Report Update 2A

Introduction:

This report evaluates the explainability of predictive models developed to assess car insurance claim behavior. The primary ethical challenge addressed is the need for transparency in decision-making in insurance, where predictions can significantly impact individuals' financial outcomes. While traditional performance metrics such as accuracy or R^2 provide insight into model effectiveness, they do not explain why a model makes a specific prediction—information that may be critical for affected individuals. To build trust, ensure fairness, and support contestability, we employ explainable AI (xAI) techniques, including feature importance and counterfactual methods. This report applies SHAP, LIME, and DiCE to the models developed in Project Update 1 and presents actionable insights and recommendations.

Data collection and preprocessing:

The data used in this project originates from a car insurance claims database and contains 10,302 instances with 21 features, including demographic, vehicle, and claim-related information. We removed sensitive features such as GENDER, MSTATUS (marital status), PARENT1 (parental status), and EDUCATION. We believe these could lead to biased predictions. Missing values were handled by imputing the median for numerical features and the mode for categorical features. For our logistic regression model, our target variable was CLAIM_FLAG (claimed filed or not), and for our XGBoost we used CLM_AMT (claim dollar amount).

Model Overview (Models we used during update 1):

The logistic regression offers a straightforward, transparent approach, allowing stakeholders to easily interpret how different variables impact predictions related to insurance claims. Its simplicity and linear structure highlight the relationships between predictors and outcomes.Conversely, the XGBoost regression model employs multiple decision trees, resulting in greater complexity and reduced interpretability. Due to its inherent complexity, this model requires supplementary explainability methods such as SHAP and DiCE, enabling clear and understandable explanations of individual predictions.

Feature Importance and Explainability:

To understand which features influenced the model's predictions most significantly, we used SHAP (SHapley Additive exPlanations). SHAP is an approach to explain individual predictions by computing the contribution of each feature to a model's output. It provides both local (individual-level) and global (dataset-level) insights. For the logistic regression model predicting CLAIM_FLAG, the SHAP summary plot and bar chart revealed that the most influential features included:

URBANICITY - Area where the driver lives or works	**BLUEBOOK** - Value of the vehicle
CLM_AMT - Cost of claim	**KIDSDRIV** - Number of driving children
REVOKED - Driver's license been revoked in the last five years?	**CLM_FREQ** - Total number of claims in the past five years
CAR_TYPE - Type of vehicle	

Rural-urban city and prior claim amount (CLM_AMT) dominated. REVOKED_Yes and ownership of sports cars also had a strong effect on claim predictions.

Ethical Implications of SHAP Output:

The SHAP results suggest that the model relies on logical, relevant indicators, like driving history (REVOKED, CLM_FREQ), car type, and vehicle value, all appropriate for assessing risk. Urbanicity might raise fairness concerns about a persons geography. If rural vs. urban status is heavily weighted, it could introduce geographic or socioeconomic discriminating, especially if rural drivers are unfairly treated due to factors beyond their control. Additionally, the model's high reliance on prior claim amount (CLM_AMT) may lead to reinforcing patterns rather than offering a fair reassessment of future risk.

Ethical Challenges and Considerations of Explainable AI

Explainable AI methods can enhances trust and fairness. However, it also presents an ethical challenges. Increased transparency may expose sensitive personal information, potentially compromising privacy. Moreover, since explanations are reliant on the underlying data, any biases present in the data could translate into biased explanations, potentially reinforcing unfair practices or discriminatory outcomes.

Another concern involves the potential for stakeholders to misunderstand or overely on XAI-generated explanations, which may lead to incorrect decisions. Communication about the strengths, limitations, and appropriate usage of XAI is critical. Transparency from XAI complicates accountability, making it challenging to attribute responsibility when AI-driven decisions result in negative outcomes. To address this, organizations must develop clear governance policies and assign responsibilities.

The use of occupation as a predictive factor demands ethical scrutiny. Although occupation is a common and predictive for insurance contexts, it poses ethical concerns regarding economic discrimination. Differentiating rates based on an individual's occupation can lead to unfair treatment and discriminatory practices. To mitigate these concerns, the model excludes non-economic like gender, marital status, and education. Should occupation remain a feature in the model, it should be accompanied by explicit, transparent guidelines designed to prevent unjustified discrimination and align with established legal and ethical standards.

From a deontological ethical perspective, grounded in Kantian philosophy, XAI methods like SHAP or Wachter's Counterfactuals align with the duty of ensuring transparency, fairness, and respect for individual autonomy. Kant emphasizes the importance of moral purity and intentions, asserting that the ethical value of actions is determined by their intent rather than outcomes. Therefore, employing XAI techniques is ethically imperative as it upholds moral integrity and promotes fairness and accountability.

Standard:

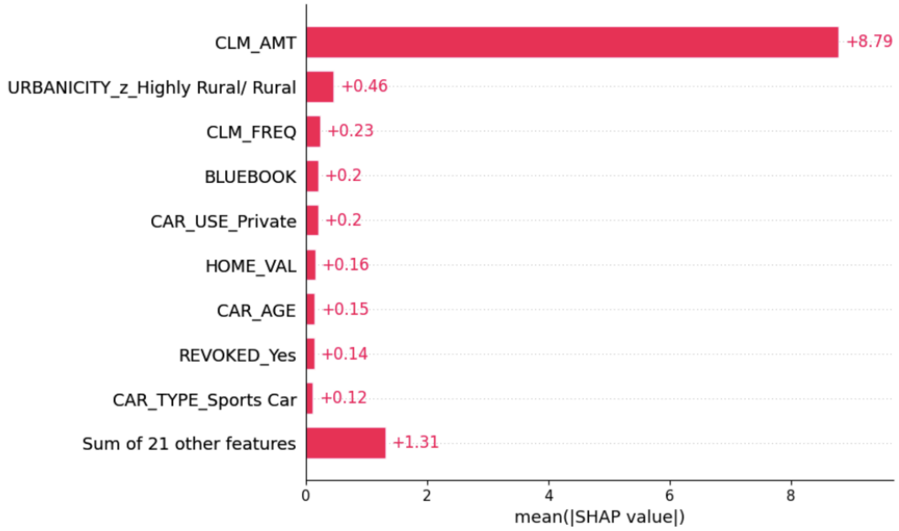
Our view is that risk assessment should depend on the way people act and not on who they are, where they live, or their social and economic circumstances. Both job, urban or rural locations from SHAP turned out to be significant in our analyses. Although we regard the frequency of claims, the age of the car, and its value were good indicators of risk. We will object to the use of occupation and urbanicity on the grounds of ethical fairness.

LIME revealed that individuals with "professional" occupation would be better able to predict claim values, all else was equal. That indicates unfairness on the basis of occupation titles alone. DiCE counterfactuals revealed that slight modifications in where a person lives or the occupation they have may alter predictions. This questions the fairness of our model—particularly for groups with a history of discrimination in the past. Redlining and racial disparity in driving records (e.g., higher MVR_PTS scores) in the real world illustrate the additional unfairness added through these factors in model predictions. This is why we do not consider these factors unless we have visible, non-sensitive behaviors.

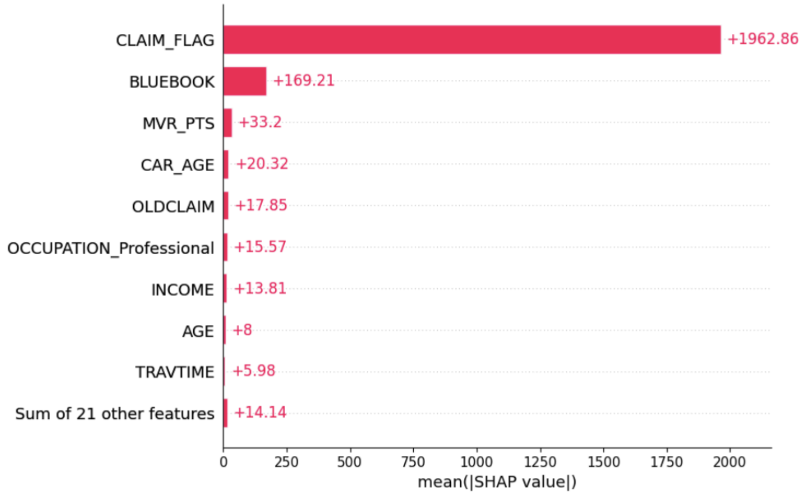
APPENDIX

Key visuals:

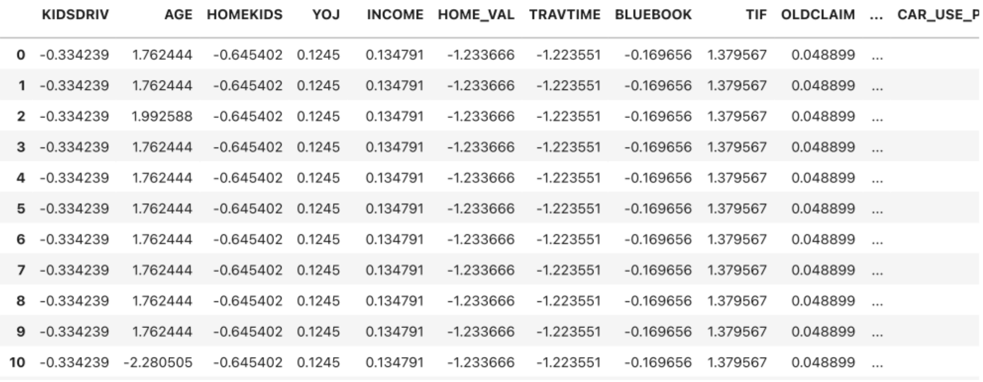
1- SHAP Plot for Logistic Regression



2- SHAP Plot for XGBoost

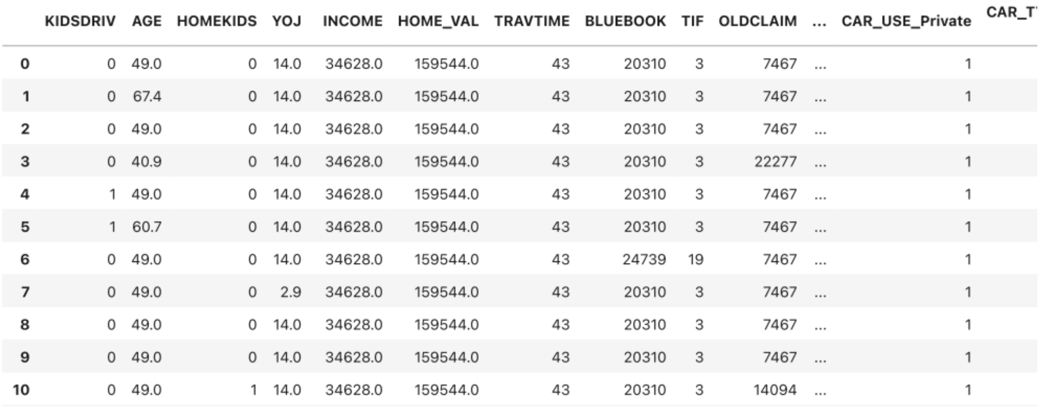


3- XAI DiCE for Logistic Regression



	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	...	CAR_USE_P
0	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
1	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
2	-0.334239	1.992588	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
3	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
4	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
5	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
6	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
7	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
8	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
9	-0.334239	1.762444	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	
10	-0.334239	-2.280505	-0.645402	0.1245	0.134791	-1.233666	-1.223551	-0.169656	1.379567	0.048899	...	

4- XAI DiCE for XGBoost



	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	...	CAR_USE_Private	CAR_T
0	0	49.0	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
1	0	67.4	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
2	0	49.0	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
3	0	40.9	0	14.0	34628.0	159544.0	43	20310	3	22277	...	1	
4	1	49.0	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
5	1	60.7	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
6	0	49.0	0	14.0	34628.0	159544.0	43	24739	19	7467	...	1	
7	0	49.0	0	2.9	34628.0	159544.0	43	20310	3	7467	...	1	
8	0	49.0	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
9	0	49.0	0	14.0	34628.0	159544.0	43	20310	3	7467	...	1	
10	0	49.0	1	14.0	34628.0	159544.0	43	20310	3	14094	...	1	