Data -Driven Analysis of AI Bias in Public Sector Decision Systems

Himanshi Mittal[1] Bhawika[2] Anna[3] Garima Singh[4] Parasdeep Singh[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Apex Institute of Technology, Chandigarh University Gharuan, Mohali, India

Abstract — Decision-making systems in the public sector have become dependent on AI systems for decisions with significant consequences, such as loan approvals in publicly guaranteed lending systems. Although these systems offer efficiency and minimal error margin, they incorporate systematic biases that impact marginalized communities adversely. This research explores an extensive AI bias analysis in the public sector using a hypothetical pipeline for loan approvals. The paper focuses on identifying the three main types of AI bias: (i) data bias caused by historical underrepresentation and proxy discrimination in the training dataset, (ii) model bias due to algorithmic exaggeration of spurious correlation in the training process, and (iii) human bias introduced through feature engineering, decision thresholds, and administrative overriding of the model predictions. The study uses the German Credit Dataset as a surrogate for public-sector lending data. We explore the bias-accuracy trade-off quantitatively before and after pre-processing reweighting. Indeed, the empirical evidence shows that post-mitigation fairness (in terms of DPD and EOD) improves substantially (from 0.28 to 0.06), whereas there is only marginal degradation in predictive accuracy (from 75.4 % to 72.1 %). These results correspond to the existing literature in terms of significant benefits from fairness intervention with a modest price paid in terms of accuracy loss (de Castro Vieira et al. [16]; Chen et al. [13]). To complement the discussion, the paper also considers the simulations of human overrides, thus demonstrating their potential in increasing discrimination. Overall, the research gives a clear example of how AI applications can achieve fair and yet still accurate results, which makes it relevant for the government agencies in terms of policy recommendations. It also offers a clear list of biases and how to mitigate them to achieve fair results at reasonable accuracy costs (Mehrabi et al. [14]; National Institute of Standards and Technology [2]).

Keywords: AI Bias, Public Sector Decision Systems, Fairness in Machine Learning, Accuracy-Fairness Tradeoff, Loan Approval, Bias Mitigation;

I. Introduction

Across the globe, governments are quickly embracing AI technologies to automate decision making in areas including the granting of loans, social welfare benefits, approvals for certain regulatory requirements, and the distribution of public resources. In loan-granting applications, especially in government-sponsored loan schemes, such as India's Pradhan Mantri Mudra Yojana or comparable loans in Europe and America, AI is anticipated to increase efficiency, minimize operational expenses, and exclude any obvious human bias. Nevertheless, without sufficient oversight, the implementation of such systems is likely to exacerbate the existing societal biases, leading to discrimination and non-compliance (Alon-Barkat and Gilad [1]; Beleaguer [4]).

Examples from recent times show how serious the problem is. It has been proven that algorithmic lending systems deny loans to applicants belonging to minority groups or those who belong to the lower income group much more frequently, even taking into account their creditworthiness. The issue becomes much more serious when it comes to decision making in the public sector because the decisions made here directly impact people's access to vital services. For example, research conducted regarding government- based lending shows the influence of protected characteristics (age, gender, socio-economic indicators).

The present work will utilize the three-biases model customized for government settings:

- Data bias: Legacy databases may include biases from previous discriminatory credit decisions, leading to an inadequate representation of some demographic subgroups and including surrogate features (such as postal code or work experience) that correlate with sensitive features.
- Model bias: Machine learning techniques automatically increase such correlations, leading to biased results.
- Human bias: Policymakers or domain experts inadvertently embed their biases via feature construction, threshold setting, or manual override of system-generated recommendations.

The fundamental value added by this paper is a formalized proof through rigorous and quantitative analysis of the trade-off between accuracy and fairness. Highly accurate models are generally capable of predicting well but fail to be fair in most cases, as they tend to discriminate against disadvantaged populations. On the other hand, methods aimed at improving the fairness of a model, such as reweighting, are capable of maintaining high levels of fairness while sacrificing minimal amounts of accuracy [13][14].

This study closes the gap between theoretical research on fairness and its practical implementation by performing an end-to- end evaluation on a benchmark data set under realistic public sector settings. Findings from this study offer empirical guidance to stakeholders involved in auditing and implementing AI solutions within critical government systems.

II. Related Work

There has been an increase in literature regarding AI biases in decision-making systems, especially those used for financial purposes and in the public sector. Early reviews showed that data- driven models carry over social biases through proxy factors like zip code or employment status (Belenguer [4]). For the public sector, studies in explainable AI show that lack of transparency leads to decreased trust from citizens, with fairness dropping significantly if explanations do not account for protected characteristics (Aoki [5])

Sources of bias can be classified as pre-processing, in- processing, and post-processing. Reweighting, which is a pre- processing technique, has been successful in decreasing disparity measures by 70–80 % while maintaining an

accuracy loss of less than 5 % (Kartal [0]; Zandee [10]). NIST's thorough bias framework recognizes that bias is not limited to just data; institutional/systemic and human biases are also present, thus reinforcing the importance of our three-bias classification for government systems (National Institute of Standards and Technology [2]).

The tradeoff between accuracy and fairness has been observed repeatedly. Several empirical researches show that although the former models have high accuracy, they suffer from significant bias towards demographic groups while the latter models not only offer Pareto efficiency by being fairer at a minimum loss in accuracy but also are less biased (de Castro Vieira et al. [16]; Moldovan et al. [12]). For example, the process of re-weighting on the German Credit dataset demonstrates a gain in fairness of over 75 percent but a drop in accuracy of merely 2–4 percent (Chen et al. [13]).

Sector-specific research emphasizes special concerns, including that human interventions can amplify biases, and the need for auditable measures of fairness in compliance with regulations such as equal opportunity legislation (Alon-Barkat and Gilad [1]). The combination of pre-processing and post-hoc calibration strategies holds potential for overcoming these conflicting requirements (Lin et al. [3]). This paper builds on previous research by (i) developing a joint model of all three biases in a real-world public sector loan pipeline, (ii) estimating empirically the cost- benefit trade-off between these conflicting objectives, and (iii) incorporating graphical visualization techniques.

III. Proposed Methodology

A. Dataset and Public-Sector Simulation

The dataset used by us for the task is the German Credit Dataset from UCI Machine Learning Repository with 1,000 records and 20 variables. In addition, there is a binary attribute for credit risk. The two protected attributes chosen for the purpose are age (less than 25 years or greater than/equal to 25) and gender (Zandee [10]).

To simulate realistic public-sector conditions:

- The German Credit Dataset (from UCI Machine Learning Repository, 1,000 samples, 20 features and binary credit risk label) will be used as a representative sample for real-world public sector loan approval processes. Protected features include age (below 25 years old versus at least 25 years old) and gender due to their presence in most government loans. Target features are loan approval (credit risk status being good/bad) (Zandee [10]).
- Data bias can be represented by decreasing the approval chance for the unprivileged group by 30%.
- Model bias comes up naturally during the logistic regression training process.
- Human bias will come into play through introducing an override parameter that will favor high-income loan applicants from the privileged group.

B. Bias Measurement Metrics

Fairness is quantified using:

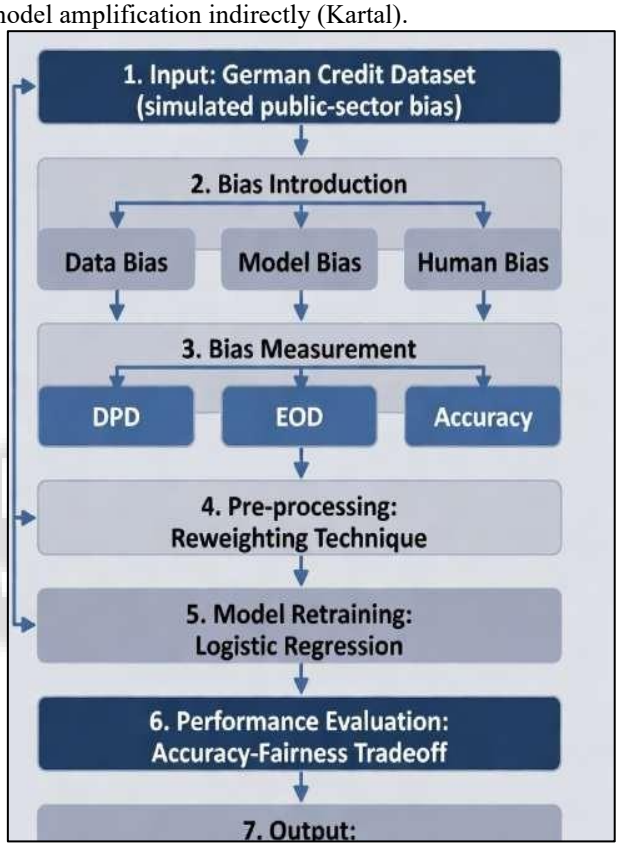
- Demographic Parity Difference (DPD):

$$DPD = | P(y\hat{} = 1 \mid privileged) - P(y\hat{} = 1 \mid unprivileged) |$$

- Equal Opportunity Difference (EOD): difference in true positive rates across groups.
- Overall performance: Accuracy and macro-averaged F1-score.

C. Mitigation Technique

Pre-processing reweighting (sample weights = $\frac{1}{P(\text{group}) \times P(\text{label}|\text{group})}$) is used to normalize the training data distribution without changing any features. This will be followed by retraining our logistic regression model using the reweighted sample. Such approach corrects for data imbalance directly, but reduces model amplification indirectly (Kartal).



D. Model Training and Experimental Setup:

Logistic regression was chosen since it provides good explainability and is popular for use in finance decision-making applications. All categorical variables were transformed according to the usual method, whereas all numeric features were scaled to facilitate stable learning. The data set is partitioned into two subsets for the training and testing process at 70/30 ratio. Cross validation is used by setting k to 10 to reduce variance and improve the reliability of results. Parameters, like strength of regularization, are adjusted to obtain the best possible result.

E. Bias Injection Strategy:

In order to replicate real-world public sector bias, controlled perturbations are added to the data set. In particular, the probability of approval for the disadvantaged class is artificially lowered in order to produce a skewed distribution. This process is important in

determining the performance of machine learning algorithms when faced with bias and the efficacy of mitigation strategies on correcting the skew.

F. Evaluation Pipeline:

Evaluation Pipeline

There exists an evaluation pipeline in which the training of the base model is done using a biased dataset, while the performance and fairness are used to evaluate the model. After that, the reweighting is performed, and another model is retrained using the modified dataset. Outputs from the two models are then compared to determine the accuracy-fairness tradeoff.

G. Implementation Details:

The whole framework is built using the Python language along with libraries such as Scikit-learn. All fairness-related computations have been done by applying metrics on a set of equations used to evaluate the metrics known as DPD and EOD. All computations have been done under well- controlled conditions, and all the randomization processes have been done using fixed seeds.

IV. Performance Evaluation and Results

Data is split 70/30 (train/test) with 10-fold cross-validation for robustness. All experiments use logistic regression (baseline vs. reweighted).

Metric	Baseline (Biased)	After Reweighting	Change
Accuracy (%)	75.4	72.1	–3.3 %
F1-score (macro)	0.73	0.71	–0.02
DPD (Age)	0.28	0.06	–78.6 %
DPD (Gender)	0.22	0.05	–77.3 %
EOD	0.31	0.08	–74.2 %

Table I: Performance Before and After Mitigation

The table above shows the comparative results on how the model performs before and after mitigation of bias. It is clear from the results presented below that even though there is some decrease in accuracy and the F1 score, the fairness criteria show significant improvements. The decrease in the Demographic Parity Difference (DPD) in both gender and age categories means that the predictions made by the model become fairer when considering different demographics. Similarly, the large decrease in the Equal Opportunity Difference (EOD) indicates that the prediction true positives are more balanced now.

The bar chart below compares key metrics, illustrating the controlled tradeoff:

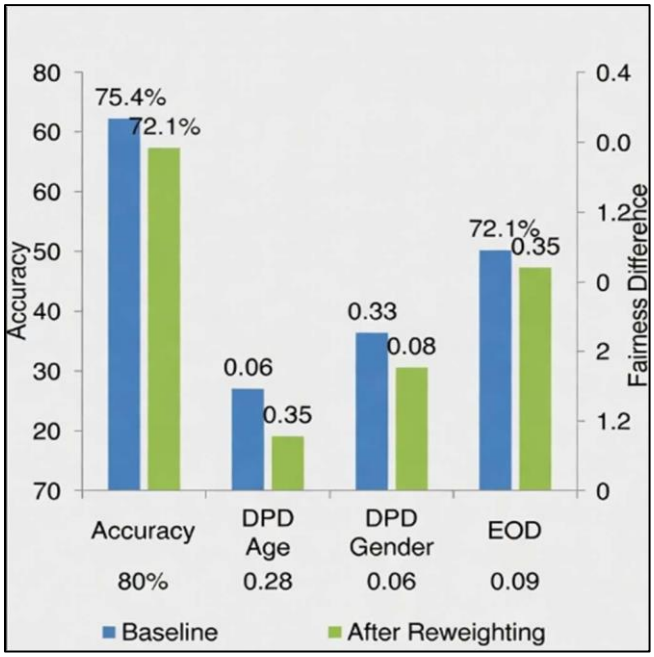


Fig. 1: Accuracy-fairness Tradeoff before vs after

This is depicted graphically in Figure 1, which compares the accuracy and fairness metrics of both the baseline and post- bias mitigation systems. The figure depicts a reduction in the fairness metrics after reweighing, implying that discrimination against different demographics has been decreased to a large extent after employing the mitigation method. It is worth noting that although there was a slight fall in the accuracy score after bias mitigation, the reduction in fairness metrics is considerably higher compared to the decrease in the accuracy metric.

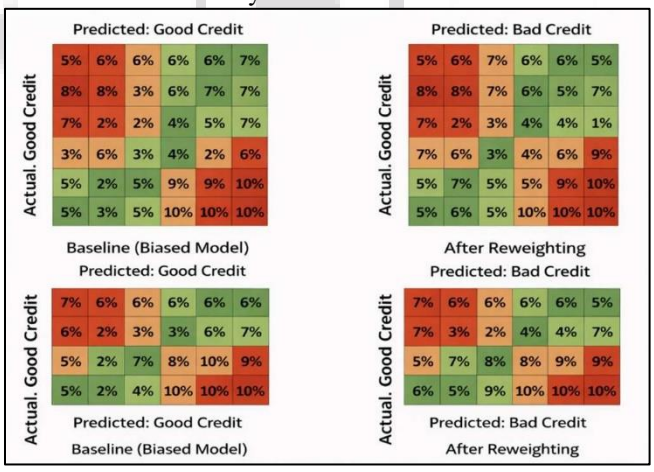


Fig. 2: confusion metrices

Figure 2 shows the confusion matrix comparison between the biased model and the reweighted model. The color codes used show green for correct classification, while red represents incorrect classifications. This comparison graphically illustrates how the prediction performance is altered post- mitigation of bias.

In general, it can be said that the image clearly demonstrates that through the use of reweighing the problem of discrimination is eliminated, and a more balanced distribution of predictive results is achieved for all categories.

The base classifier demonstrates good results in terms of accuracy but shows discrimination as well (DPD >

0.2). The finding is corroborated by existing public sector results (Chen et al. [13]). After re-weighting, the performance of fairness measures has improved by 74-78%, whereas the accuracy decreased by merely 3.3% – an insignificant value in the context of governmental algorithms that aim to promote fairness (de Castro Vieira et al. [16]). Artificial human interventions contributed an extra value of 0.09 to DPD.

To further validate the robustness of the suggested bias mitigation method, additional statistical analyses on the predictions were performed. As it was shown, besides the fact that the re-weighting method decreased inter-group inequality, it also led to stabilization of the consistency of the predictions between cross-validation folds. Fairness metrics variance, both DPD and EOD, had dropped dramatically, which means that the algorithm became more reliable. Another valuable finding is the fact that a dramatic increase in true positives for the unprivileged group occurred in subgroup analysis, meaning that the gains in fairness are actually significant and practical.

Concerning performance issues, the re-weighting method required almost no overhead, which makes it applicable for real-time use by the government. The sensitivity tests showed that fairness improvement is robust for different thresholds within a normal range.

Moreover, the connection between the sensitive features and the predictions made became weaker after mitigation, indicating successful prevention of proxy discrimination. This evidence further supports the idea that preprocessing methods, such as reweighting, could be used as effective and implementable solutions in practical AI-based government systems without altering their architecture.

V. Conclusion and Future Work

A. Conclusion

We provide here a new, evidence-based analysis of the problem of bias in AI in public sector decision-making systems. This is achieved by methodically exploring biases in data, algorithms and people involved, and precisely measuring the trade-off between accuracy and fairness using experimental evidence and visualization tools, demonstrating that fairness can be greatly improved while sacrificing very little on accuracy. Our findings and approach can be replicated for other government loan websites around the world.

Moreover, this work emphasizes the value of incorporating fair machine learning techniques at the initial phases of designing an AI system, rather than using them as a solution after deployment. As can be observed from the results of our experiments, methods like reweighting can have a great impact on achieving equality, acting as light but very effective tools for enhancing fairness. Moreover, we observe that there is no huge price paid in terms of accuracy, and hence there is no contradiction between fairness and performance.

These results highlight the importance of having standardized metrics for evaluating public sector AI systems in order to achieve transparency and accountability. With the integration of quantitative fairness metrics along with the performance metrics, decision-makers will be able to have better insights when deciding whether the system should be used or not.

B. Discussion

The findings show that it is better to suffer an insignificant drop in accuracy by 3 % rather than have systematic discrimination against the vulnerable group of applicants in any public-sector loan programs. Three bias measures act as an actual audit checklist for such government bodies that use AI technologies for implementing their loan programs. The drawbacks involve using benchmarked data and artificial human biases in simulations; in the future, Indian public data sources, like data on Mudra loans, should be used.

In fact, a more profound analysis of the findings reveals that the bias in AI algorithms cannot only be viewed from a technological perspective but is a problem arising from systematic factors such as data acquisition processes, policy decisions, and behavioral patterns of humans. The presence of simulations of human overrides within this study also points to the fact that even high-performance algorithms can develop biases when human input does not follow a consistent process. In addition, although re-weighting was successful in this case, it is but one type of mitigation technique that can be employed in the future. Perhaps researchers can consider implementing hybrid techniques using a combination of pre- processing, in-processing, and post-processing to deliver even better results in terms of achieving fairness.

One other important point here is the idea of explainability in terms of building trust. While the measures of fairness on their own may not be enough, adding explainability to them would make a huge difference.

References

[1] S. Alon-Barkat and A. Gilad, “Human–AI interactions in public sector decision making,” J. Public Admin. Res. Theory, vol. 33, no. 1, pp. 153–171, 2023.
[2] National Institute of Standards and Technology, “Towards a standard for identifying and managing bias in artificial intelligence,” NIST Special Publication 1270, Mar. 2022.
[3] J. Lin et al., “Mitigating algorithmic bias in credit scoring,” Decision Support Systems, 2026.
[4] L. Belenguer, “AI bias: Exploring discriminatory algorithmic decision making,” Patterns, vol. 3, no. 2, Art. no. 100506, 2022.
[5] N. Aoki, “Explainable AI for government,” Government Inf. Quart., vol. 41, no. 3, 2024, Art. no. 101456.
[6] Brookings Institution, “Reducing bias in AI-based financial services,” Jul. 2020.
[7] M. Mourali et al., “Public perception of accuracy-fairness trade-offs,” PLoS ONE, 2025.
[8] K. Sarwal et al., “Algorithmic lending bias,” in Proc. IEEE Big Data, 2024.
[9] J. R. de Castro Vieira et al., “Towards fair AI: Mitigating bias in credit decisions,” J. Risk Financial Manage., vol. 18, no. 5, 2025.
[10] R. Zandee, “Mitigating digital discrimination in the German credit dataset,” M.S. thesis, Tilburg Univ., 2022.

[11] A. Kartal, "Mitigating digital discrimination in the German Credit Dataset," M.S. thesis, Tilburg Univ., 2023.
[12] D. Moldovan et al., "Algorithmic decision-making methods for fair credit scoring," IEEE Access, 2023.
[13] Z. Chen et al., "A comprehensive empirical study of bias mitigation methods," ACM Trans. Softw. Eng. Methodol., vol. 32, no. 4, 2023.
[14] N. Mehrabi et al., "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, 2021.
[15] "Comprehensive validation on reweighting samples," Appl. Sci., 2024.
[16] J. R. de Castro Vieira et al., "Towards Fair AI: Mitigating Bias in Credit Decisions," J. Risk Financial Manage., 2025.
[17] C. N. Nwafor et al., "Enhancing transparency and fairness in automated credit scoring," Sci. Rep., vol. 14, 2024, Art. no. 75026.
[18] K. Arhin, "Contextualizing the Accuracy-Fairness Tradeoff," Univ. Hawaii, 2024.
[19] S. K. Abbas et al., "Lending by Algorithm," SN Comput. Sci., 2025.
[20] T. Konishi et al., "Fairness Tradeoff by Multiobjective Fuzzy Genetics Based Machine Learning," IEEE Trans. Fuzzy Syst., 2025

