPython notebook.

3. Model Training

A Random Forest Classifier was trained on the normalized feature set. Based on the notebook outputs, the model evaluated on the test data returned:

- Accuracy = **1.0**
- F1 Score = **1.0**

Since these values came directly from executed notebook prints, they are reported as exact.

It is rare to achieve perfect scores — this suggests either:

1. The Random Forest effectively captured the underlying patterns, or
2. The classification boundary was unusually clean, or
3. Class imbalance and simplicity of the decision boundary contributed to trivial classification.

Rather than assuming flawless performance, we analyze the possible reasons behind the perfect metrics.

4. Explainability Analysis

4.1 LIME – Local Explanation

One analysed output from the notebook showed:

> “The RF predicted: 0”

meaning the classifier predicted Not Popular.

LIME generated a feature importance ranking for that specific prediction. It demonstrated which song attributes pulled the prediction toward Not Popular. This was visible in the explanation weights LIME generated per instance.

4.2 SHAP – Global Explanation

SHAP analysis provided an overview of:

- which features consistently drive predictions upward
- which features drive predictions downward

From our analysis:

- high instrumentalness reduced likelihood of popularity
- high energy boosted likelihood of popularity
- acousticness generally correlated negatively with popularity
- danceability contributed positively

Unlike LIME, SHAP values can be aggregated over all instances for systemic understanding.

5. Discussion

The perfect metrics initially suggest an ideal model — but XAI shows that the model is not magical; it is pattern-exploiting. Given the slight class imbalance toward “Not Popular” songs, and the fact that musical popularity has strong statistical correlations with a few dominant features, the model may succeed due to structural simplicity rather than deep reasoning.

Explainability allows us to:

- detect bias

- uncover oversimplifications
- identify dataset limitations

For example, if speechiness is associated with lower popularity, the model may inadvertently penalize rap music. Without explainability, this bias would remain invisible.

6. Conclusion and Future Work

Our application of XAI demonstrates how perfect evaluation scores can be misleading unless supported by interpretability. Rather than celebrating a 1.0 accuracy, we scrutinize it. This study shows that:

1. XAI yields fine-grained accountability
2. Transparency improves trust
3. Interpretability supports ethical AI use

6.1 Future Research Directions

- incorporating tempo-dependent mood modeling
- generating user-customized explanations in natural language
- recognizing genre-sensitive fairness constraints
- experimenting with other models (XGBoost, Gradient Boosting, Deep Learning)
- validating with user behavioral feedback loops

7. References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?"
2. Lundberg, S. M., & Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions."
3. Breiman, L. (2001). Random Forests.
4. Kaggle Dataset Source.