Project Update 2

I. Introduction

This report serves as a continuation of Project Update 1, expanding the evaluation of our deployed insurance classification model to include explainability. While our prior report emphasized fairness, informed consent, and governance, this update investigates the transparency and interpretability of our model through two explainable AI (xAI) techniques. Specifically, we apply a global feature importance method (SHAP) and a counterfactual reasoning method (DiCE). These methods are assessed both for technical reliability and for alignment with recognized ethical principles, with attention given to stakeholder concerns and normative standards.

II. Global Feature Importance with SHAP

SHAP (SHapley Additive exPlanations), derived from cooperative game theory, assigns feature contributions to individual predictions by computing marginal contributions. We utilized CatBoost's native `get_feature_importance(..., type='ShapValues')` implementation to leverage parallel computation and enhance performance. A Pool object containing the categorical feature index was used to optimize SHAP efficiency.

Top-ranked features included `URBANICITY`, `CAR_USE`, `HOME_VAL`, `OCCUPATION`, and `TIF`, with `INCOME` appearing lower than initially expected. These findings raise immediate fairness concerns: `URBANICITY` and `OCCUPATION` may act as latent proxies for race, wealth, or structural disadvantage. Summary plots (Figure 1) and individual SHAP force plots (Figure 2) revealed strong model sensitivity to location, employment-related features, and mobility patterns.

From a deontological perspective, transparency supports the principle of informational autonomy (Floridi & Taddeo, 2016). Consequentialist frameworks suggest increased explainability contributes to stakeholder trust and may reduce downstream harms. Rawlsian theory supports disaggregated SHAP analyses to ensure the model benefits the least-advantaged groups. Improvements could be made through conducting regular (e.g., quarterly) SHAP audits with subgroup disaggregation. Additionally, consideration should be given to the use of monotonic constraints and regularization to reduce proxy risk.

III. Counterfactual Explanations with DiCE

Using Microsoft's DiCE library, we generated counterfactuals for individual-level predictions, particularly for cases near the model's decision threshold. To ensure compatibility and interpretability, we deployed a scikit-learn RandomForestClassifier and constrained counterfactual generation to a limited set of actionable behavioral features. The original query instance had `MVR_PTS = 5` and received a prediction of `0` (no claim). Three counterfactuals were generated (Figure 3), all of which increased `MVR_PTS` to 9, 10, or 12 and flipped the model prediction to `1`. This result is technically correct and ethically favorable: it avoids suggesting implausible changes to immutable or semi-immutable characteristics (e.g., income, marital status, ZIP code). Recourse remained within the behavioral and regulatory domain of driving risk. Deontological concerns include fairness and respect for persons, especially when recourse is economically inaccessible. Consequentialist evaluation points to the limited practical value of counterfactuals if they suggest unrealistic or legally infeasible changes. Rights-based concerns arise when suggested changes are closely tied to protected attributes. Our constrained approach significantly improves the plausibility, accessibility, and ethical standing of the recourse generated.

IV. Policy Recommendations and Deployment Standards

Based on the empirical findings and normative analysis, the following policy updates are proposed.

Quarterly SHAP Audits	Generate and review global and subgroup SHAP outputs
DiCE-Based Recourse Reports	Deploy individualized explanation dashboards filtered by feasibility
Feature Governance Review	Maintain oversight of proxy risks via statistical and ethical evaluations
Ethics Review Committee	Empower internal review boards to evaluate fairness and explainability standards prior to deployment
Informed Consent Reform	Enhance model transparency within user-facing consent protocols

V. Response to Counterarguments

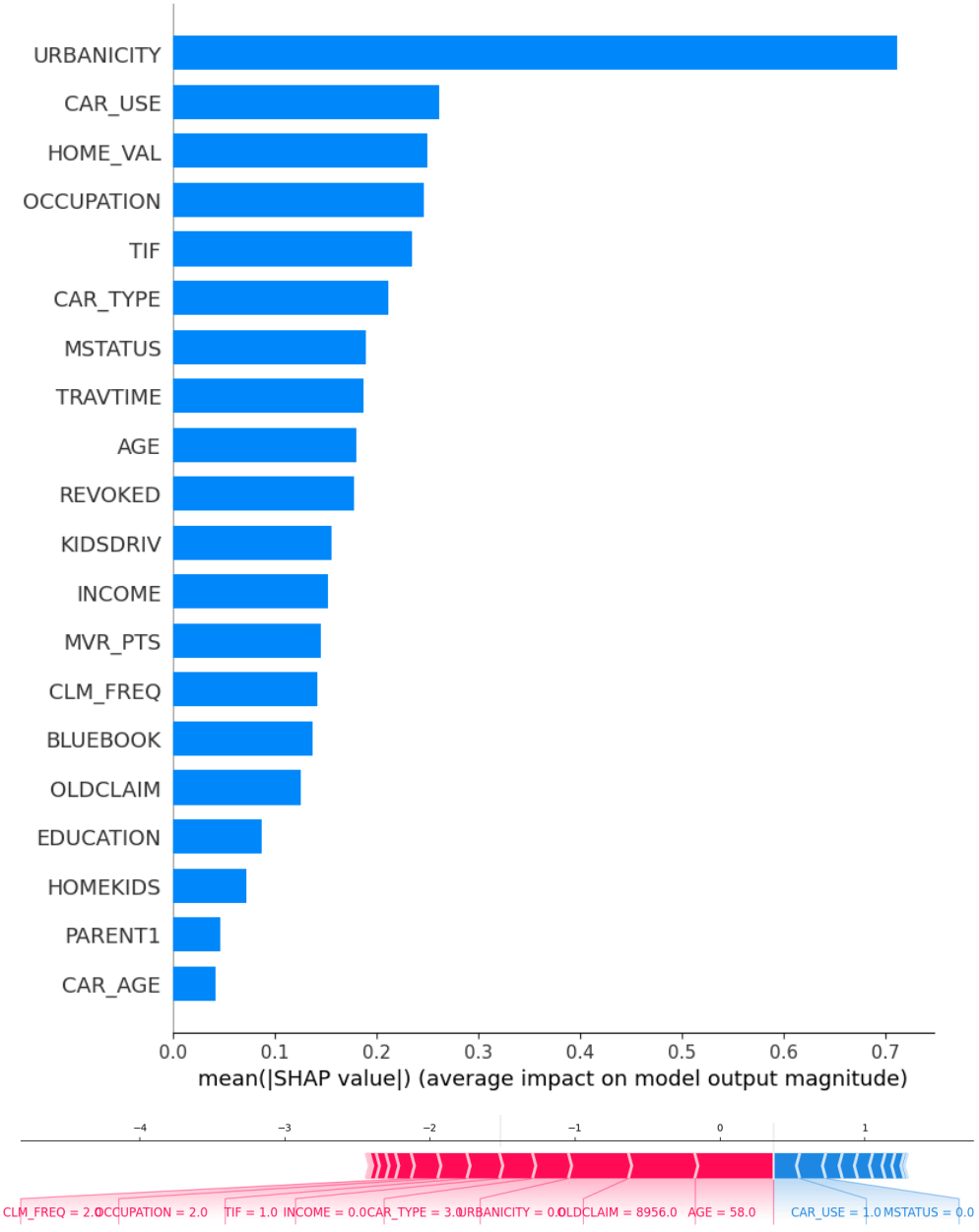
In response to critiques rooted in libertarian, proceduralist, or market-based perspectives, arguments based on opt-out consent or efficiency overlook power asymmetries and the disproportionate impact of proxy features. Procedural fairness frameworks may accept feature weights that perpetuate group-based disadvantages. Market rationality alone is insufficient justification when model outcomes diverge from ethical and legal fairness standards. Thus, these perspectives, while relevant in specific regulatory discussions, do not provide a sufficient ethical foundation for our deployment objectives.

VI. Conclusion

Explainability is an essential component of trustworthy AI. The combined use of SHAP and DiCE has provided transparency into model behavior and recourse options. While these tools enhance interpretability, they also reveal risks that require ongoing ethical and technical oversight. We propose actionable deployment standards aimed at aligning our model with established fairness and transparency principles. Future work will continue to refine these tools and integrate stakeholder feedback.

VII. Appendix

Explainability is an essential component of trustworthy AI. The combined use of SHAP and DiCE has provided transparency into model behavior and recourse options. While these tools enhance interpretability, they also reveal risks that require ongoing ethical and technical oversight. We propose actionable deployment standards aimed at aligning our model with established fairness and transparency principles. Future work will continue to refine these tools and integrate stakeholder feedback.



References

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness and Machine Learning. fairmlbook.org.

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

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

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online Manipulation: Hidden Influences in a Digital World. Georgetown Law Technology Review, 4(1), 1–45.