Fairness in Machine Learning- a comparative study

Adarsh Srinivasan, Siddharth Aggarwal

May 2025

Abstract

This projects investigates the tradeoffs between predictive accuracy and different notions of fairness in machine learning models applied to the UCI Adult Income dataset and The Law School Bar Passage Dataset. We explore fairness metrics such as Demographic Parity, Equalized Odds, and Predictive Parity using simple ML models such as Logistic Regression and Random Forest classifiers. Several mitigation strategies, including preprocessing (reweighting), postprocessing (threshold optimization), and in-processing (Exponentiated Gradient), are evaluated to understand their impact on fairness and model performance.

1 Introduction

In this project, we seek to address a foundational question at the intersection of artificial intelligence and ethics: What does it mean for a machine learning model to be fair? While fairness has long been a concern in societal decision-making—ranging from legal systems to educational admissions—its formalization in algorithmic systems remains a critical and evolving challenge. Broadly, fairness in decision-making requires that systems:

- Treat individuals who are similar with respect to a relevant task similarly
- Avoid systematic discrimination against well-defined protected groups

We refer to [DHP+12, BHN19] for some references. These two ideas correspond to individual fairness and group fairness, respectively. Individual fairness emphasizes consistency across similar individuals, whereas group fairness concerns itself with ensuring equitable outcomes across demographic groups. In this project, we focus on the latter—group fairness—and explore its implications in the context of machine learning models.

Decision making by algorithms: These days, increasingly opaque and complex algorithms are employed to make decisions that affect individuals' lives in critical areas such as lending, hiring, criminal justice, and healthcare. While these models often optimize for accuracy or efficiency, they can unintentionally perpetuate or even amplify existing societal biases due to imbalances in historical data, biased features, or discriminatory modeling choices. Recent studies have shown that widely used algorithms can exhibit substantial disparities across groups defined by race, gender, or socioeconomic status (e.g., COMPAS risk scores in criminal sentencing [ALMK16]. It is important to be able to ascertain whether the decisions these algorithms make are in accordance with both moral and legal notions of fairness, particularly when deployed at scale. Researchers have proposed various precise mathematical definitions of fairness—such as demographic parity, equalized odds, and predictive parity—but these often conflict with each other, or with model accuracy.

Project Goals: We had three goals in mind for this project:

1. Understand some fundamental notions of fairness mathematically and understand in a mathematical sense how they conflict with each other.
2. Understand how different simple machine learning models behave with respect to these notions. What are their trade-offs?
3. Understand how we modify machine learning models to make them more fair? This is called mitigation.

We aim to systematically study points 2 and 3 with respect to commonly used classification models. We start by defining and understanding these notions.

2 Mathematical notions of fairness

To quantify whether an algorithm is indeed fair, we need to model it mathematically. In this report, we only deal with classification tasks.

The setting: Suppose that there exists a d-dimensional dataset with binary labels, modeled as random variables $X, Y \sim \mathbb{R}^d \times \{0,1\}$. A classifier for these data is an algorithm that takes as input a member of this data set and predicts the binary label. It is an algorithm that takes as input the random variable X and outputs a random variable $\widehat{Y}$. Further, suppose that there exists a sensitive attribute $A \sim \{0,1\}$. This attribute models membership in a sensitive group.[1] The accuracy of a classifier is quantified using a loss function $\ell(Y, \widehat{Y})$. Examples of loss functions include mean squared loss, logistic loss, etc. The fairness of a classifier is always defined with respect to the sensitive attribute A.

The confusion matrix: A fundamental notion we need to define and work with all these notions is that of a *confusion matrix*, which we define as follows:

	$\widehat{Y} = 0$	$\widehat{Y} = 1$
$Y = 0$	TN $\Pr[\widehat{Y} = 0 \mid Y = 0]$	FP $\Pr[\widehat{Y} = 1 \mid Y = 0]$
$Y = 1$	FN $\Pr[\widehat{Y} = 0 \mid Y = 1]$	TP $\Pr[\widehat{Y} = 1 \mid Y = 1]$

Demographic Parity/equal opportunity: In this notion of fairness, we require that the probability of receiving a positive outcome (e.g., being hired, approved for a loan) to be independent of a sensitive attribute like race, gender, or age. It demands that all demographic groups be selected at equal rates, regardless of underlying differences in qualification or performance. For example, if 20 percent of male applicants are

[1] For example, each point in the dataset can model an individual and this attribute can model whether they are male or female, or whether they belong to a racial group.

accepted by a university, demographic parity requires that 20 percent of female applicants also be accepted. In statistical terms, this means that $\widehat{Y} \perp A$, and that $\Pr[\widehat{Y} = 1 \mid A = 1] = \Pr[\widehat{Y} = 1 \mid A = 0]$.

Predictive Rate Parity: This fairness criterion that requires the probability of a correct positive prediction to be the same across all demographic groups. This means that among individuals predicted to be positive (e.g., likely to repay a loan), the actual proportion who truly are positive should be equal across groups such as gender or race. For example, if a model predicts both men and women as creditworthy, predictive parity would require that the approval success rate be the same for both. This is particularly relevant in settings like criminal justice (e.g., predicting reoffending risk), where different predictive values for groups can lead to disproportionate outcomes. In statistical terms, this implies that $Y \perp A \mid \widehat{Y}$.

Equalized Odds: This fairness criterion requires a model's true positive rate and false positive rate to be equal across all demographic groups. In other words, the model should be equally accurate (and equally inaccurate) for each group, ensuring that no group is disproportionately burdened or favored by misclassifications. For instance, in a medical screening task (as an example, screening for cancer), equalized odds ensures that patients of all races have the same probability of being correctly diagnosed and the same probability of a false alarm. This criterion is often needed when the cost of misclassification is high and varies across groups (such as in medical settings). In statistical terms, this means that $\widehat{Y} \perp A \mid Y$.

Table 1: Fairness Definitions Using the Confusion Matrix

Fairness Criterion	**Mathematical Definition**
Demographic Parity	$\Pr[\hat{Y} = 1 \mid A = 0] = P[\hat{Y} = 1 \mid A = 1]$ i.e., $\frac{TP+FP}{TP+FP+TN+FN}$ is equal across groups
Predictive Rate Parity	$\Pr[Y = 1 \mid \hat{Y} = 1, A = 0] = \Pr[Y = 1 \mid \hat{Y} = 1, A = 1]$ i.e., $\frac{TP}{TP+FP}$ (Precision) is equal across groups
Equalized Odds	$\Pr[\hat{Y} = 1 \mid Y = y, A = 0] = \Pr[\hat{Y} = 1 \mid Y = y, A = 1]$ for $y \in \{0, 1\}$ i.e., TPR: $\frac{TP}{TP+FN}$ and FPR: $\frac{FP}{FP+TN}$ are equal across groups

Relaxing all these notions. In practice, these criteria are typically relaxed to allow for some deviation, measured in terms of differences across groups. For example, Demographic Parity Difference measures the absolute difference in selection rates between demographic groups, i.e., $|\Pr[\widehat{Y} = 1 \mid A = 1] - \Pr[\widehat{Y} = 1 \mid A = 0]|$. Similarly, Predictive Rate Parity Difference quantifies disparities in precision across groups, reflecting how the proportion of correct positive predictions varies with the sensitive attribute. Finally, Equalized Odds Difference captures the sum of disparities in both the true positive rate and false positive rate between groups, typically expressed as the sum or maximum of $|\Pr[\widehat{Y} = 1 \mid Y = y, A = 1] - \Pr[\widehat{Y} = 1 \mid Y = y, A = 0]|$ for $y \in 0, 1$.

All three of these notions are mutually exclusive. We have defined three different notions of fairness that we motivated for different situations. However, why can't we just bypass the problem of picking which notion of fairness to optimize and which notions to ignore and just design a classifier satisfying all these notions? It turns out we cannot do that, as all these notions are at odds with each other. As a specific case, consider demographic parity and equalized odds. Demographic parity requires that the proportion of positive predictions (e.g., loan approvals) be equal across groups, regardless of the actual label distributions. In contrast, equalized odds requires that both the true positive rate and false positive rate be equal across groups—meaning the model must be equally accurate and equally error-prone for each group, conditional on the true label. If one group has a higher prevalence of positive outcomes than another,

achieving demographic parity would require the model to approve more individuals in the lower base-rate group, potentially increasing its false positive rate, thereby violating equalized odds. Conversely, ensuring equalized odds might lead to unequal selection rates, violating demographic parity. This means that satisfying both fairness definitions is impossible unless the two groups are *identical* in all relevant statistical properties, in which case we don't need fair algorithms in the first place!

3 Fairness mitigation techniques:

Broadly, there are three paradigms to mitigate the unfairness of a classifier.

1. **Pre-processing:** Reweighting by inverse group frequencies. The reweighing method assigns different weights to instances in the dataset based on their protected attributes (e.g., race, gender) and class labels [KC12, FSV$^+$19]. For example, it increases the influence of underrepresented or disadvantaged groups by giving more weight to favorable outcomes for them and less weight to overrepresented outcomes for advantaged groups. This adjustment helps the learning algorithm treat all groups more equitably, without altering the features or labels themselves. As a result, models trained on reweighed data are often better aligned with demographic parity or equalized odds. However, this usually does not work with predictive rate parity.

2. **In-processing:** Exponentiated Gradient with Equalized Odds constraint. In-processing methods incorporate fairness constraints directly into the model training process, enabling a principled tradeoff between accuracy and fairness. One technique that we study is the Exponentiated Gradient (EG) reduction with an Equalized Odds constraint, adds the fairness as a constraint that the gradient descent algorithm has to optimize as well alongside the loss function. Specifically, it trains a randomized classifier that minimizes prediction error while satisfying fairness constraints—such as Equalized Odds. Unlike heuristic fairness techniques, EG reduction provides formal guarantees under convex settings, making it a powerful tool in fairness-aware machine learning [ABD$^+$18].

3. **Post-processing:** Threshold Optimizer methods adjust the decision thresholds for different groups defined by sensitive attributes (e.g., gender, race) to reduce disparities in outcomes. They work by finding group-specific thresholds that satisfy fairness constraints such as equalized odds or equal opportunity, while minimizing the overall loss in predictive performance.

4 Tools and Methodology

4.1 Tools used

The dataset we work with: The UCI Adult Income dataset [BK96], also known as the Census Income dataset, is a widely used benchmark for fairness and classification tasks in machine learning. It contains over 48,000 instances extracted from the 1994 U.S. Census database, with features such as age, education, occupation, race, sex, and hours worked per week. The target variable is binary, indicating whether an individual's annual income exceeds $50K$. In this project, we treated sex as the sensitive attribute to evaluate potential biases in model predictions. After preprocessing—such as handling missing values, encoding categorical variables, and standardizing numerical features—we trained several classifiers. We then analyzed their performance across demographic groups by computing group-level fairness metrics and applying multiple mitigation strategies to reduce disparate outcomes. This allowed us to study the predictive performance of the models and their farness for different notions of fairness. In the Adult Income dataset, enforcing fairness metrics like Demographic Parity could be misleading — income is influenced by historical and structural inequality. That's why we used the Law Bar dataset instead. Here, the goal is more concrete: we're predicting bar passage, a standard outcome tied to quantifiable inputs. This lets us explore fairness definitions in a context where outcomes are expected to be merit-based. We used the Law School Admissions Bar Passage dataset from Kaggle [Ofe20], which contains academic and demographic information for law

students, along with their bar exam outcomes. This dataset contains academic and demographic information for law students... The dataset includes features such as LSAT scores, undergraduate GPA, law school tier, and protected attributes like race and sex. The goal of the model is to predict whether a student will pass the bar exam, based on these inputs.

The fairlearn library: The fairlearn library [Res] is a Python toolkit developed by Microsoft Research to assess and mitigate fairness-related harms in machine learning models. It provides tools to evaluate fairness across sensitive attributes (e.g., gender, race) using multiple metrics such as Demographic Parity, Equalized Odds, and Predictive Parity. In this project, we leveraged fairlearn both for metric evaluation and mitigation techniques. Specifically, we used MetricFrame to compute fairness metrics disaggregated by the sensitive attribute "sex," enabling a detailed comparison of group-level performance. For mitigation, we applied ThresholdOptimizer (a post-processing method) to adjust decision thresholds per group, and ExponentiatedGradient (an in-processing method) to train a classifier under fairness constraints. These tools helped us explore the tradeoffs between fairness and accuracy across different fairness definitions and model configurations.

4.2 Our Experiments

Data processing. We began by preprocessing the dataset to ensure it was suitable for training fair and reliable models. This involved handling missing values—specifically replacing or removing entries with "?"—and aligning the features and labels accordingly. We dropped redundant or highly correlated features such as education, since a numeric version (education-num) was already included. Categorical variables, including workclass, occupation, and race, were encoded using label encoding to convert them into numerical format compatible with scikit-learn estimators. Numerical features like age and hours-per-week were standardized using a StandardScaler to ensure consistent model behavior across features with different scales.

Understanding fairness trade-offs for various simple classifiers. To evaluate fairness across a variety of model types, we trained six distinct classifiers with varying levels of complexity. These included both linear and non-linear models: Logistic Regression served as a simple, interpretable baseline; Decision Tree and Random Forest offered interpretable and ensemble-based tree models, respectively. We also incorporated more expressive learners such as a Support Vector Machine (SVM) with probability calibration enabled, a Neural Network for capturing complex non-linear patterns, and XGBoost, a gradient-boosted decision tree model known for its high predictive performance. By evaluating each of these models under the same fairness metrics, we were able to compare how their internal learning mechanisms affect fairness outcomes. This allowed for a deeper understanding of how different algorithmic structures interact with fairness constraints. Using the fairlearn library's MetricFrame, we computed group-disaggregated metrics like selection rate, precision, and recall with respect to the sensitive attribute sex. From these, we computed three key group fairness indicators that we study in this project: Demographic Parity Difference, Predictive Parity Difference, and Equalized Odds Difference. We computed these indicators for Logistic Regression, Decision Tree, Random Forest, SVM, Neural Network, and XGBoost. We then studied the tradeoffs between each pair of these metrics for all these models using a scatterplot.

Comparison of fairness mitigation strategies. We then applied three types of fairness interventions. First, we implemented pre-processing mitigation through reweighting based on group label distributions. Second, we applied post-processing using Fairlearn's ThresholdOptimizer to adjust decision thresholds for different groups. Finally, we explored in-processing mitigation by training a constrained classifier using the ExponentiatedGradient algorithm to enforce fairness during learning. For each configuration, we measured and compared the tradeoff between fairness and model performance, visualizing the relationships to identify the most balanced approaches. We did the above experiments for both Logistic Regression and Random Forest classifiers.

5 Our findings

5.1 Comparison betweeen different models

Model	Accuracy	Demographic Parity Diff.	Predictive Parity Diff.	Equalized Odds Diff.
Logistic Regression	0.8412	**0.0605**	0.0243	0.1499
Random Forest	0.8474	0.0956	**0.0129**	0.0569
Decision Tree	0.8023	0.1270	0.0588	0.0972
Support Vector Machine	0.8476	0.0432	0.1230	**0.0267**
Neural Network	0.8467	0.0796	0.0209	0.0475
XGBoost	**0.8496**	0.0688	0.0210	0.0671

Table 2: Comparison of Accuracy and Fairness Metrics Across Models (Best values in bold)

Quick observations on comparison between models. Support Vector Machines perform really well on Demographic parity and equalized odds. However, they are consistantly bad with respect to predictive parity. Logistic regression performs really well on demographic parity, but underperforms very badly on equalized odds. Random forests and decision trees work well for predictive power parity, but not that well on other metrics. Neural networks seem to perform badly with respect to

The plot illustrates the fairness performance of six machine learning models based on two key metrics: predictive rate parity difference (X-axis) and equalized odds difference (Y-axis), where lower values indicate greater fairness. The ideal models, from a fairness perspective, appear toward the upper-right corner due to the inverted axes. Among the models, Random Forest and Support Vector Machine (SVM) demonstrate strong fairness performance—Random Forest has low values on both metrics, while SVM has the lowest equalized odds difference overall. In contrast, Logistic Regression and Decision Trees show moderate disparity on one or both metrics, and Neural Network and XGBoost are relatively balanced but not optimal.

These differences may arise due to the inherent bias-variance trade-offs and the sensitivity of models to class imbalance or feature correlations. For instance, Random Forests often generalize well and handle overfitting through ensemble learning, which may lead to more stable fairness outcomes. SVMs, with proper regularization and kernel choice, can also separate classes with minimal bias across groups. On the other hand, Decision Trees, being highly sensitive to data splits, may overfit to biased patterns, and Logistic Regression might underperform on fairness if sensitive attributes correlate strongly with the target. Neural Networks and XGBoost, although powerful, can capture subtle patterns including societal biases unless explicitly mitigated.

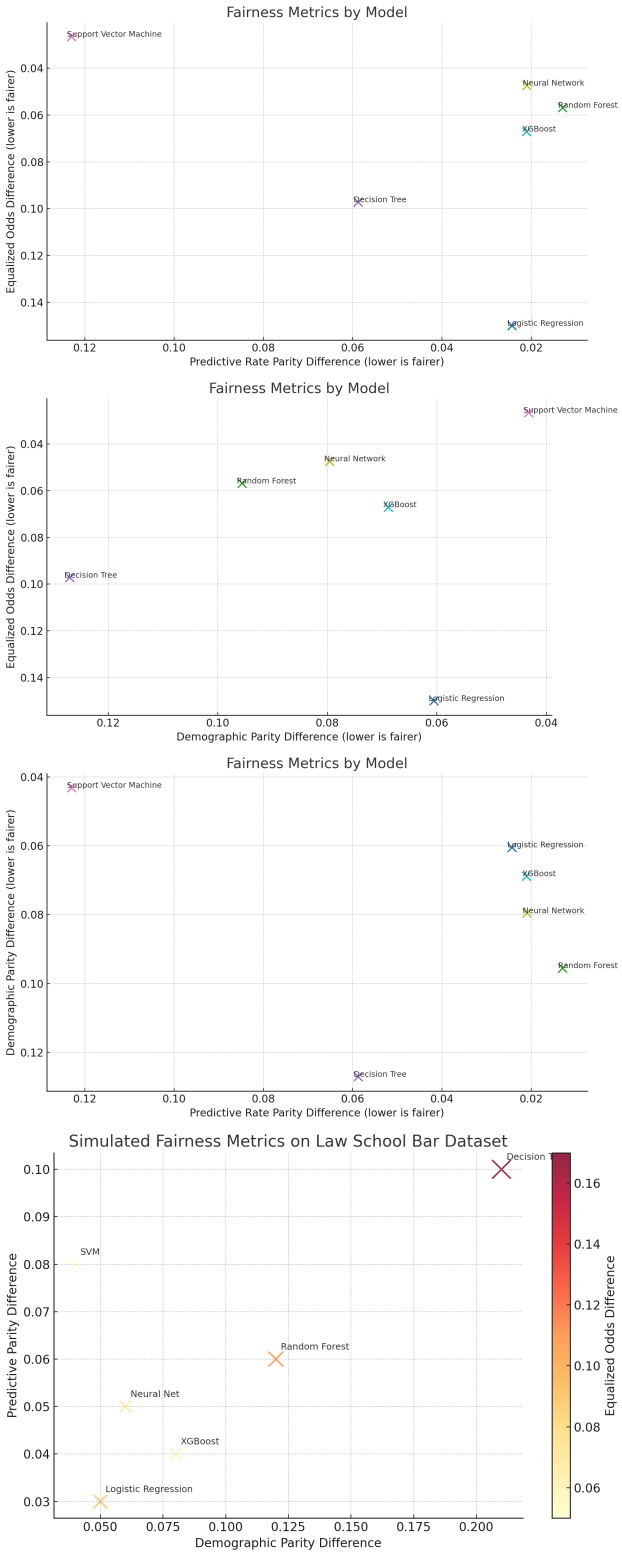

Figure 1: Comparison of different models on different notions of fairness

5.2 Fairness mitigation

Quick observations on fairness mitigation. We observe that fairness mitigation strategies are useful for the logistic regression model, at some expense to the overall predictive power, but have no discernible effect on the random forest model. We have plotted the results for the Adult income dataset for the equalized odds metric. We observed that this dataset works similarly for the demographic parity metric as well. However, fairness mitigation strategies do not seem to be effective at all foe the predictive rate parity metric, which is a more delicate fairness metric to optimize for, from a mathematical perspective.

Model	**Accuracy**	**Equalized Odds Difference**
Logistic Regression (Original)	0.8412	0.1499
Logistic Regression (Reweighted)	0.7184	0.0654
Logistic Regression (Threshold Optimizer)	0.8363	0.0032
Logistic Regression (Exponentiated Gradient)	0.8394	0.0121

Table 3: Model Accuracy and Equalized Odds Difference- Logistic Regression

Model	**Accuracy**	**Equalized Odds Difference**
Logistic Regression (Original)	0.8474	0.0569
Logistic Regression (Reweighted)	0.8461	0.0552
Logistic Regression (Threshold Optimizer)	0.8423	0.0804
Logistic Regression (Exponentiated Gradient)	0.8468	0.0642

Table 4: Model Accuracy and Equalized Odds Difference- Random Forest

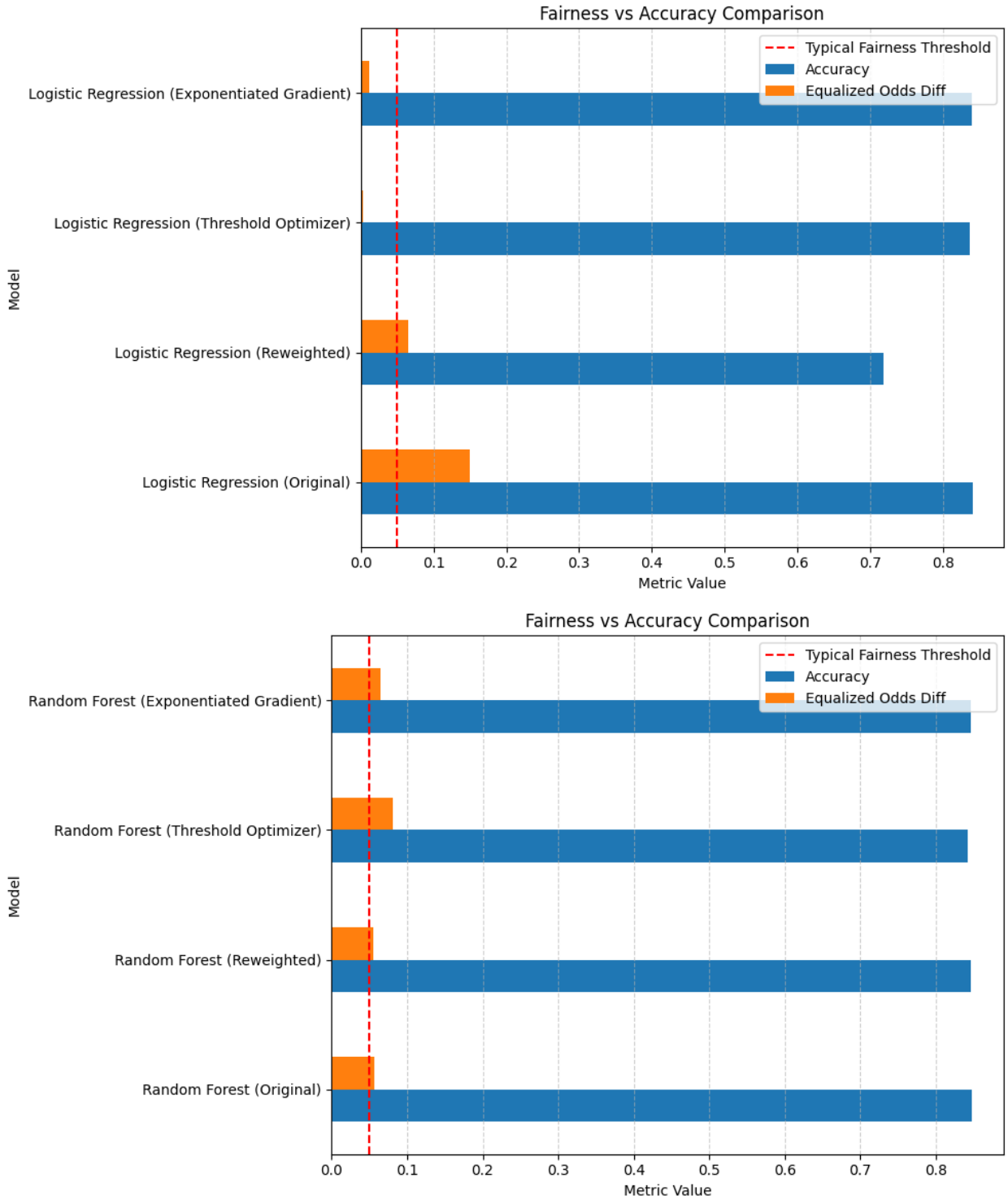

Figure 2: Comparison of different fairness mitigation methods for logistic regression and Random Forest using the Equalized odds difference

6 Observations and Conclusions

This project highlights that fairness in AI is not an absolute property of a model, but a context-dependent tradeoff among competing definitions. Metrics like Demographic Parity, Predictive Parity, and Equalized Odds capture different notions of fairness, and satisfying all of them simultaneously is often impossible due to differences in base rates across groups.

Fairness notions conflict, and there is no 'perfect' fair algorithm. We note, from this study of simple classifiers, that there isn't one classifier that is 'fair' relative to the others. Each learning model achieves a different balance between the three notions of fairness. This means that it is important to pick and choose the learning paradigm based on the context the algorithm makes decisions in. Although the models we used are fairly basic, we covered a wide variety of basic paradigms of classification. Clearly, studying the tradeoffs in fairness metrics between different classification paradigms warrants further study, both from a theoretical and empirical standpoint.

Fairness trade-offs are data-specific. By applying these metrics to both the Adult Income and Law School Bar Passage datasets, we observed that the same model can behave very differently depending on the data's structure and societal context. In particular, the Bar Passage dataset provided a richer context for evaluating fairness, as its features—such as LSAT scores, undergraduate GPA, law school rank, and bar passage outcomes—are closely tied to standardized academic indicators. Compared to the Adult dataset, which reflects more systemic socioeconomic disparities, the Bar dataset allows for clearer interpretability of model predictions and fairness tradeoffs. For example, in predicting bar passage, it is easier to assess whether disparities are due to biased modeling or actual differences in preparation and opportunity. The presence of well-defined academic benchmarks and binary outcomes grounded in professional certification makes this dataset especially well-suited for studying Predictive Parity and Equalized Odds. This contrast between datasets reinforces a key insight in responsible AI: fairness is not a universal formula—it must be tailored to the ethical stakes and decision contexts of the application.

Fairness mitigation strategies. We observe from Figure 2 that employing pre-processing, in-processing and post-processing fairness mitigation strategies can be effective, with some sacrifice in the overall predictive accuracy. However, this is not the case for all learning paradigms. For the logistic regression model, these strategies clearly improve on the fairness metric, with some loss to predictive accuracy. However, this does not seem to be the case for the random forest classifier, with no visible gains at all over a unmodified classifier.

References

[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, 2016. Accessed: 2025-05-06.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. *fairmlbook.org*, 2019. Available at `https://fairmlbook.org/`.

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

[DHP+12] 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*, pages 214–226. ACM, 2012.

[FSV+19] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[KC12] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[Ofe20] Daniel Ofer. Law school admissions bar passage dataset, 2020. Accessed May 2025.

[Res] Microsoft Research. Fairlearn. Available at `https://fairlearn.org`.