Music recommender systems and children

How demographic features impact the accuracy of recommendations

Teun Bosch[1]

Supervisor(s): Sole Pera, Robin Ungruh

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Teun Bosch
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Robin Ungruh, Masoud Mansoury

Abstract

Music-streaming platforms rely on recommender systems to help listeners navigate millions of tracks, including a growing number of children using these platforms. However, most systems are optimized for adults, often resulting in recommendations that fail to reflect preferences or needs of children. While demographic features have been shown to improve performance in models focused on adults, their impact on a child-centric recommender system remains unexplored. This study investigates whether incorporating demographic features (age, gender, and country) and profiling features (exploratoryness, concentration, and replayness) improves the quality of music recommendations for children. Using a filtered subset of the LFM-2b dataset, we evaluate a baseline model based on implicit-feedback interactions against variants extended with different combinations of demographic and profiling features. Results show that demographic features led to a reduced accuracy across most models. In contrast, profiling features significantly increase top-K accuracy, with improvements up to 18%. These findings highlight the limitations of recommender systems tuned for adults when applied to children and emphasize the value of behavioral-aware modeling in the development of more effective child-centric music recommender systems.

1 Introduction

Recommender systems have become an integral part of our everyday life, impacting our media consumption on a wide variety of different platforms. Platforms like Spotify and Apple music use these systems to recommend new music, songs and artists to users, relying on these systems to guide users through the extensive music catalog available. By doing so, they help reduce the information overload the users might face when navigating the platform by themselves [1]. Importantly, children are now highly active users of these platforms, with over 60% of those aged 9–18 using Spotify to access music [2].

Despite their widespread use by younger audiences, current music recommender systems are primarily designed and optimized for general, typically adult, audiences [3]. As a result, children often receive recommendations that do not match their developmental stage, include repetitive or inappropriate content, or fail to reflect their preferences [4]. This design bias overlooks the distinct listening behaviors and developmental needs of children. Prior work by Schedl and Bauer [5] shows that collaborative filtering systems tailored to users aged 6–18 significantly outperform general models, while adult-trained models degrade in performance for these age groups. Additionally, prior work by Spear et al. [6] highlights the significant differences among children, indicating that a model optimized for adults may not generalize for them. Moreover, previous work by Vigliensoni and Fijunaga [7] demonstrates that incorporating demographic features, such as age, country, and gender, can significantly improve the performance of recommender systems. However, this work was carried out on a broad user base, dominated by adults, and did not consider the unique characteristics of children. As a result, it remains unclear whether such demographic features are equally beneficial for children.

This research aimed to answer the following question: *How do demographic features, such as age, gender, and country, and profiling features impact the performance of music recommender systems tailored for children?* To answer this question, we used the LFM-2b dataset [8], which contains over two billion listening events along with user-provided demographic features. In addition to the available demographic features, we computed three additional profiling features that capture different aspects of listening behavior: exploratoryness, concentration, and replayness. Combined with the demographic features, these profiling features aim to provide a more complete representation of the user. We trained a baseline recommendation model using solely implicit feedback and extended it with demographic and profiling features to assess their impact. The performance of the models was assessed using standard top-K metrics including Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDCG).

Our results showed that, contrary to findings in adult-focused systems, demographic features such as age, gender, and country did not improve recommendation performance for children, and in some cases even led to a reduced accuracy. In contrast, profiling features derived from user listening behavior significantly enhanced model performance, with improvements up to 18% in top-K metrics. These findings suggest that behavioral patterns are more important than demographic traits when recommending new music for younger users. This study provides a detailed evaluation of the effectiveness of features in child-focused music recommendation, supporting the development of more accurate recommender systems tailored to children.

This report is structured as follows. Section 2 describes the methodology. It covers the preprocessing steps for the dataset, the formulas of the profiling features and a description of the recommendation algorithm used for the models. Section 3 introduces the experimental setup, including the details of the recommendation algorithm and the evaluation procedure. Section 4 presents the results from the experiment. Section 5 discusses the results and their implications. Section 6 reflects on the ethical considerations relevant to a child-focused recommender system and the reproducibility of the experiment. Finally, Section 7 concludes the report and outlines limitations and future recommendations.

2 Methodology

In order to investigate the impact of demographic and profiling features on recommender systems, we structured our methodology into three main components. First, we

required a dataset consisting of listening events from children annotated with demographic features. This dataset needed to be free of redundant or unusable data, while still providing enough data points for training and evaluating the recommender systems. The second component involved computing profiling features based on the cleaned dataset. These profiling features describe the listening behavior and provide the recommender system a more complete picture of the user. For our study, we selected three profiling features—exploratoryness, concentration, and replayness—each chosen to capture distinct aspects of user behavior. The third component of our methodology would be the recommender system itself. Utilizing the data and features gathered in the first two steps, the recommender systems were compared on different metrics. Each model used a different set of features, resulting in varying model performances. By comparing these results, we aim to identify the set of features that provides the most significant improvement.

This section follows a chronological structure, where we start with the first part implemented and end with the last part. The first subsection describes the dataset, including the filtering criteria and the retained data. The second subsection presents the profiling features and their respective formulas. The final subsection describes the algorithms used for the recommender systems.

2.1 Data

For our recommendation models, we required a dataset that included demographic features, such as country, age, and gender, as well as the listening events of users. These were necessary for computing the profiling features. For this study, the LFM-2b dataset contained the required fields at scale [8]. It included timestamped listening histories along with demographic features such as age and country. Although the demographic features are self-declared by the users, prior research shows that users tend to provide accurate information in their online profiles, as they want to be represented accurately [9]. The dataset contained listening events from over 120,000 users of the online music platform Last.fm.

2.1.1 Data cleaning

The LFM-2b dataset contained listening events from a wide variety of users, including both children and adults. First, we filtered out all the listening events from adults and retained only those recorded when a user was under 18. To further reduce noise, we filtered out all listening events except those from 2012, the year with the highest number of child interactions. We then applied a replay-based filter that retained only the songs that a user listened to at least twice, based on the assumption that repeated plays more reliably indicate user preference [10]. Finally, we found that there were users with missing data fields, such as their gender, country, or age. These users were excluded from the dataset, since we were unable to do the experiment with them.

Even after removing users with missing fields, a significant amount of unusable data remained in the dataset that required further filtering. For example, there were still users who had only listened to a small number of songs or artists. Similarly, some songs had been listened to by not more than one user. To reduce the sparsity in the dataset, we used k-core sampling. We selected a threshold $k = 5$ and we filtered out all users who interacted with fewer than five items. We then selected a threshold $k = 3$ for the items and filtered out all items with fewer than three interactions. We repeated this process until all remaining items and users met the thresholds.

An issue related to user ages arose during the finalization of our dataset. We filtered out listening events based on the listener's age at the time of the interaction. However, users who turned 18 during 2012 had all their post-birthday listening events filtered out, resulting in the removal of a substantial portion of their listening history. To prevent this, we also filtered out all users who turned 18 before the end of 2012. Additionally, instead of using the static age field provided in the dataset, we calculated an average age across a user's listening events and used that value as their age. Users who were 17 for all of their listening events were assigned an age of 17.0, while those who were 12 for all of their listening events received a value of 12.0. All other users fell between these two values, resulting in a continuous distribution of user ages, rather than a categorical value.

2.1.2 Training, test and validation sets

Splitting the dataset into training, validation, and test sets is necessary to train and evaluate the recommendation models. A common approach is an 80/10/10 split; 80% of the data is used for training, 10% for validation, and 10% for testing. However, to increase the realism of our experiments, we applied a different splitting strategy based on the timestamps of listening events. Since the dataset only contained listening events from 2012, we chose to split the dataset based on the full calendar year. To ensure that our recommender system captured a general time frame of listening activities, we excluded the last two months from the dataset. We split the dataset at the end of August, September, and October, corresponding to the training, validation and test sets, respectively.

As seen in Table 1, the final dataset contained over 10,000,000 listening events from 4,407 unique users. Additionally, these events involved more than 160,000 unique tracks, with an average of approximately 62 interactions per track. In the validation and test sets, the average number of interactions per track is lower, at around 9 interactions per track. Furthermore, not all users in the full dataset were present in each individual subset.

2.2 Profiling features

In addition to the demographic features already present in the dataset, we also computed additional profiling features to further improve the performance of the recommender. These features were intended to give the model more insight into the listening behavior of each user. For example, how much do they explore music, and how often do they replay their favorite songs? In combination with the demographic features, these profiling features provided the recommender system a more complete picture of each user. These profiling features

Table 1: Overview of the dataset statistics after filtering and splitting. The table shows the number of listening events, unique tracks, and unique users in the training, validation, and test sets, along with the total across all sets.

	Events	Tracks	Users
Training set	8,272,548	157,751	4,277
Validation set	972,550	112,989	3,076
Test set	921,957	113,204	3,060
Total dataset	10,167,055	160,423	4,407

were computed on the listening history of the user, resulting in numerical values. Since these features were computed on the listening history, they could only be computed with the training set. Using the full dataset would risk leaking information from the validation and test sets into the training process.

2.2.1 Exploratoryness

The first profiling feature we examined was exploratoryness. This feature was found to be the most impactful across a more general user base [7]. It represents how much a user explores different music compared to other users. The value is computed from the user's full listening history using a function based on the user's individual listening frequencies. This function returned a value between 0 and 1, where lower values indicated limited exploration, and values closer to 1 indicated high exploratory behavior. The function definition is as follows:

Let:

- x be a user from the dataset.
- L be the number of submitted track logs by user x.
- S be the set of unique tracks listened to by user x.
- s_i be the number of logs for the i-th most played track in user x's history.

Then the exploratoryness ex_x is computed as:

$$\text{ex}_x = 1 - \frac{1}{L} \sum_{i=1}^{|S|} \frac{s_i}{i}$$

2.2.2 Concentration

Alongside exploratoryness, which showed how much a user explored music, we also considered how balanced a user's music taste is. While some users focused heavily on a few favorite artists, others listened to many different artists equally. This feature captured that balance by computing the normalized entropy of each user's artist listening history. Since the function was normalized, it resulted in values between 0 and 1. Values closer to 1 indicated that a user listened to artists in roughly equal proportions, while values closer to 0 indicated a strong preference for a few artists. The function definition is as follows:

Let:

- x be a user from the dataset.
- L be the number of submitted artist logs by user x.
- A be the set of unique artists listened to by user x.
- a_i be the number of logs for the i-th artist in user x's history.

Then the concentration b_x is computed as:

$$b_x = -\frac{1}{\log |A|} \sum_{i=1}^{|A|} \frac{a_i}{L} \log \left(\frac{a_i}{L} \right)$$

2.2.3 Replayness

In addition to exploratoryness and concentration, we also included a feature that captures how often a user returns to their favorite songs. This feature captured the user's replay rate by measuring how frequently they listened to their most played tracks. Some users frequently replayed their favorite songs, while others preferred a wider variety without repeatedly replaying their favorite tracks. Since this function was normalized, it resulted in values between 0 and 1. Values closer to 1 indicated that a user frequently listened to their favorite tracks, while values closer to 0 indicated that the user listened to a wide variety of tracks, with little to no repetition. The function definition is as follows:

Let:

- x be a user from the dataset.
- L be the total number of submitted track logs by user x.
- T be the set of unique tracks listened to by user x.
- T_k be the set of the user's top k most played tracks, where $k = \min(100, |T|)$.
- p_i be the number of logs for the i-th track in user x's history.

Then the replayness r_x is computed as:

$$r_x = \frac{1}{L} \sum_{i=1}^{|T_k|} p_i$$

2.2.4 Distribution of profiling features

After computing the profiling feature values for all users, we visualized their distributions using box plots, shown in Figure 1. Both exploratoryness and concentration have very high medians (around 0.9) and relatively narrow interquartile ranges. They also have a small tail of low-value outliers below 0.5. In contrast, replayness is widely distributed across the entire range with values between 0 and 1, a median above 0.75 and most values above 0.5. These differences in distributions highlight that most users tend to consistently explore music and have balanced listening habits, even if their tendency to return to their favorite tracks varies more widely. This variability could be valuable for personalized recommendations.

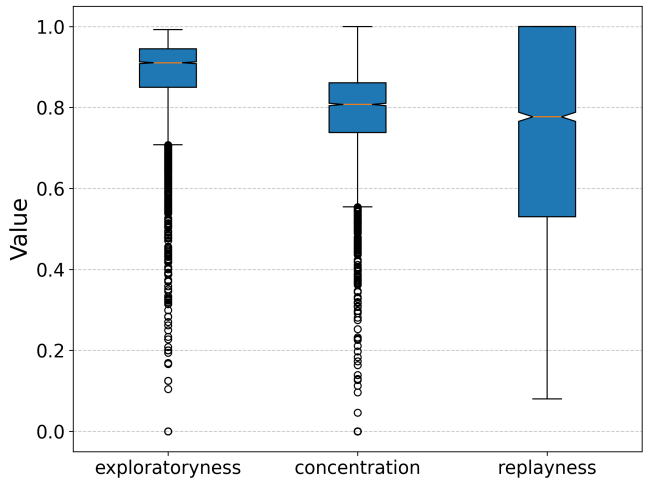

Figure 1: Distribution of the three profiling features across all users. Exploratoryness and concentration are generally high and consistent across users, while replayness shows more variation, reflecting different replay behaviors.

2.2.5 User profile vector

To create a complete user representation for our recommender systems, we combined the demographic and profiling features into a unified user profile vector. Age was normalized between 12.0 and 17.0 to fall within a 0–1 range. Categorical variables like country and gender were one-hot encoded. These, along with the profiling features, were flattened into a one-dimensional vector containing strictly numerical values.

These user profile vectors provided the recommender systems with additional context beyond user-item interactions. Demographic features can help capture regional trends, while profiling features provide insights into the listening behavior. By integrating both, we intend to assess the impact of each feature on model performance and identify the most valuable features for improving recommendation accuracy and quality.

2.3 Recommendation algorithm

As the foundation of our recommendation models, we selected Neural Matrix Factorization (NeuMF) [11]. NeuMF is a collaborative filtering framework that combines General Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) components, modeling both linear and nonlinear interactions between users and items. It was specifically designed for implicit feedback tasks, which involve datasets such as ours that contain only observed user interactions without any ratings. This ability, combined with its capacity to model both linear and non-linear user-item relations, made it a strong fit for our study.

Another advantage of NeuMF was its flexibility in incorporating additional features, such as our demographic and profiling features. The model architecture allowed for easy extension by concatenating user profile vectors to the learned user embeddings. This enabled us to directly evaluate the impact of added features on recommendation performance. By comparing the baseline NeuMF model with extended versions that included additional features, we were able to better understand the contribution of those features to personalized recommendations.

3 Experimental setup

The goal of our experiment was to evaluate the impact of demographic and profiling features on the accuracy of recommender systems for children. We used two types of data: user-item interaction data, split into training, validation, and test sets, and user profile vectors, each containing demographic and profiling features. Each profile vector included three demographic features: a normalized age value, a flattened one-hot encoded country value and a flattened one-hot encoded gender value. It also contained three profiling features: exploratoryness, concentration, and replayness, each normalized to a value between 0 and 1. We used these features to test whether their inclusion improved recommendation accuracy compared to a model solely based on interaction data.

To compare recommender systems, we first established a baseline model, which we later extended with additional features. This model was solely based on the interaction history of the user and implemented using a neural collaborative filtering approach, Neural Matrix Factorization (NeuMF) [11]. It used user-item interactions to learn latent embeddings for users and items. These embeddings learned underlying patterns in user preferences and item characteristics. At the end of each iteration, the embeddings were passed through a Multi-Layer Perceptron (MLP) to estimate the likelihood of user-item interaction. The final output of the MLP was a single sigmoid activation unit that predicted the probability of interaction. Using backpropagation, the model learned from prediction errors and updated the embeddings accordingly.

We extended the same neural collaborative filtering approach to incorporate the additional features from the user profile vector. In these models, the user embeddings were concatenated with the profile vector and passed through a dense layer, resulting in an enhanced user embedding. This design encouraged the model to incorporate information from the profile vector, even if it did not improve the performance of the resulting model. The enhanced user embedding was concatenated with the item embedding and this resulting vector was then passed through an MLP to estimate the likelihood of user-item interaction. The final output of the MLP was the prediction of the probability of a user-item interaction. Backpropagation was used to update the weights and embeddings accordingly.

As our interaction dataset only contained positive interactions, we applied negative sampling to include negative interactions. A positive interaction refers to a user interacting with a song, and a negative interaction corresponds to a song the user did not listen to. Negative sampling was used

to generate a set of unobserved user-item pairs assumed to be negative. For each positive interaction, we sampled four random negative items that the user had not previously listened to. Negative interactions were assigned a label of 0, while positive interactions were labeled 1 in the training data.

For our hyperparameter selection, we tuned embedding size E, learning rate η , and L2 regularization λ to balance convergence speed and overfitting risk. The final choices are summarized below:

- **Embedding size** $E \in \{16, 32, 64\}$;
  $E = 32$ gave the best validation performance.
- **Learning rate** $\eta \in \{1e-3, 5e-4, 1e-4\}$;
  $\eta = 1e-3$ combined with a Reduce-on-Plateau scheduler converged fastest without overfitting; η = 1e-2 was also evaluated, but that was prone to overfitting and was discarded.
- **L2 regularization** $\lambda \in \{1e-5, 1e-4, 5e-4\}$;
  $\lambda = 1e-5$ performed best after adding a dropout layer to the MLP with $p = 0.2$.
- **MLP architecture:** hidden units $[64, 32]$; unchanged during training.
- **Loss function:** we started with Binary Cross-Entropy, which was swapped for Bayesian Personalized Ranking (BPR) [12], as it directly optimized top-K ranking.
- **Optimizer:** Adam [13] was used.
- **Epochs:** the models were trained for 100 epochs; the best performing checkpoint was retained based on validation performance.

These settings were used consistently across all models during the training process.

To evaluate the performance of the recommendation models, we used Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDCG); both are commonly used in top-K recommendation tasks. Specifically, we used HR@10 and NDCG@10 in our evaluation, where 10 refers to the number of top-ranked recommendations. HR measures whether a relevant item appeared in the top-K recommendations for a given user. NDCG measures not only the presence of relevant items in the recommendations but also the position in the ranked list. HR and NDCG are often reported together, as HR reflects recommendation accuracy, while NDCG captures ranking quality.

4 Results

In total, there were 64 possible combinations of the six input features. Due to computational constraints and expected redundancy among certain combinations, we selected 28 representative models. These included all single-feature models, all possible combinations within the profiling features, and a range of cross-group combinations involving both demographic and profiling features. We focused on combinations that were most likely to provide useful insights, such as all features in isolation and combinations pairing exactly one demographic feature with one profiling feature. Redundant or low-value combinations were excluded to maintain feasibility without compromising any insights gained. All selected models can be found in Appendix A. For each model, we computed HR@10 and NDCG@10 to evaluate and compare performance.

Figure 2 summarizes the performance results of all tested recommendation models. The top graph ranks models by their mean HR@10, and the bottom graph ranks them by NDCG@10. Each bar includes a 95% confidence interval (CI), indicating the range in which a random user's metric is likely to fall. Feature combinations are abbreviated on the x-axis using the initial letter of each feature, except for concentration, which is represented as 'co'. The baseline model is labeled with a b and highlighted in red. It achieved an HR@10 value of 0.774 and an NDCG@10 value of 0.265. Models to the right of the baseline model outperformed it, suggesting an increase in recommendation accuracy when adding those features. Conversely, models shown to the left of the baseline underperformed. In particular, these underperforming models relied solely on demographic features. Models incorporating only country, age, and gender saw drops of 4.6%, 8.3%, and 9.4%, respectively, compared to the baseline. Combining all three demographic features improved results slightly but still trailed the baseline model by 2%.

In contrast, several models that incorporated profiling features consistently outperformed the baseline across both HR@10 and NDCG@10. Among them, the model using only replayness ranked among the top performers, achieving a 4.9% improvement in HR@10 and 18.1% improvement in NDCG@10 compared to the baseline model. Exploratoryness and concentration also improved performance, with increases of 11.8% and 7.7% in NDCG@10, respectively. Interestingly, concentration underperformed compared to the other two profiling features. When combined with replayness, overall performance was similar to that of the model using exploratoryness alone, despite replayness on its own performing better. The combinations of concentration with exploratoryness and exploratoryness with replayness performed nearly identically to their strongest individual feature counterparts, with less than 1.5% difference in NDCG@10. The model combining all three profiling features performed worse than the model incorporating replayness alone.

We now turn to models that combine demographic and profiling features to examine whether demographic information provided additional benefit when paired with behavioral features. As shown in Figure 2, age and country in combination with concentration yielded performances comparable to concentration alone, with differences of less than 2% for HR@10 and NDCG@10. However, gender gave a small improvement of 3.6% in HR@10 and 13.2% in NDCG@10 over the baseline. The combination of exploratoryness and age underperformed relative to other models, achieving results closer to the baseline with improvements of 0.5% and 4.7% for HR@10 and NDCG@10, respectively. Combining

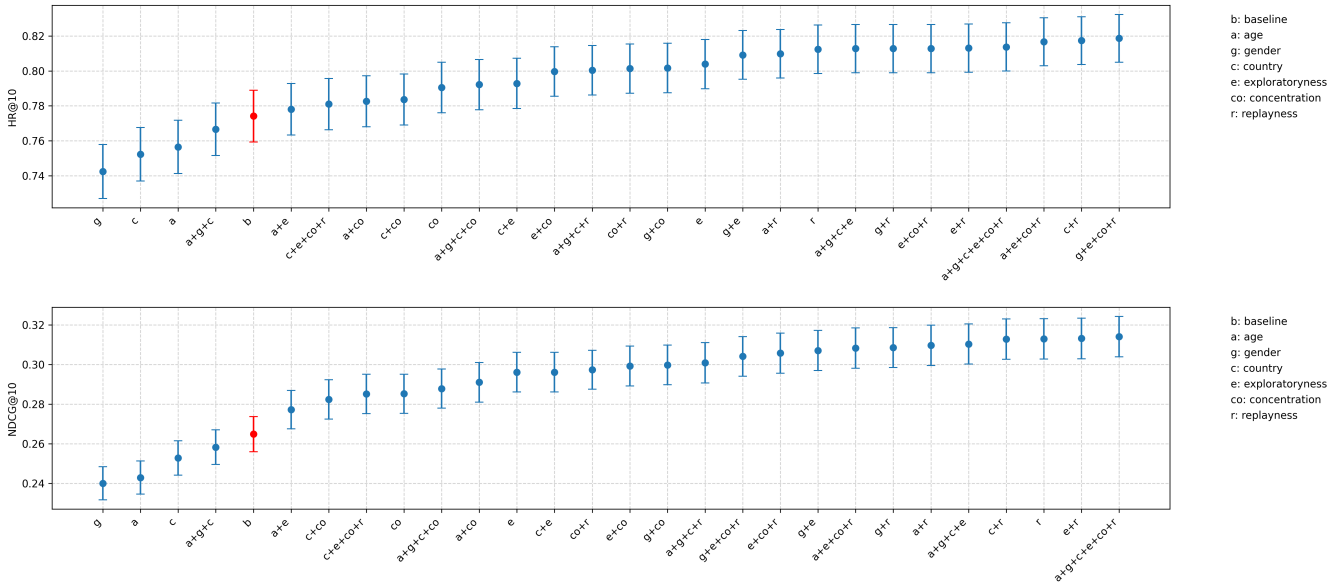

Figure 2: HR@10 (top) and NDCG@10 (bottom) with 95% confidence intervals (CI) on the test set across all experiments. The baseline model (shown in red) is consistently outperformed by models incorporating profiling features. Replayness, in particular, significantly improves the recommendation quality. In contrast, demographic features perform worse than expected for children.

exploratoryness with gender was the only combination to outperform the model that included exploratoryness alone, achieving a 4.5% and 15.9% improvement for HR@10 and NDCG@10, respectively, in comparison to the baseline. As replayness already performed strongly alone, only the addition of country yielded a further improvement. Although replayness alone and the combination showed similar NDCG@10 values, the combination achieved 0.817 in HR@10, compared to 0.812 for replayness alone.

Finally, we examine the remaining models that combine all profiling features with a single demographic feature, all demographic features with a single profiling feature, and the model that includes all six input features. Overall, none of these broader combinations provided meaningful improvements over the best-performing models that relied solely on profiling features. Interestingly, the model combining age, country, and gender with exploratoryness outperformed all models that included exploratoryness as their only profiling feature. The model using all six input features outperformed all other models on NDCG@10, achieving a value of 0.314, an increase of 18.6% compared to the baseline model. In summary, profiling features drove the largest gains, while demographic features offered inconsistent or limited benefits when added to other models.

5 Discussion

The goal of this research was to examine whether demographic features, such as age and country, and profiling features impact the performance of music recommender systems tailored for children. We specifically evaluated whether these features contributed positively to the recommendation accuracy, as measured by HR@10 and NDCG@10. Our findings show that profiling features, particularly replayness, consistently improve performance, while demographic features lead to a reduction in performance. Although combinations of demographic and profiling features yielded mixed results, several top-performing models incorporated both feature types. In the following discussion, we interpret these findings, consider possible explanations, and reflect on their implications for future child-centric recommender systems.

5.1 Demographic features

The demographic features, age, gender, and country, did not lead to an improvement in recommendation performance and, in most cases, reduced accuracy. This finding contrasts with prior research on a general audience, where such features have been shown to improve recommender systems [7]. One possible explanation for this difference is that children's listening preferences are less tied to demographic features like age or country, and more influenced by global trends. Prior research shows that children adjusted their song likability rating based on popularity cues [14], suggesting that teenagers are particularly open to such trends.

Furthermore, age itself turns out to be a surprisingly weak predictor of music preferences for children. Previous research shows that children's music tastes evolve continuously through adolescence and reflect a complex combination of personality and social factors, rather than following uniform age-based stages [15]. Importantly, the authors note that there is no singular stage at which music preference is stable. Two children of the same age can be at completely different stages in how they engage with music. As a result, demographic features such as age, country, and gender fail to capture what influences musical preferences

for children.

5.2 Profiling features

In contrast to demographic features, profiling features based on user listening behavior led to significant improvements in both performance and accuracy across all models. By capturing direct signals from user behavior, these features provide a more personalized representation of musical preference. For children, whose tastes are continuously evolving and are influenced by personality and social factors, these features more accurately reflect their current musical interests. A possible explanation for this lies in the variability of musical development among children. While demographic features cannot reliably distinguish these stages, profiling features can capture a child's current stage in that process. As a result, profiling features are better suited to account for this diversity, as clearly reflected in their performance.

In particular, replayness emerged as the most influential feature in improving recommendation performance and accuracy. This feature captures how frequently users revisit the same tracks. Prior research indicates that children aged 12–17 frequently replay their favorite songs, often engaging in repetitive listening for mood regulation and identity formation [16, 17]. Although there are limited amount of direct comparisons with adult listening habits, the role of repetitive listening as a tool for children in emotional coping appears to be especially pronounced. This may explain why replayness proves to be such a strong feature in predicting musical preference. For many children, music is not only entertainment but also a tool for emotional expression. A recommender system that incorporates replayness can therefore provide recommendations that are more suitable for children.

5.3 Combining demographic and profiling features

Since the profiling features already performed well, combining them with demographic features resulted in inconsistent and marginal gains. Across all combinations, the clearest pattern was that replayness alone achieved nearly all of the achievable performance gains. The only combination outperforming replayness alone paired it with the demographic feature country. An explanation for this outcome remains unclear, as the country feature does not exhibit any obvious correlations with replayness. Interestingly, the combination of all demographic features with exploratoryness performed relatively well, despite none of these features performing strongly on their own. Prior work shows that this feature combination outperformed all others in adult-focused models [7], highlighting a complementary relationship between those features. The model that utilized all six features performed the best, with a slight margin over models incorporating replayness. However, this advantage may result from the greater volume of features available to that model, rather than a meaningful interaction between them.

Taken together, these results confirm that behavior-based profiling features achieve the most improvements for child-centric recommender systems, whereas demographic features provide, at best, a marginal improvement and, at worst, introduce noise. Future systems designed for children should therefore treat demographics as optional and prioritize behavioral profiling features, in particular replayness.

6 Ethical aspects

As this research falls within the domain of computer science and machine learning, several important ethical considerations must be addressed. First, we discuss the ethical implications of the dataset, including concerns around bias and the implications of our findings. We then reflect on the use of data involving children and the associated ethical implications. Finally, we address the reproducibility of our research, and in particular our experiment.

An important concern lies in the biases inherent in our dataset, as well as in the music recommender systems broadly. Musical preferences are known to correlate with demographic features such as age, country, and gender. For instance, Hispanic listeners might lean toward Latin genres, while younger users often gravitate toward newer, trendier artists. Because our models were trained using these features, they learned these demographic patterns and base their recommendations on them, which may amplify the biases already present in the dataset. Although this bias amplification raises ethical concerns, our focus was on analyzing the impact of individual features on recommendation performance. However, a system using these features could reinforce existing filter bubbles, limit cross cultural discoveries, and amplify differences between musical preferences of different genders. To mitigate these risks, it is essential to carefully assess the system's impact for real-world applications to protect users.

The dataset used in this research consists of user listening data, in particular that of children, which raises specific ethical considerations. Although the LFM-2b dataset was used for the experiment, it has been taken offline due to licensing issues. While this mitigates some concerns around data privacy, the use of children's data still asks for careful ethical reflection. Since children are in a developmental stage in their life, they are particularly vulnerable to algorithmic influences. Furthermore, children are more vulnerable to manipulation, less equipped to critically assess algorithmic output, and unable to give consent by themselves.

Lastly, reproducibility is a requirement of this research, as we intend to make it possible for others to replicate our results using the same methodology. Since NeuMF is widely used in recommendation research, basic implementations are available in public repositories. Our implementation used in this study is available at the following repository: link to github. This ensures that other researchers can reproduce the results presented in this study. However, a major limitation to reproducibility is that the dataset has been taken offline due to licensing issues. While the preprocessing of the data is described, the inability to access the data does pose a challenge to replicate the findings exactly.

7 Conclusion

The goal of this research was to determine if demographic features, such as age, gender, and country, and profiling features impact the performance of music recommender systems tailored for children. Our experiments demonstrated that the choice of feature set significantly impacts the quality of recommendations. We found that demographic features did not improve recommender systems and, in most cases, reduced the accuracy of the system. We also found that profiling features improve the accuracy of recommendations, with replayness achieving the strongest individual impact. The set of features that performed the best was the set containing all six; however, the combination of replayness with country yielded nearly identical results. Interestingly, the best performing model excluding replayness combined exploratoryness with the three demographic features, indicating a complementary relationship between these features. Taken together, these findings emphasize that behavioral profiling features more effectively capture children's musical preferences than demographic features.

This research contributes to the limited research done on music recommender systems for children by systematically evaluating the impact of demographic and profiling features on recommendation accuracy. While prior research has shown the value of demographic features in adult-focused recommender systems, our findings challenge their usefulness in child-centric systems and instead highlights the importance of behavioral patterns. This research provides new insights into how personalization strategies should be adapted for younger audiences, prioritizing the development of behavior-driven modeling.

This research is subject to several limitations that may have affected model performance. Because this study focuses specifically on children, the amount of available data is smaller than that of adults. Furthermore, the number of listening events for younger children (aged 12–14) is only a fraction of that for older children (aged 15–17), leading to potential age related data issues. Another limitation is the computational infeasibility of training models for all possible combinations of input features. As the research supported the use of behavioral profiling features in child-centric recommender systems, many other behavioral patterns remain unexplored and could be mapped to additional profiling features. Each added feature doubles the number of possible models, making it infeasible to train all models, or even a well representative subset. Therefore, it is possible that the most influential behavioral patterns for children are still excluded from the evaluated set of features.

Using the findings and limitations of this study, several aspects for future research can be identified Given the data limitations in this study, the first recommendation is to use a dataset with a more balanced distribution across the full age range (12–17). This will mitigate any potential age-related data imbalances. Furthermore, we recommend evaluating additional profiling features that capture alternative behavioral patterns. This complements with the recommendation to explore additional demographic dimensions, such as education level, socioeconomic status, and religion. Incorporating these features could expand the feature space and lead to further improvements in recommendation accuracy. Finally, future research should explore alternative modeling approaches. While this study employed NeuMF, other recommendation architectures could be applied to further validate the findings. Exploring alternative models would enhance the robustness of the study's contributions.

References

[1] Dirk Bollen et al. "Understanding choice overload in recommender systems". In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys '10. Barcelona, Spain: Association for Computing Machinery, 2010, pp. 63–70. ISBN: 9781605589060. DOI: 10.1145/1864708.1864724. URL: https://doi.org/10.1145/1864708.1864724.

[2] Ofcom. *Children and parents: Media use and attitudes report 2023*. Accessed: 2025-06-22. 2023. URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/children/childrens-media-use-and-attitudes-2023/childrens-media-use-and-attitudes-report-2023.pdf?.

[3] Yashar Deldjoo et al. "Enhancing Children's Experience with Recommendation Systems". In: Aug. 2017.

[4] Esteban Gómez, Vasiliki Charisi, and Sophie Chaudron. "Evaluating Recommender Systems with and for Children: Towards a Multi-perspective Framework". In: *Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop*. Vol. 2955. CEUR Workshop Proceedings. Accessed: 2025-06-22. Amsterdam, The Netherlands, 2021. URL: https://ceur-ws.org/Vol-2955/paper2.pdf.

[5] Markus Schedl and Christine Bauer. "Online Music Listening Culture of Kids and Adolescents: Listening Analysis and Music Recommendation Tailored to the Young". In: *CoRR* abs/1912.11564 (2019). arXiv: 1912.11564. URL: http://arxiv.org/abs/1912.11564.

[6] Lawrence Spear et al. "Baby Shark to Barracuda: Analyzing Children's Music Listening Behavior". In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys '21. Amsterdam, Netherlands: Association for Computing Machinery, 2021, pp. 639–644. ISBN: 9781450384582. DOI: 10.1145/3460231.3478856. URL: https://doi.org/10.1145/3460231.3478856.

[7] Gabriel Vigliensoni and Ichiro Fujinaga. "Automatic Music Recommendation Systems: Do Demographic, Profiling, and Contextual Features Improve Their Performance?" In: *International Society for Music Information Retrieval Conference*. 2016. URL: https://api.semanticscholar.org/CorpusID:17941472.

[8] Markus Schedl et al. “LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis”. In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. CHIIR ’22. Regensburg, Germany: Association for Computing Machinery, 2022, pp. 337–341. ISBN: 9781450391863. DOI: 10.1145/3498366.3505791. URL: https://doi.org/10.1145/3498366.3505791.

[9] Scott Counts and Kristin Stecher. “Self-Presentation of Personality During Online Profile Creation.” In: *Self-Presentation of Personality During Online Profile Creation.* Jan. 2009.

[10] Pierre Hanna. “Considering Durations and Replays to Improve Music Recommender Systems”. In: *CoRR* abs/1711.05237 (2017). arXiv: 1711.05237. URL: http://arxiv.org/abs/1711.05237.

[11] Xiangnan He et al. “Neural Collaborative Filtering”. In: *CoRR* abs/1708.05031 (2017). arXiv: 1708.05031. URL: http://arxiv.org/abs/1708.05031.

[12] Steffen Rendle et al. “BPR: Bayesian Personalized Ranking from Implicit Feedback”. In: *CoRR* abs/1205.2618 (2012). arXiv: 1205.2618. URL: http://arxiv.org/abs/1205.2618.

[13] Diederik Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations* (Dec. 2014).

[14] Gregory S. Berns et al. “Neural mechanisms of the influence of popularity on adolescent ratings of music”. In: *NeuroImage* 49.3 (2010), pp. 2687–2696. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2009.10.070. URL: https://www.sciencedirect.com/science/article/pii/S1053811909011410.

[15] David Hargreaves, A. North, and Mark Tarrant. “Musical Preference and Taste in Childhood and Adolescence”. In: *The Child as Musician: A Handbook of Musical Development* (Jan. 2012). DOI: 10.1093/acprof:oso/9780198530329.003.0007.

[16] Suvi Saarikallio and Jaakko Erkkilä. “The role of music in adolescents’ mood regulation”. In: *Psychology of Music* 35.1 (2007), pp. 88–109. DOI: 10.1177/0305735607068889. eprint: https://doi.org/10.1177/0305735607068889. URL: https://doi.org/10.1177/0305735607068889.

[17] Adrian C. North, David J. Hargreaves, and Susan A. O’Neill. “The importance of music to adolescents”. In: *British Journal of Educational Psychology* 70.2 (2000), pp. 255–272. DOI: https://doi.org/10.1348/000709900158083. eprint: https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1348/000709900158083. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000709900158083.

Appendices

A Overview of Trained Models

Table 2: Overview of the 28 trained models and their associated input features.

ID	Name	Included Features
0	Base	None
1	Exploratoryness	Exploratoryness
2	Concentration	Concentration
3	Replayness	Replayness
4	AllBehavioral	Exploratoryness, Concentration, Replayness
5	Age	Age
6	AgePlusAll	Age, Exploratoryness, Concentration, Replayness
7	ExploratorynessConcentration	Exploratoryness, Concentration
8	ConcentrationReplayness	Concentration, Replayness
9	ExploratorynessReplayness	Exploratoryness, Replayness
10	AgeExploratoryness	Age, Exploratoryness
11	AgeConcentration	Age, Concentration
12	AgeReplayness	Age, Replayness
13	CountryFeatures	Country
14	CountryPlusAll	Country, Exploratoryness, Concentration, Replayness
15	CountryExploratoryness	Country, Exploratoryness
16	CountryConcentration	Country, Concentration
17	CountryReplayness	Country, Replayness
18	GenderFeatures	Gender
19	GenderPlusAll	Gender, Exploratoryness, Concentration, Replayness
20	GenderExploratoryness	Gender, Exploratoryness
21	GenderConcentration	Gender, Concentration
22	GenderReplayness	Gender, Replayness
23	AllFeatures	Age, Country, Gender, Exploratoryness, Concentration, Replayness
24	AllDemographics	Age, Country, Gender
25	ExploratorynessDemographics	Age, Country, Gender, Exploratoryness
26	ConcentrationDemographics	Age, Country, Gender, Concentration
27	ReplaynessDemographics	Age, Country, Gender, Replayness

B Generative AI tools used

B.1 Spellchecker

During the writing of the paper, the built-in spellchecker from Overleaf was used to correct any misspelled words. This ensured a high level of spelling accuracy, which would then be improved upon by the next source used.

B.2 Usage of LLM

LLM have been used during the writing process of the report as a tool to improve sections, based on formality, tone and grammar. For this, ChatGPT's 4o model was used with the following generalizable prompt structure:

> "Could you give me a review of the following section of a academic report, listing suggestions on formality, tone and grammar."

These prompts were accompanied by a section or paragraph of the report. The model provided feedback on informal language, tone inconsistencies, and grammar mistakes. The AI model also gave suggestions based on these mistakes, which were critically assessed and valuable suggestions were incorporated, while others were discarded.

In addition to these prompts, clarification about different sections of a research report were also prompted. Especially at the start of the project, some sections were still unclear. After working on the project, the latter sections were relatively easy to grasp and did not need any further clarifications. Prompts given to ChatGPT 4o were the following:

> "What would be included in the research purpose of a research paper?"
>
> "What should there be in the abstract of a research paper?"

The ideas presented by the model were carefully assessed, before working on the specific parts of them. An example would be for the research purpose. As answer, the model gave four different aspects that should be covered in the research purpose: What you are trying to discover/understand/prove, why this topic matters, what gap in the existing knowledge it addresses, and what the expected contributions are. These are all aspects of a good research purpose, which was a great guideline to follow.