Final Project Report
Yahoo Music Recommender System
EE627WS-Data Acquisition/Modeling/Analysis–Big Data Analytics
Stevens Institute of Technology
Dustin Haggett & Brianna Cirillo

Introduction

In the modern digital age, recommendation systems have become essential for delivering personalized content to users, especially in domains like music, video, and e-commerce. This project focuses on building a hierarchical music recommender system using the Yahoo! Music dataset, a large-scale real-world dataset that captures user preferences across multiple levels of music metadata. Each track is part of an album, which is linked to an artist, and artists may span multiple genres. Users in the dataset provide ratings between 0 and 100 for various items across these levels.

The goal of the project is to predict, for each user in the test set, which 3 out of 6 candidate tracks they are most likely to enjoy. This is framed as a binary classification task where three tracks must be labeled as “1” (liked) and three as “0” (not liked). The complexity of this task lies in leveraging the hierarchical structure of the data while coping with the sparsity and diversity of user interactions.

Methodology

The project began with hierarchical feature engineering to extract and organize the available user ratings. For each user-track pair in the test set, We compiled the user’s historical ratings for that specific track, as well as the associated album, artist, and all relevant genres. This multi-level feature set enabled the use of both rule-based and machine learning-inspired approaches, capturing the depth of user preferences in a structured format.

Simple Weighted Scoring

One of the first methods tested was a simple weighted scoring model, where the score for each track was calculated as the sum of the user's ratings for the associated album and artist. Tracks were then sorted by this score and labeled as "liked" or "disliked" accordingly. This baseline performed surprisingly well, achieving strong accuracy due to the high signal strength of album and artist preferences.

Weight Tuning

To further refine the model, We experimented with different combinations of weights for the album, artist, and genre features using grid search and fine-tuning scripts. Results consistently showed that giving high weight to album ratings (typically 0.8–0.95), moderate weight to artist ratings, and minimal or zero weight to genre ratings produced the best results. These findings were implemented in a more general scoring function that allowed flexible weighting.

Weighted Hierarchical Scoring

This approach used the tuned weights to calculate a weighted sum across album, artist, and genre ratings. This generalization of the earlier model allowed for precise control over the influence of each level in the hierarchy and offered clearer interpretability of the feature contributions.

Ensembling

To reduce potential variance and increase robustness, we implemented a majority-voting ensemble of the top-performing models. This is a standard practice in machine learning, often providing marginal gains in performance by aggregating diverse perspectives from different models.

Performance Evaluation

Across all methods, the performance converged around the same top score. Below is a summary of observed accuracy:

Method	Accuracy
Simple Album+Artist Scoring	0.846
Weighted Scoring (Tuned)	0.846
Ensemble of Top Models	0.846

These results highlight that the hierarchical scoring strategy, especially with well-chosen weights, captures the most predictive aspects of user preferences in the dataset. The minimal benefit from genre features likely stems from their weaker signal and higher sparsity. The ensemble did not outperform the best single model, likely due to strong similarity among their outputs.

Overall Results and Comments

The final ensemble result achieved an accuracy of 0.846, equal to the top-performing individual models. This consistency suggests that the current feature engineering and scoring approach has reached a local optimum within the constraints of the data.

This project showcases the effectiveness of hierarchical feature engineering and interpretable, rule-based models in recommendation systems. Although more complex models like collaborative filtering or neural networks could potentially yield improvements, this pipeline is efficient, scalable, and robust. Future directions could include enriching features with user-user similarity, applying dimensionality reduction, or introducing hybrid models that combine collaborative and content-based methods.