Smart Tunes for Kids

Comparing Deep Learning with Traditional Models in Music Recommendations for Children

Supervisor(s): Sole Pera, Robin Ungruh
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 23, 2024

Name of the student: Lennart Verstegen
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Robin Ungruh, Julian Urbano Merino

Abstract

The exponential growth of online content and consumer options has increased the reliance on recommender systems. Children, as a distinct user group, require tailored recommender systems different from those for adults. However, research on recommendation models for children is limited. This study evaluates deep learning recommendation models according to several performance and beyond-performance metrics on data from underage users on the Last.fm streaming platform, offering insights into optimal recommendation strategies for this demographic. Traditional non-deep learning models are used as baselines.

1 Introduction

Upon accessing a variety of social media and streaming services, users are presented with content that has been selected by an algorithmic process. This is the work of recommender systems (RS), which are now commonplace in the digital age [22]. The primary objective of these systems is to suggest content that is likely to be of interest to the user and to be interacted with. As today's children are growing up in the digital age, they also utilise online platforms to consume a diverse range of content on a daily basis [28, 24, 10]. The RS of the platforms they interact with are the primary means by which children discover their content. These RS are designed with only adults in mind however, and aim to enhance *their* user experience and engagement. This focus on adults arises because RS's effectiveness is often measured by overall user satisfaction across the entire user base, and adults constitute the largest and most active demographic on these platforms. It is important to recognise that children have specific needs and capabilities however, that most RS don't cater to. The type of content consumed by children, particularly during their formative years, has the potential to impact their development and influences their knowledge, future beliefs, and behaviour. It is therefore necessary to consider the type of content that children require, such as more factual educational content and content that is suitable to their reading and comprehension skills. As such, a different approach may be required when recommending content to this demographic [28, 8, 10, 18].

This highlights the necessity for children using computer systems to have specific needs and capabilities. These include the need for more factual educational content and content suitable to their reading and comprehension skills. The RS of the platforms they interact with are the primary means by which children discover new content. These RS are primarily designed with adults in mind, and only aim to enhance user experience and engagement without catering to children's specific needs. This focus on adults arises because RS's effectiveness is often measured by overall user satisfaction across the entire user base, and adults constitute the largest and most active demographic on these platforms. Consequently, RS are predominantly evaluated and optimized to meet the needs and preferences of adult users, rather than children. As such, a different approach may be required when recommending content to this demographic [28, 8, 10, 18].

Nevertheless, research on recommender systems in the context of children is still in its early stages, with only a limited number of studies having been conducted thus far [28, 10, 24]. This can be attributed to the numerous challenges inherent to the design and evaluation of recommender systems for children. These include the lack of available datasets, which is often due to practical and legal constraints, as well as the limited attention span or illiteracy of the target audience when attempting to conduct a survey [7]. Additionally, the multitude of stakeholders such as caretakers and teachers involved

in determining the usefulness of a recommender system presents a further challenge. The developers of recommender systems are also unable to rely on traditional mechanisms such as explicit feedback or written reviews, as underage users typically do not provide such feedback [18]. The limited research in this area means that there is still no clear understanding of how to create a 'good' recommender system that satisfies the needs of children and proposes content that will maintain their interest.

To the best of our knowledge, no research has been conducted on the application of deep learning in recommender systems for children. This approach appears to be a promising method for developing an optimal recommender system tailored to children. Deep learning-based recommender systems, due to their complex architecture, are capable of capturing intricate relationships within the data itself. They can incorporate multiple factors from external data sources, such as contextual, textual, and visual information, to make more informed recommendations about content that aligns with the needs of children while maintaining a high level of recommendation quality [30].

This paper presents a measurement of the performance of several state-of-the-art deep learning models according to a small but specific set of metrics in the context of music recommendation, that contributes to the construction of a robust music recommender system suitable for children. The metrics encompass performance, to determine whether content is being recommended that is of interest to children; diversity, to ensure that diverse viewpoints are presented and that echo chambers are avoided, thus exposing children to a broad range of viewpoints and topics, which is crucial for developing critical thinking skills and well-rounded knowledge; and novelty, to introduce children to new and unfamiliar content, catering to their innate characteristics such as curiosity and exploration, which can stimulate their learning and interest in new subjects [28, 10]. The selected models are run on data from underage users from the Last.fm[1] music streaming platform. We employ a number of traditional recommendation models as baselines in order to gain insight into the areas in which modern deep learning models demonstrate increased effectiveness.

After this evaluation, does this study provide several insights into the use of deep learning in the context of recommending music to children, provding recommendations that help towards building a optimal music RS tailored to the specific needs and capabilites of children.

Our research question is: "Do music recommendation systems using deep learning models outperform traditional models in terms of performance, diversity, and novelty when trained and evaluated on child data?"

2 Related Work

The field of recommender systems has witnessed considerable advancement over the years. A subsequent amount of new or enhanced deep learning models have been proposed over the years [32]. [6, 31] compare a number of neural models against traditional baselines utilising a multitude of well-established datasets and a wide range of metrics. [30, 5, 14, 22] conduct comprehensive surveys on the recent advancements and performances of modern deep learning approaches.

Recommender systems have also been studied in the context of music platforms. New models have been developed with a specific focus on music recommendation, with a number of these incorporating deep learning [9, 4]. [17] presents a comprehensive overview of the models employed in music recommender systems, while [23] offers an in-depth analysis of the state of the art in deep learning-based music recommendation. [26] identifies the current challenges and future

[1] https://www.last.fm/

directions for music recommender research.

While research on recommender systems tailored to children is very limited, there have been some notable studies conducted in this area. [8] investigates the impact of different demographics on the effectiveness of recommender systems, including the age group 1-17. [10] conducts a literature review on RS in the context of children and identifies key opportunities, challenges, and risks in children-centered RS evaluation, while [28] considers the relationship between recommender systems and children, with a particular focus on the potential adverse effects of the content children may encounter online.

3 Methodology

In this Section, we outline the methodology employed in our research. We begin by describing the data source, followed by an overview of the framework used to facilitate our experiments. We then present the algorithms selected for evaluation, including both traditional and deep learning models, and conclude with an explanation of the metrics and experimental procedures used to assess their performance. Elliot enables the user to define the entirety of the evaluation in a single configuration file. This file specifies the data prefiltering, recommendation models, evaluation metrics, and other experimental parameters. To ensure the reproducibility of the research, we are sharing the configuration file used for this study[2].

3.1 Data

The primary data source for training and evaluation of our models is the LFM-2b dataset [25], a collection of listening events that span from February 2005 to March 2020 from the music streaming platform Last.fm[3]. The dataset is well-established, and is used in numerous modern research projects on RS. It is the third most frequently used dataset for RS research according to a recent survey by Bauer et al. [3] This research utilizes the user IDs, track IDs, and listening counts of each user for each track. By focusing on user-track interactions we keep the scope of the project limited and allow for a more focused and clear evaluation of the selected models. The data has been filtered to exclude users who have not played a song at least five times and tracks that have not been played at least five times. These thresholds are necessary for a meaningful use of collaborative filtering algorithms [15]. Furthermore, only those user-song interactions are being retained where the user played the song a minimum of ten times. This is to remove noisy data and to focus on songs that the user has a strong preference towards. [15, 12]. Finally, we consider binary user-track interactions, where a value of 1 is assigned if the user has listened to the track at least once, and a value of 0 otherwise.

3.2 Framework

For this research, we used the Elliot framework [1]. The framework offers an extensive selection of state-of-the art models but also well-established traditional models used as baselines in research on recommender systems [3]. The decision was taken to utilise a framework, as this permits other researchers to reproduce or extend this research within a defined environment. This framework is also one of the recommended frameworks for RS research as defined by ACM RecSys.[4] [2].

3.3 Algorithms

An extensive literature review was conducted to select our models. This review was based on a number of surveys, similar research that compared models and a key reference book as main sources [30, 21, 3, 14]. The criteria were that they should be well-established and

[2]https://shorturl.at/ERNPs

[3]Last.fm

[4]https://github.com/ACMRecSys/recsys-evaluation-frameworks

high-performing RS, in order to give the best possible representation of their categories. All of the models utilize different methodological approaches. This diversity allows us to gain a deeper insight of what each category is capable of.

There are three main types of recommendation algorithms: collaborative filtering, content-based, and hybrid. Collaborative filtering algorithms recommend items liked by similar users, relying on user-item interactions, while content-based filtering algorithms recommend items similar to those liked by the user in the past, utilizing item features. Hybrid models combine these two techniques [21]. With the exception of the non-personalised models, are all of the selected models based on collaborative filtering. This facilitates the comparison of models and results.

Here we present the traditional recommendation models selected for this study, including both non-personalized and personalized algorithms. The non-personalized models, in particular, act as benchmarks, setting a standard for what even the simplest recommendation algorithms should achieve.

Traditional models:

Random A non-personalized model creating random recommendations for users.

MostPop A non-personalized model recommending the most overall popular songs to everyone.

BPRMF Bayesian Personalized Ranking with Matrix Factorization combines the strengths of matrix factorization, the most popular technique for recommender systems and the foundation of many effective algorithms [31], with a Bayesian framework to optimize for personalized item ranking [19]. It is frequently employed as a non-neural baseline in RS research and is known to perform reasonably well [30, 2].

UserKNN User k-Nearest Neighbours generates recommendations by identifying similar users based on their preferences and behaviors [20]. Despite its early development in 1994, has the model consistently shown strong performance, highlighting the effectiveness of neighborhood-based collaborative filtering methods in modern recommender systems [2].

Here we present the deep learning recommendation models selected for this study. The models were selected according to the previously defined criteria, with consideration being given to the specific context, namely, the recommendation of music for users under the age of 18. To address the unique needs of this demographic, we chose models that are adept at extracting intricate patterns from implicit feedback and are effective in handling sparse data, which is typical when dealing with children's interaction data [21, 28, 10]. In order to as comprehensively as possible cover the range of deep learning techniques utilised by recommendation models, all of the selected models were chosen from different sections as outlined in this survey on deep learning-based recommender systems [30].

Deep learning models:

NeuMF Neural Matrix Factorization employs a combination of the established technique of matrix factorization and the more recently developed technique of multi-layer perceptrons, collectively known as neural collaborative filtering. It is a new (2017) state-of-the-art model and is widely used as a neural baseline in the recent literature [2].

MultiVAE Multimodal Variational AutoEncoder is a collaborative filtering model that uses autoencoders to learn and capture complex patterns in user-item interactions. It encodes the interactions into a compressed format and then decodes them, helping to make more accurate recommendations based on these learned patterns [13]. It is also a recently developed model (2018)

and has outperformed existing non-neural baselines in an analysis by Ferrari Dacrema et al. [6, 2].

AMF Adversarial Matrix Factorization is a collaborative filtering model that uses adversarial learning. It involves a generator creating fake user-item interactions and a discriminator trying to distinguish between real and fake interactions. This adversarial process helps the model learn better representations of user preferences [11]. Besides it outperforming Bayesian Personalized Ranking according to He et al.[11], is this method of operation known to work well with sparse data [16].

NGCF Neural Graph Collaborative Filtering utilizes graph neural networks (GNN) to exploit high-order connectivity in user-item interaction graphs, effectively capturing the complex structures within the data [29]. While this is not a particularly well-known neural baseline, can it be of interest to gain an insight into GNN's and is it argued that they are "a strong candidate in learning the difficult semantics and impact of multiple types of behaviors" [14].

3.4 Metrics

Here we present the performance and beyond-performance metrics employed in our research.

Performance

One crucial property of a recommendation model is performance, which measures how many items a recommendation system suggests that are relevant to a user [21]. It typically involves assessing whether the items included in the top N recommendations match the items that the user has interacted with that were excluded from the training data.

Normalized Discounted Cumulative Gain (nDCG) evaluates the ranking quality of the recommendations by considering both the the accuracy of recommended items and their positions in the ranking list. This is a popular metric for deciding whether a model's overall recommendations suggests are of interest to the user [3]. **Mean Reciprocal Rank** (MRR) measures the average of the reciprocal (1 divided by the rank) of the first relevant item in a recommendation list. It evaluates how quickly the first relevant item appears, giving higher scores to systems that rank relevant items earlier. This is also a popular metric, which can be used to determine how quickly a model suggests an item that the user is interested in. This is of relevance in the context of children as they require faster stimuli caused by their shorter attention span when using online platforms [27].

Beyond-performance

While the primary goal of a recommender system is typically the performance of its recommendations, the research community has become aware that performance is not the only aspect of a recommendation model suitable towards children. It is important for the developer of such a model to take into account the needs and innate characteristics of children [10]. For this reason, we will employ diversity and novelty metrics to evaluate our recommendation models. These metrics will enable us to identify whether our models recommend diverse songs and new songs have not been interacted with frequently. Our beyond-accuracy metrics are **Gini Index** and **Expected Popularity Complement** (EPC), both selected because they are straightforward yet effective measures of diversity and novelty respectively. The former measures the inequality in the distribution of recommended items, with a lower value indicating more diverse recommendations across users. EPC, a novelty metric, evaluates how well a recommendation system promotes less popular items, by favoring less mainstream content.

3.5 Experiments

Here we present our exact setup when running the selected models.

Regarding model parameters, have we selected an arbitrary seed for `Random` and are none necessary for `MostPop`. For the remaining models, did we select parameters as they were given in the Elliot documentation. We did not employ hyperparameter optimization in this initial analysis. Because of this, are we able to observe the baseline performance of the models and conduct a fair comparison under standardised conditions. However, in order to gain further insight into the characteristics of deep learning models, we conducted a second iteration of our highest performing deep learning model `MultiVAE`, now applying hyperparameter tuning techniques. This is to provide a general indication of the extent to which deep learning algorithms would benefit from hyperparameter tuning and whether they are capable of outperforming traditional baselines as a result. It's results are presented as `MultiVAE HP`.

The data is initially partitioned into a test and training set via a five-fold random cross-validation approach. The models are trained and a top-20 recommendation list is then generated, this is a common cutoff threshold in recommender system evaluation [2]. Subsequently, the selected metrics are computed in accordance with the recommendation list against the test set, and paired t-tests are performed to verify whether the observed differences in metric values are statistically significant for the models in question.

4 Results

Here we present the results of our conducted experiments.

4.1 Performance

The results for each model on the performance metrics are presented in Table 1 in Appendix A and visualized in figure 1.

As can be inferred from the figure, does `UserKNN` score the highest on both performance metrics. `BPRMF`, `NeuMF` and `NGCF` perform very poorly. The deep learning models `MultiVAE` and `AMF` show moderate performance, with `MultiVAE HP` performing reasonably well.

4.2 Beyond-performance

The results for each model on the beyond-performance metrics are presented in Table 2 in Appendix A and visualized in figure 2.

As the graph indicates, do most models show very poor performance. We do not consider the high Gini index of `Random`, as it is expected to have a diverse set of recommendation but does not recommend any relevant content. All models show a very low Gini index with the only exceptions being `UserKNN` and `MultiVAE HP`, who show moderate values. Furthermore, only the EPC values of `UserKNN` and `MultiVAE HP` are of substantial amount, with `MultiVAE` and `AMF` showing moderate values.

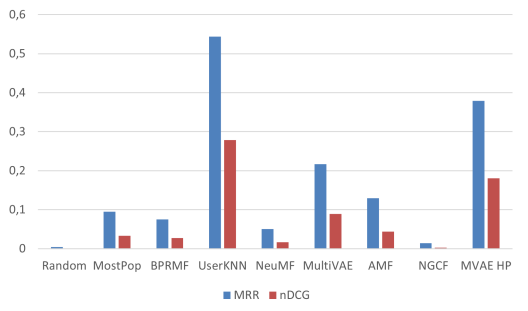

Figure 1: Results performance metrics

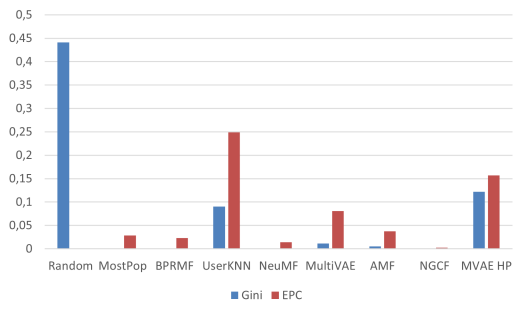

Figure 2: Results performance metrics

5 Discussion & Limitations

As can be inferred from the data, did most models perform very poorly. This is due to none of the models utilizing hyperparameters. While hyperparameter optimization (HPO) is not exclusive to deep learning algorithms, but is applicable to machine learning tasks in general, deep learning models often introduce additional hyperparameters. As stated in this survey by Zhang et al. [30], deep learning often requires extensive HPO to achieve optimal performance. Nevertheless, we continue to assess and discuss each model, as this provides a comprehensive understanding of the fundamental performance of these models, which do not require extensive tuning.

It is also important to consider the scale of the dataset. After the filtering process, the set was found to contain 1,786 underage users. This study demonstrates the performance of the models on a small dataset. However, it is essential to consider the scalability of these models when applied to larger datasets.

Performance

A key attribute of an effective recommender system is its ability to consistently perform well. While other factors may be relevant for a RS designed for children, it is still essential that the recommended content is of interest to the user.

Because of `UserKNN` scoring the highest on performance metrics, does this indicates that user-based collaborative filtering methods are capable of effectively capturing the preferences of children. The reliance of `UserKNN` on neighbouring (similar preference) users implies that children have a lot of similar tastes and are influenced by their peers. It has a particular high MRR score, which implies that users receive the desired music recommended without having to scroll through many irrelevant options, leading to increased user satisfaction.

For the deep learning models, is `MultiVAE` optimized for recommendation with implicit data [30]. This technique may be a promising option, particularly given that implicit data is often the only data available from children. Given that `AMF` is known to work well with sparse data, it also performs reasonably well in this context and may be a promising approach to recommending to children when data is severely limited. An important observation is the increase in performance from `MultiVAE HP`. While still outperformed by `UserKNN`, is it evident that even basic HPO can greatly increase deep learning models' performance.

Beyond-performance

UserKNN has the highest Gini and EPC score of the models not utilizing HPO, indicating that it recommends the most diverse and novel selection of songs. This implies the model best satisfies the subsection of the previously stated specific needs of children that we focussed on, while also assisting music streaming platforms in presenting a larger portion of their catalogue to users, rather than only popular music. While `UserKNN` has the highest values, does it still perform poorly, with only the EPC value of UserKNN being reasonably high, indicating it does sporadically recommend unpopular content. The poor performance of all models is likely due to their bias towards popular items, as they optimize solely for accu-

racy. A different optimization technique or some other way, enforcing a more diverse and novel set of recommended songs, is therefore recommended. We see a substantial increase however, in the Gini index and EPC value of `MVAE HP`. This further indicates the immense potential of deep learning algorithms when even basic HPO makes such a significant impact.

Relating this study to a similar study where they compared deep learning algorithms to traditional baselines conducted by Anelli et al., can we see some similarities although they don't evaluate on children's data. An Amazon Digital Music[5] is employed in their research and they utilize all our selected models except `AMF` and `NGCF`. While their models show generally smaller performance, is `UserKNN` also their best performing model, even with the deep learning models utilizing HPO. In terms of beyond-performance metrics, do `UserKNN` and our deep learning models show similar performance. This suggests that, with a sufficient amount of HPO, deep learning models can achieve the same results as UserKNN.

Further limitations

While the LFM-2b dataset is quite extensive, the Last.fm platform is mostly used by adults, resulting in a limited amount of data from underage users. This study also focused on data from users with a known age of under 18, meaning that any underage users with an unknown age were not considered in this research. As the Last.fm platform employs its own recommendation algorithm, it is likely that this influences user listening histories, that is consequently reflected in the dataset. Additionally, when prefiltering our data, we bi- narised it, which meant that the significance of a higher play count was lost when training and evaluating the model.

Furthermore, was this research was conducted using offline evaluation, which assesses models using historical data without real-time interactions. Online evaluation, commonly using A/B tests, however, is considered the gold standard for effectiveness[8]. Because of our offline evaluation, were we unable to measure the real-time adaptability and user engagement of our models, which can give us a more accurate assessment.

6 Conclusions

In this paper, we evaluated several state-of-the-art deep learning algorithms and compared them in terms of performance, diversity and novelty metrics. Although a large part of the models performed poorly due to them not utilizing hyperparameter optimization (HPO) , did we get an insight into their baseline performance, which enabled us to provide several recommendations towards building an optimal recommender systems suited for children and their specific needs and capabilities. We discovered, that in the specific context of which we conducted this study, deep learning models do not outperform traditional methods. To contextualize however, did we run a well-performing deep learning model with HPO. This provided us the insight that deep learning models do have the potential to outperform traditional baselines in terms of our selection of metrics, as with a very basic hyperparameter optimization did the deep learning model achieve almost similar results to the traditional baseline. Overall is there some potential to be seen in deep learning in music recommender systems for children, but is further research required to explore their full capabilities.

7 Responsible Research

We have devoted considerable attention to the ethical considerations and the integrity of our research. As our experiments were conducted via an offline evaluation, no real participants were needed, and therefore no harm could be done to them. The dataset

[5]https://music.amazon.com/

is publicly available and the privacy of the users that have contributed to the set is also preserved by only storing an incremental numeric user identified that cannot be traced back to a particular Last.fm user name. [25]. The data obtained from the experiments was directly inputted into the graphs without any alteration. The methodology section also provides a comprehensive explanation of the experiments conducted, thereby ensuring the reproducibility of the results. Furthermore, this paper properly cites all sources and gives credit to original ideas and previous research.

8 Future Work

As stated previously, does this research have several limitations. Future reseach can address these limitations and bring us closer to the optimal music recommender for children. Reiterating on this research, but employing HPO for all models will give a better insight into what these models are capable of, instead of their baseline performance. A larger dataset can give insight into the scalability of the models and research the potential of their real-world application. We utilized a small selection of features, that only covered a partition of the specific needs and capabilities of children, and future studies can evaluate according to other metrics which cover other aspects of these children's needs. Finally, it would be beneficial to conduct an online study that can obtain ethical committee approval, as this would provide a more comprehensive understanding of the way children engage with our selected models.

References

[1] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco M. Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, 2021*, 2021.

[2] Vito Walter Anelli, Alejandro BellogÃ-n, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. Top-n recommendation algorithms: A quest for the state-of-the-art. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, July 2022.

[3] Christine Bauer, Eva Zangerle, and Alan Said. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *ACM Trans. Recomm. Syst.*, 2(1), mar 2024.

[4] Hung-Chen Chen and Arbee L. P. Chen. A Music Recommendation System Based on Music and User Grouping. *Journal of Intelligent Information Systems*, 24(2):113–132, March 2005.

[5] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *CoRR*, abs/1911.07698, 2019.

[6] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. *CoRR*, abs/1907.06902, 2019.

[7] Michael D. Ekstrand. Challenges in evaluating recommendations for children. 2017.

[8] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness.

In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 23–24 Feb 2018.

[9] Ferdos Fessahaye, Luis Perez, Tiffany Zhan, Raymond Zhang, Calais Fossier, Robyn Markarian, Carter Chiu, Justin Zhan, Laxmi Gewali, and Paul Oh. T-recsys: A novel music recommendation system using deep learning. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6, 2019.

[10] Emilia Gómez, Vicky Charisi, and Stéphane Chaudron. Evaluating recommender systems with and for children: towards a multi-perspective framework. In Eva Zangerle, Christine Bauer, and Alan Said, editors, *Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2021 colocated with the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, The Netherlands, September 25, 2021*, volume 2955 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[11] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, June 2018.

[12] Oleg Lesota, Alessandro B. Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Analyzing item popularity bias of music recommender systems: Are different genders equally affected? *CoRR*, abs/2108.06973, 2021.

[13] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[14] Matteo Marcuzzo, Alessandro Zangari, Andrea Albarelli, and Andrea Gasparetto. Recommendation systems: An insight into current development and future research challenges. *IEEE Access*, 10:86578–86623, 2022.

[15] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing Management*, 58(5):102666, 2021.

[16] Huy Xuan Nguyen and Le Minh Nguyen. Attention mechanism for recommender systems. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 174–181. Waseda University, 2019.

[17] Dip Paul and Subhradeep Kundu. *A Survey of Music Recommendation Systems with a Proposed Music Recommendation System*, pages 279–285. 01 2020.

[18] Maria Soledad Pera, Jerry Alan Fails, Mirko Gelsomini, and Franca Garzotto. Building community: Report on kidrec workshop on children and recommender systems at recsys 2017. *SIGIR Forum*, 52(1):153–161, aug 2018.

[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. *CoRR*, abs/1205.2618, 2012.

[20] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and

John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. Association for Computing Machinery.

[21] F. Ricci, L. Rokach, and B. Shapira. *Recommender Systems Handbook*. Springer US, 2022.

[22] Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9, 2022.

[23] Markus Schedl. Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics*, 5, 2019.

[24] Markus Schedl and Christine Bauer. Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. *CoRR*, abs/1912.11564, 2019.

[25] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 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*, pages 337–341, 2022.

[26] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *CoRR*, abs/1710.03208, 2017.

[27] Alexander J Simon, Courtney L Gallen, David A Ziegler, Jyoti Mishra, Elysa J Marco, Joaquin A Anguera, and Adam Gazzaley. Quantifying attention span across the lifespan. *Front. Cogn.*, 2, June 2023.

[28] Robin Ungruh and Maria Soledad Pera. Ah, that's the great puzzle: On the quest of a holistic understanding of the harms of recommender systems on children, 2024.

[29] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 165–174, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1), feb 2019.

[31] Wayne Xin Zhao, Zihan Lin, Zhichao Feng, Pengfei Wang, and Ji-Rong Wen. A revisiting study of appropriate offline evaluation for top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 41(2), dec 2022.

[32] Wang Zhou, Jianping Li, Malu Zhang, Yazhou Wang, and Fadia Shah. Deep learning modeling for top-n recommendation with interests exploring. *IEEE Access*, 6:51440–51455, 2018.

A Experimental results

A.1 Performance

model	MRR	nDCG
Random	0.0043278	0.001115026
MostPop	0.094738007	0.032976927
BPRMF	0.074916199	0.027552041
UserKNN	0.543490938	0.278878022
NeuMF	0.050637891	0.01668786
MultiVAE	0.21662134	0.089239653
AMF	0.129356615	0.043758931
NGCF	0.014380967	0.002986716
MVAE HP	0.378804854	0.180790701

Table 1: Performance metric results

A.2 Beyond-performance

model	Gini	EPC
Random	0.44136378	0.001069855
MostPop	0.000314338	0.028291411
BPRMF	0.000366118	0.02323975
UserKNN	0.090044167	0.24864758
NeuMF	0.000314314	0.014464637
MultiVAE	0.01130243	0.080828934
AMF	0.005434915	0.037843418
NGCF	0.000444997	0.002849149
MVAE HP	0.121715588	0.157181419

Table 2: Performance metric results