Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-023-00827-8

RESEARCH

Exploring and mitigating gender bias in book recommender systems with explicit feedback

Shrikant Saxena[1] · Shweta Jain[1]

Received: 20 July 2022 / Revised: 1 November 2023 / Accepted: 1 November 2023

Abstract

Recommender systems are indispensable because they influence our day-to-day behavior and decisions by giving us personalized suggestions. Services like Kindle, YouTube, and Netflix depend heavily on the performance of their recommender systems to ensure that their users have a good experience and to increase revenues. Despite their popularity, it has been shown that recommender systems reproduce and amplify the bias present in the real world. The resulting feedback creates a self-perpetuating loop that deteriorates the user experience and results in homogenizing recommendations over time. Further, biased recommendations can also reinforce stereotypes based on gender or ethnicity, thus reinforcing the filter bubbles that we live in. In this paper, we address the problem of gender bias in recommender systems with explicit feedback. We propose a model to quantify the gender bias present in book rating datasets and in the recommendations produced by the recommender systems. Our main contribution is to provide a principled approach to mitigate the bias being produced in the recommendations. We theoretically show that the proposed approach provides unbiased recommendations despite biased data. Through empirical evaluation of publicly available book rating datasets, we further show that the proposed model can significantly reduce bias without significant impact on accuracy and outperforms the existing model in terms of bias. Our method is model-agnostic and can be applied to any recommender system. To demonstrate the performance of our model, we present the results on four recommender algorithms, two from the K-nearest neighbors family, UserKNN and ItemKNN, and the other two from the matrix factorization family, Alternating Least Square and Singular Value Decomposition. The extensive simulations of various recommender algorithms show the generality of the proposed approach.

Keywords Recommender system · Gender bias · Fairness

✉ Shweta Jain
shwetajain@iitrpr.ac.in

Shrikant Saxena
shrikant.saxena.here@gmail.com

[1] Computer Science and Engineering, Indian Institute of Technology, Ropar 140001, Punjab, India

Published online: 25 March 2024

1 Introduction

Recommender systems influence a significant portion of our digital activity. They are responsible for keeping the user experience afresh by recommending varied items from a catalogue of millions of items and also adapting their recommendations according to the personality and taste of the user. Therefore, a sound recommender system may go a long way in improving user experience quality, hence the user retentivity of a digital outlet.

Recommender systems have historically been judged on their accuracy (Herlocker et al., 2004; Shani & Gunawardana, 2011). When it is concerned with other factors such as novelty, user satisfaction, and diversity (Hurley et al., 2011; Knijnenburg et al., 2012), the focus continues to be just on the satisfaction of the information needs of the users. Although of immense importance to the relevance of a recommender system, these criteria do not capture the complete picture. In recent years, the public and academic community have scrutinized artificial intelligence systems regarding their fairness. It has been observed that the results generated by various recommender systems reflect the social biases that exist in human stratum (Ekstrand et al., 2018; Shakespeare et al., 2020; Boratto et al., 2019). Datasets in which user interactions are Missing Not At Random inherently carry inherent biases with them against a particular user or item group, as pointed out by Carraro and Bridge (2022). They further propose a sampling-based approach to counter the bias. Atas et al. (2021) argue that traditional recommender systems often make the false assumption that user preferences are stable and well-defined, which can become one of the sources of biasedness and inaccuracy. In reality, user preferences are complex and can be influenced by a variety of factors, such as personality traits, emotional states, and cognitive biases. Burke (2017) presents a taxonomy of classes for fair recommendation systems. The author suggests different recommendation settings with fairness requirements such as fairness for only users, fairness for only items, and fairness for both users and items. Our work falls into the fairness for only items category where bias is shown by a particular set of users against a specific set of items in the dataset. In particular, we are interested in studying and eliminating users' biasedness against the items associated with a specific gender in recommendation systems.

Bias prevention approaches can be classified according to the phase of the data mining process in which they operate: pre-processing, in-processing, and post-processing methods. Pre-processing methods aim to control the distortion of the training set. In particular, they transform the training dataset so that the discriminatory biases contained in the dataset are smoothed, hampering the mining of unfair decision models from the transformed data. In-processing methods modify recommendation algorithms such that the resulting models do not entail unfair decisions by introducing a fairness constraint in the optimization problem. Lastly, post-processing methods act on the extracted data mining model results instead of the training data or algorithm. The method presented in our work is a hybrid of a pre-processing phase and a post-processing phase.

Two prominent studies have focused on gender bias in recommender systems. The work by Shakespeare et al. (2020) establishes the existence of bias in the results of the music recommender systems, and the work by Ekstrand et al. (2018) focuses on bias shown by Collaborative Filtering (CF) algorithms while recommending books written by women authors. Both studies establish that the CF algorithms produced biased results after being fed data containing biases from various socio-cultural factors. While both works focus just on showing the existence of bias in the presence of the users' implicit feedback, we also consider the explicit feedback ratings and the bias that may arise out of it. Another pertinent work by Zehlike et al. (2017) proposes a fair top-k ranking algorithm and defines the problem as determining a “reasonable top-k ranking” that maximizes utility while accomplishing two constraints. Firstly, each contender present in the top-k list should be more qualified than the ones not present. Secondly, for any two contenders in the top-k list, the less qualified contender should be ranked below the other one. This is a post-processing technique that removes the induced bias via a re-ranking mechanism that mainly follows three criteria - ranked group fairness, selection utility, and ordering utility. Similar to our work this algorithm reduces group bias without impacting the model’s accuracy. Also, being a post-processing technique it can be applied to all the basic recommendation models. Similarly, our model handles the case when the items associated with a specific gender might have received worse feedback than they otherwise ought to receive from a set of users. While Shakespeare et al. (2020); Ekstrand et al. (2018) limit their studies to examining bias in recommendations, we go one step further and propose a model to mitigate these biases by quantifying a particular user’s bias and debiasing his or her feedback ratings. We theoretically show that the debiased ratings are unbiased estimators of the true preference of the user. Once the ratings are debiased, they are fed into the recommender algorithms as input to produce recommendations for the desired set of users. Since the recommender system is now fed with the debiased ratings, the resulting recommendations are free from the bias factor and avoid a self-perpetuating loop in the future.

The bias of an individual user reflects his or her taste. However, the KNN-based algorithms produce recommendations based on similar characteristics between a set of users and naive implementation of these algorithms reflects the bias of one user in the recommendations produced for the other user. While not directly comparing the rating history of different users or items, Matrix Factorization algorithms rely on deriving latent factors, which depend on the rating history. Both approaches make the system increasingly biased and homogenized after users interact with their biased recommendations and generate data for the next iteration. The above discussion suggests that though it is necessary to reflect the user’s preference in the recommendations produced for him or her to achieve accuracy, it is equally necessary to prevent the bias of one user from reflecting in the recommendations of another similar user. Our research focuses on this particular objective.

Our debiased ratings assure that the biases of one user do not affect other users; however, it may lead to a loss of accuracy because of not reflecting the user’s own preferences. We introduce a new step called preference correction which injects the user’s preference parameter into his/her own debiased recommendation to maintain the accuracy of the system. The novelty of our work lies in computing the user’s preference parameter which not only helps in debiasing the ratings but also in maintaining the preferences of users. On the publicly available Book-Crossing dataset (Ziegler et al., 2005) and Amazon Book Review dataset (Ni et al., 2019), we empirically show that this approach retained the significant reduction in bias and had minimal effect on the accuracy of the system. We also show that our model performs much better than the existing baseline even after preference correction. The bias reflected in the recommendations produced by the UserKNN, ItemKNN, ALS, and SVD algorithms is reduced by as much as 42.39%, 37.65%, 26.51%, and 41.43% respectively for the Amazon dataset and by 37.82%, 30.73%, 24.99%, and 32.34% for the Book-Crossing dataset. When measured with respect to Root Mean Squared Error(RMSE), the final accuracy loss in the case of the Amazon dataset comes out to be 7.8%, 11.96%, 12.49%, and 10.38% respectively for the four algorithms. In the case of the Book-Crossing dataset, the RMSE loss comes out to be 13.86%, 18.13%, 11.41%, and 12.89% respectively. In particular, the following are our main contributions.

1.1 Contributions

- We propose a model to quantify the gender bias in the recommender system when explicit feedback is present.
- We propose a principled approach to de-bias the ratings given and theoretically show that the debiased ratings represent the unbiased estimator of the true preference of the user.
- We empirically evaluate our model on publicly available book datasets and show that the approach significantly reduced the bias in the system. To show the generality of our proposed approach, we show the results on four algorithms, UserKNN, ItemKNN, ALS, and SVD. We also provide a comparison of our model with the existing baseline.
- To further enhance the accuracy of the debiased system, we propose an approach of preference correction that respects the user's own preferences towards his/her recommendations. We show that the final recommender system significantly reduces the bias in the system while not deteriorating the accuracy much.

2 Related works

The problem of gender bias and discrimination has received lots of attention in recent works (Hajian et al., 2016). Many proposals like Pedreschi et al. (2008, 2009); Ruggieri et al. (2010); Thanh et al. (2011); Mancuhan and Mancuhan (2014); Ruggieri et al. (2014) are dedicated to detecting and measuring the existing biases in the datasets while other efforts (Kamiran et al., 2010, 2012; Hajian & Domingo-Ferrer, 2013; Hajian et al., 2014a, b; Dwork et al., 2011; Zemel et al., 2013) are focused on ensuring that data mining models do not produce discriminatory results even though the input data may be biased. Most of these works focus on the classical problem of classification. Amatriain et al. (2011) discusses the application of various classification methods like Support Vector Machines, Artificial Neural Networks, Bayesian Classifiers, and Decision Trees in recommender systems. Their findings indicated that a more complex classifier need not give a better performance for recommender systems, and more exploration is needed in this direction.

When considering "fairness for only users" according to the taxonomy presented by Burke (2017); Boratto et al. (2019) and Tsintzou et al. (2018) discuss the bias with respect to the preferential recommending of certain items only to the users of a specific gender. While methodologies studied in Boratto et al. (2019) are only appropriate for implicit feedback, the Group Utility Loss Minimization proposed in Tsintzou et al. (2018) works only with respect to the UserKNN algorithm. Both papers address the issue of gender bias by employing post-processing algorithms. Though Boratto et al. (2019) and Tsintzou et al. (2018) have addressed the issue of fairness of recommender systems with respect to gender, they have done so from the perspective of recommending certain items only to users of a specific gender. The difference between their work and our study lies in the fact that we focus on the more direct issue of gender bias in recommendations shown to items associated with a specific gender.

Shakespeare et al. (2020) in their research highlights the artist gender bias in music recommendations produced by Collaborative Filtering algorithms. The work traces the causes of disparity to variations in input gender distributions and user-item preferences, highlighting the effect such configurations can have on users' gender bias after recommendation generation. Mansoury et al. (2020) discuss the biases from the perspective of a specific group of individuals (for example, a particular gender) receiving less calibrated and hence unfair recommendations. Ekstrand et al. (2018) explores the gender bias present in the book rating dataset. Our work is different from the works by Shakespeare et al. (2020); Mansoury et al. (2020) and Ekstrand et al. (2018) in primarily two factors: (i) we consider explicit feedback as opposed to implicit feedback, and (ii) we propose a principled approach to de-bias the ratings and theoretically show that the debiased ratings are unbiased estimators of true ratings.

The research by Leavy et al. (2020) focuses on algorithmic gender bias and proposes a framework whereby language-based data may be systematically evaluated to assess levels of gender bias prevalent in training data for machine learning systems.

A couple of works in fair recommender systems focus on improving the exposure of the items belonging to minority groups. They do so by upsampling the items associated with minority groups (Boratto et al., 2021), or by adding more data points to the dataset so as to achieve overall fairness (Rastegarpanah et al., 2019). On the contrary, our goal in this paper is to provide a systematic way to reduce the bias of one user affecting the recommendations to other users.

The closest work to our model is the FA*IR proposed by Zehlike et al. (2017). Similar to our model, they propose a structured approach to mitigate group bias in item rankings. Their method aims at mitigating bias on the item side in the ranked list output of machine learning algorithms. Similarly, our work is also focused on mitigating gender bias in the ranked recommendations produced by recommender systems. FA*IR proposes fair-ranking criteria to ensure group fairness in top-k items. While ensuring visibility to protected candidates, it ensures that the contenders in the final ranking should be the most qualified ones by enforcing the selection utility. The procedure considers p as a minimum target proportion of protected elements in the ranking and the ranking is declared unfair if the proportion of items from disadvantaged group falls below this threshold. Our approach on the other hand does not require any such threshold. Furthermore, while the FA*IR methodology enforces various criteria to maintain fairness on the item side, our approach is simple and ensures fairness by feeding unbiased ratings of the users to the recommender system. This direction avoids the self-perpetuating loop in the recommender system. Once such a system is deployed, there is no further need for interference in the system to ensure fairness. We believe this is a strong first step in a new direction for a fair recommender system.

Further, with close examination, we find a paucity of works that focus on the niche area of countering the group bias present in explicit ratings given by the users. Much of the work is limited to just exploration and confirming the fact that a group bias is present in the recommendations produced in the results of recommender systems. The few works that focus on countering the bias present, are limited to implicit feedback. Moreover, no existing approaches provide a theoretical framework to mitigate the gender bias from the recommender system. FA*IR, even though it produces non-personalized rankings, does attempt to counter the group bias in the final output, and hence comes closest to our work.

3 The model

Consider a recommender system having $\mathcal{U} = \{1, 2, \ldots, U\}$ users and $\mathcal{I} = \{1, 2, \ldots, I\}$ items. Let $\mathbb{D}$ and $\mathbb{A}$ denote the set of items associated with the disadvantaged group and advantaged group, respectively. For example, in a book recommender system, the books represent the items; $\mathbb{D}$ and $\mathbb{A}$ represent the set of books written by women and men authors respectively. With respect to the book recommender system, researchers have already shown that the data is biased against female authors' books (Ekstrand et al., 2018).

Let $r_{ui} \in [1, R]$ denote the rating that user u has given to the item i. As opposed to previous works, we consider explicit feedback wherein biases may not only arise from not giving a rating to the item but may also come from giving a bad rating to the item. The user profile $p_u = \{X_u, R_u\}$ represents the set of books (X_u) and the ratings ($R_u = \{r_{ui}\}_{i \in X_u}$) that user u has given to those items.

The proposed recommender system first pre-processes the data that involves: 1) finding the log-bias θ_u of each user u and 2) generating the debiased rating d_{ui} of each user u and item i using the computed bias in the first step. We then theoretically show that the debiased ratings generated are unbiased estimators of the true preferences of the user for the items rated by them. Thus, the debiased dataset can then be fed into various recommender algorithms to generate an unbiased predicted rating of a user u for the item i, denoted by $\tilde{d}_{ui}$. This debiasing step ensures that the existing biases are not boosted further in the system. Our debiasing model is independent of any recommendation algorithm. We show the performance of our debiasing model on both K-nearest neighbours-based algorithms (UserKNN, ItemKNN) as well as matrix factorization-based algorithms (Alternating Least Square and Singular Value Decomposition) to produce the recommendations.

In the next step, we use a preference corrector to reintroduce the preferences of a particular user u to his/her own recommendations. This is achieved via producing a user-specific rating $\tilde{r}_{ui}$ from the debiased rating $\tilde{d}_{ui}$. The recommendations are re-ranked according to the adjusted ratings, and the recommendations are presented to the user. This step ensures that the system does not lose accuracy for not considering the preferences of the users. Figure 1 shows the schematic diagram of our model. Consider that the ratings r_{ui} are continuous values ranging from 1 to R, then mathematically, a biased recommender system can be represented as follows:

1. Each user u, while rating an item i, scales down the maximum rating R by $e^{p_{ui}}$. p_{ui} is a random variable, drawn from a distribution function $P_u(I)$, which has a mean value of α_u. p_{ui} represents the logarithm of the true preference of the user u for the item i. For the sake of brevity, we call it the log-preference of the user u for the item i. Hence $e^{p_{ui}}$ is a representation of the true preference of user u for the item i.
2. In case the item is associated with the disadvantaged group, the user u further scales down the rating of the item by a factor $e^{q_{ui}}$. q_{ui} is a random variable, drawn from a distribution function $Q_u(I)$ having a mean value of β_u. q_{ui} represents the logarithm of the biasedness of the user u shown to the item i. For the sake of brevity, we call it the log-bias of the user u for the book i. Hence $e^{q_{ui}}$ represents the biasedness of the user u for the book i.
3. For each user u, β_u is sampled from the a distribution function $\Omega(x)$ which governs the global log-bias tendency of the users. We denote the mean value of $\Omega(x)$ by γ.

Thus, ratings r_{ui} can be expressed as:

$$r_{ui} = \begin{cases} R/e^{p_{ui}}, & \textit{ifiis associated with advantaged group} \\ R/e^{p_{ui}}e^{q_{ui}}, & \textit{ifiis associated with disadvantaged group} \end{cases} \tag{1}$$

We now present a detailed description of each of the steps.s

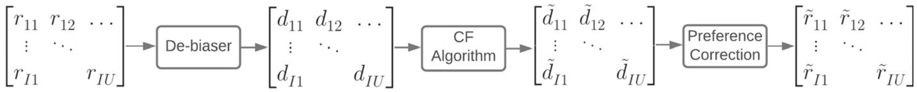

r_{11} r_{12} $\cdots$
r_{I1} r_{IU}
De-biaser
d_{11} d_{12} $\cdots$
d_{I1} d_{IU}
CF Algorithm
$\tilde{d}_{11}$ $\tilde{d}_{12}$ $\cdots$
$\tilde{d}_{I1}$ $\tilde{d}_{IU}$
Preference Correction
$\tilde{r}_{11}$ $\tilde{r}_{12}$ $\cdots$
$\tilde{r}_{I1}$ $\tilde{r}_{IU}$

Fig. 1 Model schematics

3.1 Estimating the mean value for log-bias

The geometric mean of the ratings given by a user u to the items associated with disadvantaged and advantaged groups, denoted by r_{ud} and r_{ua} respectively, are given by the following expressions:

$$r_{ud} = \left(\prod_{i \in \mathbb{D} \cap X_u} r_{ui} \right)^{1/|\mathbb{D} \cap X_u|} \quad \text{and} \quad r_{ua} = \left(\prod_{i \in \mathbb{A} \cap X_u} r_{ui} \right)^{1/|\mathbb{A} \cap X_u|}$$

Further, the log-bias in the user profile p_u is given by $\theta_u = \ln\left(\frac{r_{ua}}{r_{ud}}\right)$.

We use geometric mean to compute the average rating of a user due to the following reasons: 1) It is less biased towards very high scores as compared to the arithmetic mean (Neve & Palomares, 2019) and 2) when cold users are involved, aggregating recommendations using the geometric mean is more robust as compared to the arithmetic mean (Valcarce et al., 2020).

The below lemma shows that θ_u is an unbiased estimator of β_u.

Lemma 1 *The expectation of log-bias, θ_u in the user profile p_u represents the mean value of the log-bias, β_u of the user u.*

Proof Let us denote $m = |\mathbb{D} \cap X_u|$ and $n = |\mathbb{A} \cap X_u|$ to be the number of items associated with disadvantaged and advantaged group respectively in user profile p_u. Then,

$$\begin{aligned} \theta_u &= \ln\left(\frac{r_{ua}}{r_{ud}}\right) = \ln\left[\frac{\left(\prod_{y=1}^{m} e^{p_{uy}} e^{q_{uy}}\right)^{\frac{1}{m}}}{\left(\prod_{x=1}^{n} e^{p_{ux}}\right)^{\frac{1}{n}}}\right] (Using 1) \\ &= \frac{1}{m}\sum_{y=1}^{m} q_{uy} + \frac{1}{m}\sum_{y=1}^{m} p_{uy} - \frac{1}{n}\sum_{x=1}^{n} p_{ux} \end{aligned}$$

Taking expectation on both sides:

$$\mathbb{E}[\theta_u] = \mathbb{E}\left[\frac{1}{m}\sum_{y=1}^{m} q_{uy} + \frac{1}{m}\sum_{y=1}^{m} p_{uy} - \frac{1}{n}\sum_{x=1}^{n} p_{ux}\right] \tag{2}$$

Using linearity of expectation and some simplification, we get:

$$\begin{aligned} \mathbb{E}[\theta_u] &= \frac{1}{m}\sum_{y=1}^{m} \mathbb{E}[q_{uy}] + \frac{1}{m}\sum_{y=1}^{m} \mathbb{E}[p_{uy}] - \frac{1}{n}\sum_{x=1}^{n} \mathbb{E}[p_{ux}] \\ &= \frac{1}{m}\sum_{y=1}^{m} \beta_u + \frac{1}{m}\sum_{y=1}^{m} \alpha_u - \frac{1}{n}\sum_{x=1}^{n} \alpha_u \end{aligned}$$

Thus, $\mathbb{E}[\theta_u] = \beta_u$. □

Once we get the log-biasedness tendencies of users, we use them to produce the debiased ratings for the given dataset.

3.2 Debiasing the dataset

The debiased rating of the item i associated with disadvantaged group and rated by user u is given as $d_{ui} = r_{ui}e^{\theta_u}$ We now provide the main theorem of our paper.

Theorem 2 $\ln(d_{ui})$ *is the unbiased estimator of the log of the true rating of the item* i.

Proof $\ln(d_{ui}) = \theta_u + \ln(r_{ui}) = \theta_u + \ln R - p_{ui} - q_{ui}$. The last equality is obtained from (1). Taking expectation on both sides:

$$\begin{aligned}\mathbb{E}(\ln(d_{ui})) &= \mathbb{E}[\theta_u] + \mathbb{E}[\ln R] - \mathbb{E}[q_{ui}] - \mathbb{E}[p_{ui}] \\ &= \beta_u + \ln R - \beta_u - \alpha_u (Using Lemma 1) \\ &= \ln R - \alpha_u = \ln\left(\frac{R}{e^{\alpha_u}}\right)\end{aligned}$$

As we can see, the expected value of $\ln(d_{ui})$ contains only the term representing the true preference of the item for user u. □

Thus, instead of r_{ui}, ratings d_{ui} are fed into the recommender system to generate the predicted unbiased ratings $\tilde{d}_{ui}$. Simply removing the bias from the user's rating could severely affect the system's accuracy because the bias of an individual user reflects their taste. However, the debiasing step helps prevent the bias of one user from affecting the recommendation of other users. Next, we use preference corrections by correcting the predicted rating of the user with respect to his/her own preference parameter.

3.3 Preference correction to improve accuracy

Note that when the users are inherently biased against a group of items, $\mathcal{D}$ then showing the items from $\mathcal{D}$ naively to these users will severely affect the accuracy of the system. The goal of this work is not just to promote the exposure of the items among the two groups but is to not let the bias of one user creep into the bias of the other user. This was achieved via debiasing the dataset. Once the debiased ratings are generated, the accuracy of the system is maintained by introducing a correction factor. Although providing us with higher accuracy, the idea to re-introduce the correction factor may lead to an overall increase in individual biases. This on a prima facie may look self-defeating, but we need to note that final ratings still have significantly less bias than original ratings. If we do not introduce the correction factor, the users might flock to a substantial bias platform due to poor accuracy.

The correction is achieved by multiplying the predicted ratings of items associated with the disadvantaged group by a factor $e^{-\theta_u}$. Thus, the final recommended ratings will be given as $\tilde{r}_{ui} = \tilde{d}_{ui}e^{-\theta_u}$. Similar to the calculation of bias in the dataset, we can now compute the bias in the recommendation profile.

3.4 Bias in recommendation profile

We generate recommendations for the users in the test set $\mathcal{T}$. The recommendation profile for a user $u \in \mathcal{T}$ is denoted by $\tilde{p}_u = \{\tilde{X}_u, \tilde{R}_u\}$, which represents the set of recommended books $(\tilde{X}_u)$ for the user u and their predicted ratings $(\tilde{R}_u = \{\tilde{r}_{ui}\}_{i\in\tilde{X}_u})$. Let the set of items associated with disadvantaged and advantaged groups be denoted by $\tilde{\mathbb{D}}$ and $\tilde{\mathbb{A}}$ respectively. The average predicted ratings of the items associated with disadvantaged and advantaged groups, denoted by $\tilde{r}_{ud}$ and $\tilde{r}_{ua}$ respectively, are given by:

$\tilde{r}_{ud} = \left(\prod_{i \in \tilde{\mathbb{D}} \cap \tilde{X}_u} \tilde{r}_{ui}\right)^{1/|\tilde{\mathbb{D}} \cap \tilde{X}_u|}$ and $\tilde{r}_{ua} = \left(\prod_{i \in \tilde{\mathbb{A}} \cap \tilde{X}_u} \tilde{r}_{ui}\right)^{1/|\tilde{\mathbb{A}} \cap \tilde{X}_u|}$ where $\tilde{r}_{ui}$ is the predicted rating given to item i in the recommendation-profile generated for a user u. The log-bias in the recommendation-profile p_u, denoted by $\tilde{\theta}_u$, is then given by $\tilde{\theta}_u = \ln\left(\frac{\tilde{r}_{ua}}{\tilde{r}_{ud}}\right)$. For an unbiased recommendation-profile, $\tilde{\theta}_u = 0$. A profile biased against disadvantaged groups will have $\tilde{\theta}_u > 0$. We can then compute the overall bias of the recommender system by taking the average overall users, and this average gives us the estimated value of γ.

4 Dataset

To evaluate the proposed model, we run experiments on two publicly available book rating datasets, the Book-Crossing dataset (Ziegler et al., 2005), Amazon Book Review dataset (Ni et al., 2019) and use FA*IR (Zehlike et al., 2017) as a potential baseline to depict that our model yields the best results in terms of fairness. We further process this dataset through the following stages:

4.1 Book author identification

Their unique ISBNs identify the books in both datasets. We identified the authors of the books present in the datasets via their ISBNs using the following three API services: Google Books API (2021), ISBNdb (2021), and OpenLibrary API (2021). We could not identify the authors of some of the books. Hence we discarded those books from the dataset.

4.2 Author gender identification

We identified the genders of the authors via their first names. We used Genderize.io (2021), an API service dedicated to identifying the gender given the first name of the person. We used a minimum confidence threshold of 90% for gender identification. For books having multiple authors, we considered the name of the first author. We could not identify the gender of some of the authors. We discard the books written by those authors from the dataset.

4.3 Filtering

The Book-Crossing dataset, in addition to explicit ratings given by users to books, also contains null values(i.e. 0) depicting the instance when the user did not rate the book, giving a form of implicit feedback. Since our focus in this work is on explicit feedback, we discard such instances. We further filtered the Book-Crossing dataset to include only those books with at least 50 ratings and only those users who have rated at least 50 books. Amazon dataset was significantly larger as compared to the Book-Crossing dataset. We filtered it to include only those books with at least 100 ratings and only those users who have rated at least 100 books. We did this filtering so that recommender algorithms have much data to produce accurate recommendations. The statistics of filtered datasets are mentioned in Table 1. The number of books written by male authors is almost equal to that of female authors for both the datasets.

Table 1 Dataset details

Statistic	Amazon	Book-Crossing
Number of male-authored books	58369	829
Number of female-authored books	58220	806
Number of users	44792	376

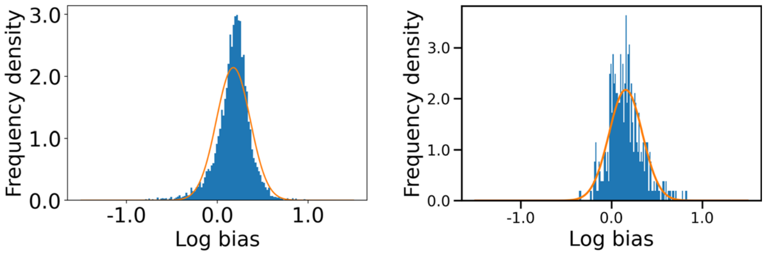
5 Experimental results

5.1 Input Bias

We show the distributions of log-bias tendency (θ_u) of the users in the Amazon dataset and the Book-Crossing dataset in Fig. 2. We observe that the mean log-bias tendency over all the users in the Amazon dataset is higher (0.176) than that of the Book-Crossing dataset (0.157)[1].

5.2 Output bias

We randomly separate 20% of users in each dataset as the test group. We generate the recommendations for the users in the test group using two K-nearest neighbours-based algorithms, UserKNN and ItemKNN, and two matrix factorization-based algorithms, Alternating Least Square and Singular Value Decomposition. These algorithms were selected because the accuracy and ranking relevance of the recommendations produced by them were among the highest values compared with other algorithms. Hence coupling our model with them would best highlight the effects brought about by the same. We calculate the estimated value of log-bias ($\tilde{\theta}_u$) accuracy, and ranking relevance in the recommendations separately for each algorithm applied to the two datasets. For accuracy, we use two error measures, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). For ranking relevance, we use two parameters, Normalized Discounted Cumulative Gains(NDCG) and Mean Reciprocal Rank(MRR).

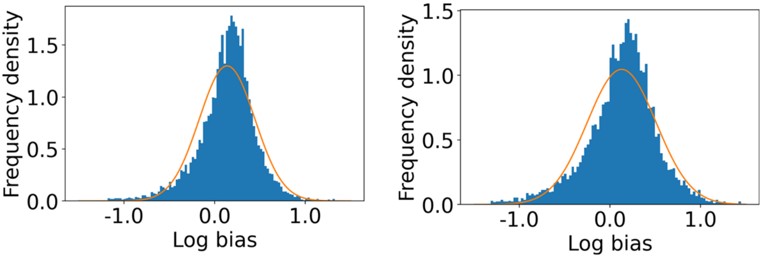


(a) Amazon dataset

(b) Book-Crossing dataset

Fig. 2 User log-bias in the original dataset

[1] code is available at https://github.com/Pyromancer11/mitigatingGenderBias



(a) UserKNN

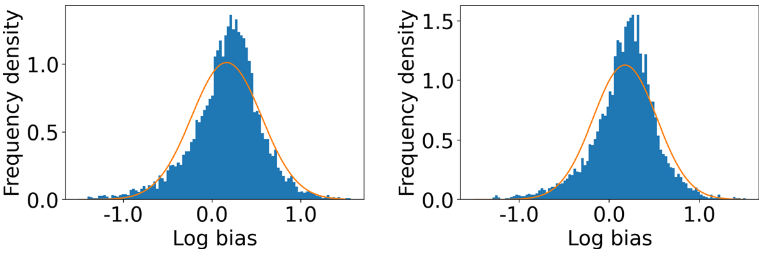
(b) ItemKNN

Fig. 3 Output log-bias in AZ dataset without employing the model under K-nearest neighbour family of algorithms

We first begin plotting the log-bias ($\tilde{\theta_u}$) distribution for the recommendations produced by the algorithms without employing our debiased model in Figs. 3 and 4 for Amazon datasets with respect to K-nearest neighbour family and matrix factorization family of algorithms. Figures. 5 and 6 similarly present the log-bias distribution for the recommendations produced by the two families of algorithms for Book-Crossing datasets respectively without employing our debiased model. We compute the log-bias by feeding biased ratings r_{ui} to the four algorithms. As can be seen from the figures, the output log biasedness was very similar to what was observed in the input data.

We next deploy our model partially. We leave out the preference correction phase and produce the recommendations using the algorithms mentioned before by feeding the debiased ratings d_{ui} to these algorithms. We estimate the mean log-bias tendency in the recommendations $\tilde{\theta}_u$ using debiased ratings produced by the algorithms $\tilde{d}_{ui}$. The log-bias ($\tilde{\theta_u}$) distribution for the recommendations produced by the algorithms after partial deployment of the model is depicted in the Figs. 7 and 8 for Amazon dataset and in the Figs. 9 and 10 for book crossing dataset. As can be seen, there is a significant reduction in log-bias tendency (64.38%) in the Amazon dataset and (53.67%) in the Book-Crossing dataset for the UserKNN algorithm. However, we also see an increase in error rates on both datasets. This is because the test data itself contains biases.

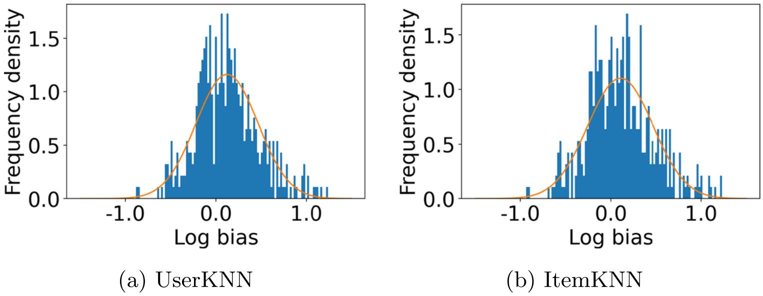
Finally, we deploy our complete model after adding the preference correction method and repeat the experiment. The log-bias ($\tilde{\theta_u}$) distribution for the recommendations produced by the algorithms after deployment of the complete model is depicted in Figs. 11, 12 for



(a) ALS

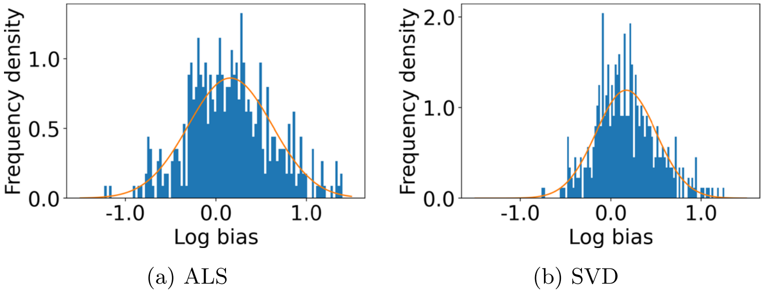
(b) SVD

Fig. 4 Output log-bias in AZ dataset without employing the model under matrix factorization family of algorithms



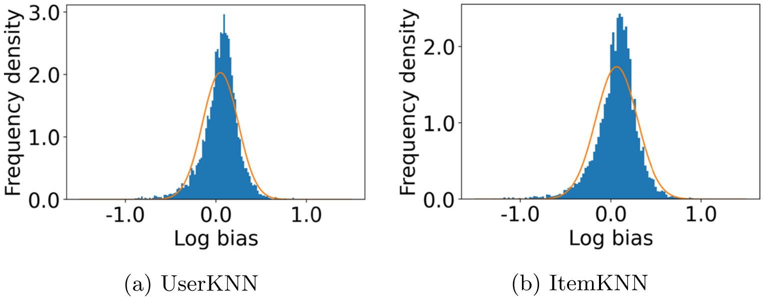
(a) UserKNN (b) ItemKNN

Fig. 5 Output log-bias in BX dataset without employing the model under K-nearest neighbour family of algorithms



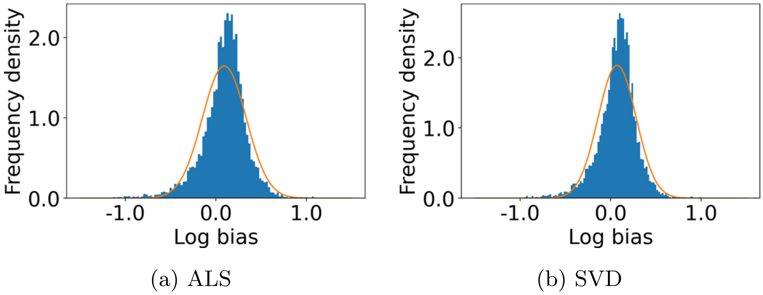
(a) ALS (b) SVD

Fig. 6 Output log-bias in BX dataset without employing the model under matrix factorization family of algorithms



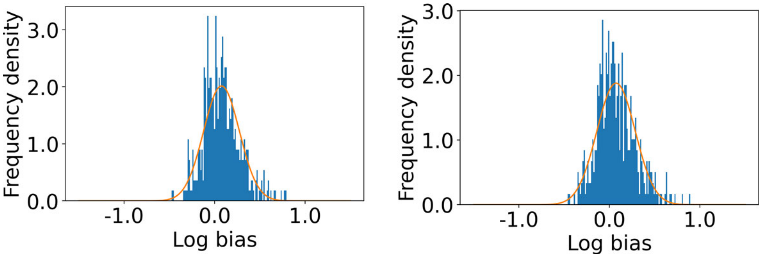
(a) UserKNN (b) ItemKNN

Fig. 7 Output log-bias in AZ dataset with debiasing under the family of K-nearest neighbour algorithms



(a) ALS (b) SVD

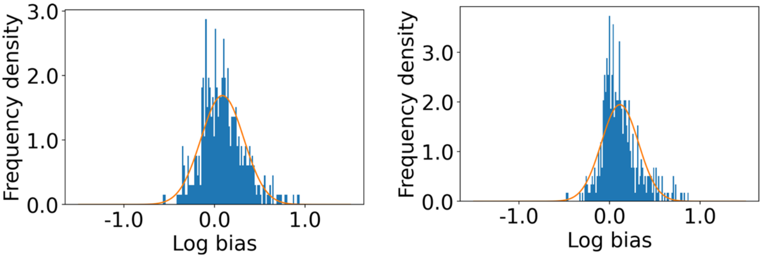
Fig. 8 Output log-bias in AZ dataset with debiasing under the family of matrix factorization algorithms



(a) UserKNN

(b) ItemKNN

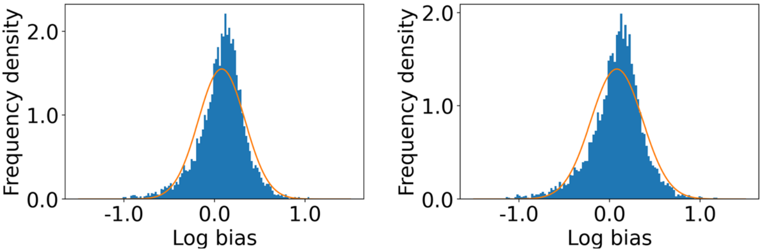
Fig. 9 Output log-bias in BX dataset with debiasing under family of K-nearest neighbour algorithms



(a) ALS

(b) SVD

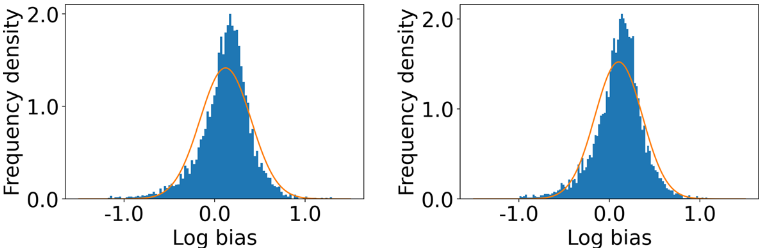
Fig. 10 Output log-bias in BX dataset with debiasing under the family of matrix factorization algorithms



(a) UserKNN

(b) ItemKNN

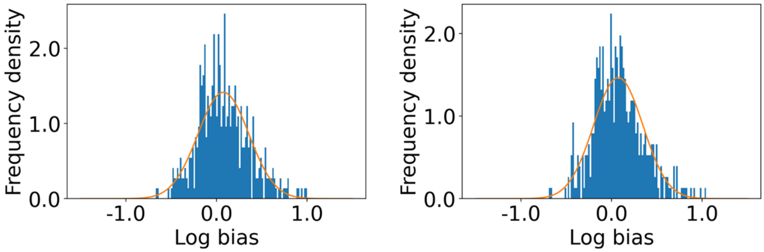
Fig. 11 Output log-bias in AZ dataset with preference correction under family of K-nearest neighbour algorithms



(a) ALS

(b) SVD

Fig. 12 Output log-bias in AZ dataset with preference correction under family of matrix factorization algorithms



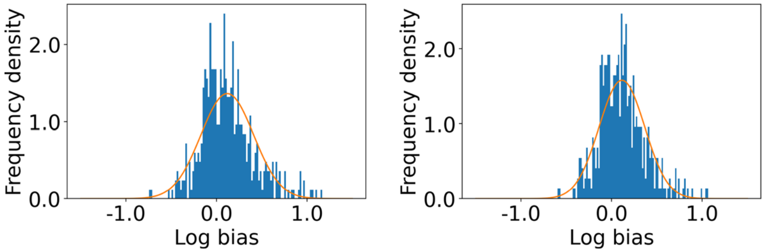
(a) UserKNN (b) ItemKNN

Fig. 13 Output log-bias in BX dataset with preference correction under family of K-nearest neighbour algorithms

Amazon dataset and in Figs. 13, 14 for book crossing dataset. The final values for all the cases are given in Table 2 for the Amzaon dataset and in Table 3 for book crossing datasets. As can be seen, there is still a significant reduction in mean log-bias tendency, which reduces by 42.39% in the Amazon dataset and by 37.82% in the case of the Book-Crossing dataset for the UserKNN algorithm. Figure 15 presents the percentage gain in bias reduction for both datasets. The percentage loss in accuracy is depicted in Figs. 16 and 17 for Amazon and Book-Crossing datasets respectively. The percentage loss in ranking relevance metrics is depicted in Figs. 18 and 19 respectively. Our model outperforms the existing baseline in terms of mean-log bias as well as ranking relevance. The accuracy loss, however, is insignificant, making this trade-off advantageous.

We next conduct significance testing to validate the log-bias reduction. Tables 4 and 5 show the p-values obtained from left-tail significance tests on the log-bias of the recommendations made for the users in the sample. We can see from the p-value for the Amazon datasets that the bias reduction is significant. For the Book-Crossing dataset, the significance of the bias reduction is less pronounced. One of the prominent reasons for this is that the test sample size for the Book-Crossing dataset was relatively small due to the small number of users in the dataset. In essence, the utility of the recommender system is maintained while reducing the log-bias tendency in the recommendations.

We further observe that the bias reduction is more in the case of UserKNN-based recommendations than the ItemKNN-based recommendations. This observation can be attributed



(a) ALS (b) SVD

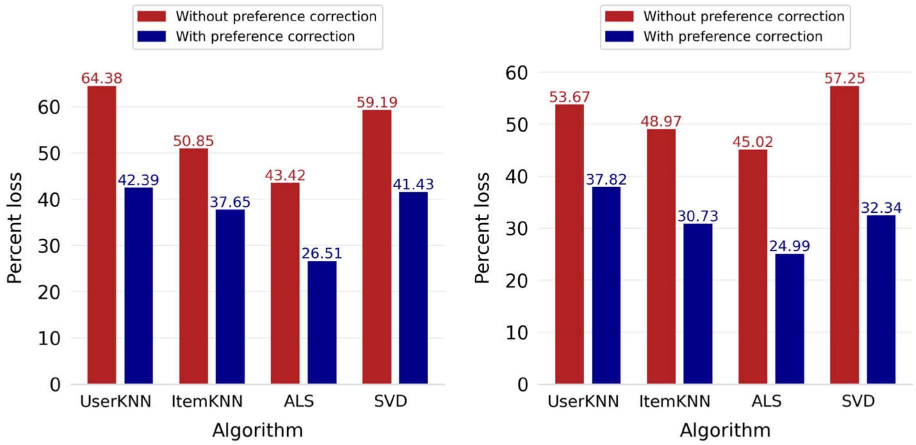
Fig. 14 Output log-bias in BX dataset with preference correction under family of matrix factorization algorithms

Table 2 Summary of Results for the Amazon dataset

Case	Algorithm	Mean log-bias	RMSE	MAE	NDCG	MRR
without model	UserKNN	0.137	0.808	0.693	0.452	0.498
	ItemKNN	0.129	0.736	0.580	0.597	0.643
	ALS	0.164	0.873	0.829	0.281	0.447
	SVD	0.175	0.790	0.753	0.342	0.471
without preference correction phase	UserKNN	0.049	1.103	0.921	0.0224	0.0278
	ItemKNN	0.063	1.076	0.873	0.0229	0.0204
	ALS	0.093	1.281	1.1211	0.0161	0.0394
	SVD	0.071	1.257	1.183	0.0138	0.0206
with preference correction phase	UserKNN	0.079	0.871	0.738	0.3982	0.4462
	ItemKNN	0.080	0.824	0.661	0.5236	0.6121
	ALS	0.121	0.982	0.903	0.2391	0.3853
	SVD	0.103	0.872	0.847	0.2989	0.4159
FA*IR	UserKNN	0.128	0.826	0.714	0.434	0.4604
	ItemKNN	0.116	0.753	0.636	0.583	0.6256
	ALS	0.153	0.915	0.854	0.267	0.4198
	SVD	0.169	0.838	0.782	0.325	0.4545

Table 3 Summary of Results for the Book-Crossing dataset

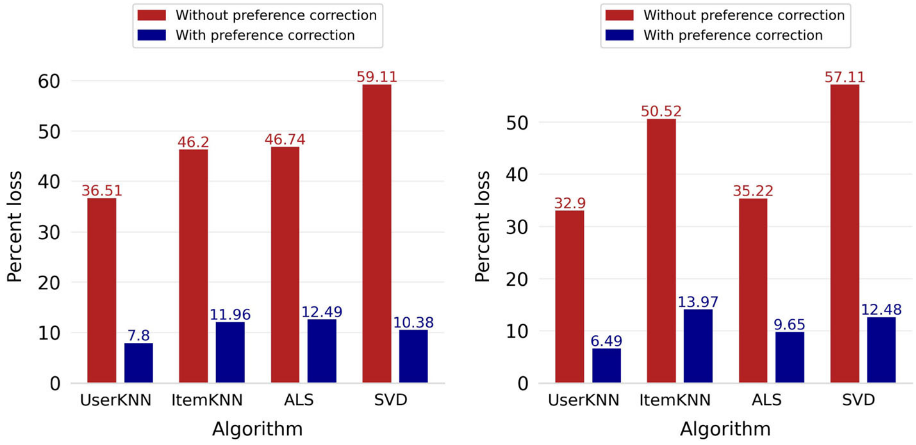
Case	Algorithm	Mean log-bias	RMSE	MAE	NDCG	MRR
without model	UserKNN	0.122	1.580	1.178	0.264	0.272
	ItemKNN	0.106	1.511	1.304	0.313	0.412
	ALS	0.158	1.815	1.642	0.235	0.370
	SVD	0.169	1.761	1.626	0.277	0.296
without preference correction phase	UserKNN	0.057	2.468	1.754	0.0232	0.0245
	ItemKNN	0.054	2.463	2.055	0.0142	0.0271
	ALS	0.087	2.752	2.175	0.0421	0.0736
	SVD	0.072	2.601	1.979	0.0261	0.0240
with preference correction phase	UserKNN	0.076	1.799	1.298	0.2099	0.2317
	ItemKNN	0.073	1.785	1.516	0.2358	0.3560
	ALS	0.119	2.022	1.768	0.1626	0.2831
	SVD	0.114	1.988	1.731	0.2258	0.2481
FA*IR	UserKNN	0.104	1.645	1.215	0.233	0.2549
	ItemKNN	0.087	1.532	1.445	0.285	0.3804
	ALS	0.139	1.934	1.698	0.209	0.3554
	SVD	0.155	1.820	1.660	0.252	0.2729



(a) AZ dataset (b) BX dataset

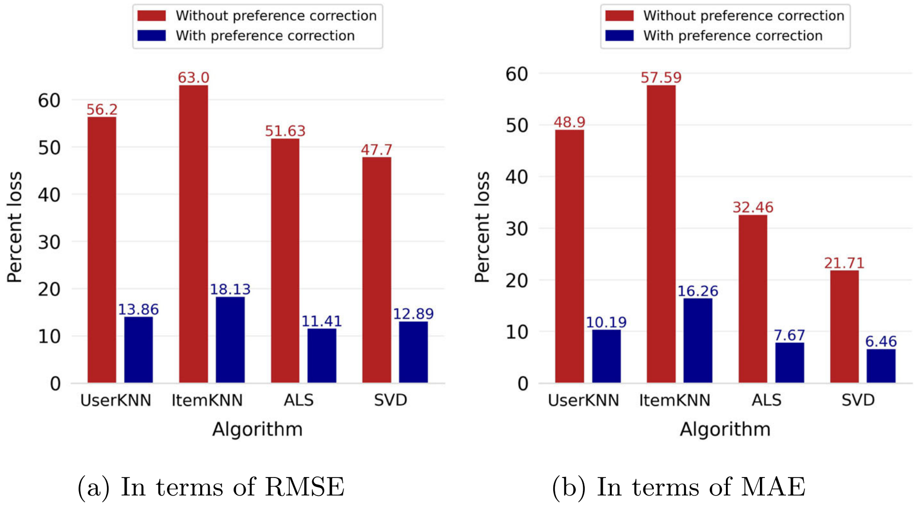
Fig. 15 Bias reduction

to the fact that our model addresses the bias originating from the distortion in ratings from the users' side. It compares the ratings of an item given by a particular user with the appropriately scaled average of ratings given by other users to that item in the dataset. It, therefore, resonates with the UserKNN algorithm, which predicts the ratings of an item for a particular user based on the ratings of that item for his or her peers. The ItemKNN algorithm, on the other hand, predicts the ratings of an item for a particular user based on the ratings given to similar items by that user. The model does not sit squarely with ItemKNN. Thus the bias reduction in UserKNN is more as compared to that in the case of ItemKNN. We



(a) In terms of RMSE (b) In terms of MAE

Fig. 16 Accuracy loss for AZ dataset

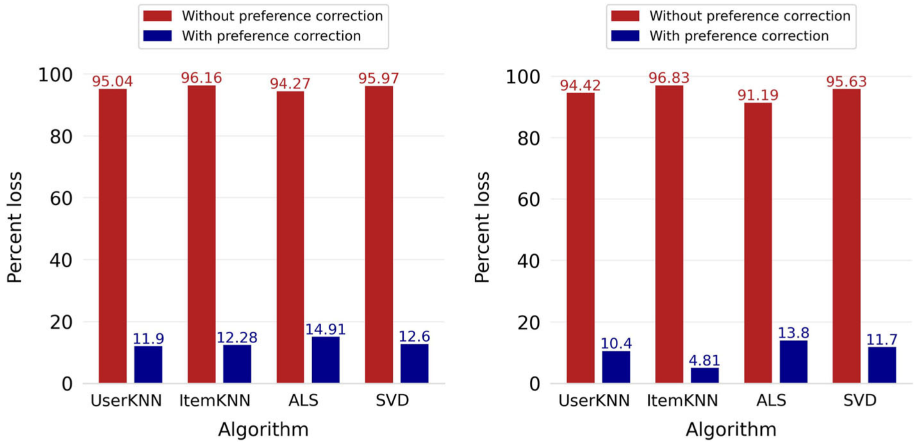


(a) In terms of RMSE (b) In terms of MAE

Fig. 17 Accuracy loss for BX dataset

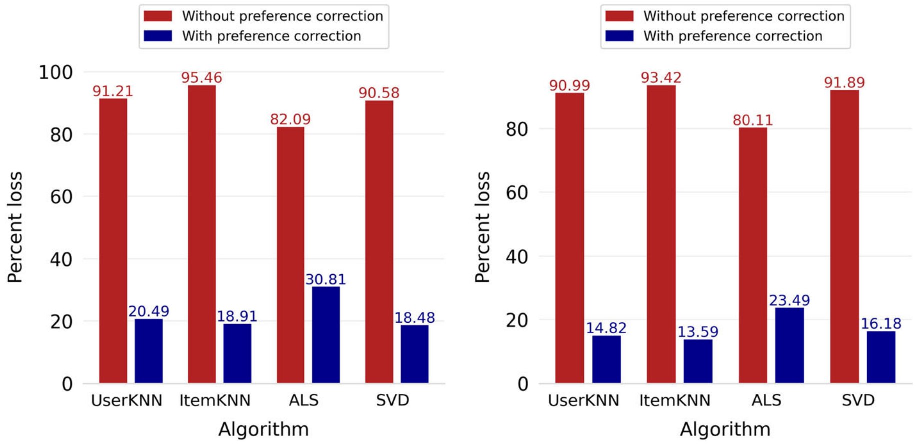
further observe that the bias reduction is more in the case of the AZ dataset as compared to the BX dataset. This observation can be attributed to the AZ dataset having a higher input mean log-bias tendency. Further, the AZ dataset has a significantly larger number of users and items which leads to a more accurate estimation of user bias scores and, therefore, more effective bias mitigation.

We observe that accuracy and ranking relevance loss is, in general, higher for ItemKNN as compared to UserKNN. This is because the model quantifies the bias of users by comparing the ratings given by them to particular items with a scaled average of ratings given by their



(a) In terms of NDCG (b) In terms of Reciprocal Rank

Fig. 18 Ranking relevance loss for AZ dataset



(a) In terms of NDCG

(b) In terms of Reciprocal Rank

Fig. 19 Ranking relevance loss for BX dataset

peers to those items. This resonates with the UserKNN algorithm, which predicts user ratings for particular items based on the ratings of similar users. Thus the model is better oriented towards the UserKNN algorithm, giving better accuracy and bias reduction in its case. In the case of matrix factorization algorithms, the accuracy and ranking relevance losses are relatively comparable. It is not clear which one of the two algorithms is more coherent with the model.

We further observe that the accuracy loss on the BX dataset is higher than that of the AZ dataset. This observation can be attributed to the fact that the user and item base of the AZ dataset is higher as compared to the BX dataset. Thus, the bias score estimates are more accurate, which provides more accurate predictions of the item scores for the users when reinserted into the recommendations.

6 Conclusion and future work

We proposed a model to quantify and mitigate the bias in the explicit feedback given by the users to different items. We theoretically showed that the debiased ratings produced by our model are unbiased estimators of the true preference of the users for the books. With the help

Table 4 Significance test results for bias reduction for Amazon Dataset

Algorithm	$\bar{x}$	μ	σ	z	p
UserKNN	0.079	0.137	0.307	-17.90	$< 10^{-5}$
ItemKNN	0.080	0.129	0.381	-12.06	$< 10^{-5}$
ALS	0.121	0.164	0.394	-10.46	$< 10^{-5}$
SVD	0.103	0.175	0.354	-19.27	$< 10^{-5}$

Table 5 Significance test results for bias reduction for Book-Crossing Dataset

Algorithm	$\bar{x}$	μ	σ	z	p
UserKNN	0.076	0.122	0.343	-1.164	0.122
ItemKNN	0.073	0.106	0.362	-0.780	0.218
ALS	0.119	0.158	0.464	-0.738	0.230
SVD	0.114	0.169	0.335	-1.413	0.079

of comprehensive experiments on two publicly available book datasets, we show a significant reduction in the bias (almost 40%) with just a 10% decrease in accuracy using the UserKNN algorithm. Similar trends were observed for other algorithms such as ItemKNN, ALS, and SVD. Our model is independent of these algorithms' choices and can be applied to any recommendation algorithm. We used the book recommender system because we were able to generate the gender information from publicly available APIs. Our model is not restricted to book recommender systems as long as protected attribute information about the items is known. We leave the extension of the model to missing protected attributes as interesting future work. It will be interesting direction to see if the ideas from fair classification literature with missing protected attributes (Coston et al., 2019) can be leveraged. We did not address the bias originating from fewer ratings for a female-authored book than a male-authored one. We leave extending the model to the bias originating from a lesser number of ratings and extensively studying the model for other recommender systems as the future directions.

Acknowledgements We would like to express our sincere gratitude to *Indian Institute of Technology, Ropar* for providing us with the opportunity to work on this project.
We are also thankful for the API services Google Books API, ISBNdb, OpenLibrary API and Genderize.io for helping us identify the authors of the books and their genders.
Our research work would not have been possible without the support of these organizations, and we truly appreciate their contributions.

Author Contributions The contributions of the authors are as follows:
- Shrikant Saxena and Shweta Jain conceived of the presented idea.
- Shrikant Saxena developed the theory and performed the computations.
- Shweta Jain wrote the manuscript.

Funding We, the authors of this research paper entitled *Exploring and Mitigating Gender Bias in Recommender Systems with Explicit Feedback*, would like to acknowledge the funding given by the Department of Science & Technology, India under grant number SRG/2020/001138 (Recipient name- Dr. Shweta Jain). The computational resources obtained from the funding were used to carry out the experiments for the project.

Availability of Supporting Data Two book-ratings datasets were used to assess the suggested model: the Book-Crossing dataset, which was initially compiled by Ziegler et al. (2005), and the Amazon Book Review dataset, which was compiled by Ni et al. (2019). Both datasets are available publicly.

Declarations

Competing interests We, the authors of this research paper entitled *Exploring and Mitigating Gender Bias in Recommender Systems with Explicit Feedback*, declare that we have no competing interests that may influence the interpretation or presentation of this manuscript.
We have no financial, personal or professional relationships with other people or organizations that could be considered potential sources of bias. Furthermore, we have no financial or personal relationships with any company or organization that could benefit from the publication of this research paper.

References

Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. (2011). Data mining methods for recommender systems. *Recommender Systems Handbook,* 39–71. https://doi.org/10.1007/978-0-387-85820-3_2

Atas, M., Felfernig, A., Polat-Erdeniz, S., Popescu, A., Tran, T. N. T., & Uta, M. (2021). Towards psychology-aware preference construction in recommender systems: overview and research issues. *Journal of Intelligent Information Systems, 57*(3), 467–489. https://doi.org/10.1007/s10844-021-00674-5

Boratto, L., Fenu, G., & Marras, M. (2019). The effect of algorithmic bias on recommender systems for massive open online courses. *Advances in Information Retrieval,* 457–472. https://doi.org/10.1007/978-3-030-15712-8_30

Boratto, L., Fenu, G., & Marras, M. (2021). Interplay between upsampling and regularization for provider fairness in recommender systems. *User Modeling and User-Adapted Interaction, 31*(3), 421–455. https://doi.org/10.1007/s11257-021-09294-8

Burke, R. (2017). Multisided Fairness for Recommendation. arXiv:1707.00093

Carraro, D., & Bridge, D. (2022). A sampling approach to debiasing the offline evaluation of recommender systems. *Journal of Intelligent Information Systems, 58*(2), 311–336. https://doi.org/10.1007/s10844-021-00651-y

Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., & Chakraborty, S. (2019). Fair transfer learning with missing protected attributes. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 91–98. https://doi.org/10.1145/3306618.3314236

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2011). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. arXiv:1104.3913, https://doi.org/10.1145/2090236.2090255

Ekstrand, M., Tian, M., Kazi, M., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. *User modeling and User-adapted Interaction,* 377–420. https://doi.org/10.1145/3240323.3240373

Genderize.io (2021). https://genderize.io/. Accessed 5 March 2021

Google Books API (2022). https://developers.google.com/books. Accessed 24 Feb 2021

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2012.72

Hajian, S., Domingo-Ferrer, J., & Farrás, O. (2014). Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-014-0346-1

Hajian, S., Domingo-Ferrer, J., & Farrás, O. (2014). Discrimination- and privacy-aware patterns. *Data Mining and Knowledge Discovery,* 29. https://doi.org/10.1007/s10618-014-0393-7

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: from discrimination discovery to fairness-aware data mining. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2125–2126. https://doi.org/10.1145/2939672.2945386

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53. https://doi.org/10.1145/963770.963772

Hurley, N., & Zhang, M. (2011). Novelty and diversity in top-n recommendation – analysis and evaluation. ACM Transactions on Internet Technology 10(4). https://doi.org/10.1145/1944339.1944341

ISBNdb (2021). https://isbndb.com/isbn-database. Accessed 27 Feb 2021

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *IEEE International Conference on Data Mining,* 869–874. https://doi.org/10.1109/ICDM.2010.50

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *IEEE International Conference on Data Mining, ICDM,* 924–929. https://doi.org/10.1109/ICDM.2012.45

Knijnenburg, B., Willemsen, M., Gantner, S., & et al. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction, 22*, 441–504. https://doi.org/10.1007/s11257-011-9118-4

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender bias in machine learning data sets. *Bias and Social Aspects in Search and Recommendation,* 12–26. https://doi.org/10.1007/978-3-030-52485-2_2

Mancuhan, K.,& Mancuhan, C. (2014). Combating discrimination using Bayesian networks. *Artificial Intelligence and Law,* 22. https://doi.org/10.1007/s10506-014-9156-4

Mansoury, M., Abdollahpouri, H., Smith, J., & et al. (2020). Investigating potential factors associated with gender discrimination in collaborative recommender systems. Proceedings of the 33rd International

Florida Artificial Intelligence Research Society Conference, FLAIRS 2020, 193–196. https://aaai.org/papers/193-flairs-2020-18430/

Neve, J., & Palomares, I. (2019). Latent Factor Models and Aggregation Operators for Collaborative Filtering in Reciprocal Recommender Systems. Proceedings of the 13th ACM Conference on Recommender Systems, pp. 219–227. https://doi.org/10.1145/3298689.3347026

Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 188–197. https://doi.org/10.18653/v1/D19-1018

OpenLibrary API (2021). https://openlibrary.org/developers/api. Accessed 02 March 2021

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568. https://doi.org/10.1145/1401890.1401959

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring Discrimination in Socially-Sensitive Decision Records. Proceedings of the 2009 SIAM International Conference on Data Mining (SDM), pp. 581–592. https://doi.org/10.1137/1.9781611972795.50

Rastegarpanah, B., Gummadi, K. P., & Crovella, M. (2019). Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 231–239. https://doi.org/10.1145/3289600.3291002

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data Mining for Discrimination Discovery. ACM Transactions on Knowledge Discovery from Data, 4(2). https://doi.org/10.1145/1754428.1754432

Ruggieri, S., Hajian, S., Kamiran, F., & Zhang, X. (2014). Anti-discrimination analysis using privacy attack strategies. *Machine Learning and Knowledge Discovery in Databases,* 694–710. https://doi.org/10.1007/978-3-662-44851-9_44

Shakespeare, D., Porcaro, L., Gómez, E., & Castillo, C. (2020). Exploring artist gender bias in music recommendation. https://doi.org/10.48550/arXiv.2009.01715

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender Systems Handbook,* 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

Thanh, B., Ruggieri, S., & Turini, F. (2011). k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 502–510. https://doi.org/10.1145/2020408.2020488

Tsintzou, V., Pitoura, E., & Tsaparas, P. (2018). Bias Disparity in Recommendation Systems. https://doi.org/10.48550/arXiv.1811.01461

Valcarce, D., Bellogín, A., Parapar, J., & Castells, P. (2020). Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal, 23*, 411–448. https://doi.org/10.1007/s10791-020-09377-x

Zehlike, M., Bonchi, F., Castillo, C., & et al. (2017). FA*IR: A Fair Top-k Ranking Algorithm. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1569–1578. https://doi.org/10.1145/3132847.3132938

Zemel, R., Wu, Y., Swersky, K., & et al. (2013). Learning Fair Representations. Proceedings of the 30th International Conference on Machine Learning 28(3), 325–333. https://proceedings.mlr.press/v28/zemel13.html

Ziegler, C.-N., McNee, S. M., Konstan, J. A. & Lausen, G. (2005). Improving Recommendation Lists through Topic Diversification. Proceedings of the 14th International Conference on World Wide Web, pp. 22–32. https://doi.org/10.1145/1060745.1060754

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.