Responsible AI Module Assessment Report: An Analysis of the UCI Adult Dataset

Blind Grade Number (BGN): 5352M
Date: December 12, 2025
Word Count: 1,357 words

1 Introduction

As machine learning models are developed by humans, they also carry with them our societal biases and can even perpetuate them further. However, as they are becoming increasingly used in many decision-making scenarios, the impacts of these biases start to add up where certain groups may become more disproportionately impacted than others. Moreover, most machine learning models nowadays usually black boxes which seemingly performs inferences without a clear understanding how they got there. Therefore, it is essential to investigate how to improve the fairness and transparency of these models.

This project explores the use of the UCI Adult dataset (Becker and Kohavi, 1996) to explore the bias a logistic regression model has to the sensitive attribute of `sex` which can lead to unfairness and performing an intervention using the exponentiated gradient algorithm (Agarwal et al., 2018). Moreover, the LIME technique (Ribeiro et al., 2016) is explored for transparency and is used to explain which features influence the prediction outcomes.

2 Dataset and Preprocessing

2.1 Dataset

The UCI Adult dataset (Becker and Kohavi, 1996) is a dataset from the UCI Machine Learning Repository that is famously used as a benchmark to measure fairness in machine learning systems. Extracted from the 1994 US Census Database, the aim of the dataset is to predict whether a person earns over $50,000 a year or not (binary classification). The dataset contains 48,842 data instances with 14 demographic and socioeconomic features (categorical and numerical), including sensitive attributes such as `race` and `sex` (male/female). In this project, fairness evaluation was conducted on the attribute `sex`, investigating whether there is a disparity in predicting income levels by sex.

2.2 Preprocessing Steps

Prior to training the classifier model, the sensitive attribute `sex` was removed from the input features to avoid the model learning explicit bias due to gender. Each categorical feature was encoded into a numerical feature by mapping each unique category to an integer value. Then, all numerical features of the dataset were scaled so that each feature has mean $\mu = 0$ and standard deviation $\sigma = 1$. This step was performed so the classifier could converge faster in training. Lastly, the dataset is split into a train and test split with a ratio of 70/30.

3 Baseline Model

3.1 Logistic Regression Setup

A logistic regression model from `scikit-learn`[1] was chosen as the baseline classifier. Apart from being fast and easy to train, its stable decision boundary and linearity makes it a perfect candidate for transparency and inner-model explanations. Moreover, the model's coefficients allows for easy analysis in interpreting relationships between input features and prediction class outcome.

For training, the model used the `lbfgs` solver with a limit of `100` maximum number of iterations to converge. In actuality, the model converged after `11` iterations.

3.2 Performance Evaluation

Table 1: Baseline classifier performance

Metric	**Value**
Accuracy	0.8249
Precision	0.7215
Recall	0.4367
AUC	0.8496

Table 1 shows the performance metrics of the baseline classifier. The model has a high accuracy of 82.49%, an expected result given the imbalanced nature of the dataset (around 75% are labelled to earn $\leq$\$50,000). A high precision score implies a low false positive rate for the positive outcome (people who earn >\$50,000). A low recall score implies a high false negative rate (people who earn >\$50,000 are often misclassified). Lastly, a high area under the ROC curve (AUC) value indicates the model's well-rounded capability to discern between positive and negative cases.

4 Fairness Analysis

4.1 Fairness Definitions

To asssess the statistical fairness of the baseline model in different gender groups, several notions are discussed.

- **Demographic parity.** Selection rate (the proportion of people that is predicted to belong in the positive class, that is people earning >\$50,000) for both gender should be the same.
- **Equal opportunity.** True positive rate (TPR) for both gender should be the same. In other words, people who has an actual income of >\$50,000 should be correctly predicted at the same rate for both male and female.
- **Equalized odds.** True positive rate (TPR) and false positive rate (FPR) for both gender should be the same. In other words, the tendency for correct prediction of people who has an actual income of >\$50,000 and incorrect prediction of >\$50,000 should be the same for both male and female.

As the goal is to have an equality of outcome for both male and female groups, therefore the focus is on group fairness.

4.2 Group Fairness Results

The `Fairlearn` toolkit (Bird et al., 2020) was used to evaluate whether each notion for group fairness was met for the baseline classifier.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

Table 2: Baseline classifier metrics by Sex

Sex	Accuracy	Selection Rate	TPR	FPR
Female	**0.8915**	0.0560	0.2694	0.0286
Male	0.7919	**0.1889**	**0.4680**	**0.0684**

Table 2 shows the metric results of the baseline classifier by `sex`, as output by Fairlearn's `MetricFrame`[2]. Based on the results, there are gender disparities across all metrics where the selection rate, TPR, and FPR is higher for males compared to females. Because these metrics are different for both male and female, therefore the baseline classifier does not satisfy demographic parity, equal opportunity, nor equalized odds. As a consequence, the model inhibits potential unfairness as it favours predicting males more than females in having an income of >$50,000 (positive class).

5 Transparency Analysis

5.1 LIME

Local Interpretable Model-agnostic Explanations (**LIME**) is an algorithm that aims to explain how a black-box classifier or regressor derives its predictions (Ribeiro et al., 2016). It does so by approximating a linearly local intepretable model. For each instance of the dataset that is being tested, LIME samples near instances and uses the black box model to predict the instances' outputs. Then, a simple linear interpretable model is trained on the near instances weighted by proximity to the original instance. This model becomes the explainer which is locally faithful to the black box model.

5.2 Local Model Explanations

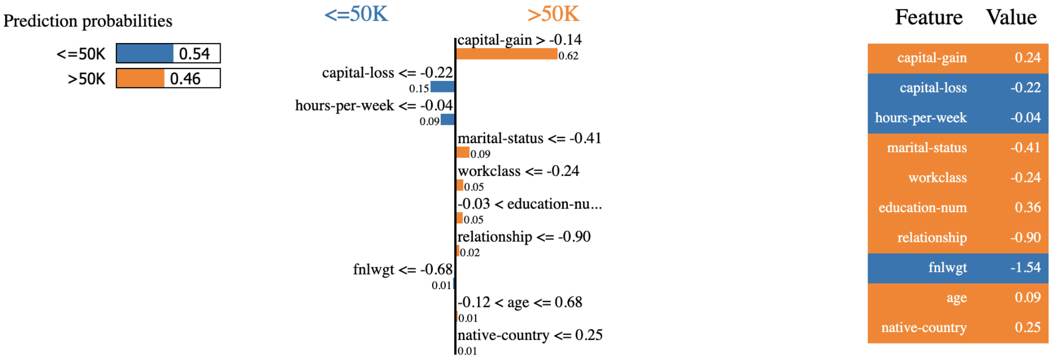

Figure 1: LIME explainer for first instance of test dataset.

Figure 1 shows the results of the LIME explainer for the baseline classifier on explaining the first instance of the first dataset. For this particular instance, features with the largest contributions to predicting income are: `capital-gain` which heavily influences the predicted income of >$50,000, followed by `capital-loss`, `hours-per-week`, `marital-status`, `workclass`, `education-num`, `relationship`, and other small influences.

Analysing the impact each feature has on influencing the outcome prediction can give insights into hidden sources of unfairness that the model may have. For example, a big influence on prediction is the `hours-per-week` feature, which can differ based on caregiving responsibilities. As a result, this feature may serve as a proxy for biases related to gender or other groupings associated with caregiving.

[2] `https://fairlearn.org/main/api_reference/generated/fairlearn.metrics.MetricFrame.html`

6 Fairness Intervention

6.1 Intervention Method

To reduce unfairness in the baseline logistic regression classifier model, a fair in-processing algorithm is introduced, that is the exponentiated gradient (EG) algorithm (Agarwal et al., 2018) constrained to demographic parity. In each iteration, the EG algorithm adjusts the weights of the logistic regression classifier (which have been previously trained) not only to improve accuracy, but also the fairness constraint (in this case, demographic parity).

6.2 Post-Intervention Evaluation

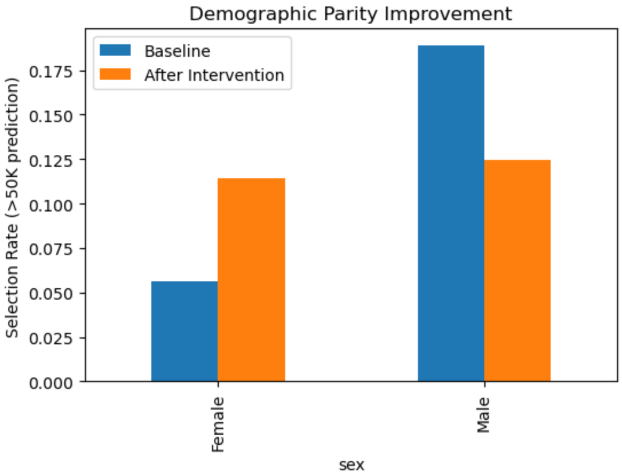

Figure 2: Demographic parity improvement of Sex for predicted income of >$50,000.

Figure 2 shows that after doing the intervention on the baseline logistic regression classifier using the EG algorithm, selection rate of females increased from 0.056 to 0.114 whereas selection rate of males decreased from 0.189 to 0.125. Even though selection rates are not perfectly the same, its gap between males and females have narrowed down significantly, meaning prediction outcome for both gender groups are close to being equal.

7 Conclusion

The baseline logistic regression classifier has shown a bias to the sex attribute when predicting income, resulting in demographic parity, equal opportunity, and equalized odds not being satisfied. However, after conducting an intervention using an exponentiated gradient with a constraint on demographic parity, disparities in selection rates between genders were significantly reduced. Moreover, LIME local explainers provided useful insights as to which features impact prediction outcomes.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

Becker, B. and Kohavi, R. (1996). Adult.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.