Factoring in What Gets Listened To

Evaluating the performance of a Factorisation Machine-based music recommender using musical features for child listeners

Konrad Barbers[1]

Supervisor(s): Sola Pera[1]**, Robin Ungruh**[1]

[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 23, 2024

Name of the student: Konrad Barbers
Final project course: CSE3000 Research Project
Thesis committee: Sola Pera, Robin Ungruh, Julian Urbano Merino

Abstract

Recommender systems play a large role on contemporary music platforms, but they tend to work less well for non-mainstream listeners such as children. Additionally, there is no one strategy to perfectly capture a listener's music preference. As children develop understanding of music in different stages, using features they respond to might make recommendations more accurate. Therefore, this study seeks to investigate the effectiveness of a recommender which utilises matrix factorisation augmented with the musical features of tempo, mode, dynamics and time signature in recommending songs a child user would like. We evaluate the quality of this recommender based on the Factorisation Machine algorithm by comparing it to a non-augmented variant of the same algorithm and similar ones using fewer of the same features. Results show that while adding features improves the quality of recommendations, adding too many or the wrong features diminishes said improvement, although more research is needed in this direction.

1 Introduction

While the topic of music recommender systems is quite thoroughly researched through a broad lens of a "normal" user [6], how recommenders perform on recommending music to child users has not been investigated as much. How well a system performs at recommending music to a user scales with the openness of the listener towards music they would usually not listen to [2]. This is less applicable to children since they tend to listen to a more narrow variety of music. Additionally, listening behaviours between different child listeners can differ greatly, making it hard to find a single system that works for all users [7]. Therefore, a new or hybrid music recommender system may have increased effectiveness when recommending music to children compared to standard approaches like user-based collaborative filtering.

It is noted that using audio features for recommending music "is advocated by MIR researchers as an alternative or complement to metadata and collaborative filtering methods (and that) Recommender systems based on audio content are not susceptible to popularity bias" [6, p. 457]. Additionally, "matrix factorization models for music rating prediction can successfully incorporate additional information such as [...] multi-level taxonomy information like genre" [6, p. 472]. Therefore, a music recommender system that leverages audio content for a matrix factorisation approach may perform better at recommending music for children popularity indications for songs among adults can be circumvented. As well as this, children can be served songs based on their own preferences and recognition of musical features according to their psychological development [8]. Hence, in this work we approach the following question:

How well does a music recommender system using matrix factorisation leveraging various audio features perform for child users?

To inform the design of such a system, the audio features which should be used for this recommender need to first be considered. Since this research concerns the creation of a recommender and not the extraction of features, and since a song's metadata on Spotify already contains multiple audio features, a selection of those was picked based on children's development in recognition of musical features [8]. To incorporate these features into the recommendation process, an algorithm that supports the addition of explicit feature data is needed. Factorisation Machine (FM) fits this mould as it is matrix factorisation-based and allows for augmentation data [4]. Next, the quality of a music recommender using these features needs to be evaluated. This can be done by comparing performance measures of this algorithm against a similar algorithm without additional features. Lastly, the effect each individual feature has on the recommendation quality should be determined; this is done by again using the same baseline algorithm and augmenting it with all-but-one of the additional features.

Following this paper's introduction, we describe the resources and tools needed for the experiment in section 3, after which we display the results in section 4. Next, we lay out ethics and reproducibility concerns in section 5 as well as limitations of the experiment in section 6, before discussing the results in section 7. Finally, section 8 concludes this paper and gives recommendations for further research.

2 Related Work

While the topic of music recommender systems is quite thoroughly researched, how they interact with child users is not as much. However, there are still a few works forming the basis of this research.

Music recommender systems. Recommender systems are well-researched, and on the topic of music recommenders, Schedl et al. [6] have investigated various recommender systems, their traits and strengths.

Child listeners and streaming. Spear et al. [7] have analysed the behaviour children of ages 6 to 17 exhibit on music streaming platforms, finding that no one single recommender stategy seems to be adequate for capturing children's listening habits. Additionally, Schedl et al. [5] have conducted research into what kinds of music children tend to listen to, noting significant differences in the music prerferences of children and adults.

Nonstandard listeners. More broadly, Kowald et al. [2] have conducted a study into recommender systems for various non-mainstream listeners, finding that the performance of music recommenders seems to increase with a listener's openness to other music they may be unfamiliar with.

3 Experimental Setup

This section details the methodology and resources necessary to conduct the experiment. It lists the datasets of listening events and song features needed, the algorithm used and how the experiment is conducted.

3.1 Data

To be able to conduct the experiment, we need data on children's listening preferences and the auditory features of songs.

Listening events. For the listening events, we use the dataset LMF-2b [3], specifically the listening counts subset. This dataset lists how many times each user has listened to every song they have interacted with. Since we are focussing on child users, only the interactions of users under 18 were kept. Additionally, all interaction with a count of less than 10 were discarded since a lower play count can mean a user likes a song less, and to reduce the size of the dataset to make the experiment less resource extensive. Lastly, only users who interacted with at least 10 songs in the set and songs which had been listened to by at least 10 users were kept for the same reasons.

Song Features. The features we use to extend the recommendation are dynamics (i.e. average loudness of a song), tempo, time signature and modality (meaning whether the song is in a major or minor key). These features were decided upon firstly because, according to Zimmerman [8], these all are features that children develop awareness of at different stages, and secondly, because these features are already parameterised in the metatadata of songs on Spotify, which make up most of the listening events in LFM-2b and can be extracted using using Spotify's API[1]. To reduce the size of this dataset for easier computation, songs which are not present in the prefiltered listening event dataset detailed above were filtered out.

3.2 Recommendation Algorithm

As the recommendation algorithm, we use Factorisation Machine (FM) [4]. This algorithm is a generalised model which leverages explicitly provided features of items or users and uses factorisation to generate recommendations. To do this, the algorithm utilises both support vector machine and matrix factorisation approaches, allowing us to use the ability of matrix factorisation to deal with sparse data but which usually only allows implicit features, and the addition of explicit features.

3.3 Conducting the Experiment

To train and evaluate our recommender system, we use Elliot [1], a customisable framework for conducting recommendation experiments with a built-in implementation of FM. All of the files used in this experiment can be found in the repository[2].

The goal of our experiment is to determine whether using the selected features improves music recommendations. For the experiment, we compare the performance of the factorisatiom machine-based recommender using an N-1 ablation study. This means that, on top of comparing the recommender to a control which does not use any additional data, we will compare it to four similar recommenders, each not using one of the four features, for a total of 6 recommenders. For each of them, we train them for 10 epochs and repeat this five times, taking the best epoch of all five folds.

We collect normalised discounted cumulative gain (nDCG) and mean reciprocal rank (MRR) metrics for each of these as performance scores. MRR is calculated by adding together the reciprocal of the rank of items recommended compared to the desired result, while nDCG calculates the relevance of a list of recommendations to a hypothetical ideal list. For both of these, a higher score means a better performance, with 1 being the theoretical best. We use both metrics to compare the performances of all iterations of the recommender in the experiment.

4 Results

Features used	nDCG	MRR
Control	0.13077	0.31161
All features	0.15900	0.32225
No loudness	0.15974	0.37828
No mode	0.15369	0.35678
No tempo	0.18303	0.49598
No time signature	0.17544	0.41342

Table 1: The best nDCG and MRR scores of all variants of the algorithm.

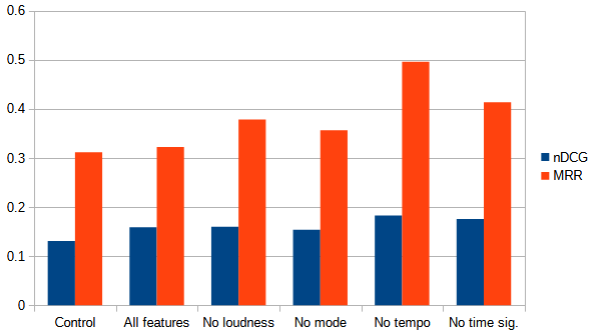

Figure 1: A visualisation of the data given in Table 1

In the experinemt, it was found that all recommenders with added features outperform the recommender with no added features. What is surprising is that, with the exception of the algorithm ignoring mode, having fewer features seems to improve the quality of recommendations.

As mentioned before, ignoring the musical mode of a song seems to have a negligible impact on the performance of the recommender, which indicates that the mode is the most important feature to factor in out of all four features used in this experiment. Curiously, while this recommender scored lower in terms of nDCG compared to using all features (0.15369 vs. 0.15900), this recommender achieved a higher MRR value (0.35678 vs. 0.32225), meaning that while this recommender ranked relevant items more highly, it did not provide as high a quantity of relevant items as the recommender with all features.

[1]https://developer.spotify.com/documentation/web-api/reference/get-audio-features

[2]https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Pera_Ungruh/kbarbers-Music-Recommender-Systems-Youngsters

As clearly visible in Figure 1, removing tempo from the features used has the biggest impact on performance, meaning that out of the four features, tempo is the one which adds to the quality of recommendations the least. This might be due to beats per minute, the unit this feature is represented it, having high internal variance.

5 Responsible Research

5.1 Ethics

For all research, it is important to reflect on potential ethical concerns, especially when said research involves people's data.

Child data safety. Since the user group we are evaluating the recommender for is children, it is especially important that their data is anonymous. Since the only identifiable characteristic in all LFM-2b sets for the users are age, gender and the country of the account, it is basically impossible to trace a user's data back to them.

Bias in selecting features. As stated in Section 3, while the features chosen for the recommender are selected due to them being associated with different developmental milestones in children, a primary reason was that these features are already present as a song's attributes on Spotify. While this was a decision made to focus on the recommender itself rather than feature extraction, which features to select and how to acquire them should be considered more in the future.

5.2 Reproducibility

It is also important to ensure that research is reproducible.

Song feature extraction. While we did process the song feature data to only contain the features needed for the experiment, we did not extract the song features via the Spotify API ourselves. This was done my one of our research group members.

Bias in user data. Since the LFM-2b dataset uses data from last.fm[3], which is a service one can voluntarily add their music streaming accounts to, there is an obvious selection bias in the users who choose to do so. However, since similar research uses the same dataset or its predecessor, LFM-1b, this bias should not affect the comparative conclusions of this research.

6 Limitations

Since using FM proved quite resource intensive, the number of interactions which could be used for running the experiment had to be reduced greatly. The specific bottleneck here is the usage of GPU memory. Beyond the base reduction of the dataset detailed in Section 3, the test ratio had to be configured as 0.999, meaning that only 0.1% of the interactions in the listening event set were used for training. This has surely caused the prediction performance to greatly differ from what it would be with a higher number of interactions used for training. It would have been possible to run the experiment at a test ration of 0.998, but this increased the time taken for training and evaluation to increase sixteenfold, making it infeasible for the limited time frame of this research project.

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

7 Discussion

In this paper, we aim to evaluate the performance of a music recommender which uses auditory features of songs for recommending music to child listeners.

The results clearly show that added features always improve the performance of the Factorisation Machine algorithm, but that using fewer than 4 features is more beneficial, suggesting a quality-over-quantity approach in the future. This might be due to some features acting in opposition to each other, which could be a further point of investigation. In addition, these features selected were quite arbitrarily chosen, as elaborated on in section 5, so other features may provide a better result, especially if consulting children on what features they like and respond to in music.

The increase of performance when removing features was quite surprising, as it was expected that using all features together would improve the result. This might be due to being forced to work with a very low training ration of 0.1% of the available user data. This also means that using this Factorisation Machine based recommender may prove too computationally expensive for regular use, although that might also be due to the implementation of it in Elliot [1].

Since the data split for training the model was so low, comparing the results to similar research is not really applicable here; rather, these results should be compared within this study. If this experiment was repeated with a more standard train-validation split, like $80:20$, the following results would be more applicable for comparison with other recommenders. Additionally, different recommenders using a similar strategy of utilising explicit features might be less expensive to use.

8 Conclusions and Future Work

With the research presented in this paper, we sought to examine whether using audio features present within songs would improve the performance of a matrix factorisation-based music recommender for children. It was found that, compared to a recommender using the same model but no features, using explicit item values for loudness, mode, tempo and time signature does create a better-performing recommender. However, it is clear that using fewer features, specifically only loudness, mode and tempo, demonstrates significantly better performance compared to using all four features. However, these results are not necessarily conclusive since we were forced to only use 0.1% of the available dataset of user-song interactions to train the model for performance reasons.

For more conclusive results, using a larger train-validation split will allow the results of an experiment like this to be comparable to studies of similar recommender algorithms. Smilarly, this experiment could be repeated with not only children but also adult listeners.

It is also advisable to investigate different features besides the ones used in this research and to find the optimal number of features to use for this.

Lastly, while this experiment used feature data that was already collected prior, different sources of data, such as signal processing, could produce a better-performing recommender.

References

[1] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2405–2414. ACM, 2021.

[2] Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex. Support the underground: characteristics of beyond-mainstream music listeners. *EPJ Data Science*, 10, 2021.

[3] Johannes Kepler Universität Linz.

[4] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000, 2010.

[5] Markus Schedl and Christine Bauer. Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. *1st International Workshop on Children and Recommender Systems (KidRec 2017), co-located with 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, 27 August*, 2017.

[6] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. *Music Recommender Systems*, pages 453–492. Springer US, Boston, MA, 2015.

[7] Lawrence Spear, Ashlee Milton, Garrett Allen, Amifa Raj, Michael Green, Michael D Ekstrand, and Maria Soledad Pera. Baby shark to barracuda: Analyzing children's music listening behavior. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 639–644, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Mary P. Zimmerman. Musical characteristics of children. *Visions of Research in Music Education*, 17, 2007.