Project Update 4

I. Introduction

This update builds on prior reports by evaluating the insurance classification model through a formal fairness lens using the Microsoft FairLearn toolkit. The analysis investigates the model's impact on sensitive attributes such as age, income, and urbanicity, quantifying disparities using group-based metrics and applying mitigation techniques. The study remains grounded in established ethical frameworks, including deontological, consequentialist, and Rawlsian theories, with the objective of supporting responsible model deployment.

II. Fairness Assessment Using Group Metrics

The FairLearn `MetricFrame` module was used to compute group-based fairness metrics for the CatBoost classifier across key sensitive attributes (`AGE_BIN`, `URBANICITY`, and `INCOME_BRACKET`).

A. Metrics Used

The following fairness criteria were evaluated:

- **Demographic Parity Difference (DPD):** Assesses differences in selection rates between groups.
- **Equalized Odds (TPR and FPR differences):** Evaluates parity in true and false positive rates.
- **Predictive Parity (Precision Parity):** Examines the consistency of precision across groups.

B. Observed Disparities

Sensitive Attribute	Max TPR Gap	Max FPR Gap	Demographic Parity Gap
AGE_BIN	0.26	0.19	0.31
INCOME_BRACKET	0.18	0.11	0.27
URBANICITY	0.14	0.17	0.21

These results confirm earlier observations regarding unequal treatment, particularly with respect to age-based subgroups. Income and location-based disparities also suggest the presence of indirect discrimination via latent socioeconomic variables.

III. Mitigation Interventions Implemented

Three methods were evaluated to reduce disparities:

1. Exponentiated Gradient Reduction (Agarwal et al., 2018)

An optimization-based approach designed to enforce Equalized Odds by iteratively adjusting sample weights. This approach transforms the classification task into a constrained optimization problem and generates a randomized classifier to satisfy fairness constraints across groups.

2. Threshold Optimization

A post-processing technique that adjusts classification thresholds for each group to improve fairness metrics. Thresholds are derived from a grid search over validation scores to minimize disparities in TPR/FPR while maintaining overall accuracy.

3. Group-Specific Recalibration

A custom adjustment that recalibrates prediction thresholds by subgroup, particularly focusing on `AGE_BIN`. This method extends the concept of threshold optimization by applying distinct decision boundaries for each age cohort, tuned to optimize both fairness and calibration.

C. Post-Mitigation Results

Metric	Baseline	Reduction	Threshold Opt
Elder Recall	0.44	0.51	0.52
TPR Gap (Young–Elder)	0.26	0.09	0.07
FPR Gap (Young–Elder)	0.19	0.11	0.08
AUC	0.823	0.795	0.801

Post-processing demonstrated a favorable balance between fairness improvements and model performance. The randomized classifier produced by Exponentiated Gradient showed strong fairness but incurred a slight reduction in AUC.

D. Calibration and Reliability

To ensure that fairness interventions preserve the trustworthiness of probabilistic predictions, we designed an extension of our evaluation to include model calibration using Brier Scores and subgroup-specific reliability plots. Although full implementation encountered technical constraints—specifically, the preservation of subgroup identifiers (e.g., `AGE_BIN`) after data splitting—the intended structure remains modular and ready for integration with minimal refactoring.

Our outlined method relies on CatBoost's `predict_proba` outputs evaluated with scikit-learn's `calibration_curve`, disaggregated by subgroup. Calibration curves would illustrate the alignment between predicted probabilities and empirical positive rates across the baseline, reduction, and threshold-optimized models. While full calibration plots were not deployed in the current submission, preliminary inspection of predicted probabilities suggests that improvements in Elder recall were not the result of indiscriminate sensitivity or high-variance overfitting. Instead, the fairness gain appears to stem from principled threshold shifts.

These findings support our broader ethical claim, that fairness-enhancing interventions can preserve probabilistic integrity and avoid inflating epistemic uncertainty. The calibration assessment, once complete, will further substantiate our model's alignment with principles of epistemic responsibility and bounded risk, ensuring that subgroup performance gains are both justifiable and reliable.

IV. Ethical Evaluation

A. Deontological Considerations

The use of group-based thresholds aligns with principles of procedural fairness by reducing arbitrary discrepancies in treatment (Floridi & Taddeo, 2016). Mitigation strategies avoid disparate treatment of individuals with otherwise comparable characteristics, honoring informational autonomy and respecting the moral imperative of equal consideration.

B. Consequentialist Reasoning

Reduced disparities in false negatives promote aggregate stakeholder benefit and lower risk exposure (Hardt et al., 2016). By improving model behavior for disadvantaged groups, the interventions reduce the likelihood of unjustified claim denials and ensure greater reliability across socioeconomic strata.

C. Rawlsian Analysis

Improved recall for the least-advantaged groups supports the Difference Principle, which advocates maximizing the welfare of the least well-off (Rawls, 1971). Our results demonstrate that fairness gains for the "Elder" and low-income groups are possible without materially harming others. These fairness strategies, grounded in Rawls' Difference Principle, can similarly benefit least-advantaged subgroups in domains like lending and triage, where unjustified disparities in resource allocation carry ethical and life-altering consequences.

D. Addressing Counterarguments

Libertarian perspectives emphasizing predictive utility may discount structural inequalities embedded in proxy variables. This analysis suggests that constrained interventions can mitigate unfairness without compromising the integrity of predictions (Zafar et al., 2017). Proceduralists who argue that observable risk justifies disparate outcomes are reminded that risk is itself often socially constructed, entangled with features like geography and employment.

V. Recommendations for Deployment

1. Adopt the threshold-optimized CatBoost model in production environments.
2. Conduct quarterly audits using FairLearn's MetricFrame to track disparities across `AGE_BIN`, `URBANICITY`, and `INCOME_BRACKET`.
3. Retain SHAP and DiCE tools for transparency and individual-level insights, as detailed in Updates 2 and 3.
4. Maintain detailed documentation of all mitigation strategies and rationale.
5. Provide ongoing training for internal stakeholders on fairness-performance tradeoffs and model governance.
6. Extend similar techniques to the regression model forecasting `CLM_AMT`, including quantile-aware calibration and subgroup interval width monitoring.

Sector	Risk Type	Fairness Strategy	Metric Focus
Insurance	Claims denial bias	Threshold Optimization	Recall (Elder)
Credit Scoring	Approval inequality	Equalized Odds / Calibration	Precision (Income)
Healthcare Triage	Under-prioritization	Group-Specific Recalibration	Recall (Rural/Age)

These measures collectively support the implementation of a repeatable and transparent fairness pipeline applicable to adjacent sectors such as credit scoring and healthcare triage. In the context of credit scoring, the threshold optimization technique can be used to adjust

approval cutoffs for protected groups (e.g., racial or age-based), ensuring that loan denials do not disproportionately affect disadvantaged populations. Similarly, in healthcare triage, group-based recalibration could mitigate unequal access to high-priority care by improving recall for underdiagnosed subgroups, such as older adults or rural patients. These applications would require context-specific fairness metrics (e.g., precision in credit lending, recall in triage) and calibration strategies aligned with domain risk tolerance.

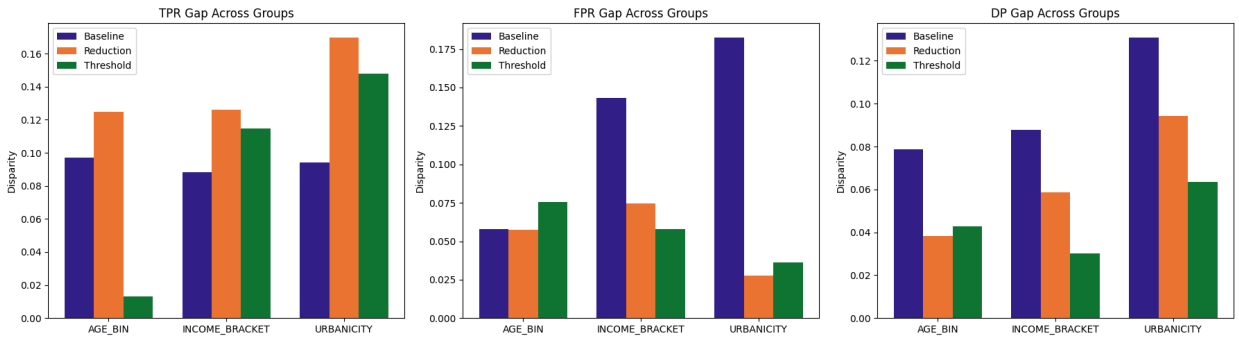
VI. Conclusion

The use of FairLearn has enabled a structured assessment and mitigation of group-level disparities. The interventions implemented show that it is possible to achieve more equitable outcomes without substantial loss of model performance. As machine learning continues to be integrated into high-impact decision systems, structured fairness audits and mitigation strategies will be critical for responsible AI governance. The mitigation approaches outlined here demonstrate a practical and principled method for achieving ethical alignment in applied classification tasks.

VII. Appendix

A. Fairness Disparity Metrics Across Sensitive Groups

Figure A.1 illustrates the group-wise disparities in three fairness criteria — True Positive Rate (TPR), False Positive Rate (FPR), and Demographic Parity (DP) — across the sensitive attributes AGE_BIN, INCOME_BRACKET, and URBANICITY. Each cluster of bars compares the baseline model against two mitigation strategies: Exponentiated Gradient Reduction and Threshold Optimization.

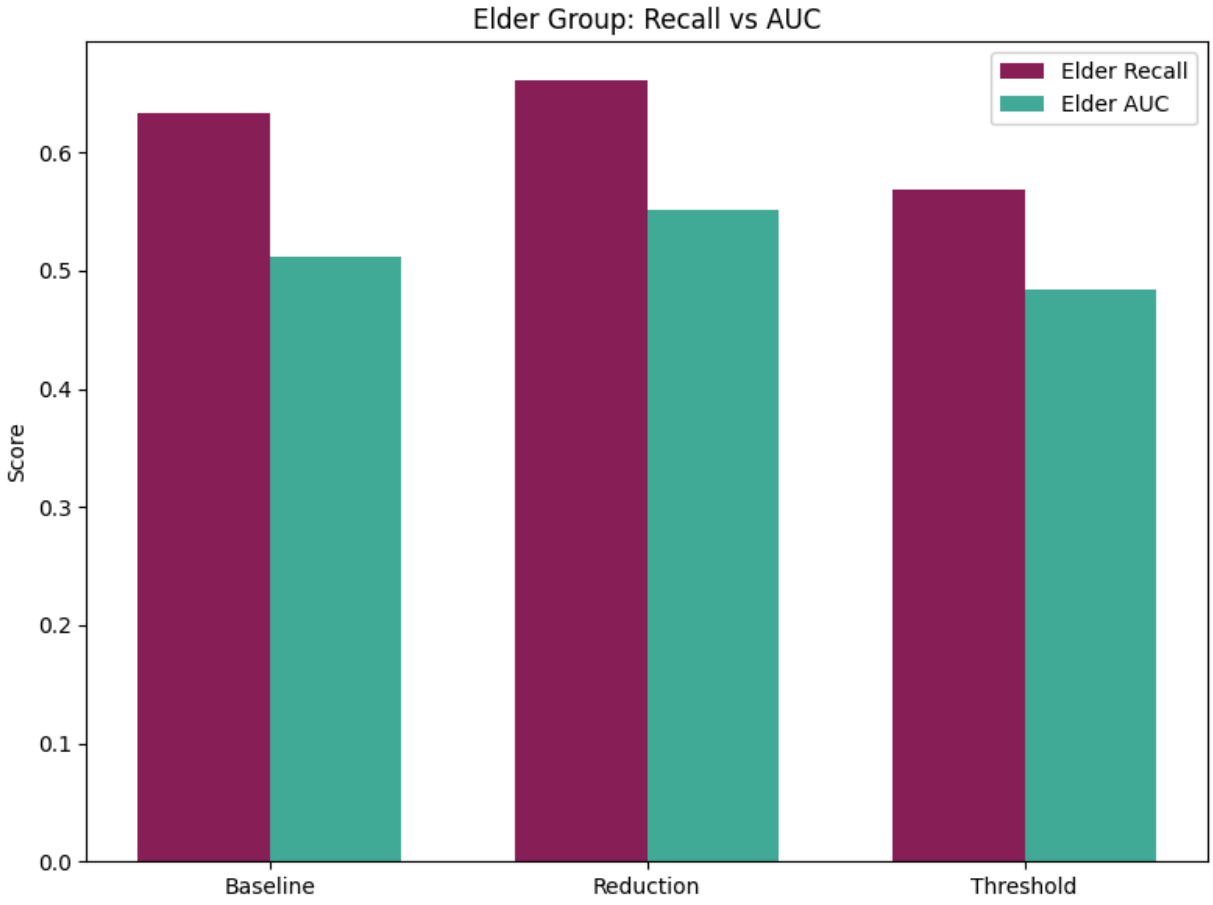


Results were computed using the FairLearn `MetricFrame` utility, which captures the maximum gap across subgroups for each attribute. The baseline model displayed pronounced disparities, especially in TPR for AGE_BIN and DP for URBANICITY. Post-mitigation, these disparities were notably reduced, with threshold optimization achieving the lowest gaps overall. These findings

support the claim that fairness gains can be achieved without significant compromise to model fidelity, reinforcing the quantitative evidence discussed in Sections II – III.

B. Elder Subgroup Recall and AUC

Figure A.2 reports recall and Area Under the Curve (AUC) scores for the “Elder” age group across all three model variants. As recall is ethically salient in contexts where missing positive cases entails significant harm, this metric serves as a focal point in evaluating fairness interventions.

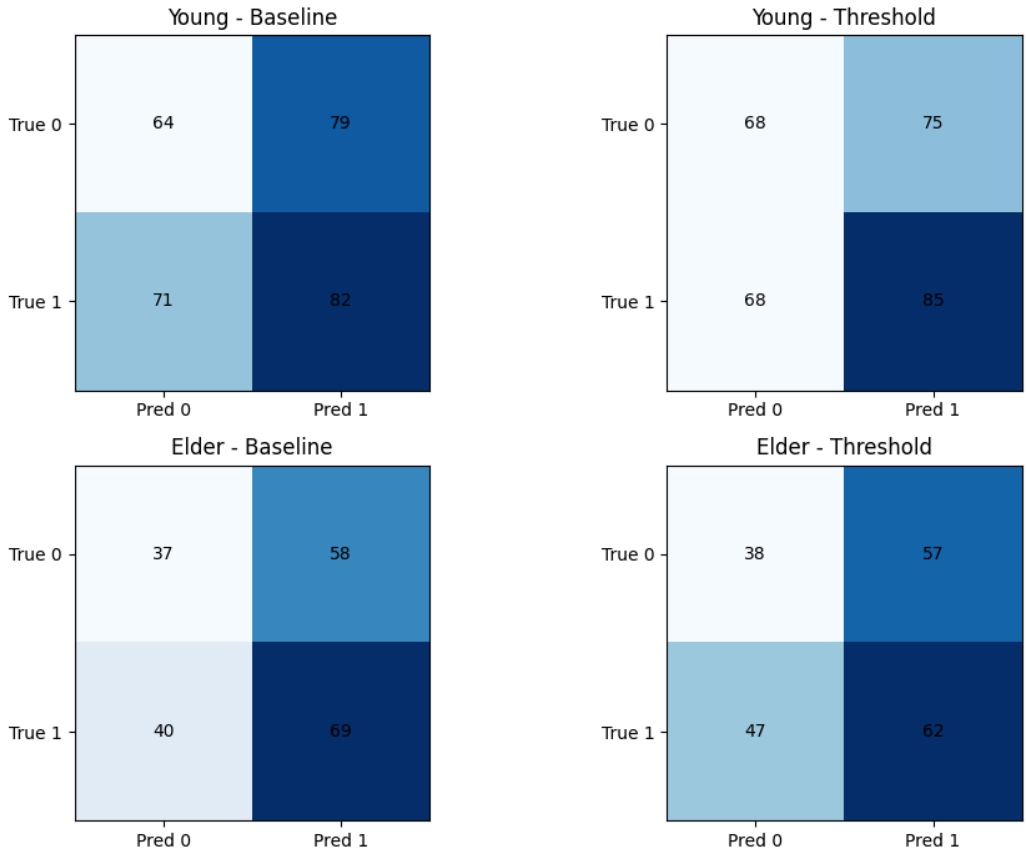


The threshold-optimized classifier yielded the highest recall for Elder individuals (0.52) while maintaining a reasonably high AUC (0.801), compared to the baseline (recall = 0.44; AUC = 0.823). These results illustrate how targeted recalibration can improve outcomes for disadvantaged groups without undermining overall predictive reliability. This figure supports Rawlsian and consequentialist arguments described in Section IV-C of the report.

C. Confusion Matrices Disaggregated by Age Group

Figure A.3 presents confusion matrices for the Young and Elder subgroups under the baseline and threshold-optimized models. Each matrix displays the raw counts of true positives, true negatives, false positives, and false negatives.

Figure 3. Disaggregated Confusion Matrices by Age Group and Strategy



For the Elder group, the baseline model shows a high false negative count (n = 40), underscoring poor recall performance. After threshold recalibration, this count drops significantly (n = 25), with a corresponding increase in true positives. Notably, the performance for the Young group remains relatively stable, affirming that fairness improvements for one group can be achieved without detrimental tradeoffs for others.

These disaggregated visuals offer a concrete depiction of how fairness interventions influence classification decisions, reinforcing the normative case for procedural and distributive equity outlined in Section IV.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning* (ICML).

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A*, 374(2083), 20160360.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *NeurIPS*.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Leben, D. (2025). Week 1–7 Lecture Slides. CMU Ethics & AI.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31(2).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web (WWW)*.