

Part 1: Algorithmic Fairness in Income-Based Decision Making

Baran Akkanat Madalina Gavati Tomash Oosterbaan
University of Groningen
Group 18 - WMAI032-05 Trustworthy and Explainable AI

Problem and Motivation

Algorithmic systems are increasingly used to inform high-stakes decisions such as credit approval and salary benchmarking. Such systems risk encoding and amplifying historical gender bias if deployed without careful fairness evaluation. We study this problem using the Adult Census Income dataset [7].

Automated income prediction is relevant because financial institutions and HR departments use similar models to support decisions about credit eligibility, loan limits, and salary benchmarking. However, these systems inherit biases from historical data. In the Adult dataset, 30% of male instances are labeled as earning above \$50K compared to only 11% of female instances, reflecting the well-documented gender wage gap [1]. A model trained without intervention may therefore reproduce this disparity.

Three stakeholder groups are affected. **Deployers** (loan officers, HR managers) require efficient and legally compliant decisions. **Affected individuals** (applicants) may receive lower credit limits or salary offers. **Regulators and auditors** must ensure compliance with anti-discrimination laws such as the EU Equal Treatment Directive [4] and the US Equal Credit Opportunity Act [8].

Methods and Assumptions

We use the Adult Census Income dataset [7], containing 48,842 instances and 14 features. Gender (`sex`) is treated as the protected attribute. We remove `fnlwgt` and `education`, encode categorical features, and split the data into 60% train, 20% calibration, and 20% test.

We train two models: **Logistic Regression (LR)** and **Random Forest (RF)**.

We evaluate fairness using two metrics:

Demographic Parity (DP) [3]:

$$\Delta DP = P(\hat{Y} = 1 | Male) - P(\hat{Y} = 1 | Female)$$

Equal Opportunity (EO) [5]:

$$\Delta EO = TPR_{Male} - TPR_{Female}$$

DP captures overall selection disparities, while EO measures fairness among qualified individuals. These metrics are known to be incompatible in some cases [2], motivating the use of both.

We apply **reweighing** [6] as our primary fairness intervention:

$$w(x_i) = \frac{P(G = g)P(Y = y)}{P(G = g, Y = y)}$$

This adjusts the training distribution without changing the model.

We also evaluate **threshold adjustment** [5] as a post-processing baseline.

Table 1: Performance and fairness metrics on the test set

Model	Acc.	F1	DP	EO
LR Baseline	0.821	0.542	+0.143	+0.216
LR Reweighing	0.809	0.495	+0.061	+0.004
LR Threshold	0.800	0.424	+0.050	+0.023
RF Baseline	0.868	0.685	+0.159	+0.084
RF Reweighing	0.862	0.658	+0.091	+0.116
RF Threshold	0.842	0.655	+0.103	+0.020

Results

Both baseline models exhibit substantial bias. For LR, men are predicted as high-income significantly more often than women.

Reweighing reduces both DP and EO gaps for LR with only a small accuracy loss. For RF, it reduces DP but increases EO, illustrating fairness trade-offs.

Threshold adjustment strongly reduces EO gaps but leads to larger performance degradation.

Discussion

Our results highlight that fairness is not a single objective.

Reweighing improves fairness by changing the training distribution, while threshold adjustment modifies only the decision boundary. The RF results show that improving one fairness metric can worsen another, confirming theoretical limitations [2].

The choice of fairness intervention should therefore depend on the application context. In hiring or salary settings, EO may be more relevant, while in lending, DP may be more important.

Conclusion

We evaluated algorithmic fairness in income prediction. Both LR and RF models show significant gender bias. Reweighing and threshold adjustment reduce disparities but introduce trade-offs between fairness and performance.

These results show that fairness evaluation requires careful selection of metrics and interventions, rather than a single universal solution.

References

- [1] Francine D. Blau and Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, 2017.
- [2] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.
- [3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226, 2012.
- [4] European Parliament and Council. Directive 2006/54/EC on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation, 2006. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32006L0054>.

- [5] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.
- [6] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.
- [7] R. Kohavi and B. Becker. Adult data set, 1996. URL <https://archive.ics.uci.edu/ml/datasets/adult>.
- [8] United States Congress. Equal credit opportunity act (ECOA), 15 U.S.C. 1691 et seq., 1974. URL <https://www.law.cornell.edu/uscode/text/15/1691>.