Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Can We Ensure Accuracy and Explainability for a Math Recommender System?

Yiling DAI[a*], Brendan FLANAGAN[b] & Hiroaki OGATA[a]
[a] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b] *Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Japan*
*dai.yiling.4t@kyoto-u.ac.jp

Abstract: Providing explanations in educational recommender systems are supposed to increase students' awareness of the recommendations, trust toward the system, motivation to adopt the recommendations. With the expectation to have a higher prediction accuracy, more and more complex recommendation models are developed, which are difficult to explain. It remains debatable that whether there exists a trade-off between the accuracy and explainability of recommender systems. In this study, we focus on the explainable math quiz recommender system--- Naïve Concept Explicit (Naïve CE) proposed in our previous work. We are interested in knowing whether the explainable Naïve CE has a good prediction accuracy compared with a powerful but less explainable model--- Matrix Factorization (MF). We also proposed a combined model CE+MF to preserve the explainability of Naïve CE and predicting power of MF. We then used a long-term quiz answering dataset to evaluate the models' accuracy as to predicting students' correctness rate of the quizzes. The results revealed that 1) The explainable model Naïve CE had a lower accuracy than the less model MF given the sparse dataset; 2) Combining two models achieved a moderate accuracy in predicting students' answers while preserving the explainability of Naïve CE. Our study served as an example of how to develop an inherently explainable educational recommender system and how to improve the accuracy by integrating more complex models.

Keywords: Recommender system, math quiz, explainability, accuracy, matrix factorization

1. Introduction

To improve learning, educational recommender systems provide recommendations based different criterion including learning activities (Afzaal et al., 2021), knowledge states (Ai et al., 2019; Tang et al., 2019), learning goals (Heras et al., 2020; Huang et al., 2019), learning styles (Heras et al., 2020; Klašnja-Milićević et al., 2011) , student profiles (Shanshan et al., 2021) and so on. Unlike consuming entertaining products such as movies or music, reading learning materials or solving quizzes requires higher levels of motivation and cognitive investment of users. Providing explanations of the recommendations to the students is considered as a solution. Some promising results of the explanations' effects have been found: increasing the students' attention towards recommended practices, the willingness to open them (Barria-Pineda et al., 2021); increasing the students' trust in hints, perceived usefulness of them, and the intention to use them again (Conati et al., 2021); increasing the students' perceived unexpectedness, novelty of recommended courses, and the interests in them (Yu et al., 2021); increasing the students' usage of the recommended quizzes (Dai, Takami, et al., 2022; Takami et al., 2022).

Unlike traditional machine learning models such as decision tree and logistic regression, recent models involving neural networks are becoming more and more difficult to understand (Khosravi et al., 2022). It is commonly considered that more complex models have better performance in terms of predicting user's behavior (Molnar et al., 2022; Rudin,

2019). To achieve model transparency, explaining complex models becomes a research field also known as “eXplainable Artificial Intelligence (XAI)” (Arrieta et al., 2020). However, there remain two open issues before suggesting researchers and practitioners to develop more complex models and then struggle to explain it:

1. It is doubted that there is a trade-off between the model’s accuracy and complexity (Molnar et al., 2022; Rudin, 2019). As Gervet et al. (2020) observed in their experiments, the superiority of deep knowledge tracing models over logistic regression models is influenced by the size and shape of the dataset.
2. Should we develop an explainable model in the first place and then improve the accuracy or develop a complex model and try to explain it afterwards (Molnar et al., 2022)?

To address these issues, we focus on a specific educational context of recommending math quizzes in this study. Previously, we proposed a simple and explainable recommender system named Naïve Concept Explicit (Naïve CE) model, which recommends quizzes based on the estimation of the students' mastery level on math concepts (Dai, Flanagan, et al., 2022). The model also provides explanations of why the students should undertake the quiz from the perspective of math concepts. We also conducted a real-life experiment where the concept-based explanations showed positive effects in motivating students to attempt the quizzes (Dai, Takami, et al., 2022). In this study, we are interested in knowing whether the explainable Naïve CE has a good prediction accuracy compared with a powerful but less explainable model--- Matrix Factorization (MF). We also proposed a combined model CE+MF to preserve the explainability of Naïve CE and predicting power of MF. We used a long-term quiz answering dataset in our learning management system to evaluate the models’ accuracy as to predicting students’ correctness rate of the quizzes. The results revealed that 1) The explainable model Naïve CE had a lower accuracy than the less model MF given the sparse dataset; 2) Combining two models achieved a moderate accuracy in predicting students’ answers while preserving the explainability of Naïve CE. Our study served as an example of how to develop an inherently explainable educational recommender system and how to improve the accuracy by integrating more complex models.

2. Related Work

2.1 Explainability of Recommender Systems

Basically, there are two approaches to generate explanations in recommender systems--- model-intrinsic and post-hoc (Zhang & Chen, 2020). In the model-intrinsic approach, the explanation explains exactly how the model generates a recommendation. In educational contexts, an example of model-intrinsic explanation can be to explain how the student’s knowledge state is estimated and why a learning item is considered preferable to improve his/her knowledge state (Dai, Flanagan, et al., 2022). Other model-intrinsic explanations include rule-based (Conati et al., 2021), keyword-based (Yu et al., 2021), concept-based (Dai, Flanagan, et al., 2022; Rahdari et al., 2020), and parameter-based (Takami et al., 2022) explanations. In contrast, the post-hoc approach allows the recommending mechanism to be a “black box” and generates the explanations afterwards. In educational contexts, a post-hoc explanation for a recommended item can be something not necessarily related to the knowledge state estimation but instrumental in motivating the student to accept the recommendation. For instance, an explanation showing how many students have attempted this item may work for students who are weak to peer pressure (Takami et al., 2023). Feature-based explanations were adopted for “black-box” models but Swamy et al. (2022) found that the explainers are not consistent on feature importance. “Black box” or “Deep” methods are commonly considered more powerful in predicting users’ behavior (Molnar et al., 2022; Rudin, 2019). However, this remains doubtful as Gervet et al. (2020) found that deep knowledge tracing models worked better with larger datasets while logistic regression models worked better with denser datasets. There is a concern of over-using

complex models while an explainable model which has a comparable accuracy is available (Khosravi et al., 2022; Molnar et al., 2022).

In the context of learning math, we consider that developing an explainable recommender system in the first place is intuitive and straightforward. We proposed a simple and explainable recommender system named Naïve Concept Explicit (Naïve CE) model, which recommends quizzes based on the estimation of the students' mastery level on math concepts (Dai, Flanagan, et al., 2022). We also conducted a real-life experiment where the concept-based explanations showed positive effects in motivating students to attempt the quizzes (Dai, Takami, et al., 2022). In this study, we aim to investigate the predicting accuracy of Naïve CE compared with other classic recommendation models. We also attempt to improve the accuracy of Naïve CE by combining it with more complex models while preserving the explainability. This study serves as an example of developing inherently explainable recommender system as suggested by Molnar et al. (2022).

2.2 Naïve CE and MF for Recommending Math Quizzes

As Birenbaum et al. (1993) suggested, identifying specific misconcepts and difficult areas is more instructive than a test score for remediation in learning Algebra. Therefore, it is important to recommend math quizzes that address the students' weak points which are readable math concepts. This motived us to propose a concept-explicit recommender system named Naïve CE (Dai, Flanagan, et al., 2022). Naïve CE assumes that solving a quiz requires the knowledge of related math concepts in the quiz. We utilized the student-quiz interactions and quiz-concept associations to estimate students' mastery levels of the concepts and the possibilities for them to answer the quizzes correctly. We then recommended the quizzes based on the possibilities and the expected learning gains in terms of the mastery level updates of the concepts. Naïve CE is inherently explainable as every step of the estimation is a shallow calculation which is understandable for human beings.

However, we also have a concern on the estimation performance of Naïve CE. In other words, how well can Naïve CE estimate students' mastery level and their probabilities to correctly answer the quizzes? Is there an inferiority in estimation performance compared with more accurate but less explainable models? As Barnes (2005) pointed out, it is debatable whether an explicit model with expert-assigned concepts models student performance better than an implicit model with latent factors. Therefore, we selected a classic recommendation model matrix factorization with latent factors as a comparative target of Naïve CE.

Matrix factorization (MF) (Koren et al., 2009; Takács et al., 2008) is a frequently used model to recommend items that users may have interests. This model assumes a user's interest towards an item comes from her/his preferences on some factors and the relatedness of the factors with the item. It then guesses the unseen user-item interactions by learning from observed user-item interactions. Khosravi et al. (2017) applied MF in their model to estimate students' knowledge gaps to answering the quizzes. To integrate the strengths of both models, Abdi et al. (2018) fed the error in Bayesian knowledge tracing model to an MF model, improving the accuracy of estimating student performance. Since MF has been verified as a useful model to estimate student performance and shares some similarities in modeling the problem with Naïve CE, we chose MF as a comparative model in discussing the estimation performance and model explainability. We also propose a method to combine two models so that the readability of concepts in Naïve CE is preserved.

3. Math Recommender Models

3.1 Problem Definition

In learning management systems, the learning activity can be modeled as a sequence of students' reactions towards learning materials. The task is to recommend learning materials that fit to an individual student's learning progress. It is common that the observed student

reactions are limited to a small set of the learning materials. As a result, predicting the student reactions on unseen learning materials is a key step. For the specific context of solving math quizzes, we formalize the problem as follows: Given a set of m students, a set of n math quizzes, and the student correctness rates on the quizzes $R \in \mathbb{R}^{m \times n}$, we want to estimate the student correctness rates on unseen quizzes $\hat{R} \in \mathbb{R}^{m \times n}$.

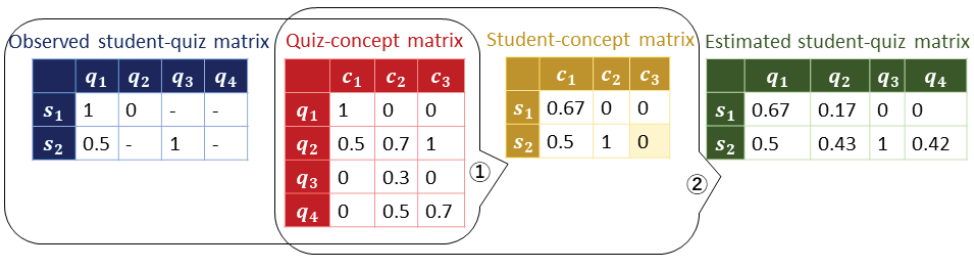
3.2 Naive Concept-Explicit Model (Naïve CE)

Suppose that we have a quiz "Let the set of all positive divisors of 12 be A. Fill in the □ with $\in$ or $\notin$. (1) 2□A (2) 7□A (3) 12□A". Solving the math quiz requires the knowledge of "set" and "positive divisor". The probability that a student can successfully solve a math quiz depends on how s/he understands the required concepts. Motivated by this intuition, we proposed Naïve CE (Dai, Flanagan, et al., 2022) and we review the mechanism of the model as shown in Figure 1:

STEP 1 When given the observed student-quiz matrix whose entries indicate the correctness rates and the quiz-concept matrix whose entries indicate the relatedness of a concept and a quiz, we calculate the students' mastery level on each concept by looking at how they successfully solved quizzes related to the concept. Note that the quiz-concept correspondence is extracted from the quiz information automatically, which is also readable concepts to students.

STEP 2 We then estimate the probability of a student successfully solving a quiz by considering how much of the required concepts has been mastered.

By doing this, the probabilities are modified by the inter-relationships between quizzes and concepts. For instance, s_1 successfully solved q_1 in the history but got an estimated success of 0.67. This is because q_1 requires the knowledge of c_1 and the student failed to solve q_2 which also requires the knowledge of c_1. However, this model falls short in coping with unseen concepts. For instance, s_2 had not attempted any quizzes related to c_3. As a result, c_3 is ignored in STEP 2.



	q_1	q_2	q_3	q_4
s_1	1	0	-	-
s_2	0.5	-	1	-

	c_1	c_2	c_3
q_1	1	0	0
q_2	0.5	0.7	1
q_3	0	0.3	0
q_4	0	0.5	0.7

	c_1	c_2	c_3
s_1	0.67	0	0
s_2	0.5	1	0

	q_1	q_2	q_3	q_4
s_1	0.67	0.17	0	0
s_2	0.5	0.43	1	0.42

Figure 1. The mechanism of Naïve CE model.

3.3 Matrix Factorization (MF)

MF decomposes the observed user-item interaction matrix $R \in \mathbb{R}^{m \times n}$ into two matrices $P \in \mathbb{R}^{m \times k}$ and $Q \in \mathbb{R}^{k \times n}$ such that $R \approx PQ$, where k is the number of latent factors. By minimizing the difference between the estimated and the observed interactions (also viewed as a machine learning process), we get full P and Q, which help us estimate the unseen interactions. The magic part of this model is that it supposes a user's interest on an item comes from the synergistic effect of the user's preferences on the latent factors and the importance of the factors to the item. This idea is somehow similar to Naïve CE except that the factors are "latent" and difficult to interpret. There is a variant of MF which considers user bias and item bias. As a user may tend to highly rate all items or an item of low quality tends to be rated low by all users, introducing bias parameters in MF helps model this situation. By doing this, the sum of PQ and bias better approximates the observed interactions but the intermediate value of PQ is harder to interpret.

3.4 Concept-Explicit Matrix Factorization (CE+MF)

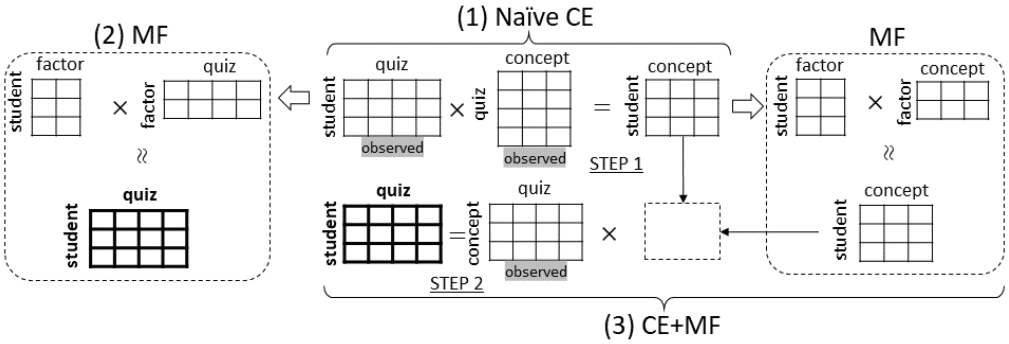


Figure 2. The mechanism of CE+MF.

In Section 3.2, we discussed that Naïve CE is easy to interpret since the concepts are predefined and the computation is simple and straightforward. However, it simply gives up guessing when encountering unseen concepts and quizzes. In Section 3.3, we described that MF is good at estimating the probable values for unseen interactions by iteratively learning from the observed interactions. However, the latent factors of the consisting matrices are difficult to interpret. To take the strengths of both models, we propose a simple hybrid model called CE+MF. As illustrated in Figure 2, Naïve CE utilized the observed student-quiz matrix and quiz-concept matrix to estimate student-concept matrix. The student-concept matrix is again used to adjust the student-quiz matrix. MF simply decomposes the observed student-quiz matrix into two matrices with latent factors. In CE+MF, we first estimate the student-concept matrix as we do in STEP 1 in Naïve CE model. Then, we adopt MF model to update the student-concept matrix where the mastery level on unseen concepts is modified. Last, we update the student-quiz matrix with the updated student-concept matrix as we do in STEP 2 in Naïve CE model. As a result, CE+MF model is supposed to have a higher predictive performance than Naïve CE model while preserving the explainability on concepts.

4. Evaluation

In this study, we aim at investigating two aspects of the recommender models--- quiz mastery estimation performance and explainability. For the quiz mastery level estimation performance, we use a historical dataset collected in a learning system to evaluate whether the models can correctly predict students' answers for unseen quizzes. For explainability, we evaluate whether each step in the model can be explained and what user-friendly explanations can be provided.

4.1 Quiz Mastery Level Estimation Performance

4.1.1 Dataset

We collected quiz answering data from our learning system (Flanagan et al., 2021) generated by the first-year students of a Japanese high school from April 2021 to March 2022. During this period, the students attempted the math quizzes in different contexts such as finishing the assignments, preparing an upcoming test, and self-oriented practicing. As they attempted a quiz, they were required to check the answer and report whether they solved the quiz successfully. Each attempt was recorded as a 0-1 score associated with the student id, quiz id, and timestamp. We computed the aggregated student-quiz correctness rate by taking the average score of all attempts throughout the period. We did not conduct any data filtering process as the temporal order of attempts and the number of attempts are not essential in this evaluation framework. Finally, we obtained a dataset consisting of 27,431 attempts for 270 unique students and 1,919 unique quizzes. Table 1 shows the

statistics of the number of attempts per student and per quiz. After converting the log data into the student-quiz correctness matrix, only 23,155 pairs of students and quizzes were observed, which indicates a very high sparsity of 95.53% ($1 - \frac{\#\ observed\ pairs}{\#\ students \cdot \#\ quizzes}$).

Table 1. *Statistics of the Dataset for Quiz Mastery Level Estimation Evaluation*

	# attempts per student	# attempts per quiz
mean	101.596	14.294
std	97.681	31.404
min	1	1
max	808	426

4.1.2 Metrics

Our main concern is to evaluate whether a model can predict a student's success probability on a quiz. Therefore, we adopt two metrics to measure the agreement between the estimated probability and true correctness rates: **Area under ROC curve (AUC)** is considered an effective metric to measure how well a model separate negative and positive samples across different decision threshold choices (Bradley, 1997). Since the true correctness rates for student-quiz pairs are real numbers between 0 to 1, we first transform the true correctness rates into 1 if it is greater than 0.5, 0 otherwise when applying AUC. **Root mean square error (RMSE)** is used to measure the absolute differences between the estimated probability and the true correctness rates.

4.1.3 Implementation

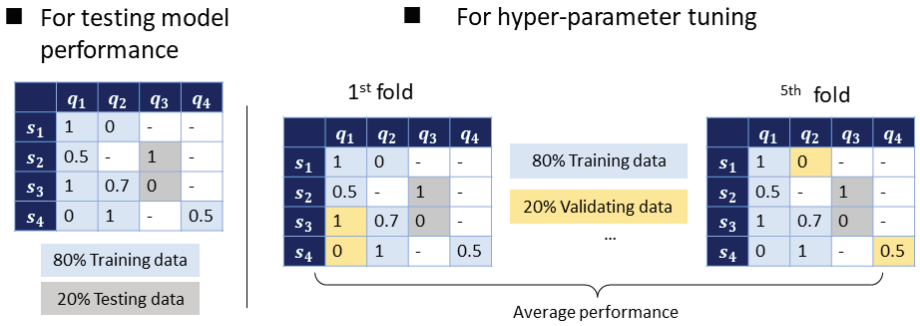


Figure 3. The data splitting process.

As illustrated in the upper part of Figure 3, we set aside 20% of the student-quiz attempts as test data and all models were blind to these data during training or tuning process. For models involving hyper-parameter tuning, we adopted a 5-fold cross validation approach to select the best combination of parameters. As illustrated in the lower part of Figure 3, in each fold, 20% of the data is used to validate the model performance and the average performance of all the folds is treated as the final performance of a combination of parameters. We adopted a grid-search approach to generate the combinations of parameters. Only the performance of the best combination of parameters will be reported in the following section.

The followings are some implementation details of the models: 1) Naïve CE. As described in (Dai, Flanagan, et al., 2022), we adopted text mining techniques to automatically extract math concepts from the quizzes. The entries of quiz-concept matrix were computed using TFIDF (Salton & Buckley, 1988) weighting scheme. 2) MF. We adopted stochastic gradient descent algorithm to obtain the matrices P and Q whose product has the minimum difference with the observed student-quiz correctness rate. We tuned three parameters--- learning rate α, regularization factor γ and number of latent factors k for MF. MF_bias and CE+MF_bias are variants with bias parameters of MF and CE+MF,

respectively. The best combination of hyperparameters is $\alpha = 0.01, \gamma = 0.1, k = 5$ and is used in all MF-related models. The code for MF-related parts was adapted from (Yeung, 2020).

4.1.4 Results

Table 2. *Quiz Mastery Level Estimation Evaluation Results*

	AUC	RMSE
Naïve CE	0.639	0.508
CE+MF	0.688	0.478
CE+MF_bias	0.692	0.464
MF	0.772	0.419
MF_bias	0.799	0.381

Table 2 shows the AUC and RMSE values for each model. Overall, Naïve CE has the lowest, MF has the highest, and CE+MF has the medium performance in both metrics. MF with bias has a better performance both in separate and hybrid models. This result is consistent with our expectation:

- Naïve CE is straightforward but ignores the unseen concepts or quizzes. From the perspective of AUC, this model can discriminate a correct or incorrect answer. However, the performance is nearly a random model from the perspective of RMSE, which means the detailed values of the correctness has a large gap with student's true mastery level.
- MF and MF_bias do a good job at approximating the observed student-quiz correctness rates and therefore the latent factors help to predict the values for unseen pairs. With the AUC value being close to 0.8, MF_bias is supposed to be practically useful to separate a correct or incorrect answer (Mandrekar, 2010). Meanwhile, the RMSE value is still high if we consider a situation where we mistake a student's correctness rate 0.98 into 0.6. However, whether the students can recognize the difference and how they perceive the estimation needs to be further investigated and discussed.
- The hybrid model CE+MF achieves better performance than Naïve CE but still has a distance to the one of MF. We consider a possible reason is the flaws in quiz-concept matrix. First, not all necessary knowledge and skills for solving a math quiz can be detected from the textual information of the quiz. Second, the relatedness of the concepts to a quiz may not be correct just judging from their occurrences in the quiz. Ingesting more elaborated domain models to Naïve CE and observe the performance improvement is one of the future directions.

4.2 Explainability

We compare the explainability of different models as shown in Table 3:

- At the lowest level, all the models can provide information about the estimated student-quiz correctness rate, which can be used to indicate the difficulty of the quiz when recommending the quiz.
- At the medium level, we try to further explain why the correctness rate is as it is. Naïve CE, CE+MF, and CE+MF_bias can provide information about the students' mastery level of concepts, which is the rationale behind the student-quiz correctness rate estimation. However, MF suffers from provide human-readable information about how the student-quiz correctness rate was estimated as the factors are latent. Fortunately, we can leverage the user bias and item bias in MF_bias to provide some extra information. Specifically, we treat the user bias as the student's general ability to solve math quizzes and the item bias as the quiz's general difficulty to all students. Note that it is potential to improve the accuracy of Naïve CE by introducing quiz bias or student bias, which will be future work.
- At the highest level, we want to explain why the concept master level or quiz general difficulty is estimated. In Naïve CE, the concept mastery level is explainable as the quiz-

concept associations is readable to human. In contrast, CE+MF or CE+MF_bias involves a matrix factorization process in estimating the concept mastery level, which is difficult to explain.

Table 3. *Explainability and Explanations of Recommender Models.*

Explainability	Model	Argument	User-Friendly Explanation
Low	Naïve CE MF MF_bias CE+MF CE+MF_bias	student-quiz correctness rate	The estimated difficulty of this quiz for you is 70%.
Medium	Naïve CE CE+MF CE+MF_bias	concept mastery level	Your mastery level of these concepts can be improved by solving this quiz: multiple, integer, proof
	MF_bias	quiz general difficulty	This quiz is difficult (84%) for most of your classmates. Let's have a try!
	MF_bias	student general ability	Your ability (23.4%) to solve quizzes is lower than the average (56.9%) of your classmates.
High	Naïve CE	quiz-concept associations	Integer is important in solving q1(50%), q2 (30%), and q3 (20%). Since you have mistaken q1 twice, you are suggested to address this problem first.

To summarize, Naïve CE is explainable until the highest level as every step in the model is a shallow computation from observed data. MF is only explainable at the lowest level as the latent factors are difficult to interpret. However, the variant MF_bias possesses some additional information about the general information of the quiz difficulty and student ability. This shows a direction to improve the accuracy of Naïve CE by introducing quiz difficulty and student ability parameters. CE+MF is explainable at the medium level as the whole framework is identical to Naïve CE, but a local step involves a MF process.

5. Discussion

As the results in Section 4.1 show, MF has the highest performance of estimating quiz mastery level while Naïve CE has the lowest, CE+MF has the medium performance. The results in Section 4.2 show that we can preserve part of the explainability of an inherently explainable model Naïve CE by combining it with a difficult-to-explain model MF. Given the fact that the quiz answering data set is sparse and the quizzes only have 14 answers on average, we did observe a trade-off between the accuracy and the explainability of different recommender models. As was explored in Gervet et al.'s work (2020), models' performance varies to the type and characteristics of the dataset and features. We think it is important to select a proper recommendation model based on the learning context and data available. Besides, we want to clarify that the explainability of the model is not necessarily equal to the explainability to the students in the practical world. For instance, it would be sufficiently explainable if the students are satisfied with the quiz general difficulty without the interest to understand how it is computed. In this case, MF_bias is not easy to explain in the sense that every step is understandable to humans but explainable to end users.

6. Conclusions and Future Work

In this study, we focused on exploring the accuracy and the explainability of math recommender systems. We took a simple and explainable recommender model Naïve CE as an example and compared its performance with a more complex but difficult-to-explain

model MF. We also attempted to combine two models so that the strengths of both models are integrated. Using a student quiz answering dataset, we found that the explainable model Naïve CE had a lower accuracy than the complex model MF, and the combined model CE+MF had a moderate accuracy. Our results also showed that it is possible to improve the accuracy of an inherently explainable model and preserve the explainability by combining it with more complex models.

Some directions of future work are: 1) Explore other models that can be integrated into the framework of Naïve CE and introduce other learning related parameters such as quiz difficulty and student ability; 2) Investigate how dataset influences the performance of the recommender models; 3) Ingesting more elaborated domain models into Naïve CE and explore the performance improvement; 4) Further explore the difference between the model explainability and the practical explainability to students. Students with different levels of motivations, information literacy, and curiosity may question the recommendations at different levels. It may be important to personalize the explanations according to different scenarios.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

References

Abdi, S., Khosravi, H., & Sadiq, S. (2018). Predicting student performance: The case of combining knowledge tracing and collaborative filtering. *Proceedings of the International Conference on Educational Data Mining*, 545–548.

Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., Li, X., & Weegar, R. (2021). Explainable AI for Data-Driven Feedback and Intelligent Action Recommendations to Support Students Self-Regulation. *Frontiers in Artificial Intelligence*, *4*.

Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-Aware Deep Knowledge Tracing and Exercise Recommendation in an Online Learning System. *Proceedings of the 12th International Conference on Educational Data Mining*.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. *Proceedings of American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, 1–8.

Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable Recommendations in a Personalized Programming Practice System. *Proceedings of International Conference on Artificial Intelligence in Education*, 64–76.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing Knowledge States in Algebra Using the Rule-Space Model. *Journal for Research in Mathematics Education*, *24*(5), 442–459.

Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503.

Dai, Y., Flanagan, B., Takami, K., & Ogata, H. (2022). Design of a User-Interpretable Math Quiz Recommender System for Japanese High School Students. *Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge*.

Dai, Y., Takami, K., Flanagan, Brendan, & Ogata, Hiroaki. (2022). Investigation on Practical Effects of the Explanation in a K-12 Math Recommender System. *Proceedings of the 30th International Conference on Computers in Education*, 7–12.

Flanagan, B., Takami, K., Takii, K., Dai, Y., Majumdar, R., & Ogata, H. (2021). EXAIT: A Symbiotic Explanation Learning System. *Proceedings of the 29th International Conference on Computers in Education*, 404–409.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing? *Journal of Educational Data Mining*, *12*(3), 31–54.
Heras, S., Palanca, J., Rodriguez, P., Duque-Méndez, N., & Julian, V. (2020). Recommending Learning Objects with Arguments and Explanations. *Applied Sciences*, *10*(10).
Huang, Z., Liu, Q., Zhai, C., Yin, Y., Chen, E., Gao, W., & Hu, G. (2019). Exploring Multi-Objective Exercise Recommendations in Online Education Systems. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1261–1270.
Khosravi, H., Cooper, K., & Kitto, K. (2017). RiPLE: Recommendation in Peer-Learning Environments Based on Knowledge Gaps and Interests. *Journal of Educational Data Mining*, *9*(1), 42–67.
Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.
Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, *56*(3), 885–899.
Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30–37.
Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.
Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 39–68). Springer International Publishing.
Rahdari, B., Brusilovsky, P., Thaker, K., & Barria-Pineda, J. (2020). Using Knowledge Graph for Explainable Recommendation of External Content in Electronic Textbooks. *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 Co-Located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, *2674*, 50–61.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.
Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, *24*(5), 513–523.
Shanshan, S., Mingjin, G., & Lijuan, L. (2021). An improved hybrid ontology-based approach for online learning resource recommendations. *Educational Technology Research and Development*, *69*(5), 2637–2661.
Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. *Proceedings of the 15th International Conference on Educational Data Mining*, 98–109.
Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem. *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)*, 267–274.
Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022). Educational Explainable Recommender Usage and Its Effectiveness in High School Summer Vacation Assignment. *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK22)*, 458–464.
Takami, K., Flanagan, B., Dai, Y., & Ogata, H. (2023). Toward Trustworthy Explainable Recommendation: Personality Based Tailored Explanation for Improving E-learning Engagements and Motivation to Learn. *The Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge*, 120–122.
Tang, X., Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 108–135.
Yeung, A. A. (2020). Matrix-factorization-in-python. In *GitHub repository*. GitHub. https://github.com/albertauyeung/matrix-factorization-in-python
Yu, R., Pardos, Z., Chau, H., & Brusilovsky, P. (2021). Orienting Students to Course Recommendations Using Three Types of Explanation. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 238–245.
Zhang, Y., & Chen, X. (2020). Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1), 1–101.