Music Popularity Prediction and Recommendation Using Machine Learning Techniques

Author: Prachi S. Vithlani

Course: MDA 620: Data-Driven Decision-Making Capstone

Table of Contents

1. Background

The music industry is a dynamic field where trends and listener preferences evolve rapidly. Accurately predicting the popularity of music tracks and providing personalized recommendations is crucial for music platforms to retain users and drive engagement. By leveraging machine learning, we can analyze audio features and user preferences to predict popularity and recommend tracks effectively. This project aims to bridge the gap between raw music data and actionable insights.

In recent years, the use of advanced machine learning algorithms has transformed how the music industry evaluates trends. These algorithms provide predictive insights into music popularity, helping artists and producers focus on features that resonate with listeners. This project explores these capabilities using data-driven techniques.

2. Business Issue

Problem Statement

Music platforms struggle to provide users with accurate popularity predictions and relevant music recommendations. The challenges include:

- Understanding which features influence a track's popularity.
- Building systems that recommend both new and relevant tracks.

Impact

Improved prediction and recommendation systems can:

- Enhance user satisfaction by providing tailored playlists.
- Drive engagement and increase streaming time.
- Empower artists to focus on creating popular tracks based on key features.

For example, understanding the role of energy and danceability in track popularity can help platforms curate playlists that align with users' moods and preferences.

3. Objective of the Project

1. Develop a machine learning model to predict music popularity based on audio features.
2. Create a hybrid recommendation system combining audio features and weighted popularity.
3. Provide actionable insights into the relationship between music features and track popularity.

4. Data Exploration and Data Visualization

Dataset Overview

- **Source**: Spotify Dataset
- **Size**: 227 records with 21 columns
- **Key Features**: Popularity, Energy, Loudness, Danceability, Acousticness, Tempo

The dataset captures essential attributes of Spotify tracks, providing insights into their musical and popularity characteristics. Key features such as Energy, Loudness, Danceability, and Acousticness are particularly influential in understanding trends in music preferences.

Data Exploration

Descriptive Statistics

- The dataset contains tracks with an **average popularity of 71.85**.
- **Energy** and **Loudness** demonstrate substantial variation, reflecting a diverse range of musical styles and intensity levels.
- Key descriptive metrics include:
    - **Energy**: Mean of 0.64 with a balanced distribution.
    - **Loudness**: Mean of -6.52 dB, centering around typical audio levels for music.
    - **Acousticness**: Mean of 0.37, with most tracks showing lower acoustic qualities.
    - **Danceability**: Mean of 0.64, indicating that many tracks are suitable for dancing.

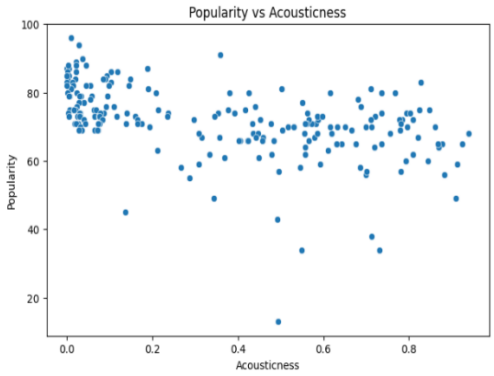
Feature Relationships

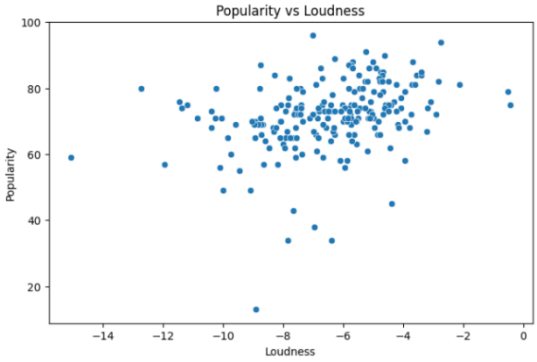
1. **Popularity and Energy**: Tracks with higher energy levels tend to achieve higher popularity, as evidenced by a moderate positive correlation of 0.25.
2. **Popularity and Loudness**: A correlation of 0.31 highlights that louder tracks are often more popular, aligning with listener preferences for impactful sound.
3. **Popularity and Acousticness**: A negative correlation of -0.43 indicates that tracks with lower acoustic content are generally more favoured.
4. **Danceability**: Moderately correlated with popularity (0.25), suggesting that tracks with rhythmic and danceable qualities resonate well with audiences.

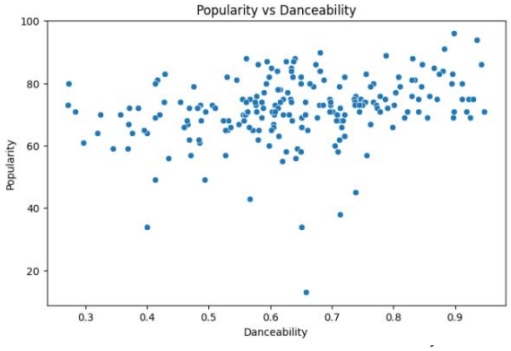
Visualizations

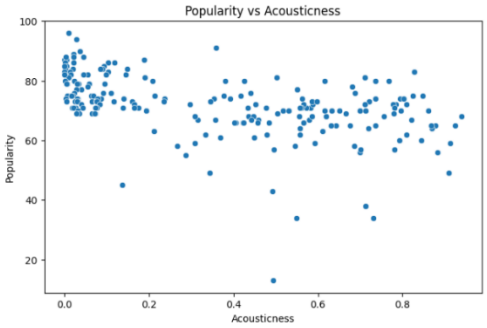
1. **Scatter Plots**

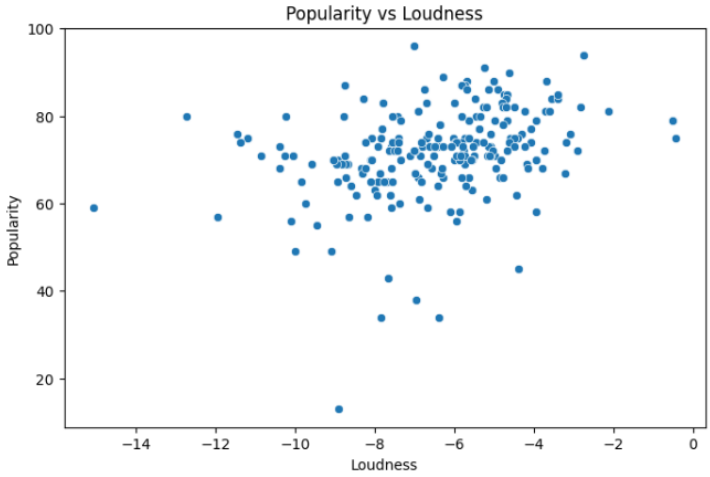
- Popularity vs. Energy: This scatter plot shows a clear upward trend, emphasizing that tracks with higher energy levels are often more popular.





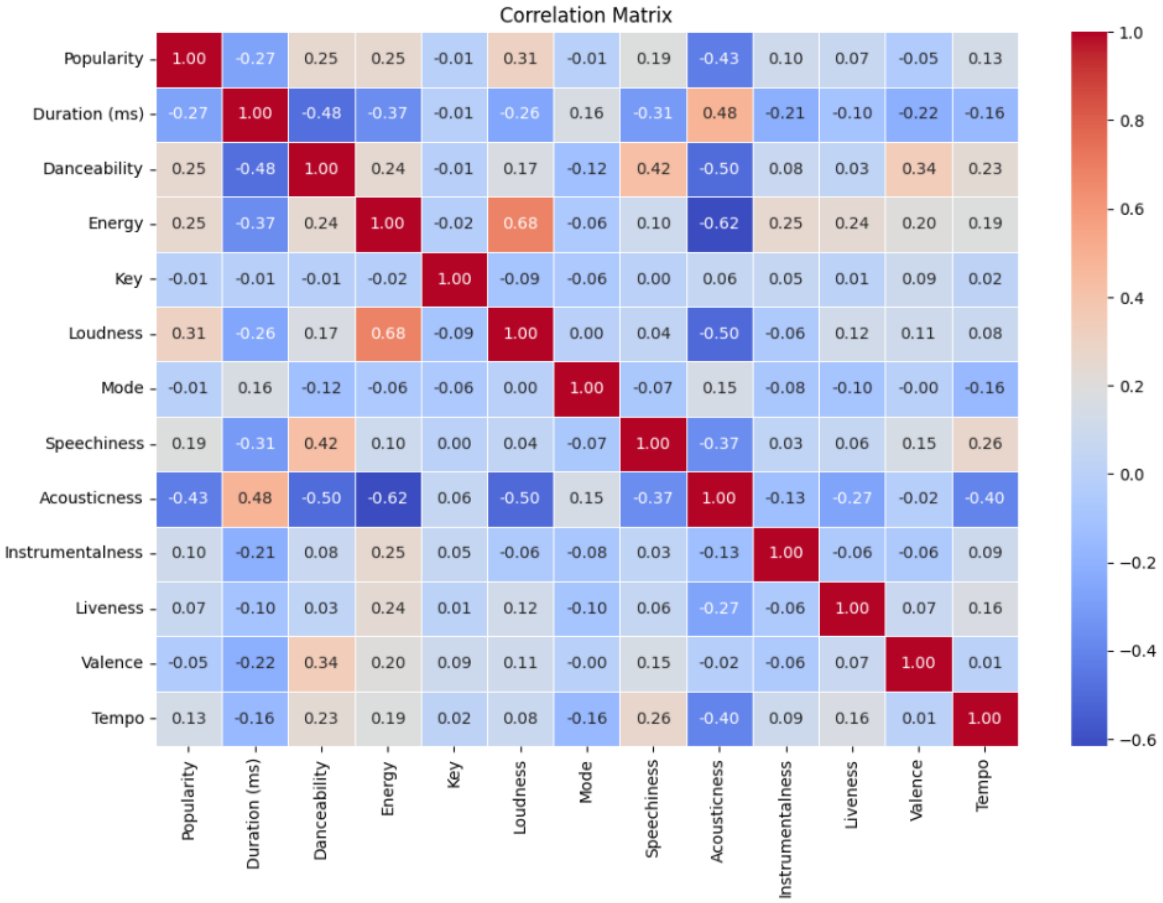






- Higher Energy levels and Danceability positively correlate with higher Popularity scores.
- Increased Acousticness and lower Loudness are associated with lower Popularity, indicating a preference for more energetic and less acoustic tracks.
- Valence (musical positivity) shows a weaker relationship with Popularity, suggesting emotional positivity alone does not strongly predict a track's popularity.

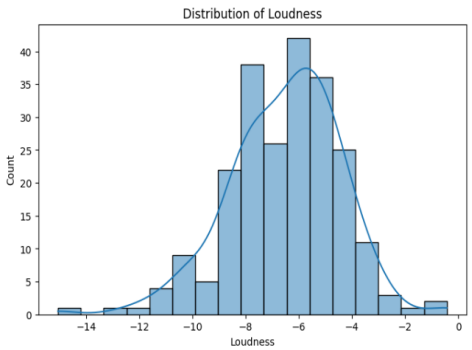
2. Correlation Heatmap

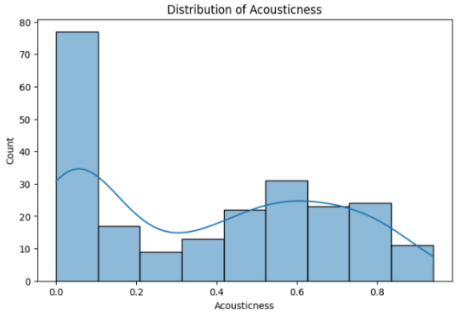


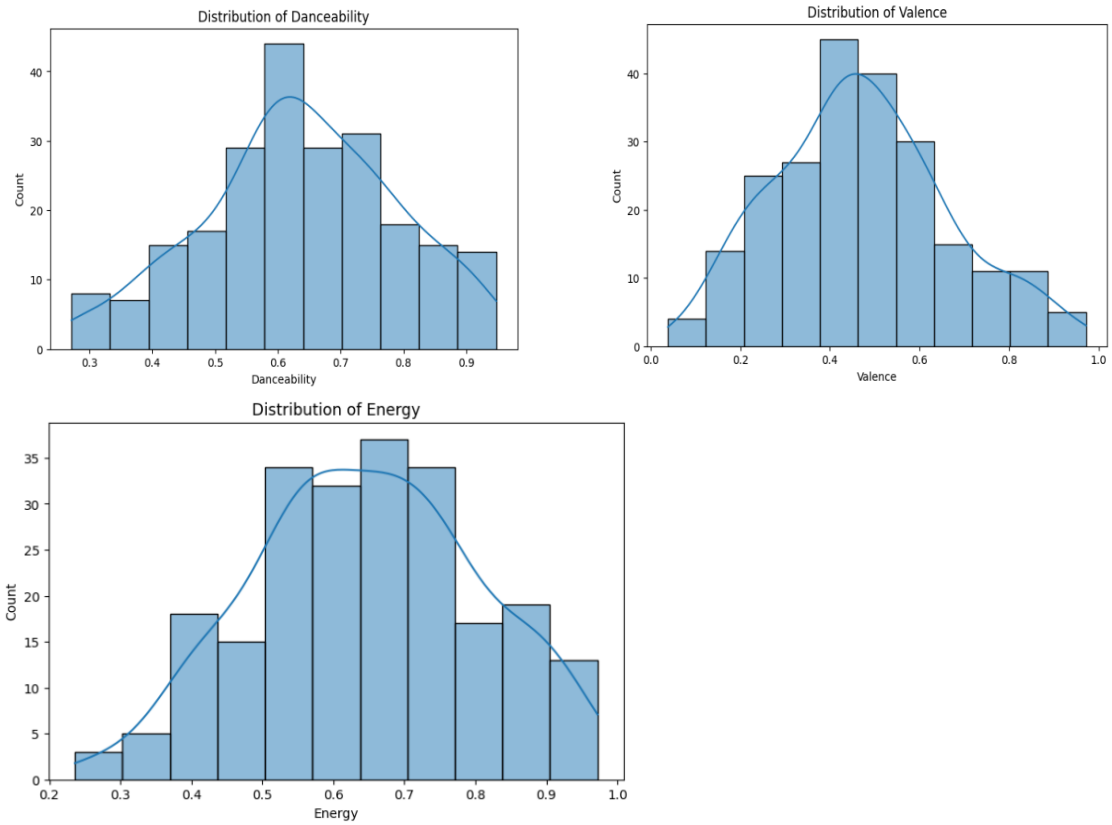
The heatmap highlights key relationships:

- Popularity is moderately correlated with Loudness (0.31) and Danceability (0.25).
- A significant negative correlation between Acousticness (-0.43) and Popularity underlines the preference for less acoustic tracks.

3. Histograms







- Energy Distribution: Displays a bell-shaped curve, suggesting a balanced range of energy levels across tracks.
- Loudness Distribution: Near-normal distribution centered around -6 dB, reflecting typical audio dynamics in the dataset.
- Acousticness Distribution: Skewed towards lower values, indicating that most tracks are not predominantly acoustic.

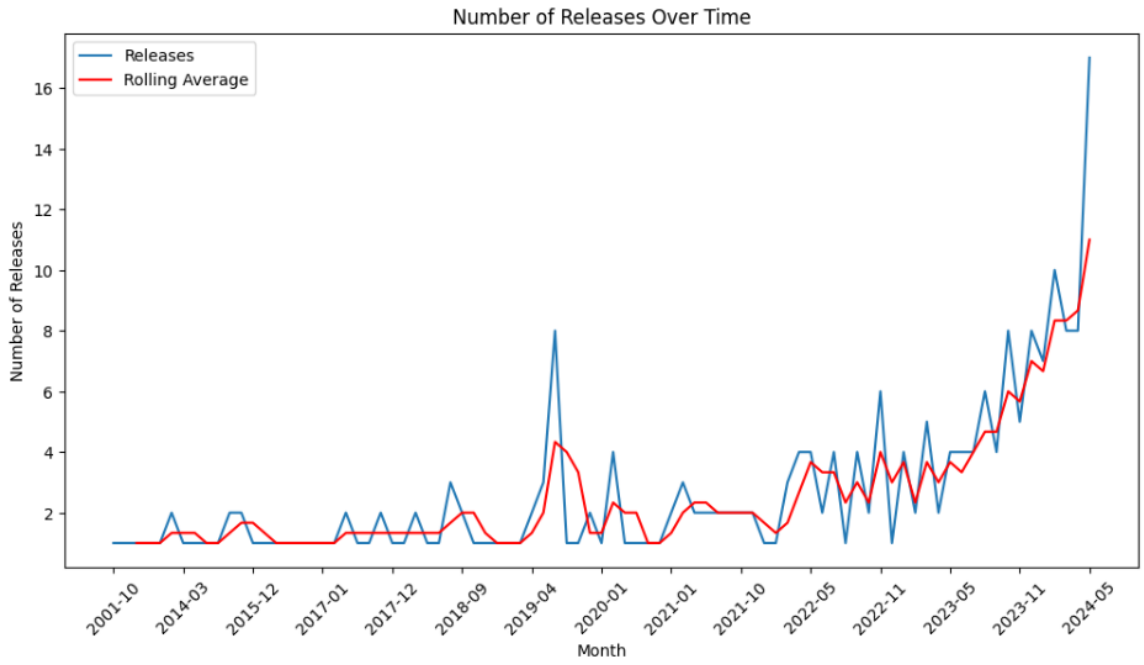
5. Data Manipulation

Preprocessing Steps

1. **Data Cleaning**:
   - Dropped irrelevant columns (e.g., Unnamed column).
   - Verified no missing values in critical features.
2. **Feature Engineering**:
   - Extracted year and month from the release date for time-series analysis.
   - Calculated rolling averages for smoother trends in popularity.
3. **Normalization**:
   - Scaled numeric features (e.g., loudness, energy) using Min-Max Scaling to bring all features to a comparable range.

Visual Analysis of Trends and Seasonality

Time Series Visualization: Number of Releases Over Time



The graph shows the number of music releases over time, with a 3-month rolling average smoothing out fluctuations to reveal broader trends.

Key Insights:

1. **Monthly Releases:**
    - The blue line represents the number of releases each month, which starts low and steady (1-2 per month) with occasional spikes.
2. **Rolling Average:**
    - The red line shows the 3-month rolling average, highlighting a gradual upward trend in the frequency of releases over time.
3. **Long-Term Trend:**
    - Recent months exhibit a sharp increase in releases, indicating a growing trend.
4. **Fluctuations:**
    - Peaks and dips are visible, likely influenced by seasonality, such as holidays or summer events.

This helps illustrate how music releases have evolved, emphasizing both growth and periodic variations.

6. Methodology/Model Building

Models Used

1. **Random Forest Regression**:
    - **Features:** Energy, Loudness, Danceability, Acousticness, Tempo, Speechiness, etc
    - **Strengths:** Handles non-linearity and feature interactions well.
    - **Weakness:** Computationally intensive for large datasets.
2. **XGBoost Regression**:
    - **Features:** Energy, Loudness, Danceability, Acousticness, Tempo, Speechiness, etc
    - **Strengths:** Gradient boosting improves accuracy by learning from errors.
    - **Weakness:** Sensitive to noise and overfitting if not tuned properly.

Training Process

Train-Test Split:

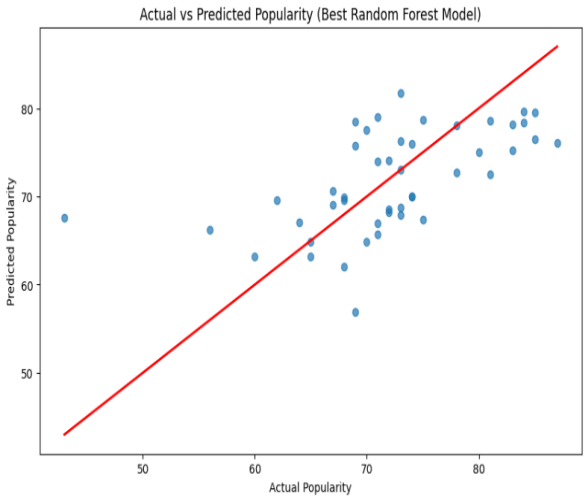
- Training data up to 2022.
- Testing data from 2023 onward to maintain temporal consistency.

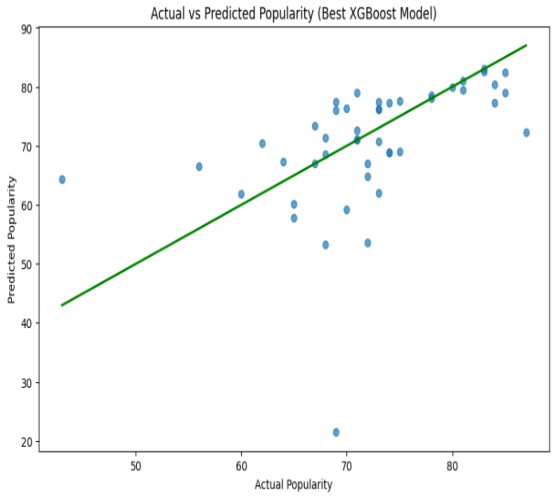
Hyperparameter Tuning:

- Random Forest: Grid search for optimal n_estimators, max_depth, and other parameters.
- XGBoost: Default parameters tested for baseline comparison.

Visualization of Model Performance

- Actual vs. Predicted popularity for Random Forest and XGBoost models.





7. Model Selection

Performance Metrics

- **Random Forest**:
    - RMSE: 6.73
    - R^2: 0.319
- **XGBoost**:
    - RMSE: 10.00
    - R^2: -0.503

Based on lower RMSE and higher R^2, Random Forest was selected for final implementation. Its ability to handle non-linear interactions made it more suitable for the dataset.

Hybrid Recommendations and Music Content Analysis

In addition to content-based recommendations, we also incorporate hybrid recommendation techniques to improve the accuracy and personalization of music suggestions. This section outlines how we combine content-based filtering with weighted popularity scores to generate more effective music recommendations.

Content-Based Recommendations

Content-based filtering relies on analyzing features of songs, such as tempo, energy, danceability, and acousticness, to provide recommendations. Using cosine similarity, we measure the similarity between songs based on these features. The higher the similarity score, the more likely the recommended song matches the characteristics of the input song. The steps involved are:

1. **Feature Extraction**: Extracting relevant features such as Danceability, Energy, Loudness, etc.
2. **Cosine Similarity**: Calculating similarity scores between songs based on their features.
3. **Recommendation Generation**: Selecting top similar songs based on similarity scores.

Hybrid Recommendations

To enhance the content-based approach, we incorporate a weighted popularity component. This involves giving higher weight to recent songs to ensure that the recommendations reflect current trends.

8. Conclusions

The analysis of music popularity prediction and recommendation systems using machine learning techniques has provided significant insights into the relationship between audio features and track popularity. Through the use of Random Forest and XGBoost regression models, along with content-based and hybrid recommendation systems, we have been able to better understand how key features such as energy, loudness, danceability, and acousticness influence music popularity. Below are the key findings:

Key Findings:

1. **Influence of Audio Features**:
    - Tracks with higher energy and louder sound tend to be more popular.
    - Danceability shows a moderate positive correlation with popularity, while acoustic tracks tend to have lower popularity scores.
2. **Model Performance**:
    - The Random Forest model outperformed XGBoost with a lower RMSE of 6.73 and a higher R^2 of 0.319, indicating better prediction accuracy and suitability for capturing non-linear relationships.
    - XGBoost, despite being sensitive to noise, provided insights but struggled with overfitting, resulting in a higher RMSE (10.00) and a negative R^2 (-0.503).
3. **Hybrid Recommendations**:
    - Incorporating weighted popularity scores into content-based recommendations improves the personalization and accuracy of music suggestions, particularly by focusing on recent trends.

9. Recommendations

Based on the findings, the following recommendations can be made:

1. **Enhance Model Performance**:
    - Continue tuning and optimizing the Random Forest model to improve prediction accuracy. Consider hyperparameter optimization and feature selection to reduce overfitting.
2. **Feature Engineering**:
    - Invest in deeper feature engineering, such as sentiment analysis and user feedback integration, to refine models beyond static audio features.

3. **Integration of Hybrid Recommendations**:
    - Develop a more robust hybrid recommendation system by incorporating more contextual factors, such as seasonal trends and artist popularity.
    - Ensure real-time data updates to reflect current musical trends accurately.
4. **User Experience**:
    - Personalize playlists by balancing audio features with user behavior analytics to enhance the user experience. Focus on creating mood-based playlists that capture dynamic trends in music consumption.

10. Future Work

Moving forward, several areas of improvement and exploration could further enhance the analysis:

1. **Advanced Machine Learning Techniques**:
    - Exploring other machine learning techniques such as Neural Networks, Autoencoders, or Variational Autoencoders to handle complex, high-dimensional datasets.
2. **Temporal Analysis**:
    - Deepen the time-series analysis to capture seasonality and short-term fluctuations in music popularity over specific time periods (e.g., holiday seasons, music festivals).
3. **User-Centric Personalization**:
    - Integrate user-specific attributes like listening history, favorite genres, and socio-demographic data to fine-tune recommendations further.
4. **Exploration of Additional Features**:
    - Consider additional features such as lyrics sentiment, mood tracking, and collaborative filtering to develop a more comprehensive recommendation system.
5. **Feedback Loop and Continuous Improvement**:
    - Establish a feedback loop system where users provide insights on recommendations to continually improve the model's accuracy and relevance over time.

By implementing these recommendations and exploring future research avenues, music platforms can offer highly personalized and accurate music recommendations that enhance both user satisfaction and engagement.

11. Bibliography/References

- Spotify Dataset: https://www.kaggle.com/datasets/vatsalmavani/spotify-dataset
- Current challenges and visions in music recommender systems research - Spotify Research. (2020, September 4). Spotify Research. https://research.atspotify.com/publications/current-challenges-and-visions-in-music-recommender-systems-research/
- Music Recommendation System Using Machine Learning. (2022, October 26). GeeksforGeeks. https://www.geeksforgeeks.org/music-recommendation-system-using-machine-learning/
- Billboard. (2024). The Hot 100 Chart. Billboard. https://www.billboard.com/charts/hot-100/
- Topic: Music in the U.S. (2019). Www.statista.com; Statista. https://www.statista.com/topics/1639/music/