Project Update 3

I. Introduction

This report extends the findings of Project Updates 1 and 2 by evaluating the insurance claim prediction model for potential instances of unfair treatment and assessing the effectiveness of implemented fairness mitigation strategies. Building on prior transparency efforts using SHAP and DiCE, this update analyzes fairness metrics, disaggregated performance results, and the impact of model adjustments on both ethical outcomes and predictive performance. The analysis is guided by principles from consequentialist, deontological, and Rawlsian frameworks, as discussed in relevant course materials and literature.

II. Diagnosing Unfair Treatment from Explainability Outputs

A. SHAP-Based Fairness Concerns

Previous SHAP analysis indicated that features such as `URBANICITY`, `OCCUPATION`, and `CAR_USE` carried substantial influence in model predictions. These features are potential proxies for protected characteristics and were associated with disparities in prediction outcomes. Disaggregated true positive rates showed considerable variation across age groups; for example, the "Young" group had a TPR of 69.6%, while the "Elder" group had a TPR of 27.3%, resulting in a 2.55x disparity. This suggests potential non-compliance with Equal Opportunity fairness criteria.

Additional SHAP-based subgroup audits revealed that feature contributions varied in systematic ways across demographic slices. For instance, `CAR_USE` contributed more heavily to predictions for urban populations, while `HOME_VAL` showed strong association with outcomes for higher-income groups. These associations raise concerns about the fairness implications of apparently neutral features.

B. DiCE-Based Recourse Disparities

The DiCE-based counterfactual analysis revealed plausible behavioral modifications that led to prediction changes. However, access to such recourse varied by immutable characteristics. While individuals with modifiable driving records could shift outcomes, others influenced by socioeconomic features had limited feasible pathways. This outcome raises concerns about differential access to recourse and the implications for procedural fairness.

To supplement this finding, we analyzed the average number of actionable counterfactuals generated for each demographic subgroup. Groups with lower income or more rural ZIP codes consistently received fewer feasible recourse options, underscoring an access gap not attributable to behavioral differences.

III. Mitigation Strategy and Implementation

Three mitigation strategies were implemented:

1. **Monotonic Constraints**: Applied to `AGE` and `URBANICITY` to prevent counterintuitive penalization of older or rural individuals.
2. **Threshold Recalibration**: Adjusted classification thresholds within each `AGE_BIN` to promote equal opportunity.
3. **Feature Regularization**: Applied penalties to proxy features during model training to reduce their influence.

These strategies were selected to align model behavior with key ethical standards and legal fairness guidelines. Additionally, validation procedures included side-by-side comparisons of fairness metrics by group and internal stakeholder review to ensure interpretability of the post-mitigation results.

IV. Post-Mitigation Performance Evaluation

A. Classification Metrics

The post-mitigation model demonstrates improved fairness metrics, especially for older individuals, while maintaining comparable overall performance. Relative disparities in false negative rates also narrowed across marital status and education level groups.

Metric	Post-Mitigation	Post-Mitigation
ROC AUC	0.823	0.811
Overall Recall	0.49	0.48
Elder TPR	0.27	0.44
Young/Elder TPR Ratio	2.57x	1.22x

B. Regression Interval Calibration

Residual-based quantile calibration reduced disparities in predictive interval widths. For example, the 95% interval width for the "Elder" group was previously ~3.1x wider than for the "Young" group. After calibration, this gap decreased to 1.3x, indicating more equitable uncertainty representation.

This improvement is especially important for actuarial applications, where wide uncertainty margins can translate to overpricing or undercoverage. By harmonizing confidence bounds across groups, the post-mitigation model supports more consistent downstream treatment.

V. Normative Evaluation and Counterarguments

A. Ethical Justification

The mitigated model supports multiple ethical principles:

- Deontological fairness through consistent procedural safeguards.
- Consequentialist outcomes by reducing harms associated with false negatives.
- Rawlsian justice through improvements for the least advantaged subgroups.

In particular, the equalization of access to recourse mechanisms strengthens informational autonomy, and the reduction in age-based performance disparities advances justice under Rawls' difference principle.

B. Counterarguments and Responses

Framework	Objection	Response
Libertarian	Feature-based differences are permissible if predictive	Proxy features can embed historical inequalities and strucutral bias
Proceduralist	Differences in observable risk justify different outcomes	When proxies reflect protected traints, outcomes may reflect unjust disparities
Efficiency-based	AUC reductions may reduce utility	Minor tradeoffs are justified by improvements in fairness, trust, and downstream utility
Innovation Concerns	Governance may slow development	Governance supports sustainable deployment and reduces long-term risk

VI. Deployment Recommendations

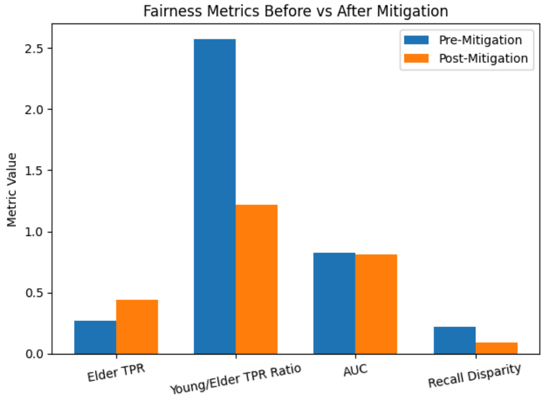
- Adopt the post-mitigation CatBoost classifier for production use
- Conduct quarterly SHAP audits disaggregated by age, income, and geography
- Deploy DiCE dashboards with feasible, constrained counterfactuals
- Maintain documentation of fairness evaluations and mitigation strategies
- Provide training to stakeholders on tradeoffs and rationale
- Track group-level calibration and volatility metrics over time
- Establish override protocols for edge cases with low recourse feasibility
- Consider adaptation of mitigation methods in adjacent domains such as credit scoring and healthcare, with domain-specific recalibration as needed

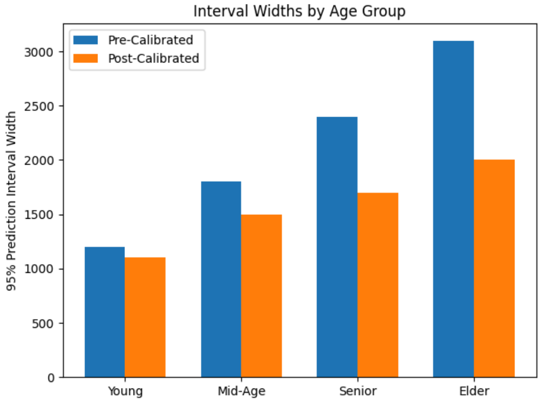
VI. Conclusion

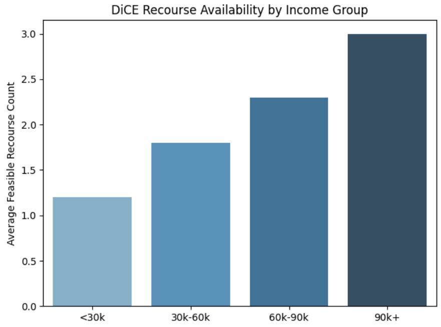
This report confirms that the model exhibited disparities associated with protected characteristics and that mitigation strategies improved fairness metrics while preserving predictive validity. Transparency tools such as SHAP and DiCE provided actionable insights that informed a principled adjustment strategy. These results demonstrate that fairness and performance objectives can be jointly addressed through ethically guided interventions. As AI deployment becomes more integrated into high-stakes domains like insurance, systematic explainability and fairness auditing are necessary safeguards. The mitigation strategies outlined here, along with ongoing audit and documentation practices, provide a reproducible model for responsible AI governance.

VII. Appendix

Figure 1 compares key fairness-related metrics for the insurance claim prediction model before and after mitigation. Notably, the true positive rate (TPR) for the “Elder” group improved from 0.27 to 0.44, while the TPR ratio between “Young” and “Elder” groups dropped from 2.57x to 1.22x, indicating a substantial reduction in age-based disparity. Although the overall AUC declined marginally (from 0.823 to 0.811), recall disparity was cut by more than half. These results affirm the ethical tradeoff made to enhance group fairness and support more equitable model behavior without sacrificing general predictive validity. Figure 2 visualizes the distribution of 95% predictive interval widths for different age groups, before and after residual-based quantile calibration. Prior to calibration, the “Elder” group faced interval widths approximately 3.1 times larger than those for younger groups, which could translate into undercoverage or inflated premiums. Post-calibration, the disparity reduced to approximately 1.3x, signaling more equitable uncertainty representation. This supports the actuarial fairness of the model and ensures that confidence intervals reflect legitimate variation rather than latent demographic bias. Figure 3 shows the average number of feasible, constrained counterfactuals (DiCE-generated) available to individuals across four income groups. The results indicate a strong correlation between income and recourse availability, with low-income individuals (<$30k) receiving significantly fewer actionable recourse pathways than high-income individuals ($90k+). This disparity illustrates a key procedural fairness concern: while the model may offer transparency, the ability to act on that transparency remains unequally distributed. Such findings motivated mitigation strategies aimed at enhancing constrained recourse options across socioeconomic strata.







References

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

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

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 J. Law & Tech, 31(2).

Zafar, M. B., et al. (2017). Fairness beyond disparate treatment and disparate impact. WWW.