Analysis of The Impact of Deep Learning-Based Recommendation Algorithms on Demographic Groups

Jiepeng Niu

DeGroote School of Business, McMaster University, Hamilton, Ontario, L8S 4L8, Canada

Abstract. This study compares the performance of TensorFlow Recommender (TFRS), Light Factorization Machine (LightFM), and Weighted Matrix Factorization (WMF) on the MovieLens 25M dataset. It focuses on accuracy and fairness across different user groups. Experiments show that TFRS achieves good accuracy and keeps fairness across gender and age, but its performance drops sharply in sparse environments. LightFM performs better in cold-start cases but shows large gaps in fairness, especially among older users. WMF shows the most consistent fairness across age and gender groups because it uses confidence-weighted feedback methods, though its accuracy is lower. In controlled tests, TFRS ranks first in recommendation accuracy, WMF ranks first in exposure balance, and LightFM ranks first in new user adaptability. These results show that each model has strengths depending on the deployment environment. TFRS is good for mobile apps with quick user updates, WMF suits systems with high fairness needs, and LightFM is good when handling new users.

1 Introduction

Demographic differences in recommender systems are an important research topic. Studies show that users of different age groups and genders often get different levels of recommendation accuracy [1]. Earlier analysis of the MovieLens dataset found that the quality of recommendations changes across groups. Older users and women usually get content with less diversity or novelty, which shows an imbalance in personalized recommendations. These results suggest that deep learning models, although good at personalization, can still create demographic biases if they are not adjusted for user diversity. To deal with demographic sensitivity, TFRS uses a dual-tower design to add age and gender features. This helps the model tell user traits apart in recommendation layers [2]. LightFM uses a hybrid design that mixes demographic information with content and collaboration signals. This approach improves results, especially for users who have little history [3]. The weighted matrix factorization model handles data imbalance by setting different confidence weights for user-item interactions, which helps to reduce bias from overrepresented or underrepresented groups [4].

niuj25@mcmaster.ca

This study uses RMSE to measure accuracy and group fairness index (GFI) to measure fairness, which comes from a fairness difference method for demographic groups [1]. RMSE looks at overall prediction performance, while GFI checks bias across different user groups, such as age or gender. Previous studies found that real-world platforms like Netflix do not depend on just one algorithm. Instead, they use several models to handle different personalization needs [5]. This setup helps balance popularity trends and individual user signals. This study compares TFRS, LightFM, and WMF on the MovieLens 25M dataset. It focuses on fairness and accuracy across gender and age groups. By studying how each model reacts to demographic differences, the study hopes to find gaps in recommendation results and give model-specific insights for fairer system design. Also, earlier research showed that methods based on user averages can hide important demographic gaps, like women and older users often getting fewer relevant or novel recommendations [6].

2 Related Work

Previous studies have looked at demographic biases in recommender systems. Wang et al. found that users from different age and gender groups get very different recommendation results [1]. Their analysis proposed fairness-aware evaluation metrics but did not compare models across different architectures. In the same way, Gomez-Uribe and Hunt built a hybrid recommendation model for Netflix that improves personalization by combining collaboration and content filtering, but they did not include fairness or demographic features [5]. Recent advances have explored incorporating demographic attributes into the recommendation process. Paranjape et al. built a classifier using gender and age to improve the matching accuracy between users and content profiles [3]. Mehta's work showed the impact of concept drift on stream classification but did not consider underrepresented user groups [7]. However, many previous evaluations focus on system-wide aggregate metrics, which may obscure the effectiveness of serving different populations. Netflix's real-world experiments show that even deep learning models can only outperform traditional methods when enhanced with rich, heterogeneous features, otherwise, the advantage is limited. This shows that evaluating models based solely on average performance without subgroup analysis may overlook critical fairness gaps [8]. Yin et al. extended this direction by embedding demographic filtering strategies into machine learning frameworks, but their work lacked standardized fairness evaluation or cross-model benchmarking [9].

To enhance large-scale personalization, Covington et al. developed a deep learning pipeline for YouTube that uses a multi-stage architecture to capture user behaviour [2]. However, such models prioritize performance over fairness, and their effectiveness across subgroups has not been tested. Lee and Noor Zuraidin applied clustering to the MovieLens dataset to group users by demographics, providing a segmentation-based recommendation approach [10]. However, their evaluation did not quantify fairness across groups. Different from previous studies, this study conducts a head-to-head comparison of three different recommendation models on the MovieLens 25M dataset: TFRS, LightFM, and WMF, evaluating the performance using RMSE and GFI, and deeply studying each model's ability to balance accuracy and demographic fairness across gender and age subgroups.

3 Methodology

This study follows a structured process that includes three steps: data preprocessing, experimental design, and metric evaluation. In the data processing stage, low-activity users will be screened, age and gender will be encoded, and a stratified training-test set split will be applied. In the experimental stage, three models will be evaluated in four scenarios: TFRS,

LightFM, and WMF, respectively, for accuracy, gender fairness, age fairness, and cold start robustness. To ensure consistency and multi-dimensional comparison, RMSE, Precision, Recall, F1, Click-Through Rate (CTR), and GFI are used to evaluate performance. The full workflow is illustrated in Fig. 1.

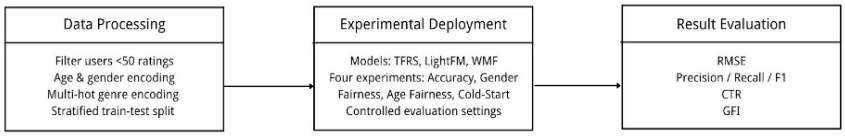

Fig. 1 The Experimental flowchart of this study (Photo credit: Original)

3.1 Data Processing

This study uses the MovieLens 25M dataset, which provides user-item ratings as well as movie metadata such as genre. To improve data quality, users with fewer than 50 ratings are excluded to reduce noise. Since the dataset does not contain demographic attributes, synthetic age and gender labels are generated through stratified sampling to simulate the real distribution for fairness analysis. Age is divided into four intervals: <18 years old (adolescents), 18-35 years old (young people), 36-55 years old (middle-aged people), and >55 years old (elderly people). Gender is binary encoded (0 = male, 1 = female). The genre of each movie is multi-hot encoded, allowing each movie to have multiple category labels. The final dataset is divided into training and test sets using stratified sampling by demographic group to ensure balanced distribution of data in each group.

3.2 Model Principles and Optimization

This study compares the performance of three recommendation algorithms with different architectures and optimization methods. Each model is built separately using the same processed dataset to make the evaluation fair.

TFRS uses a dual-tower structure. It maps user demographics (age and gender) and item types into different embedding spaces. These spaces are optimized together to predict user-item matches. This design allows personalization and keeps demographic patterns in mind.

LightFM uses a hybrid model. It mixes collaborative filtering with content-based features like user demographics and movie genres. This setup helps users with little history by using extra features and solves the cold start problem when interaction data is few.

WMF adjusts the training by giving higher confidence weights to frequent user-item interactions. This change reduces bias from groups with more interactions and helps give fairer exposure across users, even though prediction accuracy may be lower.

3.3 Evaluation Metrics

Six metrics are used to evaluate prediction accuracy and recommendation quality for different demographic subgroups. The six metrics are RMSE, GFI, precision, recall, F1 score, and CTR. The multi-metric setup is chosen to show not only the differences in rating prediction performance but also the differences in user-focused recommendation outcomes and simulated engagement. It allows a broader comparison of model performance.

Root mean square error measures the average size of rating prediction errors. It is found by taking the square root of the mean squared difference between predicted ratings and true ratings for all points. In the formula, N represents the total number of rating observations, $\hat{r}_i$ represents the predicted rating for the i-th observation, and r_i represents the true rating for the corresponding observation. This metric reflects the overall accuracy of the predicted scores, with larger errors being penalized more heavily.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_i - r_i)^2} \quad (1)$$

The group fairness index assesses fairness by examining the differences in exposure across demographic groups, such as age or gender. In the equation, G represents the total number of demographic groups, while g represents the index for each specific group. N_g is the number of users or items in group g, E(i) represents the exposure (i.e., recommendation frequency) of item i, and $\bar{E}$ corresponds to the average exposure across all groups. Lower GFI values indicate that the recommendations provided by the system are more demographically balanced.

$$GFI = \frac{1}{G}\sum_{g=1}^{G}\left/\frac{1}{N_g}\sum_{i\in g}E(i) - \bar{E}\right/ \quad (2)$$

Precision reflects the proportion of recommended items that are truly relevant to the user. The "True Positive" in the formula represents the number of relevant items that are correctly recommended, while the "False Positive" represents the number of irrelevant items that are incorrectly recommended. Precision is particularly important when the cost of recommending irrelevant content is high, as it prioritizes a shorter but highly relevant list of items.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

Recall evaluates the model's ability to retrieve all relevant items for the user. In the recall calculation, "True Positive" again represents the number of relevant items that are successfully recommended, while "False Negative" represents the number of relevant items that are missed. Recall is particularly important in applications where failure to recommend useful content can seriously affect the user experience.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

The F1 score provides a harmonic mean between precision and recall. It effectively balances the trade-off between these two metrics. The "Precision" in the formula measures the correctness of the recommended items, and the "Recall" evaluates the completeness of the recommendation. The F1 score is particularly useful when the cost of false positives and false negatives is considered equal, providing a single, unified assessment of recommendation quality.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (5)$$

CTR models user engagement by measuring the proportion of recommended items that users engage with. In the CTR formula, "Number of Clicked Items" represents the number of recommended items that users clicked or accepted, while "Number of Recommended Items" represents the total number of items recommended to the user. This metric indirectly measures user satisfaction by reflecting how often recommendations lead to engagement and reflects the regret users feel when they ignore irrelevant items.

$$CTR = \frac{Number\ of\ Clicked\ Items}{Number\ of\ Recommended\ Items} \quad (6)$$

4 Experiments and Results

4.1 Dataset

The MovieLens 25M dataset was used as the basis for the experiments, and the raw data was pre-processed in multiple steps. Users with fewer than 50 ratings were filtered. Since demographic features (age and gender) were not present in the original dataset, stratified sampling was used to generate synthetic labels. These generated attributes were then used to support fairness-related experiments. Movie genres were extracted and converted into multi-hot vectors to preserve multi-category assignments. After processing, the entire dataset was stratified into training and test sets to preserve the demographic proportions of all groups. The final cleaned dataset was used consistently in all models and experiments.

4.2 Cross-model Accuracy Evaluation (TFRS vs. LightFM vs. Weighted Matrix Factorization)

To evaluate the differences in model behaviour, three widely adopted recommendation algorithms were selected: TFRS, LightFM, and WMF. Four performance metrics were used, including RMSE for regression tasks, and precision, recall, and F1 score for classification evaluation.

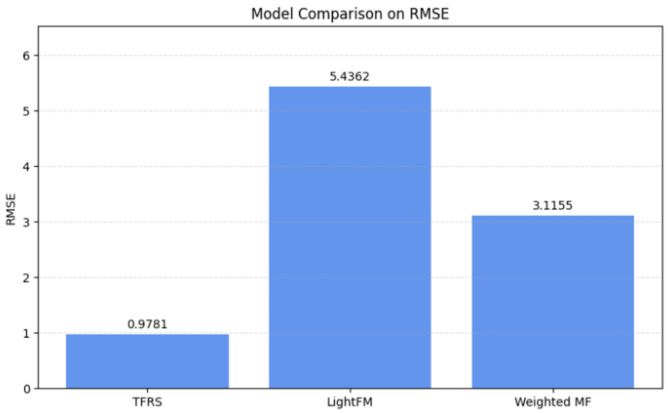

Fig. 2 Model comparison based on RMSE (Photo credit: Original)

As shown in the RMSE comparison in Fig. 2, TFRS has the highest accuracy and the lowest RMSE of 0.9781; WMF ranks second with an RMSE of 3.1155; and LightFM has the highest error of 5.4362. These results show that TFRS performs well in modelling user

preferences and rating behaviours, and its prediction error is significantly lower than the other two models.

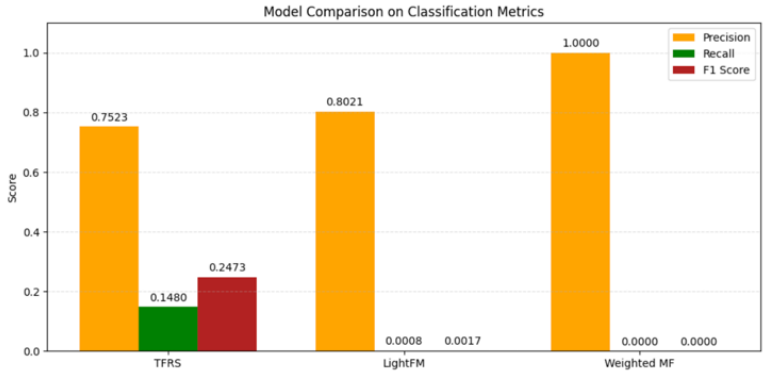

Fig. 3 Comparison of model classification metrics (Photo credit: Original)

Fig. 3 shows that among the classification metrics, TFRS presents the most balanced results in all three metrics (accuracy = 0.7523, recall = 0.1480, F1 score = 0.2473). This reflects that it performs moderately but steadily in identifying related items without overfitting. LightFM has a relatively high precision of 0.8021, but its recall (0.0008) and F1 score (0.0017) are extremely low, indicating that the model's prediction strategy is too cautious, retrieving only a small number of items and ignoring most relevant cases. Weighted MF's performance is even more extreme, with the highest precision of the three models (1.0), but completely failing in both recall (0.0000) and F1 score (0.0000). This shows that although a few predictions of WMF are technically correct, the overall recommendation coverage is very limited and lacks generalizability.

4.3 Gender Fairness Evaluation

To evaluate gender fairness in the recommendation system, this experiment calculated the CTR of male and female users, respectively, and measured the group fairness index (GFI) as the absolute difference between the two. All indicators are based on the pre-processed dataset and the prediction scores of three models: TFRS, LightFM, and WMF.

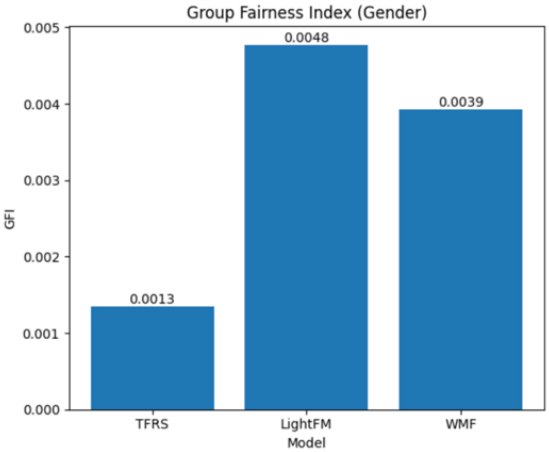

Fig. 4 GFI by Gender Across Recommendation Models (Photo credit: Original)

As shown in Fig. 4, LightFM has the highest GFI (0.0048), indicating the largest difference in exposure between male and female users, and reflects the most demographically unbalanced output of the three models. WMF is followed closely by WMF, which has a GFI of 0.0039, indicating slightly better gender balance than LightFM, but there is still a considerable gap. TFRS has the lowest GFI (0.0013), which indicates the most balanced exposure distribution between genders.

The CTR values further validate the trends. As shown in Fig. 5, LightFM has the highest CTR for male users (0.5604) and a slightly lower CTR for female users (0.5557), resulting in the largest gender difference among the three models. WMF is closely followed by WMF, with a CTR of 0.5560 for males and 0.5521 for females, with a moderate gender difference. In contrast, TFRS performs the most balanced, with a CTR of 0.5557 for males and 0.5543 for females, with the smallest gender difference. The bar charts in Fig. 5 clearly reflect these trends, with smaller differences between male and female CTR values indicating more balanced user engagement. These results show the trade-off between maximizing engagement and maintaining population fairness. TFRS prioritizes balanced exposure, LightFM tends to increase engagement at the expense of fairness, and WMF falls in between.

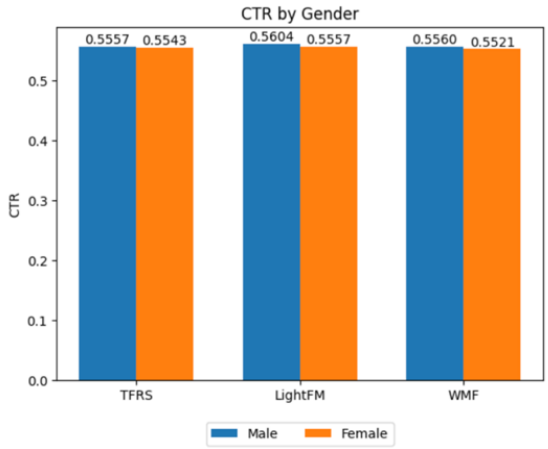

Fig. 5 CTR by Gender Across Recommendation Models (Photo credit: Original)

4.4 Age Fairness Evaluation

To evaluate the fairness in terms of user age, the recommendation effects of the three models in four age groups, namely adolescents, young people, middle-aged people, and elderly people, were compared based on the GFI and CTR.

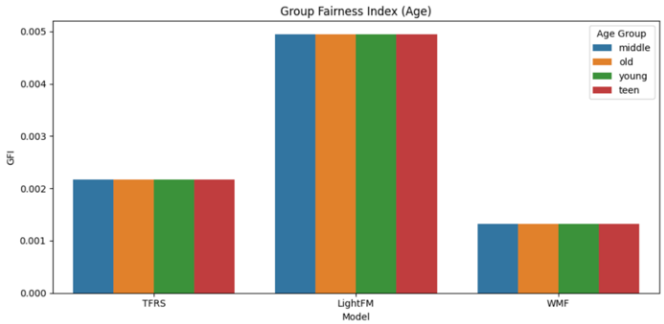

Fig. 6 GFI by Age Group (Photo credit: Original)

As shown in Fig. 6, LightFM has the largest fairness difference (0.00495), indicating that there are significant differences in the effects of different age groups. Specifically, its CTR ranges from 0.1975 (youth) to 0.2079 (teenagers), which is the largest CTR gap among all models. In contrast, WMF has the lowest GFI (0.00132) and CTRs concentrated between 0.1950 and 0.2049, indicating that its user exposure is fairer. TFRS has a medium GFI (0.00217) and CTR values ranging from 0.1997 (young) to 0.2026 (elderly), reflecting its relatively balanced performance.

The CTR distribution further demonstrates the consistency of user engagement across age groups. As shown in Fig. 7, WMF has the most stable CTR levels, with scores ranging from 0.1950 to 0.2004 across all age groups, which is consistent with its lower GFI values. On the other hand, LightFM has the most variability, with higher CTRs for teenagers (0.2079) and lower CTRs for young adults (0.1975), highlighting its high fairness variability. TFRS's CTR values range from 0.1997 for young users to 0.2026 for older users, indicating moderate stability, although slightly higher volatility than WMF. The visual comparison in Fig. 7 confirms that models with lower GFI values, such as WMF, tend to maintain more consistent engagement across age groups.

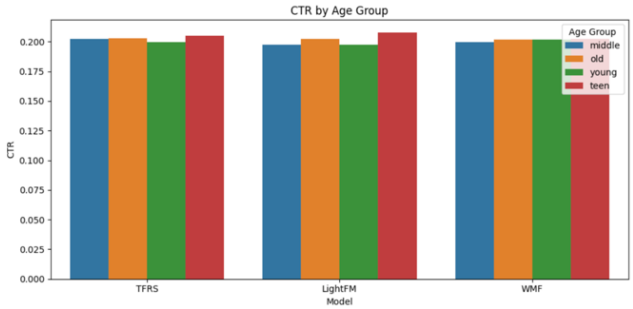

Fig. 7 CTR by Age Group (Photo credit: Original)

4.5 Cold-Start Scenario Simulation

To evaluate the robustness of the recommendation model under sparse interaction conditions, this experiment simulates a cold start scenario and filters out users with fewer than 10 ratings. These users typically lack sufficient behavioural history for traditional collaborative filtering models to generate accurate recommendations. The experiment compares the model performance on the cold start group and a randomly sampled set of 1,000 non-cold start users, using recall@5, F1 score, and CTR as evaluation metrics.

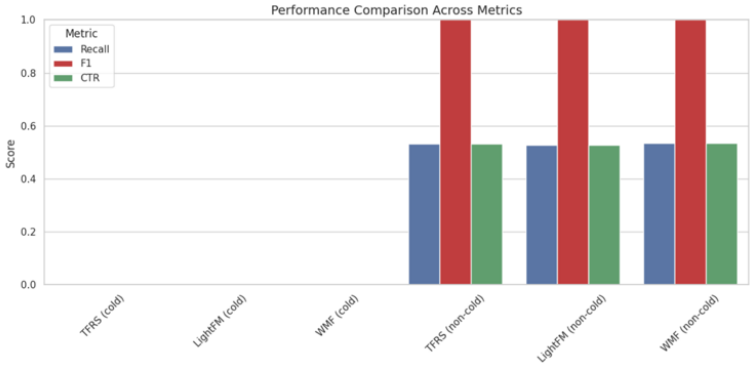

Fig. 8 Cold-Start and Non-Cold-Start Performance Across Models (Photo credit: Original)

As shown in Fig. 8, TFRS, LightFM, and WMF models all failed to produce meaningful results for the cold-start group. The recall and click-through values were undefined, while the F1 score dropped to 0. This result highlights the limitations of collaborative filtering models in the absence of sufficient interaction history. This suggests that all models are unable to effectively generalize to less engaged users. However, WMF has the highest recall (0.5132), followed by LightFM (0.5072) and TFRS (0.4988). All models achieved perfect F1 scores (1.0) and high consistency between click-through and recall, which indicates that they can successfully identify relevant recommendations given rich interaction data.

4.6 Analysis

This section critically analyses the experimental results, including data and graphs, presented in former sections, and provides a detailed analysis of patterns, inconsistencies, and model-specific behaviours on four dimensions: prediction accuracy, gender fairness, age fairness, and cold-start robustness.

In terms of accuracy, TFRS always performs better than LightFM and WMF. It gets the lowest RMSE and the most even classification results. LightFM also has high accuracy, but its recall is very low, which shows it often under-recommends and ranks items too safely. WMF has more extreme results. It shows perfect accuracy but zero recall and F1 score. This means it works only in a few cases and cannot give good recommendations in most cases. These results show the trade-off between model complexity and ability to generalize. Deep models like TFRS can better find user-item patterns, while simple methods can fail when data is sparse or uneven.

From a gender fairness view, all models show close click-through rates for male and female users, meaning gender does not change surface engagement much. But GFI scores show hidden gaps. LightFM shows the biggest gender bias, WMF shows a moderate level, and TFRS shows the best balance. This means even if the user clicks look the same, LightFM still shows strong bias inside. WMF reduces bias a little with its confidence weighting, but it does not fully solve it. TFRS shows the best balance, keeping fairness without hurting user interaction much.

Age fairness evaluation reveals more subtle dynamics. WMF unexpectedly becomes the most stable model in this dimension, with the lowest GFI and narrow CTR distribution. This suggests that despite its lower overall accuracy, the exposure of users of different ages is still balanced. TFRS performs the second best, maintaining a moderate difference. In contrast, LightFM exhibits the strongest age correlation, with a significant performance difference between teenagers and young adults. This difference may stem from the way LightFM fuses collaboration and content signals, with age groups with richer interaction histories (e.g., teenagers) benefiting more, while less active age groups (e.g., older or younger users) may suffer from sparse item bias. The results suggest that demographic fairness depends not only on model quality but also on the interaction of user attributes with the density of training signals.

Under cold start conditions, the performance of all three models deteriorates drastically, with undefined recall and click-through rate values and F1 scores dropping to zero. This flaw further highlights the core weaknesses shared by TFRS, LightFM, and WMF. They rely on the user's interaction history with items, which makes them ineffective for new or sparsely active users. Due to the lack of basic predictive signals, these models cannot produce meaningful results even with architectural differences or added auxiliary features. In contrast, the non-cold start group shows stable and effective performance. The consistency between models shows that if there is enough data, the recommendation quality tends to converge. But this also makes the flaws in sparse conditions more prominent, which highlights the need for better cold start solutions.

In summary, no single model dominates across all dimensions. TFRS provides the best prediction accuracy and maintains reasonable fairness, but degrades under cold-start conditions. LightFM is more robust to sparse data and exhibits moderate fairness performance, but its prediction accuracy suffers from overly cautious recommendations. WMF achieves the highest population fairness (especially in terms of age) but fails to provide meaningful recommendations in terms of recall and click-through rate. These trade-offs suggest that recommender system design requires coordination between objectives and model selection.

5 Conclusion

This study conducts a comprehensive comparative evaluation of three recommendation models: TFRS, LightFM, and Weighted Matrix Factorization on the MovieLens 25M dataset, focusing on prediction accuracy, demographic fairness, and cold-start environment evaluation. The results show that while TFRS achieves superior accuracy and balanced F1 scores in the standard setting, its performance drops significantly in the cold-start environment. Although LightFM has lower overall accuracy, it shows stronger robustness when the user interaction history is sparse. WMF gives the worst results in both accuracy and engagement, but achieves the most balanced exposure among different age groups. One important finding from the experiment is the trade-off between personalization and demographic inclusiveness. Deep learning models like TFRS capture small user-item patterns well but depend a lot on dense interaction data. This makes them less generalizable in real-world cold-start settings. On the other hand, simpler models like WMF can produce more stable exposure because their outputs are more conservative. These results show that the choice of algorithm should match closely with the system's application goals.

This study also has some limitations. It only tested three models with fixed hyperparameter settings. It also used only age and gender as fairness features. Besides, fairness was evaluated only based on exposure differences GFI without considering long-term satisfaction or diversity. Future work should focus on fairness-aware optimization methods, like adversarial training and reweighting strategies, to directly reduce bias during model training. It would also be helpful to expand the set of demographic factors (for example, occupation and location). Adding time-related changes and building hybrid models that combine collaborative filtering with large-scale language or vision models could also help create more inclusive and context-aware recommendations.

References

1. Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: ‘A survey on the fairness of recommender systems’, ACM Trans. Inf. Syst., 2023, 41, (3), pp. 1–43
2. Covington, P., Adams, J., Sargin, E.: ‘Deep neural networks for YouTube recommendations’, Proc. 10th ACM Conf. on Recommender Systems, Boston, MA, USA, Sept. 2016, pp. 191–198
3. Paranjape, V., Nihalani, N., Mishra, N.: ‘Design and development of a demographic-based movie recommender system using hybrid ML techniques’, Int. J. Comput. Commun. Control, 2024, 19, (4)
4. Hu, Y., Koren, Y., Volinsky, C.: ‘Collaborative filtering for implicit feedback datasets’, Proc. 8th IEEE Int. Conf. on Data Mining, Pisa, Italy, Dec. 2008, pp. 263–272

5. Gomez-Uribe, C.A., Hunt, N.: 'The Netflix recommender system: Algorithms, business value, and innovation', ACM Trans. Manage. Inf. Syst., 2015, 6, (4), pp. 1–19
6. Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., et al.: 'All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation', Proc. Conf. on Fairness, Accountability and Transparency (FAT*), New York, USA, Jan. 2018, pp. 172–186
7. Mehta, S.: 'Concept drift in streaming data classification: Algorithms, platforms and issues', Procedia Comput. Sci., 2017, 122, pp. 804–811
8. Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., Basilico, J.: 'Deep learning for recommender systems: A Netflix case study', AI Mag., 2021, 42, (3), pp. 7–18
9. Yin, L.J., Safar, N.Z.M., Kamaludin, H., Abdullah, N., Yusof, M.A.M., Supriyanto, C.: 'Adopting machine learning in demographic filtering for movie recommendation systems', J. Soft Comput. Data Min., 2023, 4, (1), pp. 1–12
10. Lee, J.Y., Safar, N.Z.M.: 'Research on the demographic filtering machine learning in movie recommendation system', Appl. Inf. Technol. Comput. Sci., 2023, 4, (1), pp. 290–307