

*C03051: Master of Analytics(Research)*

*CRICOS Code: 075277F*

*A Feasible Situation Awareness-Based Evaluation Framework for Quality of Machine  
Learning Explanations*

*March 2024*

# *Situation Awareness-Based Evaluation Framework for Quality of Machine Learning Explanations*

---

*Jamie Wu*

School of Computer Science  
Faculty of Eng. & IT  
University of Technology Sydney  
NSW - 2007, Australia



## CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Jamie Wu, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Jamie Wu]

DATE: 10<sup>th</sup> March, 2024

PLACE: Sydney, Australia



## ABSTRACT

EXplainable Artificial Intelligence (XAI) has emerged as a critical domain, with the aim of enhancing the transparency and interpretability of advanced machine learning (ML) models. As the need to introduce more complicated ML in broader industries surged, especially for industries with high sensitivity to safety, such as finance and medicine, explanations for the complexity of the models attract attention. A lack of methodology for in-context user need analysis and low-level evaluations of explanations appears to be the pain point for professionals in such industries to use advanced ML models confidently. As a case study, in one of the safety-sensitive fields, actuarial insurance pricing, this research addresses a significant gap by focussing on user-centric evaluations of XAI explanations. The study unfolds in two main parts.

The first study uses the Actuarial Control Cycle (ACC) and the Goal-Directed Task Analysis (GDTA) framework to conduct a detailed analysis of user needs in insurance pricing. Focussing on the prediction of claim counts for Motor Third Party Liability Insurance (MTPL) using Generalised Linear Models (GLM), this study establishes a robust foundation for understanding the nuanced requirements of actuarial professionals in complex pricing scenarios.

Building on the insights gained from the first study, the second study evaluates the effectiveness of XAI explanations, particularly those derived from SHAP values. A user-participated questionnaire, grounded in Endsley's 1995 Model of Situation Awareness, provides quantitative metrics to assess the Situation Awareness (SA) of users. This study delves into the user-centric evaluation of XAI techniques in the specific context of insurance pricing, contributing to the evolving landscape of XAI.

Synthesising the results of both studies, the research challenges the traditional limitations in explaining ML models and highlights the importance of aligning XAI techniques with user needs, fostering transparency, trustworthiness, and effective decision-making in the intricate field of actuarial science. The discussions underscore implications for refining XAI methodologies, improving explanations, and improving user satisfaction. The study acknowledges limitations and challenges while emphasising the need for an iterative control cycle of the effectiveness of XAI, enabling ongoing collaboration between model developers and users to refine explanations and promote a symbiotic dynamic within the actuarial control cycle.



## ACKNOWLEDGMENTS

First, I express my sincere gratitude to my principal supervisor, Jianlong Zhou, for his invaluable guidance, unwavering support, and constant encouragement throughout my research journey. I am also deeply grateful to my co-supervisor, Dr. Zhidong Li, for his insightful suggestions and constructive feedback, which have greatly contributed to the quality of this thesis.

I am grateful to the Australian Government for providing financial support through the Research Training Programme (RTP) and the University of Technology Sydney (UTS), particularly the School of Computer Science within the Faculty of Engineering and IT, for providing me with the necessary resources and a conducive environment to pursue my research. Also, I acknowledge Dr. Chandranath Adak for providing this thesis template.

A special thanks goes to the two adorable cats from my neighbour who frequently visited me during the challenging year of 2023. Their comforting presence helped me through difficult moments and brought me joy and motivation. I am forever grateful to my parents for their unconditional love and unwavering support, and to my friends who patiently listened to me and provided understanding and encouragement.

Lastly, I would like to sincerely thank the actuarial professionals from various companies who generously volunteered their time to participate in this study. Your insights and expertise have contributed significantly to the depth and relevance of this research.





## LIST OF ABBREVIATIONS

1. XAI: eXplainable Artificial Intelligence
2. GDPR: General Data Protection Regulation
3. CRR: European Union's Capital Requirements Regulation
4. MTPL: Motor Third Party Liability (insurance)
5. SHAP: SHapley Additive exPlanation
6. GLM: Generalised Linear Regression Model
7. SAGAT: Situation Awareness Global Assessment Technique
8. SPAM: Situation Present Assessment Method
9. A-GDTA: Actuarial Goal-Directed Task Analysis
10. GLM-XAI system: A system consists of GLM-based machine learning method and XAI-based explanations for the machine learning method.
11. GLM-SA system: A system consists of GLM-based machine learning method and SA-based user need of explanations for the machine learning method.
12. GLM-SA I: Level one of situation awareness in a GLM-based explaining system
13. GLM-SA II: Level two of situation awareness in a GLM-based explaining system
14. GLM-SA III: Level three of situation awareness in a GLM-based explaining system

# TABLE OF CONTENTS

<b>List of Abbreviations</b>	<b>ix</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 High-Stakes Industries . . . . .	2
1.3 Motivation and Research Question . . . . .	6
1.3.1 The Established Foundation in Explaining Machine Learning Models . . . . .	6
1.3.2 The Lack of Evaluations for Machine Learning Explanations by XAI . . . . .	7
1.3.3 Implications to the High-stakes Industries . . . . .	7
1.3.4 Research Questions . . . . .	8
1.4 Research Aims and Research Objectives . . . . .	10
1.5 Significance of the Research . . . . .	11
1.6 Structure of the Thesis . . . . .	12
<b>2 Literature Review</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 XAI From a Historical Perspective . . . . .	15
2.2.1 Evolution of Machine Learning: From Competing with Humans to Ensuring Accountability in Complex Models . . . . .	16
2.2.2 Rising Importance of Explainability in Machine Learning: Evolution, Challenges, and a Surge in XAI Research . . . . .	16
2.2.3 Summary . . . . .	19
2.3 The Concepts in XAI: Arguments and consensus . . . . .	19
2.3.1 What is XAI . . . . .	19
2.3.2 The Aim of Research in XAI . . . . .	20
2.4 Evaluation of Explanations Provided by XAI . . . . .	21
2.4.1 Exploring Connections Between Machine Learning Methods, Black-Box Explanations, and Evaluation Frameworks in XAI . . . . .	21
2.4.2 Effort in Evaluation of Explanations Towards User Need . . . . .	22
2.4.3 Situation Awareness as Evaluation Metrics in Dynamic Decision Making . . . . .	25

2.5	XAI Application in Non-Life Insurance Pricing . . . . .	26
2.5.1	Background and Scope . . . . .	26
2.5.2	Fundamental Insurance Concepts and Industrial Context . . . . .	27
2.5.3	Machine Learning in Non-life Insurance Pricing . . . . .	30
2.5.4	Literature Review Process of XAI Application in Non-life Insurance Pricing . . . . .	32
2.5.5	Systematic Review Results of XAI Application in Non-Life Insurance Pricing . . . . .	34
2.6	Conclusion . . . . .	40
<b>3</b>	<b>User needs Analysis Towards Responsible XAI</b>	<b>43</b>
	<b>Actuarial Goal-Directed Task Analysis(A-GDTA): A Case Study</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Applying the Goal-Directed Task Analysis (GDTA) Framework in High-Stakes Industries . . . . .	45
3.3	Research Objective: Analyse User Needs Using A-GDTA Framework . . . . .	48
3.4	Methodology . . . . .	49
3.4.1	Qualitative Research Method . . . . .	49
3.4.2	Data Collection . . . . .	50
3.4.3	Participants and the Use Case . . . . .	51
3.4.4	Experiment Design: Actuarial Goal-Directed Task Analysis . . . . .	53
3.4.5	Remarks: Integration with Actuarial Control Cycle . . . . .	55
3.5	Results and Analysis . . . . .	56
3.5.1	Determining the user needs: Desiderata of Effective EXplainable Artificial Intelli- gence(XAI) . . . . .	56
3.5.2	Application of Endsley 1995 Model of SA . . . . .	59
3.5.3	Categorising the User Needs: Situation-Awareness Based User Needs . . . . .	60
3.6	Evaluation of Results . . . . .	62
3.7	Extend the Outcome to A More General Context . . . . .	63
3.8	Conclusion . . . . .	64
<b>4</b>	<b>User-based Evaluation by Situation Awareness Metrics</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.2	Introduction to Explanation Methods . . . . .	69
4.2.1	Feature Importance: Permutation Graph and SHAP Value . . . . .	69
4.2.2	Global Explanations of Feature Effect: ICE Plots and SHAP Value . . . . .	70
4.3	Research Objective . . . . .	71
4.4	Methodology . . . . .	72
4.4.1	Choosing a Cross-sectional Quantitative Study Method . . . . .	72
4.4.2	Experiment Design . . . . .	73
4.4.3	Data Source . . . . .	75
4.4.4	Situation Awareness Metrics . . . . .	76
4.5	Experiment Stage 1: Building the GLM-XAI System . . . . .	78
4.5.1	The Use Case . . . . .	78

4.5.2	Building a Predictive Generalised Linear Model . . . . .	79
4.5.3	Results: Explaining the Predictive GLM Using XAI Techniques . . . . .	79
4.5.4	Summary . . . . .	83
4.6	Experiment Stage 2: User Participating Evaluation . . . . .	84
4.6.1	User-participating Questionnaire . . . . .	84
4.6.2	Hypotheses . . . . .	85
4.6.3	Results and Evaluating Metrics . . . . .	86
4.7	Remarks and Future Work . . . . .	89
4.7.1	Extend to General Context: Explanations of Good Quality . . . . .	89
4.7.2	Level 3 Model Situation Awareness . . . . .	90
4.7.3	Post Experiment Insights . . . . .	91
4.7.4	Future Work of Deeper Investigating in Level 3 User Needs . . . . .	92
4.8	Conclusion . . . . .	93
<b>5</b>	<b>Discussion</b>	<b>95</b>
5.1	Recap of Results and the Answer to Research Questions . . . . .	95
5.2	Key Findings and Comparison to Previous Research . . . . .	97
5.3	Remarks and Future Research Directions . . . . .	98
5.4	Publication Progress and Future Plan . . . . .	101
5.4.1	Accepted Conference Papers . . . . .	101
5.4.2	Proposed Journal Publications . . . . .	101
5.4.3	Alternative Journal Considerations . . . . .	102
<b>6</b>	<b>Conclusion</b>	<b>103</b>
6.1	Machine Learning, EXplainable Artificial Intelligence, and Safety-sensitive Industries . . .	103
6.2	User-Based Evaluation of Machine Learning Explanations . . . . .	103
6.3	Bring Light in Real-life Scenario . . . . .	104
6.4	Future Avenues in User-centric XAI for Safety-sensitive Industries . . . . .	105
<b>7</b>	<b>Appendix</b>	<b>107</b>
	<b>Bibliography</b>	<b>121</b>

## LIST OF FIGURES

FIGURE	Page
1.1 Relationships among ML, AL, XAI, and Insurance Industry . . . . .	5
1.2 XAI Layer in Insurance Pricing . . . . .	13
2.1 Research Outcomes Search Result Count by Year . . . . .	18
2.2 Model of SA in dynamic decision making[41] . . . . .	24
2.3 Flowchart Illustrating the Literature Review Methodology . . . . .	33
2.4 According to[98], this graph depicts a black box classification model (f) delineated by pink and blue regions. The focal point of explanation is represented by a bold red cross, surrounded by locally sampled instances denoted by red crosses and blue circles, their proximity serving as weights. The locally faithful explanation (g), portrayed as a dashed line, elucidates the model's behavior within the specified region. . . . .	38
2.5 According to[77], SHAP values attributing to each feature the change in the expected model prediction when conditioning on that feature. . . . .	38
3.1 Relationship chart illustrating the combination of ACC and GDTA into the A-GDTA framework for user needs analysis. . . . .	47
3.2 User-Based Analysis Method . . . . .	52
3.3 Two-Step Experiment Design Diagram . . . . .	56
3.4 Model of SA in dynamic decision making[41] . . . . .	60
3.5 The Categorised Three Level GLM-SA User Needs . . . . .	62
4.1 The process of explanation . . . . .	75
4.2 Flow chart of steps in questionnaire . . . . .	75
4.3 Age Distribution of Participants . . . . .	77
4.4 Gender Distribution of Participants . . . . .	77
4.5 Participants Demographics . . . . .	77
4.6 The Permutation Feature Importance Graph . . . . .	81
4.7 The Permutation Feature Importance Graph . . . . .	81
4.8 The Permutation Feature Importance Graph . . . . .	82
4.9 The ICE Graph for the Feature Town . . . . .	83
4.10 The Graph of SHAP Value by Feature Value . . . . .	83

## LIST OF TABLES

TABLE	Page
1.1 XAI Applications across High-stakes Industries . . . . .	3
1.2 Roadmap: User-Centric XAI Evaluation in Insurance Pricing . . . . .	14
2.1 AI and XAI from a Historical View . . . . .	18
2.2 Insurance Concepts and Definitions . . . . .	31
2.3 Table. XAI Methods and Actuarial Concepts as Interchangeable Key Search Terms . . . . .	32
2.4 Summary of XAI Methods Applied in Pricing Tasks . . . . .	35
3.1 Keywords on Pricing Objectives: We have referred to but not limited to the list of references: [78][29] [64][30][91][7][94][18]. We claim that several industry reports with content similar to the referenced material are available online. Due to length constraints, we do not enumerate every document we have examined. . . . .	57
4.1 Policy Data Dimensions . . . . .	79
4.2 Estimation of GLM . . . . .	79
4.3 Statistical Test Results of SELF-SA . . . . .	86
4.4 Description of Quantitative metrics for level 1 and level 2 of Situation Awareness . . . . .	87
4.5 Test Result for GLM-SA1 and GLM-SA2 . . . . .	88

## INTRODUCTION

## 1.1 Overview

XAI (eXplainable Artificial Intelligence) is a research area focused on developing AI systems that can provide explanations for their decisions and actions. The goal of XAI is to make AI more transparent, interpretable, and trustworthy to end users. Research in XAI involves developing new techniques and methods to generate explanations for ML(Machine Learning) models and designing interfaces that allow users to interact with AI systems and understand the explanations provided.

Well-known machine learning method such as XGBoost predictive classification algorithm, neural network, support vector machine, etc. have been widely accepted to be helpful in safety-sensitive areas from an academic perspective[100]. However, due to the limit in complexity and explainability, there is still a long way to go if we want to use these methods in the industry environment. This brings our attention to XAI. For example, research in XAI provides a gateway for industry users to shed light to the black box and understand how the underlying advanced machine learning method behaves and why the machine learning algorithms provide such results. Despite this growing interest in the development of XAI methods, the evaluation of the quality of explanations provided by XAI methods has not received the same attention within the XAI research community. The rapid advancement of XAI has led to exciting possibilities for bridging the gap between cutting-edge machine learning techniques and practical industry applications. However, this progress comes with a critical caveat: the absence of a comprehensive evaluation system. Without robust evaluation criteria, there is a risk of misusing the XAI results when they are not sufficiently reliable or informative for end users.

To address this gap, our research aims to develop rigorous evaluation frameworks for XAI methods. The framework will consider not only the general properties of explanation quality, but also the practical utility and impact. By conducting a real-life case study, we proposed a framework to evaluate explanations, and applied the method to a niche area within insurance pricing, which is a representative of safety-sensitive fields. By doing so, we can ensure that XAI techniques are not only scientifically sound but also practically applicable in real-world scenarios. While the field of XAI continues to evolve, it is imperative that we invest

in comprehensive evaluation mechanisms. Only through a rigorous assessment can we confidently deploy XAI solutions that empower end users and facilitate informed decision making.

In this chapter, we provide an introduction to our journey through XAI. We first provide the background of this research area and the industry context in which we will focus. Second, we discuss how the research problem of evaluating the explanations arises. Our compass then guides us toward the research aims, objectives, and probing questions. As we navigate this terrain, we recognise the importance of unraveling the black box of machine learning methods for practical industry applications. However, like any expedition, we acknowledge the limitations that eventually accompany our quest.

## 1.2 High-Stakes Industries

The reliability of methods adopted is crucial in industries such as healthcare, aviation, and finance, where the stakes are high and the impact of AI decisions can be far-reaching. In healthcare, XAI can be used to explain the predictions made by AI models in diagnosing diseases or recommending treatments. By providing clear explanations for the factors contributing to a particular diagnosis or treatment recommendation, healthcare professionals can make more informed decisions and ensure that the AI system is not perpetuating biases or making errors. Similarly, in aviation, XAI can be employed to explain the decisions made by autonomous systems, such as collision avoidance or flight path optimization, to pilots and air traffic controllers. This transparency helps build trust in the AI system and allows humans to intervene if necessary.

Before delving into insurance pricing as the focal point of this study, we first explore the broader applications of XAI within high-stakes industries. XAI has emerged as a critical tool in domains where decision-making processes must be transparent, interpretable, and accountable. High-stakes industries, such as healthcare, financial services, legal and judicial systems, and aviation, rely heavily on AI-driven solutions to optimize operations, enhance accuracy, and ensure compliance with regulatory standards. However, the complexity of AI models often obscures their decision-making processes, necessitating the integration of XAI to provide clarity and build trust among stakeholders. As summarised in Table 1.1, this section provides an overview of XAI applications across these industries, highlighting their unique requirements and challenges.

In healthcare, XAI enhances clinical decision-making by providing interpretable insights into AI-driven diagnostics. For example, saliency maps highlight critical regions in medical images, aiding in disease detection, while XAI systems explain diagnoses by linking patient data to outcomes, fostering trust. Key requirements include clinical accuracy, patient trust, and adherence to medical protocols. In financial services, XAI ensures transparency and compliance. It clarifies credit scoring models, reducing bias in lending decisions, and justifies fraud detection alerts by identifying suspicious transaction patterns. This promotes fair practices, regulatory compliance, and customer trust. In the legal and judicial sector, XAI supports risk assessment, case outcome prediction, and sentencing decisions. Transparent recidivism risk models ensure due process and fairness, while XAI insights help legal professionals make accountable decisions. In aviation and transportation, XAI improves safety and efficiency by explaining route optimization and justifying safety decisions. Real-time system behavior interpretation is critical for emergency response,



Table 1.1: XAI Applications across High-stakes Industries

Industry	XAI Applications	Key Requirements
Healthcare	<ul style="list-style-type: none"> <li>Clinical decision support systems using saliency maps for disease detection [117]</li> <li>Explanation of diagnosis recommendations based on patient data</li> </ul>	<ul style="list-style-type: none"> <li>Clinical accuracy</li> <li>Patient trust</li> <li>Compliance with medical protocols</li> </ul>
Financial Services	<ul style="list-style-type: none"> <li>Credit scoring explanations [33]</li> <li>Fraud detection alerts [26]</li> <li>Automated lending decisions with factor importance analysis [104]</li> </ul>	<ul style="list-style-type: none"> <li>Regulatory compliance</li> <li>Fair lending practices</li> <li>Customer transparency</li> </ul>
Legal & Judicial	<ul style="list-style-type: none"> <li>Risk assessment tools for recidivism prediction</li> <li>Case outcome prediction with factor analysis</li> <li>Decision support for sentencing [46]</li> </ul>	<ul style="list-style-type: none"> <li>Due process</li> <li>Algorithmic fairness</li> <li>Legal accountability</li> </ul>
Aviation & Transportation	<ul style="list-style-type: none"> <li>Route optimization explanations</li> <li>Safety decision justification</li> <li>Real-time system behavior interpretation [71]</li> </ul>	<ul style="list-style-type: none"> <li>Safety assurance</li> <li>Operational efficiency</li> <li>Emergency response readiness</li> </ul>

ensuring reliability and transparency.

Overall, XAI adoption in high-stakes industries highlights the importance of interpretability, accountability, and trust in AI systems. By addressing domain-specific needs, XAI enables effective AI utilisation while mitigating risks. This foundation sets the stage for exploring XAI in insurance pricing. While XAI has proven valuable across high-stakes industries, insurance pricing stands out as a strategic focus for deeper study. The reliance on complex AI models to assess risk and set premiums often results in opaque decisions, creating challenges for transparency and trust. By choosing insurance as our entry point, we

can explore how XAI addresses these issues in a domain where fairness, accountability, and regulatory compliance are critical, offering both technical and ethical insights into AI-driven decision-making. In the insurance industry, XAI can be applied to the process of insurance pricing. Insurance companies often use complex AI models to determine premiums based on various risk factors, such as age, driving history, and location, in the context of car insurance pricing. However, these models can sometimes produce results that are difficult to interpret or explain to customers. By incorporating XAI techniques, insurance companies can provide clear explanations for the factors that contribute to a particular premium, making the pricing process more transparent and understandable to policyholders. The insurance industry, designed to provide financial protection against unforeseen events, relies on actuarial science as an indispensable tool for its very existence. Insurance is fundamentally a business of risk, and companies assume the responsibility of compensating policyholders in the event of adverse and unpredictable events. Actuaries play a vital role by employing mathematical and statistical methods to meticulously assess, quantify, and price these risks. Their expertise ensures that insurance premiums accurately reflect the potential financial liabilities associated with various events, allowing insurers to maintain financial stability and fulfil their promises to policyholders. Actuarial science is a discipline that applies mathematical and statistical methods to assess and manage financial risks. Actuaries have been using machine learning methods to analyse data, evaluate the potential future environment, and make informed predictions about financial uncertainties. By combining machine learning modelling, statistical analysis, and financial theory, actuaries play a crucial role in helping organisations make sound decisions related to risk management, ensuring the long-term financial stability of insurance companies, pension funds, and other entities.

There are adequate studies in the area of actuarial science using machine learning methods and advanced AI techniques. The extensive survey by Richman[99] demonstrates various successful applications of AI and deep learning across multiple actuarial domains, including mortality modeling, claims reserving, non-life pricing. The survey's inclusion of practical implementations through publicly available code repositories further validates the maturity and accessibility of machine learning research in actuarial applications, reinforcing the acknowledgment of this discipline intersection. Because insurance is a highly regulated industry, there is a relatively high demand for the use of explainable methods to produce results that affect the interests of policyholders, regulators, and internal governance. In the context of insurance industry, research shows that XAI methods have been used more in the areas of claims management, underwriting, and actuarial pricing compared to other areas such as liability valuation and actuarial financial reporting. We illustrate the links in Figure1.1.

As mentioned above, insurance is a highly regulated industry. The higher level of regulatory requirements applying to all industries using data to generate results for decision making applies to insurance industry as well. As most data used by insurance industry including policyholder's information and claimants' information is categorised into personal data, it is required by a number of data legislation to provide explanations for the data analytical methods used for the decision making in the industry. Also, there are industry level of regulations such as Australian Prudential Regulation Authority (APRA), which is an independent statutory authority that supervises institutions across banking, insurance, and superannuation and promotes financial system stability in Australia. APRA is responsible for regulating the insurance industry to protect the interests of policyholders and ensure the financial soundness of

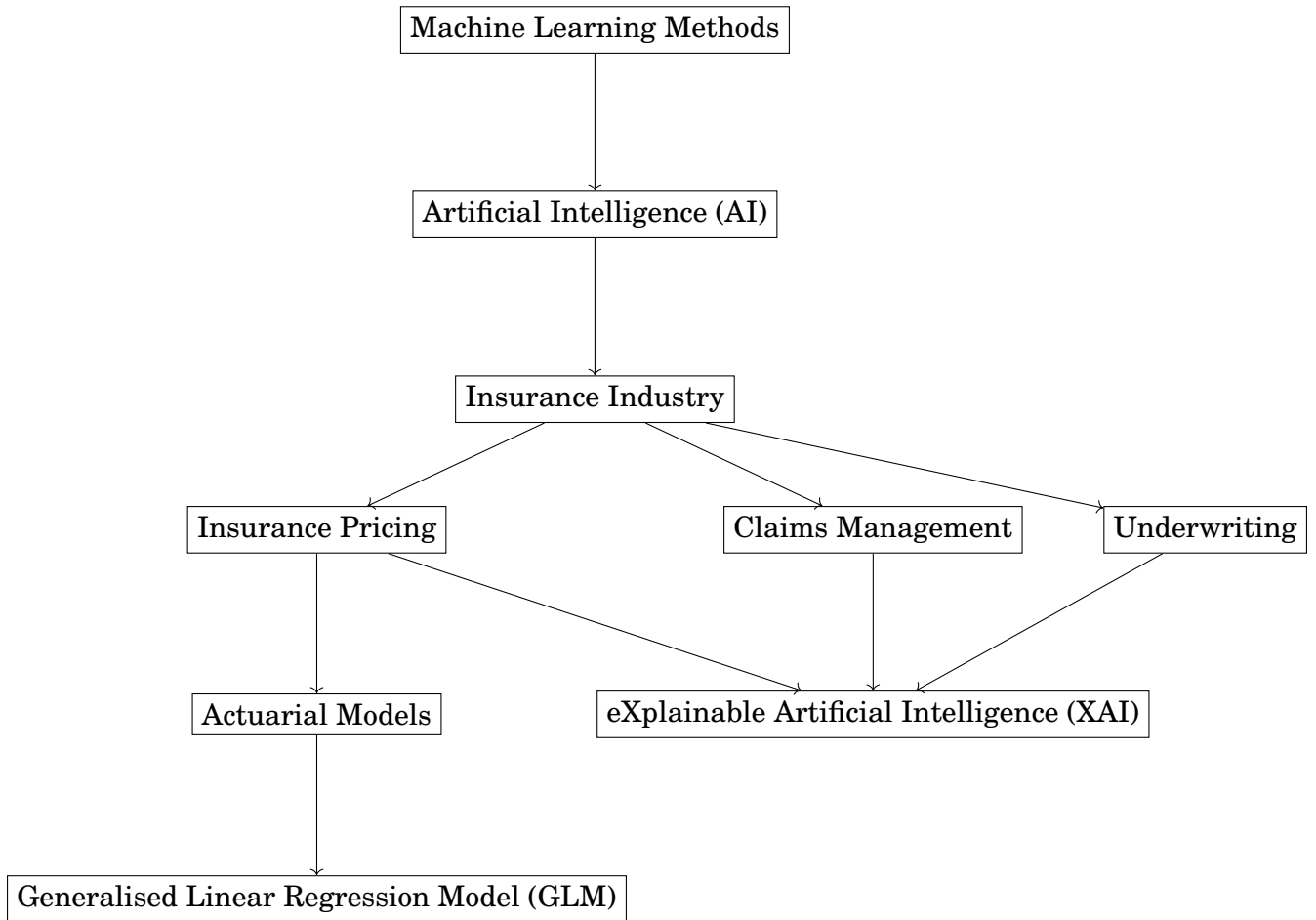


Figure 1.1: Relationships among ML, AI, XAI, and Insurance Industry

insurance companies. Insurance pricing is subject to APRA's regulation because it directly impacts the financial viability of insurance companies and the affordability and accessibility of insurance products for consumers.

There are many scenarios in insurance where commercial decision making is based on personal data. Assumptions in actuarial modelling on anti-selection or fraud may be determined by analysing fraud behaviours in individual level of claims data. This could be a material input of the pricing model that builds the price that customers need to pay. Regulatory requirements on explaining the application of machine learning methods create a demand on using explanation tools such as XAI techniques to pair with utilising advanced machine learning methods. If we broaden our perspective beyond the insurance example and zoom back to all high-stakes application scenarios, there are several regulatory systems that provide high-level guidelines for the responsible use of AI. The General Data Protection Regulation (GDPR), a comprehensive data protection and privacy regulation enacted by the European Union (EU), establishes a robust framework for the processing of personal data and grants individuals greater control over their own information. For instance, the GDPR stipulates that organizations must handle personal data transparently, ensuring consent, data security, and providing individuals with the right to access and control their personal information held by companies. Moreover, the GDPR mandates that decisions made by AI systems be explainable to individuals, thereby promoting accountability and trust in AI-driven processes.

From a global view, the EU's Capital Requirements Regulation (CRR) 2013 serves as another example

where explanations are required when employing complex AI systems for commercial decision-making. This regulation aims to ensure that financial institutions maintain sufficient capital to withstand economic stress and manage risks effectively. In the context of AI, the CRR 2013 emphasizes the importance of transparency and interpretability in AI models used for risk assessment and capital allocation. For example, if a bank uses an AI system to determine credit risk, the CRR 2013 would require the bank to provide clear explanations of the factors influencing the AI's decisions, enabling regulators and stakeholders to assess the system's reliability and fairness.

In conclusion, if there are well-developed XAI techniques that can be applied in the context of safety-sensitive industries such as insurance pricing, XAI has significant potential to play a cornerstone role in pushing the edge for the application of advanced machine learning methods to the industry, providing more solid analysis support for regulatory requirements.

## **1.3 Motivation and Research Question**

### **1.3.1 The Established Foundation in Explaining Machine Learning Models**

Several established methods exist for explaining complex machine learning models. Pairwise empirical Pearson correlations[96] are commonly used to elucidate the direction and scale of the effect of individual features on predictions. This method, widely applied in various contexts, facilitates understanding the interplay between features, guiding users in determining whether to incorporate interaction terms based on feature correlations[25]. Additionally, in systems based on Generalised Linear Models(GLMs), understanding the linearity and additivity of feature effects after rescaling is crucial. Professionals employing such models must possess a solid mathematical background to interpret the true effects of features, often calculated transparently within the regression equation. The choice of the loss function in GLM models also serves as an explanatory tool, with feature importance based on permutation calculated from test data losses helping identify significant features. However, interpretability is closely tied to the selected loss function, emphasising the need for alignment between the modelling methodology and the loss function to ensure accurate and meaningful explanations.

These existing explanation methods highlight the importance of comprehending feature relationships, linearity, additivity, and loss functions in interpreting machine learning model results. However, they also underscore the challenges inherent in dependence on specific methodologies and the potential for inappropriate explanations if the chosen methods are not well-suited. For example, feature importance calculated from permutation for models trained by different loss function is not comparable across models. This discussion underscores the need for advancements in XAI, aiming to address these limitations and enhance the interpretability of complex models. Most of the existing methods are used as traditional ways of explaining machine learning methods. In the meanwhile, for those techniques that have been developed separately, usually it is model-agnostic, and developed after the concept of XAI has been highly referred to and become a research area.

In the new field of XAI, researchers initially faced challenges of defining its scope, emphasising the importance of providing explanations to enhance system robustness and prevent bias, unfairness, and discrimination[4]. Despite the lack of consensus on a strict definition, the literature acknowledges XAI as a

domain focused on developing novel explanation methods, a perspective supported by extensive literature reviews that examine the evolution of explainability concepts in machine learning. Notably, it is distinct from delving into black-box machine learning methodologies, as intrinsic XAI methods primarily explore model dynamics and impose constraints on original machine learning methods.

### **1.3.2 The Lack of Evaluations for Machine Learning Explanations by XAI**

In the evaluation of explanations within XAI, scholars emphasise the necessity of assessing explanations for advanced Black-Box methods. A prevailing top-down structure, delineates the progression from data types to machine learning methods, XAI methods, explanation methods, and finally, evaluations of explainability. The overarching mindset in this area typically follows a sequence in which machine learning methods require explanations, prompting the development of explanation methods, and subsequently requiring the assessment of explanation quality. However, the lack of a strict definition of explainability in the XAI domain complicates the evaluation, raising the initial challenge of understanding the expectations for evaluating explainability. Two distinct types of research papers emerge: the first focusses on providing explanations for advanced Black-Box methods, often trusting explanation techniques based on theoretical backgrounds without rigorous evaluation; the second, with evaluation methods as the research objective, proposes tools for the XAI community, comparing various methods based on predefined metrics to assess the quality of provided explanations.

Despite these evaluation efforts, there is a notable gap in the assessment landscape, particularly with respect to user-based evaluation. Although some literature acknowledges the importance of communication between method providers and end users, the evaluation typically stops at the interpretable predictor result without delving into the subsequent user-centric processing. Modifying technical results for users with varying experience and interests is identified as a crucial step, but the lack of established evaluation criteria for this phase poses a challenge. As the modifying process is intricately tied to the characteristics of end users, comprehensive evaluation frameworks often fall short in addressing this user-centric gap, leaving the impact of improper processing significant. Therefore, the call for a complete evaluation framework is underscored, encompassing solutions to bridge the gap between end users and technical processes, ensuring a more holistic assessment of the XAI explanations. The ultimate goal of the explanations provided by XAI, as analysed in the existing research, is to enhance human understanding and satisfy the specific informational needs of users in performing particular tasks, and methodologies developed to study these needs open a gateway for objective measurements of whether these needs are effectively satisfied in a given scenario [35].

### **1.3.3 Implications to the High-stakes Industries**

The application of complex machine learning methods in industries that are highly sensitive to safety, finance, and ethics presents significant challenges. While these sophisticated algorithms have the potential to revolutionise decision-making processes and improve efficiency, their inherent lack of transparency raises concerns about their reliability, fairness, and accountability. The "black box" nature of many machine learning models makes it difficult for stakeholders to understand how decisions are being made, which can

lead to a lack of trust and potential risks. This opacity becomes particularly problematic when algorithmic decisions impact vulnerable populations or when there are legal requirements for non-discrimination and equal treatment[120].

XAI tools have emerged as a promising solution to shed light on the inner workings of these complex models. By providing insights into the decision-making process, XAI techniques can help to demystify the black box and increase transparency. These tools play a crucial role in identifying and mitigating algorithmic bias, as they can reveal whether models are making decisions based on protected attributes or proxy variables that may lead to discriminatory outcomes[24]. However, the lack of in-context evaluation of the explanations generated by these tools presents its own set of risks. Without a proper assessment of the quality and relevance of the explanations within the specific context of the industry and the decision at hand, there is a danger that the explanations may be misleading or do not capture the full complexity of the situation, potentially masking underlying biases or unfair practices[113].

The growing importance of algorithmic fairness and accountability has also raised significant legal and regulatory considerations. Organizations deploying machine learning systems may face liability risks if their models make discriminatory decisions or if they cannot provide adequate explanations for critical decisions affecting individuals' rights or opportunities. This has become particularly relevant in regulated industries such as banking, where laws require transparent and justifiable decision-making processes for activities like loan approvals or credit scoring. Furthermore, emerging regulations around algorithmic accountability, such as the "right to explanation" enshrined in data protection laws, mandate that organizations must be able to explain automated decisions to affected individuals[27].

To mitigate these risks and ensure the effective application of machine learning in sensitive industries, evaluating the quality of the explanations in context is crucial. This involves not only assessing the technical accuracy of the explanations but also considering their interpretability, relevance, and actionability within the specific domain. Organizations must establish robust frameworks for evaluating both the fairness of their models and the effectiveness of their explanations in identifying and addressing potential biases. By engaging domain experts and stakeholders in the evaluation process, organisations can ensure that the explanations provided by XAI tools are meaningful, trustworthy, and aligned with the industry's specific requirements and constraints. This contextual evaluation is key to building trust in machine learning models, facilitating their responsible deployment, and realising their full potential to drive innovation and improve decision-making processes while mitigating potential risks and ethical concerns. Moreover, this approach helps organizations maintain compliance with evolving regulatory requirements and establish clear lines of accountability for algorithmic decisions.

### **1.3.4 Research Questions**

High-stake industries require extensive model validation and interpretability due to their significant impact on business decisions and regulatory compliance. Taking insurance industry as an example, the technical environment of an insurance company in Australia is undergoing a transformation from centralised software such as SAS that requires more mechanical intervention to adjust the model to an open-source more efficient environment such as Python, R, etc. Basic machine learning models, such as GLM and random forest classification, have already been applied to the industry level. However, traditional

modelling tools take a long time for professionals to perform a model change, including running programmes and testing, which limited the industry to apply more advanced machine learning methods. Meanwhile, traditional mechanical modelling can hardly keep up with the pace of market changes, especially in the pricing context.

When it is time to welcome the reform, it is also the time that we have to face the challenge of explaining complicated machine learning methods. As insurance is a highly regulated industry, there is a high potential to use explainable methods to produce results that affect the interests of policyholders, regulators, and internal governance. We need to explain how the model produces the results and how to interpret the results to our stakeholder, customers, legislations, and more.

In the insurance industry, while XAI could be the potential solution when introducing more advanced black-box models, there is a challenge in grounding high-level standards of sufficiency and appropriateness for explanations into low-level application context. Given a specific use case, the need of end-users should be important criteria when evaluating whether the explanations are sufficient and appropriate. Existing research outcomes on evaluating XAI explanations either stay on discussing high-level evaluation or focus on a niche application scenario. There is a gap where a universal methodology is not available to develop an evaluation of the quality of the explanation when faced with a new application scenario. Without a structured user-based evaluation framework, using XAI techniques in actuarial work embeds the risk of misusing XAI methods, operational difficulties due to inadequate explanations, and regulatory challenges when using a black-box model system to make decisions.

Our objective is to develop rigorous evaluation frameworks for XAI methods that assess both the general properties of explanation quality and their practical utility, using insurance pricing as an initial case study while ensuring our findings are transferable to other high-stakes industries with similar needs for model interpretability and regulatory compliance. This will be achieved by conducting a comprehensive user need analysis to identify the specific requirements for XAI explanations in the insurance domain, and subsequently evaluating the explanations through a user-participating survey. A successful evaluating framework consists of defined goals and boundaries of the evaluation supported by in-context user need analysis, with quantitative metrics to evaluate whether the user need is satisfied. To frame our investigation into user needs for ML explanations in industry applications, we draw upon Endsley's 1995 Model of Situation Awareness (SA)[40]. This model, widely applied in human factors research, describes three hierarchical levels of situation awareness: perception, comprehension, and projection. We adapt this framework to the context of ML explanations, particularly focusing on how XAI tools can support different levels of user understanding in industry-level applications. This adaptation allows us to systematically analyse both the varying depths of user needs - from basic conceptual understanding to complex relationship comprehension - and to develop quantitative measures for evaluating explanation effectiveness. Within this structured framework, our aim is to answer two research questions.

- RQ1. How can we systematically determine different levels of user needs when explaining ML models in industry applications - from basic concept understanding to complex relationship comprehension?
- RQ2. How can we quantitatively evaluate whether the ML model explanations by XAI tools effectively meet these different levels of user understanding needs?

User needs may vary from basic understanding of key concepts and terminology related to machine learning models to a deeper level of understanding that enables reasoning about model behavior. These needs encompass comprehending how the model works, the input features it considers, and the general logic behind its predictions. In the context of insurance and financial services, the end users primarily consist of actuaries involved in insurance pricing and portfolio managers making pricing prediction-based decisions, each requiring different depths of model understanding: actuaries need detailed comprehension to develop and validate pricing models, while portfolio managers need sufficient understanding to make informed business decisions based on model outputs. While our proposed framework can accommodate other stakeholders such as internal auditors and regulatory workers, our primary focus is on addressing the needs of actuaries and portfolio managers. This focused approach allows us to develop and validate solutions that directly address the complex requirements of pricing model development and its practical application in portfolio management, while ensuring the framework remains adaptable for end users beyond our primary focus groups, such as regulatory workers and internal auditors.

By addressing different levels of needs, explanations can be more effective in building trust, facilitating decision-making, and promoting the responsible use of machine learning models. Tailoring explanations to specific user needs ensures that the information provided is relevant, actionable, and aligned with their goals and expectations. Different tasks that users need to perform may require different levels of explanation. Evaluating explanations level by level ensures that the provided information is suitable for the specific task at hand. Evaluating explanations at different levels, we can provide targeted feedback to improve the quality and effectiveness of explanations for specific user needs.

Considering the different levels of user needs and evaluating explanations accordingly, we ensure that explanations are meaningful, relevant, and effective in supporting users' understanding, decision-making, and trust in the model's outputs. To investigate these research questions, we will conduct a case study in the context of insurance pricing. By examining the different levels of user needs and evaluating the quality of explanations in this specific context, we aim to provide insights that are grounded in real-world challenges and requirements. The case study will involve engaging with domain experts to understand their specific needs and expectations regarding model explanations. Through this case study, we will explore how breaking down user needs and evaluating explanations level by level can contribute to building trust, and supporting informed decision-making in the insurance pricing domain. The findings from this study will contribute to the development of practical guidelines and evaluation metrics for delivering meaningful and effective explanations in machine learning applications within the insurance industry and beyond.

## 1.4 Research Aims and Research Objectives

The overall research aim is to evaluate the quality of machine learning explanations via a structured user-based framework within the landscape of real-life practices. In order to achieve this aim, we break it down into two studies. The first study aims to bridge the gap between the theoretical underpinnings of XAI and its practical implementation in real-life scenarios, specifically focussing on an actuarial application case of Motor Third Party Liability (MTPL) insurance using GLM.

- RO.1 The objective is to determine the user needs of explanations when applying ML models in



safety-sensitive industry.

- Specifically, in the context of insurance pricing case study, the objective is to determine the user needs of explanations when applying GLM based on the tasks of actuarial professionals in the context of MTPL insurance pricing, aligned with industrial guidance.

The first study lays the groundwork by examining the low-level user need of machine learning explanations in the context of actuarial pricing, focussing on GLM-based insurance scenarios. It establishes the foundation for the second study, which zooms in on the specific explanations and applies a comprehensive user-based evaluation in a insurance pricing scenario. With the case-based user need derived from the first study, we continue to perform a user-participating evaluation as an experiment, aiming to evaluate the effectiveness of explanations on whether the user need is satisfied.

- RO.2 The research objective is to evaluate the effectiveness of explanations for machine learning models by assessing the satisfaction of user needs at different levels.
- Specifically, in the context of the insurance pricing case study, the objective is to evaluate the quality of explanations provided by XAI techniques for GLM-based pricing model, by measuring the satisfaction of actuarial professionals' needs at different levels, using quantitative metrics derived from data collected through a questionnaire.

The user need analysis in the first study serves as a precursor, providing insight into the nuanced needs of actuarial professionals, while the second study further refines this understanding and extends it to evaluate the effectiveness of explanations by XAI techniques in enhancing user satisfaction and decision making. These two studies collectively contribute to answering the overarching research question by offering a holistic approach to evaluating XAI explanations within insurance pricing contexts. The first study, grounded in actuarial practices, establishes the foundation for understanding user needs in complex pricing scenarios. The second study, building upon this foundation, delves deeper into the specific methodology including a novel XAI technique called SHAP(SHapley Additive exPlanations) values, and leverages a user-participated questionnaire based on Endsley's Model of Situation Awareness. Together, they provide a comprehensive user-based evaluation framework that not only addresses the gap in XAI research but also ensures practical applicability by aligning with the cognitive processes of users in real-world decision-making contexts.

## 1.5 Significance of the Research

This study is of paramount importance within the dynamic landscape of XAI, particularly when viewed through the lens of actuarial applications. The intricate nature of the actuarial field, where risk assessment and financial decision-making play a pivotal role, necessitates a robust understanding of the underlying machine learning models. By scrutinising the case-by-case analysis of user needs (RQ1) within the actuarial domain, this study aims to tailor the XAI guidelines to the specificities of insurance and financial scenarios. The application of XAI in this context is crucial, as it not only provides transparency in complex risk models,

but also ensures that decision-makers in the actuarial field can comprehend and trust the outcomes of advanced machine learning algorithms.

Moreover, the development of metrics to evaluate explainability at varying levels of recognition (RQ2) gains added significance in the actuarial context. Insurance professionals and actuaries require nuanced interpretations of model outcomes based on specific tasks, such as predicting claim counts or assessing the impact of variables on financial risk. Our research aims to propose metrics that align with the diverse needs within the actuarial realm, offering a customised and comprehensive approach to evaluating the effectiveness of XAI models in this critical sector. We also include a Figure. 1.2 showing how evaluating XAI methods is connected to a bigger scope in the insurance pricing example.

In essence, this study not only contributes to the broader field of XAI, but also has direct implications for the actuarial profession. By providing a framework for case-specific analysis and task-dependent evaluation metrics, our research aims to enhance the interpretability of machine learning models in the actuarial context. This, in turn, can foster informed decision making, mitigate risks, and enhance the trustworthiness of AI-driven solutions in the complex landscape of insurance and financial assessments. As the actuarial field continues to evolve in the face of technological advancements, the insights generated from this study stand to play a pivotal role in shaping the responsible and effective integration of XAI methodologies within actuarial practices.

Furthermore, the study's results are anticipated to contribute not only to theoretical advancements but also to the practical implementation of XAI within actuarial frameworks. The case study is also the first empirical study in user-based evaluation for actuarial machine learning models with more than 40 industry professionals with even actuarial background participating. The implications of our research extend beyond academic discourse, offering tangible benefits to industry professionals seeking to harness the power of AI for more accurate risk assessments and strategic financial planning. In essence, this study serves as a catalyst for the convergence of cutting-edge AI methodologies with the intricate demands of the actuarial field, fostering a symbiotic relationship that propels both fields toward greater efficiency, transparency, and adaptive decision making.

## 1.6 Structure of the Thesis

We have one comprehensive literature review and two empirical studies to complete this research. In Chapter 1, the broader background of the study is introduced. Explaining the main research gap that we will focus on, research objectives and research questions have been illustrated. We also discussed the significance and limitations of the study. In Chapter 2, we provide a comprehensive literature review, spanning historical evolution, conceptual foundations, prevalent XAI techniques, emerging trends, industry context, and the evaluation of explanations provided by XAI techniques. A crucial emphasis is placed on a significant gap in the existing literature: the limited attention to user-based evaluations when applying XAI in insurance. This chapter highlights the centrality of user-centric evaluations, setting the stage for subsequent studies.

In Chapter 3, we have the first study that delves into a detailed analysis of user needs in the case of actuarial pricing. We employ the Actuarial Control Cycle (ACC) framework and the Goal-Directed Task

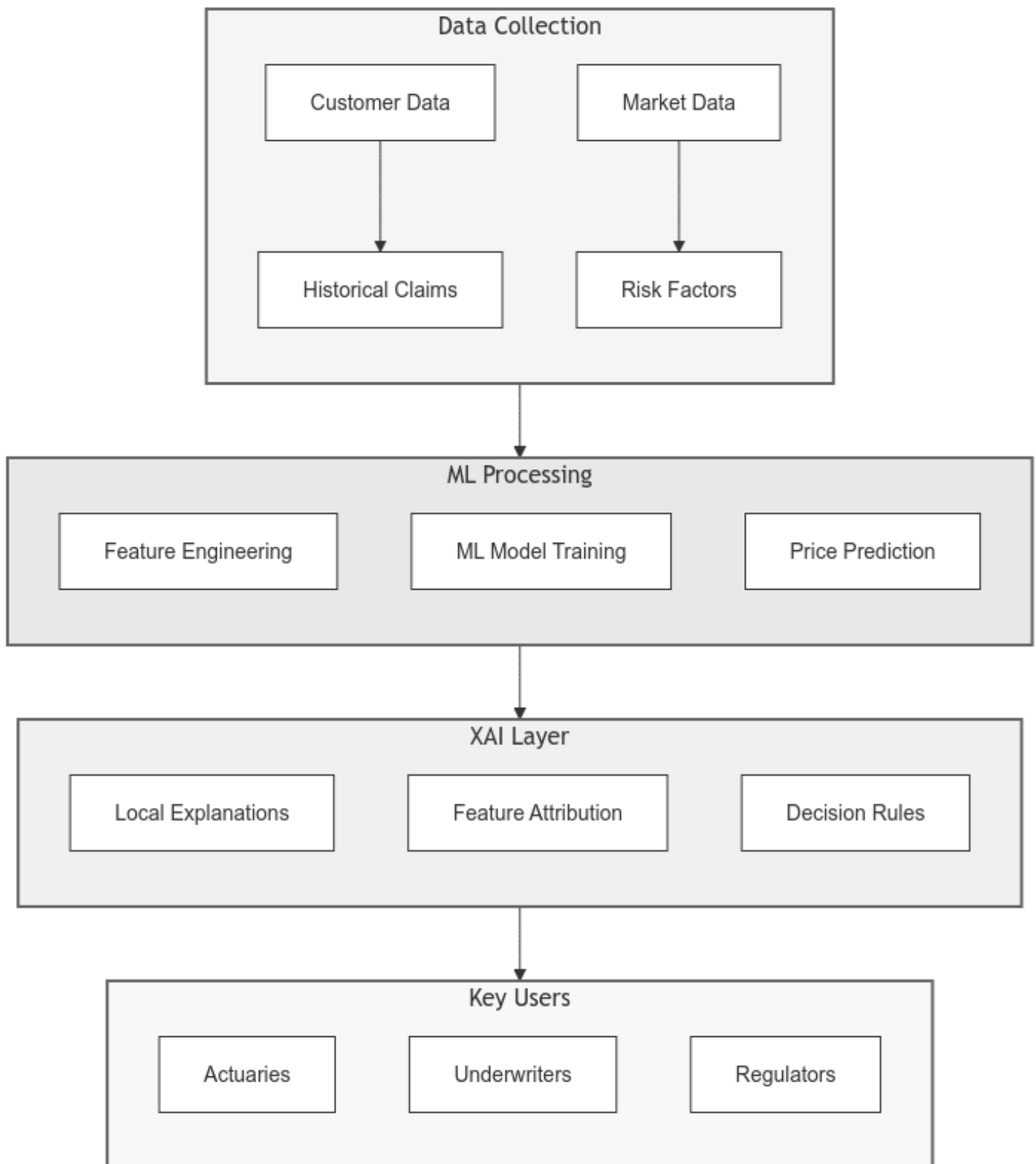


Figure 1.2: XAI Layer in Insurance Pricing

Analysis (GDTA) framework. Focussing specifically on insurance, the study addresses the central actuarial task of predicting claim counts for MTPL products using GLMs. This chapter establishes the foundation for understanding user requirements and informs subsequent evaluations in the user-centric context.

Next, in Chapter 4, building on the user needs analysis conducted in the first study, the second study aims to evaluate explanations generated by XAI, particularly those derived from SHAP values. Based on quantitative metrics derived from the SA of users, the study employs a user-participated questionnaire. Aligned with Endsley’s 1995 Model of Situation Awareness, the research provides a nuanced examination of XAI explanations, specifically in the context of insurance pricing.

Finally, we have Chapter 5 and Chapter 6 to synthesise the results of the two studies, emphasising their combined insights into user-centric perspectives in ACC framework insurance pricing. The discussions include implications for refining XAI techniques, improving transparency in informative decision-making, and improving user satisfaction. The chapter also addresses challenges encountered in the studies and future works.

Given the complex interplay between Chapters 3 and 4, which form the core empirical components of this research, we provide a table below to guide readers through our methodology and findings. This table roadmap illustrates the connections between our user needs analysis and subsequent evaluation of XAI explanations, ensuring clarity in following our research progression.

Table 1.2: Roadmap: User-Centric XAI Evaluation in Insurance Pricing

Chapter	Focus	Key Components
2	Literature Review	Research Gap Identification: User-Based Evaluation in Insurance
3	Study 1: User Needs Analysis	ACC Framework GDTA Framework User Need in Pricing with GLM
4	Study 2: XAI Evaluation	SHAP Values and alternative XAI methods Situation Awareness Metrics User-Participated Questionnaire
5	Synthesis	Discussion and Conclusions

## LITERATURE REVIEW

## 2.1 Introduction

EXplainable Artificial Intelligence (XAI) has witnessed remarkable growth and evolution, carving its niche in various domains. This review of the literature embarks on a comprehensive exploration of XAI, spanning from its historical evolution to current trends and applications, delving into its conceptual foundations and the techniques employed. The nuanced analysis unfolds across five key sections: XAI's historical trajectory, the conceptual landscape with its debates and agreements, a survey of prominent XAI techniques and emerging trends, an examination of the evaluation methodologies for XAI explanations, and a crucial focus on the application of XAI in non-life insurance pricing. Notably, our review sheds light on a significant gap in the existing body of knowledge: the limited attention to user-based evaluations when applying XAI in non-life insurance. The last section not only highlights this pivotal gap, but also emphasises its centrality, making it the focal point for our thesis. By addressing this critical lacuna, our study seeks to contribute valuable insights to the intersection of XAI and nonlife insurance, specifically emphasising user-centric evaluations as a key area for future exploration and refinement.

## 2.2 XAI From a Historical Perspective

In this section, we give a review on recent developments in machine learning methods and XAI, in the light of the responsibility of using the results of the machine learning model. We study how researchers claim that XAI could help to discharge the responsibility of applying the results of the machine learning model to reality. In recent years, the focus on model explainability has been developed both broadly and deeply, bringing XAI to a hot research area. Looking at XAI from a historical point of view is helpful in understanding the current position of this area and how it might develop in the future[28]. Thus, we looked back to the time when the work in explainability of a complex model began to be seen. Research papers started to put independent sections to illustrate the explainability of complex model when expert systems, as one of the most impactful complex models started to be developed. One example was XPLAIN[109]

which has been raised to demonstrate how an expert system can be produced with enhanced explanations of its behaviour in 1983. Taking this time as a starting point, we look at how explainability becomes an unavoidable topic when new machine learning methods are evolving and how XAI techniques started to be developed to narrow the gap between complicated machine learning algorithms and reality. Roberto et al. provided a historical overview of how explanations are conceptualised in different decision systems, including expert systems, machine learning, recommender systems, and neural-symbolic learning and reasoning (2019), so that a categorisation of explanations based on the reasoning characteristics of each underlying decision system can be reached[28]. We acknowledge the categories that make it possible to investigate explanation techniques from the top-down. We adopt a similar mindset, but we focus on the role that explanations played at different times.

### **2.2.1 Evolution of Machine Learning: From Competing with Humans to Ensuring Accountability in Complex Models**

In an earlier stage, the purpose of machine learning method was more to show to what extent machine learning can achieve a task compared to the capability of human beings and how much machine learning can compete with human beings in a given activity. One classic example is the Deep Blue chess machine that defeated world champion Garry Kasparov in 1997[22]. More recent applications include image recognition[61], automated translation[13], etc. Both the public interest and the academic interest focused on how the machine can be compatible with humans in a given scenario, rather than the reliability of the process that produces the results. Although the foundation of an academic result and strict logical deduction had to be demonstrated under the regulatory authority of the academic community, the explainability of the complex machine learning method has not drawn much attention at the early stage.

Starting around the mid-2000s, there was a resurgence in interest and progress in machine learning, and broad agreement among research has been achieved that the development of machine learning has shifted the world with massive applications across areas from finance to medicine. Meanwhile, the complexity of newly developed machine learning methods is exponentially increasing. At this stage, the role of machine learning methods is to assist in human decision-making or even to provide decisions for human beings independently. The impact of making wrong decisions based on machine learning output requires more attention, especially when it is related to finance or medical treatment. Applying complicated algorithms in safety-critical industries cannot be the same as using advanced ML methods to feed people the right advertisements. Some research pointed to the increasing requirements for a higher level of assurance provided for ML[6]because of the different role ML is playing and will be playing.

### **2.2.2 Rising Importance of Explainability in Machine Learning: Evolution, Challenges, and a Surge in XAI Research**

The significance of model outcomes intensified due to the direct impact on financial matters, human life, and health, since errors in suggestions are intricately entwined with legislative and judicial concerns of these consequential aspects. This requires the complex model to be explained so that users can make judgments about the reliability of the model. Compared to the time the expert system was developed,

new machine learning methods consist of evolving information that is more complicated than human-understandable common sense reasoning[28]. Influential machine learning methods developed over time were usually presented with specific explanation methods to demonstrate how the methods work and how the model outcome should be interpreted. According to Breiman (2001), the permutation feature importance measurement can be used to explain how much each feature contributes to the model results. This measurement was presented when random forest was presented as a new machine learning algorithm in the publication. When calculating the increase in the prediction error of the model after permuting the feature while keeping all other features unchanged, a feature of greater significance can be distinguished from a feature of smaller significance[20]. Measurement of the importance of the permutation feature not only explains to the audience how the algorithm works, but also provides a direction for parameter calibration for the model developer. There is also a separate effort from the community to focus on the explanation component. Fisher, Rudin and Dominici (2019) elaborated the permutation feature importance for random forest to a model-agnostic algorithm[47]. Another classical method to show the model dynamics of machine learning is the partial dependence plot (PDP), which shows the average of partial dependency between the predicted result and the input of the feature[92]. Goldstein et al.(2015) proposed using individual conditional expectation (ICE) plots to display one curve for each data point showing how the prediction of the instance changes when a feature changes. ICE plots refine PDP by graphing the functional relationship between the predicted response and the characteristic for individual observations[53].

When we go further to provide reliable suggestions based on machine learning techniques for stakeholders with a less technical background, we must face the challenge of the complexity of algorithms, which is technically difficult for humans to understand[82]. More explanation tools need to be developed while existing tools need to be better connected to end users. Until now, ML methods can be powerful in providing a correlation between the input and output features. However, the internal rationale behind the output of the model can hardly be presented[59]. Black-box methods became a popular name to emphasise the lack of internal logic and casual relations between input and output in ML methods. Powerful black-box tools have not been utilised enough in safety-critical industries due to lack of explainability. In fields where complex machine learning methods are applied without comprehensive elucidation, persisting with potentially inadequate explanations generated by diverse techniques poses significant risks. Continuously moving forward without clear and comprehensive understanding is akin to navigating either blindly or with limited vision, both of which entail considerable uncertainty and potential pitfalls. This prompts the rapid growth of the research area of XAI. We noticed that there has been a wave of research interest in eXplainable Artificial Intelligence since 2020. Referring to the most frequently used search engine Google scholar, we used the key word "machine learning explainable" to return the count of search results by year. The result is summarised in Figure. 2.1. From the bar chart, we see a surge in research results that shows after the year 2019 with an upward trend of seeing more publications.

When there are a variety of new explanation methods produced under the umbrella of XAI, the standard to evaluate the quality of XAI explanations are also a parallel hot topic[122]. There is a common ground among academic peers that the overall purpose of developing XAI techniques is to provide more transparency and evidence of the trustworthiness of artificial intelligence methods. In[9], it has been stated that the work on explainability consists of developing explanation methods either to enhance the

Research Outcomes Search Result Count by Year

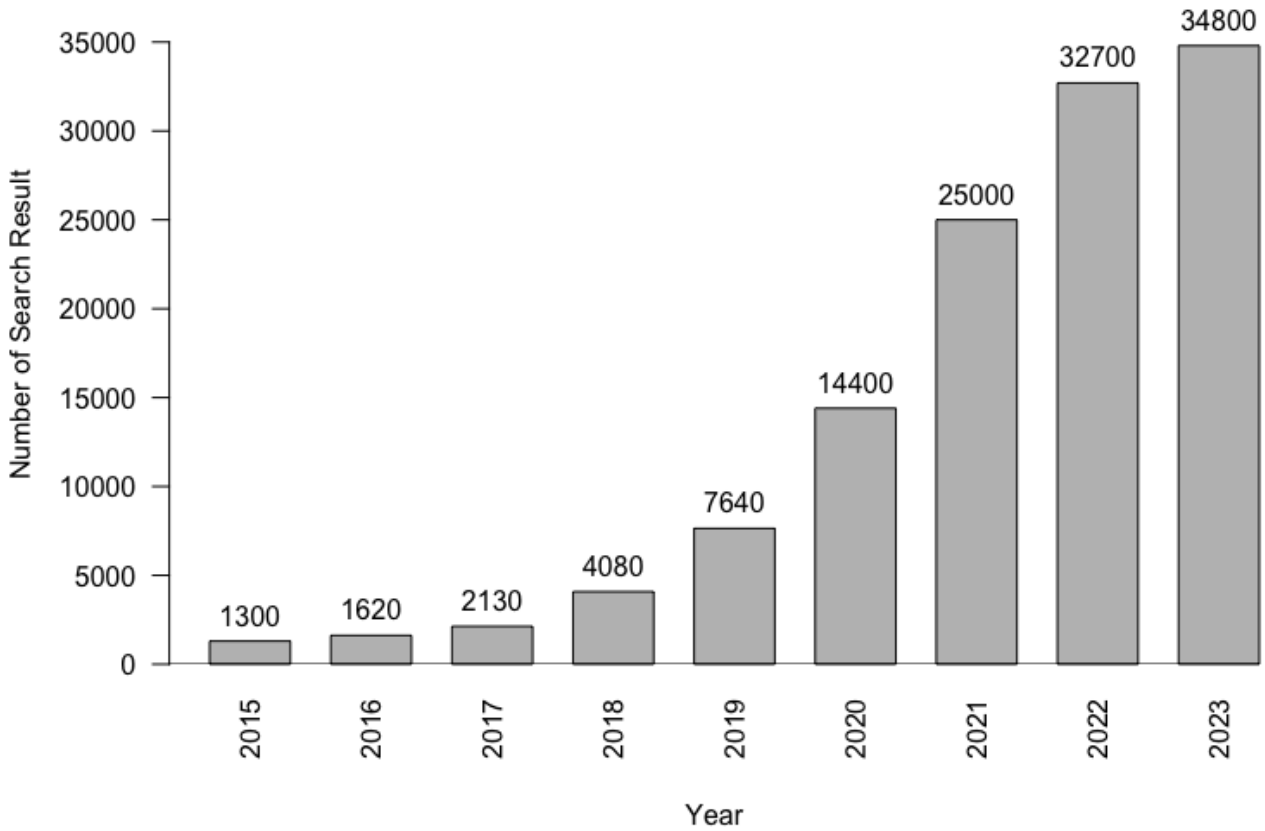


Figure 2.1: Research Outcomes Search Result Count by Year

Table 2.1: AI and XAI from a Historical View

Time	Achievements	Comments
1983 to Mid-2000s	Expert System with XPLAIN, Deep Blue chess machine, Image Recognition	Achieve human's capability, ML to compete with human beings.
Mid-2000s to 2019	Random Forests, Support Vector Machines (SVM), Neural networks and deep learning, Permutation Feature Importance	Machine Learning Boom, and arises of early explanation methods.
2020 to 2023	SHAP value, LIME, XAI	Advanced ML applied in safety-critical domains drives XAI prominence.

explainability of the model output or to enhance the explainability of the internal process of artificial intelligence methods. Recently, the expectation of good explanation methods has been raised to a higher level with an emphasis on casual explanations, in order to provide more reliable artificial intelligence algorithms in decision making[9]. At this stage, XAI starts to play a significant role in allowing users to make judgments on whether a complex machine learning method is appropriate to be used in reality. Thus the emphasis on connecting XAI to end users reached a higher level in more recent time.



### 2.2.3 Summary

This section of the literature review has delved into the historical underpinnings of the development of XAI, providing a comprehensive panorama of the evolution and key milestones in the field. The journey through historical perspectives has not only illuminated the origins of the concept of explainability, but also showed the stages where XAI has become an independent research topic, which further shaped the trajectory of advanced machine learning applications with complexities. Examining the historical work has allowed us to trace the development of key concepts and methodologies, observing how they have matured and adapted over time. By focussing on the explanation and evaluation part of the work where new machine learning methods such as random forests have been proposed, we have gained valuable insights into existing explanation methods that form the bedrock of making complicated AI algorithm explainable. Moreover, a more prospective lens after looking into history has facilitated an appreciation for the challenge of explainability for ensemble machine learning methods or methods that evolved from the basics, revealing that the explainability of advanced machine learning applications with complexity should always be challenged; existing explanation methods should be refined; new explanation methods should be developed.

We know that XAI as a rising research topic has attracted increasing interest from the research community. We also provide a nuanced perspective on the responsibility of using the results of the machine learning model. From the latter stage of the historical view, we see that the demand for developing XAI comes from the nature of modern machine learning applications implemented in more safety sensitive industries such as medicine and finance. In these areas, the outcomes of the model gain more importance, since they directly impact financial matters, human life, and health, with errors in suggestions intricately linked to judicial concerns. Considering the consequential nature, it is unlikely for us to rely on a black-box method with low explainability, thus we require the area of XAI to develop in pairs of the surge in study of machine learning methods. In addition, we highlight that the current state of the research landscape requires more attention to evaluate the explanations produced by XAI.

In essence, this historical exploration serves as a compass, guiding us through the intellectual terrain of the research area of XAI and offering valuable lessons for the road ahead. We appreciate the rich heritage of the machine learning research community, which has helped us embark on our own scholarly endeavours in a more niche research focus in evaluating explanations.

## 2.3 The Concepts in XAI: Arguments and consensus

### 2.3.1 What is XAI

When the XAI was first encountered as a new research area, researchers all hope to be able to answer the question of what XAI is. Although a strict definition was hardly agreed on, the academic community made the effort to describe what this subject is doing. XAI was defined as Artificial Intelligence (AI) that provides an explanation to improve the robustness of the system and allows diagnostics to prevent bias, unfairness, and discrimination, so that trustworthiness increases in how and why decisions are made[28]. Academics also questioned what explanation method is. Although there is general agreement on what XAI is expected

to deliver, research has struggled to agree on a clear and strict definition of explainability[35].

Within the extensive literature on the subject, it is discerned that explainable artificial intelligence constitutes a research domain focused on the generation and advancement of novel explanation methods. It is not a domain that dives deeper into black-box machine learning methodologies, despite the fact that intrinsic XAI methods primarily explore model dynamics and offer explanations by imposing constraints on the original machine learning methods[2][84].

We hope to set a boundary for the research topic on the basis of the discussions on the definition of XAI and XAI methods. In this thesis, the XAI method is paired with a machine learning method. With a clear target to explain either the result or the dynamics of the paired machine learning model, the XAI method is a method with solid theory support and mathematical foundation, explaining how the model results have been produced by the machine learning method or how the model results should be interpreted in the context of application scenario.

### **2.3.2 The Aim of Research in XAI**

What is a sufficient explanation produced by XAI? The general purpose of developing XAI is to provide more transparency and evidence of the trustworthiness of artificial intelligence methods. In[9], it has been stated that the work in explainability consists of developing explanation methods either to improve the explainability of the model output or to enhance the explainability of the internal process of artificial intelligence methods. Recently, the expectation of good explanation methods has been raised to a higher level with an emphasis on casual explanations, in order to provide more reliable artificial intelligence algorithms in decision making[9].

Taking a step back to our need for explanations in a general context rather than the XAI context, it is worthwhile to look at the concept from a social science perspective. Human users are highly selective in both the explanations we seek and those we accept as explanatory[73]. There is research focused on the interaction between explanations in the social sciences area (including psychology, philosophy, and cognitive science, etc.) and in AI.

Using a black box as a metaphor for the ML method, we have a question to answer regarding explainability: How does the black box give the output with the input? The general methodology of explanation methods is to make sense of the output given by machine learning methods by providing a relation between the input and output of the Black Box. The way to give the relation can be to approximate the behaviour of the Black Box or other ways. For example, it have been agreed on one way to classify explanation methods into global methods, local methods, and introspective methods. Global explainability aims to give an approximation of the black-box performance such that the overall output can be tracked by an interpretable dynamic. Consequently, local explainability points to a specific subset of the whole outcome. The local approximation is expected to be given by the local explanation method. Introspective methods use what-if questions to link the input and output.

Other classification of explanation methods can be seen as well. Vaishak Belle and Ioannis Papantonis incorporated a category of explainability[10] that includes simplification explanations, explanations of relevance of features, local explanations and visual explanations. Examples of local explanations include Anchors, LIME<sup>®</sup> (Local Interpretable Model-Agnostic Explanations) and Counterfactual instances. Ex-

amples of visual explanations include ICE and POP. The general methodology proposed above still fits in different method classifications. The shared core principle is to explain the relationship between the input and output of the Black Box.

In summary, under different classification of XAI methods, XAI method is developed to explain how the model results have been produced by the machine learning method or how the model results should be interpreted in the context of the application scenario.

## 2.4 Evaluation of Explanations Provided by XAI

### 2.4.1 Exploring Connections Between Machine Learning Methods, Black-Box Explanations, and Evaluation Frameworks in XAI

It has been widely agreed in the literature that evaluation of explanations of the advanced Black-Box method is necessary. So scholars varying from the traditional machine learning area to the newer explainable AI area all have great interest in the connections among machine learning methods, explanations of advanced Black-Box ML methods, and evaluations of the explanations. In [115], a pyramid graph is built in [115]. From the bottom layer to the top layer, we have data of different data types, models include ML methods, XAI methods with explainable process, explanation methods to perform the explaining, and evaluations of explainability provided. This shows us a top-down structure. We have Machine Learning methods requiring explanations first, then explanation methods are developed, and next we need to evaluate the quality of the explanation. This is a standard mindset in most of the survey papers in this area.

As the strict definition of explainability in XAI area is still an open research question under discussion [35], it can be imagined that the first challenge is to understand what is expected from the evaluation of explainability.

There are two typical types of research paper regarding the evaluation of explanations:

- For the first type, providing explanations for advanced Black Box method is the research objective. However, why the provided explainability techniques are reliable is usually treated as out of scope but meanwhile stressed as essential and worthy of great concern. In [85], SHAP values were used to explain the performance difference between Machine Learning methods and the traditional Cox Proportional Hazard (CPH) method. The authors trusted the SHAP values based on the "solid theoretical background" behind the technique rather than evaluating them.
- Second, the evaluation method itself is the research objective, and the purpose is to propose evaluation tools for the Explainable AI community. Three evaluation metrics have been developed in [72] to quantify the explainability of Explainable AI methods interacting with ground truth triggers. In this type of research, a number of Explainable AI methods are usually compared based on proposed evaluation metrics to arrive at a conclusion on the quality of explanations provided.

It could be seen that the second type of research is premise of the first type when evaluating and theoretical validation support are required before applying the intended explanation method. Not only the communication between these two types of researches but also the conversation between method providers

and the method end users is of great concern. What if the conversation happened after the explanation has been produced?

Some literature constructed the solution to the required explanation as a further processed abstraction under an explanation logic after the explainability model has produced an interpretable predictor[59]. The purpose is to modify the technical result for the end users with different experience background and different interest.

### **2.4.2 Effort in Evaluation of Explanations Towards User Need**

As the effect of explanations for ML is highly dependent on the specific need of end users, it is naturally difficult to come up with generic evaluation criteria. For this reason, the evaluation of explainability in research usually only accounts for the partial work up until the constructs of the explainer, and did not provide technical evaluation for the effect of the explanations. For example, some article shows clear performance metrics for ML modeling, while evaluating metric for explainability tools is neglected[86].

In the meanwhile, modifying process according to the evaluation of explanations should have served as last step before the explainability results consisting of all developing information flow to the end user, the impact of improper processing could be significant. In this research, we argue that a complete evaluation framework should be proposed including a solution to address the gap between end users and the technical process.

So far, we have identified two type of evaluation needs.

- To evaluate the explanation method in order to provide theoretical support for the community to comfortably use the explanations. This will require more technical methodology validation[79].
- To evaluate the effectiveness of explanations so that end users would know how to make use of the explanations as well as the Black Box method being explained[43].

It is fair to argue that proper evaluation for the quality of explanations should not be skipped. We surveyed literature to understand more reasons why researchers think we need to put more work in developing evaluation methods.

- To make a choice

In some cases, there are more than one existing explanation methods available to perform the same explanation. Same explanations here mean explanations under same principals or for the same purpose. Evaluation of explainability is then required to choose the best fit of method. For example, in order to perform a local explanation, Anchors and LIME are both potential fit for the purpose[122].

- To match the blooms of the explanation methods development

It is hard to find consensus in research on what the proper evaluation of explainability is for benchmarking purpose, while countless notions and increasing number of explanation methods have been thriving[122][35][115]. Also, there is a lack of rigor in the existing evaluations. Subjectivity and objectiveness in the evaluations is another essential concern. To summarise, the development of

evaluation methodology is lagged comparing to the blooms of the explainability research which is sparked by XAI.

- To serve further development of explanation methods

The purpose of the creation of new explanation methods or the improvement of existing explanation methods is to improve the ability to explain ML methods. Evaluation of explanations could be a feedback for further development in explanation method itself[10]. If considering prediction as a task of ML method, with a minimized cost of reducing prediction accuracy, explanations are desired to be developed to improve explainability of the prediction model.

The evaluation of explanations by XAI encompasses various perspectives, including subjective and objective evaluations, quantitative and qualitative assessments, as well as human-participating and non-human-participating approaches. Subjective evaluations, rooted in personal experiences and opinions, must strike a balance with objective assessments that rely on factual evidence. Quantitative evaluations involve measuring quantities, providing both specific metrics and supporting broader qualitative arguments. The distinction between human-participating, which involves subjective human experiments, and non-human-participating evaluations, derived from objective formal definitions, adds another layer to the multifaceted evaluation process. While non-human-participating evaluations tend to be more objective, human-participating evaluations, such as surveys, can incorporate high levels of objectiveness by gathering factual data on individuals' comprehension of explanations. Achieving a comprehensive and balanced evaluation involves navigating these diverse perspectives, considering both technical reliability and human relevance in the assessment of explainability methods.

- Subjective Evaluation and Objective Evaluation

The word subjective in dictionary is relating to the way a person experiences things in his or her own mind, based on feelings or opinions rather than facts. The word objective is relating to the way based on facts rather than feelings or opinions[34].

The possible challenge of obtaining a proper evaluation for explainability is to find a balance between suggestiveness and objectiveness, because the explanation methods should be both technically reliable and related to people who require it.

- Quantitative Evaluation and Qualitative Evaluation

The word quantitative in dictionary is relating to, or involving the measurement of quantity or amount. The word qualitative in dictionary is relating to, or involving quality or kind, a high level of value or excellence, how good or bad something is[34].

Specific quantitative evaluation can be both support of a broader qualitative evaluation argument and an independent metric to evaluate a specific property of explainability.

- Human-participating and Non-human-participating

Human-participating evaluation is evaluation involving conduction of human experiments[35]. This kind of evaluation can be time consuming. The choose of participators is subjective, how human beings experience the evaluation personally can be based on feelings and/or opinions thus subjective.

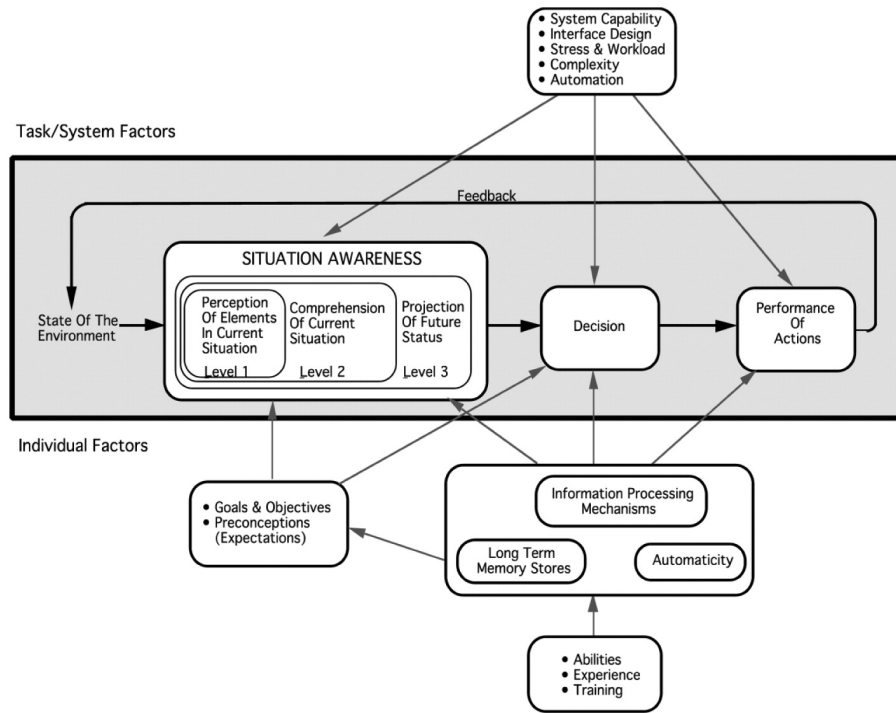


Figure 2.2: Model of SA in dynamic decision making[41]

Non-human-participating evaluation is evaluation requires no human experiments[35]. This type of evaluation is derived from formal definition within the explanation method that can be used as proxy. As long as the formal definition chosen is objective, non-human-participating evaluation is more likely to be objective.

However, human-participating evaluation can include high-level of objectiveness as well. Surveying real humans to obtain the quality of explanations is a natural and convincing way, and this is why there are a lot of human-centered evaluation being accepted. By nature, an explanation with decent quality means human beings can understand it correctly and clearly. For each survey point, whether the surveyed individual comprehends the explanation properly or not is a fact. Collecting the facts and summarizing evaluation based on facts is objective.

The ultimate goal of explanations provided by XAI is to satisfy needs of human users with a specific task that user needs to accomplish. There is an agreement among literature that human understanding should be enhanced by XAI[35]. In order to determine what the informational needs are for humans to perform in a specific scenario, researchers developed a variety of methodologies. Study of features of informational needs also open a gateway for objective measurements of whether the needs are satisfied.

There is an ongoing application in scenarios with complexity where humans have informational needs to perform a task [103]. It has been reviewed in a series of surveys [44][16] that objectively measuring SA(Situation Awareness) is important to evaluate whether users understand the information acquired in a variety of application scenarios including air traffic control, emergency management, healthcare, etc.[43][107][42]. Figure. 3.4 shows an example of using SA to evaluate and improve user satisfaction.

### 2.4.3 Situation Awareness as Evaluation Metrics in Dynamic Decision Making

SA(Situation Awareness) involves the perception of relevant elements within a defined spatial and temporal context, the comprehension of their significance and implications, and the projection of their potential state in the near future. This process enables individuals to develop a comprehensive understanding of their environment, interpret the information collected to form a coherent mental picture, and anticipate future changes or developments based on the current situation, ultimately facilitating informed decision-making and appropriate responses[41][89]. According to Patrick and Morgan(2010), the concept of Situation Awareness and models started to appear in the Google Scholar search from 1988, and there was a surge of research results from the year 1995[95]. In 2008, Wickens reviewed the progress in using Situation Awareness in measuring in different application areas[116]. In 2013, Endsley and Jones reviewed work using Situation Awareness as a basis for providing guidelines and designing systems, showing a gateway from high-level measurement to detailed metrics in specific mechanisms[106]. Among the research conducted in the past three decades, the Endsley 1995 model[39] has been highly cited[43]. We believe that there are two components in the Endsley 1995 model that are useful in our research in terms of proposing case-specific evaluating metrics rather than high-level evaluation. First, Endsley emphasised the role of goals and proposed a goal-directed task analysis method, which can be a framework for us to analyse user's need in the given application scenario. Second, the model clearly defined three levels of Situation Awareness(perception, comprehension, and projection), and we found it potential to use this definition to categorise the user's need in a given application scenario.

Regarding existing metrics to measure SA in application scenario, there is a discussion on the advantages and disadvantages of metrics in terms of objectiveness and feasibility. If a metric requires users' active input or surveyors' input, the objectiveness might be affected. For example, recording user's response time and errors are considered as more objective because there is no input required by the operator. Also, while process-based measures are informative for gaining insight into how people develop SA, this information is indirect and can only be used to infer the quality and completeness of the resulting SA, thus describing the state of knowledge acquired by the individual about the situation. Two people may arrive at completely different understandings of a situation based on the same process. In addition, these techniques provide information to measure SA by understanding what information subjects focus on and how they process this information to form situational understanding and prediction. The information is not complete enough[44][45][101].

Overall, there is a preference on objective and direct measurements from the literature. Examples are SAGAT(Situation Awareness Global Assessment Technique) and SPAM(Situation Present Assessment Method). These two technologies are all query-based measurements. SAGAT provides queries at random times in the scene, providing an objective, unbiased assessment of SA, while SPAM and real-time probing also provide queries to assess SA; however, queries can be provided verbally in real-time, while the individual is performing his or her normal during the operation task. SAGAT scores are percentages expressing the level of correctness for each query, based on operationally relevant tolerance bands. In addition to response accuracy, the time to respond to each SA probe is collected as an indicator of information availability in SPAM[43][44][63][36].

There are three levels of SA according to literature[40][44][102][31]. We provide a brief introduction to

it as below.

**Perception (Level 1 SA):** The basic perception of the status, attributes, and dynamics of features in a specific scenario. Level 1 SA involves understanding of concepts, knowing definitions of terminologies, simple reasoning, and direct recognition of the features of the scenario including situational elements (objects, events, people, systems, environmental factors) and their current states (locations, conditions, modes, actions).

**Comprehension (Level 2 SA):** Level 2 SA involves combinatorial analysis of discrete Level 1 SA elements through processes such as pattern recognition, interpretation, and evaluation. Level 2 SA has higher demand for integrating information to analyse the impact upon the individual's task.

**Projection (Level 3 SA):** The third and highest level of SA involves the capability to project future status in a specific scenario. Level 3 SA is based on the awareness on dynamics of the elements in scenario and comprehension of the situation including the acquired awareness from Level 1 SA and Level 2 SA.

The definition of SA requirements is highly adaptable to different scenarios[108]. SA has been applied in multiple areas, so it can be a reliable reference if we hope to evaluate the effectiveness of explanations perceived by human users[88].

For example, L. Sanneman and J. A. Shah claimed that SA-based informational needs can guide the design of XAI systems, ensuring they provide the necessary information about AI behaviour that aligns with the perception, comprehension and projection requirements, ultimately supporting users in making informed decisions and taking appropriate actions.[103]. An adapted version of three levels of XAI from three levels of SA is proposed as part of the evaluation framework in the underlying research. It has been discussed that SAGAT measures the discrepancies between ground truth and the SA of the person being measured. In [72], it is claimed that XAI aims to provide human users with the relevant components of their situation awareness that specifically pertain to AI behaviour, allowing them to better understand and interact with AI systems. What the human user has already understood is a subset of the informational need for the user to perform the task. The indication of the need for explanation can come from[66].

## **2.5 XAI Application in Non-Life Insurance Pricing**

### **2.5.1 Background and Scope**

Non-life insurance is also referred to general insurance in Australia. Whole life insurance, terminal illness insurance, temporary or permanent disability insurance, and income protection insurance are under the umbrella of life insurance in Australia. As the name suggests, the concept of non-life insurance covers the range of insurance products which are not life insurance. Health insurance is part of non-life insurance. Non-life insurance is a comprehensive category that encompasses two primary business lines: the commercial line and the personal line.

Commercial insurance is specifically designed to mitigate the risks inherent in business operations and corporate entities. For example, a software development company might secure a commercial insurance policy to protect against potential liabilities arising from data breaches or intellectual property disputes. However, consumer insurance, which is called a personal line, meets the personal insurance needs of individuals, protecting their assets, and providing financial security in various aspects of their lives[60][65].



The key difference between commercial and domestic insurance lies in the nature of the risks they address. While commercial insurance focusses on the complex and often large-scale risks faced by businesses and corporations, domestic insurance aims to provide personalised coverage for an individual's unique needs and circumstances. Both types of insurance play crucial roles in the non-life insurance sector, offering tailored solutions to mitigate financial losses and promote stability for businesses and individuals alike[87][5].

Since the machine learning boost from the mid-2000s, it has been an industry norm to use GLMs for non-life insurance pricing. Naturally, more frequent applications of advanced machine learning in the pricing of this area are GAMs, GBMs or neural networks. Insurers already have access to a variety of policy and claims data which can be used for advanced machine learning methods for predictive pricing. Other less structured data sources including images, geographic data, textual data, and medical evidence are also valuable training sets after appropriate data transformation[14][119].

In this section, we continue to perform a more systematic review on all peer-reviewed XAI applications in non-life insurance pricing with an extra focus on the evaluation of explainability. Eling et al. (2021) performed research assessing the impact of AI on insurance work including product development, underwriting and pricing, contract administration and customer services, claims management, etc.[37] The focus of this review was insurability of risks. Emer et al. (2022) expanded the research to assess XAI methods in AI applications with a similar scope[93]. It has been consistently mentioned in both works that areas within insurance where AI has been applied include fraud detection, claims reserving, risk assessment, etc. These two research works adopted similar inclusion and exclusion criteria.

Emer et al. (2022) classified XAI techniques into feature interaction and importance, attention mechanism, data dimensionality reduction, knowledge distillation, and rule extraction, and intrinsically interpretable models. We agree that this classification successfully avoids overlap between different XAI methods. We are specifically interested in the interaction and importance of the features among all for three reasons. First, as shown in[93], around 27 percent of the reviewed articles used techniques that can be classified into feature interaction and importance, which is the most popular among five of the categories. Second, how the feature contributes to the model results can be used in a variety of real-life scenarios including explaining the pricing principle for different stakeholders, feature selection, pricing optimising, etc. Third, it can be visualised and verbalised to be well connected to end users with different levels of technical background. Among the different niche areas in XAI applications in insurance, we zoomed in the work of pricing in non-life insurance to build our systematic review based on the work of the above-mentioned work of the two highly-referable research, with a specific focus on the evaluations of explanations produced by XAI techniques. To align with the XAI boost trend since late 2019, we review all related work between 2020 and 2023. We did not aim to update immediately for publications in 2024 as the estimated submission for this thesis is in January 2024, but a few valuable newly published results are mentioned in the review.

## 2.5.2 Fundamental Insurance Concepts and Industrial Context

Niche areas in insurance where XAI could potentially play a role include product development, claims management, underwriting, actuarial pricing, and administration in providing services. We reviewed

published articles and research results for an expanded illustration of those areas of work, for the purpose of sharing insights for researchers with little insurance background, and providing industrial context for this research. In the end of the section, we present a table of relevant insurance concepts with a brief definition and examples.

*Insurance product development, Claims management, Fraud detection, Underwriting, and Actuarial pricing* Development of insurance products includes five key activities: (1) Planning of product changes, (2) Idea exploration (3) Screening and evaluation, (4) Physical development, (5) Launch the product[62]. These five activities were structured in 1993, which is still valid for modern insurance companies. However, the method of physical development and launch may involve operational information technology systems and administration systems in the decades, where natural language processing (NLP) or other data mining methods might be relevant. Claims management in insurance involves the process of handling and overseeing insurance claims from the initial report to the final resolution. It includes activities such as assessing the validity of claims, determining coverage, and coordinating settlement or payment. Effective claims management is crucial for insurers to ensure timely and fair resolution while minimising fraud and operational defects. Fraud detection is an important area for claims management, which has a significant impact on the insured risk. Both the administrative work in claims handling and fraud claims detection may be benefited by applying machine learning and XAI techniques.[114][1][80] Insurance underwriting is the process of evaluating and assessing the risk associated with the insured (a person, a property or an entity). It involves analysing various factors of the insured to investigate the potential for loss or damage. The goal of underwriting is to determine the appropriate premium and terms for coverage based on the perceived level of risk. Underwriting professionals usually work closely with the pricing team[78][29]. Actuarial pricing in insurance involves using statistical and mathematical models to assess risk and set appropriate premium rates for insurance policies. Consider factors such as demographic data, past claims data, and economic trends to determine the likelihood of future losses. In life insurance, pricing often focusses on mortality and longevity risks, while in general insurance, it encompasses a broader range of risks such as property damage, liability, and other nonlife contingencies. In general insurance, underwriting and pricing are closely related processes that work in tandem. Underwriting involves assessing the risk associated with insuring a particular individual, property, or entity. The information collected during the underwriting is then used to determine the appropriate price of the insurance policy. Essentially, underwriting provides the data and analysis necessary for pricing decisions, ensuring that the premiums set adequately reflect the level of risk associated with the coverage. The goal is to establish a balanced and competitive pricing structure that allows the insurer to cover potential losses while remaining attractive to customers[112][97].

*Insurance premiums, Profitability, Loss and Loss ratio* The insurance premium, within the context of developing an AI system for premium calculation, constitutes the financial contribution paid by an individual or business to an insurance company in exchange for coverage against specific risks. It serves as the quantitative representation of the cost associated with the insurance policy and is intricately tied to the insurer's ability to manage risks effectively. In the intricate realm of non-life insurance premiums, the financial considerations are shaped by a myriad of factors. These include the characteristics of the insured entity, such as the type and condition of property in property insurance or the driving history in motor

insurance. The interplay of data, analysed using advanced machine learning algorithms, unveils nuanced risk profiles, influencing the calculation of premiums and ensuring the cost of coverage considering the portfolio of potential expenditure for an insurance company[64][30][91].

The profitability of a non-life insurance company hinges on a delicate equilibrium between accurately assessing risk through premium calculation and efficiently managing claims. An intriguing aspect of the insurance business lies in the fact that premiums are earned when future claims remain unknown. This uncertainty is two-fold, stemming from the uncertainty of both the timing and the total cost of future claims. The crux of the profitability of a non-life insurance company resides in its ability, similar to machine learning predictions, to accurately estimate these impending claims[30]. At the same time, the expenses associated with claim handling are intricately linked to factors such as the volume, timing, and complexity of the claims, all of which are inherently uncertain. The performance of the pricing algorithm is profoundly influenced by the quality of the claim data which is a product of the claims management[58][110]. In navigating this intricate landscape of uncertainties, the synergy of advanced technologies and strategic management becomes paramount for sustained profitability in the non-life insurance sector.

To discuss the maintenance of a target profitability, the concept of loss is highly referred to, because it is linked to the establishment of actuarial pricing and the calculation of the loss ratio. The premiums collected serve as a financial reservoir to cover potential losses, representing the anticipated cost of future claims. As mentioned above, the challenge lies in accurately predicting these future losses, which encompass both the timing and the magnitude of claims. In actuarial terms, the loss ratio, calculated as the ratio of incurred losses to earned premiums, serves as a critical metric. A proficient estimate of expected losses, coupled with efficient claims management, is imperative to maintain a favourable loss ratio, thus aiming to achieve the target profitability[30][121][118].

#### *Fairness in actuarial pricing and Regulatory environment in Australia*

Fairness in actuarial pricing refers to the just and equitable treatment of insurance policyholders in the actuarial pricing process. It involves ensuring that the premiums charged accurately reflect the risk assessed and that pricing practices do not result in unjust discrimination or disparate treatment among different demographic groups[51]. For example, using GLMs in private motor insurance pricing analyses various variables that impact claim outcomes, such as driver age, vehicle type, and historical claims data. The goal of fair pricing is to create a nuanced predictive model that considers a multitude of factors to determine the risk profile of a policyholder, while different segments of customers would not be treated discrimination[90].

Fairness in actuarial pricing is highly regulated. The Australian Financial Complaints Authority (AFCA) plays a significant role in overseeing fairness in actuarial pricing by providing a dispute resolution mechanism for consumers. AFCA is an independent body that handles complaints and disputes between consumers and financial service providers, including insurance companies. AFCA does not directly establish actuarial pricing regulations, but provides a platform for consumers to voice concerns about unfair or discriminatory pricing practices. The regulatory oversight primarily falls under the purview of the Australian Prudential Regulation Authority (APRA) and the Australian Securities and Investments Commission (ASIC). Insurers are expected to demonstrate that their actuarial pricing models consider a broad range of relevant factors and avoid unjust or discriminatory practices. The Insurance Contracts

Act of 1984 and the Australian Securities and Investments Commission Act of 2001 are among the key legislation that outlines the legal framework for insurance contracts and pricing practices. These regulatory bodies provide guidelines and standards that insurers must adhere to when developing and implementing actuarial pricing models[100][111]. The principles of fairness, transparency, and nondiscrimination are stressed to safeguard the interests of consumers and maintain the integrity of the insurance market[57].

Furthermore, regulatory oversight often involves reviews and audits to ensure insurers' adherence to established regulations, particularly in instances where advanced machine learning methods are employed in pricing. In such cases, insurers must provide detailed explanations that elucidate how prices are determined, the factors considered, and the underlying logic. It is crucial to be able to demonstrate to regulatory authorities that pricing is equitable, aligning with regulatory requirements and ethical standards. The appropriate XAI techniques for achieving this goal enable the insurance company to adopt more complicated pricing algorithms[100].

### **2.5.3 Machine Learning in Non-life Insurance Pricing**

Actuarial pricing is usually performed by a team of actuarial professionals in an insurance company. For nonlife insurance products, actuarial professionals have access to data from a variety of sources, including underwriting policyholder information, historical loss data from claims management, additional risk details, including geographic information, studies of natural disasters in an insured area, and any other information influencing pricing[48]. Economic factors such as inflation and interest rates also affect the indexation of the claims value or the earned premium. There may be additional data from reinsurance. External environment changes such as regulatory changes, competitor behaviour, insurance market conditions can also be unstructured information but essential for pricing decisions.

Actuaries use historical loss data for claim prediction models. Machine learning algorithms can help actuaries predict the likelihood of future claims for different policyholders. Specifically, regression models are usually used to predict the severity of future claims and estimate the cost of claims, helping insurers understand the potential financial impact. Having information on policyholders, machine learning clustering algorithms can group policyholders into segments depending on their similarity, helping insurers with segmentation before pricing. Machine learning models can also help with ad hoc risk analysis associated with specific properties. For example, in property insurance, image recognition algorithms may be used to evaluate property conditions. Similarly, geographic data might be a meaningful input to machine learning models to identify patterns and trends related to natural perils such as floods, storms, bushfires, etc.[56].

Compared to traditional methods or classic GLMs, Christopher et al. pointed out that advanced machine learning methods can learn non-linear transformations and interactions between risk factors from the data without specifying them or knowing the pattern before hand. In non-life interference pricing, decision trees, random forests, neural networks, and support vector machines are the most widely used machine learning methods[14].

However, before entering a more sophisticated pricing methodology based on advanced machine learning methods, being able to explain the model dynamics and the model results is essential with the consideration of being responsible for satisfaction of customers, profitability of the business, and requirements of regulations. Furthermore, the optimal method to be used can vary among insurers, depending on the

Table 2.2: Insurance Concepts and Definitions

<b>Insurance Concept</b>	<b>Definition</b>	<b>Examples</b>
1. Insurance Product Development	The process of creating and designing new insurance policies or coverage options to meet evolving market needs.	Change the renewal frequency.
2. Claims Management	The systematic handling of insurance claims from the initial report to the final settlement, ensuring efficient and fair resolution.	Settle a claim with payment.
3. Fraud Detection	Utilising tools and techniques to identify and prevent fraudulent activities within the insurance system.	Identify fake claims and reject the claims.
4. Underwriting	Risk Assessment associated with the insurance of a person, property, or entity, determining the coverage terms and premiums.	Advise a renewal premium with coverage limits and deductible amounts.
5. Actuarial Pricing	The use of statistical models to evaluate risk factors and set appropriate premium rates for insurance policies.	Set motor insurance premium rates based on risk factors.
6. Insurance Premiums	The amount of money policyholders pay to an insurance company in exchange for coverage against specific risks.	The payment a policyholder makes to buy a motor insurance.
7. Profitability	The measure of an insurance company's financial success, balancing premium income and claims expenses.	Unable to pay the claims as they fall due means poor profitability.
8. Loss and Loss Ratio	Loss represents the financial setback due to claims, and the loss ratio is the ratio of incurred losses to earned premiums.	Loss: Total claim cost of a product during the policy period. Loss ratio: Total claim cost divided by total earned premium.
9. Fairness in Actuarial Pricing	Ensuring just and equitable treatment of policyholders in the pricing process, avoiding discriminatory practices.	Pricing differently for female against male or the elders against the young may be discriminatory.
10. Regulatory Environment in Australia	The legal and regulatory framework that governs insurance practices in Australia, including standards for fairness and transparency.	APRA, ASIC, AFCA, The Insurance Contracts Act of 1984

XAI Methods	Actuarial Concepts
Variable importance	Non-life insurance pricing
Permutation-based variable importance	Property and Casualty (P&C) insurance pricing
SHAP value	General insurance pricing
Feature importance ranking measure	Predicting claims frequency
Surrogate model	Predicting claims severity
LIME	Predicting cost
Partial dependence and partial dependence plots	Predicting loss
Friedman H-statistic	

Table 2.3: Table. XAI Methods and Actuarial Concepts as Interchangeable Key Search Terms

business stage and risk exposures. Having a black-box machine learning method that is explainable can be a milestone for advanced technology to be rooted in the actuarial industry. In the next section, we will show our results of a systematic review of the published XAI applications using feature importance and interactions in non-life pricing from 2020 to 2023.

### 2.5.4 Literature Review Process of XAI Application in Non-life Insurance Pricing

In conducting the literature review for this thesis, an extensive exploration was performed using Google Scholar as the primary search engine, complemented by queries on analogous academic platforms such as Web of Sience, IEEE Xplore, and Scopus. The search criteria were meticulously defined to include research published within a time frame spanning from 2020 to 2023. The focus of the investigation was on the area of application of actuarial pricing, specifically within the non-life insurance sector, also known as Property and Casualty (P&C) insurance or general insurance. Within this niche, the review focused on the emerging field of XAI, with a specific emphasis on understanding feature interaction and importance. This meticulous approach ensured the retrieval of the most recent and pertinent literature, facilitating a comprehensive analysis of advancements and trends in XAI methodologies applied to feature interpretation within the dynamic landscape of actuarial pricing in the non-life insurance domain.

Table. 2.3 outlines the interchangeable keywords we put in the search engine. Having a set of search terms is a search method adapted from the literature review method used in [93] and [37].

Furthermore, by taking advantage of contemporary advances in publication databases, which offer tools to showcase interconnected papers and suggest related works, a meticulous approach was adopted to handle recommended papers. The initial step involved a comprehensive scan of the suggested list, from which relevant publications were manually selected. Subsequently, the same criteria were applied, taking into account the publication year, alignment with the research topic, relevance to the designated application area of non-life insurance, and alignment with the specified XAI stream focused on feature interaction and importance. In particular, a discerning strategy was implemented to exclude highly similar works that did not substantially contribute to the understanding of XAI methodologies in actuarial pricing. For example, works that apply identical XAI methods to data sets of lower quality or employ machine learning methods with inferior performance were excluded, ensuring that the literature review encompasses high-quality and insightful contributions to the field.

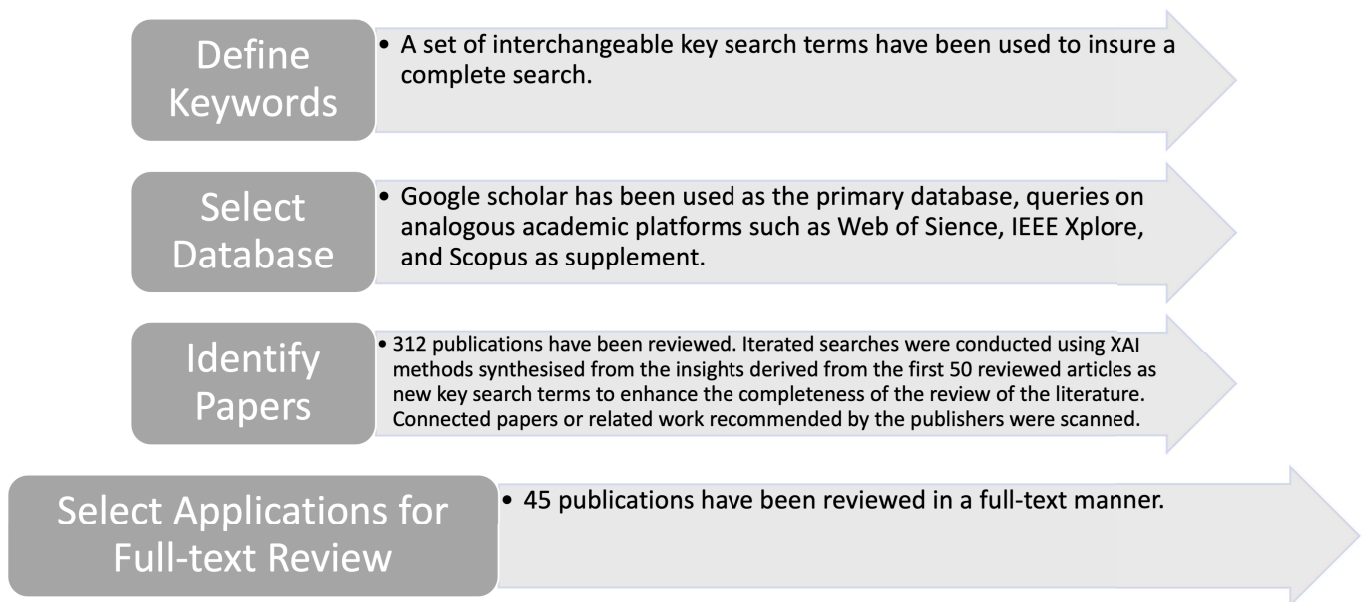


Figure 2.3: Flowchart Illustrating the Literature Review Methodology

In addition, a comprehensive examination of highly relevant survey-type publications was conducted to acknowledge existing arguments and conclusions on the broader topic of XAI in insurance. Despite an extensive review, it was noted that there is a conspicuous absence of a review paper specifically addressing the niche direction of feature interaction and importance within the context of actuarial pricing for non-life insurance. Consequently, in the absence of a dedicated literature review in this precise domain, insights from pertinent arguments proposed in existing survey-type papers were extrapolated and meticulously elaborated to form a foundational understanding of the niche direction under investigation. This strategic approach ensures that the literature review not only assimilates existing knowledge, but also actively contributes to advancing the discourse within the specialised realm of XAI methodologies applied to feature interpretation in actuarial pricing within the specified insurance business line.

Finally, a series of iterated searches was conducted to enhance the completeness of the review of the literature. XAI methods, synthesised from the insights derived from the first 50 reviewed articles, were used as keywords in subsequent Google Scholar inquiries. This iterative process aimed to uncover additional relevant studies and refine the understanding of the landscape surrounding feature interaction and importance in actuarial pricing within the specified time frame (2020-2023) and business domain of non-life insurance. To maintain rigour and relevance, high-level studies and qualitative discussions lacking empirical studies were excluded. This iterative search strategy served as a dynamic mechanism for capturing emerging perspectives and methodologies within the nuanced field of XAI applied to actuarial pricing.

In Figure. 2.3, the search methodology is elucidated using a flow chart, illustrating the meticulous process in which 312 articles were examined for relevance to our specific niche research area. As a result, we went through the 45 pertinent applications selected for a full-text review.

### **2.5.5 Systematic Review Results of XAI Application in Non-Life Insurance Pricing**

The systematically selected published research results are first reviewed with the focus of the XAI method used with the associated actuarial pricing task. We aim to find out which XAI tools have been chosen more frequently than others, in the context of the application scenario. Second, we reviewed the selected applications again to investigate how the atoms convince the audience that the explanations are sufficient and have met the need. We rephrase this step as investigating the evaluation of explanations produced by the XAI method in each study. Recall that our criterion for the line of insurance business is non-life insurance, and the XAI methods we are interested in are methods explaining Feature Interaction and Importance(FII).

Among the selection process, we have chosen 45 application studies for full text review. We list examples of literature in Table. 2.4 to show a slice of our work.



Table 2.4: Summary of XAI Methods Applied in Pricing Tasks

Citation & Published year	XAI method	Pricing task	End user(s)	Evaluation of explanations
[76], 2023	Variable importance, Permutation-based variable importance, SHAP value, Feature importance ranking measure	Various	Not specified	Quantitative: Alignment with expert intuition, consistency across different instances, Efficiency, Fairness and bias, Regulatory compliance
[81], 2021	Surrogate model	Predicting claim frequency, Measuring the importance of an extreme event, Identify profile prone to claim	Not specified	Not specified
[54], 2020	SHAP value	Predicting renewal behaviour	Insurers	Comparative study with other methods
[70], 2021	SHAP value	Monthly cost predictions for renewal business groups	Insurers	Not specified
[19], 2021	Contextualising SHAP value	Communicating feature importance in car insurance pricing	Non-expert users	Quiz-based evaluation
[100], 2021	Surrogate model	Predicting claim frequency for car insurance	Not specified	Not specified
[55], 2021	Permutation feature importance, Partial dependence and partial dependence plots, Friedman H-statistic for feature interaction strength, Accumulated Local Effect (ALE) plots	Explaining Combined Actuarial Neural Network(CANN)	Not specified	Not specified
[17], 2022	SHAP value, LIME	Predicting the cost of health insurance	Domain experts	Not specified

Table 2.4 continued

Citation & Published year	XAI method	Pricing task	End user(s)	Evaluation of explanations
[12], 2022	Various, not specified	Classify insurance claims (e.g., home or car damage) by NLP	Domain experts	Qualitative: Interviews and contextual inquiry

### *XAI Definitions and Problem Formulation*

The definition of what method should be included in the area of XAI and what does not is a complex and nuanced endeavour. The ambiguity arises from the fact that certain methods employed to explicate opaque or black-box algorithms predate the development of the XAI area. Techniques such as feature importance analysis and sensitivity analysis have long been used to shed light on the decision-making processes of machine learning models. Furthermore, an additional layer of complexity emerges when examining published research outcomes that, while underscoring the significance of explainability in the context of machine learning, may not explicitly reference the notions and concepts in the XAI framework. In our review, we extend the paper search to machine learning applications even if we did not see the term explainable AI or interpretable ML mentioned. We then scan the paper to find whether the work on showing explainability is done. This lack of terminological alignment does not diminish the crucial role that these methods play in enhancing the interpretability of the model. As such, the indistinct demarcation between traditional interpretability approaches and formally recognised XAI methodologies shows the evolving nature of the field and the imperative to acknowledge the multifaceted roots of explainability within artificial intelligence. Some review articles shared the argument that appropriate methods and explanation forms depend on the application itself and the context in which it is embedded[93][105][67].

A customised formulation to the case of the XAI method turned out to be more effective. For example, in [76], four of the explanation methods including variable importance, permutation-based variable importance, SHAP value, and feature importance ranking measure have been defined in a universal format before applying them to four different machine learning applications featuring different machine learning methods and actuarial pricing tasks. Evaluation metrics are proposed in a similar structure. In [54], the SHAP value as the XAI method explaining the importance of the features was illustrated with a summary graph using the use case dataset, in the context of the contribution of each variable to the propensity to renew the insurance policy. The following example embraces the definition format in context, a format that is more recommended.

### *XAI Methods and Application Scenarios*

Across the applications we reviewed, about 30% of the chosen XAI methods included surrogate modelling or the extraction of a surrogate model from the fitted machine learning model. In[81], an equivalent tree with a tail tree surrogate has been used as the extracted rule to explain tree-based models, including the regression tree and random forests. LIME is a prominent XAI method introduced by Ribeiro et al. in 2016[98]. As shown in Figure. 2.4, LIME generates localised explanations for individual predictions by approximating intricate global models with simpler, interpretable models. This approach facilitates human understanding by providing sets of interpretations that can be visually presented and analysed, promoting transparency, accountability, and trustworthiness in machine learning systems. LIME as a surrogate technique generates locally faithful and simplified local explanations around specific instances, allowing actuaries to understand the impact of input features on pricing predictions[32]. LIME has been mentioned frequently in papers that choose a surrogate model as an explanation. When several different XAI methods are applied for a comparative study, LIME is usually applied as part of the comparison. We recognise that LIME is one of the most popular surrogate techniques of XAI.

SHAP, proposed by Lundberg and Lee in 2017, is another influential XAI tool[77]. This method attributes

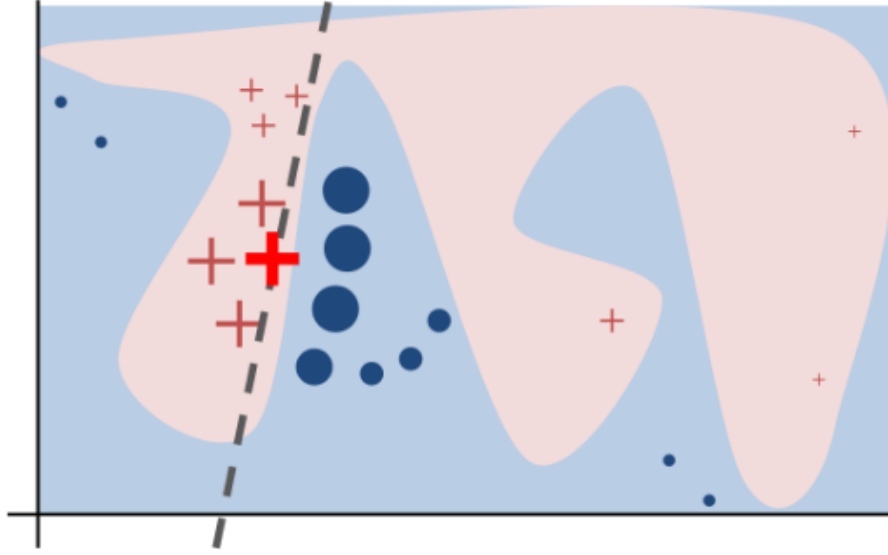


Figure 2.4: According to[98], this graph depicts a black box classification model ( $f$ ) delineated by pink and blue regions. The focal point of explanation is represented by a bold red cross, surrounded by locally sampled instances denoted by red crosses and blue circles, their proximity serving as weights. The locally faithful explanation ( $g$ ), portrayed as a dashed line, elucidates the model's behavior within the specified region.

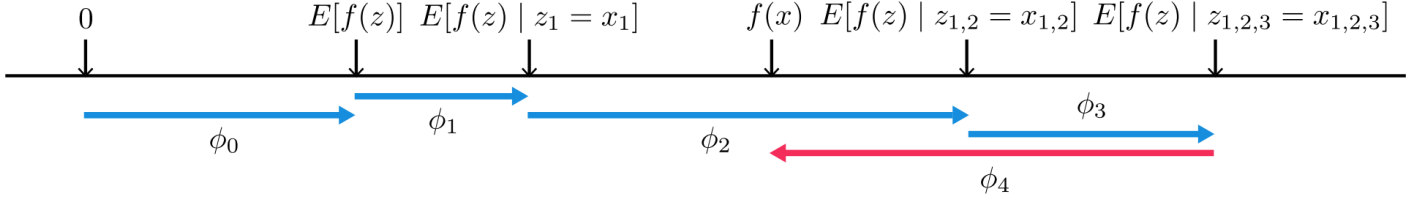


Figure 2.5: According to[77], SHAP values attributing to each feature the change in the expected model prediction when conditioning on that feature.

to each feature the change in the expected model prediction when conditioning on that feature. In Figure. 2.5 It is illustrated how we can get from the base value  $E[f(z)]$  that would be predicted if we did not know any features to the current output  $f(x)$ . This diagram shows a single ordering. However, when the model is non-linear or the input features are not independent, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the  $\phi_i$  values across all possible orderings.

Based on the principle shown in the graph, the SHAP values provide both local and global explanations for the model predictions. SHAP assigns values to features in input data based on the Shapley value from cooperative game theory. In particular, SHAP is versatile, accommodating various machine learning models and input data types, including tabular, image, and text data. Its significance lies in generating not only local explanations, but also a comprehensive global understanding of how different features interact and contribute to model predictions. In contexts like insurance pricing, where risk relativity is crucial on a global scale, SHAP's ability to provide a broader view aligns well with the diversified nature of the industry.

The solid theoretical foundation and proven applicability in fields analogous to actuarial modelling further enhance SHAP's appeal over LIME in certain scenarios.

As another highly referred XAI method, it has been used to compare with the traditional permutation-based variable importance method to rank risk factors in terms of the influence power on the predicted results. About 50% of the reviewed applications have adopted SHAP-value-based visualisation in their explanations. SHAP value usually shows results comparable to those of traditional explanation methods such as permutation-based variable importance, other feature importance ranking measure, and feature partial dependence[55][92].

Although the LIME and SHAP value are both popular XAI methods according to the articles we reviewed, LIME focuses on generating locally faithful and interpretable explanations by approximating the model's behaviour around a specific instance through the use of locally trained interpretable models. On the other hand, SHAP values provide a holistic and theoretically sound approach based on cooperative game theory, offering a consistent framework for explaining the contribution of each feature to a model's output across all instances. Although both methods have their merits, SHAP is considered superior because of its mathematical foundation, addressing some of the limitations associated with LIME, such as its sensitivity to the choice of perturbed instances. SHAP value also shows robustness, while it was claimed in some studies that close entries can lead to significantly different interpretations when using LIME. SHAP's ability to provide a consistent and globally meaningful interpretation of feature importance contributes to its widespread adoption and preference in various machine learning applications[32].

We also summarised actuarial application scenario from the pricing tasks that have used above-mentioned XAI tools. Approximately 40% of the reviewed articles are applying XAI techniques to predict the claim pattern so that the claim cost of the product can be estimated. This includes predicting the frequency and severity of claims, classifying types of claims, and identifying profiles prone to claims. Additionally, around 20% work focused on communicating feature importance in the pricing model to stakeholders in the insurance company. For example, various XAI techniques have been used to explain the functionality of advanced models such as CANN and the usage of natural language processing (NLP)[12][55]. We found that multiple XAI techniques are usually used together in the case where the study is user-orientated. Ad hoc analysis, such as analysis of policyholder's renewal behaviour and measurement of the significance of extreme events can also be benefited from the XAI techniques. Periodical work such as estimating the costs for a specific policy group monthly or quarterly needs the support of XAI techniques to reduce the time of manual process of generating explanations.

To our surprise, less than 10% of the studies that underwent full-text review clearly specified the target users. When the specifications are clear, most of users are domain experts, either actuarial professionals or professionals with both insurance background and technical background. Users were vaguely mentioned as insurers or insurance companies in half of the reviewed studies.

#### *Absence of an Evaluating Framework*

The absence of specified end users in the applications indicates a challenge in assessing the effectiveness of explanations generated by XAI methods. Without a clear identification of the intended audience for these explanations, it becomes difficult to measure their impact and relevance in real-world scenarios. Some studies use comparative results from different XAI methods to make the final decisions, which indirectly

evaluate the explanations by the consistency among different methods. However, consistency does not guarantee effectiveness.

Defining and involving end users in the evaluation process is significant to ensure that XAI explanations meet their intended purposes and provide meaningful insight. The application of XAI methods in the absence of a well-developed evaluation system poses multifaceted challenges to legislation and regulatory frameworks. The lack of clear evaluation criteria complicates the establishment of standardised guidelines, creating uncertainties in the treatment of accountability, fairness, and transparency issues. In the insurance market, this ambiguity affects risk assessment, pricing strategies, and claim processing, potentially disrupting market dynamics. Customers may struggle to understand and trust AI-driven decisions, which affects trust in the industry. Navigating these challenges requires a collaborative effort involving policymakers, regulators, and industry experts to ensure the responsible and ethical integration of XAI into the insurance sector.

In this context, user-based evaluation is of significant importance, but the landscape of user-based studies remains largely unexplored. Current studies have been limited by small samples, which limits the generalisability of the findings. Furthermore, the evaluation methods used primarily involve the extraction of simple interview scripts for qualitative analysis, underscoring the need for more comprehensive and diverse user-based research to gain a nuanced understanding of the effectiveness and impact of applying XAI methods in the insurance business line.

## 2.6 Conclusion

A comprehensive literature review that focuses on XAI applications in insurance has been conducted, with particular emphasis on identifying research gaps that will guide ongoing studies in insurance as an entry point for understanding XAI implementation in high-stakes industries.

The Actuarial Control Cycle (ACC) is a continuous and iterative process within actuarial science that encompasses problem identification, model solution development, monitoring, and model refinement. It serves as a dynamic framework for actuaries to adapt pricing and risk management work based on changing needs arising from emerging patterns, market conditions, and evolving risk profiles.

When XAI techniques are applied to explain factors influencing individual predictions, actuaries must adjust explanation methods or pricing strategies in response to end-user demands. This iterative process ensures that the ACC remains responsive to the dynamic needs of the insurance landscape, improving pricing accuracy and actuarial decision-making.

Through our review, we have identified several critical research gaps:

- The effectiveness of XAI explanations lacks comprehensive evaluation due to the absence of user-based studies. Current research has not adequately addressed how end-users interact with and understand these explanations in practical settings.
- There is insufficient integration of end-user demand and satisfaction metrics in evaluating XAI explanations' effectiveness, which is crucial for a robust monitoring step in the ACC process. Existing studies have primarily focused on technical metrics rather than user-centric evaluation measures.

- The field lacks a structured framework for aligning explainability approaches with specific user needs in non-life insurance pricing. While the importance of user-centric design is acknowledged, there is limited research on how to systematically assess and incorporate user requirements into XAI implementations.

The desiderata for explainability in non-life insurance pricing must stem from a thorough analysis of user needs. This involves a nuanced understanding of the application scenario and careful consideration of the target audience. By aligning explanations with user requirements, insurers can improve transparency, build trust, and facilitate meaningful engagement with stakeholders.

Our research suggests that Endsley's 1995 Situation Awareness Model can serve as a theoretical foundation for developing quantitative metrics to evaluate the effectiveness of SHAP value explanations in well-defined use cases. This approach provides a promising direction for bridging the identified gaps between technical implementation and practical utility in insurance pricing explainability.





## USER NEEDS ANALYSIS TOWARDS RESPONSIBLE XAI

### 3.1 Introduction

Artificial Intelligence (AI) and industries with high sensitivity to safety and ethics are converging. Those industries, including banking and finance, insurance, and medicine, require professionals to perform high-stakes decision making with potential consequences of errors or biases. This introduces a realm of possibilities and challenges. The increasing use of complex machine learning (ML) models in safety-sensitive decision-making requires explainability and transparency, leading to the adoption of eXplainable Artificial Intelligence (XAI) techniques. Although numerous XAI methods have been proposed to elucidate complex ML results, evaluating their quality remains a critical challenge, particularly when putting the explanations in specific use cases. Our research embarks on a nuanced exploration, focussing on the low-level evaluation of the explanatory power of XAI tools within the intricate realm of safety-sensitive industries. We chose insurance pricing as a case study.

Choosing insurance pricing as a case study for my research on evaluating XAI tools in safety-sensitive industries is a strategic decision that can yield valuable insights and practical applications. Insurance pricing is a complex and high-stakes domain that heavily relies on accurate risk assessment and decision-making, making it an ideal testbed for the proposed framework. By focusing on this specific use case, we can develop a tailored approach that addresses the unique challenges and requirements of the insurance industry, with industry-level professional decision making and context. We integrated industry view by introducing Actuarial Control Cycle(ACC) workflow as part of the user analysis framework. This is transferable when performing a similar user need analysis in other safety-sensitive areas where ML can play a role. Professional guidance of the chosen area should be studied so that an alternative for ACC in the chosen area can be introduced or developed.

Moreover, the lessons learned and best practices derived from this research in the insurance pricing context can be readily transferred and adapted to other safety-sensitive areas, such as banking, finance, healthcare, medicine, and any area where ML methods can play a role and safety is a significant concern. These industries share similar concerns regarding the explainability and trustworthiness of AI-driven

decisions, as well as the need for robust evaluation methodologies. My framework, which categorizes user needs into different levels of situation awareness, provides a structured and comprehensive approach to assessing the explanatory power of XAI tools. This framework can serve as a valuable resource for practitioners and researchers across various safety-sensitive domains, enabling them to better understand and evaluate the effectiveness of XAI techniques in their specific contexts. By demonstrating the transferability and generalisability of my research outcomes, I underscore the broader impact and significance of my work in advancing the responsible and transparent use of AI in critical decision-making processes.

When techniques associated with XAI have been proposed in numerous research outcomes to provide a solution to explain machine learning results, we face the challenge of evaluating the effectiveness of XAI techniques in the context of each use case. Our research interest is to evaluate the effectiveness of the explanations produced by the XAI tools on a low level of each use case. The research objective is to determine the desiderata of XAI explanations based on the need of industry professionals in a use case of GLM-based non-life insurance pricing. Ultimately, our objective is to identify a bridge capable of mitigating the disparity between the XAI theories and the practical implementation.

The central thrust of our investigation is to bridge the gap between the theoretical underpinnings of XAI and its practical implementation in insurance pricing settings. In the insurance pricing domain, actuarial professionals play a crucial role as industry practitioners. They are responsible for developing sophisticated models that assess risk and determine appropriate premiums for various insurance products. These models often involve complex statistical and mathematical techniques, such as generalised linear models (GLMs) and machine learning algorithms. Actuaries must carefully design, test, and validate these models to ensure their accuracy, fairness, and compliance with regulatory requirements.

Proposing the Actuarial Goal-Directed Task Analysis (A-GDTA) framework, we harmonise the ACC(Actuarial Control Cycle) framework and GDTA(Goal-Directed Task Analysis) framework for an in-context user needs analysis. In 1985, ACC framework was introduced by Jeremy Goford, who was a British actuary served as the President of the Institute of Actuaries from 1984 to 1986[52]. The GDTA work flow is a key component of Endsley's broader theory of situation awareness, which emphasises the importance of understanding the goals, perceptions, and comprehension of individuals in dynamic decision-making contexts. Mica Endsley is a human factors psychologist and former Chief Scientist of the U.S. Air Force. She first introduced the concept of Goal-Directed Task Analysis in her 1993 paper presenting GDTA as a method to identify critical information requirements to maintain situation awareness in dynamic environments[38], and elaborated further on the GDTA in her work in later years describing GDTA as a process to determine the goals, decisions, and information needs of operators in complex systems[40][41]. The GDTA framework, tailored by the ACC framework, to address industry-specific concerns and monitor practical experience, presents a promising approach for conducting systematic user needs analysis within the dynamic industry environment, effectively bridging the gap between theoretical concepts and real-world applications.

We explore the intricacies of user needs specific to GLMs based non-life insurance pricing, combining established actuarial methodologies with advanced XAI techniques. As XAI becomes more prominent in providing interpretable information, it is crucial to evaluate its efficacy. Our research focusses on evaluating explanations generated by XAI tools, with the aim of determining their effectiveness in addressing the nuances of each use case, specifically Motor Third-party Liability Insurance (MTPL) pricing using GLMs.

Our aim is to define the essential criteria for effective XAI explanations tailored to the needs of actuarial professionals. Our research seeks tangible results that resonate with the challenges faced by actuarial professionals, enhancing the transparency of decision-making and bridging the gap between XAI theories and their practical implementation. We contribute to the responsible and effective integration of AI within the dynamic landscape of actuarial practices. The research aligns with the evolving landscape where actuarial professionals incorporate advanced ML models in non-life insurance pricing. The literature review reveals a gap in the evaluation criteria specific to the end users, actuarial professionals in our case. The absence of a robust evaluation framework that addresses the needs of practitioners hinders the responsible deployment of XAI techniques. We address this gap by employing the A-GDTA framework to align XAI explanations with the informational needs of actuarial professionals.

Using MTPL insurance as a specific case adds practical relevance to our investigation. Predicting claim counts for such policies using GLMs is a fundamental actuarial task. By focussing on this widely used method, our research remains grounded in established actuarial practices, ensuring that our findings are applicable and adaptable within the industry. Our research establishes a bridge between traditional actuarial practices and the evolving landscape of XAI. Integrating XAI techniques, such as SHAP values, provides transparent and interpretable explanations for complex models. Aligning XAI explanations with the informational needs of actuarial professionals is crucial to enhance decision-making processes, increase transparency, and foster trust in the application of advanced machine learning methods.

Our research provides a comprehensive framework for evaluating the effectiveness of XAI explanations within the complex landscape of actuarial pricing. By combining actuarial methodologies with XAI techniques, we bridge the gap between theory and practice. We determine the information need using the GDTA study and categorize it into three levels of user needs under the Situation Awareness (SA) system. Our ultimate aim is to evaluate the quality of XAI explanations based on how effectively they meet user needs, empowering actuarial professionals with transparent and interpretable information for responsible AI integration.

## **3.2 Applying the Goal-Directed Task Analysis (GDTA) Framework in High-Stakes Industries**

Industries such as healthcare, finance, and insurance are particularly sensitive to the application of AI due to the potential safety, privacy, ethical, and financial consequences of errors or biases in decision-making processes. Insurance pricing serves as an excellent case study to examine the challenges and opportunities of AI integration due to its complex nature, reliance on sophisticated models, and direct impact on both insurers and policyholders.

The actuarial profession plays a crucial role in assessing and managing financial risks in various domains, such as insurance, pensions, and investments. To effectively navigate the complexities of these industries, actuaries rely on a systematic approach known as the ACC framework. The ACC framework serves as a comprehensive framework that guides actuaries in identifying, analysing, and managing risks while ensuring the financial stability and solvency of the organisations they serve[11].

At its core, the Actuarial Control Cycle consists of several key stages, including problem definition,

solution design, implementation, monitoring, and refinement. Each stage involves a series of tasks and decision-making processes that allow actuaries to accurately assess risks, develop appropriate solutions, and adapt to changing circumstances. For instance, in the context of general insurance pricing, the ACC framework can be applied to determine the appropriate premiums for various insurance products.

Consider a scenario in which an insurance company wants to introduce a new insurance product. Using the ACC framework, the actuarial team would begin by defining the problem, which involves understanding the specific risks associated with insuring homes, such as natural disasters, theft, and property damage. They would then proceed to gather relevant data, such as historical claims information and market trends, to inform their analysis.

The actuaries would then design a pricing model that incorporates the identified risks and aligns with the company's business objectives. This stage involves selecting appropriate statistical techniques, such as GLMs, to estimate the expected claims costs and determine the optimal premium rates. The implementation stage would involve integrating the pricing model into the company's operational systems and processes, ensuring that the new product is accurately priced and ready for launch. Throughout the product's lifecycle, the actuarial team would continuously monitor its performance, comparing actual claims experience against the assumptions used in the pricing model. If deviations occur, the actuaries would refine the model and adjust the premiums accordingly, ensuring that the product remains profitable and sustainable over time.

While the Actuarial Control Cycle provides a robust framework for managing risks, it is equally important to consider the user needs and decision-making processes involved in actuarial applications. This is where GDTA comes into play. GDTA is a methodology that focusses on understanding the goals, decisions, and information requirements of users in the context of their work environment.

In the realm of user needs analysis for dynamic decision-making, GDTA can be applied to identify key decision points, the information needed to support those decisions, and the goals that drive the decision-making process. For example, consider an actuarial team tasked with developing a pricing model for a new car insurance product. By conducting a GDTA, the team can gain insights into the specific needs and goals of the underwriters who will use the model to make pricing decisions.

The GDTA process would involve identifying the primary goal of the pricing scenario, which may be to accurately assess the risk profile of potential policyholders and determine the appropriate premium rates. The analysis would then break down this high-level goal into sub-goals and decision points, such as evaluating driving history, vehicle characteristics, and demographic factors. For each decision point, the GDTA would specify information requirements, such as access to historical claim data, industry benchmarks, and predictive models.

By combining the Actuarial Control Cycle and Goal-Directed Task Analysis, there is an opportunity to develop a tailored user needs analysis methodology specifically designed for actuarial applications. This integrated approach would allow actuaries not only to manage risks effectively, but also to ensure that the tools and models they develop align with the goals and decision-making needs of the end-users.

In summary, the ACC framework provides a comprehensive analysis for managing risks in actuarial applications, while the GDTA framework offers a structured approach to understanding user needs and decision-making processes. As shown in Figure 3.1, using the strengths of both methodologies, actuaries

can develop a robust and user-centric approach to risk management, which ultimately leads to more effective and efficient actuarial solutions.

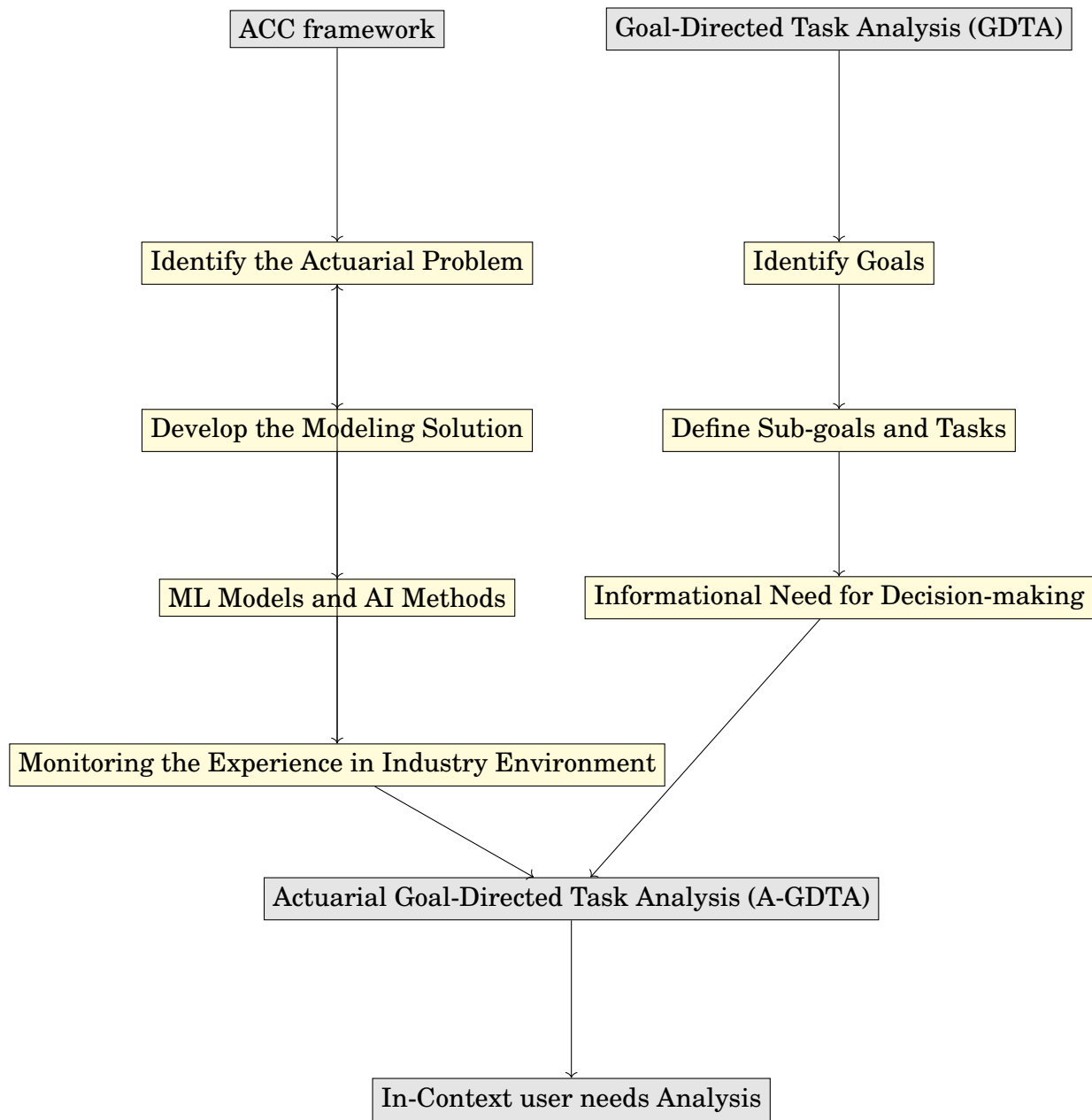


Figure 3.1: Relationship chart illustrating the combination of ACC and GDTA into the A-GDTA framework for user needs analysis.

The integrated approach of combining the Actuarial Control Cycle and Goal-Directed Task Analysis for user needs analysis in insurance pricing can be readily transferred to other safety-sensitive industries. For example, in healthcare, this methodology could be applied to develop clinical decision support systems that align with the specific needs and goals of healthcare professionals. By identifying the key decision points, information requirements, and goals involved in patient diagnosis and treatment, the resulting tools and models would be more effective and user-centric. Similarly, in the finance industry, this approach could be used to create risk assessment and management systems that cater to the specific needs of financial analysts and portfolio managers. By understanding the goals and decision-making processes involved in financial risk management, the developed solutions would be better equipped to support the users in

making informed and accurate decisions. Ultimately, the transferability of this integrated user needs analysis methodology to various safety-sensitive industries demonstrates its versatility and potential to enhance the development of effective and user-centric solutions across diverse domains.

### **3.3 Research Objective: Analyse User Needs Using A-GDTA Framework**

The primary research objective is to determine the user needs of explanations when applying ML models in safety-sensitive industry. We perform a detailed analysis of user needs in the area of actuarial pricing combining the ACC framework and GDTA framework. This investigation specifically focusses on the domain of non-life insurance pricing, with the central actuarial task being the prediction of claim counts for MTPL insurance using GLMs.

This research objective is important for several reasons. First, actuarial pricing involves complex decision-making processes that require a deep understanding of various factors, including risk assessment, the risk significance of each risk factor, and complex interactions between variables. By employing A-GDTA, the study aims to uncover the goals and sub-goals associated with actuarial pricing, providing a comprehensive overview of the informational needs of actuarial professionals. In the course of our extensive literature review, a notable gap has been identified in the evaluation of explanations generated by XAI methods. While advanced machine learning methods hold significant potential for various application scenarios, the lack of transparency in black-box models necessitates the application of XAI techniques. However, a critical deficiency arises from the absence of systematic evaluations to determine the efficacy of the explanations provided by XAI. This gap is particularly pronounced in the context of user-based evaluations, which are crucial for understanding the practical utility of these explanations in real-world scenarios. The root cause of this dearth lies in the absence of standardised and scientifically rigorous methods for assessing user needs specific to given application scenarios.

Recognising the importance of addressing this void, our research places emphasis on constructing a robust scientific framework for thorough user needs analysis. The ultimate objective is to narrow the existing gap by providing a systematic and evaluative approach to understanding and applying XAI in complex scenarios. Achieving this research objective also serves our ultimate research interest in this thesis, which is to evaluate the effectiveness of explanations produced by XAI. Understanding the user needs informs the design of metrics to assess the effectiveness of explanations generated by various XAI methods. This critical insight serves as the foundational step preceding the evaluation process afterwards, enabling us to tailor metrics that align precisely with the nuanced requirements of users. By comprehensively elucidating user needs, we ensure that the subsequent evaluation is contextually relevant and robust.

The choice of MTPL insurance as the specific product use case adds practical relevance to the investigation. MTPL insurance is a common and critical non-life insurance product, and predicting claim counts for such policies is a fundamental task in actuarial practices. The use of GLMs reflects a traditional, yet widely used method in the actuarial field, ensuring that research remains grounded in established practices.

Furthermore, the research aims to bridge the gap between traditional actuarial practices and the evolving landscape of XAI. Integrating XAI, particularly through techniques such as SHAP values, seeks

to provide transparent and interpretable explanations for complex models. The alignment between XAI explanations and the genuine informational needs of actuarial professionals is crucial for enhancing decision-making processes, increasing transparency, and fostering trust in the application of advanced machine learning methods.

The devised two-step workflow exhibits versatility, and the applicability extends beyond the actuarial realm to different sectors within the insurance industry where ML techniques are employed, provided that the actuarial component is suitably adjusted. We encourage the active exploration and extension of this study by the actuarial research community with the aim of establishing a solid foundation for the responsible integration of AI in the actuarial domain.

In summary, the research objective of analysing user needs through A-GDTA in the context of actuarial pricing is valid, as it addresses the intricacies of a key actuarial task, incorporates practical relevance through the chosen insurance product, and opens the gate to enhance transparency in decision-making processes by integrating XAI techniques.

## 3.4 Methodology

### 3.4.1 Qualitative Research Method

We intend to conduct a thorough analysis of user needs tailored to the intricate requirements of an application scenario in the actuarial domain. This involves exploring the real-world actuarial context to identify the nuanced needs and preferences of end users when using XAI techniques that help to achieve actuarial tasks. Furthermore, to ensure the relevance and applicability of the results, we calibrate the analysis by involving the perspective of domain experts, namely, actuarial professionals, whose insights and expertise are indispensable in refining the XAI explanations to align with the intricacies of actuarial practices.

We have a two-step workflow to analyse user needs and calibrate the analysis. We first construct an A-GDTA process combining the ACC framework used broadly in the industry, and the GDTA framework introduced as a method to identify and analyse situation awareness (SA) requirements by Endsley. With a central actuarial task, a modelling solution will be proposed to solve an identified actuarial problem; goals and sub-goals are analysed. Following the top-down process, we derived a set of informational need that actuarial professionals should need to achieve the sub-goals and the goals. The analysis extended the reference from the section of "XAI Application in Non-life Insurance Pricing" in the literature review<sup>2</sup>, from the perspective of analysing goals and sub-goals to fulfil the task of actuarial pricing.

To gain detailed industrial insight in analysing user needs in the context of applying XAI in predicting claim count using GLM, we designed an in-depth individual interview process to collect textual contents of goals, sub-goals, and informational needs to achieve the goals. We also aim to gain industrial concern that we might have missed in the first step of A-GDTA.

Choosing a qualitative research method is particularly advantageous for our task of analysing user needs in the realm of applying XAI in predicting claim counts using GLM within the actuarial domain. The primary advantage is the nuanced understanding we can derive from the real-world actuarial context. By employing qualitative methods, we can explore the intricate requirements and preferences of end

users in a detailed manner, allowing us to uncover subtle nuances that quantitative approaches may overlook. Additionally, qualitative research facilitates a deeper engagement with domain experts, in our case, actuarial professionals. This enables us to obtain insights that go beyond mere statistical patterns, incorporating expert judgments and contextual considerations that are crucial to refining XAI explanations to align seamlessly with the intricacies of actuarial practices. The qualitative approach is inherently flexible, allowing us to adapt and refine our analysis based on emerging insights, ensuring the relevance and depth of our findings in the dynamic actuarial landscape.

### **3.4.2 Data Collection**

#### ***Part 1: Text Scanning and Analysis***

In the initial phase of the A-GDTA framework, marked as Step 1, we conducted a detailed textual analysis focused on actuarial pricing tasks. This exploration benefited from insights from authoritative actuarial textbooks, publicly available pricing guides, and relevant industry reports. We first shortlisted all relevant industry materials in the first 30 results using the each key word combination including actuarial pricing, insurance pricing, general insurance pricing, non-life insurance pricing, and casualty insurance pricing when searching in Google. We prioritised materials published by the main chartered actuarial professional body with influence such as Institute of Actuaries(Australia), the Institute and Faculty of Actuaries(IFoA), and Casualty Actuarial Society (CAS). We read and select content directly describing pricing objectives. Then we use text search tool to find similar description in other materials. If similar description can be found in at least one more different text source, we summarise the pricing objective and and shortlist it. Using textual content as the primary data source, we progressively move from high-level objectives to more granular details. Identified objectives were then categorized based on their focus, encompassing aspects such as feature definition, correlations, feature importance, and feature interactions.

Subsequently, we synthesised the extracted text to discern the informational needs inherent in actuarial pricing. These needs were further classified and assigned to three different levels of situation awareness. This meticulous process ensured a complete understanding of the nuances within actuarial pricing tasks. The results of Step 1 will be presented in a tabular format, including attributes such as keywords on pricing objectives, high-level/low-level classification, and focus, in the results and discussion section. The ultimate outcome is an A-GDTA short report where objectives are formatted into goals, sub-goals, and mapped with Situation Awareness.

#### ***Part 2: Inputs from Actuarial Professionals***

The data collection process for Step 2 of A-GDTA involves gathering insights from domain experts through semi-structured interviews in a thematic format. The deliberate choice of this approach is based on three strategic considerations. First, the clarity obtained from Step 1 of A-GDTA allows the formulation of structured prepared questions, ensuring systematic extraction of required information during interviews. Second, the inclusion of exploratory elements is facilitated through open-ended questions with the aim of eliciting more in-depth insights from an industrial perspective. The combination of structured prepared questions and open-ended inquiries forms the foundation of our semi-structured interview design.

In addition, the decision to conduct individual interviews, as opposed to group sessions, is motivated



by the goal of obtaining independent views from actuarial professionals, minimising potential influences of social engagement with other participants. The selection criteria are stringent, inviting actuarial professionals with standard actuarial education from a university accredited by the Actuaries Institute in Australia and a minimum of one year of industry experience.

Regarding the data transcription process, we adopt a detailed note-taking approach during interviews, prioritising key content over verbatim transcriptions. Post-interview, these notes undergo refinement and formatting before being shared with interviewees via a follow-up email. Actively seeking corrections or agreements on the notes, the finalised notes serve as the collected data from the interviews. This iterative process ensures the accuracy and reliability of the information collected, aligning with the meticulous nature of the qualitative analysis in our study.

In summary, the data collection process involves a systematic two-step workflow, blending A-GDTA text analysis with semi-structured interviews, ensuring a comprehensive understanding of actuarial pricing tasks and requirements. This qualitative approach not only explores the real-world actuarial context, but also engages domain experts, enhancing the depth and relevance of our findings for the application of XAI in non-life insurance pricing.

### 3.4.3 Participants and the Use Case

To systematically identify and validate the understanding needs in GLM applications, we employed a dual-approach methodology combining text-based scanning analysis with domain expert interviews, as illustrated in Figure 3.2. The resultant framework categorises these needs into three levels of situation awareness: perception of data and model components, comprehension of model implications, and projection capabilities for future scenarios. This structured approach enables us to analyse both the theoretical underpinnings and practical requirements for effective model understanding in insurance pricing contexts.

Three actuarial professionals, each possessing a minimum of one year of industry experience, who are members of the Australian Actuaries Institute with a standard actuarial education background, are invited as domain experts for this study. They participate as interviewees with a standard actuarial education background promising a basic understanding of machine learning methods and a sufficient understanding of GLM. To mitigate potential confirmation bias, participants are deliberately selected from teams unrelated to the main researcher’s industrial work, preventing unconscious interpretation of information that aligns with pre-existing expectations. This precaution not only protects against familiarity-induced biases, but also serves to eliminate any potential conflict of interest. For example, by involving actuaries who are not directly affiliated with the researcher, the study ensures unbiased perspectives of XAI methods without the influence of previous working relationships.

We choose SHAP-value-based explanations as an example for XAI techniques, highlighting that this does not restrict the application of XAI techniques to SHAP values exclusively. The selected actuarial application scenario involves predicting the claim count for pricing purposes. GLM is a traditional method widely used in predicting models for non-life insurance pricing. In addition, most actuarial professionals possess proficiency in this method, acquired through standard actuarial education or professional experience. Hence, by choosing a scenario that actuaries are familiar with - predicting claims for MTPL insurance using GLMs, we make it feasible for them to discuss the actuarial tasks and informational need readily.

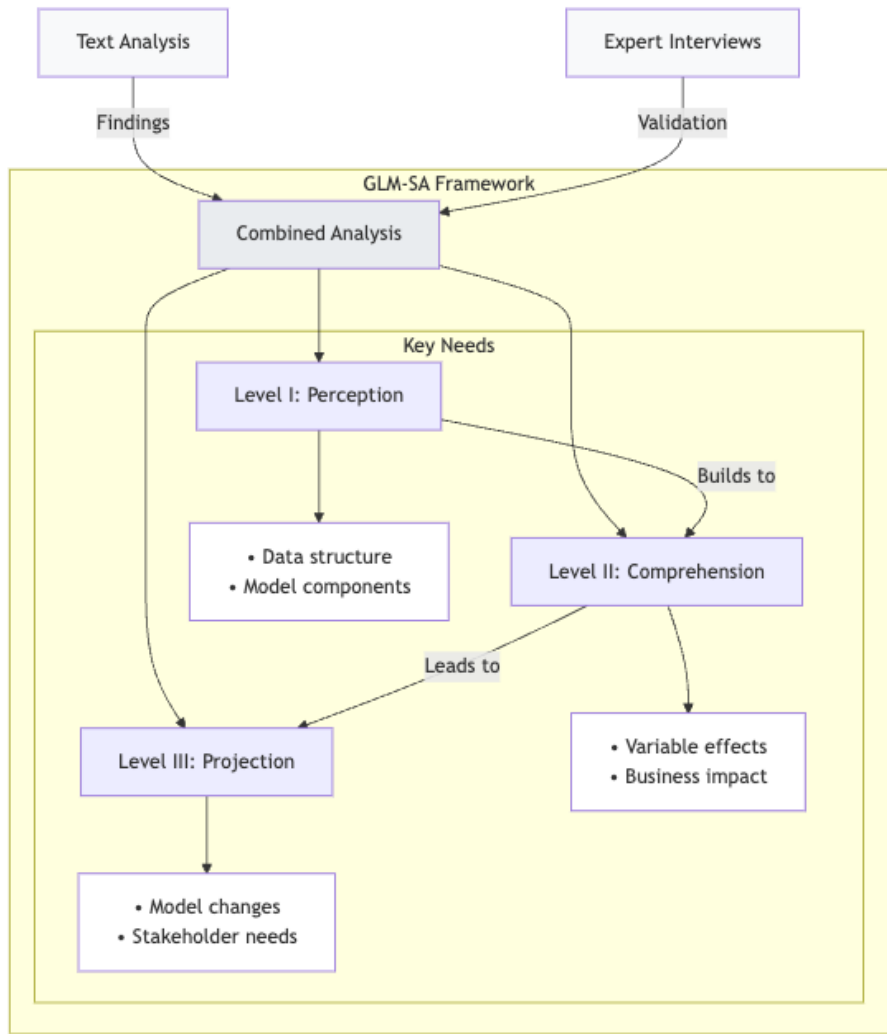


Figure 3.2: User-Based Analysis Method

This aligns with our research goal of figuring out the user-based desiderata of explanations from XAI methods in the context of setting prices for non-life insurance.

In the industrial case of MTPL insurance, the central actuarial task revolves around predicting claim counts for this specific non-life insurance product using GLMs. To achieve this task, actuarial professionals working in the insurance domain must align with the pricing task in the traditional actuarial pricing procedure, which, at the product level, involves presenting assessment on the risk relativity of each risk factors. This goal can further be dissected into sub-goals, predominantly requiring actuarial judgments and pricing decisions based on the GLM model results considering the features of MTPL insurance.

For each of these sub-goals, actuarial professionals necessitate a comprehensive understanding of the model results and the mechanisms by which these results are derived. Building on this foundational understanding, we recognise that the classic GLM pricing model, which forms the basis of the simplified use case, can be extended to more advanced variants tailored to the unique intricacies of MTPL insurance. For example, with a Gradient Boosting Machine (GBM) which captures complex patterns in the data, followed by a GLM fit on residuals, a more accurate prediction might be achieved. We use GLM as a simplified use case, with the potential to leverage this case study to more complicated machine learning algorithms.

### 3.4.4 Experiment Design: Actuarial Goal-Directed Task Analysis

#### 3.4.4.1 Experiment Preparation

##### *Scenario Specification*

Before the task analysis process, we performed a minimum scenario specification process referring to one of the classical scenario analysis procedures[69] [3], and summarised as follows.

- **Scope:** What task is the explanation to accomplish? The task is to help actuarial professionals understand the predictive model suggested for the claims count, so that risk-related judgments can be clearly made, which serves the pricing of the product.
- **Who:** Actuarial professionals are the intended users.
- **When:** The time in which the evaluation would be performed. The evaluation will be performed before the GLM is accepted for pricing use.
- **Where:** The venue where the evaluation would be performed. The evaluation will be performed in the form of an online questionnaire. There is a minor requirement for the environment in which participants complete the online survey. The environment should be generally quiet with minimal distractions. The typical environment will be a standard work-from-home work station or an office environment.
- **What:** What are the main explanatory activities? A hybrid of text-based and graphical explanations will be presented while participants will be required to respond to the quiz.

The specification of the standard scenario provides a selection of relevant details of the context for us to perform a complete GDTA study and an effectiveness evaluation afterwards. The work of evaluating effectiveness is performed with criteria developed via GDTA study, and thus will be presented in the next chapter. Following the scenario specified above, we followed the top-down identification principle to perform the user needs analysis directed by a major goal. An interview with domain experts is required so that the user's need is engaged with a practical perspective.

##### *Standard Goal-Directed Task Analysis (GDTA)*

A standard Goal-Directed Task Analysis (GDTA) is a cognitive task analysis technique that places a focus on comprehending an organisation's goals and the tasks essential for their accomplishment, particularly within the framework of Situational Awareness (SA). This method involves a systematic analysis, progressing step by step, to unravel the decisions and information requirements crucial for end-users to make informed and effective decisions.

##### *Linking GDTA to Actuarial Control Cycle*

General insurers need robust product governance processes to support the delivery of pricing promises to consumers. Effective product governance must be implemented throughout the product life cycle and must be supported by robust controls.[7]

In the context of actuarial practices, the GDTA process aligns seamlessly with the ACC framework, a comprehensive framework widely used in the industry. Combining the two frameworks allows for the effective control required by the Australian Securities and Investments Commission(ASIC). The Actuarial

Control Cycle, consisting of stages such as problem identification, solution development, implementation, and monitoring, provides a theoretical basis for constructing the GDTA work flow in the context of actuarial tasks. Specifically, GDTA corresponds to the problem identification and monitoring phases, unraveling critical information needs and decision-making processes in actuarial scenarios. The task analysis work flow is guided by combining the two framework is defined as A-GDTA.

#### **3.4.4.2 Two-Step Experiment Design**

##### ***Step 1: A-GDTA Text Analysis for Actuarial Pricing Task***

The first step involves the identification of an actuarial problem within the Actuarial Control Cycle. For our chosen product, MTPL insurance, the central actuarial task revolves around predicting claim counts. By applying A-GDTA, we delve into the informational needs associated with this central task. For instance, the sub-goals could include understanding the influence of various features like driver age, vehicle type, or geographical location on claim frequencies. This initial stage serves as a crucial foundation for proposing a targeted solution.

##### ***Step 2: Semi-Structured Interviews with Actuarial Professionals***

In the second step, we move into the solution development phase of the ACC, utilising XAI techniques to provide explanations aligned with the informational needs identified in A-GDTA. For MTPL insurance, this might involve the use of GLMs to predict claims. The objective is to offer solutions that address the identified actuarial problem to interviewees representing potential users. Subsequently, through in-depth interviews with actuarial professionals, we gain feedback. This feedback represents the monitoring phase in the ACC, where the effectiveness of the proposed solution is assessed in the context of MTPL insurance.

During the interview, we went through the GDTA process[103] so that the views of professionals can be collected as an important support to determine the information need in the case study. There were structured and unstructured components in the interview. Open questions were asked after the GDTA structured steps without a fixed frame. The interviewee was instructed to raise as many concerns as possible until no more questions could be considered during the interview session.

The use of the GLM result to guide the model of claim count for the pricing of MTPL insurance has been agreed by the interviewees. The functionality of GLM in the industry, especially in the pricing context, has been stated to help professionals understand the drivers of claims. And GLM is suitable for modelling claim count purposes.

We have designed a comprehensive semi-structured interview process, ensuring a systematic and informative interaction with our interviewees. Before the interview, participants will receive a detailed document that outlines the process and sets clear expectations for the discussion. Associated with this is a carefully draughted consent form, explicitly addressing ethical considerations, maintaining confidentiality, and emphasising the anonymity of the participants. The consent form will provide a detailed overview of the research objectives, the voluntary nature of participation, and the measures in place to protect the privacy and confidentiality of participants. Our goal is to create a transparent and respectful environment, promoting open communication while prioritising the ethical principles that guide our research. This thoughtfully designed interview process aims to elicit valuable insights while prioritising the well-being and confidentiality of our participants.

- Section 1. Introducing the research
- Section 2. Introducing the A-GDTA process in context

Three important components of GDTA, including goals, decision-making, and information need, will be illustrated. The major goal will be proposed to use the result of GLM to guide the modelling of claim count for MTPL insurance pricing. Subgoals are proposed as examples, but interviewees were also asked to give their views on what the subgoals should be. A combination of agreed sub-goals from both the interviewer and the interviewee were recorded as text-based notes.

- Section 3. We allow the interviewee to ask questions, answer questions, and take notes on the conversation.

The questions asked by the interviewee are aimed at showing the main actuarial concerns. we account for the concerns in the explanatory content production.

- Section 4. Discussion on Decision Making

We ask whether the interviewee agreed with the goal setting and why. For each sub-goals, we asked the question of what main actuarial decisions or judgement calls would be involved.

- Section 5. Ask for advice on what questions need to be answered to provide enough information for the decision-making discussed above.
- Section 6. This step is optional. We asked the question to spark a free discussion on the main challenges in terms of explainability when applying GLM in the working environment.

We illustrate the two-step experiment design in Figure 3.3. The design involves identifying an actuarial problem using the ACC framework and the GDTA framework to determine the informational needs associated with the central task. In the second step, XAI techniques are proposed to provide explanations aligned with the identified informational needs, and semi-structured interviews with actuarial professionals are conducted to gather feedback on the effectiveness of the proposed solution.

### 3.4.5 Remarks: Integration with Actuarial Control Cycle

The feedback obtained in Step 2 becomes the input to the start of the Actuarial Control Cycle. This iterative process ensures a continuous refinement of our understanding of actuarial problems and informational needs. If new insights emerge from the interviews, indicating a previously unidentified problem (e.g., a feature not considered in the initial analysis), it is incorporated into the problem identification phase of the ACC. Similarly, if the existing understanding of informational needs is challenged or corrected, it becomes another accurately identified problem within the cycle. This adaptive approach aligns with the dynamic nature of actuarial tasks and facilitates a more responsive and effective application of XAI techniques in non-life insurance pricing.

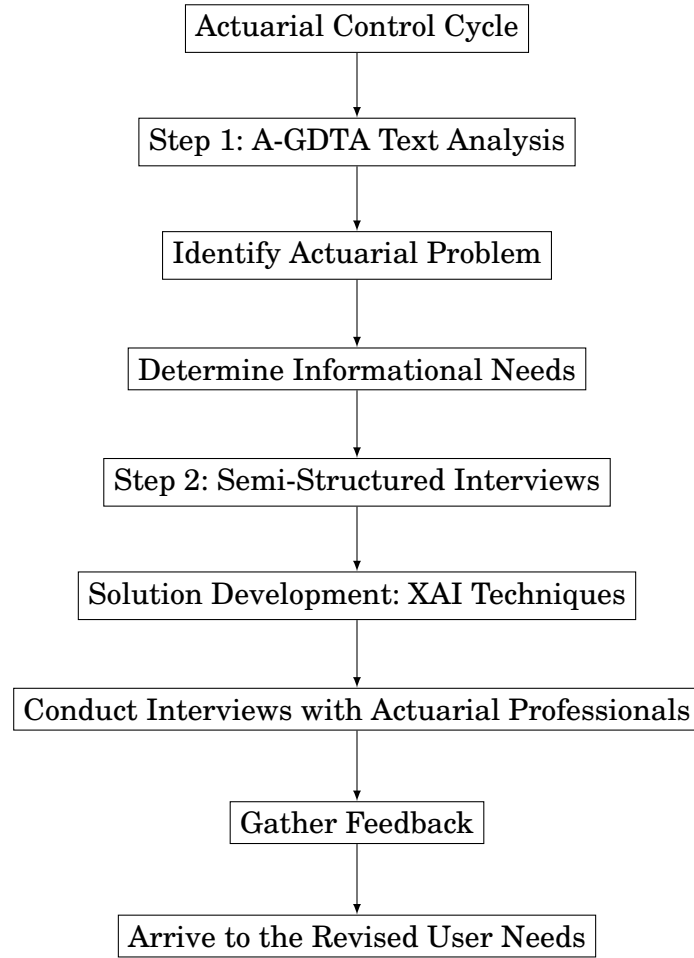


Figure 3.3: Two-Step Experiment Design Diagram

## 3.5 Results and Analysis

### 3.5.1 Determining the user needs: Desiderata of Effective EXplainable Artificial Intelligence(XAI)

#### *Result of A-GDTA Text Analysis for Actuarial Pricing Task*

We scanned and analysed authoritative actuarial textbooks, publicly available pricing guides, and relevant industry reports, with the range of non-life insurance(or general insurance), in the area of insurance pricing. We summarised the keywords on pricing objectives as in Table 3.1

The results of the scanning process yield the following insight: At a high level, actuarial tasks primarily involve elucidating the intricacies of the pricing process and premium components to various stakeholders, including portfolio managers, the underwriting team, claims management team, and regulators. Actuarial professionals must have the ability to provide accurate and comprehensive explanations of how the price is determined globally. On a more detailed level, actuaries need to have the ability to emulate the premium for a specific instance, essentially providing localised explanations. This becomes crucial in instances where a customer disputes the correctness or fairness of the price, necessitating the involvement of regulators. In the context of actuarial utilisation during the pricing phase, actuarial analysts are typically required to justify the inclusion of specific risk factors in the pricing model. In addition, they should be adept at breaking down the impact of each risk factor, corresponding to the features of the machine learning

Table 3.1: Keywords on Pricing Objectives: We have referred to but not limited to the list of references: [78][29] [64][30][91][7][94][18]. We claim that several industry reports with content similar to the referenced material are available online. Due to length constraints, we do not enumerate every document we have examined.

Keywords on Pricing Objectives	Level	Focus
Determine the coverage against general insurance risks	Low	Risk coverage
Determine financial protection against uncertainties	Low	Claim cost estimation
Determine the coverage against natural disasters	Low	Risk coverage
Determine the coverage damages to the	Low	Risk coverage
Determine the coverage against car theft	Low	Risk coverage
Determine the coverage against lost luggage	Low	Risk coverage
Determine insurance premiums	Low	Premium calculation
Estimate the claim cost for asset protection	Low	Claim cost estimation
Provide pricing information to assist claim process	High	Pricing explanation
Provide pricing information to assist underwriting process	High	Pricing explanation
Estimate the coverage of repair or replacement costs	Low	Claim cost estimation
Estimate the cost of claim for liability	Low	Claim cost estimation
Explain the pricing process	High	Pricing explanation
Determine risk coverage	Low	Risk coverage
Clarify policy terms	High	Policy clarification
Provide pricing information to assist loss payment procedures	High	Pricing explanation
Determine circumstances for payment	Low	Claim cost estimation
Determine premium factors	Low	Pricing factors
Consider underwriting concerns for pricing	High	Underwriting
Determine car insurance premiums	Low	Premium calculation
Estimate the impact of age, sex, and driving record on premiums	Low	Pricing explanation
Emulate the premiums given location, car type and car age	Low	Premiums calculation
Review the pricing promises to customers	High	Pricing explanation
Conduct effective remediation if there is a correction on price	High	Premium calculation

algorithm. This logical breakdown ensures a thorough understanding and justifiable incorporation of risk factors into the pricing model.

The actuarial technical premium is an estimate of claims costs and other business costs. More accurately, the cost-based premium formula is

$$(3.1) \quad \text{Premium} = \text{Loss} + \text{Expense} + \text{Underwriting Profit}.$$

The *Loss* in equation (3.1) is the expected claim cost, which involves randomness from the claim frequency and the claim severity.

In the case where the number of policies in a collection,  $n$ , is large, then the average provides a good approximation of the expected loss. The pure premium is then defined as  $E(X) \approx \frac{\sum_{i=1}^n X_i}{n} = \frac{\text{Loss}}{\text{Exposure}} = \text{Pure Premium}$ , that is, the sum of losses divided by the exposure. If we introduce the claim count, we get  $\text{Pure Premium} = \frac{\text{claim count}}{\text{Exposure}} \times \frac{\text{Loss}}{\text{claim count}} = \text{frequency} \times \text{severity}$ [68]. According to the definition and the pure premium formula, accurately estimating the claim cost, which includes claim count and claim severity, serves as a fundamental determinant in the pricing formulation process.

Applying the conceptual constructs to our use case: determining the price of MTPL insurance entails a crucial focus on estimating the claim cost, which stands out as the most pivotal element in the pricing

process. This estimate involves a comprehensive analysis of various factors, including the frequency of claims, which represents the frequency of claims expected to occur, and the severity of claims, which denotes the average cost per claim. We use a GLM predictive model for the claim count so that both the frequency and severity of the claim can be estimated when we know the total exposure to the policy and the average size of the claim. The actuarial pricing task is then determined as follows:

**Using GLM to predict the claim count for MTPL insurance, ensuring that the premiums are aligned with the associated risks and potential financial liabilities.**

Next, leveraging the extracted keywords and the textual content analysis, we extend our comprehension to the specific product within our use case, namely MTPL insurance. In this context, we have organised the identified objectives into a sequential list that outlines the essential pricing activities integral to routine business operations as below.

- Assess risk factors: Evaluate the risk associated with different types of vehicles.
- Determine premium rates: Calculate appropriate premium rates based on risk modeling.
- Estimate claim cost: Predict the potential cost of claims for bodily injury and property damage.
- Evaluate individual risks: Analyze the driving records and history of insured individuals.
- Adjust premium rates: Modify premium rates based on claim history and market trends to remain competitive.
- Define policy terms: Clearly specify terms and conditions for coverage.
- Establish pricing policies: Align pricing strategies with regulatory requirements.
- Review for in-force and renewals: Regularly review and update premium rates.
- Monitor industry experience: Stay informed about industry trends and adjust pricing strategies accordingly.

Across traditional actuarial pricing methods, the GLM predictive model emerges as a versatile tool that addresses pivotal objectives such as risk factor assessment, premium rate determination, and claim cost estimation. For the assessment of risk factors, our aim is to apply XAI techniques to explain these factors. With XAI tools, a system consists of GLM-based machine learning method and XAI-based explanations for the machine learning method can be developed in determining premium rates. The GLM-XAI system can be used not only to find the relationship between risk factors and premiums but also to explain the relationships as needed. In conclusion, we believe that our use case is appropriate to investigate the influence of XAI techniques in explaining model results to reach the goals we derived from the qualitative analysis.

### ***Results of Semi-Structured Interviews with Actuarial Professionals***

During the interviews, we navigated the A-GDTA process to gather valuable insights from actuarial professionals, supporting the determination of information needs in our case study. We provided a simplified use case, with prepared examples of both traditional explanation methods including permutation-based



feature importance and Individual Conditional Expectation(ICE) graph, and SHAP explanations, but only presenting examples upon explicit request from the interviewees. This approach was strategically aligned with the objective of incorporating industrial views on informational needs.

The interview format encompassed structured and unstructured components. Following the structured steps of A-GDTA, open-ended questions were introduced without imposing a rigid framework. The interviewees were encouraged to express their concerns, insights, and concerns freely, promoting a dynamic and comprehensive exploration of information needs. This approach aimed to extract a thorough understanding of the goals and sub-goals associated with actuarial pricing tasks, providing a solid foundation for subsequent analysis and synthesis.

After the interviews, the gathered notes undergo a meticulous process of refinement and formatting. Subsequently, these refined notes are shared with the interviewees via a follow-up email. The interviewees are required to provide corrections or confirmation of their agreement with the notes. The finalised notes, incorporating any necessary adjustments, then serve as the collected data derived from the interviews.

Combining the proposed sub-goals and the opinions of the interviewee, we identified three sub-goals that we aim to achieve by XAI techniques:

- Understand the main contributors of the predicted number of claims;
- Rationale the parameters of the model by comparing to industry consensus of features that may contribute to more or less claims;
- Identify any requirements for adjusting the GLM model, including non-linearity and correlation among features.

### 3.5.2 Application of Endsley 1995 Model of SA

Endsley 1995 Model of Situation Awareness (SA) showed in Figure.3.4 is a comprehensive framework for understanding and enhancing performance in professional tasks. The model consists of three levels of SA: perception, comprehension, and projection, which emphasises the dynamic interaction between goals, attention, and information processing. In the actuarial context, applying this model involves recognising the systematic and random components of the GLMs, comprehending the implications of predictor variables, and projecting potential future scenarios.

Perception at level 1 of situation awareness involves understanding the dimensions and nature of data used in the GLM. To reach level 2 of situation awareness, actuaries must grasp the effects of predictor variables, comprehend correlations, and discern pricing implications. Level 3 of situation awareness is the highest within the Endsley 1995 model. It focusses on projection, necessitating the ability to foresee the impact of altering model components and effectively use explanation methods, such as SHAP explanations, in future communication with other stakeholders.

We apply Endsley 1995 Model of SA to the user needs we have extracted from the A-GDTA process to build a GLM-SA system. The system consists of GLM-based machine learning method and SA-based user needs of explanations for the machine learning method. Building the GLM-SA system acknowledges the active role of actuaries in developing their SA, considering factors such as attention, communication, and

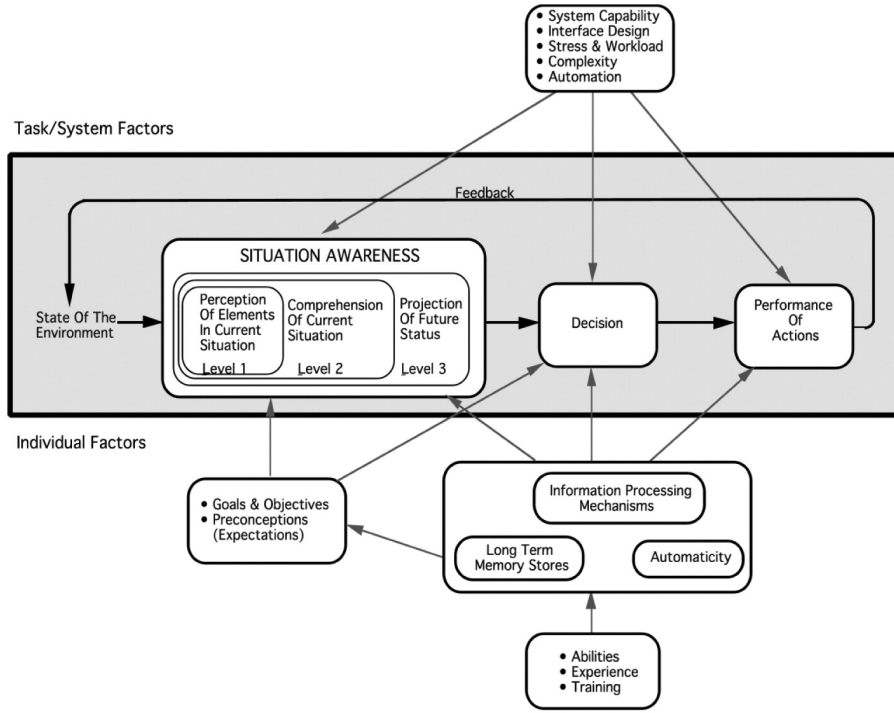


Figure 3.4: Model of SA in dynamic decision making[41]

tool manipulation. GLM-SA system provides a structured user needs category with the potential to build SA-based metrics to evaluate the effectiveness of explanations produced by XAI techniques serving a GLM model. The metrics should evaluate whether the SA improved after perceiving the explanations or whether a threshold of SA is achieved for a specific actuarial task. The evaluation results can then become the monitoring guide for actuaries to navigate the complexities of pricing decisions, addressing key factors such as data perception, model comprehension, and future projection within a dynamic actuarial landscape.

### 3.5.3 Categorising the User Needs: Situation-Awareness Based User Needs

Decision making is a typical goal in other research where A-GDTA is used for similar purposes as we do; however, we agreed in the interview that being able to make correct actuarial judgements is more appropriate as a target result of perceiving the explanations. The accuracy of judging whether a statement regarding the features and the predictive results is true or false could be the criteria to evaluate whether the user is benefited from the explanations when they are to justify the model results.

The goal-directed informational need is then driven by the information that users may need to make correct true-or-false judgements on statements around the sub-goals. We categorised the informational need under the Situation Awareness theory as follows.

#### GLM-SA I: Perception of data and model components

The domain experts emphasised the fundamental significance of understanding the relationship between predictor variables and target variables in GLM modelling. They articulated that essential considerations for model development include the semantic representation of predictors, their distributional characteristics, and their empirical validity within the model framework. The experts further underscored

the critical importance of comprehending the underlying training data's characteristics and structure, noting that these elements constitute foundational knowledge for effective model implementation.

- Understanding the systematic component and the random component of GLM,
- Understanding the target variable and the predictor variables,
- Understanding the basic interpretation of GLM as a model in the given context,
- Understanding the data being used including dimensions of data and nature of simulated data.

#### **GLM-SA II: Comprehension of the implication given by the model**

The domain experts emphasised the critical importance of understanding the magnitude and directionality of predictor variables' effects on the target variable. They articulated particular interest in the hierarchical influence of factors, specifically examining the positive or negative correlations between predictor variations and target variable responses. Furthermore, they highlighted the value of identifying both dominant and minimal-impact factors, as this enables validation against domain expertise and facilitates potential model refinement through variable selection optimisation.

- Understanding the implied effect of each predictor variable,
- Understanding the correlation or interaction between two predictor variables,
- Understanding the pricing implication of the predictor variables.

#### **GLM-SA III: Projection of the near future**

The domain experts emphasised the need for iterative model refinement capabilities, particularly highlighting the importance of impact assessment for model modifications. As one expert noted, practitioners must be equipped to evaluate the consequences of variable exclusion and the introduction of interaction terms, as these decisions frequently arise from regulatory changes or emerging business requirements. The experts also stressed the pragmatic aspects of model communication, emphasising that actuaries should possess the flexibility to leverage both traditional and contemporary explanation methods, such as SHAP values, depending on the stakeholder audience. This adaptability in explanation approach was deemed particularly crucial when engaging with non-technical stakeholders such as underwriting teams or regulatory bodies.

- Understanding the possible impact if removing one predictor variable from the existing model,
- Understanding the possible impact if adding a interaction term to the existing model,
- Understanding how to make use of explanation methods including existing methods and SHAP explanations if communication with other stakeholders is required.

As illustrated in Figure 3.5, applying the Endsley 1995 Model of SA, the goal-directed informational needs for users of GLMs are categorized into three levels: GLM-SA I (Perception of data and model components), GLM-SA II (Comprehension of the implication given by the model), and GLM-SA III (Projection of

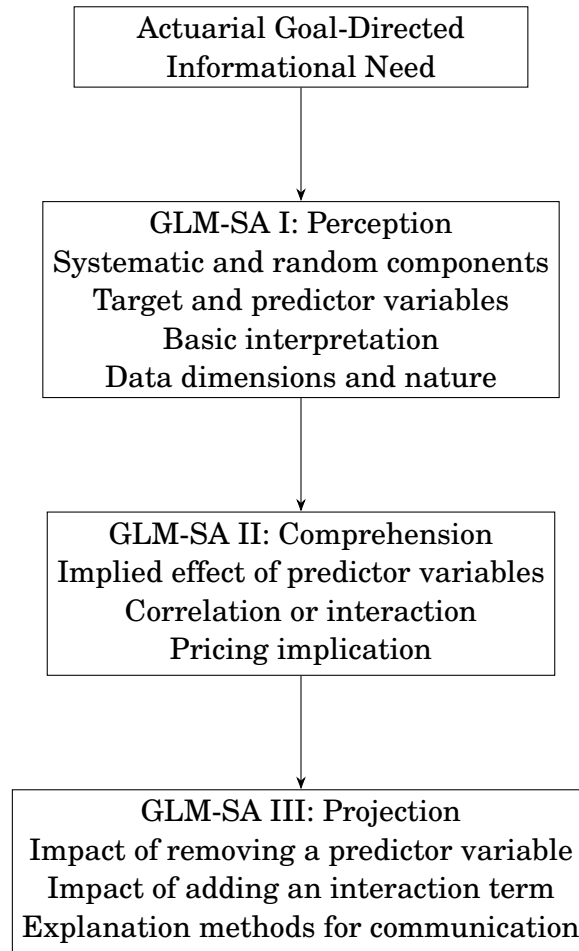


Figure 3.5: The Categorised Three Level GLM-SA User Needs

the near future). GLM-SA I focuses on understanding the basic components and data of the GLM, while GLM-SA II involves comprehending the implied effects and interactions of predictor variables and their pricing implications. GLM-SA III encompasses understanding the potential impacts of modifying the model, such as removing or adding variables, and utilising explanation methods for stakeholder communication.

### 3.6 Evaluation of Results

In this study, the A-GDTA framework, amalgamating the ACC framework and GDTA frameworks, was employed to comprehensively analyze user needs in the domain of actuarial pricing, with a specific focus on predicting claim counts for MTPL insurance using GLMs. The primary motivation behind adopting A-GDTA was rooted in addressing the intricate decision-making processes inherent in actuarial pricing, necessitating an in-depth understanding of multifaceted factors such as risk assessment, significance of risk factors, and complex variable interactions.

The qualitative analysis approach started with a text scanning-based analysis, with the aim of capturing the intricacies of pricing tasks in the chosen use case. Subsequently, the A-GDTA workflow was introduced, incorporating the standard GDTA process and the ACC framework. This systematic methodology facilitated the generation of actuarial pricing tasks, unveiling the goals and sub-goals associated with this complex domain. The scientific rigour of the method lies in its ability to provide a structured and reproducible

approach to understanding user needs in actuarial pricing.

To enhance the robustness of the user needs analysis, an in-depth interview with a domain expert was conducted. This step served to integrate industry perspectives and further tailor the extracted user needs to the practicalities of actuarial professionals. The systematic integration of industry views ensures that the user needs extracted are not only theoretically grounded but also align with the real-world challenges and requirements faced by actuarial practitioners. The scientific validity of incorporating expert opinions in qualitative analysis is underscored by the richness and depth of the insights obtained, which contribute to a more holistic understanding of the needs of the users.

To categorise and structure the derived user needs, the Endsley 1995 model of Situation Awareness (SA) was applied. This addition significantly improved the readability and organisation of the results. Endsley's model, with its focus on three levels of SA, perception, comprehension, and projection, provided a clear framework to classify informational needs systematically. This application made the results more accessible and laid a solid foundation for constructing SA-based evaluation metrics in subsequent user-based evaluation work.

The reliability of the results is fortified by the systematic and scientific nature of the methodologies employed. The A-GDTA framework, with its roots in established actuarial frameworks and task analysis methodologies, ensures that the user needs identified are contextually relevant and aligned with the intricacies of actuarial pricing. The Endsley 1995 model, a well-established and widely cited framework in the domain of SA, adds further credibility to the results by offering a recognised structure for categorising user needs.

### **3.7 Extend the Outcome to A More General Context**

In addition to the insurance pricing domain, the A-GDTA framework and the systematic approach employed in this study can be readily adapted to other similar areas beyond the actuarial profession. For instance, in the field of medicine, where healthcare professionals are faced with complex diagnostic and treatment decisions, the GDTA methodology can be applied to understand the specific goals, sub-goals, and informational needs associated with these critical processes. By analysing the decision-making points and corresponding user needs in the context of patient care, the GDTA approach can contribute to the development of more effective and user-centric clinical decision support systems. After a thorough study of professionalism guidance, an alternative to ACC in medicine context may be introduced to complete the framework.

To extend the user need analysis methodology to more general context, the transferability of this approach extends to other industries where complex decision-making processes are involved. In the aviation industry, for example, the GDTA framework can be employed to analyse user needs in the context of pilot decision-making, air traffic control, or aircraft maintenance. In the legal sector, the methodology can be adapted to understand the informational needs of legal professionals in the context of case management, legal research, or contract analysis.

The study presented here serves as a compelling example of how to perform an in-context user needs analysis tailored to industry-specific perspectives. The combination of the ACC and GDTA framework,

expert interviews, and the application of the Endsley 1995 model of Situation Awareness demonstrates a comprehensive and scientifically rigorous approach to understanding user needs in complex domains. The systematic nature of the methodology ensures that the results obtained are reliable, contextually relevant, and aligned with the practical challenges faced by professionals in their respective fields.

By showcasing the effectiveness of this approach in the insurance pricing domain, this study paves the way for future research and applications in other areas where a deep understanding of user needs is crucial for developing effective decision support tools and systems. The transferability of the methodology highlights its potential to contribute to the advancement of user-centric design and evaluation across various industries, ultimately leading to the development of solutions that are better aligned with the needs and goals of the users they serve.

### **3.8 Conclusion**

The systematic integration of Explainable Artificial Intelligence (XAI) techniques, particularly through state-of-the-art methods like SHAP values, underscores the unwavering commitment to enhancing transparency and interpretability in complex decision-making processes. By bridging the gap between traditional actuarial practices and the cutting-edge advancements in XAI methods, this groundbreaking study positions itself at the forefront of ensuring responsible and informed integration of advanced machine learning techniques in the realm of insurance pricing and other safety-sensitive domains. The meticulous approach employed in this research sets a new standard for the ethical and reliable implementation of AI-driven solutions, paving the way for a future where the power of machine learning is harnessed while maintaining the utmost integrity and accountability.

The proposed two-step workflow, firmly rooted in the A-GDTA framework, demonstrates remarkable versatility and adaptability, making it applicable to a wide range of scenarios. Its inherent flexibility allows the study findings to be rigorously validated and extended across various contexts, including pricing algorithms beyond GLMs, sophisticated ensemble machine learning models, and the continuously evolving landscape of XAI methods. Moreover, by replacing the Actuarial Control Cycle (ACC) framework with practice guidance specific to other industries, the workflow can be seamlessly extended to different domains that share a similar high sensitivity to safety concerns as the insurance industry, where intricate machine learning techniques are increasingly employed. This adaptability underscores the far-reaching impact and potential of the study, as it lays the groundwork for the responsible integration of advanced AI technologies across a broad spectrum of industries.

In conclusion, the application of the A-GDTA framework, complemented by in-depth expert interviews and structured through the renowned Endsley 1995 model, has yielded a comprehensive and scientifically grounded understanding of user needs within the context of actuarial pricing. This systematic approach ensures the contextual relevance and practical applicability of the identified user needs, laying a robust foundation for subsequent evaluations and advancements in the seamless integration of XAI techniques into established actuarial practices. The study's unwavering commitment to transparency, reliability, and scientific rigour positions it as an invaluable contribution to the rapidly evolving landscape of actuarial decision making. By setting a new benchmark for the responsible adoption of AI-driven solutions, this

research opens up exciting avenues for future exploration and collaboration, fostering a culture of trust, accountability, and innovation within the actuarial community and beyond.

In a more general context, the findings and methodologies presented in this study have far-reaching implications that extend beyond the realm of actuarial pricing and the insurance industry. The successful application of the A-GDTA framework, coupled with the integration of cutting-edge XAI techniques, serves as a powerful testament to the potential of this approach in addressing the challenges of transparency and interpretability in complex decision-making processes across a wide range of domains. From healthcare and finance to transportation and energy, industries that grapple with the responsible deployment of advanced machine learning models can draw valuable insights from this study. By adapting the workflow to their specific contexts and incorporating domain-specific best practices, researchers and practitioners can leverage the power of this approach to ensure the ethical and reliable implementation of AI-driven solutions in their respective fields. This study not only contributes to the advancement of actuarial science but also serves as a guiding light for the broader AI community, demonstrating the importance of rigorous user needs analysis, contextual relevance, and the seamless integration of explainable AI techniques in fostering trust, accountability, and transparency in the age of artificial intelligence. As we navigate the uncharted territories of AI adoption, the lessons learned from this study will undoubtedly shape the future landscape of responsible AI deployment, empowering decision-makers across industries to harness the full potential of these technologies while upholding the highest standards of integrity and fairness.





## USER-BASED EVALUATION BY SITUATION AWARENESS METRICS

### 4.1 Introduction

In the literature review conducted in Chapter 2, a discernible research gap emerged, revealing a deficiency in practical evaluations of explanations generated by XAI. This gap poses a significant impediment to the broader adoption of XAI in industry environment, particularly within the safety-sensitive industry. Although considerable research efforts have been dedicated to the development of novel methods and the application of XAI techniques to diverse pricing models, the lack of robust evaluations hinders their practical utility. The quality of an explanation is crucial for real-life applications, where considerations of financial impact and regulatory requirements weigh heavily.

Effectiveness, in the context of evaluating the explanations produced by XAI techniques for ML models, is a crucial measure of the quality and utility of these explanations. An effective explanation is one that successfully conveys the underlying reasoning, the decision-making process, and the key factors influencing the model's output to the intended users. It enables users to understand, trust, and appropriately rely on the model's predictions or recommendations.

When evaluating the quality of explanations, effectiveness serves as a comprehensive and user-centric criterion. Although other properties such as simplicity, robustness, adaptability, completeness, and faithfulness are important, the effectiveness of an explanation encapsulates the extent to which it meets the specific needs and requirements of the users within their given context. Users may prioritise certain properties over others based on their use case, and an effective explanation should align with these priorities.

In our chosen niche application area of non-life insurance pricing, actuarial professionals engaged in pricing activities contend with various stakeholders, including underwriting personnel, portfolio management teams, and operational units. However, despite the strides made in XAI, the lack of a reliable evaluation framework leaves uncertainty about the sufficiency of explanations for practical implementation. To enhance the coherence of this introduction, a logical link can be established by emphasising the critical need for reliable evaluation methods to bridge the gap between XAI advances and their effective deployment in safety-sensitive domains such as insurance pricing.

In the domain of evaluating XAI methods from a stakeholder perspective, Martin et al. [82] have introduced a framework that features a feedback loop. This framework ensures an ongoing refinement process for XAI systems, leveraging stakeholder feedback to improve system performance. This iterative feedback mechanism showed an impact in tailoring XAI systems to better align with the evolving needs and expectations of stakeholders. Acknowledging the significance of co-creation between method developers and end users, we build upon the previously cited paper where stakeholders are engaged in the evaluations. Based on this approach, we adopt a mindset centred on evaluating the quality of explanations through user feedback. The user in our case is actuarial professionals, rather than stakeholders. However, the need for communication with stakeholders is part of the user need, which has been identified through our previous study in Chapter 3.

Stepping from the user need study we have performed and the insight of evaluating based on user need, we are aspiring to formulate a comprehensive framework for evaluating the quality of XAI explanations in the context of non-life insurance pricing. This shift to user-driven evaluations not only aligns with emerging trends in XAI research, but also addresses the practical concerns raised in the insurance industry. By incorporating the perspectives of actuarial professionals and leveraging their insights, our objective is to contribute to the development of a versatile and applicable evaluation framework that can enhance the reliability and acceptance of XAI explanations in this specific domain.

The SHAP value (SHapley Additive exPlanation), as a popular XAI method, holds promise in providing explanations tailored to the high to low level of user needs. In our quest to generate explanations for subsequent evaluation, we draw inspiration from a case study that implements SHAP values as a novel technique to gauge the importance of features in interpreting the results of machine learning models [75]. Furthermore, we incorporate insights from a tutorial that uses SHAP values to explore various ways of explaining GLM predictions, particularly in the context of actuarial considerations [83]. We created a use case as adaptation from the two sources, allowing us to integrate SHAP values into our explanations generation, and then perform a user-based evaluation on the explanations.

To assess how XAI, particularly SHAP values, can enhance users' understanding of model results compared to existing explanation tools, we have enlisted the expertise of 40 volunteered actuarial professionals for a user-based evaluation. The following sections will detail the background knowledge of the case study, with a summary of the user-need analysis on three levels of Situation Awareness(SA) from our earlier study, followed by an empirical evaluation of the explanations produced by both existing methods and SHAP values. Finally, we will provide a comprehensive analysis of the empirical data, which represents the feedback from users. This user-based evaluation process sheds light on the cooperation of Machine Learning(ML) and XAI for better decision making in the context of actuarial pricing.

Our study proudly stands as the inaugural user-based evaluation within the niche of non-life insurance actuarial pricing, solidifying its role as a critical link between advanced theoretical underpinnings and the pragmatic intricacies of industry practice. Serving as a bridge between sophisticated theoretical techniques and applications in the actuarial profession, this research represents a notable advance in understanding the role of XAI in real-world decision-making. Our work holds the promise of reshaping the landscape of applied artificial intelligence, offering profound implications for future developments in XAI evaluation methodologies in actuarial science.

## 4.2 Introduction to Explanation Methods

### 4.2.1 Feature Importance: Permutation Graph and SHAP Value

In the field of machine learning, understanding the relative significance of input variables is essential. Feature importance provides a quantitative measure of how much each input feature contributes to the model output, allowing researchers and practitioners to gain insights into the underlying factors driving the predictions[49]. Provides valuable insights into the underlying relationships between the features and the target variable, helping to interpret the model, feature selection, and decision-making processes. Two prominent methods for measuring feature importance adopted in our experiment are permutation importance and SHAP (SHapley Additive exPlanations) importance. Although both techniques assess the significance of features, they approach the task from different perspectives, offering complementary insights into the model's behaviour.

Permutation importance is a model-agnostic method that evaluates the impact of each feature on the model's performance by measuring the decrease in performance when the feature's values are randomly shuffled. The process involves iteratively permuting the values of a single feature while keeping the other features unchanged and evaluating the resulting change in the model's performance metric, such as accuracy or F1 score. If the permutation of a feature leads to a significant deterioration in the model's performance, it indicates that the feature plays a crucial role in the model's predictions. Permutation importance provides a global perspective on feature importance, assessing the overall impact of each feature on the model's performance across the entire dataset. However, it does not offer insights into the direction or magnitude of a feature's influence on individual predictions.

On the other hand, SHAP importance is based on the concept of Shapley values from cooperative game theory. It assigns importance scores to each feature for a particular prediction by calculating the feature's contribution to the prediction. SHAP values represent the magnitude and direction (positive or negative) of a feature's impact on the prediction, providing a more granular understanding of feature importance at the individual prediction level. The SHAP importance of a feature is determined by the average absolute value of its SHAP values across the dataset. By considering the absolute values, SHAP importance captures the overall magnitude of a feature's influence, regardless of the direction of its impact. SHAP importance is model-agnostic and can be applied to any machine learning model, but it requires access to the model's predictions to calculate the SHAP values[84].

SHAP value explains the prediction of an instance  $x$  by computing the contribution of each feature to the prediction. It uses Shapley values from coalitional game theory, which tell us how to fairly distribute the "payout" (similar to the concept of prediction in ML context) among the features. A feature can be an individual value (for tabular data) or a group of values (e.g., superpixels for images). SHAP represents the explanation as an additive feature attribution method in the form of a linear model:

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j$$

where  $g$  is the explanation model,  $z' \in 0, 1^M$  is the coalition vector,  $M$  is the maximum coalition size, and  $\phi_j \in \mathbb{R}$  is the feature attribution for feature  $j$ , i.e., the Shapley values. We can calculate the SHAP value for

each of the features in an ML model. For a local explanation, a larger absolute SHAP value represents stronger feature importance. For a global explanation, mean absolute SHAP value are usually used to show the feature importance globally as the formula shown below.

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}|$$

The main difference between permutation importance and SHAP importance lies in their underlying approaches and the information they provide. Permutation importance focuses on the global impact of features on the model's performance, assessing the decrease in performance when a feature is randomly shuffled. It offers a high-level understanding of feature importance, identifying the features that have the most significant overall impact on the model's predictions. In contrast, SHAP importance delves into the individual feature contributions for specific predictions, providing a more detailed and localised understanding of feature importance. It reveals the magnitude and direction of each feature's influence on a particular prediction, enabling a deeper interpretation of the model's decision-making process. Although permutation importance emphasises the global importance of features, SHAP importance offers a more nuanced and interpretable analysis of feature contributions at the individual prediction level.

In conclusion, permutation importance and SHAP importance are valuable techniques for measuring feature importance in machine learning models. Permutation importance assesses the global impact of features on the model's performance, while SHAP importance is capable to provide a more granular understanding of feature contributions for individual predictions. Both methods offer complementary insights into the model's behaviour and can be used in conjunction to gain a comprehensive understanding of feature importance.

## 4.2.2 Global Explanations of Feature Effect: ICE Plots and SHAP Value

An Individual Conditional Expectation (ICE) graph is a visualisation technique that shows how the model's prediction changes as a single feature varies while holding all other features constant. It provides a way to understand the relationship between a specific feature and the predicted outcome, considering the interactions with other features. ICE graphs are created by plotting the predicted outcome on the  $y$ -axis against the values of a selected feature on the  $x$ -axis, with separate lines for each instance in the dataset. We will show graph examples in our experiment in later section.

ICE graphs offer a detailed view of how the model responds to changes in a specific feature, taking into account the individual characteristics of each instance. By examining the shape and slope of the ICE lines, users can identify patterns, non-linearities, and interactions in the model's behaviour. ICE graphs can reveal important insights, such as the presence of thresholds, plateaus, or sudden changes in the predicted outcome as the feature value varies. They provide a more nuanced understanding of the model's response to individual features compared to global feature importance measures.

On the basis of the SHAP value introduced earlier, we can also plot SHAP values by feature as a visualisation that shows the impact of each feature on the model's prediction for a specific instance. The graph displays the SHAP values for each feature, which represent the feature's contribution to the predicted outcome. Positive SHAP values indicate that the feature contributes positively to the prediction, while

negative values indicate a negative contribution. The magnitude of the SHAP value represents the strength of the feature’s influence on the prediction.

The SHAP value by feature graph provides a clear and interpretable way to understand how each feature affects the model’s prediction for a given instance. It allows users to identify the most influential features and their direction (positive or negative impact) at a glance. By examining the SHAP values, users can gain insights into the model’s decision-making process and understand which features are driving the prediction for a specific instance. This information can be valuable for model debugging, feature engineering, and explanatory purposes[84].

While both SHAP value by feature graphs and ICE graphs provide insight into the model’s behaviour, they focus on different aspects. SHAP value by feature graphs show the contribution of each feature to the prediction for a specific instance, considering the interactions among features. On the other hand, ICE graphs focus on the relationship between a single feature and the predicted outcome, showing how the prediction changes as the feature value varies for each instance. Together, these two visualisation techniques complement each other and provide a comprehensive understanding of the model’s behaviour and the impact of features on predictions.

### 4.3 Research Objective

The motivation for our research lies in developing a standardised work flow to evaluate the effectiveness of XAI explanations, with heavier focus on those generated from SHAP values, in addressing user informational needs within the context of insurance pricing. Our central research question posits whether these novel explanations can significantly enhance user satisfaction by catering to specific sub-goals identified through a detailed user need analysis we performed before this study and presented in Chapter 3.

Our research objective is to evaluate the effectiveness of explanations for machine learning models by assessing the satisfaction of user needs at different levels. Specifically, in the context of the insurance pricing case study, the objective is to evaluate the quality of explanations provided by XAI techniques for GLM-based pricing model, by measuring the satisfaction of actuarial professionals’ needs at different levels, using quantitative metrics derived from data collected through a questionnaire.

The effort to make this research objective possible required a comprehensive user-need analysis. We had a detailed illustration of the work and the results in Chapter 3. This user need analysis, inspired by Endsley’s 1995 Model of Situation Awareness (SA), classifies informational needs into three levels. Perception (GLM-SA I), Comprehension (GLM-SA II), and Projection (GLM-SA III). By applying Endsley’s model to a use case of predicting claim count for Motor Third Party Liability Insurance (MTPL), our research creates a structured framework for evaluating the impact of XAI explanations on users’ situation awareness. This approach, rooted in a deep understanding of user needs, serves as a solid foundation for evaluating the effectiveness of SHAP values for better decision making in insurance pricing. Through a systematic process, we identified three critical sub-goals aligned with the levels of Situation Awareness. These sub-goals include understanding the main contributors to predicted claims (GLM-SA I), rationalising model parameters by industry consensus (GLM-SA II), and projecting the near future with the model (GLM-SA III).

This detailed analysis allows us to categorise user needs in a way that directly corresponds to Endsley’s model, providing a nuanced and granular evaluation framework with practicability. By explicitly linking the sub-goals to the user needs and Endsley’s model, we ensure an alignment of our research objectives with users’ cognitive processes in an insurance pricing context.

The significance of this research objective comes from its pioneering contribution to addressing a notable gap in the current research community. Despite the growing importance of XAI, there has been a conspicuous lack of comprehensive evaluations, especially regarding the effectiveness of explanations produced by XAI tools. The root cause of this gap lies in the limited exploration of user needs in real-life scenarios, which hinders the development of robust evaluation methodologies. Our work bridges this critical gap by narrowing down the scope of user needs and aligning them with a well-established model of situation awareness. This alignment not only makes the evaluation of XAI explanations feasible, but also offers a novel and practical contribution to the broader discourse on transparent and interpretable AI systems. Consequently, our research provides a significant step forward in understanding and implementing effective XAI in real-world decision-making contexts.

## **4.4 Methodology**

### **4.4.1 Choosing a Cross-sectional Quantitative Study Method**

Capturing a snapshot of user perceptions is essential for addressing the research objectives outlined for several reasons. First and foremost, it provides a clear and focused lens by concentrating on a specific point in time, and we can isolate and assess the impact of the explanations on users’ situation awareness without the confounding influence of temporal factors.

Moreover, a cross-sectional approach captures user perceptions at a single point in time, which allows us to gather data that directly collect targeted feedback from participants with the practicability and applicability by focusing on data directly relevant to the current state of the insurance pricing model and the XAI techniques employed. This relevance is particularly important given the rapidly evolving nature of the field, where new advancements and techniques are continually emerging.

In addition, the snapshot approach is essential to maintain the integrity and reliability of the data collection process. By administering a questionnaire at a single point in time, we can ensure that all participants are exposed to the same set of explanations and operate under similar conditions. This consistency is crucial for minimising potential biases and confounding variables that could arise from collecting data over an extended period. By capturing a snapshot of user perceptions, we have confidence in the robustness and validity of your quantitative metrics, which form the basis for the evaluation of the explanations’ effectiveness.

Based on the cross-sectional research methodology, our design for this study includes evaluations on the Perception (GLM-SA I) and Comprehension (GLM-SA II) level in the SA-categorised user need in our Actuarial Goal-Directed Task Analysis(A-GDTA). Meanwhile, Projection (GLM-SA III) as the highest level of Situation Awareness(SA) is linked to the ability to forecast the future application of the explanation method, including projecting alternative scenarios such as effectively communicating model results to diverse stakeholders. Testing in GLM-SA III might require the collection of time-series data,

where increased attention must be devoted to managing uncertainties over time. In addition, vigilance is warranted with respect to the monitoring of long-term memory and the interplay between the perception of explanations and the ongoing accrual of professional experience.

For example, study participants may experience varying workloads during the observation period, leading to disparate levels of exposure to certain tasks. Professionals engaged in domains where XAI could be particularly beneficial are likely to manifest increased attention and interest in our research. This divergence in exposure introduces a potential bias in observations, as users with greater exposure to work scenarios conducive to XAI may exhibit distinctive patterns of interaction with the explanation method. This requires careful consideration to avoid skewing the findings through the inadvertent introduction of observational bias.

This type of design is particularly useful for our research because we aim to examine user perception and attitudes via their submitted judgements within the target users who are actuarial professionals at a given point. It is less time-consuming and resource-intensive compared to longitudinal designs, where data is collected over an extended period. Although cross-sectional studies provide valuable information on the current state of a population, they may not capture changes over time or establish causal relationships. We argue that in the case of more complicated applications which require a longer learning curve and a focused user need on Projection (GLM-SA III), extended study should be performed to collect data over time with appropriate control of variables that might depend on time.

Taking into account the resources and time allowed for our research, our decision is to employ a cross-sectional quantitative study methodology for data collection using an online questionnaire. This approach aims to establish situation awareness-based evaluation metrics, with a specific emphasis on the GLM-SA I and GLM SA II levels. Given the precision demanded in actuarial science, we consider it imperative to employ a quantitative approach for systematic data collection and analysis. This method facilitates the derivation of metrics based on Situation Awareness(SA), ensuring measurable insights. These metrics play a key role in fostering a more rigorous and objective evaluation of user perceptions.

#### 4.4.2 Experiment Design

The aim of the study is to evaluate XAI methods, specifically focussing on the popular SHAP value (Shapley additive explanation). Drawing inspiration from the framework introduced by Martin et al. [82], which emphasises a feedback loop for the ongoing refinement of XAI systems, our research design aligns with the stakeholder perspective. Recognising the significance of co-creation between method developers and end-users, we extend our study mindset to focus on evaluating XAI explanations via user feedback. This approach aims to propose a generic workflow for the evaluation of XAI explanations, ensuring their alignment with the evolving needs and expectations of stakeholders. Our study uses SHAP values to provide explanations tailored to different levels of user needs, as indicated by our user-need analysis on three levels of Situation Awareness (SA). Inspired by a case study [75] and a tutorial [83], we create a use case that integrates SHAP values into our explanation generation process, setting the stage for a user-based evaluation. This cross-sectional design, embedded in the context of non-life insurance actuarial pricing, serves as a pioneering effort in bridging the gap between theoretical advancements and the practical intricacies of industry practice, contributing to the evolving discourse on XAI evaluation methodologies in

actuarial science.

We believe that SHAP value(Shapley Additive Explanation) can provide explanations according to the three levels of user need. To produce explanations for subsequent evaluation, we propose a use case adapted from (1) a case study using the SHAP value as a novel technique to provide importance to the features of interpretation of the results given by machine learning models[75], and (2) a tutorial using SHAP value to explore different ways to explain the GLM predictions for a variety of purposes in an actuarial context[83].

To assess how XAI may improve user's understanding of model results on the foundation of existing explanation tools, we have collected 40 effective responses from volunteered actuarial professionals to do a user-based evaluation. Instead of employing a conventional dual-group design, which typically involves an experimental group and a control group to conduct a comparative study, our research methodology adopts a singular group of participants. This unique approach involves the evaluation of both established industry explanation methods and innovative SHAP value explanations within the same participant group. It is pertinent to note that the constraints on participant recruitment prevented the formation of distinct experimental and control groups. Despite this limitation, we prioritised achieving a substantial and diverse pool of participants to ensure robust subjective perceptions.

The decision to avoid group division with a smaller sample size was grounded in the aspiration to enhance the representativeness of our findings for the broader population. Smaller sample sizes are susceptible to sampling bias, where the characteristics of the sample diverge significantly from those of the population, potentially leading to unreliable results. In our pursuit of a more comprehensive and unbiased understanding, we opt for a larger, unified sample that includes all potential participants.

This methodology aligns with real-world working scenarios in which existing methods are routinely employed alongside new techniques, with the latter typically introduced after the application of established methodologies. This sequential approach facilitates a systematic comparison of results. Consistent with this pragmatic approach, our study commenced by presenting established explanation methods, followed by the introduction of innovative SHAP value explanations.

Regarding the design of the questionnaire, our objective was to articulate the explanations generated by the chosen methods with precision while capturing user perceptions through data collection. To achieve this, we adapt the explanation process described in [59], as depicted in the accompanying figure. This modified process not only aligns with our dual goals, but also proves to be an efficient means of attaining them. The refined explanation process is shown in Figure. 4.1, serves as the foundation for our questionnaire design. The modification was carried out with a deliberate focus on clarity in presenting the various explanations generated by the methods used. Simultaneously, our aim was to create a framework conducive to collecting data that accurately reflects user perceptions. By integrating insights from [59] and tailoring the process to our specific study context, we ensured that our questionnaire design optimally serves the dual purpose of elucidating explanation methods and eliciting meaningful user feedback. In Figure. 4.2, we illustrated the 14 steps participants will go through in the survey, including three main sections and several times of check-in nodes of subjective self-reported perception. During each time of subjective perception check, the questions that participants need to answer are the same.



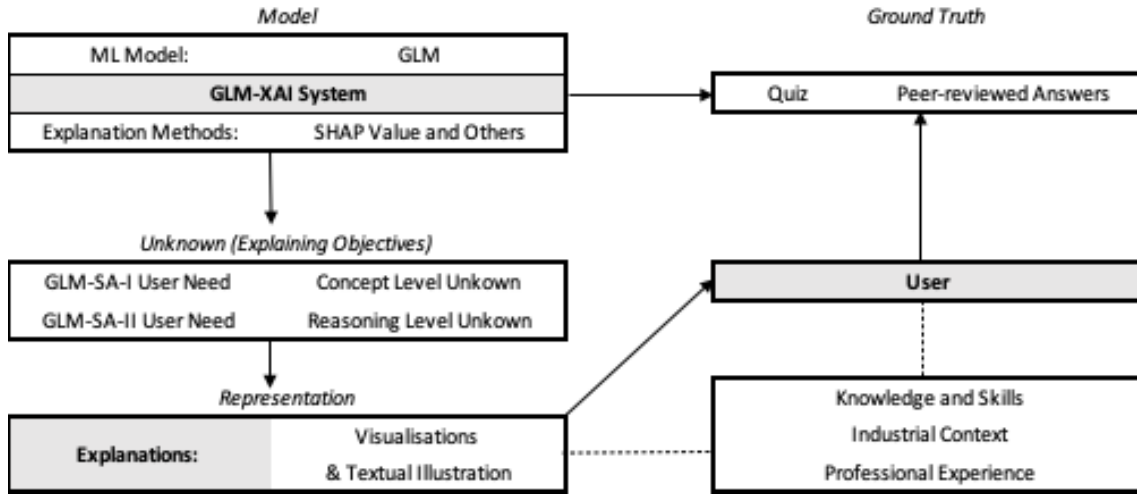


Figure 4.1: The process of explanation

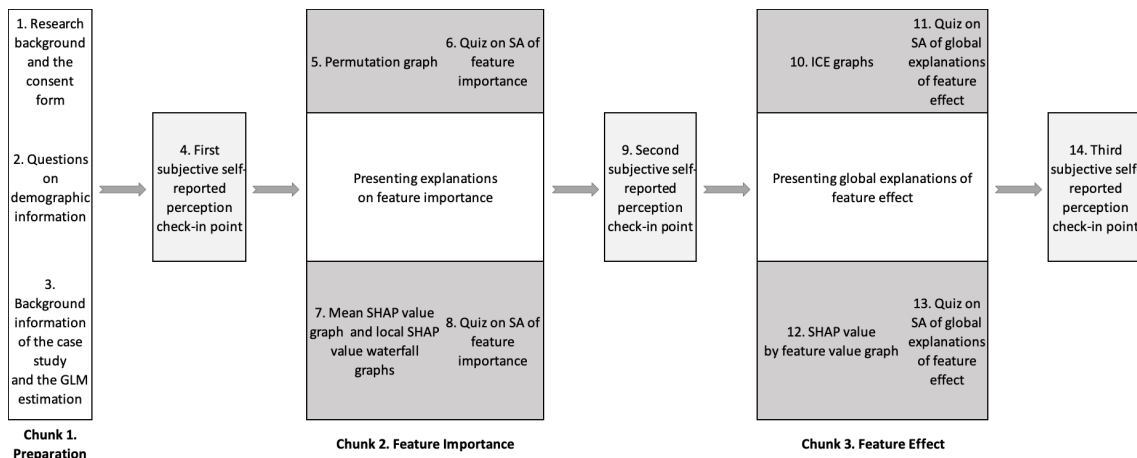


Figure 4.2: Flow chart of steps in questionnaire

### 4.4.3 Data Source

For the initial phase of the study, focused on the development of a Machine Learning (ML) model and the subsequent generation of XAI explanations, we utilised an open-source simulated claims dataset specific to Motor Third Party Liability Insurance (MTPL). This dataset is part of a collection of datasets, originally for the book 'Commutational Applied Research with R' edited by Arthur Charpentier[23], and can be access via OpenML with a data identification number 45106L. Using this data set, we constructed a predictive GLM to capture the relationship between risk factors and the predicted claims count. Subsequently, to introduce better explainability of the model and the results, we applied various explanation methods, including permutation-based feature importance, Individual Conditional Expectation (ICE) plots, and SHAP (SHapley Additive exPlanations) values.

In the second phase, we sought to perform the user-based evaluation through direct engagement with end users. To achieve this, we administered an online questionnaire, obtaining first-hand information and perceptions from individuals within the target demographic. These user-generated data serves as a user perception source, providing a quantitative dimension to our investigation by incorporating the user's insurance pricing judgements making performance after perceiving the explanations.

#### Participants

We sent around 60 formal invitations to volunteered subjects for our user participation survey, and 40 effective responses have been collected. Invitations were sent in a variety of ways, including email, text messaging, virtual meetings, and phone calls, where a short verbal or text induction was provided once the volunteer accepted the invitation and was ready for more information. The candidates are all actuarial professionals with a minimum of one year of industry experience. All of them have received standard actuarial education in accredited education programmes. We excluded actuarial professionals who are in the same working group as the author of this thesis, who works in actuarial pricing, so that conflict of interests and behaviour bias can be avoided.

As shown in Figure. 4.5, among the 40 participants, the gender split is acceptable, even though there are slightly more females. More young actuarial professionals who work closer to hands-on analytical work are expected to accept the invitation. However, we continue to collect responses until we also collected responses from actuarial professionals from senior management who have longer post-qualification experience. This makes our survey more complete and shows insights from senior actuaries with broader industry insights.

#### *Ethical Concern*

Ethical concern of the study has been addressed within UTS ethics board. According to UTS ethical principles, participation in this study is entirely voluntary. If individuals opt to participate, they will receive invitations to participate in a series of research-designed decision-making tasks. It is essential to emphasise that this study does not have physical risks. However, participants may encounter unfamiliar topics during the tasks and a mild reminder is provided that they may need to answer questions under time constraints. The time allocated for the study has been carefully determined to be more than sufficient to mitigate potential concerns about stress or anxiety. In addition, participants will be required to provide basic demographic information, such as sex and age range. All information collected is anonymised, ensuring that no personally identifiable details are disclosed. Any data presented in publications will be carefully structured to prevent the identification of individual participants.

Furthermore, it is emphasised that non-participation in the study will not have any adverse effects on the relationship between participants and researchers or the University of Technology Sydney. A key ethical consideration is the option for participants to withdraw from the study at any time, without the need to provide a reason. If a participant chooses to withdraw, any collected samples will be destroyed promptly. The consent form serves as an agreement between the participant and the research team, in which the participants consent to the collection and use of personal information for the specific research project. All information collected is treated with the utmost confidentiality. In addition, participants are informed that their information can be stored for future use in research projects related to or extending from the current study, with a continued commitment to confidentiality in all cases.

### **4.4.4 Situation Awareness Metrics**

We include both subjective metrics and objective metrics in our study.

#### **Subjective Metrics: Self-reported Situation Awareness**

We disseminate explanations to users through an online questionnaire, prompting them to indicate to what extent they agree with statements regarding their subjective perceptions of explanations. Participants will be able to choose from a scale from 0 to 10, indicating their view from "not agree at all" to "completely

0.45

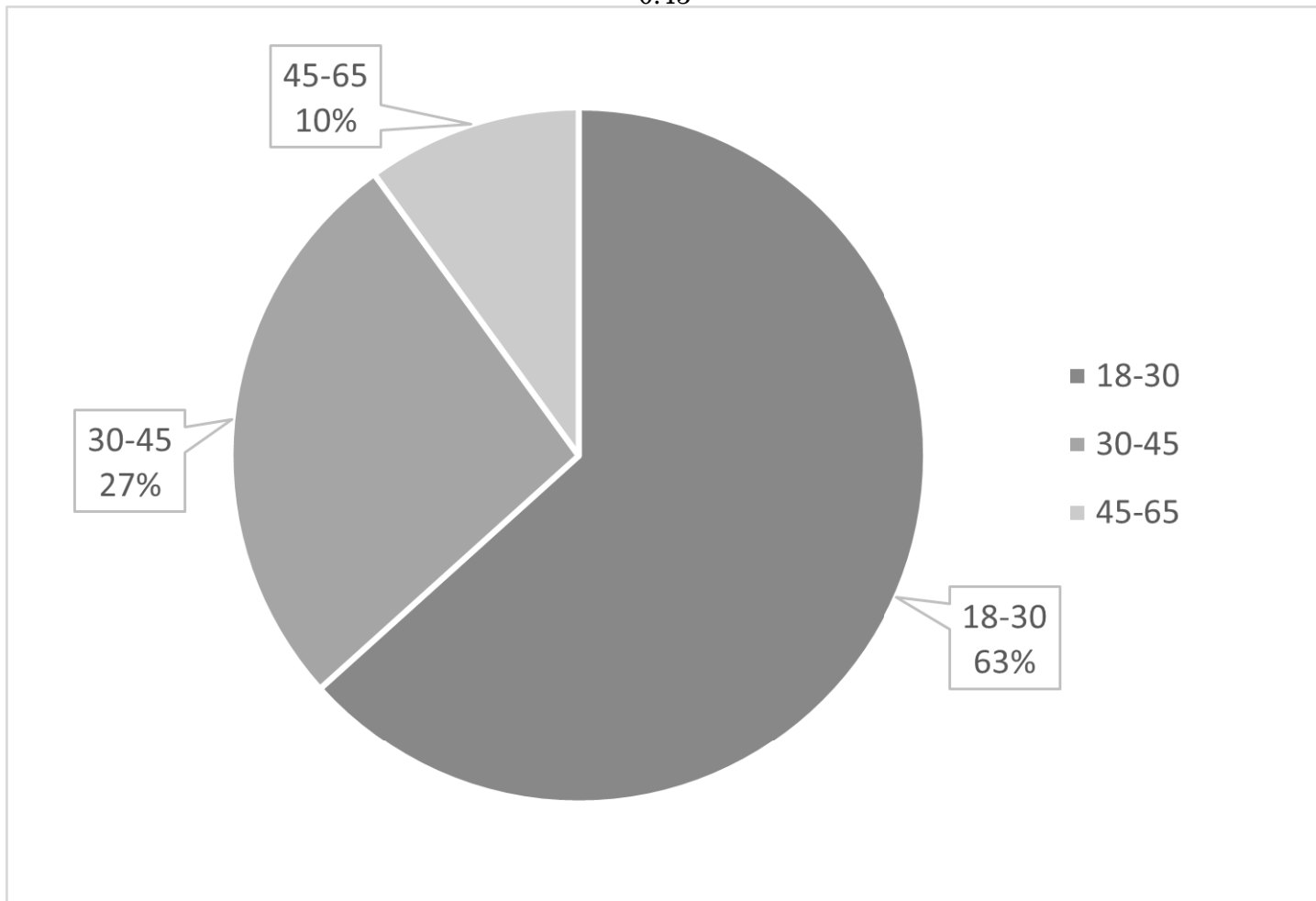


Figure 4.3: Age Distribution of Participants

0.45

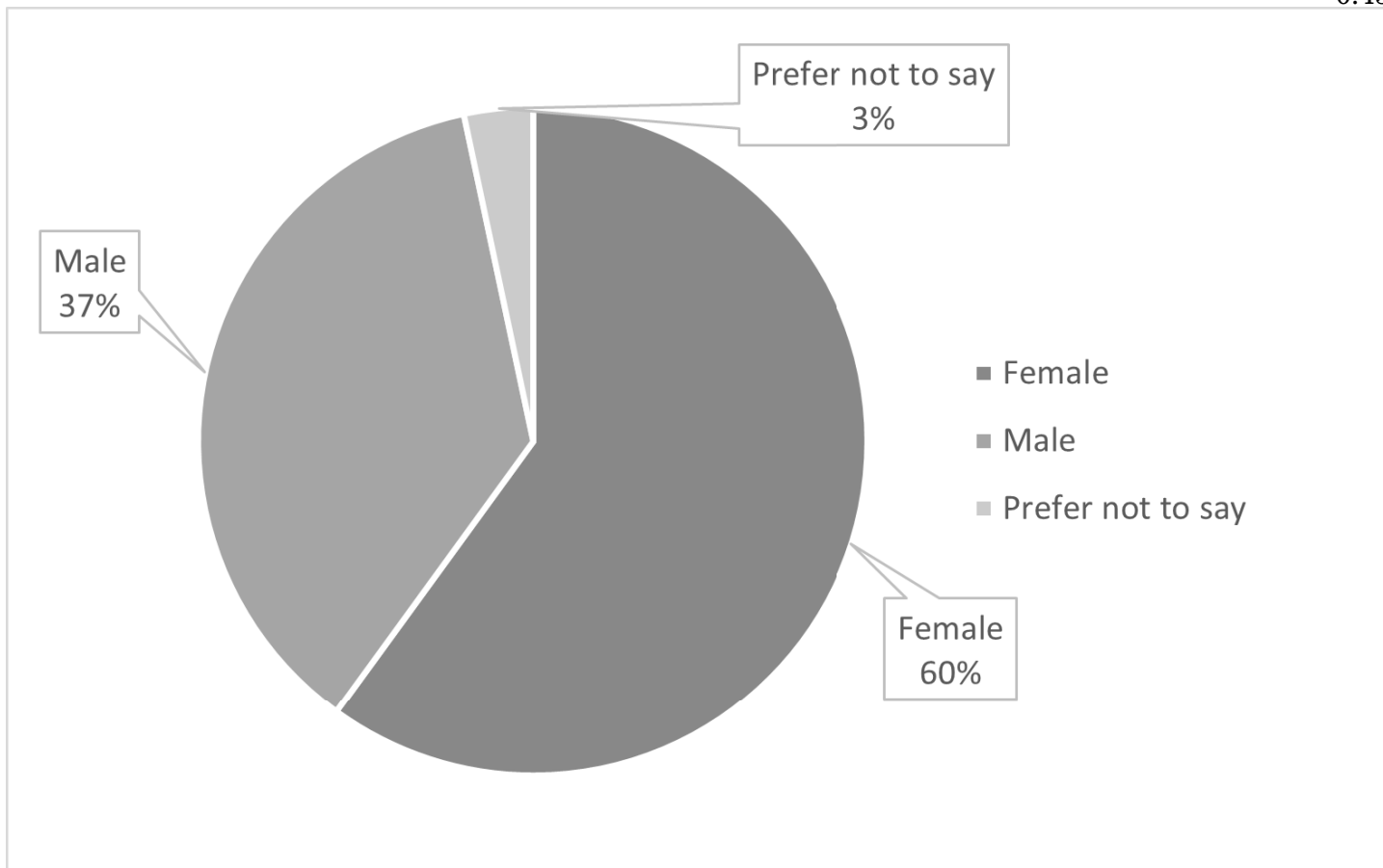


Figure 4.4: Gender Distribution of Participants

Figure 4.5: Participants Demographics

agree". This approach serves as a data source for analysing users' attitudes towards the efficacy of explanation methods, with a specific focus on their stance towards the SHAP value as a novel XAI method. The method employed for collecting users' subjective perceptions is considered both efficient and reliable, given its reliance on scale-based agreement.

Similarly, the System Usability Scale (SUS), introduced by John Brooke in 1986, has been a stalwart tool for evaluating the usability of new systems for over three decades [21]. This scale shares a fundamental similarity with our effectiveness evaluation methodology. However, it is important to note that SUS has faced criticism because of its subjective nature. Although it is efficient and provides a quick assessment, its power diminishes when used without an accompanying objective analysis. Therefore, our chosen method of user perception assessment, which incorporates a scale-based agreement system, aims to address this limitation and improve the reliability of our findings. We will mitigate this by performing an objective quiz-based evaluation for objective situation awareness metrics.

### **Objective Metrics: Situation Awareness Derived From Query Scores**

In the earlier chapter, there are three levels of user need in the GLM-based application scenario: GLM-SA I, GLM-SA II, and GLM-SA III. As GLM-SA III concerns the projection of future use of the explanation methods, which requires time series data collection, we bound our study within the user need test in GLM-SA I and GLM-SA II with the decision to use a cross-sectional quantitative analysis.

In the online questionnaire, participants are required to answer a list of "true or false" questions, which test level 1 and level 2 of Situation Awareness in the GLM-based application scenario (GLM-SA I, GLM-SA II). The accuracy of participants' responses is the source of their perception of the explanations. There is a set of correct answers prepared by the research team of this study, exploiting both the actuarial expertise from the authour of the thesis and the knowledge in XAI from the team. The answers have been reviewed and agreed on by the three interviewees who participated in the earlier A-GDTA user need study. Each question asked in the questionnaire focused on one level of situation awareness, so the raw data collected from the questionnaire is on GLM-SA I or GLM-SA II. The objective metrics will be constructed through a recalculation based on the collected data, which will be formulated in the experiment section.

## **4.5 Experiment Stage 1: Building the GLM-XAI System**

### **4.5.1 The Use Case**

We built a predictive model on GLM in the context of Motor Third Party Liability Insurance(MTPL). Then we chose (1) existing explanation methods such as ICE graphs and permutation-based feature importance, and (2) SHAP value(Shapley Additive Explanation) as the novel XAI techniques to generate explanations based on user need analysed by A-GDTA method. We propose to categorise user need into three levels of Situation Awareness, and conducted an empirical study on whether explanations by SHAP value effectively improved users' Situation Awareness.

In the context of our research, the MTPL insurance requires a revised rating system. This revision involves estimating the expected number of claims that a policyholder might file throughout the policy period. To achieve this, we employ a GLM for modelling the expected claim count. The interpretability of these model results is crucial, allowing actuarial professionals to make informed judgements and decisions

Table 4.1: Policy Data Dimensions

Variable Name	Description	Example Value or Range	Median
year	Policy year under insured	2018, 2019	N/A
town	Whether the primary driving area is in town	1, 0	N/A
driver_age	The age of the driver	[35, 56]	45
car_weight	The weight of the car in kilograms	[1090, 1460]	1240
car_power	The power of engine in kilowatts	[86, 156]	116
car_age	The age of the car in years	[1.0, 6.0]	3.0
claim_nb	The number of claims made	[0, 5]	0

about risk assessment in the pricing process. Our objective is to investigate the potential enhancement in explanatory effectiveness brought about by SHAP, a novel XAI method, in comparison to previously developed explanations before the advent of XAI.

Our primary audience comprises actuarial professionals, and as part of the user-participating evaluation, they will be tasked with reviewing the provided explanations and responding to a quiz designed to assess their perception of the explanations. This evaluation process aims to gauge the utility and comprehensibility of the explanations in a practical work setting.

### 4.5.2 Building a Predictive Generalised Linear Model

The data we used is a data set consisting of both policy and claims information simulated by a known GLM model on MTPL insurance. The dimension of the data is one million rows by seven columns. Each row represents information of an individual insurance policy. The information of the variables and the range of values are shown in Table 4.1.

The estimation of parameters in the fitted GLM model is shown in Table 4.2

Table 4.2: Estimation of GLM

Feature name	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-18.1182	14.3142	-1.27	0.2056
year	0.0076	0.0071	1.07	0.2858
town	0.3577	0.0077	46.48	0.0000
driver_age	-0.0044	0.0003	-16.19	0.0000
car_weight	-0.0000	0.0000	-2.18	0.0293
car_power	0.0041	0.0001	43.45	0.0000
car_age	-0.0216	0.0011	-19.83	0.0000

### 4.5.3 Results: Explaining the Predictive GLM Using XAI Techniques

We provide two types of explanation. The first type is feature importance, illustrating the extent of a feature's effect on the model result, i.e. how important a feature is for the predicted result given by the GLM.

The second type is the model dynamics, which shows how a feature can affect the predicted result. For each type of explanation, we have chosen one commonly used method from existing explanation methods, which have been developed along with machine learning methods, and SHAP value from novel XAI explanation methods. Hence, there are four variations of explanations.

#### 4.5.3.1 Feature Importance: Permutation Graph and SHAP Value

The permutation graph is an existing explanation method which was first proposed when Breiman introduced random forests[20]. The permutation feature importance is defined to be the decrease in the accuracy of the model when a single feature value is randomly shuffled. The textual illustration provided to users is as follows: If I randomly shuffle the value of a single risk factor in the data, for example, car power, leaving the claim number and all other factors in place, how would that affect the accuracy of predictions in the shuffled data?

SHAP values show how much a given factor changed the prediction compared to if we had made that prediction at a baseline value of that factor. SHAP values can be positive or negative, and the SHAP value can be calculated based on an individual record. We use the mean absolute value of the graph as a global explanation. We also show SHAP value waterfalls for 3 random data points for user, so that user can validate whether it is true for a specific case. Textual illustration of what SHAP value is and the illustration of the mean absolute SHAP value are provided for the users.

Listing 4.1: Packages and Functions in R Environment for Permutation Feature Importance Graph and SHAP-Based Feature Importance Graph

```
# Packages
library(kernelshap)
library(shapviz)
library	flashlight)

# Functions
# Define the log-linked GLM model in the flashlight environment
flashlight(the model, label = "GLM", predict_function)
multiflashlight(list(), data, y)
multiflashlight(fl_glm, linkinv = log)

# Functions
# Calculate the permutation feature importance and plot
light_importance()
plot()

# Functions
# Calculate the kernel shap values and visualise SHAP-based feature importance
kernelshap()
shapviz()

# Calculate the local shap values and plot waterfall chart
sv_waterfall()
```

The packages and functions used to calculate permutation feature importance and SHAP-based feature importance have been provided in Listing 4.1. The graph of importance of permutation characteristics is

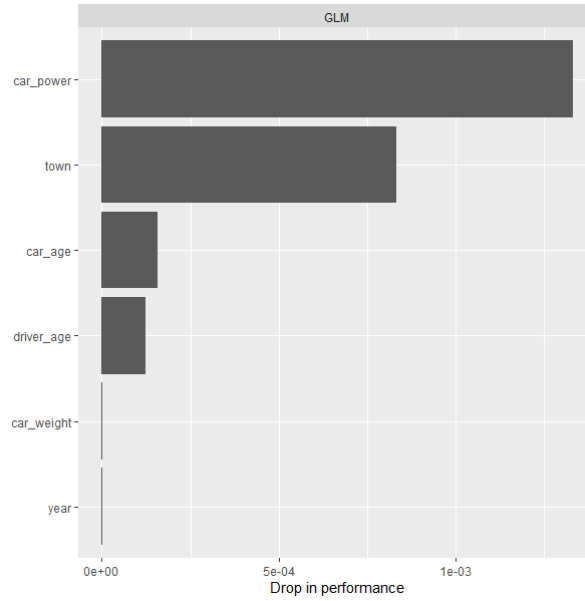


Figure 4.6: The Permutation Feature Importance Graph

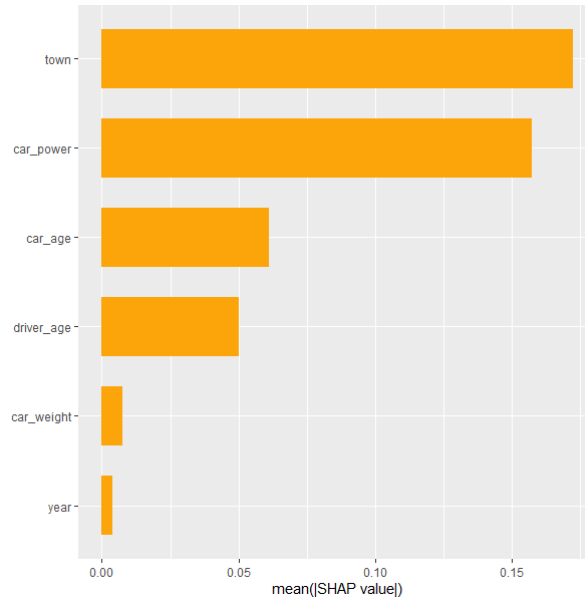


Figure 4.7: The Permutation Feature Importance Graph

shown in Figure. 4.6; the mean absolute SHAP value is shown in Figure. 4.7; one of the local examples provided to the user is Figure. 4.8. Due to the difference in the explanation method, the ranking of feature importance in these two graphs is not necessarily the same. We provided the local example to show the breakdown of the SHAP value for the value of each feature and hope to see whether this could make the mean SHAP value more convincing. We asked this question in the survey.

#### 4.5.3.2 Global Explanations of Feature Effect: ICE Plots and SHAP Value

The ICE(Individual Conditional Expectation) plots tell us that for each factor, how the prediction of the instance changes when the value of this factor changes. We provide the ICE graph of town with a textual explanation as an example to let the user know about the ICE graph. For the ICE plot for the town feature shown in Figure. 4.9, with all other factors fixed, the expected claim number(on a canonical scale) can

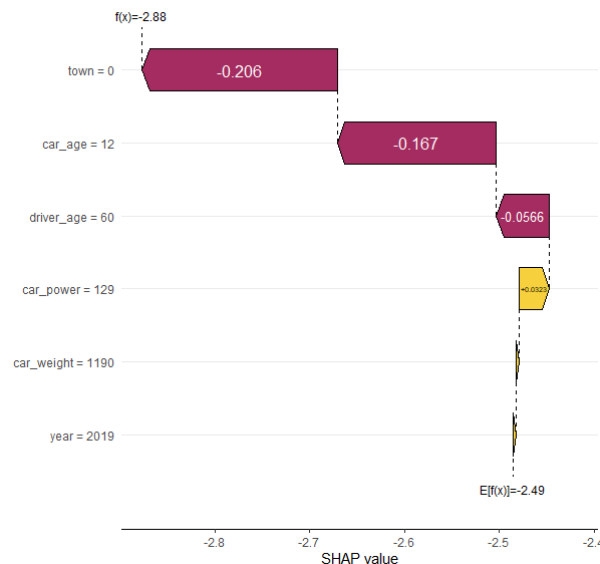


Figure 4.8: The Permutation Feature Importance Graph

change by a range from around 0.02 (that is, 2%) to around 0.06 (that is, 6%) as the town value changes from 0 to 1. Each line in the graph represents one record of the claims data.

We also provide the graph: SHAP by feature value as in Figure. 4.10 and provide a textual explanation of the feature of "town". For "town", a darker colour means low value, which should be 0. A lighter colour means a higher value, which should be 1. For a lighter colour, the corresponding SHAP value is around 0.13. It means positive effect. It is closer to 0 compared to the SHAP value of the lighter colour, which means a less intense effect. In addition, the packages and functions used to calculate ICE values, mean absolute SHAP values, and visualise results have been provided in Listing 4.2.

#### Listing 4.2: Packages and Functions in R Environment for ICE Graph and Mean Absolute SHAP Value Graph

```
# Packages
library(kernelshap)
library(shapviz)
library	flashlight)

# Functions
# Calculate the ICE values and plot
light_ice()
plot()

# Functions
# Calculate the kernel shap values and visualise SHAP-based feature importance
kernelshap()
shapviz()
sv_importance()
```



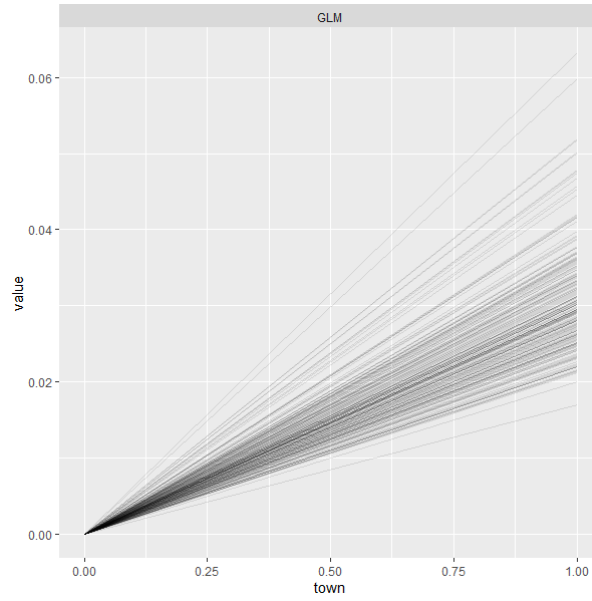


Figure 4.9: The ICE Graph for the Feature Town

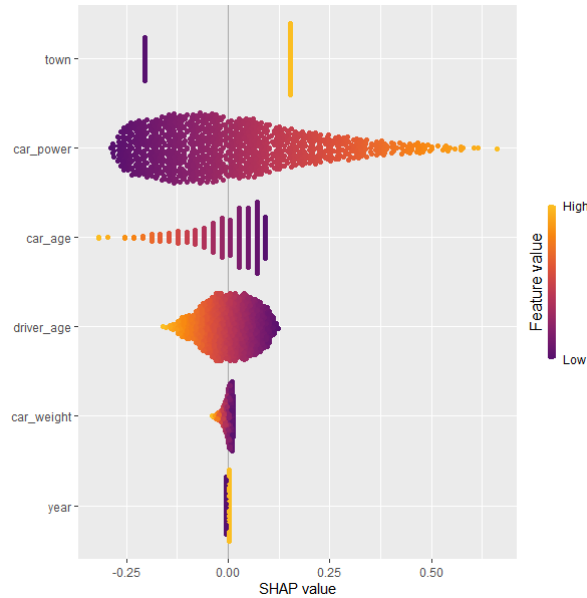


Figure 4.10: The Graph of SHAP Value by Feature Value

#### 4.5.4 Summary

In the initial phase of our experiment, a predictive GLM was employed to address actuarial pricing concerns. Based on insights gathered from a previous study, a comprehensive list of user needs relevant to the application of predictive GLM in actuarial pricing was formulated. These user needs were further refined through collaboration with domain experts during interviews, resulting in a calibrated set of requirements.

The earlier study also categorised the identified user needs into three levels of Situation Awareness, providing a framework for understanding their significance in practical applications.

In the current stage of the experiment, we have two sets of explanations of feature importance by permutation graph and mean absolute SHAP value, and two sets of explanations of feature impact by ICE graph and SHAP value by feature. These explanations were systematically aligned with the previously

categorised user needs, allowing for a nuanced examination of the model's interpretability in relation to user expectations.

Moving forward, the second half of the experiment aims to conduct a user-based participatory human evaluation. The objective is to assess whether the SHAP value, recognised as a powerful XAI method by academic peers, effectively enhances user satisfaction in meeting the identified user needs. The evaluation will employ Situation-Awareness-Based metrics to gauge improvements, providing valuable insights into the practical utility and impact of SHAP values in the actuarial domain.

## **4.6 Experiment Stage 2: User Participating Evaluation**

### **4.6.1 User-participating Questionnaire**

When understanding the explanations, users are expected to develop Situation Awareness (SA). The authors of [50] believe that Situation Awareness can support decision makers when they need to acquire an improved awareness of operational and mission-critical situations; hence we propose that SA-based metrics can be used to evaluate whether the explanations effectively improved user's ability to achieve the goal set in the GDTA study.

According to Clara Bove et al.(2021)[19], there might be potential cognitive biases when users use their existing understanding to answer the questions in the questionnaire. We believe that our target users have even a knowledge background as actuarial education in Australia is standardised and this professional area is niche. Furthermore, actuarial professionals are trained to provide the best estimate using objective judgment and ground truth, without subjective judgment[15][8], so they are less likely to use purely observation in their experience to answer the question while giving model results. Given the characteristics of the target users stated above, incorporating the understanding of explanations into judgments organically is plausible, because this is closer to the actual work situation in reality. Hence we do not design the experiment under the set-up of experimental group and control group.

In the experiment, we first provide background information about the GLM model and then provide an explanation in the order of explanation of feature importance by permutation graph, feature importance explanation by SHAP values, global explanations of feature effect by ICE graph, and global explanations of feature effect by SHAP values. The first content chunk presents the feature importance explanation for all 6 features, then test user's perception by a quiz on randomly chosen features. Users are required to respond with true, false, or not sure of the given statement on feature importance. Example statement is "Car weight does not make a difference in predicting the claims count." The second content chunk presents the global explanations of the feature effect for all six features and then tests the user's perception by a quiz on randomly chosen features. Similarly, users are required to respond with true, false, or not sure for the given statement but on how the value of the feature influences the model results. One example statement is "As the car weight increases, it is likely to have more claims."

### 4.6.2 Hypotheses

The insights we drew from the literature review and the interview with domain experts prompt us to posit that the integration of SHAP-based explanatory methodologies within industrial frameworks potentially necessitates a prolonged period for comprehensive assimilation.

In the investigation, we focus on SHAP-based explanations. By narrowing the scope of your investigation to SHAP, we provide a more targeted and in-depth analysis of how this specific XAI technique impacts users' Situation Awareness (SA) in the context of insurance pricing. According to the insights gleaned from interviews with domain experts in the study conducted in Chapter 3, other explanatory methods such as permutation feature importance graph and ICE graphs have already been integrated into real-life applications. By focussing on SHAP, we explore the potential benefits and challenges of introducing a novel XAI tool into an established industrial framework, offering valuable insights into the process of adopting emerging technologies in industry practice, especially in the new wave of advances in AI.

Within the considerations above, we have two compelling hypotheses that warrant empirical examination concerning this premise.

- H1: SHAP values as a commonly used emerging XAI tool, when used alongside existing XAI methods, may not necessarily improve users' ability to (a) understand basic feature information or (b) comprehend feature relationships in ML models.
- H2: SHAP-based explanations may differ in their effectiveness between helping users understand basic feature information versus helping them comprehend feature relationships.

The hypotheses demonstrate a keen understanding of the nuanced ways in which SHAP-based explanations may influence users' perception and decision-making processes. By examining the effectiveness of SHAP at different levels of SA (H1) and considering the potential differences in its impact on Level 1 and Level 2 SA (H2), we can provide a comprehensive and granular assessment of how this XAI technique shapes users' understanding and situational awareness. This multi-faceted approach is crucial for capturing the complex interplay between explanatory tools and human cognition, enabling us to draw meaningful conclusions about the real-world applicability and usefulness of SHAP in insurance pricing contexts. Answering RQ1 reveals structured user needs for ML explanation, while RQ2 develops metrics to evaluate explanation effectiveness. Our hypotheses then translate industry concerns into testable propositions, demonstrating our answer to RQs via case study.

Answering RQ1 reveals structured user needs for ML explanation, while RQ2 develops metrics to evaluate explanation effectiveness. Our hypotheses then translate industry concerns into testable propositions, demonstrating our answer to RQs via case study. We test the hypotheses in both the subjective and objective sense. We asked for self-reported perception from users, which builds up the self-reported Situation Awareness metrics, and tested the user's ability to perform accurate judgement accordingly after the explanations via queries. The query scores build up the quantitative evaluating metrics based on Situation Awareness, showing the effectiveness regarding whether the SA-based user needs have been satisfied.

### 4.6.3 Results and Evaluating Metrics

Based on the collected data, we used two metrics: self-reported situation awareness (SELF-SA) and situation awareness on query scores (GLM-SA) to provide complementary insights into assessing the effectiveness and utility of XAI explanations.

#### 4.6.3.1 Self-reported Situation Awareness

To understand how explanation affects the subjective experiences of the participants, we designed a self-reported assessment that participants could complete at the beginning and after each section of perception of explanation and the corresponding quiz. We asked participants to score how they agree with the following statements. Statement 1 and Statement 2 represent level 1 Situation Awareness, while Statement 3 represents level 2 Situation Awareness.

- Statement 1: I understand the main contributors of the predicted number of claims.
- Statement 2: I can tell the important factors against the less important factors. I can tell the effect of each factor.
- Statement 3: I can explain the effect of each factor on the results.

From the literature, we draw two significant advantages of SHAP value as an explanation tool and design specific statements for a self-reported perception of participants as below, such that we know whether participants acknowledge the advantages in their perception.

- The samples illustrated by SHAP value waterfalls helped me to trust SHAP value explanations more.
- SHAP value by feature having all feature in one graph is favourable for less workload.

For each response, we recalculate the scores provided into SELF-SA1-imp and SELF-SA2-imp.

$$\text{SELF-SA1-imp} = \sum(\text{post explanation SA1 level self-reported scores}) - \sum(\text{pre-explanation SA1 level self-reported scores})$$

$$\text{SELF-SA2-imp} = \sum(\text{post explanation SA2 level self-reported scores}) - \sum(\text{pre-explanation SA2 level self-reported scores})$$

Table 4.3: Statistical Test Results of SELF-SA

	Estimate	Statistic	P.value	Parameter	Conf.low	Conf.high
SELF-SA1-imp	16.76	1.45	0.1580	28.00	-6.90	40.42
SELF-SA2-imp	5.93	0.90	0.3747	28.00	-7.54	19.40

We performed an one sample T-test with the null hypothesis that the true mean of SELF-SA improvement is zero. Table. 4.3 shows the results. The P values of both tests are greater than 0.05, so we do not reject the null hypothesis that the true mean of SELF-SA improvement is zero. This means we do not have significant statistical evidence to support the statement that participants report Situation Awareness improvement with the use of explanations based on SHAP values at both level 1 and level 2 of SA.

#### 4.6.3.2 Situation Awareness on Query Scores

In the step of presenting explanations, participants are required to answer a list of "true-or-false" questions based on the explanations given. The questions test level 1 of Situation Awareness or level 2 of Situation Awareness. If the response is consistent with the correct answer from the peer reviewed, the participant gains 10 points on this question; otherwise, 0 point is gained. We recalculated the scores based on the accuracy of the entries and have GLM-SA1-PG, GLM-SA1-SP, GLM-SA2-ICE, GLM-SA2-SP. We provide a description of these quantitative metrics in Table. 4.4. Next, we explain how the metrics are calculated.

Each quantitative metric is calculated based on the responses to relevant questions that were asked after the presentation of the relevant explanations. Considering the requirement of similar workload, we have the same number of questions for each metric, with similar average finishing time. For metric  $N$ , we have:

$$(4.1) \quad \text{Metric}_N = \sum_i (\text{Score on } N)_i$$

**$N = \text{Model} - \text{SA Level} - \text{XAI Method}$**  is a defined combination of

- Model: which machine learning model is explained
- SA Level: which level of SA is targeted
- XAI Method: which explanation tool it relies on

For example,  $N = \text{GLM} - \text{SA1} - \text{PG}$  is the metric to evaluate SA1 level of the explanations for GLM-based machine learning method provided by permutation feature importance method. Note that it is transferable to be used in any XAI method application in explaining machine learning model, with emphasis to any level of SA.

Table 4.4: Description of Quantitative metrics for level 1 and level 2 of Situation Awareness

Metrics	Description
GLM-SA1-PG	Level 1 GLM Situation Awareness by permutation feature importance graph.
GLM-SA1-SP	Level 1 GLM Situation Awareness by SHAP value feature importance.
GLM-SA2-ICE	Level 2 GLM Situation Awareness based on ICE graph.
GLM-SA2-SP	Level 2 GLM Situation Awareness by SHAP value by feature.

We first calculated the sample difference between GLM-SA1-PG, GLM-SA1-SP, and the sample difference between GLM-SA2-ICE and GLM-SA2-SP for each response. We noticed that there are more responses with a negative sample mean in the former case and a positive sample mean in the latter case. We then performed paired T-test with null hypotheses that

- Test 3: The true mean difference of GLM-SA1 is greater than zero;
- Test 4: The true mean difference of GLM-SA2 is less than zero;

Table. 4.5 shows the results. For test 3, we have a p-value of 0.0616 which is greater than 0.05, so we do not reject the null hypothesis that the true mean difference of GLM-SA1 is greater than zero. We

Table 4.5: Test Result for GLM-SA1 and GLM-SA2

	Estimate	Statistic	P.value	Parameter	Conf.low	Conf.high	Method
Test 3	-4.14	-1.59	0.0616	28.00	-Inf	0.29	Paired t-test
Test 4	5.86	3.64	0.0005	28.00	3.12	Inf	Paired t-test

don't have enough statistical evidence to say SHAP-based explanations have a significant impact on the improvement of GLM-SA I level of perception according to the GLM-SA I user need. This aligns with our H1: Explanations by SHAP values do not necessarily improve user's level 1 and level 2 of Situation Awareness as an additional tool to existing explanation methods.

Given that we have a medium size of sample, we argue that sample size should not be the main reason why significant results are not obtained. The absence of significant results from the statistical tests on the impact of SHAP-based explanations on the improvement of the GLM-SA I level of perception can be attributed to several factors. First, the presentation of explanations as a one-off demonstration and test may not fully capture the potential benefits of SHAP values. Users may require more time to familiarise themselves with the novel explanatory approach and to fully grasp its implications for their decision-making processes. Additionally, the effectiveness of SHAP-based explanations may be influenced by the specific way in the visualisation they are presented and integrated into the user interface. In real life, it is more natural to allow users to try and ask questions as they perceive the new tool. The learning process should include users making mistakes and being able to correct them afterward. As we have been using a structured user screening criteria aligning the actuarial qualification system in Australia, we have reduced bias from individual differences in users' prior knowledge, cognitive abilities, and experience with XAI techniques, which is a contribution to this research area.

For test 4, we have a p-value of 0.0005 which is less than 0.05. We reject the null hypothesis and conclude that we have enough statistical evidence to state the true mean difference of GLM-SA2 is less than zero. This aligns with our H2: The effectiveness of SHAP-based explanations for level 1 and level 2 Situation Awareness can be different.

The finding that the effectiveness of SHAP-based explanations differs between Level 1 and Level 2 Situation Awareness highlights the complex nature of how users perceive and process explanatory information. This result suggests that the impact of SHAP-based explanations is not uniform across different levels of situational awareness, and that users may have varying capacities to interpret and apply the insights provided by SHAP values depending on the depth and complexity of their understanding. This observation underscores the importance of considering the multifaceted nature of situation awareness when designing and evaluating XAI techniques, and emphasises the need for explanatory approaches that can adapt to the diverse needs and abilities of users at different levels of cognitive processing.

#### 4.6.3.3 Summary of Results

The SELF-SA metric captures the subjective self-reported evaluation of users on their perceived improvement in situation awareness after being exposed to the XAI explanations. This subjective feedback offers a direct window into the user experience and the perceived value of the explanations. However, self-reports can be susceptible to biases and may not always align with objective measures of understanding and task

performance.

The GLM-SA metric addresses this potential limitation by objectively evaluating users' comprehension of the explanations through their accuracy on related query tasks. By testing situation awareness at different levels (level 1 for basic perception and level 2 for deeper comprehension and projection), this metric provides a more granular assessment of how well the explanations foster different aspects of situational understanding. The finding that SHAP-based explanations had a clearer positive impact on level 2 situational awareness highlights their potential value to promote deeper reasoning, while being less effective for more surface level situational awareness.

By considering both subjective SELF-SA and objective GLM-SA metrics in tandem, a more holistic perspective on the effectiveness of explanation can be gained. Discrepancies between self-reported and query-based metrics may reveal situations where users' perceived understanding diverges from their actual comprehension abilities. In contrast, when metrics align, it strengthens confidence in the validity of the assessments. This multifaceted approach accounts for the complex cognitive factors involved in processing explanations and translating them into meaningful situation awareness. Ultimately, leveraging these complementary metrics can guide the iterative refinement of XAI systems to better calibrate explanations to meet users' needs across different levels of situational understanding.

## 4.7 Remarks and Future Work

### 4.7.1 Extend to General Context: Explanations of Good Quality

In the broader context of industries sensitive to safety, the integration of XAI has immense potential to improve decision-making processes and ensure transparency and accountability. Beyond the realm of insurance pricing, industries such as healthcare, finance, and transportation grapple with the challenges of deploying advanced machine learning models while maintaining the trust and understanding of stakeholders. The need for reliable evaluation frameworks to assess the quality and sufficiency of XAI explanations is paramount across these domains. Drawing from key insights from our study on insurance pricing, such as the importance of user-driven evaluations, the incorporation of domain-specific considerations, and the promise of techniques such as SHAP values, researchers and practitioners can develop comprehensive evaluation methodologies tailored to their specific contexts. This user-centric approach to the evaluation of XAI, which prioritises the needs and perspectives of professionals within each industry, can contribute to the development of more effective and accepted XAI systems. As we navigate the complexities of AI adoption in safety-sensitive domains, the lessons learnt from our inaugural user-based evaluation in insurance pricing can serve as a guiding light, paving the way for responsible and transparent AI implementation across a wide range of industries.

It is important to note that while users may not explicitly express concern for certain properties such as robustness within the scope of a single task, properties of explanations that are not fully evaluated in user-based evaluation can also be crucial for establishing the scientific reliability and overall quality of the explanation method. Robustness, for example, ensures that the explanations remain consistent and accurate across different inputs and variations in the data. Lack of robustness could lead to misleading or inconsistent explanations, undermining the trustworthiness of the XAI technique. Therefore, it is essential

to consider and evaluate these properties alongside user-based effectiveness to ensure that the explanations are not only useful, but also grounded in sound scientific principles.

User-based evaluation plays a pivotal role in assessing the effectiveness of explanations, as it directly captures the perspectives and experiences of the intended users. It provides valuable insights into how well the explanations facilitate understanding, supports decision-making, and enhances user trust in the ML model. However, it is crucial to acknowledge that user-based evaluation alone may not be sufficient to comprehensively assess the quality of explanations. As mentioned, other relevant properties should also be evaluated using established methods such as model-agnostic evaluations, which assess the explanations' faithfulness to the underlying model, or expert-grounded evaluations, which involve domain experts validating the explanations' technical correctness and coherence.

Ultimately, a comprehensive evaluation framework that combines user-based effectiveness with assessments of other relevant properties is necessary to ensure the high quality of the explanations produced by the XAI techniques. User-based evaluation serves as a critical component, as it validates the practical utility and impact of the explanations in real world settings. It acts as the final touchpoint before the XAI application is approved for deployment, ensuring that the explanations meet the needs and expectations of the intended users.

By incorporating user-based evaluation as a strong pillar within a holistic evaluation framework, we can provide a robust and reliable assessment of the quality of explanations. This approach recognizes the importance of user perspectives while also acknowledging the need for scientific rigor and validation of other essential properties. Through this comprehensive evaluation, we can instill confidence in the explanations produced by XAI techniques, promoting their responsible and effective integration into real-world ML applications.

In summary, the effectiveness of explanations, as determined through user-based evaluation, is a key indicator of their quality and utility. However, it should be complemented by assessments of other relevant properties to ensure the explanations are not only useful to users but also scientifically reliable and robust. By adopting a comprehensive evaluation framework that combines user-based effectiveness with model-agnostic evaluations and other complementary methods, we can provide a strong foundation for the development and deployment of high-quality explanations in XAI for ML models.

### **4.7.2 Level 3 Model Situation Awareness**

In our study presented in this chapter, we have not yet included the level 3 Situation Awareness test in our experiment, especially for users' awareness of possible impact of adding an interaction term to the existing model and understanding how to choose explanation methods between existing methods and novel explanations using SHAP values if communication with other stakeholders is required. This might be able to test via what-if type of statements and statements in a simulated communication scenario. However, since the estimated GLM does not include the interaction term, asking what if questions on interactions might introduce cognitive bias mentioned in the experiment design, we did not conduct this test. Regarding one item of GLM-SA III communication test with stakeholders which was results from the previous study in Chapter 3, we believe that the textual content of the simulated scenario may outweigh the effectiveness of the explanation itself, which may make the responses less reliable. Also, choosing stakeholders may



introduce bias as well; for example, explaining to actuarial professionals in the valuation area might be easier than explaining to product managers. This means that less than three scenarios might not be enough for us to conclude, while extended length of the survey increases information load, which might distort the results. Furthermore, SHAP value is a powerful explanation tool, and more complicated explanations can be produced by it, however, comprehending it requires a learning curve rather than a one-off presentation. Hence in the current experiment setting, it is not appropriate to include those explanations.

We did not show more complicated explanations generated from SHAP value because the control of average complexity for an online survey. However, those explanations may empower deeper explanations that existing tools could hardly provide, especially in meeting user needs of GLM-SA III. For example, the SHAP value by feature could be made two-way, which could allow us to explore how the predicted results are impacted by two features as the feature values go to different directions. This can be helpful when we hope to investigate interaction of two features or even more than two features. It can also embark the hypotheses of non-linearities.

In short, the results of level 3 Situation Awareness in our user needs analysis in Chapter 3 were not tested in this study. It requires more complicated experiment settings, and may require iterations of survey with careful control of external aspects that may influence the results. Deeper study on level 3 model SA could be conducted on the foundation of this study.

### 4.7.3 Post Experiment Insights

We performed casual post questionnaire interviews with a few participants. The interviewees were randomly chosen and the interview was via either message or a call with a length of 5 to 10 minutes.

Before the post questionnaire interviews, we already had a belief that the one-off presentation of SHAP-value based explanations may not be sufficient to show the benefits of this new tool. Having discussed this, even though our SHAP value by feature is a basic one-way graph, some participants mentioned that they spent relatively long time in the graph of SHAP value by feature and this reduced their confidence on the new method. From the post questionnaire interview, we also realise that when information extracted from different explanation methods is inconsistent, some participants tend to feel confused and not trust their understanding of the given explanations. Therefore, adding another explanation tool without allowing enough learning time to allow users to gain familiarity can reduce the accuracy of users' response even though the additional explanation tool is of good quality of information.

There are a few expected advantages of the SHAP value by feature comparing to the existing ICE graphs. First, the efficiency of showing information is higher in SHAP value by feature. As it shows how SHAP values vary for all features in one compact graph, it is especially convenient when we want to compare among behaviours of different features. Second, we can easily tell the distribution of values from low to high for each feature, in the meanwhile, the graph also tells us the distribution of SHAP values for each features. The combination of these two benefits should have a favourable explaining effect of analysing that as the feature value increases, how the feature impact on the predicted results changes. We believe that those benefits are not fully received by participants because of the required learning time of SHAP values.

#### 4.7.4 Future Work of Deeper Investigating in Level 3 User Needs

As discussed above, we argue that deeper work in evaluating level 3 user needs can be conducted on the strong foundation of our study. To address this, an iterative XAI effectiveness control cycle can be developed. According to Tania(2023), users reported a greater curiosity about the answer to a question when they expect to learn more useful information. Especially when the explanation prompts a question of "why" type showing curiosity in the search for an explanation, there is a positive effect on reporting greater explanatory satisfaction[74].

In our context, the explanation provided by the XAI system should be highly relevant to the task users need to complete, so users have active expectations to seek useful information rather than being passive in information taking. This can be ensured by the Goal-Directed Task Analysis (GDTA) process. However, in the situation where users do not show significantly higher Situation Awareness in performing the target tasks after perceiving the explanations, allowing users to ask questions may continue to provoke users' curiosity. If the questions can be passed back to the developer side, the second submission of an explanation has the potential to show high efficiency.

Developing an iterative query process between model developers and users can play a pivotal role in fostering user need and subsequently improving the effectiveness of explanations. This iterative exchange functions as a dynamic feedback loop, catalysing mutual understanding and co-creation between developers and users. Taking queries from users as input to adjust the machine learning model itself or the XAI techniques is aligned with user-centric design principles and may provide iterative refinement of machine learning models.

One fundamental benefit of this iterative approach is its ability to stimulate user curiosity. As users actively participate in refining the XAI system through queries, they become more engaged in the process, driven by an inherent curiosity to comprehend and influence the system. The engagement allows users to expand their knowledge about the black-box method, thus improving the transparency and trustworthiness. Each query-response cycle presents an opportunity for users to delve deeper into the model's functionalities, understand its limitations, and explore its potential, thereby nurturing a sense of curiosity and exploration.

Furthermore, this iterative exchange nurtures a sense of empowerment among users. As their queries contribute to model improvements, users perceive a tangible impact on the system, fostering a sense of ownership and agency. This increased involvement ignites curiosity as users recognise their ability to shape and refine the technology with which they interact. The sense of ownership should build up user's confidence and should be able to be reflected in the subjective Situation Awareness improvement. This may lead to the compelling hypothesis that:

*Several iterations between model developers and end users can improve the subjective situation awareness of users compared to a presentation of one-off explanations.*

In reality, to deal with more complex machine learning methods and XAI explanations pointing to different objectives, the overall comprehension by end users may not be completed in a one-off explanation.

We can open the machine learning method for improvement if we arrive at the conclusion that the explainability of the machine learning method is not sufficient according to the situation awareness of users and improve the model for better explainability. We can also open up the XAI techniques for refinement if we believe the machine learning model has sufficient explainability, but the explanations produced by the

XAI techniques are not satisfying. For example, we know that LIME (Local Interpretable Model-agnostic Explanations) is a technique used for explaining the predictions of machine learning models on individual instances by approximating the model’s behaviour in the local vicinity of that instance. The approximation for individual instances may be tuned to be closer to the actual machine learning model. In some cases, the XAI technique itself may not need to be changed, but the way we present the results can be improved by improved visualisation. Here, we define an XAI system as a complete system consisting of the underlying machine learning methods and paired XAI techniques producing explanations for users. We may have another hypothesis to test:

*Refinement of the XAI system based on user feedback can improve the objective situation awareness of users compared to a one-off explanation presentation.*

Designing an experiment to test the two hypotheses mentioned above involves controlling the impact of long-term memory and the cognitive pattern of users. Controlling long-term memory involves managing the storage and retrieval of information based on its importance in the ongoing task. Relevant information should be retained for a longer period, aligning with the demands of situational awareness. Surveying the same participants after each iteration is challenging if the out-of-experiment activities of the participants have an impact on their memories of the information perceived during the experiment. The SPAM framework (Saliency-Based Progression of Attention and Memory), or similar cognitive models, can be instrumental in controlling long-term memory during an experiment regarding situation awareness. These frameworks offer structured methodologies for understanding and controlling cognitive processes, including memory, in the context of situational awareness studies. Finding an appropriate framework to deal with the control of cognitive pattern in the break of each iteration may move the study further to a case study of similar complexity to the application we may have in reality.

Incorporating iterative exchange in the evaluation of effectiveness of XAI explanations not only improves the explainability and usability of advanced machine learning methods, but also cultivates a healthy dynamics among method developers and end users. By actively involving users in the evolution of models, developers foster an environment where users are encouraged to ask questions, seek understanding, and explore the system’s capabilities, thereby significantly enhancing user satisfaction. In the actuarial control cycle context, queries raised by end users should be considered as new problems identified after the monitoring of experience, bringing the cycle back to the problem-identifying step, until no new problem is identified in the current situation. Metrics based on Situation Awareness can be designed to be the criteria for the cycle to close.

## 4.8 Conclusion

The investigation of the efficacy of SHAP-based explanations in fostering Situation Awareness (SA) represents a pivotal contribution to the evolving landscape of XAI research within industrial contexts. By rigorously evaluating the impact of SHAP on users’ SA through a multifaceted assessment approach, this study establishes a robust methodological framework that can be adapted and extended to scrutinise the effectiveness of other emerging XAI techniques as they are introduced into practical applications. This contribution is profoundly significant given the rapid pace of innovation within the XAI domain,

where novel methods and approaches are continually being developed to enhance the interpretability and transparency of complex machine learning models.

The study's focus on SHAP as an exemplar XAI tool serves as a valuable proof-of-concept, demonstrating how to systematically evaluate the effectiveness of explanatory methods in promoting informed decision-making and enhancing user comprehension. By unveiling the nuanced effects of SHAP explanations on different levels of SA, from basic perception to deeper comprehension and projection, the research highlights the multifaceted nature of situational understanding and the need for explanatory approaches that can adapt to the diverse cognitive needs of users.

In particular, the divergent findings between subjective self-reported SA and objective query-based SA metrics underscore the importance of employing a comprehensive evaluation framework that accounts for both user perceptions and objective measures of understanding. This holistic approach not only strengthens the validity and reliability of the assessments, but also provides valuable insights into potential discrepancies between users' perceived and actual comprehension abilities, informing the iterative refinement of XAI systems to better calibrate explanations to meet the needs of diverse user groups.

By demonstrating a rigorous methodology to evaluate the effectiveness of SHAP in improving informed decision making in an industrial context, this study establishes a foundation upon which future research can be based. As novel XAI techniques continue to emerge, the methodological framework presented here can be adapted and applied to assess their effectiveness across various domains, contributing to the broader discourse on the adoption and evaluation of XAI technologies in real-world practical settings.

## 5.1 Recap of Results and the Answer to Research Questions

In this research, we performed two empirical studies in actuarial pricing context answering two research questions regarding evaluating the quality of explanations of GLM-based machine learning model. Across these two studies, there is a consistent emphasis on the context and practicality when applying eXplainable Artificial Intelligence(XAI). The shared focus on connecting theoretical knowledge of XAI tools to user need in the area of actuarial pricing practices integrates the results of the two studies to reach the overall research aim of evaluating the effectiveness of machine learning explanations provided by XAI techniques via a structured user-based framework.

In the user need analysis towards Responsible XAI performed in Chapter 3, we first conducted a text scanning across published industry reports and authorised academic resources including textbook in use, and we obtained the insight that the need for actuarial professionals to provide explanations for the machine learning models they used can be summarised as: Actuarial professionals must possess the ability to provide accurate and comprehensive explanations of how the price is determined globally. On a more detailed level, actuaries need to have the ability to emulate the premium for a specific instance, essentially providing localised explanations. Next, we summarised the actuarial pricing task in the context of the use case as: Using GLM to predict the claim count of Motor Third Party Liability Insurance (MTPL), ensuring that the premiums are aligned with the associated risks and potential financial liabilities by being able to explain the risk premium components both locally and globally. Combining the elements in Actuarial Control Cycle(ACC) and Goal-Directed Task Analysis(GDTA), we reached a framework guiding us to obtain a comprehensive list of critical pricing activities, which encompasses the following eight essential sub-goals: Assess risk factors, Determine premium rates, Estimate claim cost, Evaluate individual risks, Adjust premium rates, Define policy terms, Establish pricing policies, and Review for in-force and renewals.

We verified the derived sub-goals by conducting a semi-structured interviews with actuarial professionals. Combining the proposed sub-goals and the opinions of the interviewee, we identified three most important sub-goals, and defined them with more details. After the interview, we analysed the interview

scripts and determined the goal-directed informational need as the information that users may need to make correct true-or-false judgements on statements around the sub-goals. We applied Endsley 1995 Model of Situation Awareness (SA) and categorised the informational need under the Situation Awareness theory as follows.

- GLM-SA I: Perception of data and model components
- GLM-SA II: Comprehension of the implication given by the model
- GLM-SA III: Projection of the near future

This achieves the first research objective:

- RO.1 The objective is to determine the user needs of explanations when applying ML models in safety-sensitive industry.
- Specifically, in the context of insurance pricing case study, the objective is to determine the user needs of explanations when applying GLM based on the tasks of actuarial professionals in the context of MTPL insurance pricing, aligned with industrial guidance.

Our first research question is then answered by using a case study to show a method that analyses user need systematically with contextualising higher-level guidelines. This can be leveraged for any given use case, considering the practical intricacies of the tasks. The actuarial element of the analysis ensured the smoothness of bringing user need analysis to actuarial science which is a relatively niche industry area.

In the second study, we have two stages of experiment. In the initial phase, we generated two sets of feature importance explanations using permutation graphs and mean absolute SHAP values, along with two sets of feature impact explanations through ICE graphs and SHAP values by feature. Our study aimed to test two hypotheses: H1 posited that SHAP-based explanations would not necessarily enhance users' level 1 and level 2 Situation Awareness (SA) on top of existing methods, while H2 suggested that the effectiveness of SHAP-based explanations for these two SA levels might differ. In the second phase, we conducted subjective and objective evaluation via collecting data from user-participating questionnaires, incorporating self-reported metrics and query-based evaluations. The results revealed that, from a self-reporting perspective, participants did not perceive a significant advantage in Situation Awareness improvement with SHAP-based explanations at both SA levels, aligning with H1. Objectively, there is no significant statistical evidence indicating that SHAP-based explanations impact the improvement of GLM-SA I level of perception, in line with H1. In addition, we found sufficient evidence to support H2, suggesting a differential impact of SHAP-based explanations on level 1 and level 2 Situation Awareness.

In summary, our findings do not provide significant statistical support for the enhancement of Situation Awareness with SHAP-based explanations, both subjectively and objectively, at level 1 and level 2. The discrepancy in the impact on these SA levels supports the notion that the effectiveness of SHAP-based explanations may vary across different levels of Situation Awareness.

This achieves the second research objective:

- RO.2 The research objective is to evaluate the effectiveness of explanations for machine learning models by assessing the satisfaction of user needs at different levels.

- Specifically, in the context of the insurance pricing case study, the objective is to evaluate the quality of explanations provided by XAI techniques for GLM-based pricing model, by measuring the satisfaction of actuarial professionals' needs at different levels, using quantitative metrics derived from data collected through a questionnaire.

Our second research question is then answered by developing a user-participating method that can be employed to formulate metrics to effectively evaluate the explanations produced by XAI tools across various levels of recognition required by the specific tasks users are required to undertake.

Together, they provide a comprehensive user-based evaluation framework that not only addresses the gap in XAI research, but also ensures practical applicability by aligning with the cognitive processes of users in real-world decision-making contexts.

## 5.2 Key Findings and Comparison to Previous Research

In this section, we will first discuss how the key findings of the research compare to previous research, then provide remarks on essential considerations in our research, and the limitations. Finally we discuss the outlook for this area, and a few future research directions. The initial challenge we encountered when aiming to focus on the evaluation of explanations for machine learning methods was the uncertainty surrounding what aspects to assess. Despite numerous properties of explanations, such as fidelity, simplicity, and robustness, suggested by the literature, it remained unclear whether we should evaluate all of them. In existing literature, researchers typically concentrate on a single property, proposing methods to assess it and subsequently determining the quality of explanations based on that particular property. However, the relevance of this approach may vary depending on the application scenario and the genuine concerns of users. Our primary finding is that once we identify a use case, the high-level user need can be grounded in reality and summarized as content linked to actionable tasks. Our research shifts the evaluation focus from high-level properties of explanations to the user's concerns. We posit that a comprehensive user needs analysis will guide us towards a subset of the properties identified in the literature, as users prioritize certain properties over others. Utilizing the A-GDTA method, we can deconstruct the goal behind the targeted property into sub-goals, aiding in navigating from high-level definitions to criteria embedded in the current use case. Depending on the user's objectives with the aid of explanations from XAI tools, we can compile a list of relevant user needs rather than adhering to high-level defined properties. Users may require one or more properties among the existing set, effectively bridging the gap between high-level evaluation methods and real-life application cases.

Our second discovery emphasises the practicality of categorizing user needs, closely linked to the tasks they must perform, into a well-developed framework, as opposed to defining explanations' properties at a high level and formulating theoretical metrics, which many current studies in this field attempt. We opted for Endsley's 1995 Model of Situation Awareness due to its extensive applicability in dynamic decision-making scenarios. While other frameworks may prove equally effective, the key is to ensure the chosen framework is systematic, encompassing all levels of user needs, and serves as a guide to pinpoint inadequacies in revealing certain aspects of user needs. This approach allows for an enhanced user needs analysis before the evaluation phase, contributing to the improvement of the overall evaluation process.

Regarding the chosen novel XAI method, SHAP values, our research is guided by the question: does this new method genuinely contribute value? This inquiry aligns with the initial motivation behind our study, aiming to evaluate the effectiveness of explanations generated by XAI method in enhancing users' utilisation of underlying machine learning methods, particularly those lacking inherent explainability. We posit that explanations prove effective if users perceive added value, coupled with an objective enhancement in their awareness of the application situation. In particular, within the niche realm of actuarial science, there exists a dearth of studies involving empirical user-based evaluations with a substantial sample size of actuarial professionals at work. To address this gap, our research undertakes a user-participating evaluation to gauge the effectiveness of explanations provided by SHAP values, permutation feature importance, and individual conditional expectation. We specifically delve into the potential impact of SHAP values as a novel method, assessing whether it adds value to the understanding and interpretation of machine learning models within the context of actuarial applications.

## 5.3 Remarks and Future Research Directions

This research offers several significant implications for the broader field of XAI and AI research:

### **Methodological Advancements**

- Establishes a user-centric evaluation framework that bridges theoretical XAI capabilities with practical industry requirements
- Demonstrates how Situation Awareness principles can be effectively applied to evaluate and improve XAI methods
- Introduces quantitative metrics for assessing qualitative aspects of explanations, providing a more robust evaluation approach

### **Industry Implementation**

- Provides a template for adapting XAI evaluation frameworks to high-stakes industries while maintaining generalizability
- Shows how to align XAI methods with regulatory compliance needs while preserving technical accuracy
- Demonstrates the importance of matching explanation complexity to user expertise levels in practical applications

### **Future Research Directions**

- Identifies key considerations for developing more user-friendly XAI methods that balance technical accuracy with comprehensibility
- Highlights the need for industry-specific adaptation of XAI frameworks while maintaining theoretical rigor



- Suggests approaches for making XAI more accessible to non-technical stakeholders while satisfying regulatory requirements

These implications contribute to both the theoretical advancement of XAI research and its practical implementation in regulated industries, particularly where transparency and accountability are crucial for decision-making processes. The implications outlined above emerge from our comprehensive research findings and highlight the broader impact of this work on XAI and AI research. These implications are further contextualized by specific observations and learnings from our empirical studies, particularly regarding the effectiveness of different explanation methods and their practical implementation in actuarial contexts. Our detailed analysis reveals several important considerations and opportunities for future development in this field.

Reflecting on the integration of SHAP values, permutation feature importance, and individual conditional expectation as explanation tools, we recognize that each method's advantages may manifest differently based on user needs. Using ICE graphs and the SHAP value by feature graph as examples, we explore potential strengths and considerations regarding the effectiveness of explanations in various user contexts.

ICE graphs provide detailed insights into the model's behavior for individual instances, which is particularly valuable when a feature's effect varies across instances. For discrete variables with a limited set of distinct values, the results may be less intuitive compared to continuous variables, potentially impacting the effectiveness of explanations. In contrast, the SHAP value by feature graph exhibits a consistent pattern for both continuous and discrete variables. ICE graphs aid in detecting interactions between features, showcasing how predictions for each instance respond to changes in a feature's value. The SHAP value by feature graph efficiently provides equivalent information for all features in a compact format, whereas an ICE graph represents a pair of features. With sufficient experience, participants can easily navigate the SHAP value by feature graph, proving advantageous when comparing potential interactions among features. However, in simpler cases where users need to focus on understanding the interaction for one or two pairs of features, ICE graphs may provide better clarity, potentially enhancing effectiveness.

Our use of a basic GLM model without interaction terms presents an opportunity for future research, particularly in investigating the impact of the SHAP value by feature graph when reading multiple ICE graphs becomes tedious or when comparing ICE graphs of multiple feature pairs. This direction is especially relevant considering the importance of accurately modeling interactions in actuarial pricing tasks. The decision to exclude a more complex GLM in this study stems from the learning curve associated with advanced explanations using SHAP values. To fully capitalize on the advantages of SHAP value explanations, future research could explore the learning curve through iterative surveys while monitoring long-term memory effects or by extending the length of a one-off questionnaire to ensure participants reach a threshold of understanding advanced explanations before conducting valid statistical testing. Resource constraints in terms of time and funding present an opportunity for future studies to delve deeper into this area.

Tracking emerging literature trends, we note studies exploring the development of multiplicative SHAP values, while current mature packages in technical environments like R and Python primarily

showcase the classic SHAP value. In non-life insurance pricing, multiplicative risk relativity involves adjusting a base premium to account for varying risks associated with specific policyholder characteristics. Identifying relevant risk factors and assigning multiplicative factors based on statistical analyses and actuarial considerations allows for an adjusted premium that accurately reflects the insured's unique risk profile. Acknowledging the actuarial interest in multiplicative risk relativity, essentially the effect of a risk factor in the machine learning model, we recognize the potential for future research efforts to investigate multiplicative SHAP values, connecting multiplicative risk relativity to multiplicative explanation tools.

The first study provides a comprehensive exploration of user needs through the Actuarial Goal-Directed Task Analysis (A-GDTA) framework, enriched by expert interviews and structured with Endsley's 1995 model. This approach offers a robust foundation for understanding user requirements in actuarial pricing. The qualitative nature of the study, based on scripting in expert interviews, allows for a nuanced understanding of user needs. However, it is important to acknowledge the potential for varying interpretations, highlighting the need for future research to validate and expand upon the findings.

The study's focused scope on the Motor Third Party Liability Insurance (MTPL) use case within non-life insurance pricing presents an opportunity for future research to explore the generalizability of the findings to a broader range of actuarial tasks and insurance domains. While the targeted approach allows for a detailed examination of specific user needs, future studies could investigate how these needs may vary in other actuarial scenarios, contributing to a more comprehensive understanding of user requirements across different contexts.

Our second study's experimental design and evaluation of the XAI system's effectiveness present opportunities for future research to address potential biases and explore more complex scenarios. The Level 3 Situation Awareness test, particularly in understanding the impact of adding an interaction term to the model and choosing between existing and novel explanation methods using SHAP values in a simulated communication scenario, could be further investigated in future studies. Additionally, the simplicity of the explanations presented in the one-off online survey highlights the need for future research to explore more complex scenarios that may require additional learning efforts.

The study's findings emphasize the importance of an iterative XAI effectiveness control cycle, promoting ongoing exchange between model developers and users to refine explanations. This iterative approach stimulates user curiosity, increases engagement, and empowers users by involving them in the improvement of the XAI system through queries. Future research could focus on addressing challenges related to long-term memory control, cognitive pattern management, and the design of metrics to guide the cycle's closure, further enhancing the effectiveness of XAI explanations and fostering a collaborative dynamic between developers and users within the actuarial control cycle.

In summary, our research provides a solid foundation for understanding user needs and evaluating the effectiveness of XAI explanations in the context of actuarial pricing. The combination of the A-GDTA framework, expert interviews, and the application of Endsley's 1995 model offers a comprehensive approach to assessing user requirements. The experimental design and evaluation of the XAI system's effectiveness contribute valuable insights into the potential strengths and considerations of different explanation methods. The identification of future research directions, such as exploring multiplicative SHAP values, investigating the generalizability of findings to broader actuarial tasks, and addressing potential biases

and complex scenarios, highlights the rich opportunities for further advancements in this area. By building upon the foundation established in this research, future studies can continue to refine and enhance the effectiveness of XAI explanations, ultimately leading to more transparent, trustworthy, and user-centric machine learning models in actuarial practice.

## 5.4 Publication Progress and Future Plan

### 5.4.1 Accepted Conference Papers

- **Early Application Study (2023)**
  - Conference: Injury & Disability Schemes Seminar 2023
  - Focus: Preliminary exploration of SHAP values in actuarial applications
  - Significance: Established initial framework for understanding actuarial needs in XAI
- **Multiplicative SHAP Values Study (2025)**
  - Conference: General Insurance Data Science and AI module, All Actuaries Summit 2025
  - Focus: Advanced interpretable machine learning in insurance pricing
  - Presentation: Scheduled for June 2025

### 5.4.2 Proposed Journal Publications

1. **User-Centric XAI Framework Paper** (Target: Q3 2025)
  - Target Journal: Expert Systems with Applications (Impact Factor: 8.665)
  - Focus: Comprehensive framework for user-based XAI evaluation in high-stakes industries
  - Content: Chapter 3 findings on user needs analysis
2. **Empirical Evaluation Study** (Target: Q1 2026)
  - Target Journal: Insurance: Mathematics and Economics (Impact Factor: 2.614)
  - Focus: Quantitative evaluation of XAI effectiveness in actuarial applications
  - Content: Chapter 4 findings on SHAP evaluation metrics
3. **Synthesis Paper** (Target: Q3 2026)
  - Target Journal: Artificial Intelligence Review (Impact Factor: 8.714)
  - Focus: Comprehensive review and synthesis of user-centric XAI evaluation in regulated industries
  - Content: Integration of complete thesis findings with broader implications

### 5.4.3 Alternative Journal Considerations

- Journal of Risk and Insurance (Impact Factor: 2.281)
- European Journal of Operational Research (Impact Factor: 5.334)
- Decision Support Systems (Impact Factor: 5.809)

This publication plan aims to establish the research profile in both the XAI and actuarial science domains, with journals selected based on their impact factors and relevance to the research focus. The timeline allows for thorough manuscript preparation and potential revision cycles.

## CONCLUSION

## 6.1 Machine Learning, EXplainable Artificial Intelligence, and Safety-sensitive Industries

The advent of machine learning has revolutionised various safety-sensitive industries, including the field of insurance pricing, by enabling more accurate and efficient decision-making processes. However, the complexity and opacity of these models have raised concerns about their interpretability and transparency, particularly in safety-sensitive domains such as insurance pricing. Explainable Artificial Intelligence (XAI) has emerged as a critical area of research, aiming to bridge the gap between the predictive power of machine learning models and the need for human understanding and trust in their outcomes. This thesis contributes to the growing body of knowledge in XAI by focusing on the evaluation of explanation effectiveness in the context of actuarial pricing, specifically through the lens of user needs and situational awareness.

The research presented in this thesis is motivated by the recognition that the effectiveness of XAI explanations should be assessed based on their ability to satisfy the informational needs of users in their specific decision-making contexts. We argue that a comprehensive evaluation framework, grounded in a deep understanding of user requirements and cognitive processes, is essential for ensuring the practical utility and impact of XAI techniques. To this end, we conducted two empirical studies that collectively address the research objectives of determining user needs and evaluating the effectiveness of explanations provided by XAI methods, using the case study of Generalized Linear Models (GLMs) in motor third-party liability insurance pricing.

## 6.2 User-Based Evaluation of Machine Learning Explanations

The first study (Chapter 3) employed the Actuarial Goal-Directed Task Analysis (A-GDTA) framework to systematically analyze user needs in the context of actuarial pricing. By integrating insights from industry

reports, academic resources, and semi-structured interviews with actuarial professionals, we identified three critical sub-goals: understanding the main contributors to predicted claims, rationalizing model parameters by industry consensus, and projecting the near future with the model. Applying Endsley's 1995 Model of Situation Awareness, we categorized the informational needs associated with these sub-goals into three levels: perception of data and model components (GLM-SA I), comprehension of the model's implications (GLM-SA II), and projection of the near future (GLM-SA III). This study demonstrates the value of contextualizing user needs within the practical intricacies of actuarial tasks and aligning them with established cognitive frameworks, providing a solid foundation for the subsequent evaluation of explanation effectiveness.

Building upon the insights gained from the user needs analysis, the second study (Chapter 4) focused on evaluating the effectiveness of explanations generated by XAI techniques, with a particular emphasis on SHapley Additive exPlanations (SHAP) values. We conducted a two-stage experiment, comparing SHAP-based explanations with permutation feature importance and individual conditional expectation (ICE) graphs. The study tested two hypotheses: (H1) SHAP-based explanations would not necessarily enhance users' level 1 and level 2 Situation Awareness compared to existing methods, and (H2) the effectiveness of SHAP-based explanations might differ between these two levels. Through a user-participating questionnaire, we collected both subjective and objective metrics to assess the impact of explanations on users' situational awareness. The results revealed no significant statistical evidence supporting the superiority of SHAP-based explanations in improving situational awareness at either level, aligning with H1. However, we found support for H2, indicating that the effectiveness of SHAP-based explanations varied between level 1 and level 2 Situation Awareness. This study highlights the importance of considering the differential impact of XAI techniques on various levels of cognitive processing and decision-making.

The key findings of this research contribute to the advancement of XAI evaluation in several ways. Firstly, we demonstrate the importance of shifting the focus from evaluating explanations based on high-level properties to assessing their effectiveness in addressing user concerns within specific application scenarios. By grounding the evaluation in a comprehensive user needs analysis, we ensure that the assessed properties are directly relevant to the users' objectives and decision-making processes. Secondly, our research underscores the value of categorizing user needs using well-established cognitive frameworks, such as Endsley's 1995 Model of Situation Awareness, as opposed to relying solely on theoretical metrics derived from high-level definitions of explanation properties. This approach enhances the practicality and applicability of the evaluation process, ensuring that the assessed explanations align with users' cognitive processes in real-world decision-making contexts.

## **6.3 Bring Light in Real-life Scenario**

The findings of the research shed light on the potential strengths and limitations of SHAP-based explanations in the context of actuarial pricing. While our study did not provide significant statistical evidence for the superiority of SHAP values in enhancing situational awareness, it highlighted the importance of considering the learning curve associated with novel XAI techniques and the need for users to develop familiarity with the explanation formats. Furthermore, our research identified opportunities for future

investigations, such as exploring the impact of SHAP values when comparing multiple ICE graphs, examining the generalizability of findings to broader actuarial tasks, and addressing potential biases and complex scenarios in XAI evaluation.

The insights gained from this research have important implications for both academia and industry. For the academic community, our work contributes to the growing body of knowledge in XAI evaluation, demonstrating the value of user-centric approaches and the integration of cognitive frameworks in assessing explanation effectiveness. The proposed evaluation framework, combining the A-GDTA method, expert interviews, and Endsley’s 1995 Model of Situation Awareness, serves as a valuable reference for future studies aiming to assess the impact of XAI techniques in various domains. By emphasizing the importance of contextualizing user needs and aligning them with established cognitive models, our research encourages a more nuanced and practical approach to XAI evaluation, moving beyond the reliance on theoretical metrics and high-level properties.

For the safety-sensitive industry such as insurance, our findings offer guidance on the adoption and integration of XAI techniques in pricing practices. The systematic analysis of user needs, grounded in the practical realities of actuarial tasks, provides valuable insights into the informational requirements and cognitive processes of actuarial professionals. This understanding can inform the development and refinement of XAI tools tailored to the specific needs of the actuarial community, enhancing the transparency, interpretability, and trustworthiness of machine learning models in insurance pricing. Moreover, our research highlights the importance of considering the learning curve and user familiarity when introducing novel XAI techniques, emphasizing the need for adequate training and support to ensure their effective utilization in practice.

## **6.4 Future Avenues in User-centric XAI for Safety-sensitive Industries**

Looking ahead, the findings of this thesis open up several avenues for future research. Firstly, the exploration of multiplicative SHAP values presents an exciting opportunity to align XAI techniques with the concept of multiplicative risk relativity in non-life insurance pricing. By investigating the potential of multiplicative SHAP values to capture the effects of risk factors in machine learning models, future studies can contribute to the development of more interpretable and actionable explanations for actuarial professionals. Secondly, extending the evaluation framework to a broader range of actuarial tasks and insurance domains can provide valuable insights into the generalizability of our findings and the potential variations in user needs across different actuarial contexts. Such research can contribute to the development of more comprehensive and adaptable XAI evaluation frameworks, applicable to a wide range of actuarial applications.

Furthermore, future research could delve deeper into the iterative nature of XAI effectiveness evaluation, exploring the dynamics of user engagement, long-term memory effects, and the evolution of explanations through ongoing feedback and refinement. By investigating the challenges and opportunities associated with implementing an iterative XAI effectiveness control cycle, future studies can contribute to the development of more robust and user-centric evaluation processes, fostering a collaborative and

continuous improvement approach to XAI integration in actuarial practice.

In conclusion, this thesis makes significant contributions to the field of XAI evaluation in the context of actuarial pricing. By emphasizing the importance of user needs analysis, cognitive frameworks, and practical applicability, our research provides a solid foundation for assessing the effectiveness of explanations generated by XAI techniques. The insights gained from the empirical studies, combining the A-GDTA framework, expert interviews, and Endsley's 1995 Model of Situation Awareness, offer valuable guidance for the development and integration of XAI tools in actuarial practice. The identification of future research directions, such as exploring multiplicative SHAP values and extending the evaluation framework to broader actuarial contexts, highlights the rich opportunities for further advancements in this area. As the actuarial industry continues to embrace machine learning and XAI techniques, the findings of this thesis serve as a catalyst for more transparent, interpretable, and user-centric approaches to decision-making, ultimately fostering greater trust and accountability in the use of complex models in insurance pricing and beyond.





# Sample of Questionnaire Process

## Explainability of GLM

---

Start of Block: Default Question Block

PARTICIPANT INFORMATION **Welcome! This is a study where we will look at a few explanation tools to see if they help you understand machine learning model results better. First here is information and consent form.**

### CONSENT FORM

(Content skipped for not showing candidates' identity.)

- ☐ I agree with the content above. (1)
- ☐ I don't agree and hope to withdraw my participation. (2)

---

Intro Hello, and welcome to a survey about GLM and its explainability. I believe you are an actuarial professional with basic understanding of GLM. That's everything we need for this survey. We will provide you some content to comprehend, and ask you questions. Some questions might be a little bit challenging. It's normal to be unsure about your answer. Actually, your "unsure" response will be a meaningful input to our study as well. If you are ready, let's go!

- ☐ Yes I'm ready! (1)

---

Age First let's get to know you. Which age range you are sitting in?

- ☐ 18-30 (1)
- ☐ 30-45 (2)
- ☐ 45-65 (3)

Gender Are you a male, female, or non-binary gender?

- ☐ Male (1)
  - ☐ Female (2)
  - ☐ Non-binary (3)
  - ☐ Prefer not to say (4)
- 

Bridge - 1 Thank you. Next we will tell you some context of the study, and we will know where GLM is used.

- ☐ OK! (1)
- 

GLM We are using GLM to model the claim counts for a car insurance, based on features include year, town, driver age, car weight, and car power.

The data summary table below shows the range of possible values.

For example, as for the variable "town", it is either 1 or 0. There are 600 130 records out of 1 000 000 has a value of 1 for "town", which accounts for 60% out of total.

Another example for driver age, the range of age is from 35 to 56, with a median of 45.

You are not required to memorise any of those information :) But do spend a few more minutes to understand the context if you need!

---

GLM We have fitted a GLM model as below using standard machine learning modelling process, with a proper performance test.

You can trust that this model is ok to use.

In this study, we focus on the model explanation rather than the modelling itself.

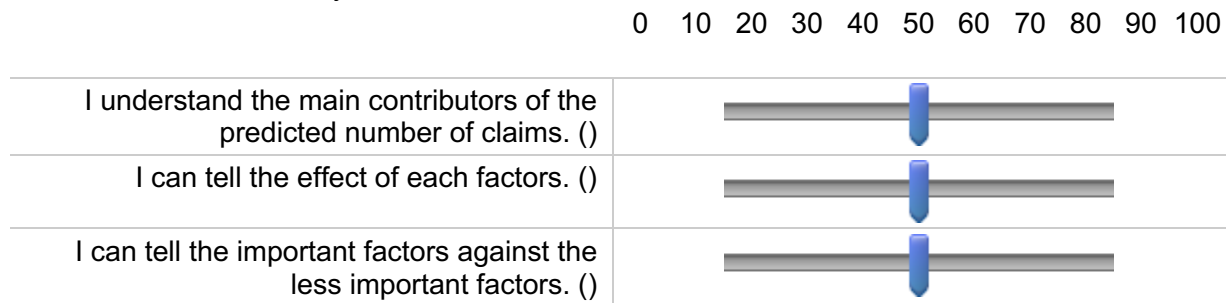
You are not required to take any notes. We will recall this estimation if it's relevant.

---

Initial perception Let's have a record of your first impression of the model.

Please rate: to what extent you agree with the statement. 0 means totally not agree, 100 means totally agree, 50 means neutral.

Remember, it's normal to be "not sure", and that's important input to this study! If feel you have no idea of the statement, you should rate it as 0.



Factor importance This is a permutation graph. We can treat it as the first tool to explain the fitted GLM.

If I randomly shuffle the value of a single risk factor in the data, for example, car\_power, leaving the claim number and all other factors in place, how would that affect the accuracy of predictions in that now-shuffled data?

In this graph, it is in order of influence power of each factors. So car\_power is of the biggest influence when we do this permutation test.

The next question will depend on your understanding of this piece of explanation.

**When the questions are presented, you will have 3 minutes to answer them. You can't change them once you submit.**

You don't need to take any notes. The graph will be recalled.

Page Break






Factor importance QD Let's have a record of your perception of the factor importance after the explanation.

The same permutation graph is here for your reference.

It's a simple true or false type of quiz with 3 minutes allowed.

You will be able to see submit button after 2 minutes.

Read the statement, if you think it's true, slide it to 1. If you think it's false, leave it as 0. If you don't know, tick the "NOT SURE" box.

	0	NOT SURE	1
Car age does not make a big difference to the result. ()			
Year does not make any difference to the result. ()			
Driver age does make a big difference to the result. ()			
Car power is the biggest contributor to the result. ()			
Car weight is the least important factor for the result. ()			

Q18 Timing  
First Click (1)  
Last Click (2)  
Page Submit (3)  
Click Count (4)

Page Break

SHAPQ16 Now we have another explanation tool. It's based on SHAP value.

SHAP values show how much a given factor changed our prediction comparing to if we made that prediction at some baseline value of that factor.

SHAP values can be positive or negative. SHAP value can be calculated based on one individual record.

We use mean absolute value for the graph below, for a global explanation. We also show SHAP value waterfalls for 3 random data points, so that you can validate whether it is true for a specific case.

The next question will depend on your understanding of this piece of explanation.

**When the questions are presented, you will have 3 minutes to answer them. You can't change them once you submit.**

You don't need to take any notes. The graph will be recalled.

One example on how to read the waterfall chart:






For the first data point, town = 1, and this represents a positive contribution to the result quantified by a SHAP value of 0.152.

-----  
Page Break

Q17 Let's have a record of your perception of the factor importance after the explanation.  
The same SHAP value graph is here for your reference.

**It's a simple true or false type of quiz with 3 minutes allowed.**  
**You will be able to see submit button after 1.5 minutes.**

Read the statement, if you think it's true, slide it to 1. If you think it's false, leave it as 0. If you don't know, tick the "NOT SURE" box.

	0	NOT SURE	1
Car age does not make a big difference to the result. ()			
Year does not make any difference to the result. ()			
Driver age does make a big difference to the result. ()			
Car power is the biggest contributor to the result. ()			
Car weight is the least important factor for the result. ()			

Q19 Timing  
First Click (1)  
Last Click (2)  
Page Submit (3)  
Click Count (4)

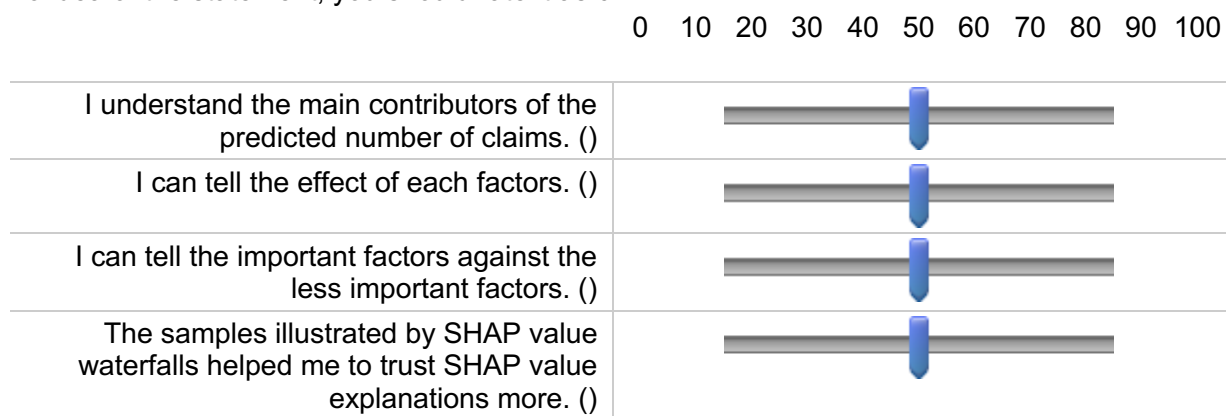
Page Break

Q20 We appreciate your hard work. It's half way of the journey!

Impression 2 Let's have a record of your refreshed impression of the model.

Please rate: to what extent you agree with the statement. 0 means totally not agree, 100 means totally agree, 50 means neutral.

Remember, it's normal to be "not sure", and that's important input to this study! If feel you have no idea of the statement, you should rate it as 0.



## ICE

Now, we will have a look at each and every risk factor.

The ICE(Individual Conditional Expectation) plots tell us that for each factor, how the instance's prediction changes when the value of this factor changes.

Example: For the plot for town(1st one), with all other factors fixed, the expected claim number(on a canonical scale) can change by a range from around 0.02 (i.e.2%) to around 0.06 (i.e. 6%) as the value of town changes from 0 to 1. Each line in the graph represents one record of the claims data.

We recall the data summary you've seen earlier, so that you know the possible range of value for each factor.

Next question will depend on your understanding of these graphs. Do spend a few minutes to read the graphs now. When the questions are presented, you will have 3 minutes to answer them. You can't change them once you submit.



Graphs will be recalled so you don't need to memorise anything.

---

Page Break

---




ICE-Q Let's have a record of your perception of the effect of the factors after the explanation.

The relevant ICE plots are here for your reference.

**It's a simple true or false type of quiz with 3 minutes allowed.**

**You will be able to see submit button after 2 minutes.**

Read the statement, if you think it's true, slide it to 1. If you think it's false, leave it as 0. If you don't know, tick the "NOT SURE" box.

	0	NOT SURE	1
As the car weight increases, it's likely to have more claims. ()			
As the car power increases, it's likely to have more claims. ()			
There might be more claims incurring from older drivers. ()			

Q26 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Page Break

SHAP Now let's look at an alternative SHAP-value based explanation tool.

The definition of SHAP value is still the same SHAP value you've seen earlier. Recall the content as below:

SHAP values show how much a given factor changed our prediction comparing to if we made that prediction at some baseline value of that factor. SHAP values can be positive or negative. SHAP value can be calculated based on one individual record.

This graph represents the SHAP value by feature value.

Example: For "town", darker colour means low value, which should be 0. Lighter colour means high value, which should be 1. For lighter colour, the corresponding SHAP value is around 0.13. It means positive effect. And it is closer to 0 comparing to the darker colour's SHAP value, which means less intense effect.

**Do spend a few minutes to read the graph now. When the questions are presented, you will have 3 minutes to answer them. You can't change them once you submit.**

Graphs will be recalled so you don't need to memorise anything.

---

Page Break




SHAP-Q Let's have a record of your perception of the effect of the factors after the explanation.

The same SHAP value by feature is here for your reference.

**It's a simple true or false type of quiz with 3 minutes allowed.**  
**You will be able to see submit button after 2 minutes.**

Read the statement, if you think it's true, slide it to 1. If you think it's false, leave it as 0. If you don't know, tick the "NOT SURE" box.

*Example recall: For "town", darker colour means low value, which should be 0. Lighter colour means high value, which should be 1. For lighter colour, the corresponding SHAP value is around 0.13. It means positive effect. And it is closer to 0 comparing to the lighter colour's SHAP value, which means less intense effect.*

	0	NOT SURE	1
As the car weight increases, it's likely to have more claims. ()			
As the car power increases, it's likely to have more claims. ()			
There might be more claims incurring from older drivers. ()			

Q29 Timing  
First Click (1)  
Last Click (2)  
Page Submit (3)  
Click Count (4)





Page Break

Q30 Nice! You've answered all timed questions! Last but not least, let's have a final record of your impression of the model after all explanations you've perceived.

---

Impression - final Please rate: to what extent you agree with the statement. 0 means totally not agree, 100 means totally agree, 50 means neutral.

Remember, it's normal to be "not sure", and that's important input to this study! If feel you have no idea of the statement, you should rate it as 0.

	0	10	20	30	40	50	60	70	80	90	100
I understand the main contributors of the predicted number of claims. ()											
I can tell the effect of each factors. ()											
I can tell the important factors against the less important factors. ()											
SHAP value by feature having all feature in one graph is favourable for less workload. ()											

End of Block: Default Question Block

---



## BIBLIOGRAPHY

- [1] Abdul-Malak, M. A. U., El-Saadi, M. M. & Abou-Zeid, M. G., 2002, 'Process model for administrating construction claims', *Journal of management in engineering*, vol. 18, no. 2, pp. 84–94.
- [2] Adamson, G., 2020, 'Explainable artificial intelligence (xai): A reason to believe?', *Law Context: A Socio-Legal J.*, vol. 37, p. 23.
- [3] Amer, M., Daim, T. U. & Jetter, A., 2013, 'A review of scenario planning', *Futures*, vol. 46, pp. 23–40.
- [4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. et al., 2020, 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', *Information fusion*, vol. 58, pp. 82–115.
- [5] Arumugam, M. & Cusick, K., 2008, 'General insurance 2020: insurance for the individual', *Sydney: Institute of Actuaries of Australia*.
- [6] Ashmore, R., Calinescu, R. & Paterson, C., 2021, 'Assuring the machine learning lifecycle: Desiderata, methods, and challenges', *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–39.
- [7] Australian Securities and Investments Commission, 2023, 'When the price is not right: Making good on insurance pricing promises', Report REP 765, Australian Securities and Investments Commission, <<https://download.asic.gov.au/media/lnxpj0uu/rep765-published-23-june-2023.pdf>>.
- [8] Bantounas, I., 2019, 'Actuarial models for estimating non life risks', .
- [9] Beckh, K., Müller, S., Jakobs, M., Toborek, V., Tan, H., Fischer, R., Welke, P., Houben, S. & von Rueden, L., 2021, 'Explainable machine learning with prior knowledge: an overview', *arXiv preprint arXiv:2105.10172*.
- [10] Belle, V. & Papantonis, I., 2021, 'Principles and practice of explainable machine learning', *Frontiers in big Data*, p. 39.
- [11] Bellis, C., 2014, 'Actuarial control cycle', *Wiley StatsRef: Statistics Reference Online*.
- [12] Benk, M., Weibel, R. & Ferrario, A., 2022, 'Creative uses of ai systems and their explanations: A case study from insurance', *arXiv preprint arXiv:2205.00931*.

- [13] Besacier, L. & Schwartz, L., 2015, 'Automated translation of a literary work: a pilot study', *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pp. 114–122.
- [14] Blier-Wong, C., Cossette, H., Lamontagne, L. & Marceau, E., 2020, 'Machine learning in p&c insurance: A review for pricing and reserving', *Risks*, vol. 9, no. 1, p. 4.
- [15] Blum, K. A. & Otto, D. J., 1998, 'Best estimate loss reserving: an actuarial perspective', *CAS Forum Fall*, , vol. 1p. 101.
- [16] Bolstad, C. & Endsley, M., 1990, 'Single versus dual scale range display investigation', *Hawthorne, CA: Northrop Corporation*.
- [17] Bora, A., Sah, R., Singh, A., Sharma, D. & Ranjan, R. K., 2022, 'Interpretation of machine learning models using xai-a study on health insurance dataset', *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, IEEE, pp. 1–6.
- [18] Boucher, J.-P. & Charpentier, A., 2014, 'General insurance pricing', *Computational actuarial science with R*, pp. 475–510.
- [19] Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C. & Detyniecki, M., 2021, 'Contextualising local explanations for non-expert users: An xai pricing interface for insurance', *Joint Proceedings of the ACM IUI 2021 Workshops*, .
- [20] Breiman, L., 2001, 'Random forests', *Machine learning*, vol. 45, pp. 5–32.
- [21] Brooke, J., 1986, 'System usability scale (sus): a quick-and-dirty method of system evaluation user information', *Reading, UK: Digital equipment co ltd*, vol. 43, pp. 1–7.
- [22] Campbell, M., Hoane Jr, A. J. & Hsu, F.-h., 2002, 'Deep blue', *Artificial intelligence*, vol. 134, no. 1-2, pp. 57–83.
- [23] Charpentier, A., 2016, *Computational Actuarial Science with R*, Chapman and Hall, Boca Raton, FL.
- [24] Chen, X., 2024, 'Algorithmic proxy discrimination and its regulations', *Computer Law & Security Review*, vol. 54, p. 106021.
- [25] Cheung, S. F., 2000, *Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices*, The Chinese University of Hong Kong (Hong Kong).
- [26] Cirqueira, D., Helfert, M. & Bezbradica, M., 2021, 'Towards design principles for user-centric explainable ai in fraud detection', *International Conference on Human-Computer Interaction*, Springer, pp. 21–40.
- [27] Cobbe, J. & Singh, J., 2021, 'Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges', *Computer Law & Security Review*, vol. 42, p. 105573.



- [28] Confalonieri, R., Coba, L., Wagner, B. & Besold, T. R., 2021, 'A historical perspective of explainable artificial intelligence', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391.
- [29] Conrad, A., Mostert, F. & Mostert, J., 2009, 'The underwriting process of motor vehicle insurance', *Corporate Ownership & Control*, vol. 6, no. 3.
- [30] David, M., 2015, 'A review of theoretical concepts and empirical literature of non-life insurance pricing', *Procedia Economics and Finance*, vol. 20, pp. 157–162.
- [31] de Villiers, J., Hobbs, K. & Hollebrandse, B., 2014, 'Recursive complements and propositional attitudes', *Recursion: Complexity in cognition*, pp. 221–242.
- [32] Delcaillau, D., Ly, A., Papp, A. & Vermet, F., 2022, 'Model transparency and interpretability: survey and application to the insurance industry', *European Actuarial Journal*, vol. 12, no. 2, pp. 443–484.
- [33] Demajo, L. M., Vella, V. & Dingli, A., 2020, 'Explainable ai for interpretable credit scoring', *arXiv preprint arXiv:2012.03749*.
- [34] Dictionary, M.-W., 2002, 'Merriam-webster', *On-line at <http://www.mw.com/home.htm>*, vol. 8, p. 2.
- [35] Doshi-Velez, F. & Kim, B., 2017, 'Towards a rigorous science of interpretable machine learning', *arXiv preprint arXiv:1702.08608*.
- [36] Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D. & Manning, C. A., 1998, 'Situation awareness as a predictor of performance for en route air traffic controllers', *Air Traffic Control Quarterly*, vol. 6, no. 1, pp. 1–20.
- [37] Eling, M., Nuessle, D. & Staubli, J., 2021, 'The impact of artificial intelligence along the insurance value chain and on the insurability of risks', *The Geneva Papers on Risk and Insurance-Issues and Practice*, pp. 1–37.
- [38] Endsley, M. R., 1993, 'A survey of situation awareness requirements in air-to-air combat fighters', *The International Journal of Aviation Psychology*, vol. 3, no. 2, pp. 157–168.
- [39] Endsley, M. R., 1995, 'Innovative model for situation awareness in dynamic defense systems', *Human Factors*, vol. 37, no. 1, pp. 32–64.
- [40] Endsley, M. R., 1995, 'Measurement of situation awareness in dynamic systems', *Human factors*, vol. 37, no. 1, pp. 65–84.
- [41] Endsley, M. R., 1995, 'Toward a theory of situation awareness in dynamic systems', *Human factors*, vol. 37, no. 1, pp. 32–64.
- [42] Endsley, M. R., 2001, 'A model of inter-and intrateam situational awareness: implications for design, training and measurement', *New trends in cooperative activities*, pp. 46–68.

- [43] Endsley, M. R., 2015, 'Situation awareness misconceptions and misunderstandings', *Journal of Cognitive Engineering and Decision Making*, vol. 9, no. 1, pp. 4–32.
- [44] Endsley, M. R., 2021, 'A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of sagat and spam', *Human factors*, vol. 63, no. 1, pp. 124–150.
- [45] Farley, T. C., Hansman, R. J., Amonlirdviman, K. & Endsley, M. R., 2000, 'Shared information between pilots and controllers in tactical air traffic control', *Journal of Guidance, Control, and Dynamics*, vol. 23, no. 5, pp. 826–836.
- [46] Fazel, S., Burghart, M., Fanshawe, T., Gil, S. D., Monahan, J. & Yu, R., 2022, 'The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies', *Journal of Criminal Justice*, vol. 81, p. 101902.
- [47] Fisher, A., Rudin, C. & Dominici, F., 2019, 'All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.', *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81.
- [48] Frees, E. W., Derrig, R. A. & Meyers, G., 2014, *Predictive modeling applications in actuarial science*, vol. 1, Cambridge University Press.
- [49] Friedman, J. H., 2001, 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, pp. 1189–1232.
- [50] Gaeta, A., Loia, V. & Orciuoli, F., 2021, 'A comprehensive model and computational methods to improve situation awareness in intelligence scenarios', *Applied Intelligence*, vol. 51, no. 9, pp. 6585–6608.
- [51] Gatzert, N. & Schmeiser, H., 2008, 'Combining fair pricing and capital requirements for non-life insurance companies', *Journal of Banking & Finance*, vol. 32, no. 12, pp. 2589–2596.
- [52] Goford, J., 1985, 'The control cycle: financial control of a life assurance company', *Journal of the Staple Inn Actuarial Society*, vol. 28, pp. 99–114.
- [53] Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E., 2015, 'Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation', *journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65.
- [54] Gramegna, A. & Giudici, P., 2020, 'Why to buy insurance? an explainable artificial intelligence approach', *Risks*, vol. 8, no. 4, p. 137.
- [55] Gustafsson, A. & Hansén, J., 2021, 'Combined actuarial neural networks in actuarial rate making', .
- [56] Hassani, H., Unger, S. & Beneki, C., 2020, 'Big data and actuarial science', *Big Data and Cognitive Computing*, vol. 4, no. 4, p. 40.

- [57] Holland, C. P., Mullins, M. & Cunneen, M., 2021, 'Creating ethics guidelines for artificial intelligence (ai) and big data analytics: The case of the european consumer insurance market', *Available at SSRN 3808207*.
- [58] Holzheu, T., Tamm, K., Lechner, R. & Fan, I., 2018, 'Profitability in non-life insurance: Mind the gap', .
- [59] Holzinger, A., Carrington, A. & Müller, H., 2020, 'Measuring the quality of explanations: the system causability scale (scs)', *KI-Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198.
- [60] JAMAL, A. Z. W. W., 2012, *Service Quality in General Insurance Industry*, Ph.D. thesis, UNIVERSITI TEKNIKAL MALAYSIA MELAKA.
- [61] Javidi, B., 2002, *Image recognition and classification: algorithms, systems, and applications*, CRC press.
- [62] Johne, A., 1993, 'Insurance product development: managing the changes', *International Journal of Bank Marketing*, vol. 11, no. 3, pp. 5–14.
- [63] Jones, D. G. & Endsley, M. R., 2000, 'Can real-time probes provide a valid measure of situation awareness', *Proceedings of the human performance, situation awareness and automation: user-centered design for the new millennium*, Savannah, GA.
- [64] Jovanovic, S., 2007, 'Non-life insurance premium', *Ins. L. Rev.*, p. 22.
- [65] Kelly, D. S. L. & Ball, M. L., 1991, *Principles of insurance law in Australia and New Zealand*, Butterworths.
- [66] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. et al., 2018, 'Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)', *International conference on machine learning*, PMLR, pp. 2668–2677.
- [67] Koster, O., Kosman, R. & Visser, J., 2021, 'A checklist for explainable ai in the insurance domain', *International Conference on the Quality of Information and Communications Technology*, Springer, pp. 446–456.
- [68] Kotsiurba, O., 2022, 'Actuarial calculations', .
- [69] Kovalenko, I., Davydenko, Y. & Shved, —., 2019, 'Development of the procedure for integrated application of scenario prediction methods', *East European Journal of Advanced Technologies*, , no. 2 (4), pp. 31–38.
- [70] Kshirsagar, R., Hsu, L.-Y., Greenberg, C. H., McClelland, M., Mohan, A., Shende, W., Tilmans, N. P., Guo, M., Chheda, A., Trotter, M. et al., 2021, 'Accurate and interpretable machine learning for transparent pricing of health insurance plans', *Proceedings of the AAAI Conference on Artificial Intelligence*, , vol. 35pp. 15127–15136.

- [71] Laña, I., Sanchez-Medina, J. J., Vlahogianni, E. I. & Del Ser, J., 2021, 'From data to actions in intelligent transportation systems: A prescription of functional requirements for model actionability', *Sensors*, vol. 21, no. 4, p. 1121.
- [72] Lin, Y.-S., Lee, W.-C. & Celik, Z. B., 2020, 'What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors', *arXiv preprint arXiv:2009.10639*.
- [73] Liquin, E., Callaway, F. & Lombrozo, T., 2020, 'Quantifying curiosity: A formal approach to dissociating causes of curiosity', *CogSci*, .
- [74] Lombrozo, T. & Liquin, E. G., 2023, 'Explanation is effective because it is selective', *Current Directions in Psychological Science*, vol. 32, no. 3, pp. 212–219.
- [75] Lorentzen, C. & Mayer, M., 2020, 'Peeking into the black box: An actuarial case study for interpretable machine learning', *Available at SSRN 3595944*.
- [76] Lozano-Murcia, C., Romero, F. P., Serrano-Guerrero, J. & Olivas, J. A., 2023, 'A comparison between explainable machine learning methods for classification and regression problems in the actuarial context', *Mathematics*, vol. 11, no. 14, p. 3088.
- [77] Lundberg, S. M. & Lee, S.-I., 2017, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, vol. 30.
- [78] Macedo, L., 2009, 'The role of the underwriter in insurance', *Primer Series on Insurance*, vol. 1, no. 8, pp. 13–29.
- [79] Mahalle, P. N., Bhapkar, H. R., Shinde, G. R., Sable, N. P. et al., 2023, 'Improving explainable ai interpretability: Mathematical models for evaluating explanation methods.', .
- [80] Mahlow, N. & Wagner, J., 2016, 'Evolution of strategic levers in insurance claims management: an industry survey', *Risk management and insurance review*, vol. 19, no. 2, pp. 197–223.
- [81] Maillart, A., 2021, *Some explainability methods for statistical learning models in actuarial science*, Ph.D. thesis, Université de Lyon.
- [82] Martin, K., Liret, A., Wiratunga, N., Owusu, G. & Kern, M., 2021, 'Evaluating explainability methods intended for multiple stakeholders', *KI-Künstliche intelligenz*, vol. 35, no. 3, pp. 397–411.
- [83] Mayer, M., Meier, D. & Wuthrich, M. V., 2023, 'Shap for actuaries: Explain any model', *Available at SSRN*.
- [84] Molnar, C., Casalicchio, G. & Bischl, B., 2020, 'Interpretable machine learning—a brief history, state-of-the-art and challenges', *Joint European conference on machine learning and knowledge discovery in databases*, Springer, pp. 417–431.
- [85] Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. & Geleijnse, G., 2021, 'Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival', *Scientific Reports*, vol. 11, no. 1, pp. 1–13.

- [86] Moreno-Sanchez, P. A., 2023, 'Improvement of a prediction model for heart failure survival through explainable artificial intelligence', *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1219586.
- [87] Murat, G., Tonkin, R. S. & Jüttner, D. J., 2002, 'Competition in the general insurance industry', *Zeitschrift für die gesamte Versicherungswissenschaft*, vol. 91, no. 3, pp. 453–481.
- [88] Neerincx, M. A., van der Waa, J., Kaptein, F. & van Diggelen, J., 2018, 'Using perceptual and cognitive explanations for enhanced human-agent team performance', *Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 15*, Springer, pp. 204–214.
- [89] Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L. & Nahavandi, S., 2019, 'A review of situation awareness assessment approaches in aviation environments', *IEEE Systems Journal*, vol. 13, no. 3, pp. 3590–3603.
- [90] Ohlsson, E. & Johansson, B., 2010, *Non-life insurance pricing with generalized linear models*, vol. 174, Springer.
- [91] Ohlsson, E. & Lauzenings, J., 2009, 'The one-year non-life insurance risk', *Insurance: Mathematics and Economics*, vol. 45, no. 2, pp. 203–208.
- [92] Orji, U. & Ukwandu, E., 2024, 'Machine learning for an explainable cost prediction of medical insurance', *Machine Learning with Applications*, vol. 15, p. 100516.
- [93] Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J. & Castignani, G., 2022, 'Explainable artificial intelligence (xai) in insurance', *Risks*, vol. 10, no. 12, p. 230.
- [94] Parodi, P., 2023, *Pricing in general insurance*, Chapman and Hall/CRC.
- [95] Patrick, J. & Morgan, P. L., 2010, 'Approaches to understanding, analysing and developing situation awareness', *Theoretical Issues in Ergonomics Science*, vol. 11, no. 1-2, pp. 41–57.
- [96] Pearson, K., 1895, 'Vii. note on regression and inheritance in the case of two parents', *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242.
- [97] Pichler, A., 2014, 'Insurance pricing under ambiguity', *European Actuarial Journal*, vol. 4, pp. 335–364.
- [98] Ribeiro, M. T., Singh, S. & Guestrin, C., 2016, 'Model-agnostic interpretability of machine learning', *arXiv preprint arXiv:1606.05386*.
- [99] Richman, R., 2018, 'Ai in actuarial science', *Available at SSRN 3218082*.
- [100] Rogan, P., 2021, *The Insurance and Reinsurance Law Review*, Law Business Research Limited.
- [101] Rose, J., Bearman, C., Naweed, A. & Dorrian, J., 2019, 'Proceed with caution: Using verbal protocol analysis to measure situation awareness', *Ergonomics*, vol. 62, no. 1, pp. 115–127.

- [102] Salmon, P. M., Stanton, N. A., Walker, G. H., Baber, C., Jenkins, D. P., McMaster, R. & Young, M. S., 2008, 'What really is going on? review of situation awareness models for individuals and teams', *Theoretical Issues in Ergonomics Science*, vol. 9, no. 4, pp. 297–323.
- [103] Sanneman, L. & Shah, J. A., 2020, 'A situation awareness-based framework for design and evaluation of explainable ai', *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, Springer, pp. 94–110.
- [104] Schmitt, M., 2024, 'Explainable automated machine learning for credit decisions: Enhancing human artificial intelligence collaboration in financial engineering', *arXiv preprint arXiv:2402.03806*.
- [105] Schotman, E. & Iren, D., 2022, 'Algorithmic decision making and model explainability preferences in the insurance industry: A delphi study', *2022 IEEE 24th Conference on Business Informatics (CBI)*, , vol. 1IEEE, pp. 235–242.
- [106] Schulz, C. M., Endsley, M. R., Kochs, E. F., Gelb, A. W. & Wagner, K. J., 2013, 'Situation awareness in anesthesia: concept and research', *The Journal of the American Society of Anesthesiologists*, vol. 118, no. 3, pp. 729–742.
- [107] Smith, K. & Hancock, P. A., 1995, 'Situation awareness is adaptive, externally directed consciousness', *Human factors*, vol. 37, no. 1, pp. 137–148.
- [108] Sreedharan, S., Srivastava, S. & Kambhampati, S., 2018, 'Hierarchical expertise level modeling for user specific contrastive explanations.', *IJCAI*, pp. 4829–4836.
- [109] Swartout, W. R., 1983, 'Xplain: A system for creating and explaining expert consulting programs', *Artificial intelligence*, vol. 21, no. 3, pp. 285–325.
- [110] Tomar, P., Sainy, M. & Gupta, R., 2019, 'Profitability analysis of insurance companies: A case of private non-life insurers', *UNNAYAN: International Bulletin of Management and Economics*, vol. 10, pp. 318–330.
- [111] Tooth, R., Li, W. & McWha, V., 2020, 'National insurance project–final report', *NSW*, vol. 507, p. 0001.
- [112] Tsanakas, A. & Desli, E., 2005, 'Measurement and pricing of risk in insurance markets', *Risk Analysis: An International Journal*, vol. 25, no. 6, pp. 1653–1668.
- [113] Van Stein, B., Vermetten, D., Caraffini, F. & Kononova, A. V., 2023, 'Deep bias: Detecting structural bias using explainable ai', *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pp. 455–458.
- [114] Vidogah, W. & Ndekugri, I., 1998, 'A review of the role of information technology in construction claims management', *Computers in industry*, vol. 35, no. 1, pp. 77–85.
- [115] Vilone, G. & Longo, L., 2021, 'Notions of explainability and evaluation approaches for explainable artificial intelligence', *Information Fusion*, vol. 76, pp. 89–106.

- [116] Wickens, C. D., 2008, 'Situation awareness: Review of mica endsley's 1995 articles on situation awareness theory and measurement', *Human factors*, vol. 50, no. 3, pp. 397–403.
- [117] Wollek, A., Graf, R., Čečátka, S., Fink, N., Willem, T., Sabel, B. O. & Lasser, T., 2023, 'Attention-based saliency maps improve interpretability of pneumothorax classification', *Radiology: Artificial Intelligence*, vol. 5, no. 2, p. e220187.
- [118] Wongsuwatt, S., Thaothampitak, W., Kongjam, N., Ruttanapibool, J., Apacuppakul, R. & Koedkaeo, T., 2020, 'The influence of loss ratio on profitability of non-life insurance companies in thailand: The moderating roles of company type', *Journal of Community Development Research (Humanities and Social Sciences)*, vol. 14, no. 1, pp. 46–60.
- [119] Wu, J., 2023, 'A feasible situation awareness-based evaluation framework for introducing explainable ai to claims management', <https://actuaries.logicaldoc.cloud/download-ticket?ticketId=e38b0297-05b6-4966-ab2a-69b11bbd1151>, accessed: 2023-05-22.
- [120] Xenidis, R. & Senden, L., 2020, *EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination*, Kluwer Law International.
- [121] Zahi, J. et al., 2021, 'Non-life insurance ratemaking techniques', *International Journal of Accounting, Finance, Auditing, Management and Economics*, vol. 2, no. 1, pp. 344–361.
- [122] Zhou, J., Gandomi, A. H., Chen, F. & Holzinger, A., 2021, 'Evaluating the quality of machine learning explanations: A survey on methods and metrics', *Electronics*, vol. 10, no. 5, p. 593.