
DO FAIRNESS INTERVENTIONS CHANGE WHAT MODELS LEARN?

A JOINT FAIRNESS AND EXPLAINABILITY STUDY OF INCOME-BASED DECISION MAKING

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

ABSTRACT

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 through a joint fairness and explainability analysis of income prediction on the Adult Census Income dataset. Our **central research question** is: *do fairness interventions change a model's internal feature reasoning, or do they merely adjust prediction outputs?* We train Logistic Regression (LR) and Random Forest (RF) baselines, apply reweighing (pre-processing) and threshold adjustment (post-processing), and first evaluate each variant with Demographic Parity and Equal Opportunity metrics. We then use SHAP and LIME to probe whether the interventions alter the explanations the models produce. We find that reweighing genuinely shifts feature reliance, while threshold adjustment achieves comparable fairness gains without changing internal reasoning, a meaningful distinction for practitioners who need to justify decisions to regulators and affected individuals.

1 INTRODUCTION

Automated income prediction is used by financial institutions and HR departments to inform decisions about credit eligibility, loan limits, and salary benchmarking. These systems offer efficiency and consistency, but they inherit the biases present in historical data. In the Adult Census Income dataset (Kohavi and Becker, 1996), 30% of male instances are labelled as earning above \$50K compared to only 11% of female instances, a gap that reflects the persistent gender wage gap documented in the labour economics literature (Blau and Kahn, 2017). A model trained without fairness intervention will learn and reproduce this disparity, systematically under-predicting high income for women.

Three stakeholder groups are affected. **Employers** (loan officers, HR managers) use model outputs for efficiency; they need decisions to be defensible and legally compliant. **Affected individuals** (applicants, employees) face real economic consequences, suppressed credit limits or lower salary offers, if the model is biased. **Regulators and auditors** must verify compliance with anti-discrimination legislation such as the EU Equal Treatment Directive (European Parliament and Council, 2006) and the US Equal Credit Opportunity Act (United States Congress, 1974).

Measuring fairness at the output level is necessary but insufficient. A model can satisfy standard group-fairness criteria while still relying on features that act as gender proxies (e.g., `marital_status`, `occupation`). Explainability is therefore not merely a technical add-on but a practical requirement: loan officers must justify rejections, applicants must be able to contest outcomes, and auditors must be able to inspect the reasoning behind predictions.

This paper builds directly on our fairness analysis from Part 1. There, we showed that reweighing and threshold adjustment can reduce output disparities, but with different trade-offs across fairness metrics and predictive performance. Here, we ask the complementary question of whether those same interventions also change the model's internal reasoning.

Our central research question is: **do fairness interventions change a model’s internal feature reasoning, or do they only modify prediction outputs?** Two supporting questions are: (RQ2) which features drive predictions, and do attributions differ systematically between male and female applicants? and (RQ3) are local LIME explanations stable enough to support individual contestation?

2 METHODS

2.1 DATASET AND PREPROCESSING

We use the Adult Census Income dataset (Kohavi and Becker, 1996) (48,842 instances, 14 features). Gender (`sex`) is the protected attribute (binary: Male/Female as encoded in the source data). We drop `fnlwgt` (census sampling weight, not available at decision time) and `education` (redundant with `education_num`), ordinal-encode remaining categorical features, and stratify-split into 60% train / 20% calibration / 20% test. A separate calibration set prevents data leakage from the post-processing intervention into the final evaluation.

2.2 MODELS

We train two classifiers that differ in transparency and capacity. **Logistic Regression** (LR, $C = 1.0$, `max_iter=1000`) is a linear model that is naturally interpretable via its coefficients, and serves as a lower-complexity baseline. **Random Forest** (RF, 200 trees, `max_depth=12`) is a non-linear ensemble that typically achieves higher predictive performance but requires post-hoc explanation. LR uses standardised features; RF is scale-invariant and uses raw features.

2.3 FAIRNESS CONTEXT FROM PART 1

To connect XAI to the same decision-making context, we retain the fairness setup from Part 1. We evaluated the models using Demographic Parity difference (DP) and Equal Opportunity difference (EO), and we applied reweighing (pre-processing) and threshold adjustment (post-processing). This fairness context is necessary because the purpose of the XAI analysis is not to explain a generic classifier, but to determine whether fairness interventions change how the model reasons.

2.4 XAI METHODS

SHAP (Lundberg and Lee, 2017) provides theoretically grounded feature attributions via Shapley values. We use **TreeSHAP** (Lundberg et al., 2020) for RF and **LinearExplainer** for LR. We also compute an *explanation fairness gap*: the mean absolute difference in per-feature mean $|\text{SHAP}|$ between male and female subgroups. For threshold-adjusted variants, SHAP is computed on the underlying base model, since threshold adjustment never modifies weights or tree splits; this directly tests our central research question.

LIME (Ribeiro et al., 2016) approximates the model locally with a ridge-regression surrogate, producing instance-level explanations. We explain five representative instances and evaluate *explanation stability* (Alvarez-Melis and Jaakkola, 2018) by adding Gaussian noise ($\sigma = 0.05$) to 50 test instances ($N = 10$ perturbations each) and measuring cosine similarity between original and perturbed explanation vectors.

3 RESULTS

3.1 FAIRNESS CONTEXT

The fairness analysis from Part 1 provides the starting point for this paper. Both baseline models exhibit substantial fairness gaps. For LR, the baseline DP difference is +0.143 and the EO difference is +0.216. Reweighting reduces both gaps substantially for LR (DP: +0.061, EO: +0.004), while threshold adjustment achieves strong EO reductions for both models (LR: +0.023, RF: +0.020). These output-level improvements motivate the main XAI question: do these interventions change what the model learns, or only how its scores are converted into final decisions?

3.2 EXPLAINABILITY ANALYSIS

Global feature importance. Figure 1 shows mean $|\text{SHAP}|$ for all RF variants. `capital_gain`, `age`, and `education_num` are the top three features and this ranking is stable across all variants, consistent with our hypothesis that threshold adjustment does not alter internal reasoning. For LR, `marital_status` and `relationship` appear more prominently, suggesting the linear model relies on social-structural features that may act as gender proxies.

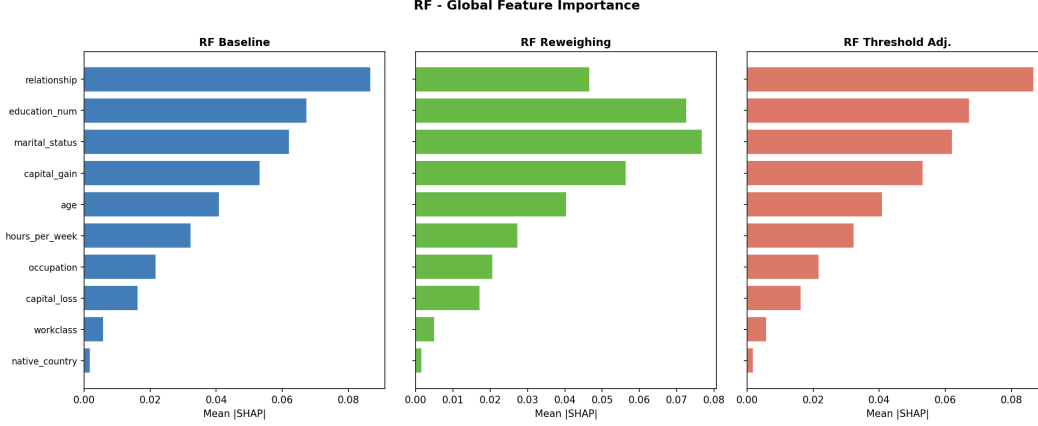


Figure 1: Mean $|\text{SHAP}|$ for RF baseline, reweighed, and threshold-adjusted variants. The feature ranking is stable across variants.

Explanation fairness gap. Table 1 and Figure 2 report the scalar explanation fairness gap across all variants.

Table 1: Explanation fairness gap (lower = more similar reasoning across genders).

Model	Explanation Fairness Gap
RF Baseline	0.0074
RF Reweighing	0.0111
RF Threshold Adj.	0.0074
LR Baseline	0.0693
LR Reweighing	0.0574
LR Threshold Adj.	0.0693

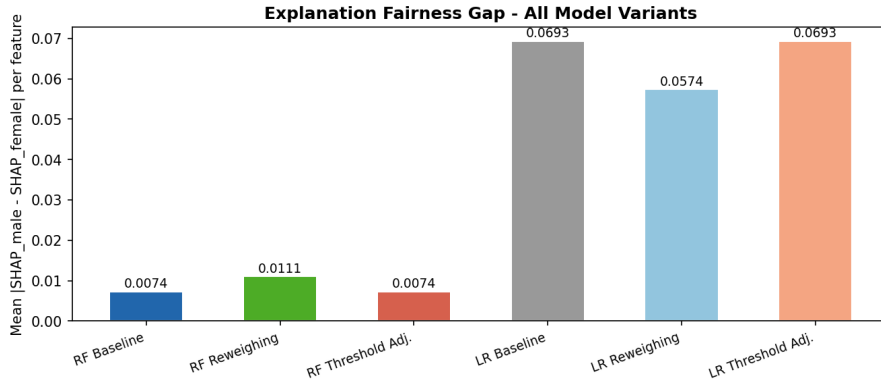


Figure 2: Scalar explanation fairness gap across all model variants. Threshold adjustment leaves the gap unchanged; reweighing reduces it for LR.

For RF, threshold adjustment leaves the gap identical to the baseline (0.0074), confirming it only moves the decision boundary. Reweighting increases the RF gap slightly (0.0111), reflecting genuine redistribution of feature weights. For LR, reweighting reduces the gap from 0.0693 to 0.0574 while threshold adjustment leaves it at 0.0693. **This is our key finding: reweighting changes internal reasoning; threshold adjustment does not.**

Subgroup SHAP analysis. Figure 3 compares male and female SHAP distributions for the RF baseline. `capital_gain` and `education_num` show similar attribution magnitudes for both groups, while `marital_status` shows a larger disparity, a potential proxy channel for gender.

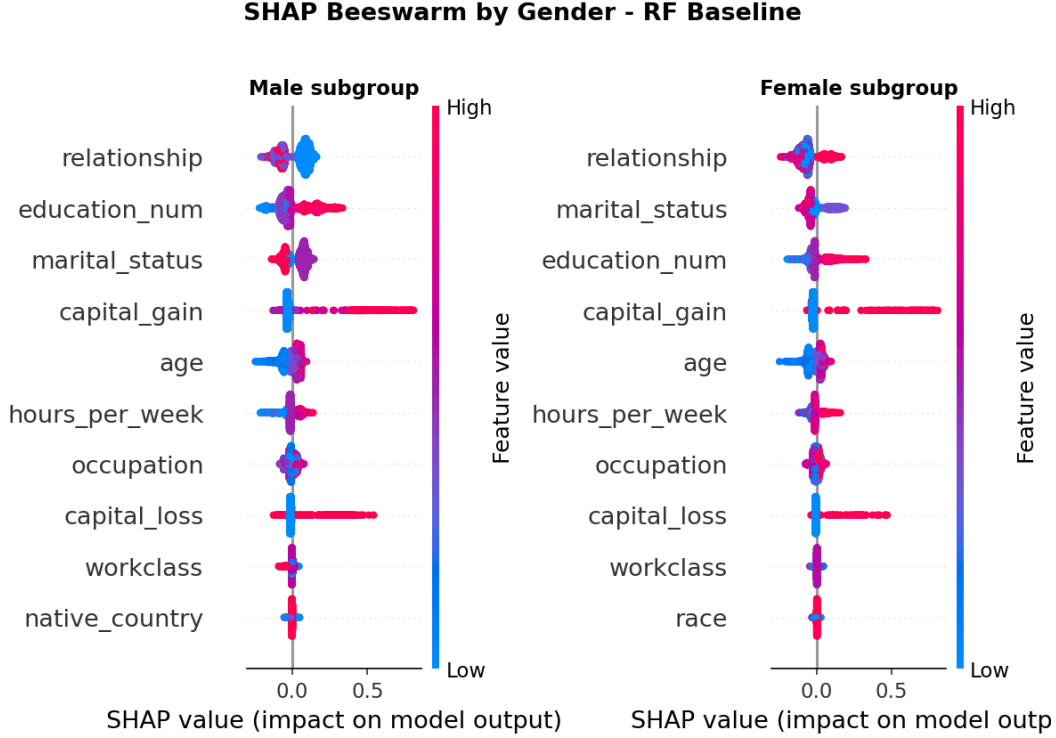


Figure 3: SHAP beeswarm plots for male (left) and female (right) subgroups under the RF baseline.

LIME stability. Figure 4 shows cosine similarity stability scores. All variants score low: RF Baseline 0.175, RF Reweighed 0.172, LR Baseline 0.210, LR Reweighed 0.232. Two applicants with nearly identical profiles could receive substantially different explanations, undermining individual accountability.

4 DISCUSSION

Reweighting vs. threshold adjustment. Our results reveal a fundamental asymmetry. Reweighting modifies the training distribution, forcing the model to learn different feature relationships, visible in both the fairness metrics from Part 1 and the explanation fairness gap in this paper. Threshold adjustment achieves comparable EO improvements by post-hoc boundary shifting, but leaves internal reasoning entirely unchanged. For regulators, this distinction matters: threshold adjustment does not address proxy-driven reasoning that could manifest in other deployment contexts.

The proxy problem. `marital_status` and `relationship` carry higher attribution weight for LR than for RF. These features are strongly gender-correlated in the Adult dataset, suggesting the linear model partially encodes gender through proxies even though `sex` is excluded, a fairness concern that output-level fairness metrics alone would not surface.

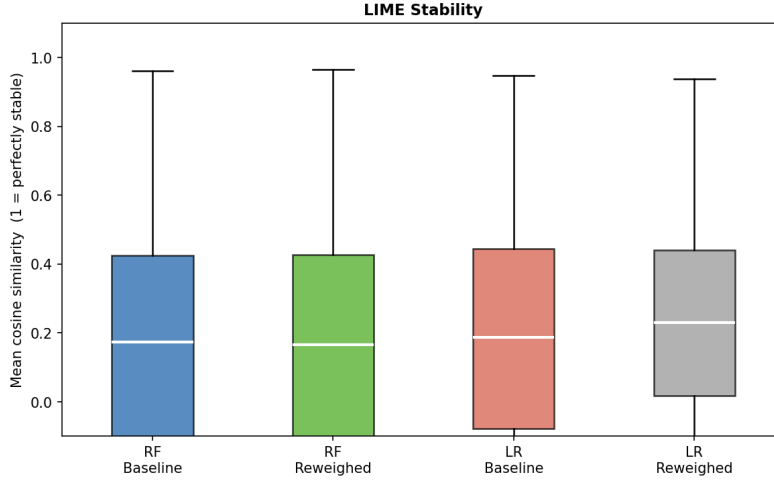


Figure 4: LIME stability (cosine similarity) under Gaussian perturbations. All variants show low stability, cautioning against using LIME for individual accountability.

LIME instability and accountability. The low LIME stability scores (≈ 0.17 – 0.23) mean that LIME explanations should be used for internal auditing only, not for communicating decisions to affected individuals.

Limitations. The dataset encodes gender as binary. The 1994 census data may not reflect current income patterns. We do not evaluate in-processing methods such as adversarial debiasing, which could offer different trade-offs. In addition, our explanation fairness gap is a useful aggregate diagnostic, but it does not capture every possible dimension of explanation quality. This demonstrates that explanation-based evaluation provides additional insight beyond standard fairness metrics, revealing differences in model behavior that would otherwise remain hidden.

5 CONCLUSION

We presented a joint fairness and explainability analysis of income prediction in a concrete deployment context with clearly identified stakeholders. The fairness analysis in Part 1 showed that reweighing and threshold adjustment can both reduce output disparities, but with different trade-offs. Building on that result, the present paper shows that reweighing genuinely alters what the model learns, while threshold adjustment only moves the decision boundary. SHAP-based explanation fairness gaps provide a practical diagnostic for detecting this difference. Low LIME stability scores caution against using local explanations for individual-level accountability. Together, these findings argue for treating fairness evaluation and explainability as complementary and inseparable components of responsible AI deployment.

REFERENCES

- David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods, 2018. URL <https://arxiv.org/abs/1806.08049>. ICML Workshop on Human Interpretability in Machine Learning.
- Francine D. Blau and Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, 2017.
- 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>.

R. Kohavi and B. Becker. Adult data set, 1996. URL <https://archive.ics.uci.edu/ml/datasets/adult>.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

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>.

APPENDIX: CONTRIBUTIONS

- **Baran Akkanat:** Fairness metric implementation, reweighing, visualisation, XAI analysis.
- **Madalina Gavut:** Data preprocessing, model training, threshold adjustment, Part 1 report.
- **Tomash Oosterbaan:** Report writing, experimental design, LIME stability analysis, code review.

All members contributed to the analysis, discussion of results, and presentation.

APPENDIX: GENERATIVE AI USAGE

We used Claude (Anthropic) to assist with code structure and report formatting. All analytical decisions, metric choices, experimental design, and interpretations of results were made by the group. We fully understand all code and are prepared to answer technical questions during the presentation.