Unwanted Bias in Machine Learning Algorithms

1st Mehrdad Soltani
Alfred Lerner Business School, University of Delaware
Institution of Financial Service Analytics
mehrdads@udel.edu

2nd Shihao Xi
Civil and Environmental Engineering
University of Delaware
xishihao@udel.edu

3rd Zherui Xu
Geography
University of Delaware
zxu@udel.edu

4th Corey Zhang
Data Science, University of Delaware
MSc, West Chester University of Pennsylvania
coreyz@udel.edu

***Abstract*—This study investigates the impact of machine learning models on income prediction and gender bias using the UCI Adult dataset. We compared the performance of three models: Logistic Regression, Decision Tree, and XGBoost. Our findings indicate that while XGBoost achieved the highest accuracy (87.26%) in predicting income levels above $50,000, it, along with the other models, exhibited significant gender bias, with accuracy rates for predicting gender based on income predictions exceeding 50%. To quantify variable importance and assess bias, we applied the Cohort Shapley technique, revealing substantial inherent biases in the models. These results underscore the necessity for advanced bias mitigation techniques in machine learning algorithms to ensure fair and equitable outcomes.**

***Index Terms*—Fairness in Machine Learning, Fairness, Gender Bias**

I. Introduction

In an era where machine learning algorithms play a pivotal role in decision-making across various sectors, the integrity and fairness of these automated systems have come under scrutiny. This paper delves into the critical examination of machine learning models, with a particular focus on their ability to predict income levels and the potential perpetuation of gender bias inherent in their predictions. Utilizing the UCI Adult dataset [1], we embark on a comparative analysis of three prominent models—Logistic Regression, Decision Tree, and XGBoost—to not only assess their predictive accuracy but also to scrutinize the extent to which they may inadvertently encode and amplify societal biases.

Our investigation is rooted in the broader context of fairness in machine learning, a field that grapples with the ethical implications of algorithmic decision-making. We explore the nuanced definitions of fairness, from group parity to individual and subgroup equity, and apply these frameworks to our analysis. By employing the Cohort Shapley technique, we aim to dissect the contribution of individual variables to the models' outcomes, offering a transparent lens through which we can observe and quantify bias.

This study is driven by the dual objectives of evaluating the models' performance in income prediction and identifying the presence of gender bias within their predictions. Through rigorous experimentation and analysis, we seek to contribute to the ongoing discourse on machine learning fairness, providing insights that could inform the development of more equitable algorithms. Our findings are intended to serve as a stepping stone for future research aimed at mitigating bias and fostering trust in machine learning applications.

This introduction sets the stage for the paper by outlining the research objectives, the importance of the study in the context of fairness in machine learning, and the methods used for analysis. It also hints at the potential implications of the findings for future research and algorithm development.

Identify applicable funding agency here. If none, delete this.

II. Related Work

A. Methods for Fair Machine Learning

Fairness in machine learning can be addressed at different stages: before training (pre-process), during training (in-process), or after training (post-process). The pre-process approach involves transforming data to remove underlying discrimination, such as removing features correlated with protected classes, as discussed by Calmon et al. [2]. The in-process approach modifies the learning algorithm to reduce bias, like adding a penalty in regression to discourage learning biased patterns [3]. The post-process approach adjusts predictions after training, effectively debiasing the output, as seen in debiasing word embeddings [4].

The research focuses on in-processing approaches, particularly for classification problems like credit approval and healthcare, where fairness issues frequently arise [5], [6]. Methods such as linear regression and logistic regression are commonly used to enhance fairness in these contexts [3]. Additionally, Samadi et al. introduced Fair Principal Component Analysis (PCA), a polynomial-time algorithm for finding a fair, low-dimensional data representation [7].

Other significant approaches include creating fair data using autoencoders [8] and adversarial debiasing algorithms, which introduce an adversarial component to promote fairness [9]. FairGAN, for example, generates unbiased synthetic data, serving as a pre-processing algorithm to prepare data for learning [10].

B. Definitions of Fairness

Fairness can be defined in terms of demographic group parity or individual constraints. Group fairness aims to ensure that statistical measures are equal across groups [11], [12], while individual fairness requires that fairness constraints apply to pairs of individuals [13]. Group fairness does not guarantee protection for individuals or subgroups, which can be problematic in high-stakes applications like medical care [14]. Individual fairness, while more targeted, is computationally costly and requires a suitable metric for comparison [15], [16].

Subgroup fairness combines aspects of both group and individual fairness to achieve better outcomes. It involves checking if group fairness constraints, such as equalizing false positives, hold over various subgroups [17]. This hybrid approach aims to address the shortcomings of traditional fairness definitions and improve fairness across both individuals and structured subgroups.

III. Data and Methods

A. Dataset

The dataset for this project comes from the UCI Adult dataset, credited to Ronny Kohavi and Barry Becker, derived from the 1994 United States Census Bureau data. It involves using personal information, such as education level, to predict whether an individual's annual income will exceed \$50,000. The dataset contains 14 input variables: a mix of categorical, ordinal, and numerical data types, as shown in Table I below. There are 48,842 rows of data, with 3,620 rows containing missing values, and missing values are denoted by a question mark (?). The binary classification task involves two class values: '$>$50K' (approximately 25%) and '$\leq$50K' (approximately 75%).

TABLE I
Summary of Dataset Features

Feature	Type	Description
age	Cont	Age of the instances
capital_gain	Cont	Capital gains
capital_loss	Cont	Capital losses
education_num	Cont	Highest education level
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes ¿ \$50K annually
marital status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
work class	Cat	Employer type

B. Models

Given the complexity of the input variables data type and the binary outcomes, three main models are adopted to build the connections between variables and output:

- **Logistic regression:** It's a statistical method for binary classification that models the probability of a binary outcome based on one or more predictor variables. It is a linear model that uses the logistic (sigmoid) function to map predicted values to probabilities.
- **Decision Tree:** It's a non-parametric supervised learning algorithm used for classification and regression. It builds a model in the form of a tree structure, where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents a class label (in classification) or a continuous value (in regression).
- **XGBoost:** Based on multiple decision trees, XGBoost is an optimized and highly efficient implementation of gradient boosting designed to be fast and performant. It is widely used for supervised learning tasks, including classification and regression, and has been a key algorithm for many winning entries in machine learning competitions.

Each model's parameters are optimized through grid search and cross-validation. Grid Search is a foundational technique for hyperparameter optimization in machine learning, offering a straightforward and systematic way to tune model parameters. When training the decision tree model and XGboost model, grid search was used to increase the performance and accuracy. Cross-validation is also used to assess the performance of a model in a more reliable way by dividing the data into multiple subsets (folds) and ensuring that each fold is used both for training and validation. In our project, grid Search incorporates cross-validation to evaluate the performance of each hyperparameter combination. This ensures that the hyperparameter selection is based on a robust estimate of model performance rather than on a single train/test split.

C. Experimental setup

The whole workflow of our project is shown in. First of all, three machine learning models are adopted to predict an individual's income level. Before training, an additional constraint is imposed to ensure that the prediction remains uninfluenced by a specific variable or set of variables (Z), which are designated for protection against bias. Here, (Z) represents the "protected variable[1]," playing a pivotal role in this context. Throughout the training process, the predictor ($\hat{Y} = f(X)$) is learned from a dataset comprising input data, the corresponding outcomes, and without the associated values of the protected variable.

For the subsequent step, an attempt is made to predict the value of the sensitive attribute from the predicted outcome using a new supervised machine learning method. If the value of the sensitive attribute can be predicted from the predicted outcome, it may indicate that the base prediction model was not fair, as the results could be biased. This approach will be applied to various machine learning methods, including XGBoost, logistic regression, and similar algorithms.

[1]In our case, sensitive attributes or protected variables are "Sex", "Race", and "relationship." Although we just focused on gender bias, we removed race also from our predictors. The relationship attribute has significant information about gender, as a result, we consider it as a sensitive attribute and we remove it.

D. Cohort Shapley Values

Mase and his co-authors introduced the cohort Shapley value, a variable importance measure rooted in the Shapley value from cooperative game theory (Shapley 1953), to quantify the impact of individual input variables on a black box function. Unlike other measures, the cohort Shapley value avoids creating unrealistic predictor combinations by using only observed data points. It forms a similarity cohort by including or excluding subjects similar to the target subject based on a variable and applies the Shapley value to the cohort averages. This method decomposes global sensitivity analysis (like ANOVA-based variance explained by Shapley) into components focused on uncertainty quantification, making it suitable for model auditing and understanding even when the model is not available or only single prediction runs are present [18].

The key distinction between SHAP [19] and cohort Shapley lies in their computation of conditional expectation Shapley; SHAP uses a joint distribution assuming feature independence, while cohort Shapley uses the empirical distribution. This difference, along with the axioms defined in SHAP and Chouldechova's definition of calibration [20], enables the application of cohort Shapley in fairness evaluations. The introduction of a squared cohort Shapley value, which splits previously studied Shapley effects over subjects, further enhances its applicability in assessing and ensuring model fairness.

IV. Results

A. Model Performance

The first objective of this project was to determine the accuracy of our prediction of incoming more or less than $50,000/yr. Based on the results, this objective was accomplished. Table II displays the performance of models on income prediction. The accuracy is 87.26%, 85.61%, and 84.86% for models of XGBoost, Decision Tree, and Logistic Regression. Thus, the model used by XGBoost has the highest accuracy among the three, due to its effectiveness in handling data and better performance on classification tasks. The accuracy of 87.26% shows that the model is able to correctly predict the income in the majority of cases. The accuracy of the Decision Tree model is slightly lower than the XGBoost model since the Decision Tree model is a simpler model that can be prone to overfitting. In addition, the model based on Logistic Regression has the lowest accuracy due to the limitation of the assumption of linearity between the feature variables and target variables. Overall, all predictions of income from the three models are generally great, indicating that models perform well across various scenarios.

In addition, the second objective of this project was to assess and mitigate gender bias in our UCI adult dataset. From Table II, the accuracies of sensitive attribute prediction for all three models are higher than 50%, 65.45% for XGBoost, 64.53% for Decision Tree, and 65.39% for Logistic Regression. It means that these predictions are better than random guessing, which would have an accuracy of 50% in predicting gender, and XGBoost has the highest accuracy. Therefore, we can predict the gender with an accuracy of 65%.

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON INCOME PREDICTION AND GENDER BIAS DETECTION

Model	Income Prediction	Sensitive Attribute Prediction
XGBoost	87.26%	65.45%
Decision Tree	85.61%	65.52%
Logistic Regression	84.86%	65.45%

B. Shapley Analysis

The Cohort Shapley value analysis revealed clear evidence of gender bias in the ground truth data, illustrated in Figure 1, where males are more likely to have an income greater than $50K. This bias is evident from the higher average Cohort Shapley values for males, indicating a higher probability of income above $50K. This inherent bias in the dataset is not only transferred but also exacerbated in the models' Shapley values. The charts show that all three models—XGBoost(Figure 4), Decision Tree (Figure 2), and Logistic Regression (Figure 3)—amplify this bias, further highlighting the disparity in predicted income levels based on gender.

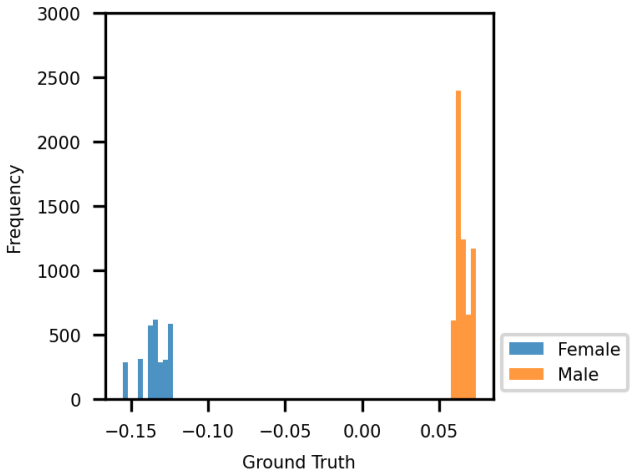

Fig. 1. Cohort Shapley Values for ground Truth

C. Bias Evaluation

The evaluation of machine learning models for predicting income levels above $50,000 revealed significant gender bias, particularly highlighted by the Cohort Shapley values. The XGBoost model showed the highest accuracy for income prediction (87.26%) but also indicated substantial gender bias with a 65.45% accuracy in predicting gender based on income. The Cohort Shapley method further confirmed this bias by demonstrating clear discrimination against females in both the dataset and the model's predictions. These findings emphasize the need for more advanced bias mitigation techniques and the application of models to diverse datasets to ensure fairness and robustness in predictions.

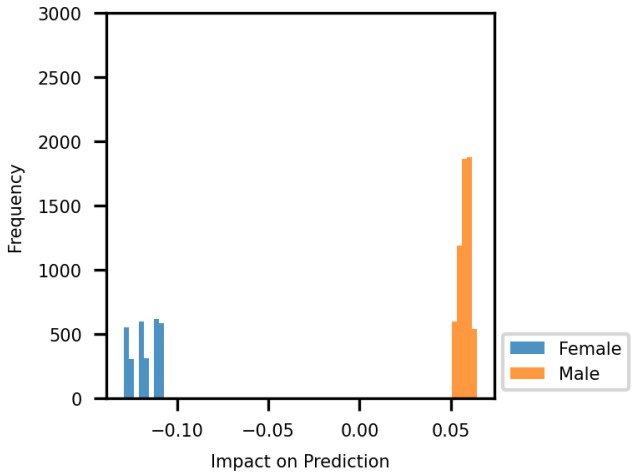

Fig. 2. Cohort Shapley Values for DT

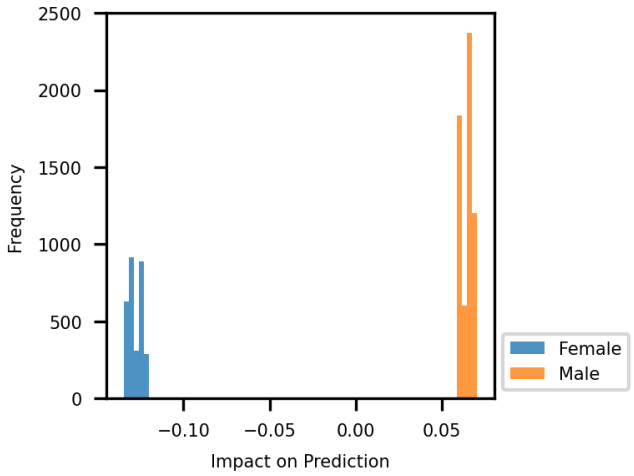

Fig. 3. Cohort Shapley Values for LR

V. Discussions and Conclusion

A. Implications of the Study

The results of our study find a couple of issues:

- **Presence of Social Bias:** All models we tested are able to guess the gender based on their income, we think they're picking up on social biases. This is a significant consideration for using these models in real-world applications.
- **Accuracy and Fairness:** XGBoost is the most accurate model we tested, but it also shows bias.

B. Future Directions

Based on what we have learned, the future research we can look into:

- **Advanced Bias Mitigation:** Look into more sophisticated methods to cut bias during model training, which might keep results accurate. Broader Dataset Analysis: Testing these models on different datasets might help us confirm how reliable the results are, to make sure they apply more than just one specific dataset.
- **Comprehensive Income Analysis:** We also plan to examine a wider range of income levels or analyze actual income for a better understanding of the trends we're observing.

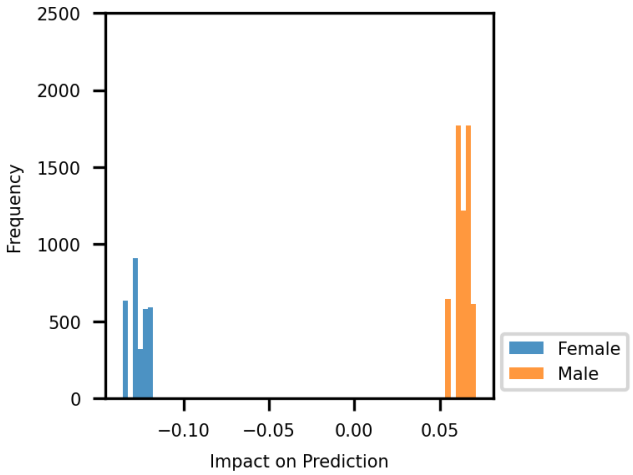

Fig. 4. Cohort Shapley Values for XGBoost

C. Conclusion

The project evaluated the predictive performance and fairness of Logistic Regression, Decision Tree, and XGBoost models in forecasting income levels above $50,000. XGBoost emerged as the most accurate model; however, all models demonstrated significant gender bias, as evidenced by high accuracy in predicting gender based on income predictions. The Cohort Shapley method provided insights into variable importance and highlighted the models' underlying biases. Despite XGBoost's superior accuracy, the pervasive gender bias across all models emphasizes the critical need for incorporating and enhancing bias mitigation strategies. Future research should explore more advanced in-processing techniques, apply models to diverse datasets, and expand income prediction scopes to ensure the robustness and fairness of machine learning models.

Acknowledgment

This research received support during the Math637 course by instructor, Professor Vu Dinh, and we really appreciate his support.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [Internet]. Irvine, CA: University of California, School of Information and Computer Science. Available from: https://archive.ics.uci.edu/ml/datasets/adult

[2] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems, 30.

[3] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods.

[4] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[5] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference.

[6] Wu, Yongkai, Lu Zhang, and Xintao Wu. "Fairness-aware classification: Criterion, convexity, and bounds." arXiv preprint arXiv:1809.04737 (2018).

[7] Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., & Vempala, S. (2018). The price of fair PCA: One extra dimension. In Advances in Neural Information Processing Systems.

[8] Jaiswal, Ayush, et al. "Unsupervised adversarial invariance." Advances in neural information processing systems 31 (2018).

[9] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference.

[10] Xu, D., Yuan, S., Zhang, L., Wu, X., & Wu, X. (2018). FairGAN: Fairness-aware generative adversarial networks. In 2018 IEEE International Conference on Big Data.

[11] Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012.

[12] Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma. "Fairness-aware learning through regularization approach." 2011 IEEE 11th international conference on data mining workshops. IEEE, 2011.

[13] Kusner, Matt J., et al. "Counterfactual fairness." Advances in neural information processing systems 30 (2017).

[14] Diana, Emily, et al. "Minimax group fairness: Algorithms and experiments." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021.

[15] Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).

[16] Kearns, Michael, et al. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness." International conference on machine learning. PMLR, 2018.

[17] Kearns, Michael, et al. "An empirical study of rich subgroup fairness for machine learning." Proceedings of the conference on fairness, accountability, and transparency. 2019.

[18] Mase, Masayoshi, Art B. Owen, and Benjamin Seiler. "Explaining black box decisions by Shapley cohort refinement." arXiv preprint arXiv:1911.00467 (2019).

[19] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[20] Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." Big Data 5.2 (2017): 153-163.