DA 626 Project

Paper : Improving Recommendation Fairness via Data Augmentation

https://doi.org/10.1145/3543507.3583341

Team 29

Name: Kaki Hephzi Sunanda

Roll Number: 210150018

Abstract:

This project explores the paper titled “Improving Recommendation Fairness via Data Augmentation”. The paper proposes a framework to improve fairness in recommendations while maintaining recommendation quality. Briefly, the framework involves Fairness-aware Data Augmentation (FDA) to generate synthetic interactions that imitate real word user-item interaction scenarios and integrate these synthetic interactions into the training process to alleviate demographic disparities.

Through the knowledge of the above framework, three key objectives were explored in this project:

1. Assess fairness on a non-binary sensitive attribute (age).
2. Evaluate fairness across multiple sensitive attributes (gender and age).
3. Analyse implicit dataset biases through user-item interaction distributions and performance metrics across groups.

Evaluation metrics such as Demographic Parity (DP), Equality of Opportunity (EO), Normalized Discounted Cumulative Gain (NDCG), and Hit Rate were used to measure fairness and recommendation performance.

Introduction:

The paper “Improving Recommendation Fairness via Data Augmentation” proposes the Fairness-aware Data Augmentation (FDA) framework to help address the issue of amplified unfairness in recommendation models due to inherent bias in training data. Fairness-aware frameworks, such as Fairness-aware Data Augmentation (FDA) brought forward in this paper, aim to mitigate these biases by introducing synthetic interactions to balance these sensitive group recommendation disparities.

This project implements and evaluates the FDA framework on fairness metrics such as Demographic Parity (DP) and Equality of Opportunity (EO) and recommendation metrics such as Normalized Discounted Cumulative Gain (NDCG) and Hit Rate, as originally used in the paper. The project employs the Bayesian Personalized Ranking (BPR) model and explores the generalizability of FDA across sensitive attributes.

Objectives

This project aims to explore the effectiveness of the Fairness-aware Data Augmentation (FDA) framework through the following objectives:

1. Non-Binary Sensitive Attribute Analysis:
    - Evaluate fairness metrics on a non-binary sensitive attribute (such as 'age') to understand the framework's impact across various demographic groups.
2. Multiple Sensitive Attributes:
    - Assess fairness metrics when considering multiple sensitive attributes together (gender and age) to address complex forms of unfairness.
3. Implicit Dataset Bias:
    - Analyse the distribution of user-item interactions and assess model performance across demographic groups to identify implicit biases.
    - Evaluate the extent to which the FDA framework mitigates these biases

Dataset

The dataset used for this project is the MovieLens dataset, which contains user-item interactions in the form of movie ratings. It includes:

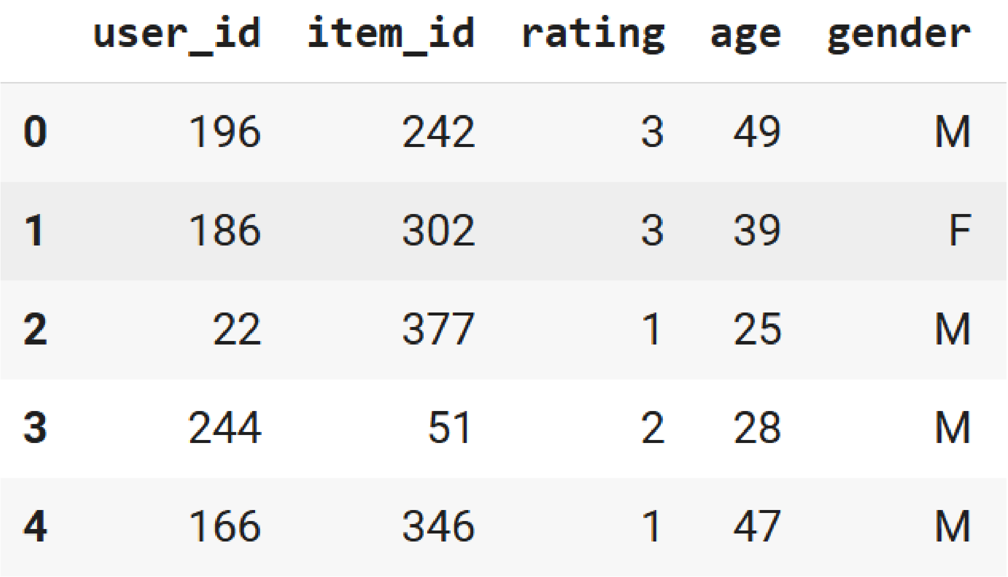
- 100,000 user-item interactions, indicating users' ratings for movies.
- 1,682 unique items (movies).
- 943 unique users.

Each interaction in the dataset is a rating ranging from 1 to 5, with higher values indicating stronger user preferences. In addition to user-item interactions, the dataset also contains user and item metadata, with user metadata including sensitive attributes like gender and age.

Data Preprocessing

The dataset is composed of three separate data frames that contain user-item interactions, user metadata, and item metadata. For this analysis, the user-item interaction data serves as the base, to which the user metadata (containing age and gender as sensitive attributes) is merged. This integration enables fairness analysis based on user demographics.

Example of the merged data:



	user_id	item_id	rating	age	gender
0	196	242	3	49	M
1	186	302	3	39	F
2	22	377	1	25	M
3	244	51	2	28	M
4	166	346	1	47	M

To distinguish between different levels of user preference, interactions are categorized as positive or negative. Ratings greater than 3 are considered positive interactions (represented as 1), while ratings of 3 or lower are considered negative interactions (represented as 0).

The age attribute from the user metadata is categorized into non-binary groups to enable a finer analysis across age demographics. The age groups are defined as follows:

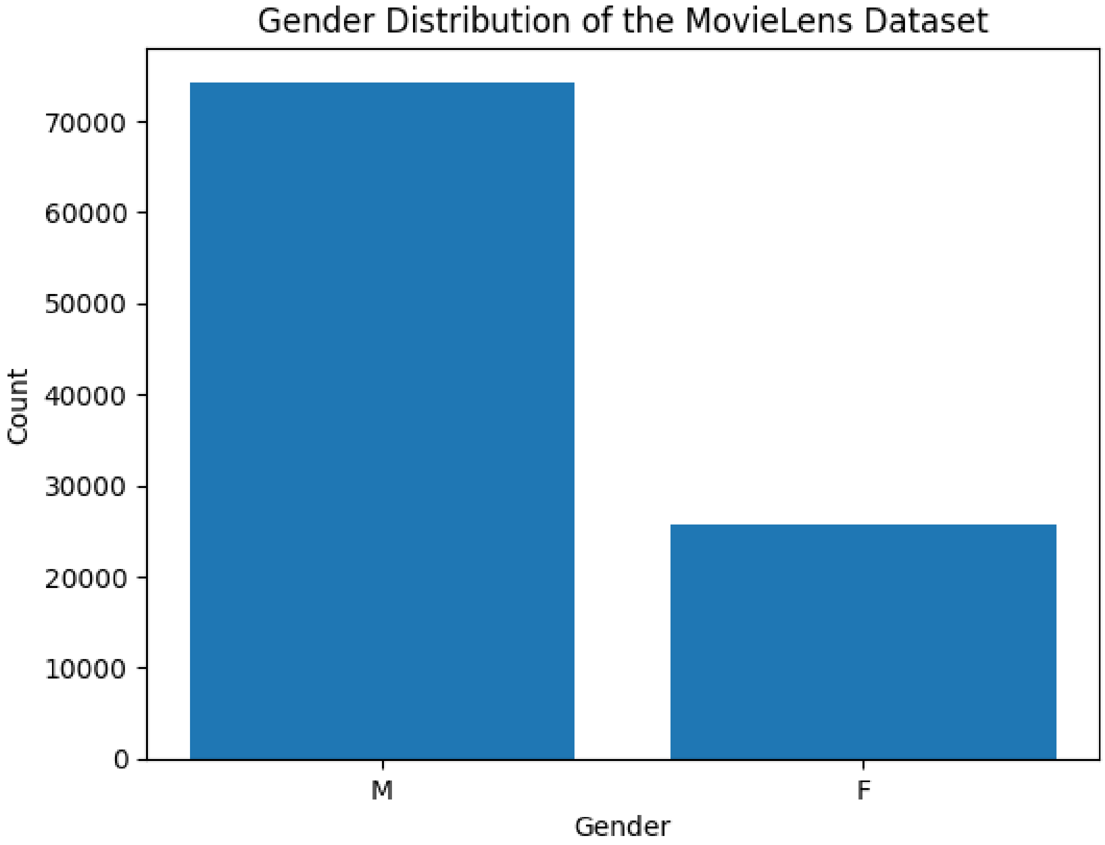
- 0: 0-18 years
- 1: 19-25 years
- 2: 26-35 years
- 3: 36-45 years
- 4: 46-55 years
- 5: 56-65 years
- 6: 66-100 years

Data Exploration:

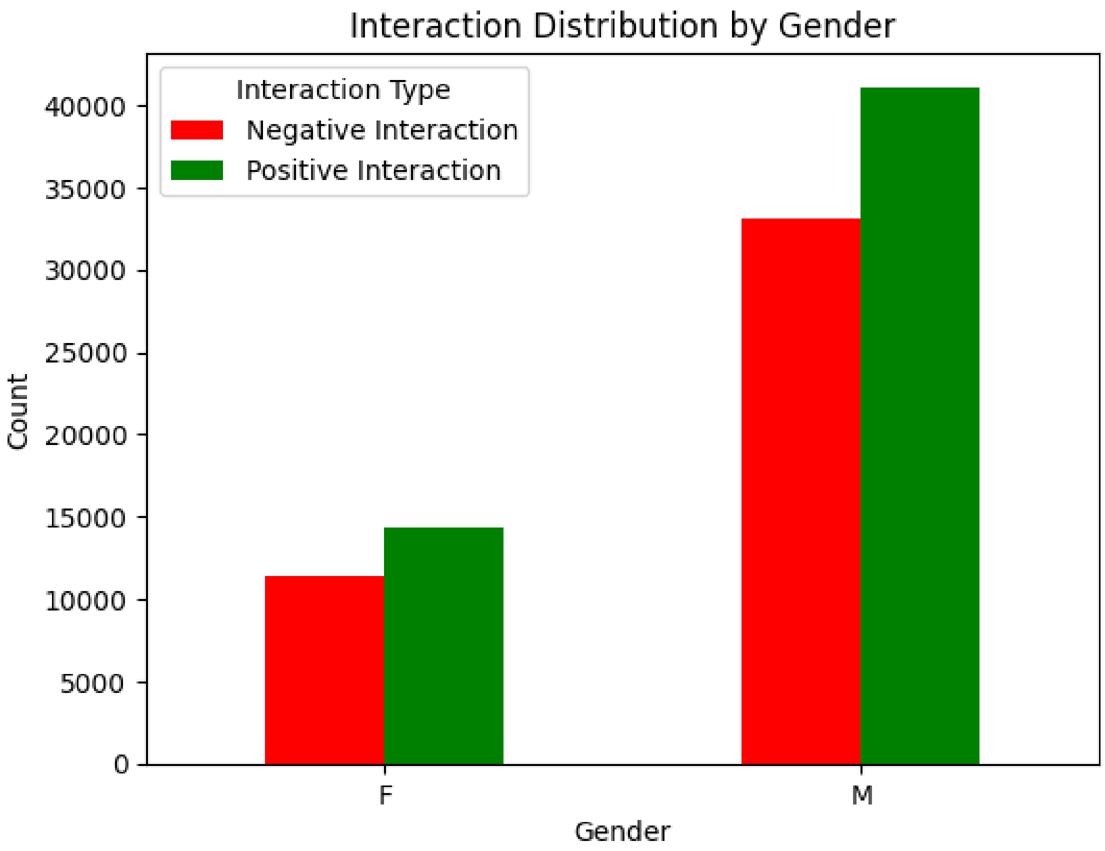
An exploratory analysis was conducted on the MovieLens dataset to understand the demographic distribution and interaction behaviour across different sensitive attributes. This analysis provides insight into any inherent imbalances within the data that could impact recommendation fairness.

A bar chart was used to visualize the distribution of genders in the dataset. The majority of interactions come from male users, with a significantly lower number from female users.

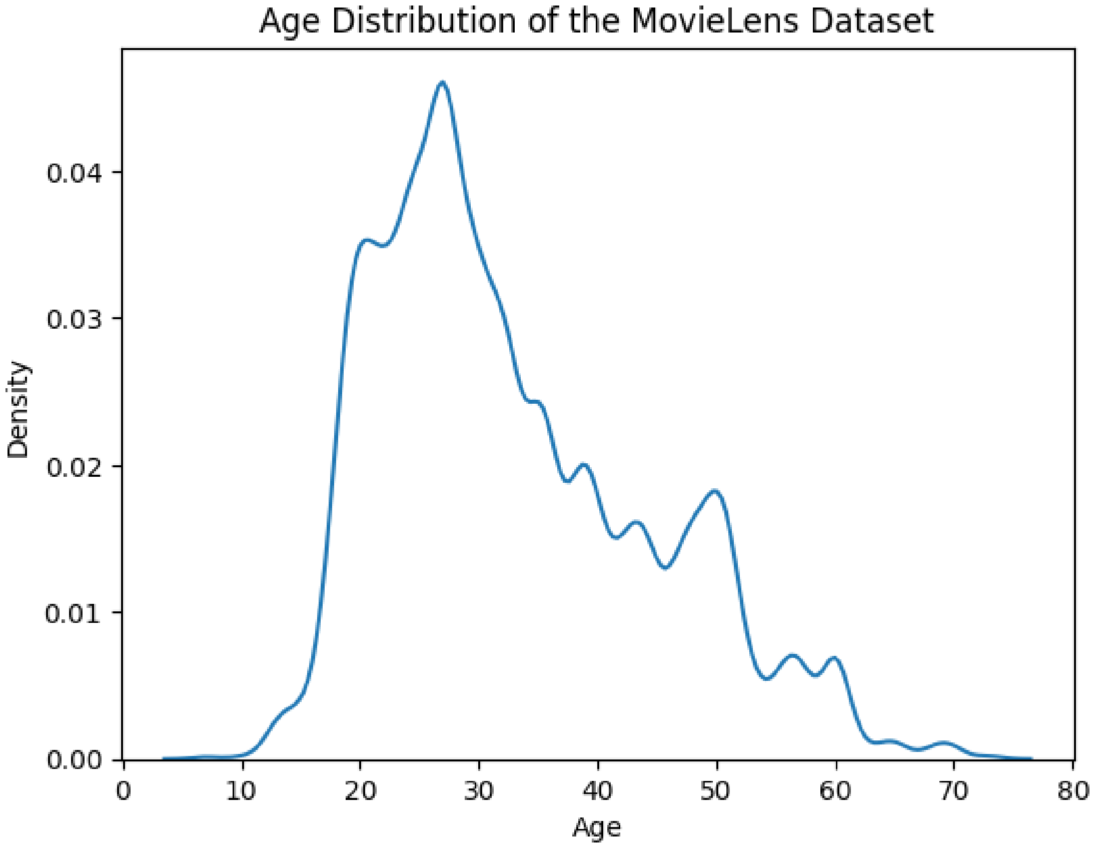
This gender imbalance indicates a potential source of bias in the recommendations generated by models trained on this dataset.

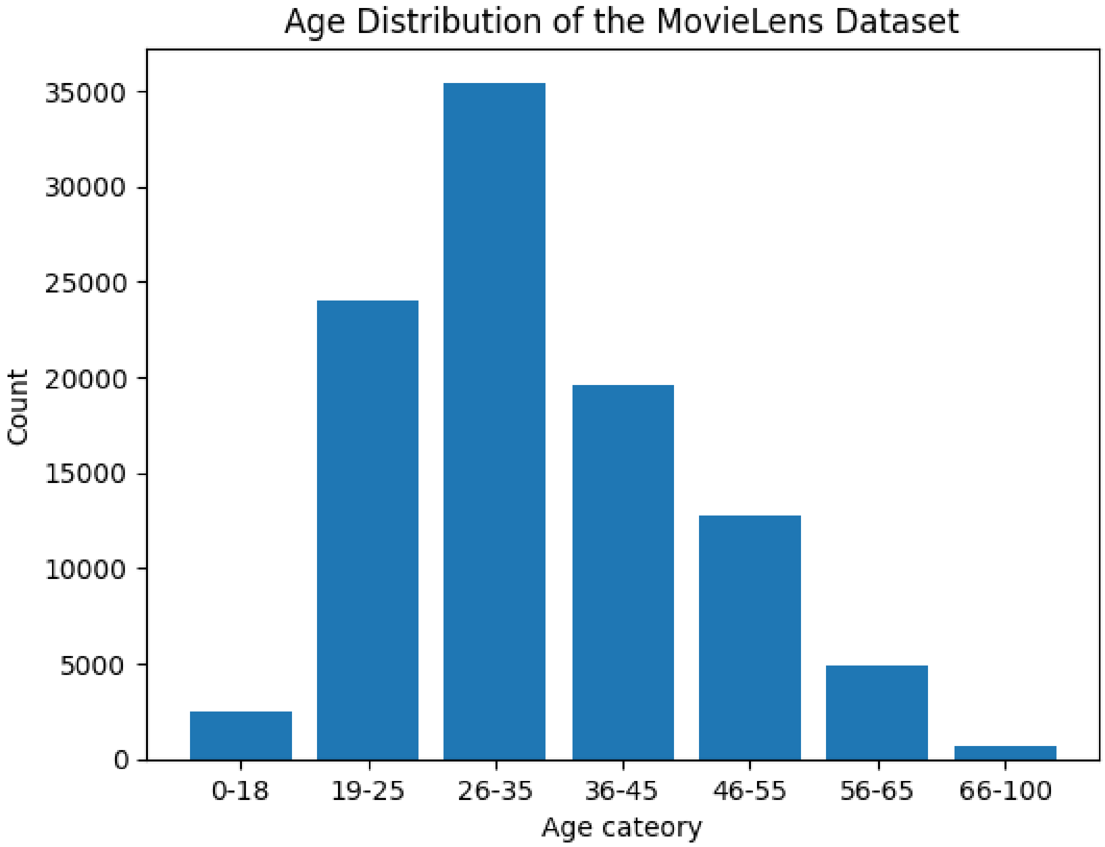


To further understand interaction patterns, the dataset was examined for positive and negative interactions by gender. The chart shows that both male and female users have a higher count of positive interactions, but the absolute number of interactions from male users is much larger, reinforcing the gender imbalance in interactions.

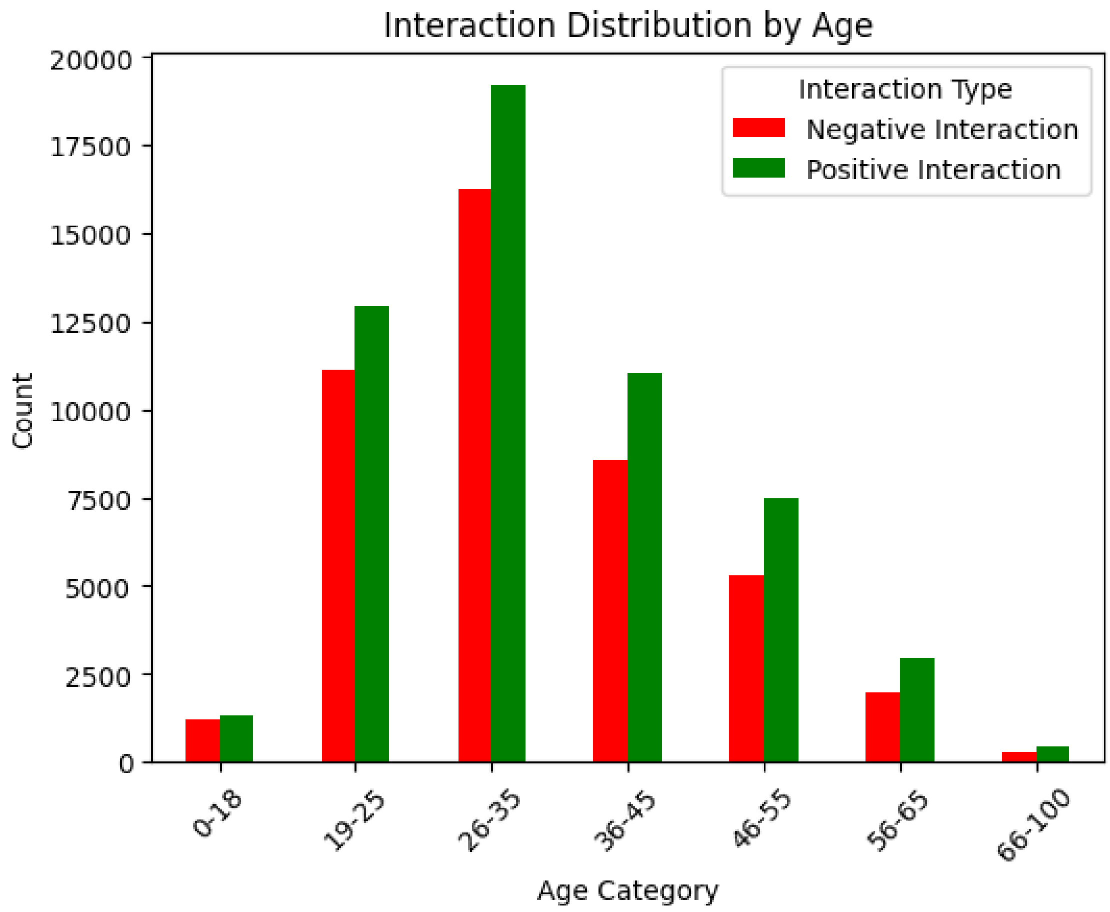


The age distribution of users was categorized into distinct age groups to assess whether certain age ranges dominate the dataset. The 26-35 age group has the highest number of interactions, followed by the 19-25 and 36-45 groups. This indicates a higher concentration of interactions from younger users.


Age Distribution of the MovieLens Dataset
Density
0.00
0.01
0.02
0.03
0.04
0
10
20
30
40
50
60
70
80
Age


Age Distribution of the MovieLens Dataset
Count
0
5000
10000
15000
20000
25000
30000
35000
0-18
19-25
26-35
36-45
46-55
56-65
66-100
Age cateory

Positive and negative interactions were also analysed across age categories. While the 26-35 and 19-25 age groups had the highest number of interactions, most age groups showed more positive interactions than negative ones, except for older age groups (56-65 and 66-100), which had relatively few interactions overall.



Data was split into training (80%), validation (10%), and testing (10%) sets.

Stratification ensured proportional representation of sensitive attributes.

Analysis:

This project uses a Bayesian Personalized Ranking (BPR) model to balance the fairness and recommendation by applying the Fairness-aware Data Augmentation (FDA) framework. The methodology involved training two models:

1. Baseline BPR model
2. BPR model with FDA to incorporate synthetic interactions aimed at improving fairness.

- Data Splitting
  - The dataset was divided into:

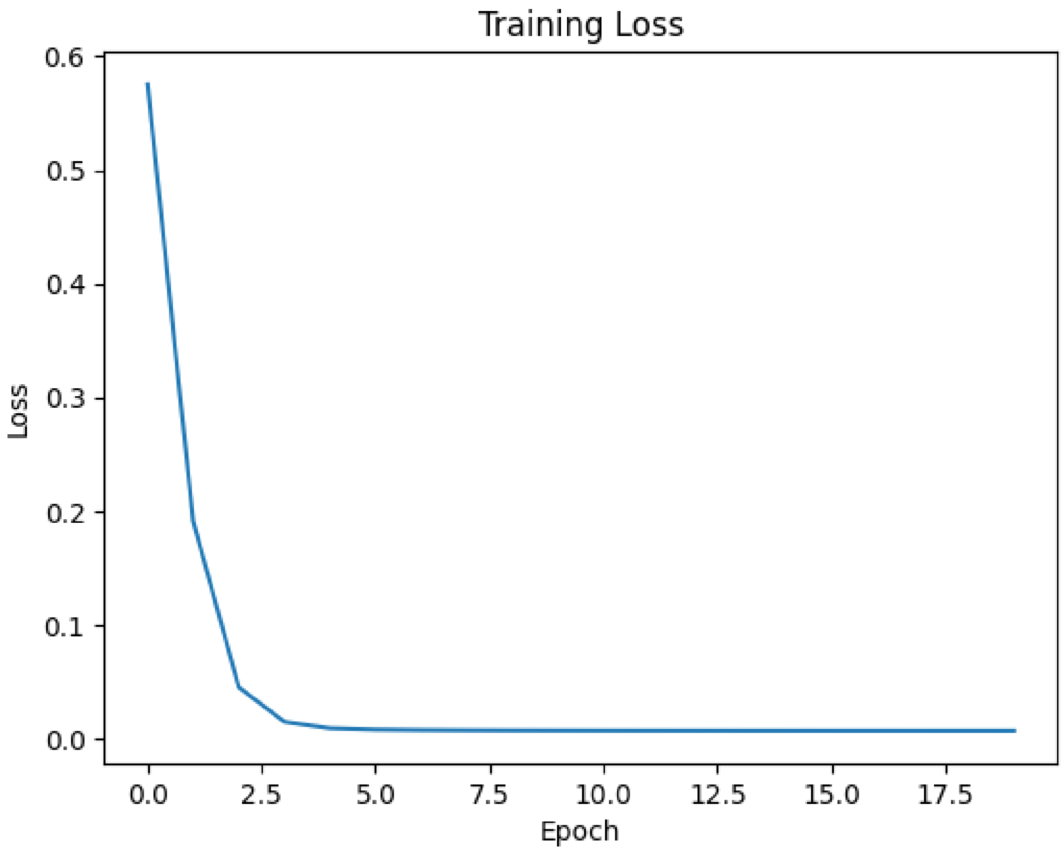
  - Training Set (80%)
  - Validation Set (10%)
  - Test Set (10%)

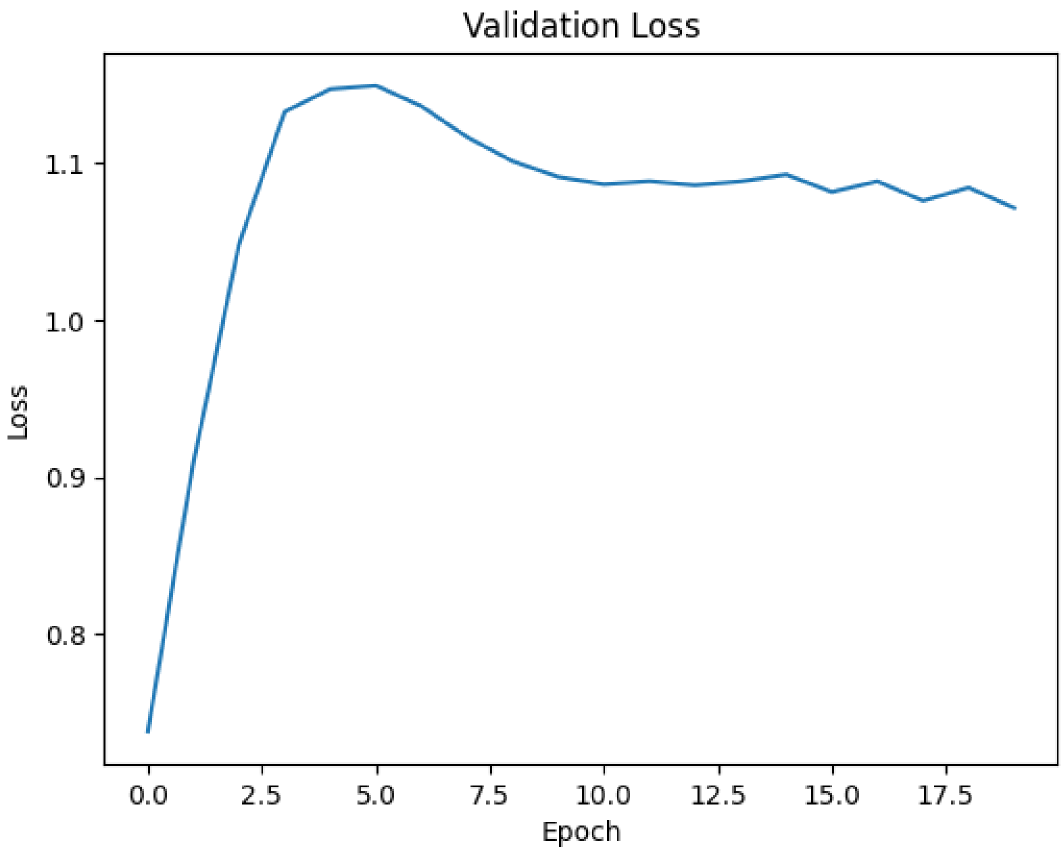
For each user, tuples of the type (user, positive item, negative item) were generated to train the BPR model. These pairs allowed the model to learn from both positive and negative interactions.

The generated pairs were used to train the baseline BPR model, aiming to maximize the likelihood of positive items over negative items for each user.

- The baseline BPR model was trained on real interactions without any synthetic data augmentation. It uses latent embeddings for each user and item, which was initialized randomly.
- The model was optimized using a binary cross-entropy loss function that compares positive and negative interactions for each user.
- L2 regularization was applied to prevent overfitting by penalizing large values in the embeddings.
- The Adam optimizer was used, and the model was trained for a specified number of epochs.

The training and validation loss were observed as follows:

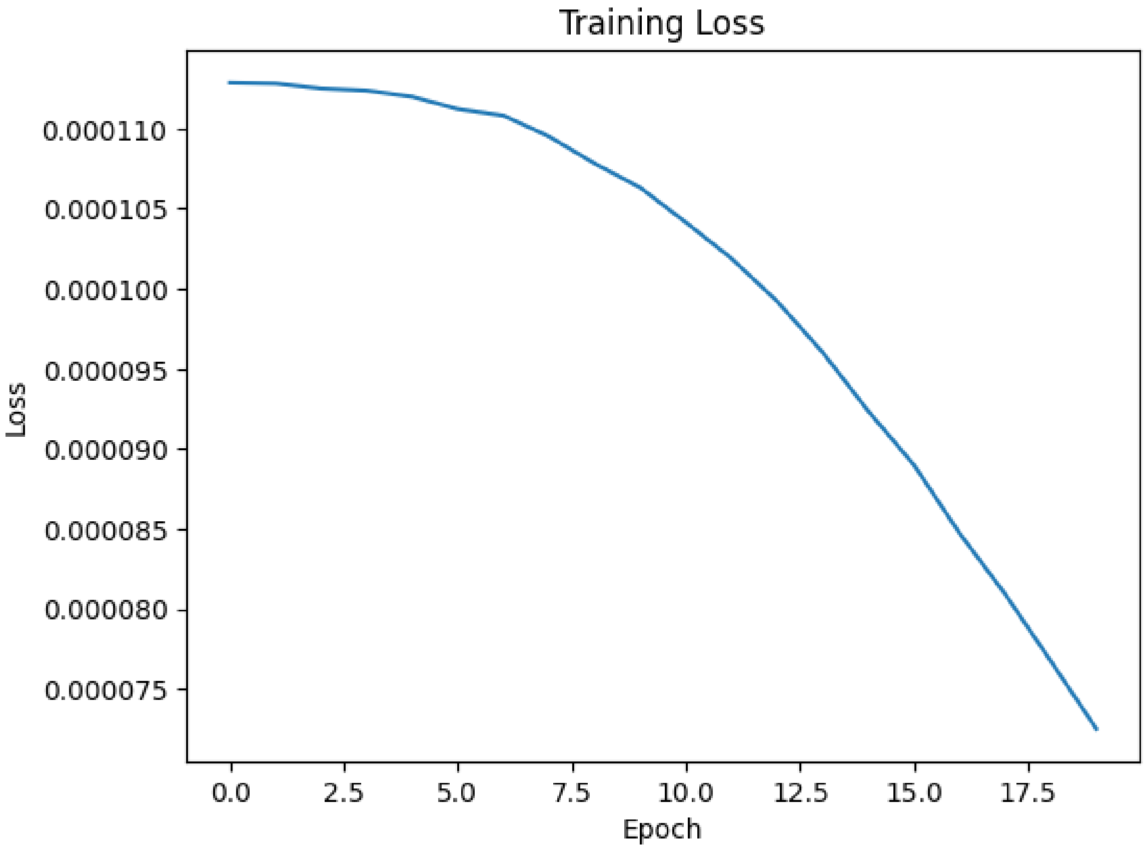


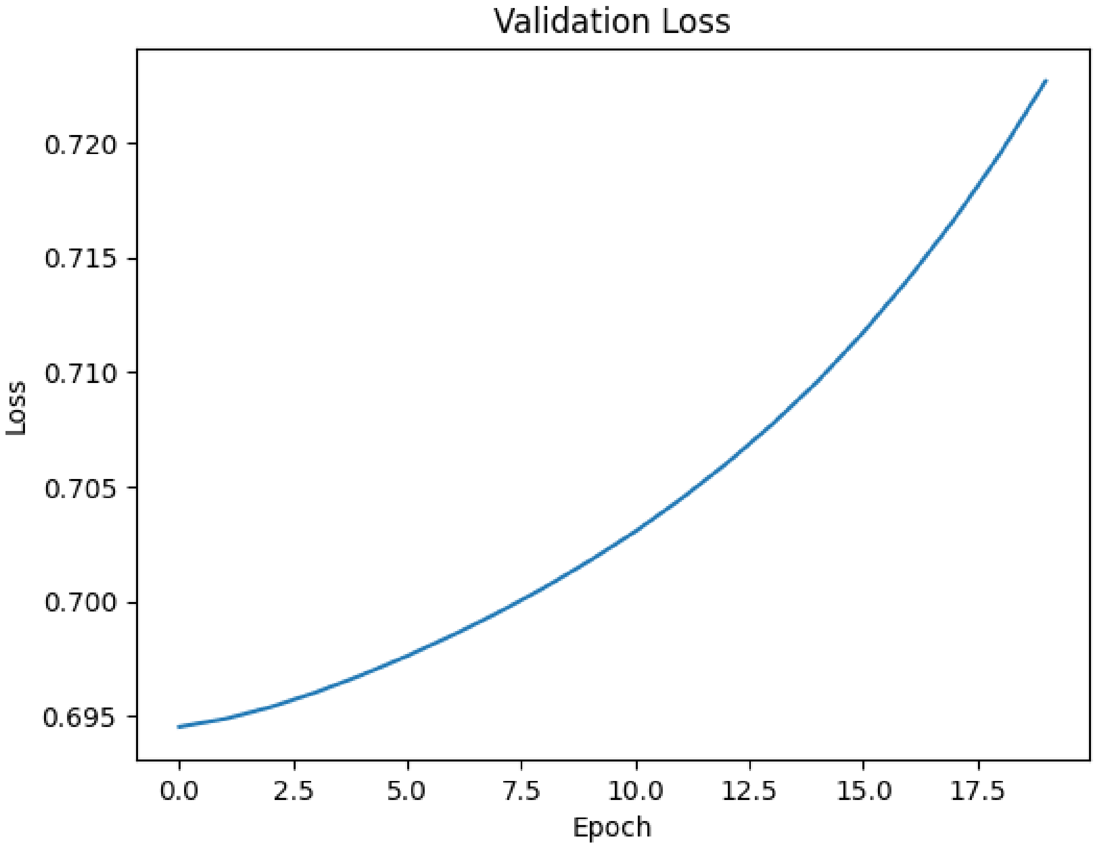


To improve fairness in recommendations, the FDA framework is implemented by modifying the BPR model to include synthetic interactions.

- For each positive user-item interaction, a synthetic item embedding was generated by adding slight noise to the item embedding. This generated synthetic pairs that mimic potential real world user interactions, addressing potential biases in the dataset.
- These synthetic interactions were appended to the original training data, creating a richer dataset with more balanced user-item interactions across sensitive groups.
- The BPR loss function was adjusted to incorporate the additional synthetic interactions.
- The FDA-BPR model was trained similarly to the baseline BPR model but on the augmented dataset.

The model's training and validation loss over epochs were observed:


Training Loss
0.000110
0.000105
0.000100
0.000095
0.000090
0.000085
0.000080
0.000075
Loss
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epoch


Validation Loss
0.720
0.715
0.710
0.705
0.700
0.695
Loss
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epoch

The results of the performance of the BPR model and BPR model with FDA framework are summarized as follows:

Metric	k=10	k=20	k=30	k=40	k=50
BPR Model without FDA					
Average NDCG	0.0056	0.0074	0.0102	0.0125	0.0138
Average Hit Rate	0.0263	0.0505	0.0788	0.0990	0.1172
Demographic Parity (DP)	0.0089	0.0086	0.0161	0.0276	0.0300
Equality of Opportunity (EO)	0.0133	0.0102	0.0233	0.0364	0.0420
BPR Model with FDA					
Average NDCG	0.0016	0.0053	0.0066	0.0081	0.0093
Average Hit Rate	0.0061	0.0303	0.0505	0.0727	0.0869
Demographic Parity (DP)	0.0118	0.0167	0.0138	0.0306	0.0343
Equality of Opportunity (EO)	0.0134	0.0180	0.0143	0.0316	0.0421

The BPR model without FDA framework generally performs better on recommendation quality metrics (NDCG and Hit Rate) compared to the BPR model with FDA. As k increases, both models improve in NDCG and Hit Rate. The FDA framework helps improve fairness metrics across all values of k. Demographic Parity (DP) and Equality of Opportunity (EO) are generally higher in the FDA model compared to the BPR model without FDA.

Objectives:

Objective 1: Fairness Analysis with Non-Binary Sensitive Attribute (Age Category)

The aim of this objective is to evaluate the FDA framework on fairness across a non-binary sensitive attribute, specifically, age categories. The original data contains user age as a continuous variable. To facilitate non-binary fairness analysis, this feature has been categorized into multiple groups.

The results are observed as follows:

Model	k	Average NDCG	Average Hit Rate	Demographic Parity (DP)	Equality of Opportunity (EO)
BPR without FDA	10	0.7131	0.0263	0.0289	0.0333
	20	0.7120	0.0505	0.1429	0.2000
	30	0.7100	0.0788	0.1429	0.2000
	40	0.7118	0.0990	0.1429	0.2000
	50	0.7115	0.1172	0.1012	0.1760
BPR with FDA	10	0.7116	0.0061	0.0111	0.0116
	20	0.7162	0.0303	0.0278	0.0320
	30	0.7143	0.0505	0.0455	0.1111
	40	0.7154	0.0727	0.0909	0.1111
	50	0.7156	0.0869	0.1364	0.1538

For the BPR model without FDA, both Average NDCG and Hit Rate increase with higher values of kk, indicating that as we consider more items (larger kk), the model provides more accurate and successful recommendations. Demographic Parity (DP) and Equality of Opportunity (EO) show relatively higher values with this implementation.

When applying the FDA framework, the Average NDCG is comparable. The FDA framework substantially reduces DP and EO, reflecting improved fairness and reduced age-based bias.

Objective 2: Fairness-aware Data Augmentation (FDA) framework across multiple sensitive attributes (gender and age category)

This objective helps to understand the framework's ability to handle multiple forms of sensitive attributes. The impact on recommendation quality and fairness metrics should be observed to evaluate the model's generalizability.

For this objective, a combined sensitive attribute group was created, by pairing gender and age category for each user. For each unique group, NDCG and Hit Rate was computed as recommendation metrics. Average of these metrics across all sensitive attribute groups was also computed to evaluate model performance. DP and EO were used again as the fairness metrics.

The results are summarized below:

Model	k	Average NDCG	Average Hit Rate	Demographic Parity (DP)	Equality of Opportunity (EO)
BPR without FDA	10	0.7131	0.0263	0.0538	0.0556
	20	0.7120	0.0505	0.1667	0.2500
	30	0.7100	0.0788	0.1667	0.2500
	40	0.7118	0.0990	0.2500	0.3333
	50	0.7115	0.1172	0.2500	0.3333
BPR with FDA	10	0.7116	0.0061	0.0111	0.0116
	20	0.7162	0.0303	0.0278	0.0320
	30	0.7143	0.0505	0.0455	0.1111
	40	0.7154	0.0727	0.0909	0.1111
	50	0.7156	0.0869	0.1364	0.1538

The Average NDCG and Hit Rate metrics are generally consistent for both models, with a slight improvement in Hit Rate as k increases for both models. However, the BPR model without FDA has slightly higher Average Hit Rates.

The FDA framework significantly reduces both Demographic Parity (DP) and Equality of Opportunity (EO) metrics across multiple sensitive attributes, demonstrating its effectiveness in alleviating bias and improving fairness in recommendations.

Conclusion:

This project explored the generalization of the Fairness-aware Data Augmentation (FDA) framework on a Bayesian Personalized Ranking (BPR) model, with a focus on achieving fairer recommendations across sensitive attributes.

The analysis was conducted across two objectives: evaluating the model's fairness and performance with non-binary sensitive attributes (age categories) and with multiple sensitive attributes (gender and age category). Key metrics such as NDCG and Hit Rate were used to assess recommendation quality, while fairness was evaluated through Demographic Parity (DP) and Equality of Opportunity (EO).

The results demonstrated that while the baseline BPR model without FDA showed higher Hit Rates, it also showed significant bias across sensitive attributes, as indicated by DP and EO values. On the other hand, the BPR model with FDA achieved substantial improvements in fairness, evidenced by lower DP and EO values across all values of k. These improvements in fairness were achieved with only minimal reductions in NDCG and Hit Rate metrics.

In conclusion, the FDA framework proved effective in increasing fairness in recommendation systems and reducing biases across sensitive attributes without compromising recommendation accuracy.