Mitigating Director Gender Bias in Movie Recommender Systems

Michael Garcia-Perez
mig009@ucsd.edu

Christine Deng
cydeng@ucsd.edu

Emily Ramond
eramond@deloitte.com

Parker Addison
paddison@deloitte.com

Greg Thein
gthein@deloitte.com

Abstract

The underrepresentation of female directors in the film industry has been extensively studied in sociology and media studies literature. Studies over several decades consistently reveal a significant gender gap, with female directors comprising only a small proportion of overall directors. Despite the recognition of this gender disparity, its impact on recommendation systems remains relatively unexplored. This project investigates whether biases stemming from gender disparities in the film industry are embedded in recommendation models. Furthermore, while existing bias mitigation tools primarily target classification tasks, their application to mitigate biases in recommendation systems has received limited attention. This project aims to bridge this gap by extending bias mitigation techniques from classification tasks to recommendation systems. The ultimate goal is to develop a fair movie recommender system that reduces biases associated with the gender of the director.

Code: https://github.com/michael-garciaperez/DSC180B-Capstone-Project

1 Introduction 2
2 Methods 3
3 Results 17
4 Conclusion & Discussion 19
References 20
Appendices A1

1 Introduction

In this paper, we delve into the impacts and consequences of data science, focusing on the role of recommendation systems in perpetuating the gender gap in directors within the film industry. Past research in sociology and media studies has extensively examined the gender gap prevalent in the film industry, particularly concerning the underrepresentation of female directors. A thorough analysis of over 2000 films released between 1994 and 2016 revealed that only 5 percent of directors were female, highlighting a significant gender disparity (Karniouchina et al. (2023)). Similarly, an examination of the gender composition of directors in top-grossing films spanning from 2007 to 2021 demonstrated a striking ratio of 11:1 male to female directors (Smith, Pieper and KHAN (2022)). While these studies support the evidence of a gender gap in movie directors, the implications of this disparity on recommendation systems have received limited attention. Given the prevalence of recommendation models in content distribution platforms, especially major distributors like Netflix, is it crucial to identify and mitigate biases associated with director gender in these systems. Bias in recommender systems impacts various stakeholders, ranging from users of content recommendation platforms, to film distributors. Recommendation systems shape the content users see and consume, so systems perpetuating gender bias in films limit the range of content users see. This also hinders the opportunities for female filmmakers to reach wider audiences and limits their films' visibility. Film distributors and production companies' use recommendation systems to promote their films and engage audiences. If these systems prioritize movies directed by men, male-dominated films receive more visibility and resources, further marginalizing female directors and limiting the diversity of content available to audiences.

Additionally, widely adopted bias mitigation tools (such as IBM's AI Fairness 360 (AIF360) toolkit) are optimized for classification tasks. Research into ethical biases within recommender models remains limited, with most studies focusing on statistical biases like popularity bias. This study aims to investigate these gaps and develop strategies to foster fairer movie recommendation systems by minimizing biases linked to director gender.

1.1 Literature Review

Research on fairness in recommendation systems delves into fairness through a statistical lens. Li, Yunqi and et al.'s recent research in 2023, their work *In Fairness in Recommendation: Foundations, Methods and Applications*, acknowledges that unfairness in gender may be caused by biases in the training data, but do not go into results from bias mitigation methods (Li et al. (2023)). They do disclose in the methods that pre-processing is a method that aims to mitigate the bias in the data so that any model learned from that data will perform more fairly. However, they strictly denote that such methods are classification bias mitigation techniques, despite the paper being about recommender systems. The authors also disclaim that it is possible the accuracy may worsen after using pre-processing techniques.

Research that examines demographic bias, like Ekstrand et al.'s *All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness*

analyzes it through the users of the recommender model, rather than the items (Ekstrand et al. (2018)). Breaking down the age and gender of the users, the paper looks at their recommender system's performance across female versus male users, and the different age groups. Yet, there is no consideration in how the items being recommended may be biased. There is no examination in the different groups of movies (e.g., genres, years, directors) but rather the user of the model. Additionally, the paper uses utility metrics to measure performance across different groups of users, and does not consider fairness metrics.

1.2 Description of Relevant Data

The ratings come from MovieLens, which contains user ratings and demographic data and movie metadata collected from the MovieLens website, a platform dedicated to personalized movie recommendations through user feedback and ratings of movies. These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined the platform in 2000. The ratings data includes unique numerical identifiers for each movie, unique numerical identifiers for each user, and a discrete number ranging from 1-5 to represent star ratings. The dataset also has another file dedicated to the movie data, including the title of the movie, year of release, list of genres, and its corresponding unique numerical identifier.

Because the MovieLens dataset contains no information about the directors of the movie, Northwestern University's lab data and research on "U.S. movies with gender-disambiguated actors, directors, and producers" has the complete gender breakdown for members such as cast, directors, and producers for over 10,000 U.S.-produced movies released between 1894 and 2011. AmaralLab's data contains the name of each director as well as their gender (as mentioned in their individual biographical pages), which is identified by its IMDb unique alphanumeric identifier. To link the gender information back to the MovieLens dataset, this IMDb identifier must be present, so it is necessary to use IMDb's title basics dataset. This comprehensive collection of metadata for various titles available on the IMDb platform includes the alphanumeric unique identifier of the title and the name of the title.

2 Methods

The methodology of this project was done in the several steps outlined below.

2.1 Exploratory Data Analysis (EDA)

2.1.1 Feature Creation

We created 'director_gender_proportion' to denote the proportion of directors for that movie. Movies can have more than one director, so we cannot simply assign each movie to whether

it is directed by a male or female. We create a new column that denotes the proportion of directors for each movie that are male. A movie directed fully by males will have a 'male_director_proportion' of 1.0, and a movie directed fully by females will have a 'male_director_proportion' of 0.0.

To prepare the dataset for model development of prediction ratings, 'male_director_proportion' is converted to a binary representation. Binarization involves categorizing the variable into two distinct groups based on a specified threshold. To determine this threshold, we examined the distribution of the 'male_director_proportion' across the data and saw that most movies were fully male-directed (i.e., had a 'male_director_proportion' of 1.0). Binarizing the variables consists of converting the proportions to binary values of either 0 or 1. We considered rounding the proportion, but that would result in only a few data points being 0. There would be a strong imbalance between 0s and 1s due to the overrepresentation of 1s. We convert the value into 1.0 if the movie is fully directed by males, and 0.0 if the movie involves at least one female director. This results in a less imbalanced distribution of 0s and 1s, compared to rounding the proportions.

To reiterate, 'male_director_proportion' was the ratio of male directors out of the total directors for that movie. 'all_male_director' is a value of 0.0 or 1.0: 1.0 if the movie is entirely male-directed, 0.0 if not (i.e., female directors are involved).

2.1.2 Distributions of Rating Data

The recommender system looks at how users compare across movie ratings by filtering out items that a user might like based on similar users. The distribution of rating scores shows users' tendencies when rating movies. Most ratings are on the higher side with a score of 4. There are less low ratings of 1 and 2 in the dataset.

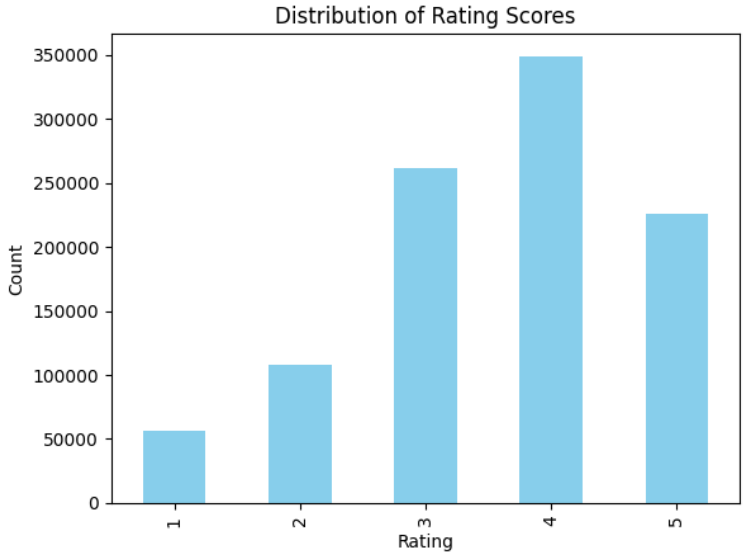

Figure 1: Distribution of Rating Scores

We further separate this by the director's gender breakdown to compare the proportion of ratings between male versus female directors. Movies with all male directors (i.e., a 'male_director_proportion' of 1.0) received a greater proportion of 5 star ratings. This leads us to define the rating of 5 stars as a high rating, using this to identify bias in the later sections to determine whether fully male-directed movies are advantaged by our model to receive a prediction of 5 stars.

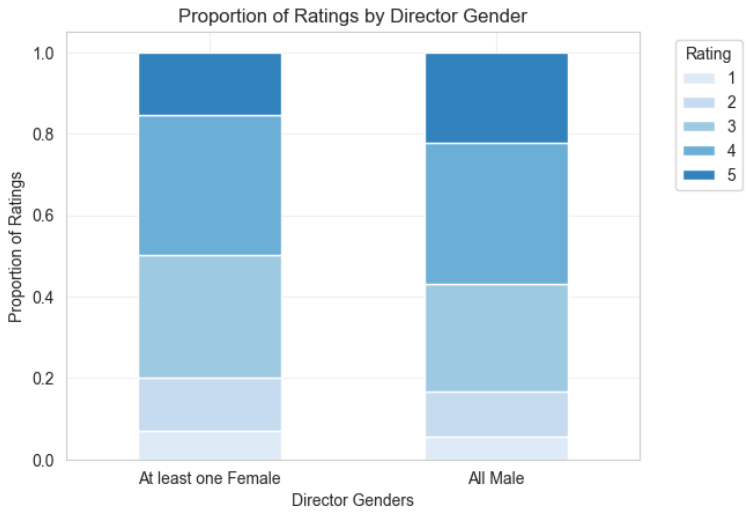

Figure 2: Distribution of Rating Scores by Director Gender

2.1.3 Distributions of Director Gender

By examining the distributions of the directors' genders in the dataset, we examine if there is any gender imbalance. Given societal context, it is expected for most movies to be directed by males. This is reflected in the dataset with a severe imbalance between the proportion of movies directed by males versus females.

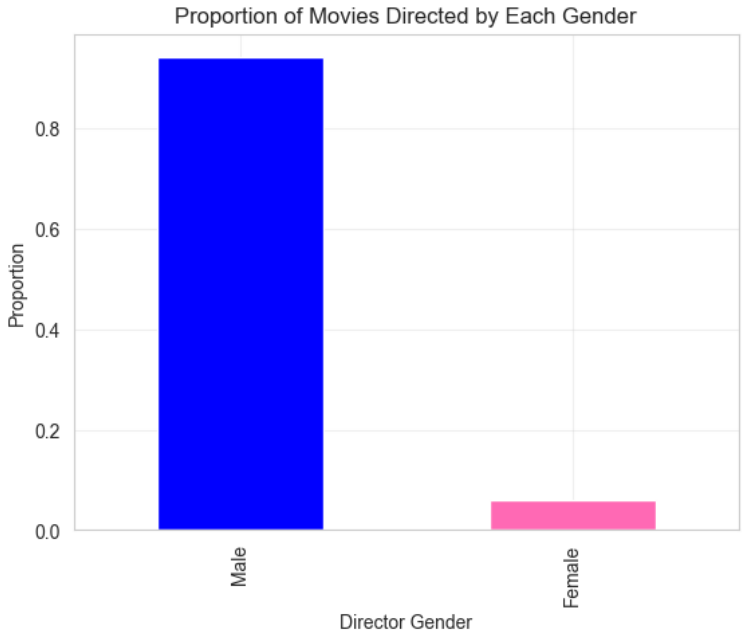

Figure 3: Proportion of Movies Directed by each Gender

While it is important to acknowledge the strong class imbalance within the dataset, training a linear regression model to predict movie ratings did not reveal a substantial discrepancy in accuracy between male-directed and female-directed movies.

The model demonstrated comparable performance in predicting ratings for both director genders. The mean squared error (MSE) for predicting ratings of female-directed movies was only slightly lower than that for male-directed movies. There is not a significant difference in accuracy between the different director genders.

Because of this similar performance in model accuracy across director genders, plus the time-constraints of the project and limited resources, we decide not to correct the class imbalance. This imbalance does not significantly impact the predictive capabilities of the linear regression model.

Our model utilizes features such as the genre of the movie and year to determine whether a user is likely to give that movie a high rating, based on their previous ratings for movies of similar features. So, we examine the distribution of how movie genres and years compare across the different director genders.

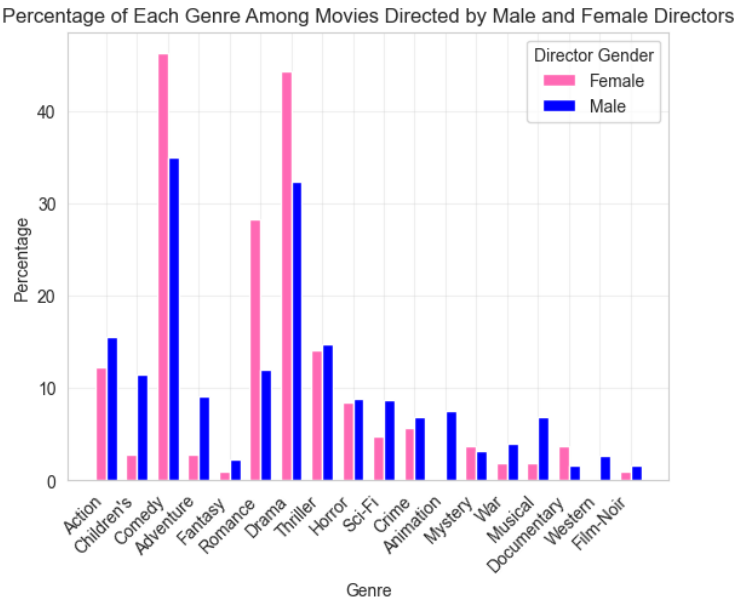

Figure 4: Percentage of Each Genre by Director Gender

Certain genres, such as animation and western, are dominated by male directors. Users who predominantly rate movies within these genres are likely to receive higher predicted ratings for male-directed movies due to the disproportionate representation of male directors.

2.2 Fairness Evaluation

We designate a perfect star rating of 5 as the favorable outcome, see Figure 2. Movies that are fully directed by males received a greater proportion of 5 star reviews compared to the other proportions, so we acknowledge movies fully directed by males as the privileged group within our analysis.

Our Disparate Impact value of 0.6936795435950334 (before mitigation) shows an imbalance in the favorable outcome between male and female directors within the dataset. Fully male-directed movies are more likely to achieve the favorable outcome compared to movies with female directors, as movies with female directors are getting 5-star ratings about 69 percent as often as fully male-directed movies. The 3/4ths rule is a standard threshold used for disparate impact analysis. According to this rule, if the favorable outcome rate for a group is less than 75 percent (also know as 3/4 as a fraction), there is a strong indication of bias being present in the dataset.

Our Statistical Parity Difference value of -0.06810877956209874 (before mitigation) indicates a difference in the proportion of favorable outcomes between fully-male directed movies and movies with female directors. The negative value suggests that movies with female directors experience a lower rate of favorable outcomes compared to fully-male directed movies. It implies that fully male-directed movies receive a higher proportion of 5-star ratings than female-directed movies. A Statistical Parity Difference close of 0 shows

no difference in the proportion of favorable outcomes between different groups though, so our value does not show a discernible bias.

2.3 Model Development

2.3.1 Classification

Given that AIF360, IBM's widely used bias detection and mitigation tool, is used for classifiers, we create a classification model. We use a Random Forest Classifier to predict whether a user will give a movie a perfect rating (5 stars) based on various movie attributes, including the year of release, genre, and whether or not the movies is fully male-directed. The model is trained and evaluated on this binary classification task, and additional focus is placed on assessing potential biases related to the protected attribute 'all_male_director' to ensure fair and unbiased predictions. We perform hyperparameter tuning, calculate fairness metrics, and report on both the model's predictive accuracy and fairness considerations.

The ROC-AUC Curve visualizes the trade-off between true positive and false positive rates and ranges from 0 to 1.

A higher ROC-AUC indicates better separation between the positive and negative classes. A typical rule of thumb for interpretation maps the following values:

- 0.5-0.6: Poor
- 0.6-0.7: Fair
- 0.7-0.8: Good
- 0.8-0.9: Very good
- 0.9-1.0: Excellent

Our AUC of 0.72 suggests a moderate discriminatory power of our model.

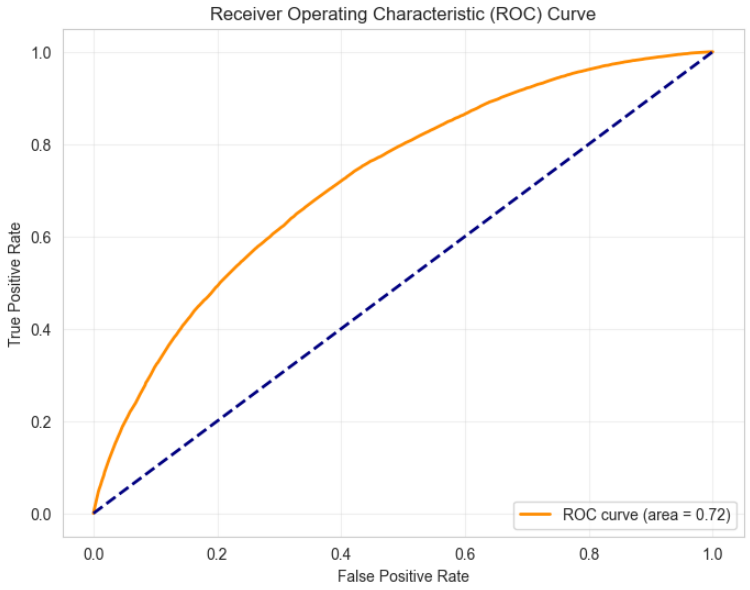

Figure 5: Receiving Operating Characteristic (ROC Curve)

The Precision-Recall Curve shows the precision-recall trade-off, which is often useful for imbalanced datasets, and the score also ranges from 0 to 1. Precision-Recall AUC is particularly useful when dealing with imbalanced datasets, focusing on the positive class.

Our AP of 0.42 suggests that the precision-recall trade-off is not very high.

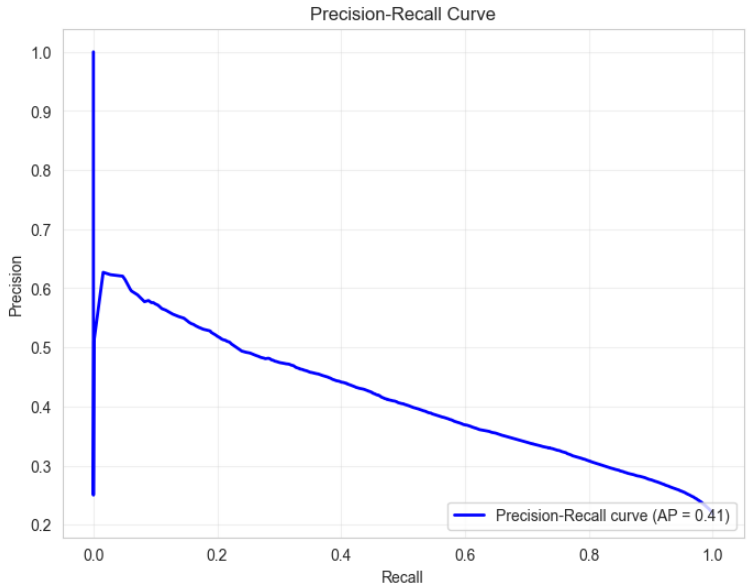

Figure 6: Precision-Recall Curve

The Feature Importances Bar Plot displays the importance of each feature in the trained Random Forest classifier model. The most important feature when predicting rating was

the year the movie was released.

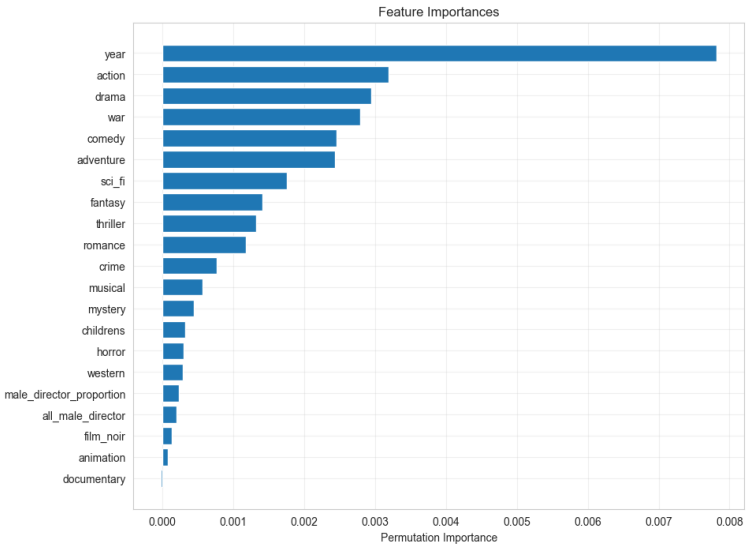

Figure 7: Feature Importance

To assess fairness metrics, we look at the classification model's Disparate Impact and Statistical Parity Difference, as we did above for the dataset.

Our Disparate Impact of 0.66751005405058 suggests that the unprivileged group is at a disadvantage compared to the privileged group, as they are less likely to receive favorable outcomes.

Our Statistical Parity Difference of -0.07432062114365237 is not a strong indicator of bias due to this value being close to 0.

The model's utility is measured through accuracy, which receives a fairly high score of 0.7866022619775142.

2.3.2 Recommendation through Similarity

Our research is in the domain of gender bias in recommender systems, so we create several recommender systems with the same task of recommending movies in our dataset. Using different similarity metrics like Jaccard Similarity, Cosine Similarity, and Pearson Correlation. These metrics are commonly used in recommendation systems to gauge the similarity between user preferences or item attributes.

Essentially, we generate three distinct recommender models, each leveraging a different similarity metric. Despite their differences in approach, all models are trained on the same dataset of user ratings, enabling a comparative analysis of their performance and the potential influence of gender bias in their recommendations.

To examine patterns in the director gender breakdown of recommended movies, we look at the recommendations for several different users. We start with looking at the top 10 movies

recommended for User 1.

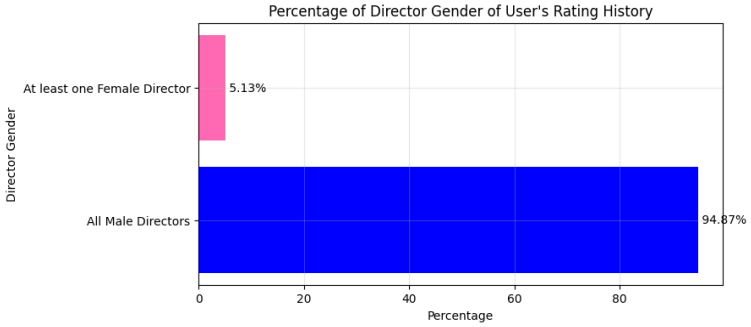

Figure 8: Director Proportion of User 1's Watched Movies

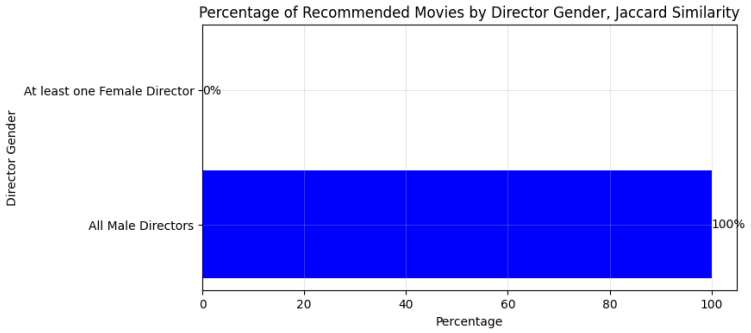

Figure 9: Director Proportion of Jaccard Similarity Recommended Movies

The recommender system using Jaccard Similarity had a lack of diversity. The user had a small portion of their movie ratings be movies featuring a female director, but the top movies recommended by the system have absolutely zero female directors.

The recommender system using Cosine Similarity featured the same results.

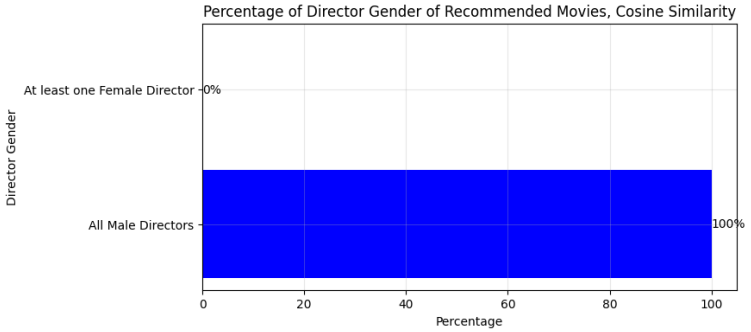

Figure 10: Director Proportion of Cosine Similarity Recommended Movies

Recommendations generated by the system using Pearson Correlation exhibited a slight increase in diversity concerning the gender of movie directors compared to the other similarity metrics. Specifically, the recommended movies exhibited an average proportion of

female directors of 20 percent, a notable increase compared to the 5 percent average observed within User 1's watch history. This suggests that the recommender system that uses Pearson Correlation yielded a more inclusive selection of movies, showcasing a greater representation of films directed by women.

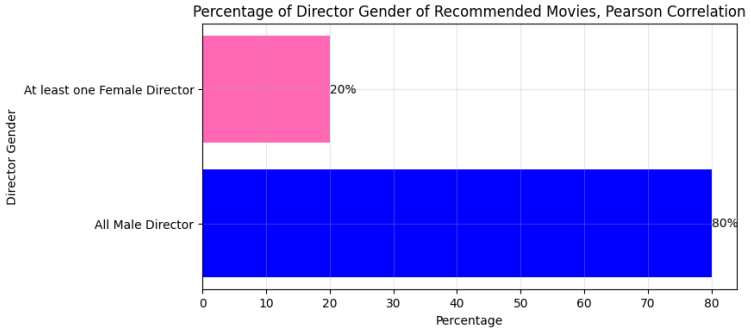

Figure 11: Director Proportion of Pearson Correlation Recommended Movies

While the findings for User 1 suggest that recommender systems using Pearson Correlation yield a more inclusive selection of movies, we examine other users to see if this is a typical trend. To determine which user to examine next, we select the user who has rated the greatest proportion of movies with female directors.

User 2908 had the greatest representation of female directed movies in the dataset, as approximately 44 percent of the movies they have rated featured female directors.

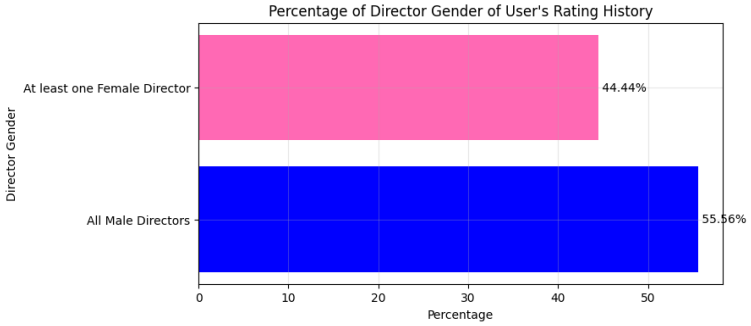

Figure 12: Director Proportion of User 2908's Watched Movies

The recommender system using Jaccard Similarity had absolutely no movies with female directors. Every movie recommended was entirely male-directed.

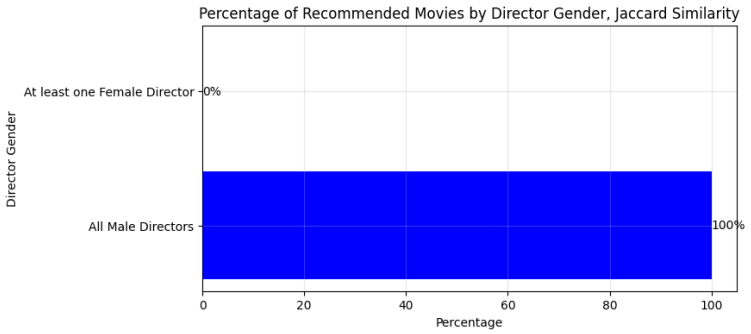

Figure 13: Director Proportion of Jaccard Similarity Recommended Movies

The recommended movies using Cosine Similarity were more diverse than the recommended movies using Jaccard Similarity, but still had a lot less movies featuring female directors than User 2908's rating history.

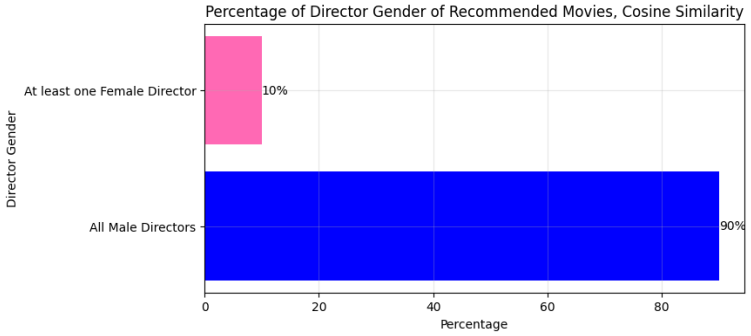

Figure 14: Director Proportion of Cosine Similarity Recommended Movies

The recommended movies using Pearson Correlation performed the same as Jaccard Similarity with no movies featuring female directors.

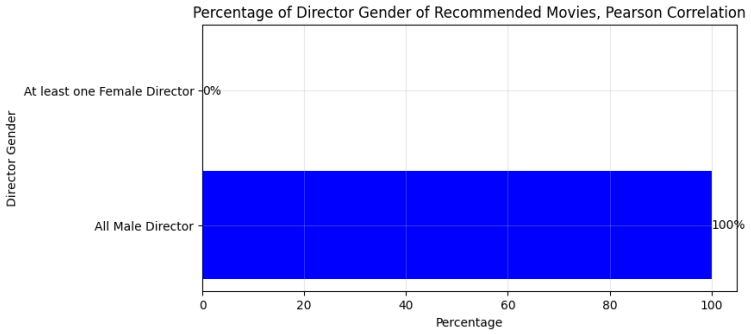

Figure 15: Director Proportion of Pearson Similarity Recommended Movies

For User 1, the recommender system using Pearson Correlation yielded the greatest percentage of female directed movies (20 percent), whereas it was Cosine Similarity (10 percent) for User 2908. User 1 received more diverse recommendations even though they had

watched less female directed movies than User 2908. It is inconclusive which similarity function yields in the least biased recommendations since it varies per user.

Looking at the matrix that the recommender systems use, they only take into account user ID, movie ID, and the user rating. It does not take into account the director's gender, hence why we investigate in the next section whether bias mitigation techniques will work on recommender systems using these similarity metrics.

2.3.3 Recommendation through Singular Value Decomposition

From the Surprise machine learning toolkit, Singular Value Decomposition (SVD) is a Matrix Factorization-based algorithm that often captures latent factors underlying user preferences. The recommender model uses an SVD object, which is trained on the training dataset to optimize predictive performance using user-item interactions.

To assess the predictive accuracy and generalization capability of the model, we use cross-validation. We perform k-fold cross-validation, with k set to 5 for robustness. The model's performance is evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These metrics provide insights into the model's ability to accurately predict user ratings and its overall performance across different folds of the dataset.

The model's Disparate Impact value of 0.3693759561648962 indicates a significant disparity in favorable outcomes between movies directed by males and those directed by females, suggesting unequal representation in the predicted ratings (recommended) items. The model's Statistical Parity Difference of -0.05460790391117668 fails to indicate any bias.

To measure model utility, we look at (Root Mean Squared Error) and (Mean Absolute Error). RMSE tends to penalize large errors more than smaller ones due to the squaring operation. It is useful for comparing models and understanding the spread of errors, but it is not very intuitive for human interpretation.

The model's RMSE value of 0.7146405963914735 before mitigation signifies the average discrepancy between predicted and actual ratings, reflecting the prediction accuracy of the model in prioritizing entirely-male directed content.

MAE, on the other hand, provides a more straightforward and human-friendly interpretation. It represents the average magnitude of errors without considering their direction (i.e., whether they are overestimations or underestimations). In other words, MAE gives the average absolute error, making it easier to understand/interpret in real-world terms. For example, if the MAE is 5, it means, on average, that the model's predictions are off by 5 units from the actual values.

With an MAE value of 0.5604532156766929 before mitigation, the model's average deviation from actual ratings is approximately 0.56 units, allowing insight into prediction accuracy without considering the direction of errors.

To convert the predicted rating scores into recommendations, we recommend the user the movies that are predicted to receive 5 star ratings.

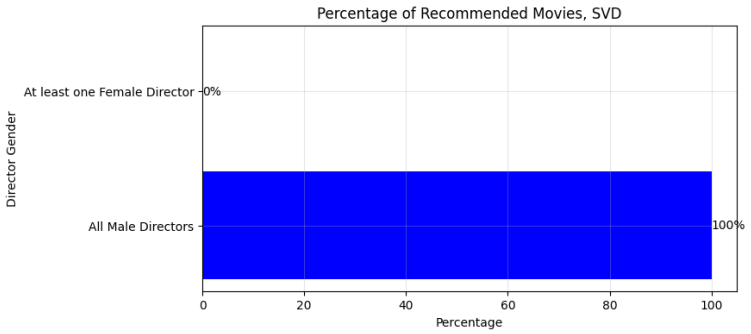

Figure 16: Percentage of Recommended Movies from SVD for User 1

Here, we observe that our model does not recommend any movies directed by females. This is because no movies directed by females were given a 5-star prediction in our SVD algorithm for this user. We would need to analyze more users to get a better picture of bias in our model's recommendations, but for now, we can observe that there is room for bias mitigation. It is also vital to note that this user's watch history only includes 5 percent female-directed movies.

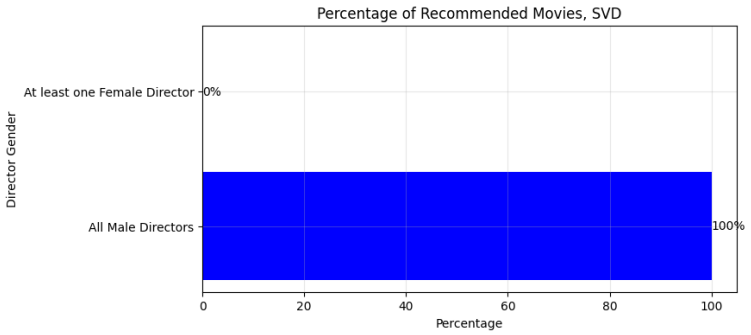

Figure 17: Percentage of Recommended Movies from SVD for User 2908

We also observe here that our model does not recommend any movies directed by females. Initially, we expected that this user would be given fairer recommendations, considering that their watch history includes 44 percent female-directed movies. However, when considering that SVD tries to predict ratings based on how other users have predicted the movie, we can see that bias from various users can influence what is recommended to other users, as seen with user 2908's diverse watch history, ultimately leading to highly biased predictions.

Note that if no 5-star predictions are available, then we would recommend 4-star rated movies to the user based on the predicted ratings from the SVD algorithm. For a more efficient recommendation, we believe that it would also be a good idea to suggest movies where the predicted rating is not rounded. For example, we would have predicted value ratings like 4.9, 4.8, etc., and would only recommend the highest top 10 ratings to users. This would be valuable since it would allow us to prioritize recommendations.

2.4 Bias Mitigation Techniques

2.4.1 Bias Mitigation in Classification

To mitigate bias, the Reweighing technique from IBM's AIF360 toolkit was applied to the dataset used in the Random Forest model.

This pre-processing technique adjusts the weights of instances in the dataset to ensure fairness across different groups. Reweighing can ensure our models are not biased towards the majority class, which improves the training process and overall performance for the minority class. In addition to class imbalance, Reweighing also factors in protected attributes and adjusts instance weights to mitigate biases related to that attribute, which in our dataset, is director gender.

Initially, a Random Forest Classifier is trained on the original dataset, and its predictions are evaluated for bias using metrics like Disparate Impact and Statistical Parity Difference. Subsequently, the dataset is reweighed using the Reweighing technique, which adjusts the instance weights to mitigate bias. A new Random Forest Classifier is then trained on the reweighed dataset, and its performance and bias metrics are assessed. This approach aims to address potential bias in the original model's predictions by reweighing instances based on the protected attribute 'all_male_director' before retraining the model.

2.4.2 Bias Mitigation in Recommendation

AIF360's reweighing technique helps mitigate bias present in the original dataset by adjusting the weights of instances based on sensitive attributes such as director gender. By doing so, it aims to ensure fairness by reducing the influence of biased features on the learning process.

The original dataset exhibits bias towards male-directed content through the underrepresentation of female directors, leading to a disproportionate representation of such movies in the recommendations. AIF360's reweighing technique helps mitigate this bias by adjusting the weights of movie instances based on director gender, ensuring a more balanced representation of both male and female-directed content.

The similarity metrics used in our recommender models above (Jaccard, Cosine, and Pearson correlation) rely on the dataset to calculate similarities between movies. Since the dataset exhibits bias towards male-directed content, these similarity metrics also favor male-directed movies in recommendations. By applying reweighing, the transformed dataset aims to mitigate such bias, ensuring that similarity calculations consider a more balanced representation of movies directed by individuals of different genders. As such, we develop the models again using the same similarity metrics on the transformed dataset.

Like the other similarity metrics above, SVD relies on user-item interaction data to make recommendations. Because the data exhibits bias towards male-directed movies, collaborative filtering models reinforce this bias in recommendations. By using a reweighted dataset, which adjusts the representation of male-directed and female-directed movies, collabora-

tive filtering models can mitigate this bias, ensuring that recommendations are based on a more equitable distribution of director genders.

3 Results

Table 1: Random Forest Classifier Results

	Disparate Impact		Statistical Parity Difference		Accuracy	
	Before	After	Before	After	Before	After
Random Forest Classifier	0.67	1.00	-0.07	0.00	0.79	0.79

After retraining on the reweighed data, the Random Forest Classifier with 100 estimators achieved an accuracy of approximately 0.79, which did not increase or decrease the accuracy compared to the original accuracy of approximately 0.79.

The bias metrics improved significantly, with the Disparate Impact being approximately 1 and the Statistical Parity Difference close to zero, indicating a fair model.

This suggests that the reweighing technique successfully mitigated bias, resulting in a fairer model while maintaining similar predictive accuracy.

Reweighing as a pre-processing bias mitigation technique was successful for the Random Forest Classifier, but remains limited in its application to recommender systems.

Despite applying the reweighted dataset to the recommender models utilizing Jaccard and Cosine Similarity, and Pearson Correlation, there was no difference in the diversity of movie recommendations. The top movie recommendations continued to exhibit a dominance of male-directed content, indicating that the reweighing technique did not effectively mitigate gender bias in these particular models. It appears that reweighing the dataset did not impact the calculation of the similarity metrics, resulting in the exact same top 10 movie recommendations before and after bias mitigation in the model.

Table 2: SVD Predictions Results

	Disparate Impact		Statistical Parity Difference		RMSE		MAE	
	Before	After	Before	After	Before	After	Before	After
SVD	0.37	0.77	-0.05	-0.00	0.71	1.21	0.56	0.99

However, reweighing the predictions using SVD was able to yield a Disparate Impact closer to the value of 1, effectively mitigating some bias in the original model. The model utility decreases though, making reweighing not the ideal choice due to lower predictive performance.

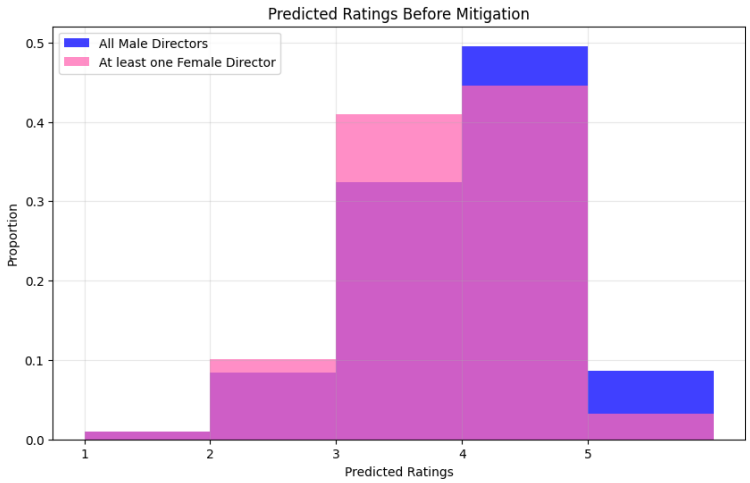

Figure 18: Predicted Ratings using SVD before Mitigation

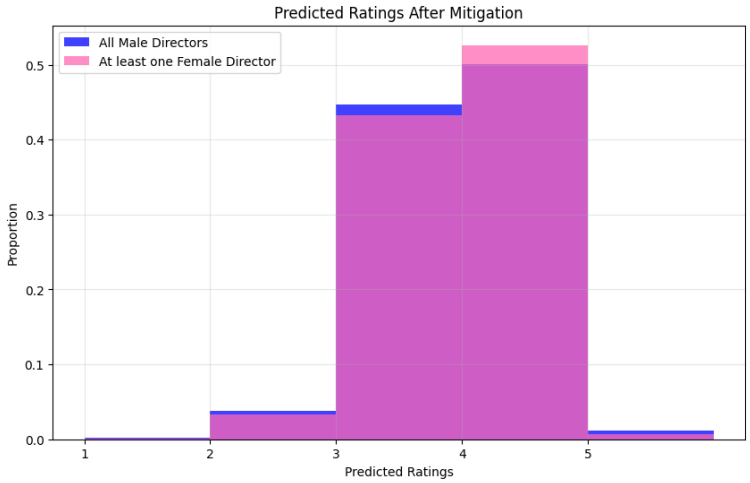

Figure 19: Predicted Ratings using SVD after Mitigation

After applying Reweighing, the amount of 5 star ratings (perfect ratings, the favorable outcome) became more equal between entirely-male directed movies and movies featuring female directors.

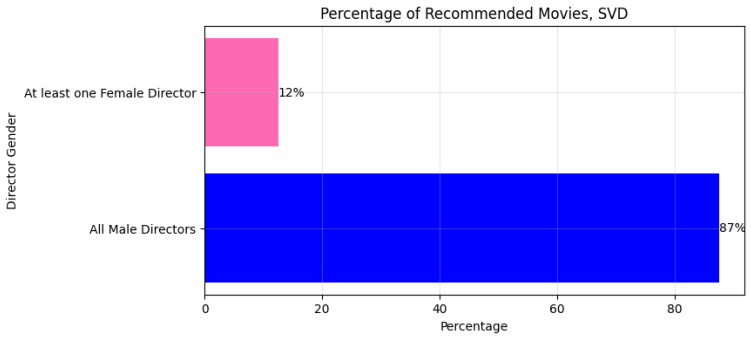

Figure 20: User 1's Recommendations using SVD after Mitigation

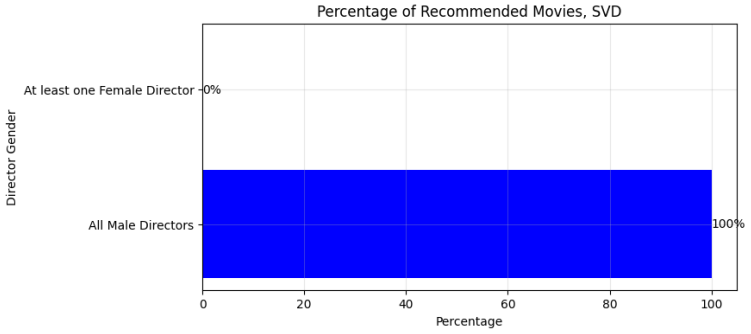

Figure 21: User 2908's Recommendations using SVD after Mitigation

Yet, after these predictions are converted to recommendations, there appears to still be a gender bias that favors entirely male-directed movies. Interestingly, User 1 received a greater proportion of movie recommendations featuring female directors, even though User 2908 has a greater proportion in their watch history. Also note that due to the nature of the SVD algorithm, predictions vary each time the code is ran, so the 12 percent is not a fixed percentage.

4 Conclusion & Discussion

The efficacy of AIF360's Reweighing technique on bias mitigation in recommender systems is subject to ongoing discourse. The topic of recommender systems spans many algorithms and various implementations, and although Reweighing made no difference in the similarity metric based recommender systems, the predictions from SVD was able to see a substantial bias mitgiation. Yet, when the SVD predictions are converted to recommendations, there continues to be a dominance of entirely male-directed films being recommended. This means that there is potential for AIF360 as a bias mitigation tool in recommender systems, although the success of bias mitigation techniques depends on the underlying calculations of the recommender system.

The basis of reweighing lies in addressing dataset imbalances. Because of the underrepresentation of female film directors in the dataset, reweighing adjusts the distribution of instances across different groups. This may lead to a reduction in the sample size of the majority group to achieve parity, which loses data and can limit the model's ability to capture the full complexity of the dataset.

The definition of "bias" is also arbitrary in our decision to use Disparate Impact and Statistical Parity Difference as the metrics of measurement. As seen in the above sections, the Statistical Parity Difference remained close to 0 in all models, failing to indicate bias. However, the Disparate Impact was able to capture the bias due to failing the 3/4ths rule. Certain metrics are better at capturing bias than others depending on the dynamics of the dataset. So, when it comes to bias mitigation, it depends on the context to decide which bias measures to look at. Even though Disparate Imapct was able to become more fair in the SVD algorithm after mitigation, it is possible other fairness metrics do not.

Additionally, simply examining the director's gender in movie recommender systems is not enough to consider for a model that can impact so many different stakeholders. It is essential to consider how gender can intersect with other demographic factors. For example, a movie recommender system can exhibit different biases based on both the gender and race of the director. The recommender may perform differently for a movie by a Black male director versus a movie by a white male director, despite both movies featuring a male director. As such, intersectionality with gender must be examined as different demographic groups may result in different bias mitigation strategies being the most effective.

Future research in bias mitigation in recommender systems should explore other popular recommender models, like libraries and algorithms such as implicit feedback, PyTorch, or TensorFlow. Ultimately, the goal is to achieve effective bias mitigation in a model with good to exceptional utility. From our research, we believe that bias is not typically mitigated in recommender systems because of how it decreases utility, as seen in the SVD algorithm. Nonetheless, we believe our research is sufficient to continue exploration in using more bias metrics, other models, and other bias mitigation techniques like in-processing and post-processing.

Lastly, while the bias mitigation techniques in this project were able to achieve a "fair" model in terms of statistical fairness metrics, these techniques do not fully address the underlying systemic biases present in society. The underrepresentation of female-directed movies in the dataset stems from gender biases present in the film industry, where women face more discrimination and receive less support. Technosolutionism refers to the reliance on technological solutions to address complex social issues, ignoring broader societal, ethical, and political implications. Bias mitigation in recommender systems is an important step in achieving fairness and equity in movie recommendation systems, but bias mitigation alone cannot effectively address biases without the work of fields such as sociology, cultural, and media studies to work towards changing the dynamics that influence film production, distribution, and consumption.

References

Ekstrand, Michael D, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness." In *Conference on fairness, accountability and transparency*. PMLR

Karniouchina, Ekaterina V, Stephen J Carson, Carol Theokary, Lorien Rice, and Siobhan Reilly. 2023. "Women and minority film directors in Hollywood: performance implications of product development and distribution biases." *Journal of Marketing Research* 60(1): 25–51

Li, Yunqi, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. "Fairness in Recommendation: Foundations, Methods, and Applications." *ACM Transactions on Intelligent Systems and Technology* 14(5): 1–48

Smith, SL, K Pieper, and AB KHAN. 2022. "Inclusion in the Director's Chair: Analysis of Director Gender & Race/Ethnicity Across 1,500 Top Films from 2007 to 2021."

Appendices

A.1 Bias mitigation A1

A.2 Pre-processing A2

A.3 Disparate Impact A2

A.4 Statistical Parity Difference A3

A.5 Recommender System A3

Definitions for several techniques used throughout the project can be found below. Please refer to this section throughout the paper if needed.

A.1 Bias mitigation

Bias mitigation refers to the process of identifying and addressing biases present in the data, algorithms, or models. A bias is an unfair treatment or disadvantage for certain individuals or groups. Bias can arise due to various factors, including historical inequalities, data collection methods, or algorithmic design. The goal of bias mitigation is to detect and rectify bias so that models can produce fair and unbiased results.

A.1.1 Protected and Non-Protected Groups

Protected and non-protected groups are the different demographic categories within a dataset, often defined by characteristics such as race, gender, age, or ethnicity. It is important to identify these groups to assess for fairness.

A protected group consists of individuals who share a particular demographic characteristic that exhibits a need to be protected from discrimination. Individuals in the protected group typically have experienced disadvantages or unfair treatment. Examples of demographics in protected groups include racial minorities, women, older adults, and individuals with disabilities.

A non-protected group represents the majority or dominant demographic within a given population and does not face the same discrimination. For example, individuals in a non-protected group may identify as male, white, young adults, or able-bodied.

A.2 Pre-processing

Pre-processing is part of IBM's AI Fairness 360 (AIF360) toolkit. Pre-processing is a set of techniques used to prepare the dataset for use in machine learning. Pre-processing techniques aim to ensure that datasets used for training machine learning models are representative, balanced, and free from biases that could lead to unfair outcomes in the model.

A.2.1 Reweighing

Reweighing is a pre-processing technique from AIF360. It adjusts the weights of samples in the data to mitigate biases and ensure fair treatment across different demographic groups. Reweighing balances the distribution of sensitive attributes within the data to reduce the impact of such attributes on the model's outcomes.

A.3 Disparate Impact

Disparate impact compares the selection rates of different groups. The selection rate is the proportion of individuals from a specific group who receive the favorable (desirable) outcome from the model. By comparing the selection rates of a protected group (e.g., minority group) to those of a majority group, disparate impact identifies disparities in treatment between groups.

Disparate impact is usually formatted as a ratio, known as the Disparate Impact Ratio (DI). The Disparate Impact Ratio is calculated as follows:

$$DI = \frac{\text{Selection Rate of Protected Group}}{\text{Selection Rate of Majority Group}}$$

Note that the Selection Rate of Protected Group is the proportion of individuals from the protected group who receive the favorable outcome of the model, and Selection Rate of Majority Group is the proportion of individuals from the majority group who receive the favorable outcome of the model.

A DI value less than 1.0 indicates that the selection rate for the protected group is lower than that of the majority group, suggesting bias.

A.3.1 3/4ths Rule

The 3/4ths rule is a standard threshold used to assess for bias in data science systems. If the DI is less than 0.75 (also known as 3/4), there is evidence of bias. This suggests that the the selection rate for the protected group is less than 75 percent of the selection rate of the majority group, indicating unfair treatment.

A.4 Statistical Parity Difference

Statistical Parity Difference (SPD) measures the difference in predictions between different demographic groups.

To calculate the SPD, the the probability of receiving a favorable outcome given membership in the non-protected group is subtracted from the probability of receiving a favorable outcome given membership in the protected group.

If the SPD is 0.0, it indicates no disparity in the outcome across different groups. Positive SPD values indicate favorable treatment of the protected group, while negative SPD values indicate favorable treatment of the non-protected group.

A.5 Recommender System

A recommender system is an algorithm that filters information (e.g., user and item data) to provide personalized recommendations or suggestions to users. By analyzing past user behaviors, preferences, and interactions, recommender systems aim to predict the likelihood that a user will be interested in a particular item, recommending users relevant content with a high predicted interest.

A.5.1 Jaccard Similarity

Jaccard Similarity measures similarity between two sets of items through the degree of overlap between the two sets.

Jaccard Similarity can be used to measure the similarity between two users or items based on their patterns of interaction. For example, the Jaccard Similarity between two users can be calculated based on the items they have both rated. Jaccard Similarity between two items can also be calculated based on the users who have interacted with both items.

The Jaccard Similarity $J(A,B)$ between sets A and B is calculated as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

- $|A \cap B|$ denotes the number of elements common to both sets A and B.
- $|A \cup B|$ denotes the total number of distinct elements in sets A and B.

In recommender systems, Jaccard Similarity is often used to identify similar users or items, which can then be used to make personalized recommendations.

A.5.2 Cosine Similarity

Cosine Similarity measures the similarity between two vectors in a multidimensional space. It calculates the cosine of the angle between the two vectors to measure their directional similarity.

Cosine Similarity can be used to measure the similarity between two users or items based on their feature vectors. For example, each item can be represented as a feature vector, and the Cosine Similarity can be calculated between these vectors to assess their similarity.

The cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is calculated as follows:

$$\text{Cosine\_Similarity}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Where:

- $\mathbf{A} \cdot \mathbf{B}$ denotes the dot product of vectors $\mathbf{A}$ and $\mathbf{B}$.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ denote the Euclidean norms of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively.

Cosine similarity calculates the similarity between user or item vectors, allowing the algorithm to recommend relevant items to users.

A.5.3 Pearson Correlation

The Pearson Correlation coefficient is a statistical measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between two variables.

Pearson Correlation can be used to measure the similarity between two users or items based on their rating patterns. For example, the Pearson Correlation between two users can be calculated based on their rating profiles for shared items.

The Pearson correlation coefficient r between variables X and Y is calculated as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i denote the ratings or scores of users i for items X and Y, respectively.
- $\bar{X}$ and $\bar{Y}$ denote the mean ratings of items X and Y, respectively.
- n denotes the number of items rated by both users for which ratings are available.

A.5.4 Singular Value Decomposition

Singular Value Decomposition (SVD) from the popular Surprise library is a matrix factorization technique commonly used in recommender systems to reduce the dimensionality of user-item interaction matrices. SVD is able to uncover latent factors underlying user preferences and item characteristics. A latent factor is an underlying abstract feature that is not directly observable but influences the interactions between users and items.

SVD decomposes the user-item interaction matrix $\mathbf{R}$ into three different matrices: a user matrix $\mathbf{U}$, a diagonal matrix $\mathbf{\Sigma}$ containing the singular values, and an item matrix $\mathbf{V}^T$. The goal is to approximate $\mathbf{R}$ as closely as possible. $\mathbf{R}$ is approximately defined as follows:

$$\mathbf{R} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Where:

- $\mathbf{U}$ represents the users in terms of latent factors.
- $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values, which represent the importance of each latent factor.
- $\mathbf{V}^T$ represents the items in terms of latent factors.

By reducing the dimensionality of the user-item interaction matrix, SVD captures the underlying patterns and dependencies in the data, enabling more efficient computation and bet-

ter recommendations. The latent factors discovered through SVD can be used to generate personalized recommendations for users based on their preferences and past interactions.