

Fairness Analysis of Machine Learning Models on the UCI Adult Income Dataset

Youssef Sallam Allan Martinez

Fall 2025

Abstract

Machine learning systems are increasingly used in high-stakes decision-making processes, raising concerns about fairness and bias. In this project, we analyze gender-based bias in income prediction models trained on the UCI Adult Income dataset. We evaluate a baseline logistic regression model using fairness metrics including demographic parity and equal opportunity. We then apply a data rebalancing mitigation strategy and assess its impact on both model performance and fairness outcomes. Our results show that high predictive accuracy does not guarantee fair outcomes and that simple mitigation techniques may be insufficient to address structural bias.

Section ownership:

Introduction, Background, Dataset, Ethics, Conclusion — Allan
Methodology, Fairness Metrics, Results, Mitigation, Discussion — Youssef

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems have been increasingly integrated into high-stakes decision-making processes that fundamentally shape human opportunities, affecting hiring, lending, criminal justice, healthcare, etc. As these systems continue

to scale, the risk of amplifying historical societal inequities (often encoded in training data) is a critical risk that must be addressed. This project analyzes gender-based bias in income prediction models trained on the UCI Adult Income dataset, which is a standard benchmark in fairness research. A primary goal was to evaluate how predictive performance interacts with fairness outcomes. Specifically, we investigated whether a standard logistic regression model disproportionately penalizes female applicants and also whether our implemented mitigation strategies successfully reduced these disparities without compromising performance/model utility.

2 Background and Related Work

Bias in ML systems are typically emerged from three stages of the modeling pipeline:

- **Historical Bias:** This occurs when the training data reflects past societal prejudices or systemic inequalities. For example, if past hiring decisions favored men due to structural sexism, a model trained upon that history will learn to replicate those preferences, which effectively automates discrimination
- **Representation Bias:** This happens when the development data fails to accurately represent the diversity of the target population. If a dataset under-samples minority groups, the model will struggle to generalize to them, leading to higher error rates for those populations
- **Algorithmic Bias:** This is introduced by the model itself, through optimization functions that prioritize global accuracy at the cost of minority group performance. A model may treat minority groups as “noise” to minimize overall loss, sacrificing fairness for marginal gains in accuracy.

Two primary metrics were used for this project:

- **Demographic Parity (DP):** Requires that positive decisions (like loan approval) be assigned at equal rates across groups. While intuitively appealing, this can also be criticized for failing to account for legitimate differences in personal qualification.

- **Equal Opportunity (EO):** Focuses on error rates, requiring that qualified individuals (positive true label) have an equal probability of being correctly classified as positive across groups.

Upon researching mitigation techniques with these metrics and biases in mind, mitigation strategies such as pre-processing, in-processing, and post-processing came up. This project primarily focused on pre-processing via gender-based oversampling to test whether addressing representation bias alone is sufficient to correct outcome disparities.

3 Dataset Description

We utilized the UCI Adult Income Dataset, extracted from the 1994 U.S. Census database [1]. This dataset was chosen due to it being a benchmark in the fairness literature and its clear demographic attributes and known imbalances. The dataset contained approximately 48,842 samples.

The binary classification task was to predict whether an individual's annual income exceeds \$50,000. The dataset includes 14 features divided into categorical and numerical:

- **Categorical:** workclass, education, marital_status, occupation, relationship, race, sex, native_country
- **Numerical:** age, fnlwgt, education_num, capital_gain, capital_loss, hours_per_week

The dataset exhibited significant gender and class imbalances. Males constitute approximately 67% of the samples, while females made up 33% (2:1 ratio). There was also a stark disparity in the target variable: approximately 30% of men in the dataset earned >\$50K, compared to only 11% of women. This structural inequality makes the dataset ideal for analyzing how models can amplify historical economic disparities.

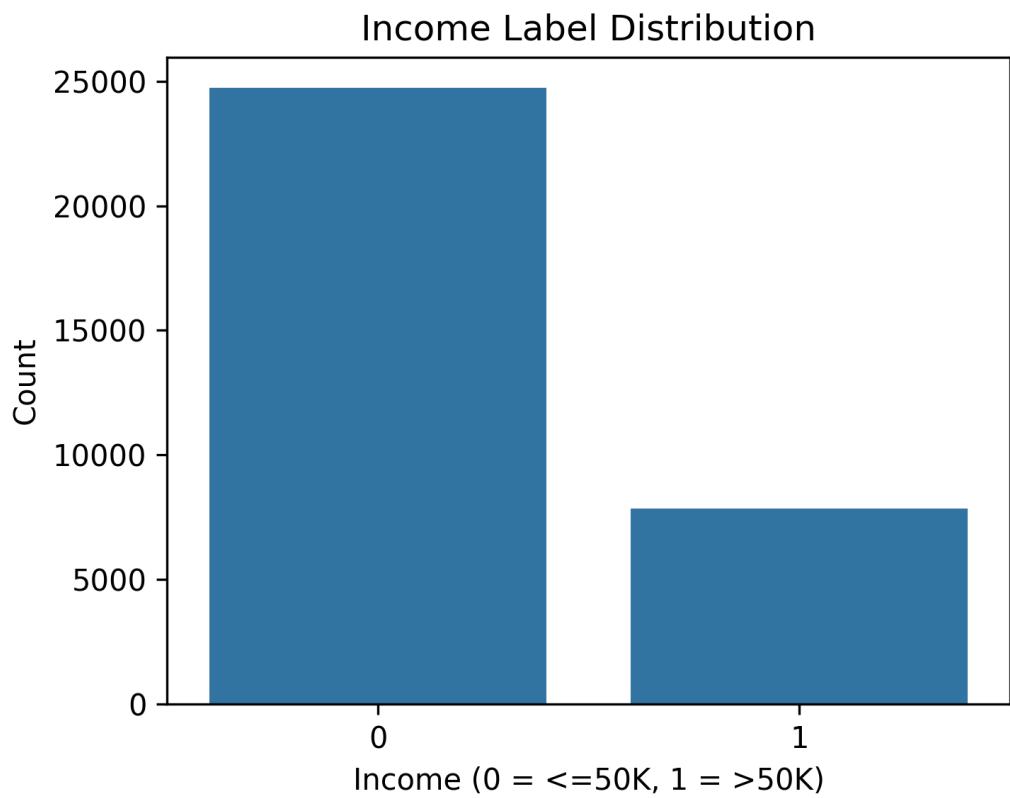


Figure 1: Distribution of income classes in the UCI Adult dataset, showing a substantial imbalance between individuals earning less than or equal to \$50K and those earning more than \$50K annually.

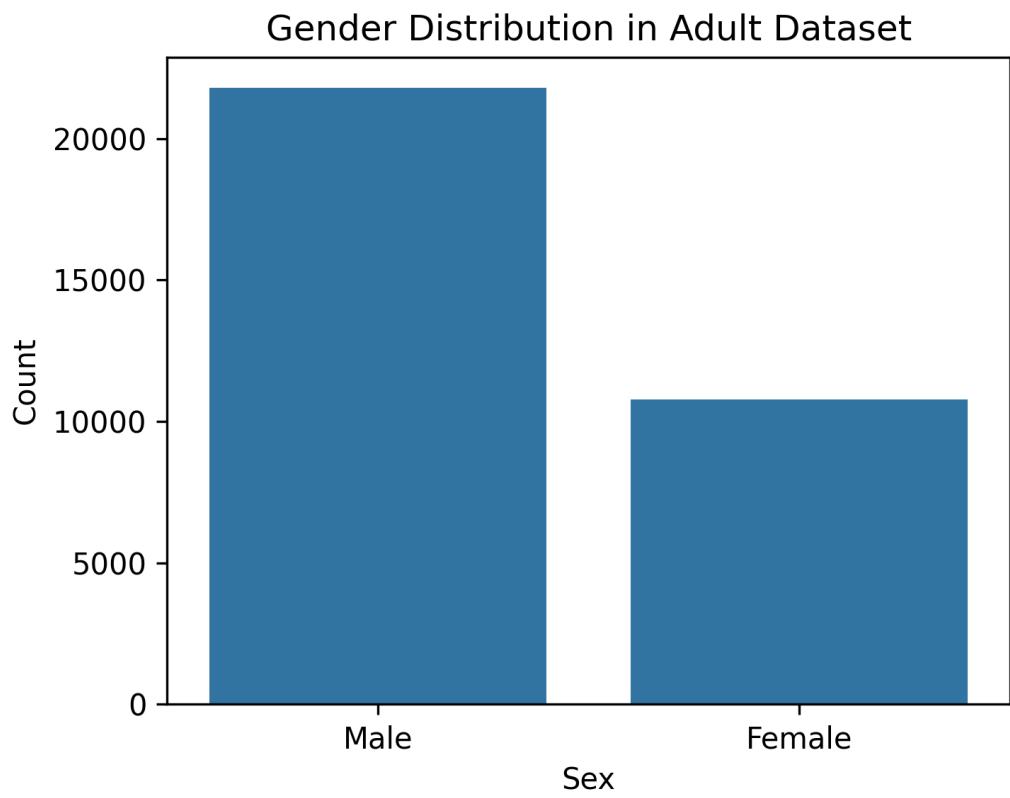


Figure 2: Gender distribution in the UCI Adult dataset, illustrating a significant representation imbalance with male samples comprising approximately two-thirds of the dataset.

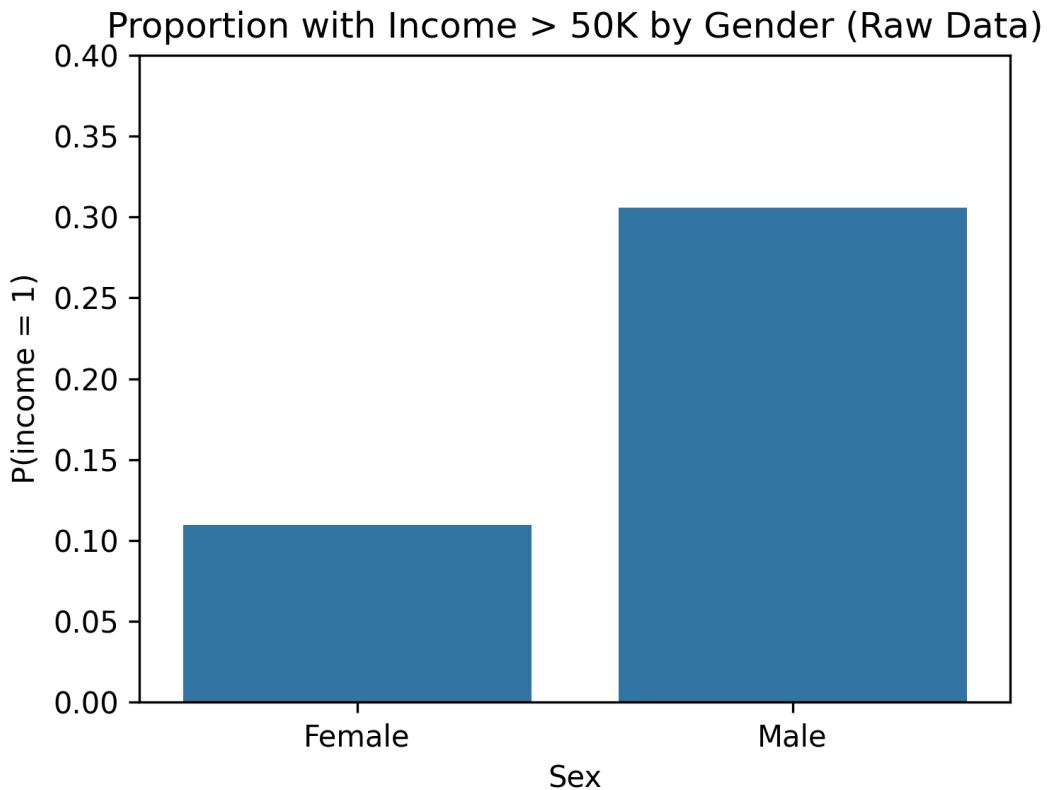


Figure 3: Proportion of individuals earning more than \$50K by gender in the raw dataset, highlighting pronounced structural income disparities prior to model training.

4 Methodology

Since this study evaluates gender-based bias in income prediction models using the UCI Adult Income dataset, we based our methodology on three major components. Those components consist of a standardized pre-processing pipeline, a baseline classification model optimized for predictive accuracy, and finally, a controlled evaluation framework designed to enable fairness analysis across demographic groups.

4.1 Data Pre-processing

The raw dataset from UCI included a vast majority of missing values that were represented by placeholder symbols, additionally, there was a mixture of categorical and numerical fea-

tures. Thus, all instances containing missing values were removed to ensure consistency across samples, categorical features were trained as well by trimming extraneous whitespace and standardized prior to encoding. Additionally, categorical variables including work class, education, occupation, relationship, race, sex, and native country, were transformed using one-hot encoding. For the numerical features such as age, educational level, capital gains + capital losses, and hours worked per week were all normalized via standard scaling. Through this preprocessing pipeline we were able to guarantee that features are on comparable scales and that categorical information is preserved without the introduction of ordinal assumptions.

4.2 Model Architecture

When choosing the baseline classification model we decided that logistic regression would be the best choice due to its interpretability and widespread use as a benchmark in fairness research. Furthermore, logistic regression provides a linear decision boundary and allows for transparent analysis of prediction behavior without the added complexity of non-linear models. We split the dataset into training and testing sets using an 80/20 ratio, stratification was also applied in order to preserve class balance across splits. Moreover, we prevented any data leakage between training and testing phases by ensuring all pre-processing steps were applied within a unified pipeline. Additionally, all models were implemented using the scikit-learn machine learning library [3].

```
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, stratify=y, random_state=42  
)  
  
model = LogisticRegression(max_iter=300)  
model.fit(X_train, y_train)
```

4.3 Evaluation Procedure

we evaluated the model performance using standard classification metrics that include accuracy, precision, and recall. These metrics provided us with insight into the overall predictive quality but they do not capture disparities across demographic groups. Thus, we evaluated fairness-specific metrics separately as described in the following section.

5 Fairness Metrics

In order to be able to asses whether the baseline model exhibits gender-based bias or not, we evaluated fairness using two commonly adopted group fairness metrics, them being: demographic parity and equal opportunity. These metrics were vital in capturing different dimensions of fairness as well as being commonly used in algorithmic auditing and fairness research.

5.1 Demographic Parity

Demographic parity's main role is to measure whether a model assigns positive predictions at equal rates across different demographic groups. In a formal sense, a model satisfies demographic parity if the probability of predicting a positive outcome is independent of group membership.

$$P(\hat{Y} = 1 \mid A = \text{Male}) = P(\hat{Y} = 1 \mid A = \text{Female})$$

In the scope of our analysis, demographic parity evaluates whether male and female individuals are predicted to earn over \$50,000 at similar rates. A significant difference between these probabilities signifies outcome disparity, regardless of true income labels.

5.2 Equal Opportunity

Equal opportunity focuses on the prediction quality for qualified individuals, specifically, it requires true positive rates to be equal across demographic groups [2]. Via this metric we

ensure that individuals who truly belong to the positive class have an equal chance of being identified correctly by the model.

$$P(\hat{Y} = 1 \mid Y = 1, A = g)$$

Where g denotes a demographic group. In our analysis, equal opportunity compares the true positive rates of male and female individuals whose true income exceeds \$50,000.

5.3 Metric Interpretation

Demographic parity evaluates disparities in predicted outcomes, on the other hand, equal opportunity evaluates disparities in prediction accuracy for qualified individuals. Together these two metrics compliment each other in providing different perspectives on fairness. Meaning, a model may perform well according to one metric while failing under the other, this truly depicts the vital importance evaluating multiple fairness criteria.

visualizations: Results for both metrics are visualized in figures 4 and 5, which compare baselines and post-mitigation fairness gaps.

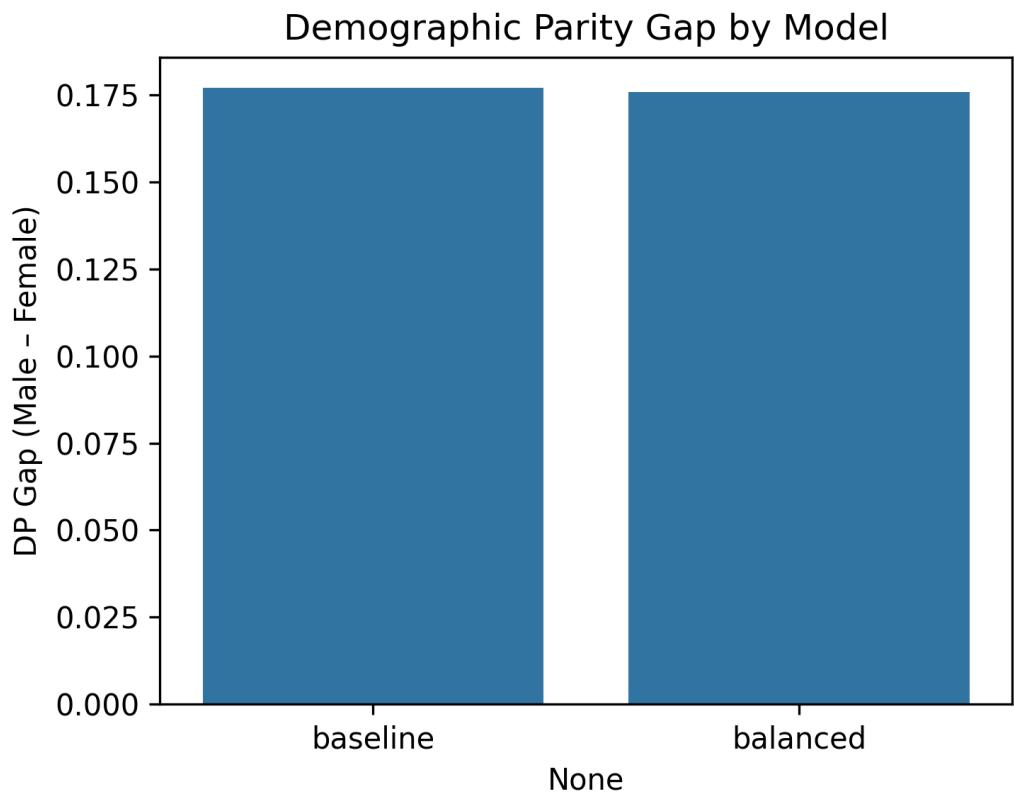


Figure 4: Comparison of demographic parity gaps for the baseline and post-mitigation models, showing limited reduction in outcome disparity after applying gender-based oversampling.

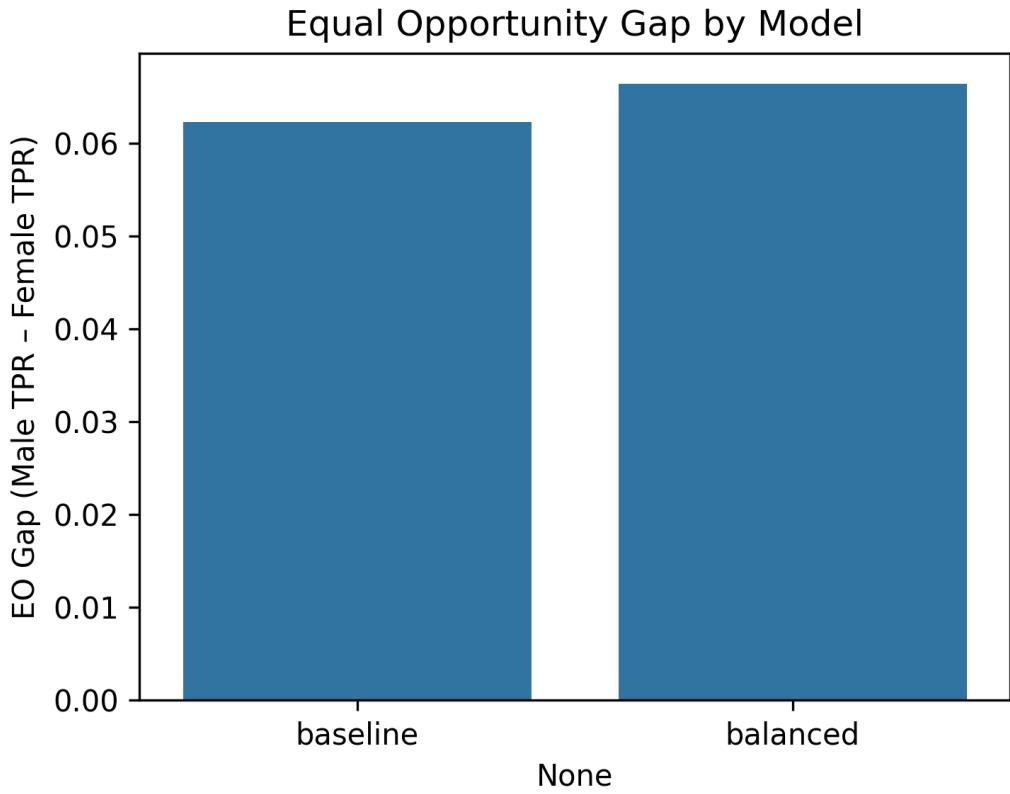


Figure 5: Comparison of equal opportunity gaps between the baseline and post-mitigation models, indicating persistent differences in true positive rates across genders.

6 Experimental Results

This section is for presenting the predictive performance and fairness evaluation of the baseline income prediction model. All results are reported on a held-out test set in order to ensure unbiased evaluation.

6.1 Baseline Model Performance

The baseline logistic regression model was able to achieve strong predictive performance on the income classification task. Table 1 summarizes the primary evaluation metrics.

Table 1: Baseline Model Performance on Test Set

Metric	Accuracy	Precision	Recall
Baseline	0.856	0.740	0.620

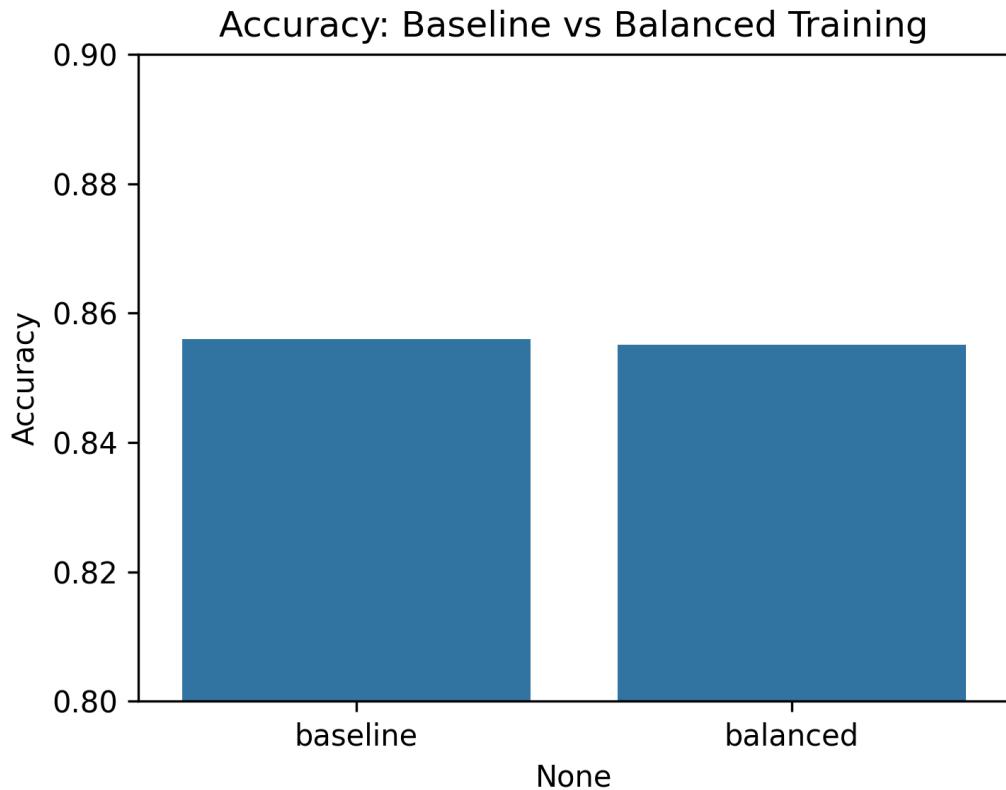


Figure 6: Model accuracy for baseline and post-mitigation training, demonstrating that predictive performance remains largely unchanged following bias mitigation.

The model’s accuracy displayed above tells us that it performs well at predicting income level overall. However, we must note that predictive performance alone does not capture whether errors and outcomes are distributed equally across different demographic groups.

6.2 Baseline Fairness Evaluation

In order to asses fairness, we used both demographic parity and equal opportunity with respect to gender for evaluation. the results displayed that the baseline model exhibited

substantial disparities across both metrics. Through the **demographic parity** analysis we were able to see that male individuals were predicted to earn over \$50,000 at a rate of approximately 26%, on the other hand, female individuals were predicted to earn over \$50,000 at a rate of approximately 8%. This results in a demographic parity gap of roughly 17.7%, highlighting that there is a strong disparity in predicted outcomes between genders. And through the **equal opportunity** analysis we were able to see that among individuals whose true income exceed \$50,000, male individuals were more likely to be correctly identified by the model than females. The true positive rate for males surpassed that of females by approximately 6.2%, showing unequal recall performance across groups. Even though these results showed high predictive accuracy, the baseline model displays serious gender-based fairness violations, which also indicates the limitation of accuracy-focused evaluations and promotes the application and the need for bias mitigation strategies.

7 Bias Mitigation and Analysis

In order to investigate whether or not fairness disparities could be reduced, we implemented a bias mitigation strategy based on data re-balancing, that aims to address the representation imbalance via increasing the presence of underrepresented demographic groups during training.

7.1 Mitigation Strategy

We applied a mitigation strategy that involved gender-based oversampling within the training dataset. Meaning, female sample were over-sampled with replacement until the number of female and male training instances were equal. This results in reducing representation bias without having to modify the model architecture or loss function. Additionally, the test dataset remained unchanged in order to guarantee fair comparison between baseline and post-mitigation results.

7.2 Post-Mitigation Performance

Following the training of the logistic regression model on the balanced dataset, we re-evaluated predictive performance on the same held-out test set. Table 2 summarizes post-mitigation performance metrics.

Table 2: Post-Mitigation Model Performance

Metric	Accuracy	Precision	Recall
Balanced Model	0.855	0.737	0.619

The overall predictive performance remained nearly unchanged compared to the baseline model, indicating that the mitigation strategy was not able to significantly reduce accuracy or classification quality.

7.3 Fairness Impact Analysis

In spite of our efforts to balance gender representation in the training data via bias mitigation, fairness metrics showed minimal improvement. Specifically, the demographic parity gap remained largely unchanged, additionally, the equal opportunity gap increased only slightly compared to the baseline model. Figures 4 and 5 compare baseline and post-mitigation fairness gaps. The results show us that simple data re-balancing alone is not enough to eliminate gender-based bias in income prediction. This outcome tells us that fairness disparities are driven not only by representation imbalance but also by deeper structural correlations in features such as occupation, education, and labor participation. Thus, simple mitigation strategies might be able to preserve performance but fail to meaningfully improve fairness.

7.4 Interpretation

The very minimalist effectiveness of the mitigation strategy indicates that there is a truly critical challenge in real-world fairness interventions. Meaning, bias can persist even when demographic representation is equalized, emphasizing the vital need for more advanced fairness-aware modeling techniques.

8 Ethical and Real-World Implications

The technical findings of this report have profound ethical implications for real-world deployment. If the biases observed in our baseline model (the 17.7% gap in DP) were deployed in a hiring algorithm, it would result in significantly fewer women receiving job offers compared to equally qualified men. An algorithm that optimizes solely for accuracy while ignoring fairness metrics can be actively harmful, potentially violating anti-discrimination laws and reducing social trust.

9 Discussion and Future Work

The results of this study indicated several fundamental problems with machine learning systems. Firstly, the study showed that even though a model can have strong predictive performance, it does not imply fairness in the model. For example, although the baseline logistic regression model achieved high accuracy, both demographic parity and equal opportunity analyses revealed that there were substantial gender-based disparities. The findings of this study support the vital importance of fairness evaluation as a distinct and required component of any model assessment.

The bias mitigation strategy we implemented that aimed to balance genders via oversampling of the training data successfully maintained the predictive performance but failed in having a significant impact on reducing fairness gaps. This suggests that the imbalance of gender representation alone was not the primary driver of bias in this task. Rather, fairness violations seem to be rooted in deeper structural correlations within the dataset such as but not limited to disparities in occupation types, education levels, and labor participation patterns. In cases where features such as the ones mentioned before act as proxies for protected abilities, simple data-level intervention is rendered insufficient.

These results also point towards a broader challenge in real-world fairness mitigation, that being, interventions that do not specifically account for the way a model uses correlated features may fail to address the main source of bias in the model. Although, data re-balancing is at times useful and is also fairly easy to implement, it does not guarantee eq-

uitable outcomes when historical and socioeconomic inequalities are embedded in the feature space.

Future work could explore more advanced fairness-aware techniques that would directly incorporate fairness objectives into the learning process. Potential approaches can include techniques such as reweighting schemes that adjust the influence of training samples, adversarial debiasing methods that explicitly remove protected attribute information from learned representations, and constrained optimization techniques that enforce fairness constraints during training. Furthermore, evaluating multiple protected attributes at the same time such as gender and race, could provide a much more comprehensive understanding of intersectional bias.

Finally, expanding this analysis to other model architectures may grant us further insight into the tradeoff between fairness and accuracy. Additionally, non-linear models such as decision trees or ensemble methods may display different fairness behaviors and call for separate evaluation. Overall, this study sheds a light on the complexity of achieving fairness in machine learning and emphasizes the vital need for careful multi-dimensional evaluation in applications that have significant impacts with it's decisions.

10 Conclusion

This project evaluated the fairness of income prediction models on the UCI Adult dataset, revealing that “fairness” requires deliberate intervention rather than just high accuracy. Our baseline logistic regression model achieved a strong predictive accuracy of 85.6% but exhibited a significant Demographic Parity gap of 17.7%, indicating a strong bias toward predicting higher income for men over women. Our mitigation strategy of gender-based oversampling proved largely ineffective. While it maintained accuracy at 85.5%, it reduced the DP gap by a negligible amount (to 17.6%) and slightly increased the Equal Opportunity Gap (from 6.2% to 6.6%). These results lead to the vital conclusion that representation bias is not the only source of unfairness. Simply showing the model more examples of women did not correct for Historical Bias encoded in the features themselves (such as occupation and education levels). Real-world fairness cannot be solved by simple data augmentation alone,

it would require more sophisticated in-processing techniques or policy-level interventions. Accuracy is a necessary but insufficient metric for responsible AI. Without rigorous fairness analysis, we risk building systems that are highly accurate at being unfair.

References

- [1] Dheeru Dua and Casey Graff. Uci machine learning repository: Adult data set. <https://archive.ics.uci.edu/ml/datasets/adult>, 2017. Accessed: 2025.
- [2] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.
- [3] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.