Report Update 2B

Introduction

This report uses XAI techniques (DiCE, SHAP, LIME) to evaluate the fairness of machine learning models for predicting insurance claims. While accuracy is important, we aim to avoid bias, protect customer trust, and minimize unfair outcomes. Our goal is to improve the models to ensure fair treatment for all customers.

Fairness Evaluation and Methodology

We understand and define unfairness as having fairness features that might create an adverse impact. For example, gender might unfairly treat one gender worse than the other, likely giving them higher premiums. Our initial concern is that "neutral" features, like URBANICITY and occupation, might be a proxy to discover protected attributes (socioeconomic background) that risk perpetuating societal biases. We use XAI, SHAP, and Lime to understand feature importance and how they contribute to driving predictions. Additionally, we use DiCE to examine how changes in inputs affect outcomes. Doing so reveals potential unfairness for individuals near decision thresholds. The combined approach of XAI methods enables a robust assessment that aims to develop fair and ethical models.

Fairness Analysis based on XAI Results

To ensure ethical alignment, we proactively excluded GENDER, MSTATUS, PARENT1, and EDUCATION as they reflected sensitive personal characteristics and bias. However, our fairness evaluation shows that even features that seem neutral—such as URBANICITY and OCCUPATION—can contribute to unfair treatment when used in predictive models.

- Analysis of URBANICITY: URBANICITY, especially the “Highly Rural/Rural” category, strongly influenced CLAIM_FLAG predictions. DiCE showed that small location changes could flip outcomes, even with other factors unchanged. Since location often reflects socioeconomic or racial factors beyond individual control, its use risks unfairly penalizing certain groups.

- Analysis of OCCUPATION: In the XGBoost model predicting CLM_AMT, SHAP and LIME showed that OCCUPATION, especially "Professional", had a strong impact. While it may reflect lifestyle, it also correlates with income and access. DiCE revealed that changing only occupation could shift predictions, raising fairness concerns around economic discrimination.

- Analysis of OTHER INFLUENTIAL FEATURES: SHAP and LIME showed that INCOME, HOME_VAL, AGE, and KIDSDRIV also influenced predictions. INCOME and HOME_VAL may unfairly impact lower-income individuals. AGE, though legally used, risks age-based bias if not tied to driving experience. KIDSDRIV and HOMEKIDS, while not directly unfair, could indirectly relate to protected characteristics.

Discussion and Ethical Synthesis

The XAI analysis reveals significant fairness risks in our models due to reliance on proxy variables like INCOME, AGE, and HOME_VAL. Although predictive, these variables risk reinforcing economic and age-based biases, raising ethical concerns about fairness and discrimination.

From a **deontological perspective**, relying on these non-driving proxies violates the ethical duty to treat individuals respectfully and fairly, focusing solely on relevant factors like driving behavior. Kantian ethics particularly emphasizes the moral importance of intent and the

imperative to avoid discrimination based on unrelated personal attributes, highlighting that ethical actions are those driven by fairness and respect for individual autonomy, emphasizing the need for algorithmic transparency.

From a **utilitarian perspective**, consequentialists focus on maximizing immediate predictive accuracy, while rule utilitarians emphasize fairness rules that ensure long-term societal trust and reduce discrimination risks.

Prioritizing fairness may reduce predictive accuracy, especially by removing variables like URBANICITY and OCCUPATION that pose discrimination risks. If occupation remains, clear guidelines must prevent economic discrimination. Thus, fairness and transparency ethically justify accuracy trade-offs, promoting trust and societal acceptance.

Standard and Recommendations

After evaluating our models with SHAP, LIME, and DiCE, we've decided to exclude both **URBANICITY** and **OCCUPATION** from any production model. While these features technically improve accuracy, they do so by reinforcing unfair advantages. Just because someone isn't a "professional" or doesn't live in a certain neighborhood can statistically mean they're more likely to file a claim — but we don't think that's something we can justify to customers in a way that feels fair.

Our standard is that risk should be assessed based on what people **do**, not who they **are**. Behavior-based features like **CLM_FREQ**, **TRAVTIME**, and **REVOKED** make sense. But location and job titles are often tied to deeper structural disadvantages. Even if these seem neutral, they act as proxies for things like wealth, access, or privilege — which is exactly what we're trying to avoid.

SHAP and DiCE made it clear these features carry real weight in predictions. DiCE showed that small changes in occupation or location alone could flip outcomes, even if everything else stayed constant. That's a red flag for fairness.

So our 3 recommendations are:

- Exclude **URBANICITY** and **OCCUPATION**
- Apply strict review before using **INCOME, AGE, or HOME_VAL**
- Provide users with guidance on what they **can actually control** — not factors out of their hands

Consent also needs to evolve. We don't think most users fully understand how data like ZIP code or occupation might affect their premiums. We suggest clearer summaries, transparency about feature use, and optional opt-outs in our journey to build trust with our customers that will reap benefits later and that tradeoff is worth it.

Conclusion:

Our XAI-based fairness evaluation revealed that features like URBANICITY and OCCUPATION, though predictive, introduce unfairness due to their ties to socioeconomic status and historical bias. We recommend excluding them from production models. We propose a standard that models should assess risk based on behavior, not demographic or socioeconomic proxies. Additionally, consent practices should be more transparent—users should clearly understand how data like ZIP code or job title affect their premiums. These steps promote fairness, accountability, and long-term trust in our predictive systems.