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 boost user engagement on streaming platforms, but their internal decision making process often opaque and users receive suggestion with no explanation, which is a growing problem that lack of trust on algorithmic systems. Here, we employ Explainable AI frameworks (LIME and SHAP) to explain the behaviour of a Random Forest model that we trained on a real-world song dataset. Our IPython notebook execution shows that the model achieves an F1-score of 1.0 and a perfect Accuracy of 1.0 on the test set. Although these metrics are undoubtedly strong, we must also take into consideration the characteristics of the dataset, such as the small class imbalance in favour of less popular songs.

We both analyse model prediction and explanation for them. While SHAP analysis uncovers global trends in feature importance. LIME explanations provide local reasoning for individual predictions. They also discuss the ethical concerns of popularity-based metrics, such as the dangers of algorithmic homogenization and genre suppression and limitations of these metrics. Finally, this work demonstrate how XAI techniques yield interpretable, accountable, and user-trustworthy AI and advocates for transparent recommendation systems.

1. Introduction

1.1 Background and Motivation

With sheer volume of content available modern streaming platforms like Spotify, Apple Music, and YouTube Music (all which have catalogues of over 100 million songs), manual exploration is nearly impossible, and recommendation systems have become a necessity for guiding users to discover music within these huge libraries. This means that recommendation systems are often black boxes that deliver personalized suggestions but do not reveal the logic that led to those suggestions, so users do not know which features, behaviours, or historical preferences caused a recommendation. This is due to lack of transparency makes it difficult to trust the recommendation system, to hold it accountable, or to ensure it fair as its logic is hidden and users cannot identify whether there are biases that might be amplifying certain genres and artists over others. Given that these systems impact what people experience and how they listen to music at a massive scale this lack of transparency is not only a technical issue but also cultural one.

Many users express sentiments such as:

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

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

1.2 The Imperative of Explainability

In the context of recommending music, where algorithms can shape cultural trends, listening behaviours artist exposure and even financial outcomes for musicians and platforms relying solely on predictive accuracy fundamentally inadequate. A model can statistically successful

to introduce hidden discrimination, where entire genres or artist communities are quietly deprioritized because the model has learned biased patterns from historical data, for instance, genres with high speckiness and unconventional rhythmic structures might be considered less popular, causing a systematic filtering out effect that influences public consumption without the user noticing. As a result, we need transparency about how recommendations are made, and Explainable AI translates decision logic from the machine level to forms that humans can understand and evaluate. By using XAI, we can examine why a model is recommending a particular song and what biases it might be incorporating to generate that recommendation.

2. Dataset and Preprocessing

We use dataset of about 170,000 songs that sourced of Kaggle. Numerical characteristic like energy, tempo, acousticness, instrumentalness, danceability, etc. are included in every song.

2.1 Target Classification

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

- Popular (1)
- Not Popular (0)

2.2 Data Splitting and Scaling

The dataset was split using a 80/20 train test split, feature standardization was applied with mean = 0 and standard deviation = 1 (to give each feature equal scaling influence during modelling) and these preprocessing steps were executed exactly as demonstrated in the IPython notebook (to be methodologically consistent and fair in the comparison of feature contributions through the model decision process).

.

3. Model Training

We trained Random Forest Classifier on normalized feature set and as printed in the IPython notebook, the model had

Accuracy = 1.0 and

F1 Score = 1.0.

which we report exactly printed. These scores of near perfect performance may initially appear to indicate a very well-trained model but these types scores are very rare for real world classification tasks, so we ask the following: Did the Random Forest really find strong structural patterns in the dataset or was classification boundary between Popular and Not Popular songs exceptionally clear or was an imbalance in classes in the dataset combined with a strong correlation between musical features and popularity.

4. Explainability Analysis

4.1 LIME – Local Explanation

LIME gave local interpretation for individual predictions. whereas aggregated explanations only provide general description of model behaviour. LIME decomposed the decision for single test instance:

RF predicted: 0 (the model classified the song as Not Popular) by assigning explanation weights to the features most responsible for this decision (lower energy, higher acousticness, or certain rhythmic traits) highlighting which characteristics caused prediction to pull toward the negative class (Not Popular) which is especially valuable because we can trace this reasoning for each song to potentially diagnose where model misinterpreting or biased, as opposed to aggregated explanations that only tell us what the model predicted.

4.2 SHAP – Global Explanation

Since SHAP is a global analysis, we can look at feature influence over the entire dataset rather than individual predictions. Aggregating SHAP values over all samples to see which musical attributes consistently push predictions toward popular and not popular, our results suggested that high instrumentalness tended to decrease the probability of popularity, meaning that instrumental-heavy tracks may have a harder time achieving mainstream popularity compared vocal-driven songs. Higher energy levels increased the probability of being classified as popular, which is consistent with the observation that energetic, upbeat tracks perform better in general listening environments. Acousticness consistently correlated negatively with popularity and danceability made positive contribution to popularity prediction.

5. Discussion

Initially, the model performs perfectly, so this perfect performance metric might give appearance of model that perfectly correct but XAI uncover that the system is not thinking it finding statistical shortcuts. The mild class imbalance towards "Not Popular" song and the strong feature correlations with a few dominant features can result in high accuracy due to structural predictability rather than a deep understanding of music. Explainability allows us to identify when a feature disproportionately contributes to predictions, when feature dependencies are oversimplified, or when the dataset is limited by genre imbalance or sampling bias for example, if features such as speechiness continue to have lower popularity labels, the model will be penalizing rap or spoken-word genres, which will cause genre bias that is masked without XAI. In the end, explainable insights remind us that "perfect model" rarely perfect.

6. Conclusion and Future Work

Our conclusion further supports fact that even though our model had extremely high almost perfect evaluation metrics. we cannot rely on such scores to evaluate performance without interpretability. XAI techniques provide us with a deeper understanding of why recommendations are made, not just what is predicted, which enables us to uncover biases determine importance of features and justify recommendation. we find that:

XAI provides fine grained accountability, transparency builds trust and fosters adoption by making recommendation reasoning explicit, and interpretability is critical for ethical and responsible AI particularly in domains that involve cultural and emotional user preferences, such as music. As more robust explanations, fairness constraints, and user-feedback driven refinement are incorporated, the legitimacy, inclusiveness, and real-world reliability of music recommendation systems continue to improve.

6.1 Future Research Directions

Incorporating tempo-dependent mood modelling generating personalized natural like language explanations, ensuring genre-sensitive fairness, experimenting with advanced models such as XGBoost, Gradient Boosting, and deep learning, and validating recommendations through behavioural feedback loop capturing implicit signals like skips replays, and session duration can collectively advance explainable music recommendation systems making more intuitive for user.

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.