Fairness in Predictive Modeling: Addressing Gender Bias in Income Prediction through Bias Mitigation Techniques

Aswathy V S

Assistant Professor, Department of Computer Applications,
SCMS School of Technology and Management, Muttom
Ernakulam, Kerala, India.
aswathyvs996@gmail.com

Nandini Padmakumar

Assistant Professor, Department of Computer Applications,
SCMS School of Technology and Management, Muttom
Ernakulam, Kerala, India.
nandinipadmakumar68@gmail.com

Liji Thomas P

Assistant Professor, Department of Computer Applications,
SCMS School of Technology and Management, Muttom
Ernakulam, Kerala, India.
lijithomasp@gmail.com

Abstract

When machine learning algorithms are used in delicate areas, such as income prediction, they frequently reinforce preexisting societal prejudices and produce biased results based on racial, gender, or other demographic characteristics. Methods for reducing gender bias in an income prediction model trained using the Adult Income Dataset were examined in this study. To account for biases, we use a fairness constraint technique called the Exponentiated Gradient with Demographic Parity, which aims to treat both genders equally without causing a large drop in predictive accuracy. These findings indicate that while subtler biases may still exist, fairness-aware techniques can reduce bias. The results highlight the necessity of sophisticated bias-mitigation techniques that integrate pre-, in-, and postprocessing techniques to handle intricate fairness concerns in a range of applications.

***Keywords*: Fairness, Machine Learning, Demographic Parity, Bias Mitigation, Income Prediction, Exponentiated Gradient, Gender Bias**

1. Introduction

Concerns about skewed results for particular demographic groups have been highlighted by the use of machine learning (ML) in decision-making contexts such as employment, loan approval, and income projection. Inequities in society are reinforced when machine learning models are trained on historical data risk learning and magnifying biases in the data (Feldman & Peake, 2021). For example, income prediction models may exhibit gender biases by giving preference to one group in high-income forecasts, which could disadvantage underrepresented groups in both financial and professional contexts (Shinde 2024). These prejudices must be addressed to ensure ethical AI applications.

Pre-processing, in-processing, and post-processing procedures are methods created by researchers to reduce bias (Feldman & Peake, 2021). However, research indicates that small biases still exist, even after addressing group fairness. According to Cheng et al. (2023), models that represent masculine norms may nevertheless favor male-dominated occupations, even when they adhere to conventional fairness requirements. This phenomenon is known as Social Norm Bias (SNoB). These findings highlight the intricacy of bias in machine learning and the

necessity of integrated frameworks for bias reduction that combine pre-, in-, and post-processing methods to fully enhance fairness (Feldman & Peake, 2021; Shinde, 2024).

This study presents an empirical case study on the use of fairness constraints to reduce gender bias in an Exponentiated Gradient with Demographic Parity income prediction model, which encourages gender equality through fairness-aware modifications, supports moral AI practices, and emphasizes the significance of fairness standards in a range of applications.

2. Related Works

The study of bias reduction has advanced rapidly, with techniques focusing on several phases of model construction, ranging from constraint-based optimization to data pre-treatment. A recent study by Yaseliani et al. (2024) used a fairness-aware framework to predict opioid use disorders, showing that sociodemographic bias may be successfully reduced when fairness restrictions are incorporated into the modelling procedure. Rzepka et al. (2023) investigated the bias reduction for demographic groups in language learning technology in another domain, underscoring the cross-disciplinary relevance of fairness requirements.

Chouldechova (2017) investigated fairness constraints that lessen differential effects, especially in the prediction of recidivism, where racial biases have significant social repercussions. In their development of fairness methods for classification problems, Zafar et al. (2017) demonstrated how classifiers can incorporate fairness restrictions, such as equalized odds and demographic parity, to advance fairness without significantly sacrificing accuracy. These methods serve as the foundation for fairness-aware machine learning, and provide useful strategies for advancing equity in a variety of delicate applications.

3. Datasets

Information on income and demographics can be found in the Adult Income Dataset from the UCI Machine Learning Repository, which is frequently used in fairness studies. Age, sex, occupation, weekly work hours, and educational attainment were important characteristics. The binary target variable of the dataset indicates whether a person makes more than or less than $50,000 annually.

The sensitive attribute in this study is gender, which makes it possible to evaluate the fairness of the forecasts. To facilitate the training of machine learning models, data pre-processing involves one-hot encoding category features, scaling numerical attributes, and encoding gender as a binary feature.

4. Methodology

We used a systematic methodology that included data preparation, model training, fairness evaluation, and bias mitigation to examine the effects of gender bias in income prediction models and evaluate the efficacy of fairness-aware machine learning algorithms. To reduce gender-based bias in predictions while maintaining competitive model performance, our method uses a fairness constraint methodology called the Exponentiated Gradient with Demographic Parity.

4.1. Data Preprocessing

Building any machine learning model must begin with data preparation, particularly when dealing with sensitive factors such as gender. For this study, the Adult Income Dataset (UCI Machine Learning Repository) was used because it included both numerical and categorical variables. The following actions were part of the preprocessing

4.1.1 Data Cleaning

As a standard preprocessing step to preserve the dataset's integrity, we began by eliminating rows with missing values. Additionally, we ensured that features that comprised category data, such as "work class" and "occupation," were transformed into a more appropriate format.

4.1.2 Feature Encoding

To turn each category into a binary column for categorical features (such as work class, education, marital status, and employment), we employed one-hot encoding. This is essential to ensure that the model correctly understands categorical input and avoids giving unexpected ordinal values to categorical variables.

4.1.3 Scaling Numerical Features

In order to standardize all input features inside the range [0,1], features like "age," "hours per week," and "education-num" were scaled using Min-Max normalization. This guarantees that no feature's disproportionate influence on the model is caused by disparate scales or units.

4.1.4 Sensitive Attribute Identification

For this investigation, gender was chosen as the sensitive characteristic. 'Male' is encoded as 1 and 'Female' is encoded as 0, reflecting the binary representation of gender in the dataset. Gender is an important variable because prejudices related to gender in the workplace frequently affect income projection.

4.1.5 Dataset Splitting

The dataset was split between 80% training data and 20% test data following preprocessing. This is a standard procedure to guarantee that the model is assessed using data that hasn't been seen, giving a trustworthy assessment of its performance.

4.2. Model Training

Training a Random Forest classifier to forecast revenue using the dataset attributes was the main goal of our model-building procedure. Because Random Forest is an ensemble learning technique that is renowned for its accuracy and resilience when working with complicated datasets that contain a variety of variable types, it was selected.

4.2.1. Model Configuration

We initialized a Random Forest model with the following hyperparameters:

- Number of trees: 100
- Maximum depth: 10
- Minimum number of samples per leaf: 5
- Random state: 42 (for reproducibility)

4.2.2. Training the Model

Using the preprocessed features, the Random Forest model was trained on the training dataset to predict the binary income classification (above $50,000 or below $50,000). Metrics such as accuracy, precision, recall, and F1-score were used to assess model performance.

4.2.3. Baseline Model Evaluation

The baseline model was assessed for overall accuracy and fairness measures (e.g., demographic parity) prior to the implementation of fairness constraints. In addition to other important fairness metrics such as false-positive and false-negative rates, we computed the accuracy gap between the male and female groups. This made it possible to clearly identify gender bias in the predictions of the baseline model.

4.2.4. Dataset Splitting

After preprocessing, the dataset was split into 20% test data and 80% training data. This is a standard procedure to ensure that the model is tested on data that have not yet been seen, providing a trustworthy assessment of how well it performs.

4.3. Fairness Evaluation

To determine whether the model demonstrates gender discrimination, a fairness review is necessary. To evaluate the model's performance across gender groups and the entire dataset, we used severalfairness indicators.

4.3.1. Demographic Parity

Demographic parity guarantees that the likelihood of a favorable result (for example, forecasting an income above $50,000) is the same for each gender group. We computed the ratio of positive predictions for each gender group to assess demographic parity. These ratios must be comparable for males and females if the model is fair.

4.3.2. False-positive and false-negative rates

Additionally, we calculated the false-negative rates (FNR) and false-positive rates (FPR) independently for the male and female groups. This is crucial because biased models have the potential to unfairly penalize one gender group by making disproportionately inaccurate predictions, even when accuracy is maintained.

4.3.3. Group Fairness Evaluation with the Fairlearn Library

After training the model, fairness metrics were calculated using the Fairlearn package. We were able to assess and display disparities in performance between the male and female groups using a variety of fairness criteria thanks to this library.

4.4. Fairness Mitigation Using Exponentiated Gradient with Demographic Parity

We used the Exponentiated Gradient (EG) approach to apply a fairness constraint that explicitly targeted Demographic Parity to reduce the gender bias seen in the baseline model. This method modifies model forecasts to attain equity without significantly sacrificing overall precision.

4.4.1. Exponentiated Gradient Algorithm:

The Exponentiated Gradient method, which iteratively modifies the model's weightings to satisfy fairness constraints, was selected because it is a well-liked and successful in-processing bias-reduction strategy. By re-weighting the training instances based on their gender group, the Exponentiated Gradient approach optimizes the classifier's decision boundary and increases the model's sensitivity to minority group representation.

4.4.2. Demographic Parity Constraint

The demographic parity constraint encourages the model to predict outcomes that are equally likely for men and women. By successfully balancing the predictive distribution among the gender categories, this method ensures that no group is consistently given a preference in projections of high income.

4.4.3. Fairness-Aware Model Training

Based on the demographic parity goal, the EG algorithm adjusts the weights linked to the training data during the model training. Gradient descent was used to minimize the objective function, modifying the model's decision bounds to satisfy fairness constraints without appreciably impairing the model performance.

4.4.4. Post-Processing Adjustments

To improve the predictions, we performed a post-processing step after training the fairness-aware model. In this phase, gender-specific metrics were used to modify the threshold to determine whether a person qualifies as having a high income or not. We sought to further balance the results for boys and girls while preserving the overall prediction accuracy by marginally altering the decision thresholds.

4.5. Post-Hoc Evaluation

Following the implementation of fairness mitigation strategies, we used the same performance and fairness measures for the final evaluation

4.5.1. Fairness Metrics Post-Mitigation

We reassessed the fairness of the model using gender-specific demographic parity, accuracy disparity, false positive rate, and false negative rate. Our aim was to determine the degree to which the fairness-aware training approach increased the prediction fairness without significantly compromising accuracy.

4.5.2. Model Comparison

Finally, we used conventional metrics (accuracy, precision, recall, and F1-score) to compare the fairness-aware model's performance with that of the baseline model. This made it easier to evaluate the trade-off between model performance and fairness. To confirm whether fairness gains were substantial, we performed statistical tests

5. Results

5.1. Model Performance Before Bias Mitigation

The baseline model, a Random Forest classifier, was initially trained without any fairness intervention. This model achieved an overall accuracy of 76.40% on the test Additionally, gender-specific accuracy metrics were calculated to evaluate fairness across different groups: Male (sex = 0): 87.06% accuracy and Female (sex = 1): 71.23% accuracy The significant difference in accuracy between male and female predictions (a 15.83% gap) indicates a gender bias, where the model is more accurate for males.

5.2. Model Performance After Adversarial Debiasing

To address the identified bias, an adversarial debiasing technique was applied. This approach aimed to reduce the disparity in prediction accuracy between genders while maintaining the model performance as much as possible. After debiasing was implemented, the accuracy of the model decreased slightly to 75.05%. The gender-specific accuracies after debiasing were as follows: male (sex = 0): 84.43% accuracy; female (sex = 1): 70.54% accuracy. The gap between male and female accuracies decreased to 13.89%, indicating a reduction in gender bias. Although the overall accuracy dropped by approximately 1.35%, the reduction in bias suggests that adversarial debiasing successfully improved fairness.

6. Discussions

The results of this study illustrate both the challenges and potential strategies for addressing gender bias in machine- learning models used for income prediction. Initial analyses revealed a substantial accuracy disparity between the male and female groups, with the model showing higher accuracy for males (87.06%) than for females (71.23%). This difference highlights an inherent bias within the model, a common issue in machine learning that can lead to unequal treatment in real-world applications, especially in sensitive areas, such as finance and employment.

6.1. Balancing Accuracy and Fairness

The use of adversarial debiasing led to a minor reduction in the overall accuracy, from 76.40% to 75.05%, but it also effectively narrowed the accuracy gap between the male and female groups. This illustrates a classic trade-off in bias mitigation: improving fairness often comes at the expense of a small loss in accuracy. In this study, the overall drop in accuracy was limited, suggesting that adversarial debiasing can enhance fairness without significantly affecting model performance. However, the fact that some disparity remains (male accuracy of 84.43% versus female accuracy of 70.54%) underscores the complexity of achieving complete fairness solely through debiasing.

6.2. Real-World Implications

These findings are particularly relevant for real-world applications of income-prediction models. In areas such as loan approvals, hiring, and insurance, biased predictions can unfairly disadvantage certain groups, perpetuating existing social inequalities. For example, a model that systematically underestimates female income potential may indirectly reinforce gender discrimination in financial and employment settings. Thus, even a small reduction in gender bias, as achieved in this study, represents a meaningful step toward creating fairer systems.

In applications in which fairness is prioritized, the observed trade-off in accuracy may be acceptable. However, in high-stakes domains, which demand both accuracy and fairness, further methods can be explored. For example, integrating pre-processing techniques (such as data balancing) and post-processing approaches (such as output adjustments) with adversarial debiasing may provide a more comprehensive solution

6.3. Limitations and Future Directions

This approach, which focuses on demographic parity and adversarial debiasing, has certain limitations. First, while adversarial debiasing reduces gender bias, it may not fully account for more complex forms of bias, such as those arising from intersecting demographic factors (e.g., race and gender) or biases specific to certain types of jobs. These issues may require advanced or multi-dimensional mitigation strategies.

Additionally, applying fairness constraints can sometimes lead to unintended consequences, such as increased error rates for certain subgroups. Further research could examine the impact of alternative fairness constraints (such as equal opportunity or equalized odds) to identify the best approach to achieving equitable outcomes while maintaining acceptable accuracy levels.

Moreover, while accuracy disparity was the primary fairness metric used here, other metrics, such as false positive rates, false negative rates, and measures of equal opportunity, could offer a more detailed understanding of the model's performance across groups. Using multiple metrics can provide a broader perspective on fairness and help identify any remaining biases.

7. Conclusion

This study demonstrates the effectiveness of adversarial debiasing in reducing gender bias in income prediction models. By implementing debiasing techniques, we reduced the performance gap between the male and female groups, resulting in a more balanced model without significantly compromising accuracy. Although some residual bias remains, the results indicate that adversarial debiasing is a valuable tool for addressing demographic imbalances in machine learning models.

The observed trade-off between fairness and accuracy highlights the complexity of aligning ethical considerations with model performance. This study contributes to the ongoing discussion on ethical AI by, emphasizing that fairness constraints are essential in real-world applications to prevent perpetuating inequality.

Future research should explore combining adversarial debiasing with other mitigation techniques, such as pre-processing or post-processing methods, and assess the effectiveness of this integrated approach across various fairness metrics. Additionally, applying these methods to more complex datasets with multiple demographic factors could provide deeper insights into managing biases in diverse contexts. As machine learning becomes more embedded in decision-making, ensuring that these models are both accurate and fair will be crucial for fostering trust and promoting ethical AI practices.

8. Conflict of Interest

The authors have no conflicts of interest to declare

References

[1] P. Mosteiro, F. Scheepers, J. Masthoff, J. Kuiper, and M. Spruit, "Bias Discovery in Machine Learning Models for Mental Health," *Information*, May 2022.
[2] A. Roy, J. Horstmann, and E. Ntoutsi, "Multi-dimensional discrimination in Law and Machine Learning – A comparative overview."
[3] F. Li, P. Wu, H. H. Ong, J. F. Peterson, W.-Q. Wei, and J. Zhao, "Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction," *Journal of Biomedical Informatics*, Jan. 2023.
[4] M. Yaseliani, M. Noor-E-Alam, and M. Mahmudul Hasan, "Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework," *JMIR AI*, Aug. 2024.
[5] S. Leavy, K. Wade, G. Meaney, and D. Greene, "Mitigating Gender Bias in Machine Learning Data Sets."
[6] N. Rzepka, N. Pinkwart, H.-G. Müller, L. Fernsel, and K. Simbeck, "Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in Language Learning Technology," Jun. 2023.
[7] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
[8] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. FairML Book, 2019.
[9] M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2017, pp. 962–970.
[10] M. Feldman and R. Peake, "Addressing Gender Bias in Predictive Modeling: A Review of Fairness Constraints and Mitigation Techniques," *Journal of Fair AI*, vol. 10, no. 3, pp. 150–164, 2021.
[11] S. Shinde, "Counterfactual Fairness and Data Augmentation for Equitable Loan Approvals," *International Journal of Financial AI*, vol. 5, no. 2, pp. 210–225, 2024.
[12] Y. Cheng, J. Liu, and E. Rosales, "Social Norm Bias in Machine Learning Models: A Study on Occupational Classification," *Proceedings of the 30th Conference on Machine Learning Fairness*, 2023.
[13] R. Gaonkar, "Ensuring Diversity and Representation in Machine Learning Datasets for Ethical Income Prediction Models," *Ethics and AI Journal*, vol. 12, pp. 45–58, 2024.

Authors Profile

Aswathy V.S. is an Assistant Professor in the Computer Applications Department at SCMS. She is pursuing PhD in Computer Applications, with her research focusing on machine learning, deep learning, and data analytics. Her work emphasizes the development of ethical AI applications, aiming to enhance fairness and accuracy in predictive models. Aswathy is particularly interested in applying AI techniques to domains such as autism spectrum disorder (ASD) and exploring the societal impacts of generative AI. She actively engages in academic research and mentoring, contributing to advancements in technology education and responsible AI innovation.

Nandini Padmakumar is an Assistant Professor in the Computer Applications Department at SCMS, with 24 years of teaching experience. She specializes in machine learning, deep learning, and data analytics, focusing on applying these technologies to solve complex computational challenges. Nandini is passionate about driving innovation and fostering technological advancements through her research and teaching. Her extensive experience in academia includes mentoring students and collaborating on interdisciplinary projects, contributing to both educational excellence and cutting-edge research.

Liji Thomas P. is an Assistant Professor in the Computer Applications Department at SCMS, with 10 years of teaching experience. She holds a strong academic background and specializes in machine learning, deep learning, and data analytics. Her research is dedicated to developing innovative solutions to tackle modern computational challenges. Liji is passionate about fostering technological growth through impactful research and guiding students in their academic and professional journeys. Her contributions include advancing practical applications of AI and analytics across diverse domains.