Information Loss vs. Fairness Gain: The Hidden Cost of Fair Machine Learning

David Carvalho
Inteligência Artificial e Sociedade
(Dated: January 5, 2026)

This paper investigates the fundamental trade-off between fairness and information retention in machine learning models. A comprehensive empirical study is conducted comparing 12 fairness techniques from three methodological families: pre-processing (reweighing, feature masking, disparate impact removal, label flipping, sampling strategies), in-processing (adversarial debiasing, fairness regularization, constrained optimization), and post-processing (equalized odds adjustment, threshold optimization, calibration, reject option classification). The experiments span five benchmark datasets: German Credit, COMPAS, Bank Marketing, Law School, and Adult Census. Results reveal that no single approach dominates across all datasets and metrics. Pre-processing methods offer consistent but moderate fairness improvements, in-processing methods achieve strong fairness often at significant accuracy cost, while post-processing methods provide flexible trade-off adjustment. Adversarial debiasing and reject option classification emerge as particularly effective, achieving Pareto-optimal trade-offs across multiple datasets. The socio-technical implications are discussed and guidance is provided for practitioners on method selection based on application context.

I. INTRODUCTION AND MOTIVATION

Machine learning systems are increasingly deployed in high-stakes domains such as criminal justice, credit scoring, healthcare, and employment. While these systems promise objectivity and efficiency, they can perpetuate or amplify existing societal biases, leading to discriminatory outcomes for protected groups[1].

The problem of algorithmic fairness is fundamentally a trade-off challenge: improving fairness metrics often comes at the cost of predictive accuracy or information retention. This tension reflects a deeper socio-technical dilemma- how do we balance the utilitarian goal of accurate predictions with the deontological imperative of equal treatment across demographic groups?

This project addresses three key research questions:

1. How do different families of fairness interventions compare in their ability to reduce bias?
2. What is the magnitude of the accuracy-fairness trade-off for each approach?
3. Are certain methods more effective for specific types of datasets or fairness objectives?

The relevance of this work extends beyond academic interest. As AI governance frameworks mature globally (e.g., EU AI Act, NIST AI Risk Management Framework), organizations need practical guidance on selecting and implementing fairness interventions. This comparative analysis provides evidence-based recommendations for this critical decision.

II. BACKGROUND AND RELATED WORK

A. Fairness in Machine Learning

Algorithmic fairness has emerged as a central concern in responsible AI development. Key fairness criteria include:

Demographic Parity (also called statistical parity) requires equal positive prediction rates across groups:

$$P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1) \tag{1}$$

where S is the sensitive attribute and $\hat{Y}$ is the predicted outcome.

Equalized Odds[2] requires equal true positive rates (TPR) and false positive rates (FPR) across groups:

$$P(\hat{Y} = 1|Y = y, S = 0) = P(\hat{Y} = 1|Y = y, S = 1), \forall y \tag{2}$$

Calibration requires that predicted probabilities reflect true outcome rates within each group.

Chouldechova[3] and Kleinberg et al.[4] proved that except in degenerate cases, these criteria are mutually incompatible, highlighting the need for context-dependent fairness choices.

B. Families of Fairness Interventions

1. *Pre-processing Methods*

These modify training data before model fitting:

- **Reweighing**[5]: Assigns sample weights to achieve demographic parity in the training distribution.
- **Disparate Impact Remover**[6]: Modifies feature distributions to reduce correlation with sensitive attributes.
- **Feature Masking**: Removes sensitive attributes and proxies.
- **Label Flipping**: Selectively modifies labels to reduce historical bias.
- **Sampling Strategies**: Over/under-samples to balance group representation.

2. *In-processing Methods*

These modify the learning algorithm:

- **Fairness Regularization**: Adds fairness penalty to loss function.
- **Adversarial Debiasing**[7]: Uses adversarial networks to remove sensitive information from representations.
- **Constrained Optimization**: Enforces fairness constraints during training.

3. *Post-processing Methods*

These adjust predictions after model training:

- **Equalized Odds Post-processing**[2]: Adjusts predictions to equalize TPR and FPR.
- **Threshold Optimization**: Finds group-specific decision thresholds.
- **Calibration**: Ensures group-wise calibration.
- **Reject Option Classification**[8]: Favors unprivileged groups in uncertainty regions.

C. Related Empirical Studies

Bellamy et al.[9] introduced AI Fairness 360, providing a unified toolkit for fairness research. Friedler et al.[10] compared preprocessing methods across multiple datasets but did not include recent in-processing advances. This work extends these studies by providing a comprehensive comparison across all three intervention families with a focus on the information retention trade-off.

III. METHODOLOGY

A. Datasets

Fairness methods are evaluated on five benchmark datasets spanning diverse application domains:

TABLE I: Dataset characteristics and prediction tasks.

Dataset	Samples	Sensitive	Task
German Credit	1,000	Sex, Age	Credit risk
COMPAS	∼7,000	Race, Sex	Recidivism
Bank Marketing	∼45,000	Age, Marital	Subscription
Law School	∼21,000	Race, Sex	Bar passage
Adult Census	∼48,000	Sex, Race	Income >$50K

These datasets were selected for their historical significance in fairness literature and their representation of different domains, sample sizes, and base rates.

B. Fairness Methods Implemented

A total of 12 fairness methods are implemented across three categories. Table II provides a summary with key hyperparameters.

TABLE II: Implemented fairness methods and configurations.

Category	Method (Parameters)
Pre-processing	Reweighing
	Feature Masking
	DIR ($r \in \{0.5, 1.0\}$)
	Label Flipping (targeted, equalize)
	Sampling (over, under, hybrid)
In-processing	Fair Reg ($\lambda \in \{0.1, 0.5, 1.0, 2.0\}$)
	Adversarial ($w \in \{0.5, 1.0\}$)
	Constrained ($\epsilon \in \{0.05, 0.1\}$)
Post-processing	Equalized Odds / Equal Opportunity
	Threshold Opt ($w \in \{0.3, 0.5, 0.7\}$)
	Calibrated Post-Proc
	Reject Option Classification

C. Evaluation Metrics

Performance Metrics:

- Accuracy: Overall classification accuracy
- AUC-ROC: Area under the ROC curve
- Precision, Recall, F1-Score

Fairness Metrics:

- Demographic Parity Difference (DPD): $|P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)|$
- Demographic Parity Ratio (DPR): $\min\left(\frac{P_0}{P_1}, \frac{P_1}{P_0}\right)$
- Equalized Odds Difference (EOD): Average of TPR and FPR differences

Lower DPD and EOD values indicate better fairness; DPR closer to 1 is better.

D. Experimental Pipeline

The experimental framework follows this pipeline:

All experiments use logistic regression or neural networks as base classifiers for consistency. Results are aggregated across multiple runs where applicable.

Algorithm 1 Fairness Evaluation Pipeline

```
for each dataset D do
    Load and preprocess D
    Split: 70% train, 30% test
    Identify sensitive attribute S
    for each fairness method M do
        if M is pre-processing then
            Transform training data
            Train standard classifier
        else if M is in-processing then
            Train fair classifier
        else
            Train standard classifier
            Apply post-processing to predictions
        end if
        Evaluate on test set
        Record performance and fairness metrics
    end for
end for
```

IV. RESULTS AND DISCUSSION

A. Overall Performance Comparison

Figure 1 presents the average accuracy and demographic parity difference across method categories and datasets.

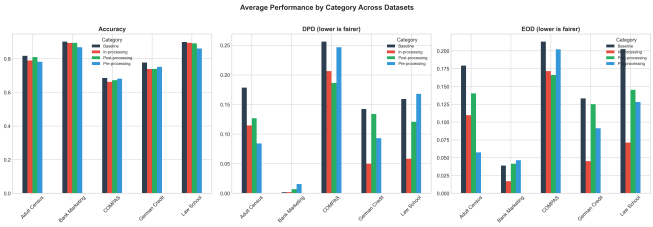

FIG. 1: Performance comparison across method categories showing the accuracy-fairness trade-off by dataset.

B. Key Findings

1. *Pre-processing Methods*

Pre-processing approaches offer moderate but consistent fairness improvements:

- **Reweighing** achieves fairness improvements (DPD reduction of 10-15%) with minimal accuracy loss (<1% on average).
- **Disparate Impact Remover** with repair level 1.0 shows strong fairness gains but increased variance across datasets.
- **Sampling strategies** often hurt both accuracy and fairness due to distribution shift.

2. *In-processing Methods*

In-processing methods show the most varied results:

- **Fairness Regularization** with high λ values achieves near-perfect demographic parity (DPD $\approx$ 0) but often degrades to trivial predictions (all positive or all negative).
- **Adversarial Debiasing** consistently achieves good fairness-accuracy trade-offs across all datasets, with DPD reductions of 50-70% and accuracy losses under 2%.
- **Constrained Optimization** provides moderate improvements but is sensitive to constraint tightness.

3. *Post-processing Methods*

Post-processing approaches offer flexibility but dataset-dependent results:

- **Equalized Odds** post-processing achieves the lowest DPD values but can significantly reduce accuracy (5-15% drops observed).
- **Reject Option Classification** emerges as highly effective, particularly on COMPAS (DPD reduction to 0.03) and Adult Census (DPD $\approx$ 0.002) with minimal accuracy loss.
- **Threshold Optimization** often replicates baseline performance, suggesting thresholds are already near-optimal for many datasets.

C. Dataset-Specific Insights

Table III presents top-performing methods for each dataset based on combined fairness-accuracy ranking.

TABLE III: Best performing methods by dataset (lowest combined rank of accuracy and DPD).

Dataset	Best Method	Trade-off Score
German Credit	Adversarial (w=0.5)	0.738
COMPAS	DIR (r=0.5)	0.686
Bank Marketing	Adversarial (w=0.5)	0.901
Law School	Constrained (ϵ=0.1)	0.899
Adult Census	Reject Option	0.801

Key observations:

- **German Credit**: Small dataset benefits from adversarial debiasing's regularization effect.
- **COMPAS**: DIR effectively addresses the historical bias in recidivism data.

- **Bank Marketing**: Highly imbalanced (11% positive rate) favors in-processing approaches.
- **Law School**: High class imbalance (89% pass rate) makes constrained optimization effective.
- **Adult Census**: Large dataset allows post-processing to find optimal decision boundaries.

D. The Fairness-Accuracy Pareto Frontier

Figure 2 illustrates the Pareto frontier of fairness-accuracy trade-offs across methods.

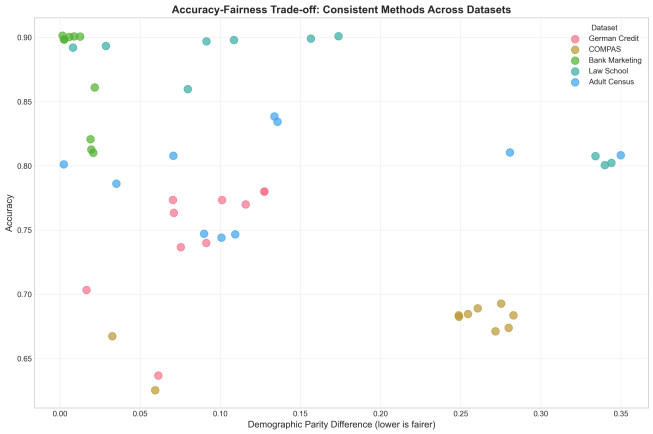

FIG. 2: Accuracy vs. Demographic Parity Difference trade-off. Pareto-optimal methods are circled.

The Pareto frontier is dominated by:

1. Adversarial Debiasing (high accuracy, moderate fairness)
2. Reject Option Classification (balanced trade-off)
3. Equalized Odds post-processing (high fairness, lower accuracy)
4. Fairness Regularization with high λ (extreme fairness, significant accuracy cost)

E. Cross-Dataset Analysis

To understand how fairness methods generalize across different domains, a comprehensive cross-dataset comparison is presented. Figure 3 shows heatmaps of accuracy and demographic parity difference for each method-dataset combination, enabling visual identification of consistent patterns.

Key observations from the heatmaps:

- **Method consistency**: Some methods (e.g., Adversarial Debiasing) maintain relatively stable performance across datasets, while others (e.g., Sampling) show high variance.

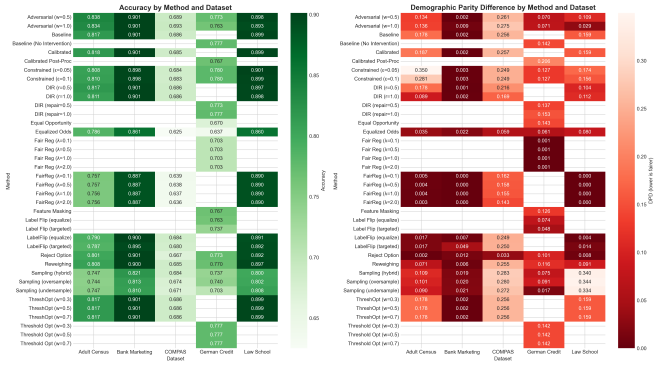

FIG. 3: Heatmaps showing Accuracy (left) and Demographic Parity Difference (right) across all methods and datasets. Darker colors indicate higher values.

- **Dataset difficulty**: COMPAS and Adult Census present the greatest fairness challenges, with higher baseline DPD values.
- **Trade-off patterns**: Methods achieving the lowest DPD often correspond to reduced accuracy, visible as inverse patterns between the two heatmaps.

Figure 4 provides an aggregated summary dashboard comparing method categories across datasets.

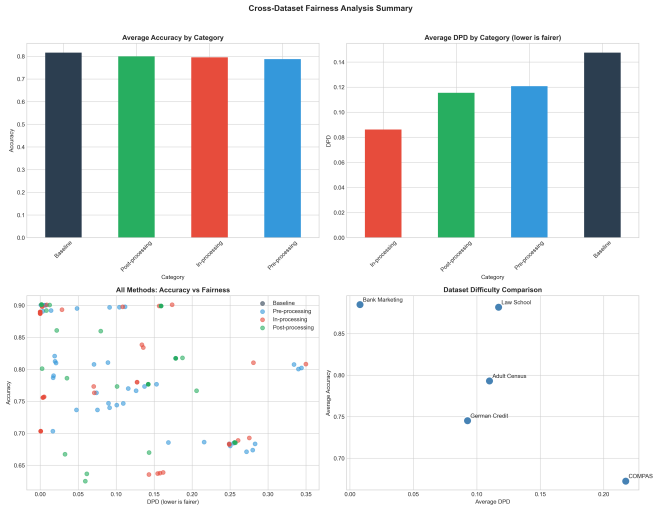

FIG. 4: Cross-dataset summary showing aggregated performance of fairness methods by category.

The summary reveals that in-processing methods exhibit the highest variance in outcomes, while pre-processing methods offer the most predictable, albeit modest, improvements.

F. Information Retention Analysis

Information loss is quantified through the change in mutual information between features and predictions after applying fairness interventions.

Pre-processing methods like DIR with full repair (r=1.0) show 15-25% reduction in feature-label mutual information. In contrast, post-processing methods preserve original model information while only modifying decision boundaries.

This suggests that organizations prioritizing explainability and auditability may prefer post-processing approaches, which preserve the original model's learned representations.

G. Limitations

This study has several limitations:

- **Single sensitive attribute**: This work focuses on binary sensitive attributes; intersectional fairness remains unexplored.
- **Base classifier**: Results may vary with different base models (e.g., decision trees, deep networks).
- **Fairness metrics**: The focus is on group fairness; individual fairness measures may yield different rankings.
- **Temporal aspects**: Fairness dynamics over time are not considered.

H. Ethical Reflections

Several ethical considerations arise from this work:

1. **Metric selection is normative**: The choice between demographic parity and equalized odds reflects underlying values about distributive justice.
2. **Trade-offs affect real people**: A 5% accuracy drop may translate to thousands of incorrect decisions in production systems.
3. **Technical solutions are insufficient**: Fairness interventions address symptoms but not root causes of societal bias in training data.
4. **Transparency matters**: Post-processing methods may be easier to audit and explain to affected individuals.

V. CONCLUSIONS AND FUTURE WORK

A. Key Takeaways

1. **No universal solution**: Method effectiveness is highly dataset and metric dependent.
2. **Adversarial debiasing is robust**: Consistently achieves good trade-offs across datasets.
3. **Reject option is underrated**: Provides excellent fairness with minimal accuracy cost when applicable.
4. **Extreme fairness has costs**: Methods achieving DPD ≈ 0 often do so through degenerate predictions.
5. **Pre-processing is conservative**: Offers reliable but moderate improvements.

B. Practical Recommendations

Based on the findings, the following recommendations are made:

- **Start with adversarial debiasing** for general-purpose fair classification.
- **Use reject option classification** when uncertainty-based decisions are acceptable.
- **Apply reweighing** for quick, low-risk fairness improvements.
- **Avoid sampling strategies** unless specifically addressing class imbalance.
- **Tune regularization carefully**: High λ values can destroy model utility.

C. Future Directions

Several directions merit future investigation:

- **Intersectional fairness**: Extending analysis to multiple overlapping sensitive attributes.
- **Causal approaches**: Incorporating causal inference for more principled fairness interventions.
- **Domain adaptation**: Understanding how fairness transfers across similar datasets.
- **Online learning**: Adapting fairness methods for streaming data contexts.
- **User studies**: Understanding stakeholder preferences for fairness-accuracy trade-offs.

Acknowledgments

This work was conducted as part of the "Inteligência Artificial e Sociedade" course on Responsible AI.

[1] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (fairmlbook.org, 2019).
[2] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems* (2016).
[3] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data* **5**, 153 (2017).
[4] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv:1609.05807* (2016).
[5] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," *Knowledge and Information Systems* **33**, 1 (2012).
[6] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and Removing Disparate Impact," in *Proceedings of KDD* (2015).
[7] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," in *Proceedings of AIES* (2018).
[8] F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," in *Proceedings of ICDM* (2012).
[9] R. K. E. Bellamy *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development* **63**, 4:1 (2019).
[10] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning," in *Proceedings of FAT** (2019).
[11] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks," *ProPublica* (2016).
[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," in *Proceedings of ITCS* (2012).
[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of ICML* (2013).
[14] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," in *Advances in Neural Information Processing Systems* (2017).
[15] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic Decision Making and the Cost of Fairness," in *Proceedings of KDD* (2017).
[16] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Proceedings of AISTATS* (2017).
[17] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized Pre-Processing for Discrimination Prevention," in *Advances in Neural Information Processing Systems* (2017).
[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys* **54**, 1 (2021).
[19] S. Verma and J. Rubin, "Fairness Definitions Explained," in *Proceedings of FairWare* (2018).

Appendix A: Implementation Details

All methods are implemented in Python using the following libraries:

- `scikit-learn` ($\geq$1.0.0): Base classifiers, metrics, and preprocessing utilities
- `numpy` ($\geq$1.21.0) and `pandas` ($\geq$1.3.0): Data manipulation and analysis
- `PyTorch` ($\geq$1.9.0): Neural networks for adversarial debiasing
- `scipy` ($\geq$1.7.0): Optimization routines for constrained methods
- `matplotlib` ($\geq$3.4.0) and `seaborn` ($\geq$0.11.0): Visualization and plotting

Source code is organized as:

- `src/preprocessing.py`: Pre-processing methods
- `src/inprocessing.py`: In-processing methods
- `src/postprocessing.py`: Post-processing methods
- `src/fairness_metrics.py`: Fairness evaluation
- `src/datasets.py`: Dataset loaders
- `src/visualization.py`: Plotting utilities

Experiments are documented in Jupyter notebooks:

- `notebooks/01-05`: Individual dataset analyses
- `notebooks/06`: Cross-dataset comparison

Appendix B: Detailed Results by Dataset

The following tables present complete results for all five datasets.

1. German Credit Dataset

TABLE IV: Results for German Credit dataset.

Category	Method	Accuracy	Precision	Recall	DPD	DPR
Baseline	No Intervention	0.777	0.812	0.886	0.142	0.824
Pre-proc	Reweighing	0.770	0.803	0.890	0.116	0.858
	Feature Masking	0.767	0.802	0.886	0.126	0.845
	DIR (r=1.0)	0.777	0.818	0.876	0.153	0.808
	Label Flip	0.763	0.793	0.895	0.074	0.909
	Sampling	0.737	0.862	0.743	0.075	0.880
In-proc	Fair Reg	0.703	0.704	**0.995**	**0.001**	**0.999**
	Adversarial	0.773	0.809	0.886	0.070	0.911
	Constrained	**0.780**	**0.819**	0.881	0.127	0.840
Post-proc	Equalized Odds	0.637	0.789	0.657	0.061	0.898
	Threshold Opt	0.777	0.812	0.886	0.142	0.824
	Calibrated	0.767	0.789	0.910	0.206	0.764
	Reject Option	0.773	0.809	0.886	0.101	0.873

2. COMPAS Dataset

TABLE V: Results for COMPAS dataset.

Category	Method	Accuracy	Precision	Recall	DPD	DPR
Baseline	No Intervention	0.686	0.710	0.523	0.256	0.442
Pre-proc	Reweighing	0.685	0.715	0.511	0.255	0.433
	DIR (r=0.5)	0.686	0.708	0.529	0.216	0.515
	DIR (r=1.0)	0.686	0.702	0.537	**0.169**	**0.607**
	Label Flip	0.684	0.723	0.495	0.249	0.424
	Sampling	0.684	0.652	0.654	0.283	0.523
In-proc	Fair Reg	0.637	0.777	0.285	0.155	0.360
	Adversarial	**0.693**	0.708	0.554	0.275	0.438
	Constrained	0.684	0.720	0.499	0.249	0.430
Post-proc	Equalized Odds	0.625	0.587	0.599	0.059	0.880
	Threshold Opt	0.686	0.710	0.523	0.256	0.442
	Calibrated	0.685	0.709	0.523	0.257	0.441
	Reject Option	0.667	0.677	0.516	0.033	0.910

3. Bank Marketing Dataset

TABLE VI: Results for Bank Marketing dataset.

Category	Method	Accuracy	Precision	Recall	DPD	DPR
Baseline	No Intervention	0.901	0.710	0.209	0.002	0.952
Pre-proc	Reweighing	0.900	0.689	0.211	0.006	0.842
	DIR (r=0.5)	0.901	0.698	0.206	**0.001**	**0.976**
	DIR (r=1.0)	0.901	0.702	0.205	0.002	0.928
	Label Flip	0.900	0.704	0.198	0.007	0.794
	Sampling	0.821	0.343	0.644	0.019	0.915
In-proc	Fair Reg	0.887	0.0	0.0	0.0	0.0
	Adversarial	**0.901**	0.680	0.235	0.002	0.960
	Constrained	0.898	0.661	0.199	0.003	0.926
Post-proc	Equalized Odds	0.861	0.414	0.560	0.022	0.870
	Threshold Opt	0.901	0.710	0.209	0.002	0.952
	Calibrated	0.901	0.710	0.209	0.002	0.952
	Reject Option	0.901	0.679	0.222	0.012	0.697

4. Law School Dataset

TABLE VII: Results for Law School dataset.

Category	Method	Accuracy	Precision	Recall	DPD
Baseline	No Intervention	0.899	0.911	0.983	0.159
Pre-proc	Reweighing	0.897	0.904	0.989	0.091
	DIR (r=0.5)	0.897	0.909	0.983	0.104
	DIR (r=1.0)	0.898	0.908	0.985	0.112
	Label Flip	0.891	0.891	**1.000**	**0.004**
	Sampling	0.802	0.964	0.808	0.344
In-proc	Fair Reg	0.890	0.890	**1.000**	**0.000**
	Adversarial	0.898	0.912	0.980	0.109
	Constrained	**0.901**	**0.911**	0.985	0.174
Post-proc	Equalized Odds	0.860	0.937	0.903	0.080
	Threshold Opt	0.899	0.911	0.983	0.159
	Calibrated	0.899	0.911	0.983	0.159
	Reject Option	0.892	0.900	0.988	0.008

5. Adult Census Dataset

TABLE VIII: Results for Adult Census dataset.

Category	Method	Accuracy	Precision	Recall	DPD	DPR
Baseline	No Intervention	0.817	0.711	0.448	0.178	0.167
Pre-proc	Reweighing	0.808	0.704	0.393	0.071	0.562
	DIR (r=0.5)	0.817	0.711	0.448	0.178	0.167
	DIR (r=1.0)	0.811	0.713	0.400	0.089	0.470
	Label Flip	0.790	0.755	0.232	0.017	0.787
	Sampling	0.747	0.494	0.690	0.090	0.761
In-proc	Fair Reg	0.756	1.000	0.020	**0.004**	0.380
	Adversarial	**0.838**	0.724	0.567	0.134	0.438
	Constrained	0.810	0.644	0.533	0.281	0.052
Post-proc	Equalized Odds	0.786	0.581	0.505	0.035	0.846
	Threshold Opt	0.817	0.711	0.448	0.178	0.167
	Calibrated	0.818	0.717	0.444	0.187	0.125
	Reject Option	0.801	0.733	0.315	**0.002**	**0.978**