Report

Master's Programme in Data Science

Fair Income Prediction: Mitigating Gender Bias With Fairness-Aware Machine Learning

Anton Bogun, Sanish Gurung

December 15, 2025

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Anton Bogun, Sanish Gurung

Työn nimi — Arbetets titel — Title

Fair Income Prediction: Mitigating Gender Bias With Fairness-Aware Machine Learning

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Report	December 15, 2025	15

Tiivistelmä — Referat — Abstract

This report focuses on measuring fairness and adopting fairness-aware machine learning models. In this project, we used publicly available 'Adult / Census Income' dataset which contains various demographic and sensitive information of an individual.

We show the unfairness found when training base models on our dataset, and use Exponentiated Gradient technique to train fair models. The resulting models achieve a good accuracy with only a few percent loss compared to base models, while achieving much better fairness. We conclude that this is a good showcase of the importance of fairness considerations in decision-making ML models, especially when the underlying data is suspect of historical biases.

Avainsanat — Nyckelord — Keywords

Fairness, Gender, Demographic Parity, Bias, Machine Learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Use of Artificial Intelligence (AI): I have used AI tools in the following ways:

1. I checked spelling errors and english fluency check like "something ..., something" to see if they sounded better or to correct my spelling. Moreover, I used keywords or expressions to check for grammatical correctness.

2. I used for formatting the Project template.

Contents

1. Introduction

In today's world, machine learning has been increasingly used in decision-making tasks such as promotions, salary prediction, loan approval, and hiring candidates. Although extensive research has been conducted in the field of machine learning and fairness, deploying these systems in the real world poses risks because these models can produce unfair or biased outcomes, leading to detrimental consequences for individuals and society [Spi23].

The main reason for unfair decisions is that many datasets contain historical or societal biases. If the machine learning models are trained on such data, they start to incorporate these biased patterns if no human corrections are included in the machine learning pipeline. However, as datasets get large, organizations may lack the sufficient human resources to do so, or the companies may not employ a sufficient amount of skilled statisticians and data scientists capable of detecting and addressing these fairness issues. [Zli17a]

Generally, when practitioners work with demographic or sensitive information, they should be fully aware of the relevant fairness concerns. Using aggregate metrics such as accuracy or precision does not always reveal biases against minority groups [Goo25]. Avoiding unfairness consists of a two stages, with the first stage involving measuring of the fairness metrics applicable in context, and the latter stage involving developing and utilizing various fairness-aware techniques. Adopting fairness may come with some accuracy loss, meaning that the model will perform poorly compared to the unrestricted and unfair model. Thus, adjusting fairness thresholds while maintaining practical utility is crucial. [Zli17b]

The project report is structured as follows: In Chapter 2, we introduce the dataset and explain the rationale behind selecting key features. In Chapter 3, we highlight existing imbalances that may produce unfair decision-making. Chapter 4 presents the fairness metrics for evaluating model behaviour and fairness. Chapter 5 describes the methods adopted to implement fairness-aware machine learning. Finally, Chapter 6 concludes the report by summarizing project findings and discussing potential future work.

2. Datasets and Feature Overview

This chapter introduces the dataset and explains the rationale for selecting key features used to measure fairness and build fairness-aware machine learning models.

For this project, we used the well-known 'Adult / Census Income' dataset, available at https://archive.ics.uci.edu/dataset/2/adult [BK96]. The dataset includes the individual's demographic and sensitive information. Initially, dataset (adult.data and adult.test) contains 48,843 records, but in our project, we ignored the adult.test and only used adult.data (32,561 records). Since our focus is on measuring fairness and adopting fairness-aware machine learning models, we need a test partition of the dataset to check how accurate and fair the model is after training on the main adult.data partition. The primary task that the models optimize within this dataset is predicting whether an individual earns over $50,000 or less annually.

2.1 Features

The *Adult / Census Income* dataset contains 15 features. Table 2.1 produces all features along with their descriptions and the number of unique values. All the features highlighted in green colours were used in our work, while those in red colours were not used. The selection was based on whether a feature is demographic or sensitive, and if it is likely to be predictive of income.

The reasons for excluding the red-colored features are:

- **Fnlwgt**: A sampling weight from the Census Bureau, providing no predictive information about income and potentially introducing unrelated bias.
- **Education-num**: This is a numeric encoding of *Education* feature which already exists in the datasets. Including it again could cause redundancy and multicollinearity.
- **Native-country**: The vast majority of individuals are from the U.S., so the few non-U.S. entries will add sparsity without meaningful predictive power and could skew fairness metrics.

In our experiments, **Income** was set as the target variable and the remaining selected features were used as input variables for both predicting income and measuring fairness. For adopting fairness-aware machine learning, **Sex** was treated as the sensitive attribute to evaluate and enforce fairness in the models.

Feature	**Description**	**Unique Values**
Age	Age of the individual	73
Workclass	Type of employer (e.g. government, private, self-employed)	9
Fnlwgt	Final sampling weight used by Census	21648
Education	Highest level of education attained	16
Education-num	Numerical representation of education level	16
Marital-status	Marital situation (married, single, divorced, etc.)	7
Occupation	Job category (tech support, executive managerial, etc.)	15
Relationship	Family role (wife, husband, own-child, etc.)	6
Race	Ethnicity (White, Black, Asian-Pac-Islander, etc.)	5
Sex	Biological sex	2
Capital-gain	Income from capital gains	119
Capital-loss	Income lost through capital	92
Hours-per-week	Hours worked per week	94
Native-country	Country of origin of the individual	42
Income	Binary target: <=50K / >50K	2

Table 2.1: Features Present In The Adult Census Income dataset

3. Exploratory Data Analysis

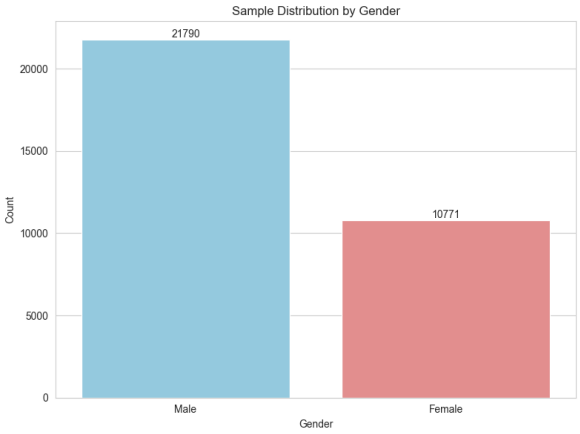
This chapter describes imbalances in the dataset by analyzing record and income distributions across genders.

It is essential to understand these imbalances and identify features where adjustments could mitigate bias because if the models are trained on historically biased data, it is often the case that models produce unfair patterns in the dataset.

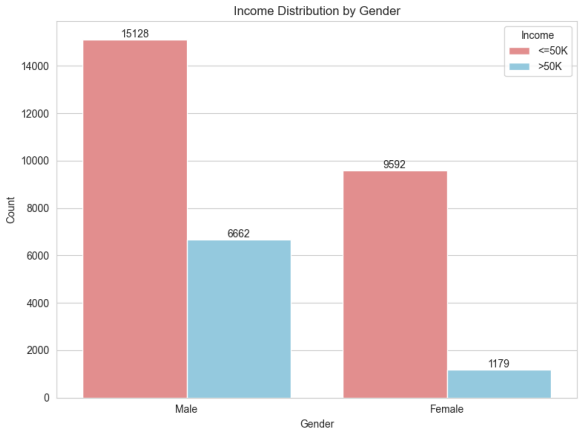
Fig. 3.1a illustrates the overall gender distribution. We can see that male records are roughly two times more common compared to female records, which highlights an imbalance in the dataset and may lead to unfairness if the model is trained naively.

Fig. 3.1b depicts the income distribution by gender, showing whether individuals earn more than $50,000 or less. It is evident that male records dominate both income (high and low) categories. For the high-income group, it is 5.65 times more males than females and for the low-income group, it is 1.58 times more males than females. A straightforward optimization for accuracy would utilize this to naturally discriminate against females, since that is the behaviour of the underlying dataset.

At a broader level, Fig. 3.1 emphasizes the implications of these imbalances. Training models on such data can propagate disparities: for example, the overrepresentation of males in the high-income category may lead a model to associate being male with higher income more often than justified by other variables. This demonstrates how historical bias can translate into algorithmic bias if not addressed carefully.



(a) Sample Distribution by Gender



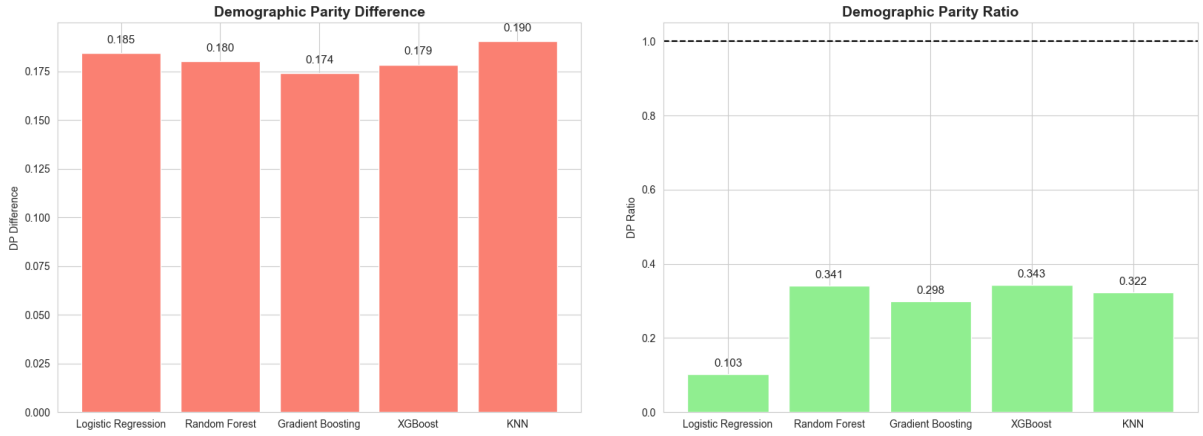
(b) Income Distribution by Gender

Figure 3.1: Exploratory Data Analysis: Income and Gender

4. Measuring Fairness

This chapter evaluates fairness in the *Adult / Census Income* dataset using five machine learning models: KNN, XGBoost, Random Forest, Logistic Regression, and Gradient Boosting, without applying fairness-enhancing techniques. Fairness is measured using Demographic Parity which assess whether groups receive similar positive predictions overall. We also looked into False Positive and False Negative Rates by gender, positive prediction rates, and test accuracy, so that we were able to draw a comprehensive view of model performance and fairness.

Figure 4.1 shows the Demographic Parity (DP) metrics which would help in understanding whether different demographic groups receive positive predictions at similar rates, regardless of their earnings. Fig. 4.1a shows an average DP difference (0.18) across models, which highlights a clear disparity between genders. Fig. 4.1b gives an average DP ratio of approximately (0.28), meaning that the protected group is likely to receive positive predictions at only 28% of the rate of the unprotected group. Notably, Logistic Regression produces the lowest DP ratio (0.17), exhibiting the largest imbalance among the evaluated models.



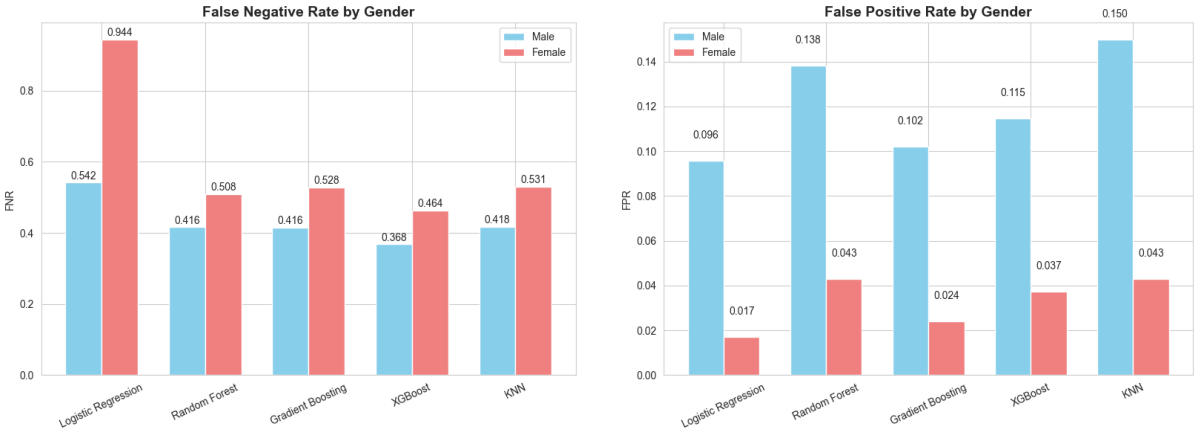
(a) Demographic Parity Difference **(b)** Demographic Parity Ratio

Figure 4.1: Demographic Parity Metrics

Figure 4.2 presents error rates such as False Negative Rate (FNR) and False Positive Rate (FPR) by gender. This helps reveal how prediction errors are distributed across

demographic groups, which is critical for understanding fairness. Fig. 4.2a shows the FNR by gender, which is necessary to analyze how frequently individuals who earn over $50,000 are incorrectly predicted as false, i.e. classified as earning less than $50000. For all models, the FNR is higher for females than for males, which implies that qualified women are systematically misclassified to earn less than equally qualified men. Fig. 4.2b shows the FPR by gender, which is similar to FNR, but this time it shows the proportion of individuals incorrectly predicted as high-income earners.

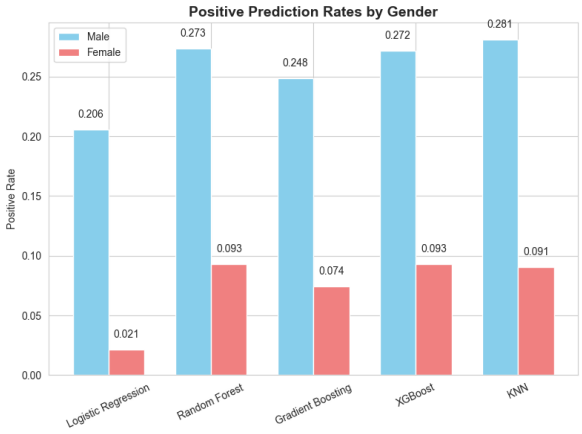
Across the models, the FPR is higher for males than females, which is exactly the opposite case compared to the FNR Fig. 4.2a. We can conclude that men are frequently positively mispredicted, i.e. claiming to earn more $50000 per year despite their qualifications suggesting otherwise.



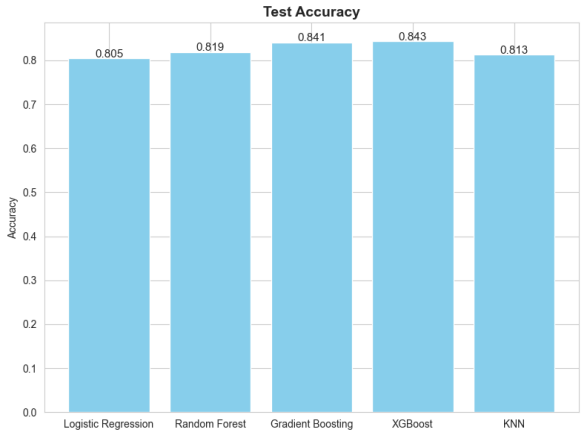
(a) False Negative Rate by Gender **(b)** False Positive Rate by Gender

Figure 4.2: Error Rates by Gender

Fig. 4.3a shows positive prediction rates of income-earning by gender, which reflects differences in decision thresholds and model behavior. From the figure, males are more likely to be predicted as high-income earners than females. Finally, Fig. 4.3b shows that the test accuracy varies across models, with the average accuracy being approximately 0.82. This shows that while our models are good at predicting the income, the FNR (4.2a) and FPR (4.2b) figures showcase the clear disparities and unfairness between the genders.



(a) Positive Prediction Rates



(b) Test Accuracy Across Models

Figure 4.3: Prediction Rates and Model Accuracy

5. Adopting Fairness-Aware Machine Learning

This chapter discusses adopting fairness-aware machine learning models using two models: Logistic Regression and Random Forest. Since gender bias is a significant issue in predictions, we employed in-processing methods, specifically learning with constraints. We used Exponentiated Gradient algorithm with the restriction of enforcing Demographic Parity, which helped the models produce fairer predictions while still achieving high accuracy.

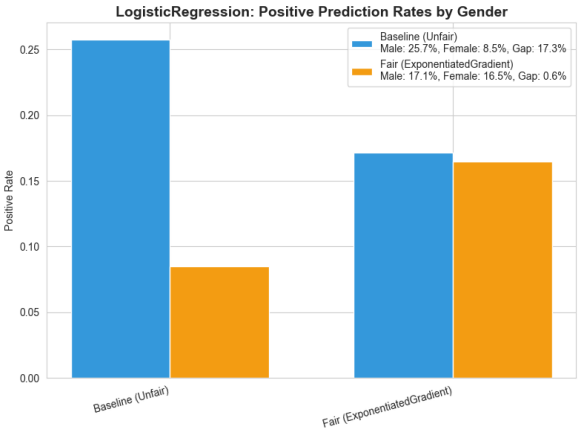
5.1 Logistic Regression

This section presents the results obtained using the Logistic Regression model. We compared the baseline model with the fairness-aware version trained using the Exponentiated Gradient method.

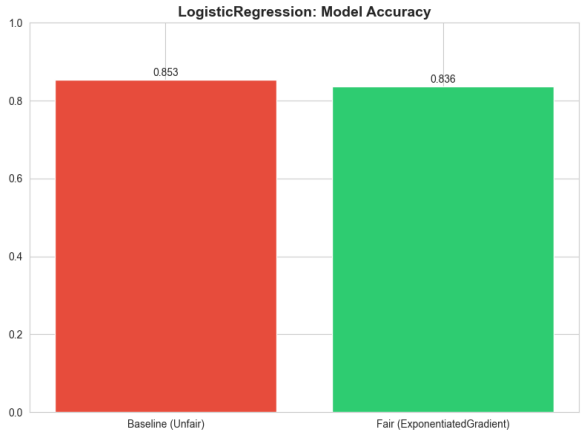
Fig. 5.1a shows that the positive prediction rates are almost balanced, meaning we are able to achieve a minimal gap of 0.6% with the exponentiated gradient from the baseline model, where the initial gap was 17.3%. In terms of accuracy, Fig. 5.1b shows that there is only a slight decrease of 0.02 compared to the baseline.

Fig. 5.1c shows the Demographic Parity Difference, and we can see that there is a substantial reduction in gender bias, dropping to 0.006 from 0.175. Fig. 5.1d depicts the Disparate Impact Ratio improving to nearly 1, which highlights that we can obtain almost perfect parity.

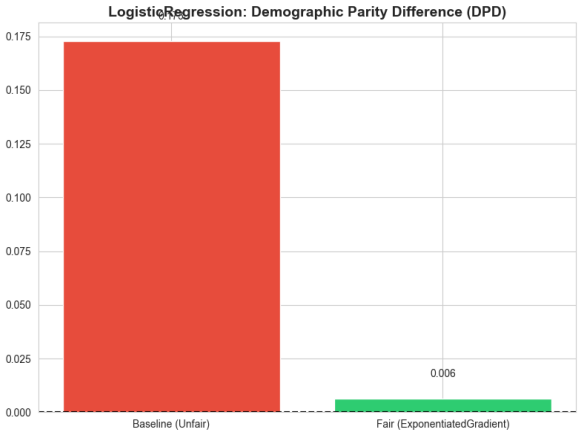
Finally, Fig. 5.1e illustrates the fairness versus accuracy trade-off, it being evident that the fairness was greatly improved, as we were successful in reducing DPD by 96.2%, and improving DPR by 94.3%. Meanwhile, the decrease in model accuracy remained minimal at only around 1.7%.



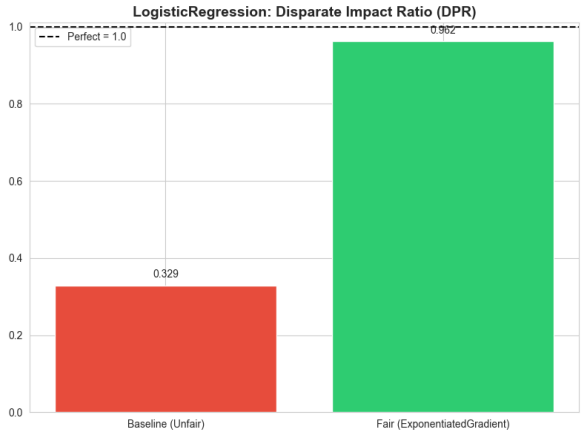
(a) Positive Prediction Rates



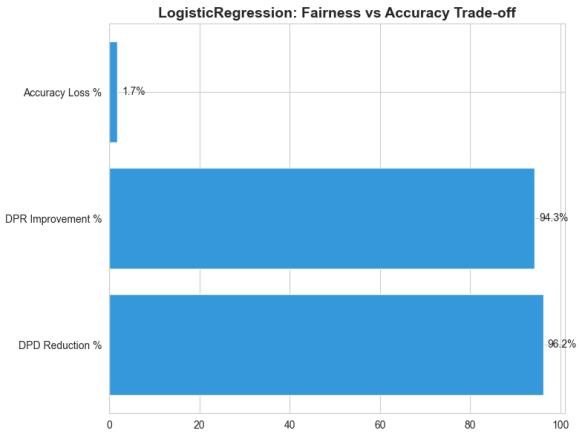
(b) Test Accuracy



(c) Demographic Parity Difference (DPD)



(d) Disparate Impact Ratio (DPR)



(e) Fairness vs Accuracy Trade-off

Figure 5.1: Comparison of baseline and fairness-aware model metrics for Logistic Regression.

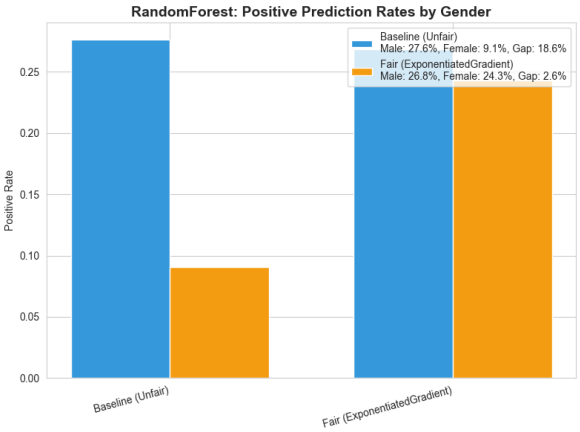
5.2 Random Forest

Similarly, in this section, we present the results obtained using the Random Forest model. We compared the baseline model with the fairness-aware version trained using the Exponentiated Gradient method.

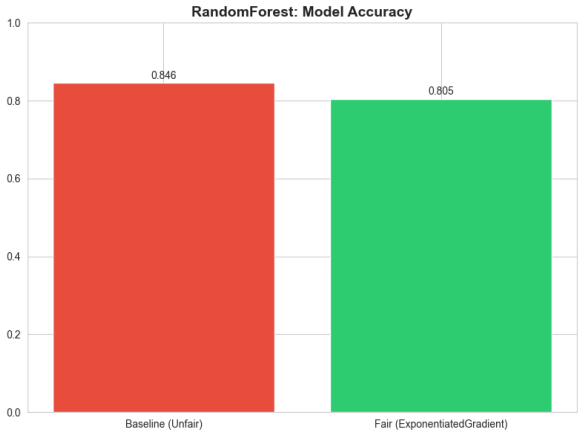
Fig. 5.2a shows that the positive prediction rates are nearly balanced across genders, obtaining a minimal gap of 2.6% with the exponentiated gradient compared to the baseline case of 18.6% minimal gap. Fig. 5.2b shows only a slight decrease of 0.041 in model accuracy compared to the baseline, which is similar to what we saw in the Logistic Regression case (Fig. 5.1b).

Fig. 5.2c highlights the major reduction in gender bias in Demographic Parity Difference, dropping to just 0.026, and Fig. 5.2d depicts the Disparate Impact Ratio (5.2d) improving to 0.905, which indicates almost perfect parity.

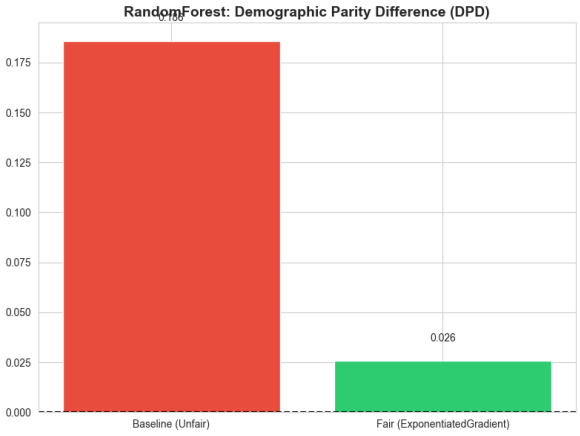
Finally, Fig. 5.2e illustrates the fairness versus accuracy trade-off. We can see that fairness has greatly improved by reducing DPD to 86.2%, and improving DPR to 85.8%. However, model accuracy dropped by 4.1%, which is worse compared to the performance achieved by Logistic Regression (Fig. 5.1e).



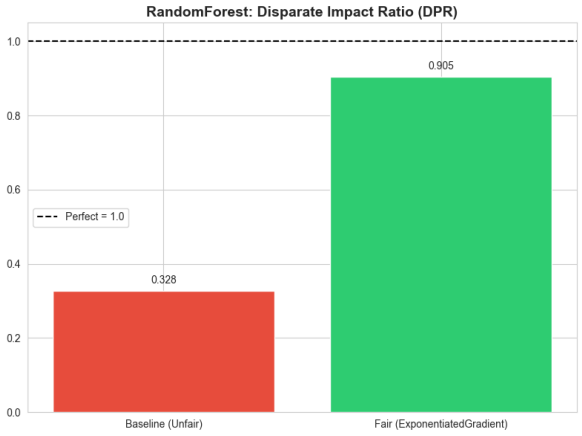
(a) Positive Prediction Rates



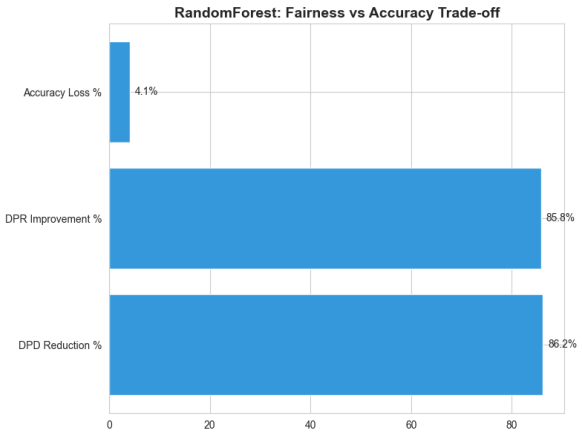
(b) Test Accuracy



(c) Demographic Parity Difference (DPD)



(d) Disparate Impact Ratio (DPR)



(e) Fairness vs Accuracy Trade-off

Figure 5.2: Comparison of baseline and fairness-aware model metrics for Random Forest.

6. Conclusions

In this report, we have shown that fairness is an important variable to consider when designing machine learning systems. Existing datasets may not actually be fair, despite being often referred to as the 'ground truth', and may incorporate undesirable historical biases. Without accounting for such biases and unfairness, dangerous feedback loops may be created that reinforce the unfairness and discrimination that are already present.

We have shown that the baseline unfair models are heavily biased in favor of males in the case of our dataset. We have analyzed this disparity by showcasing the Gender Distribution (Fig. 3.1a) and Income Distribution by Gender (Fig. 3.1b), which revealed that males dominate both the overall dataset (2:1 ratio) and particularly the high-income category (5.65:1 ratio). When measuring fairness the presented baseline models, we observed a consistent gender bias and a Demographic Parity difference of 0.18, and a Disparate Impact Ratio of just 0.28, which indicates that females received positive predictions at only 28% of the rate of the males.

To address these disparities, we implemented fairness-aware machine learning using the Exponentiated Gradient algorithm on two models (Logistic Regression and Random Forest). Both models demonstrated substantial improvements in fairness metrics, with Logistic Regression achieving the best results. Random Forest also showed significant fairness improvements, at the cost of slightly more accuracy loss than Logistic Regression.

These results demonstrate that it is possible to achieve substantial fairness improvements with relatively modest accuracy trade-offs.Of course, there is significant future work that could be done in relation to fairness in such contexts, as the environment we have considered is relatively simple and does not deal with advanced and more nuanced forms of hidden bias in the data.

To conclude, this project demonstrates that fairness-aware machine learning is not only necessary but also practically achievable given careful consideration. As ML systems continue to be used in increasingly high-stakes decision making, the integration of fairness considerations must become standard practice rather than an afterthought. The techniques and insights presented in this report provide a foundation for developing more just and fair automated decision systems.

References

[BK96] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. 1996.

[Spi23] Paul Spiegelhalter. *Fairness and Bias in Machine Learning*. https://www.pythian.com/blog/fairness-and-bias-in-machine-learning. 2023.

[Zli17a] Indre Zliobaite. *Fairness-aware machine learning: a perspective*. 2017. arXiv: 1708.00754 [cs.AI]. URL: https://arxiv.org/abs/1708.00754.

[Zli17b] Indre Zliobaite. "Measuring discrimination in algorithmic decision making". In: *Data Mining and Knowledge Discovery* 31.4 (July 1, 2017), pp. 1060–1089. ISSN: 1573-756X. DOI: 10.1007/s10618-017-0506-1. URL: https://doi.org/10.1007/s10618-017-0506-1.

[Goo25] Google Developers. *Fairness: Evaluating for bias*. https://developers.google.com/machine-learning/crash-course/fairness/evaluating-for-bias. Accessed: 2025-12-14. 2025.