MSc Business Information Technology (BIT)

Beyond Generalized Linear Models: Advancing Insurance Pricing through Interpretable and Explainable Machine Learning

Author: Jens Reil

Supervisor(s): Marcos R. Machado, João R. Moreira

September, 2024

Faculty of Behavioural, Management and Social Sciences (BMS)
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
University of Twente

Acknowledgements

This master's thesis in the actuarial science domain was completed in September, 2024. This was the final part of the Master of Science (MSc) in Business Information Technology (BIT). It also represents the culmination of my time as a student. With all the unexpected occurrences, new acquaintances, and (nearly) limitless freedom, these years have been fantastic. I genuinely cherished these experiences, yet I am eager and thrilled to embark on my professional career. I express my gratitude to you, the reader, for dedicating your time to read this report.

I would like to extend my appreciation to Triple A - Risk Finance for their support and collaboration. Their depth of knowledge and insightful contributions have enriched the understanding of the topic, providing valuable perspectives that have significantly enhanced the overall quality of this thesis. I am particularly grateful to Axel Pison, my primary supervisor from the company, for the hours he dedicated to refining my thesis and for offering valuable perspectives on the topic. His assistance was crucial in shaping the paper into its current form. Additionally, I would like to express my personal gratitude to Pieter Stel for facilitating my entry into the company and assisting me in collecting data from a real-world context, which significantly enriched the thesis. Lastly, I want to thank all my other colleagues, specifically the Data Analytics Team, that made my experience at Triple A - Risk Finance one to never forget.

Secondly, I would like to express my sincere appreciation to the team at Turien & Co., especially Arno Schipper, Bram van Empelen, Björn Jalving, and the Data Management team. Providing real-world data made this thesis far more engaging, as it gave me the sense that my contributions could truly make a difference. Arno's guidance, support, and collaboration has been invaluable throughout the process. His expertise and dedication made a significant impact, for which I am truly grateful.

Support did not only come from the companies mentioned. Specifically, I would like to express my sincere gratitude to Marcos R. Machado for his invaluable supervision and guidance throughout the course of thesis project. His expertise, encouragement, and quick constructive feedback have been instrumental in shaping the quality and direction of this paper. The meetings were always insightful and enjoyable, exactly the way I envision ideal supervision. I would also like to extend my personal appreciation to João R. Moreira for his role as a second supervisor in enhancing the quality of this paper.

While my thesis was my main focus in the past months and my studies were the focus for the past years, the external activities happening outside of my academic development were the most important to me. I want to thank my parents, my sister, all my friends, and everyone who helped me, for all the support, sincere advise, good talks, and the great adventures.

Jens Reil
Enschede, September, 2024

Abstract

This thesis explores the integration of machine learning (ML) techniques into car insurance premium pricing, traditionally handled by generalized linear models (GLMs). While GLMs offer transparency and meet regulatory requirements, the potential for enhanced predictive accuracy has made the insurance industry to explore ML alternatives. This research examines whether ML can improve premium pricing predictions while maintaining the interpretability and explainability essential for industry adoption. The study applies various ML algorithms, including LightGBM, XGBoost, and a Multi-Layer Perceptron (MLP), comparing them against traditional GLMs using various performance metrics for both severity and frequency data. To address concerns about the explainability and interpretability of ML models, explainable artificial intelligence (XAI) techniques such as SHAP and LIME are employed. The results indicate that ML models outperform GLMs in predictive power for both severity and frequency claims data. However, ensuring model explainability, especially in regulatory and compliance contexts, remains a challenge, as highlighted through interviews with industry stakeholders. The research proposes a hybrid approach, where ML models complement GLMs, combining their strengths to improve accuracy without sacrificing transparency. This study contributes to the existing body of research on AI-driven insurance pricing, advocating for a gradual integration of ML models supported by explainability tools such as SHAP.

Keywords: (Non-life) Insurance Pricing, Artificial Intelligence (AI), Machine Learning (ML), Explainable Artificial Intelligence (XAI), SHapely Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME)

Contents

Acknowledgements

Abstract

List of Acronyms

1 Introduction **1**
- 1.1 Company Overview . . . 2
  - 1.1.1 Triple A - Risk Finance . . . 2
  - 1.1.2 Turien & Co. . . . 3
- 1.2 Problem Identification . . . 4
- 1.3 Motivation . . . 5
- 1.4 Research Questions . . . 5
- 1.5 Research Method . . . 6
- 1.6 Thesis Structure . . . 6

2 Context Description **7**
- 2.1 Fundamentals of Car Insurance . . . 7
  - 2.1.1 About Car Insurance . . . 7
  - 2.1.2 Policy Claim Types . . . 7
  - 2.1.3 Policy Coverage Structures . . . 8
- 2.2 Insurance Pricing . . . 9
  - 2.2.1 Customer Premium Determination . . . 9
  - 2.2.2 Pure Premium Determination . . . 10
    - 2.2.2.1 Frequency & Claim Severity . . . 10
    - 2.2.2.2 Frequency-Severity Model Principle . . . 11
- 2.3 Industry Model Standard in Insurance Pricing . . . 12
  - 2.3.1 Generalized Linear Models (GLMs) . . . 12
  - 2.3.2 Limitations of GLMs . . . 12
- 2.4 Adoption of Machine Learning (ML) in Insurance Pricing . . . 13
  - 2.4.1 Transition from GLMs to ML . . . 13
  - 2.4.2 Current State of ML Applications . . . 14
  - 2.4.3 Challenges for the Application of ML . . . 15
- 2.5 Explainable Artificial Intelligence in Insurance Pricing . . . 16
  - 2.5.1 About Explainable Artificial Intelligence (XAI) . . . 16
    - 2.5.1.1 Establishing Common Understanding of XAI Terminology . . . 17
    - 2.5.1.2 The Trade-Off between Accuracy and Interpretability . . . 18
  - 2.5.2 Taxonomy of XAI Approaches . . . 19
    - 2.5.2.1 Transparent Models & Post-Hoc Explainability . . . 19

2.5.2.2 Model-Specific & Model-Agnostic Explanation Methods . . 20
2.5.2.3 Global & Local Explanation Methods 21
2.5.3 Integration of XAI in Insurance Pricing 22
2.6 Regulatory Considerations in AI-Driven Insurance Practices 23
2.6.1 Regulatory Frameworks Governing AI in Insurance 23
2.6.1.1 General Data Protection Regulation (GDPR) 24
2.6.1.2 Artificial Intelligence Act (AI Act) 24
2.6.2 Regulatory Consequences for Insurance Companies & End Consumers 25
2.7 Summary 26

3 Methodology **28**
3.1 CRISP-DM 28
3.2 Machine Learning Techniques 29
3.2.1 Generalized Linear Model (GLM) 29
3.2.2 Light Gradient Boosting (LightGBM) 31
3.2.3 Extreme Gradient Boosting (XGBoost) 32
3.2.4 Multi-Layer Perceptron Neural Network (MLP) 34
3.3 XAI Methods 35
3.3.1 SHapely Additive exPlanations (SHAP) 36
3.3.2 Local Interpretable Model-agnostic Explanations (LIME) 37
3.4 Validation Methods 38
3.4.1 Evaluation Metrics 38
3.4.2 Cross-Validation (CV) 40
3.4.3 Semi-Structured Interviews 40
3.5 Summary 41

4 Experimental Setup **42**
4.1 Approach 42
4.1.1 Experimental Framework 42
4.1.2 Coding Environment 42
4.2 Business Understanding 43
4.2.1 The Three Lines of Defence (3LoD) Model 43
4.2.2 Process Stakeholders 44
4.3 Data Understanding 45
4.3.1 Naming Conventions 46
4.3.2 Data Confidentiality 46
4.3.3 Data Exploration 47
4.3.3.1 Severity Data 47
4.3.3.2 Frequency Data 51
4.4 Data Preparation 54
4.4.1 Cleaning 54
4.4.2 Train-Test Split 56
4.4.3 Encoding 56
4.4.4 Scaling 57
4.5 Model Exploration 57
4.5.1 H2O AutoML: Automatic Machine Learning 57
4.5.2 AutoML: Severity Model 58
4.5.3 AutoML: Frequency Model 59
4.6 Modeling 60
4.6.1 Baseline GLM & ML Models 60

4.6.2 Hyper-Parameter Tuning . . . 61
4.6.3 Model Validation . . . 61
4.6.3.1 Metrics . . . 62
4.6.3.2 Actual Predicted Performance (APP) Plots . . . 62
4.7 Evaluation . . . 63
4.7.1 XAI Techniques . . . 63
4.7.2 Interviews . . . 63
4.7.2.1 Approach . . . 63
4.7.2.2 Sample Size . . . 64
4.7.2.3 Process Overview . . . 64
4.8 Summary . . . 66

5 Results & Discussion **67**
5.1 Model . . . 67
5.1.1 Performance . . . 67
5.1.1.1 Severity . . . 68
5.1.1.2 Frequency . . . 69
5.1.2 Validation . . . 70
5.1.2.1 Severity . . . 71
5.1.2.2 Frequency . . . 72
5.2 Explanations . . . 74
5.2.1 Severity . . . 74
5.2.1.1 Global Explanations: SHAP Beeswarm & Bar Plots . . . 74
5.2.1.2 Local Explanations: SHAP Waterfall & Force Plots, LIME 77
5.2.2 Frequency . . . 80
5.2.2.1 Global Explanations: SHAP Beeswarm & Bar Plots . . . 80
5.2.2.2 Local Explanations: SHAP Waterfall & Force Plots, LIME 83
5.3 Interview Evaluation . . . 84
5.4 Summary . . . 87

6 Conclusion **88**

A Additional Context Description **100**
A.1 Application of GLMs . . . 100

B Additional Methodology **103**
B.1 CRISP-DM & CRISP-ML . . . 103
B.2 Example SHAP: Global & Local Explanations . . . 104
B.2.1 Example Beeswarm & Bar Plots . . . 104
B.2.2 Example Waterfall & Force Plots . . . 106
B.3 Example LIME: Local Explanation . . . 107

C Additional Experimental Setup **108**
C.1 Framework (Roadmap) . . . 108

D Additional Results & Discussion **109**
D.1 Model Validation . . . 109
D.1.1 APP Plots Severity . . . 109
D.1.2 APP Plots Frequency . . . 112
D.2 Model Evaluation . . . 116

D.2.1 Severity Evaluation . . . 116
D.2.1.1 Beeswarm Plots . . . 116
D.2.1.2 Waterfall Plots . . . 117
D.2.1.3 Force Plots . . . 118
D.2.1.4 LIME Plots . . . 119
D.2.2 Frequency Evaluation . . . 120
D.2.2.1 Beeswarm Plots . . . 120
D.2.2.2 Waterfall Plots . . . 121
D.2.2.3 Force Plots . . . 122
D.2.2.4 LIME Plots . . . 123

E Interviews **124**
E.1 Exploration Interview . . . 124
E.2 Validation Interviews . . . 126
E.2.1 Interview #1: Individual A . . . 128
E.2.2 Interview #2: Individual B . . . 129
E.2.3 Interview #3: Individual C . . . 130
E.2.4 Interview #4: Individual D . . . 132
E.2.5 Interview #5: Individual E . . . 133

List of Figures

1.1 Company overview of Triple A - Risk Finance, an independent and innovative consultancy company, and some of its clients [14]. . . . 3
1.2 Company overview of Turien & Co., both an insurer as well as an insurance underwriter [15]. . . . 4

2.1 Scope of Explainable Artificial Intelligence (XAI) [55]. . . . 17
2.2 Visualizing the trade-off between model accuracy and model interpretability, and a representation of the area of improvement where the potential of XAI techniques and tools resides [52]. . . . 18
2.3 Conceptual diagram illustrating various post-hoc explainability approaches for a ML model M_φ [52]. . . . 20
2.4 Summary of XAI challenges its impact on the principles for Responsible AI [52]. . . . 23
2.5 The AI Act defines four levels of risk for AI systems [66]. . . . 25

3.1 CRISP-DM, a framework for data mining projects [22]. . . . 28
3.2 Illustration of a generalized linear regression model, showcasing the relationship between a predictor variable x and response variable y [70]. . . . 30
3.3 LightGBM, which stands for Light Gradient Boosting Machine, is a distributed, high-performance implementation of the gradient boosting framework. [72]. . . . 32
3.4 XGBoost, which stands for Extreme Gradient Boosting, is an advanced implementation of the gradient boosting machine (GBM) algorithm [78]. . . . 33
3.5 MLP, which stands for Multi-Layer Perceptron, falls under the category of feedforward algorithms, which is type of NN [82]. . . . 35
3.6 5-Fold iteration CV [102]. . . . 40
3.7 An overview of structured, semi-structured, and unstructured interviews [106]. As visualized, semi-structured interviews are a combination of both structured and unstructured interview concepts. . . . 41

4.1 The three lines of defence (3LoD) model visualized [111]. . . . 43
4.2 An overview of the process for a model based on the 3LoD model and information provided by `the insurance company`. For this thesis, one could look at this visualization for a premium pricing model deployment process. This is a conceptual representation, making this potentially differ compared to other companies. . . . 44
4.3 An overview of the claim (severity) data over the years at `the insurance company`. . . . 48
4.4 An overview of the distribution of claim amounts over the years at `the insurance company`. . . . 48

4.5 An overview of the distribution of claim amounts over the years at the insurance company, focused on the gross number of claims. 49
4.6 An overview of the various car brands within the the insurance company's portfolio. 49
4.7 Boxplots of selected features for the reparation severity dataset at the insurance company. 51
4.8 An overview of the distribution of fuel type of cars over the years at the insurance company. 52
4.9 Boxplots of selected features for the reparation frequency dataset at the insurance company. 53
4.10 An overview of the NAs with the provided datasets at the insurance company. 54
4.11 The different hyper-parameter tuning methods visualized (GS, RS, and BO) [118]. 61

5.1 APP plots for the four models for Feature 170 for the reparation severity dataset at the insurance company, where the distribution of the feature bins is provided as well. 71
5.2 APP plots for the four models for Feature 162 for the reparation frequency dataset at the insurance company, where the distribution of the feature bins is provided as well. 73
5.3 Beeswarm plot made for the (best-performing) MLP model based on the severity data of the insurance company, showing global feature contributions for the respective features within the model. 75
5.4 Bar plots made for the four models based on the severity data of the insurance company, showing global feature contributions based on SHAP values for the respective features within the model. 76
5.5 Waterfall plot made for the MLP model based on the severity data of the insurance company, showing local feature contributions based on SHAP values for the respective features within the model. 78
5.6 Force plot made for the MLP model based on the severity data of the insurance company, showing local feature contributions based on SHAP values for the respective features within the model. 79
5.7 Example of a LIME explanation for the MLP model based on the severity data of the insurance company, showing local feature contributions for the respective features within the model. 80
5.8 Beeswarm plot made for the XGBoost model based on the frequency data of the insurance company, showing global explanations for the respective features within the model. 81
5.9 Bar plots made for the four models based on the frequency data of the insurance company, showing global feature contributions based on SHAP values for the respective features within the model. 82
5.10 Waterfall plot made for the XGBoost model based on the frequency data of the insurance company, showing local feature contributions based on SHAP values for the respective features within the model. 83
5.11 Force plot made for the XGBoost model based on the frequency data of the insurance company, showing local feature contributions based on SHAP values for the respective features within the model. 84

A.1 Overview of both Poisson and Gamma statistics of the case study [20]. . . . 101

A.2 An analysis of parameter estimates of the case study [20]. 102

B.1 An overview of the CRISP-DM and CRISP-ML methodology respectively, showing the added *Monitoring and Maintenance* step [132]. 104
B.2 Example: Global explanation plots with SHAP based on *The Boston Housing Dataset* [92]. 105
B.3 Example: Local explanation plots with SHAP based on *The Boston Housing Dataset* [92]. 106
B.4 Example: LIME output of a local observation. For this example, the *Rain in Australia Dataset* from Kaggle was used [133]. 107

C.1 A roadmap for the project based on CRISP-DM, detailing all the phases conducted within the thesis. 108

D.1 APP plots for the four models for `Feature 62` for the reparation severity dataset at `the insurance company`. 109
D.2 APP plots for the four models for `Feature 88` for the reparation severity dataset at `the insurance company`. 110
D.3 APP plots for the four models for `Feature 6` for the reparation severity dataset at `the insurance company`. 110
D.4 APP plots for the four models for `Feature 171` for the reparation severity dataset at `the insurance company`. 111
D.5 APP plots for the four models for `Feature 145` for the reparation frequency dataset at `the insurance company`. 112
D.6 APP plots for the four models for `Feature 165` for the reparation frequency dataset at `the insurance company`. 113
D.7 APP plots for the four models for `Feature 168` for the reparation frequency dataset at `the insurance company`. 113
D.8 APP plots for the four models for `Feature 54` for the reparation frequency dataset at `the insurance company`. 114
D.9 APP plots for the four models for `Feature 76` for the reparation frequency dataset at `the insurance company`. 114
D.10 APP plots for the four models for `Feature 172` for the reparation frequency dataset at `the insurance company`. 115
D.11 Beeswarm plots for the four models for the reparation severity dataset at `the insurance company`. 116
D.12 Force plots for the ML models for the reparation severity dataset at `the insurance company`. 117
D.13 Force plots for the ML models for the reparation severity dataset at `the insurance company`. 118
D.14 LIME local explanation plots for the ML models for the reparation severity dataset at `the insurance company`. 119
D.15 Beeswarm plots for the four models for the reparation frequency dataset at `the insurance company`. 120
D.16 Waterfall plots for the ML models for the reparation frequency dataset at `the insurance company`. 121
D.17 Force plots for the ML models for the reparation frequency dataset at `the insurance company`. 122
D.18 LIME local explanation plots for the ML models for the reparation frequency dataset at `the insurance company`. 123

List of Tables

2.1 Car insurance policy structures [26]. 8
2.2 Terms in Explainable Artificial Intelligence (XAI) [52]. 18

4.1 Mapping the taxonomies of claim causes (identified by Section 2.1.2) with the naming conventions used by `the insurance company`. Multiple "..."'s within a column mean that this cause type was divided into multiple datasets. 46
4.2 Descriptive statistics for some of the features within the reparation severity dataset at `the insurance company`. 50
4.3 Descriptive statistics for some of the features within the reparation frequency dataset at `the insurance company`. 53
4.4 Row information of the reparation dataset for severity and frequency after determining the selected data (SD) gathered from the initial full dataset (ID), where also the size of the dataset after the removal of the year 2024 is indicated (YR). 55
4.5 Final row and column information of the datasets per claim type for severity and frequency before cleaning (BC) and after cleaning (AC). 55
4.6 Model performances of the reparation severity dataset from `the insurance company` found via AutoML. *Model ID* is abbreviated due to space issues. Values of interest are highlighted in blue. 58
4.7 Model performances of the reparation frequency dataset from `the insurance company` found via AutoML. *Model ID* is abbreviated due to space issues. Values of interest are highlighted in blue. 59
4.8 An overview of the various interviews conducted for this thesis, including essential background information about each interviewee and the context of the interviews. 65

5.1 An overview of the results of the various ML techniques for the reparation severity dataset. 68
5.2 An overview of the results of the various ML techniques for the reparation frequency dataset, indicating the best values with a green color. 69

B.1 CRISP-DM process model descriptions [22]. 103

List of Acronyms

3LoD	Three Lines of Defence
AI	Artificial Intelligence
APP	Actual Predicted Performance
AutoML	Automatic Machine Learning
BO	Bayesian Optimization
CART	Classification And Regression Tree
CRISP-DM	Cross Industry Standard Process for Data Mining
CRISP-ML	Cross Industry Standard Process for Machine Learning
CV	Cross-Validation
DL	Deep Learning
DRF	Distributed Random Forest
DT	Decision Tree
DNN	Deep Neural Network
EV	Explained Variance
GAM	Generalized Additive Model
GBM	Gradient Boosting Machine
GDPR	General Data Protection Regulation
GLM	Generalized Linear Model
ICE	Individual Conditional Expectation
IQR	Interquartile Range
LE	Label Encoding
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LM	Linear Model

LR	Linear Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMS	Min-Max Scaling
MAE	Mean Absolute Error
MRM	Model Risk Management
MSE	Mean Squared Error
NA	Not Available
NLP	Natural Language Processing
NN	Neural Network
OHE	One-Hot Encoding
PDP	Partial Dependence Plot
$\mathbf{R}^2$	R-Squared: Coefficient of Determination
RF	Random Forest
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
RMSLE	Root Mean Squared Logarithmic Error
RS	Robust Scaling
SHAP	SHapely Additive exPlanations
SS	Standard Scaling
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
XRT	Extremely Randomized Trees

Chapter 1

Introduction

The insurance industry comprises companies that provide risk management through insurance policies [1]. The fundamental principle of insurance is that the insurer agrees to compensate for losses from specific uncertain future events. In return, the insured or policyholder pays a periodic fee, known as a premium, to the insurer for this coverage against potential risks. The primary goal of insurance companies is to safeguard the property rights of entities who enter into insurance contracts, protecting them in the case of insurable events [2]. While insurance operations inherently involve managing policyholder risks, the unique aspect of the insurance business is that these companies also shoulder their own risks. These risks include potential financial losses due to economic downturns, ineffective investment strategies like shifts in security interest rates, and the high likelihood of losses coming from asset value fluctuations and the aggregation of liabilities.

In an insurance portfolio of an insurance company, the risk levels of the policyholders vary. To account for this diversity in risk, insurance companies adjust premiums according to the risk profiles of the policyholders. This means that those with higher risk profiles pay more for their insurance than those with lower risks [3]. The way these risk levels of certain policyholders are determined is called pure premium pricing. The pure premium, or risk premium, is mainly composed of the combination of the effects of frequency and average cost for each risk group or insured and the selected guarantee [4].

In actuarial pricing, the possible risk of providing clients with insurance is evaluated, and the price ranges that allow for this risk and yet turn a profit are determined [5]. For the past three decades, pricing processes within the insurance industry have seen little change [6]. Traditionally, actuarial pricing relied heavily on statistical models and mathematical frameworks to quantify and manage risks [7]. Actuaries utilize these mathematical models to estimate the likelihood of events to ensure insurance companies can set aside sufficient funds for potential claims. For instance, analyzing mortality rates among specific age groups aids insurers in predicting when they might need to disburse for life insurance policies [8].

However, in recent years, the integration of artificial intelligence (AI) techniques, such as machine learning (ML)[1], has emerged as a transformative force within insurance pricing [6]. This advancement provides actuaries and pricing teams with the tools to enhance decision-making, allowing for quicker and more effective choices. Modern insurers opt for ML for various reasons, such as efficiently handling and analyzing the growing volume of diverse data from claims, customer interactions, market trends, and advancing technologies, and improving speed and accuracy in tasks like the internal processing of incoming claims

[1] ML is a subset of AI that automatically enables a machine or system to learn and improve from experience [9].

and underwriting [10]. ML could help insurers assess claims more quickly and accurately by identifying patterns in large datasets, which could ensure fairer payouts by improving risk assessment and premium pricing.

With the rapid adoption of AI in the insurance industry, regulators and unions are proactively seeking ways to enhance the trustworthiness of models through ongoing monitoring and the enforcement of compliance measures [11]. The integration of ML techniques in insurance pricing models poses a multifaceted challenge, encompassing technical nuances, the impact of explainable AI on transparency, and the critical intersection of evolving technologies with regulatory compliance [12]. Initiatives like the EU's General Data Protection Regulation (GDPR) and the recent Artificial Intelligence Act (AI Act) underscore the necessity of transparency in industries like insurance where ML plays a pivotal role [12]. Addressing these complexities is crucial to optimize model selection, ensure regulatory adherence, and foster a holistic understanding of the challenges and opportunities in this rapidly evolving space.

1.1 Company Overview

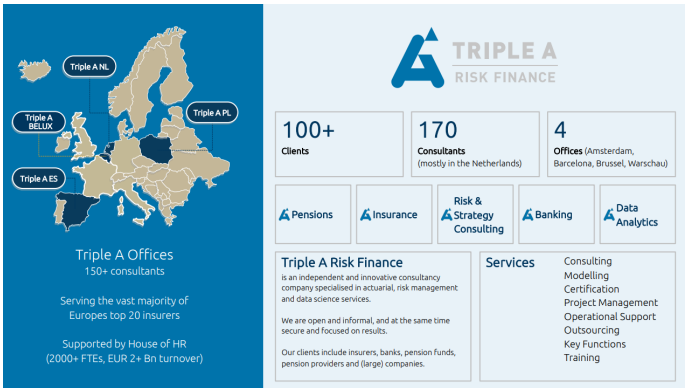
This section provides a comprehensive overview of the companies that were stakeholders with the thesis. Understanding the different roles of these companies within the context of this thesis is essential for understanding their contribution towards the final outcome of the study. In this case, the focus is specifically on two firms: Triple A - Risk Finance[2] and Turien & Co.[3], both of which play significant roles in their respective areas of expertise within financial (insurance) services.

In the following sections, information about both companies is provided, as well as its specific role within this thesis. This section aims to provide an understanding about both firms and which role these companies fulfilled.

1.1.1 Triple A - Risk Finance

Triple A - Risk Finance is an independent and innovative consultancy firm, distinguished by its specialized focus on actuarial services, risk management, and data science. In 2014, the company joined forces with Talent&Pro, the largest secondment provider in the financial services sector [13]. This resulted in Redmore Group where Triple A and Talent&Pro operate independently and serve clients together [13]. Triple A - Risk Finance operates a network of offices strategically located across various European regions, such as the Netherlands, Belgium, Spain, and Poland. With over 150 consultants within its organisation, the firm employs a significant workforce capable of providing a diverse range of services across various business lines, such as pensions, insurance, risk & strategy consulting, banking, and data analytics.

[2] https://aaa-riskfinance.nl
[3] https://turien.nl



(a) Company overview of Triple A - Risk Finance.

(b) Client overview of Triple A - Risk Finance.

Figure 1.1: Company overview of Triple A - Risk Finance, an independent and innovative consultancy company, and some of its clients [14].

The services provided by Triple A - Risk Finance are specialized in various areas, such as consulting, modelling, certification, and project management, as shown in Figure 1.1a. The firm also extends its services to operational support, outsourced solutions, key function services, and training, thereby providing an all-encompassing suite of services designed to fit to the various dimensions of risk finance. Triple A - Risk Finance works for over 100 different clients, ranging from insurers, pension funds, pension administrators, banks, and other large-scale corporations. This shows the wide-reaching influence of the company across multiple financial sectors. The organization boasts a considerable footprint in the European consultancy sector, prominently serving the vast majority of Europe's top 20 insurers, as can be seen in Figure 1.1b.

The topics of this thesis are explored in collaboration with the department (business line) *Data Analytics* within Triple A - Risk Finance, comprising a total of 14 employees. The goal of the Data Analytics team is to deliver value to the respective client organisation by making better decisions after a thorough analysis of available data. The involvement of Triple A - Risk Finance in the thesis will ensure that the findings are practical and directly applicable to real-world business challenges. Furthermore, it offers a unique opportunity to apply academic concepts in a professional setting, thereby validating theoretical frameworks with empirical data and contributing to the advancement of knowledge in actuarial science. From now on, Triple A - Risk Finance will be referenced as `the consultancy firm` due to anonymization and confidentiality within the thesis.

1.1.2 Turien & Co.

Turien & Co. is a multifaceted company operating both as an insurer and an insurance underwriter [15]. The company distinguishes itself through a combination of Dutch and international mandates, and it also manages Ansvar, a sustainable and personal insurance company. This unique setup allows Turien & Co. to be flexible and robust in the insurance market, often adopting an independent approach. Turien & Co. focuses on providing fast, high-quality, and distinctive insurance solutions and services [15]. The company aims for product leadership and utilizes both internal and external risk carriers to offer the best products possible, making it a reliable partner for insurance advisors, with a strong emphasis on personal contact and long-term relationships.

In the annual report for 2023, the premium auto segment remained the largest segment

for Turien & Co. (Ansvar), with a portfolio premium of €76 million, accounting for approximately 38% of the total portfolio premium [16].

(a) Company logo of Turien & Co.

(b) Turien & Co. holds the largest position in car insurance within its own portfolio [17].

Figure 1.2: Company overview of Turien & Co., both an insurer as well as an insurance underwriter [15].

Turien & Co. Assurandeuren have provided a part of their car insurance data for the research of this paper, playing an important role in the thesis by supplying the necessary data, positioning them as a stakeholder in the research project. Their involvement is essential not only for the access to accurate and comprehensive data but also for the insight and expertise they offer regarding the insurance industry. This will enhance the overall quality and credibility of the research findings. Additionally, Turien & Co.'s participation may also allow for a practical evaluation of theoretical concepts within a real-world context, thereby bridging the gap between academic research and industry practice. This collaborative approach will significantly contribute to the relevance and applicability of the thesis subject. From now on, Turien & Co. will be referenced as `the insurance company` due to anonymization and confidentiality within the thesis.

1.2 Problem Identification

The integration of ML into insurance pricing presents a promising yet challenging assignment. Traditionally, insurance companies have relied on generalized linear models (GLMs) (see Section 2.3.1 for more information) for pricing policies. While GLMs offer robust predictability and simplicity, these traditional statistical models may not fully capture complex patterns in data as effectively as more advanced ML techniques [18]. This thesis seeks to explore the potential of ML to enhance the pricing models beyond what GLMs can offer.

This thesis addresses the lack of practical, implementable solutions that apply ML to insurance pricing models, with a specific focus on improving explainability and interpretability of these models. Furthermore, there is a lack of comprehensive frameworks that address the integration of ML models that separately and jointly consider frequency and severity of policy (insurance) claims, as GLMs are still considered to be industry standard [19]. The industry is searching for innovation within the pricing of insurance premiums, and ML could potentially be a solution. This problem identification sets the stage for exploring how ML can be tailored to meet the demands of the insurance industry, focusing on overcoming the challenges associated with replacing traditional GLM approaches.

1.3 Motivation

Currently, most insurers are using GLM approaches to determine insurance premiums [19]. The main reasons for using GLMs is that the pricing models should be explainable to the different stakeholders, simple to program, and adaptable to marketing demands and benchmark studies where competitors are analyzed. This is why more advanced models such as ML models are not yet the industry standard within insurance pricing.

There is a rising interest for insurance firms as well as for the field of actuarial science in the adoption of AI in insurance. According to literature, there are possibilities for the implementation of ML models in insurance pricing based on the frequency-severity model approach [20]. The goal would be to determine insurance premiums more accurately, while still being able to interpret the results found. The research objective tries to change the approach to insurance premium pricing from using traditional GLM approaches to the usage of ML, while doing this in such a way that interpretability an explainability are taken into account for stakeholders[4]. Achieving this goal will hopefully lead to a more accurate insurance pricing model for insurance products.

1.4 Research Questions

The central focus of this thesis is to explore the integration of ML techniques in insurance pricing as a replacement for the current industry standard Generalized Linear Models (GLMs), with a particular emphasis on looking into interpretability and ensuring explainability. This investigation also aims to develop validation method that supports this transition. The main research question for this thesis is:

"How can ML algorithms be integrated into insurance pricing to replace traditional statistical models (GLMs) in such a way that interpretability and explainability are taken into account?"

To comprehensively address the main research question, this thesis will explore several subquestions that highlight various aspects of the transition from traditional GLMs to ML-based models in insurance pricing, which are mentioned below.

1. *How does a GLM work, and how is explainability ensured in this context?*
   GLMs are the current industry standard in insurance pricing techniques [19]. This question seeks to delve into the operational mechanisms of GLMs and their application in the insurance domain, focusing on their inherent interpretability and explainability. Special attention will be given to the frequency-severity model principle, which represents a method for premium determination in the literature [21].
2. *How do ML compare to GLM approaches and how can ML models be integrated into the insurance pricing process?*
   Despite the theoretical research into the integration of ML in insurance pricing, practical applications remain limited with GLMs still predominating [19]. This subquestion will explore potential ML approaches that can be defined and applied to model frequency and severity.
3. *How can one validate the explainability outcomes through the usage of XAI methods in ML-based insurance pricing models?*

[4]Stakeholders within the project can be defined as insurers (responsible for making insurance policies available to clients), actuaries (responsible for applying the insurance policies), and data scientists (responsible for creating the insurance policies). This is further elaborated in Section 4.2

This question addresses the need for evaluation methods to represent the trustworthiness of explanations provided by ML models in insurance pricing. Given the sparse and often uncritical existing literature on explainable artificial intelligence (XAI) in this field, this subquestion aims to look into possibilities for assessing explainability, ensuring that outcomes from these models are neither blindly accepted nor misunderstood.

By addressing these subquestions, the thesis aims to contribute valuable insights to the field of insurance pricing, adding research for more reliable and transparent use of ML technologies.

1.5 Research Method

The thesis is based on the CRISP-DM methodology, which is elaborated upon in Section 3.1. The CRISP-DM (*CR*oss *I*ndustry *S*tandard *P*rocess for *D*ata *M*ining) project defines a process model which provides a framework for carrying out data mining projects which is independent of both the industry sector and the technology used [22].

The CRISP-DM framework is an established methodology for conducting data mining projects, detailed in the guide by Schröer et al [23]. It outlines six sequential phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase involves specific tasks, which will be handled more elaborately in Section 3.1.

1.6 Thesis Structure

Chapter 1 begins by providing an overview of the problem context for this thesis. Following this, Chapter 2 presents the literature review, focusing on car insurance, its policy structures, and the current methods used in insurance pricing. The chapter emphasizes the industry-standard GLMs and explores the challenges associated with integrating ML into insurance pricing. Additionally, it discusses various studies that employ GLMs and ML for determining risk premiums. Chapter 3 outlines the CRISP-DM framework used in the research, covering phases such as business understanding, data understanding, data preparation, modeling, and evaluation. The chapter addresses various ML techniques and explainable algorithms (XAI) that could be used within the thesis, as well as several validation methods used. Chapter 4 describes the approach to data handling, model exploration, and validation. Then, Chapter 5 presents the model performance outcomes and validation through both technical analysis, explanation plots (for model evaluation), and interviews. Finally, the thesis concludes with Chapter 6, summarizing the findings, discussing limitations, and offering suggestions for future research.

Chapter 2

Context Description

2.1 Fundamentals of Car Insurance

Insurance can be categorized into life and non-life. Non-life insurance covers any other risk except for life-risk of the person injured. Life insurance covers only the life-risk of the person insured [24]. According to the Dutch Motor Insurance Liability Act (Dutch: "Wet aansprakelijkheidsverzekering motorrijtuigen (WAM)"), car insurance is classified under the non-life insurance category [25]. As this thesis will handle car insurance data retrieved by `the insurance company` (Section 1.1.2), this section will outline the policy structures of (non-life) car insurance in accordance with Dutch law.

2.1.1 About Car Insurance

In the Netherlands, car insurance has been mandatory since the Wet Aansprakelijkheidsverzekering Motorrijtuigen (WAM) came into effect in 1963 [26]. This law requires all car owners to have at least third-party liability insurance, known as WA insurance (further evaluated in Section 2.1.3). This basic insurance covers damages to third parties only and is the least expensive option. Additionally, car owners can opt for more comprehensive coverage such as the WA+ or allrisk insurance policy structures. The allrisk insurance, as explained in Section 2.1.3, covers everything included in the WA+ policy plan plus damages caused by the car owner.

2.1.2 Policy Claim Types

Various types of car insurance claims can be made to cover the costs associated with vehicle issues, as detailed below along with their respective descriptions [26]:

1. *Damage to others.* This claim type addresses any damages you may cause to other people's property or personal injuries inflicted on others in an accident. It includes compensation for medical expenses, vehicle repairs, and property damage that you are legally responsible for.
2. *Theft and break-ins.* This coverage handles the loss or damage of your vehicle resulting from theft or unauthorized entry. It includes not only the theft of the vehicle itself but also damage sustained during a break-in and the loss of personal items or car accessories that were stolen.
3. *Window damage.* This type of claim covers the costs associated with repairing or replacing your car's windows if they are cracked, shattered, or otherwise damaged.

It typically includes damage from accidents, vandalism, or environmental factors like falling debris.

4. *Weather / fire damage.* This claim type provides coverage for damage to your vehicle caused by natural events such as severe weather conditions or fire. It includes incidents like hailstorms, floods, lightning strikes, and wildfires, which can cause significant damage to your car.
5. *Damage caused by yourself.* This coverage addresses any damage to your own vehicle resulting from accidents or incidents that you are responsible for. It includes collisions with other vehicles or objects, as well as accidents caused by losing control of the vehicle.
6. *Vandalism.* This type of claim covers damages inflicted on your vehicle through intentional acts of destruction or defacement by others. It includes actions such as smashing windows, or other forms of deliberate damage.

Like many other countries, the Netherlands employs a no-claims bonus system (known as "Bonus-malus") for car insurance [27]. Under this system, drivers can receive increasingly significant discounts on their insurance premiums for every year they do not file a claim. While the specifics can vary among different insurers, it is possible for drivers in the Netherlands to achieve up to an 80% reduction in their annual insurance costs after 10-15 years of claim-free driving [27].

2.1.3 Policy Coverage Structures

In the Netherlands, there are three primary types of car insurance available to vehicle owners [26]. These three policy coverage structures are summarized in Table 2.1 and will be further evaluated in this section. The claim types mentioned below are explained in Section 2.1.2.

Table 2.1: Car insurance policy structures [26].

Claim Type	WA	WA+	Allrisk
Damage to others	X	X	X
Theft and break-ins		X	X
Window damage		X	X
Weather / fire damage		X	X
Damage caused by yourself			X
Vandalism			X

The first and most basic is the WA (Dutch: "Wettelijk aansprakelijkheid") insurance, which is mandatory for every car owner [25]. This coverage ensures that you are insured for damages caused to third parties, making it the least comprehensive and consequently the most affordable option [26]. For those seeking additional coverage, the WA+ insurance (in Dutch also known as: "Beperkt Casco") or limited comprehensive coverage, expands beyond mere third-party liability. This plan covers damages to your own vehicle as well, including window damage, theft or break-in damage, and damages resulting from storms or fires. The most comprehensive option available is the Allrisk (Dutch: "Volledig Casco") insurance, generally recommended for new cars [26]. This type of insurance covers all the aspects of the limited comprehensive plan and also includes compensation for damages to

your own car that you may have caused. This makes it the most inclusive car insurance option, offering extensive protection for your vehicle.

The cost of the policy commercial premium (also known as customer premium, handled in Section 2.2.1) is influenced by several factors including the type of coverage one seeks, the age, where one lives, the type of car one drives, the annual distance covered, and specific driving habits. Typically, the more accidents one is involved in, the higher the premium will be.

2.2 Insurance Pricing

Insurance pricing is a crucial aspect of the insurance industry, representing the determination of premium amounts that policyholders pay in exchange for coverage against potential risks [3]. The context of insurance pricing is deeply rooted in the fundamental principles of risk management and financial stability. Actuaries play a important role in this context, employing historical data, statistical models, and other relevant factors to estimate the likelihood and severity of potential losses. This risk assessment (estimating the likelihood that a policyholder will file a claim and the potential severity of that claim) is fundamental in determining the appropriate premium amounts that policyholders must pay for their insurance coverage. The following subsections will explore the concepts of claim frequency and severity, as well as the principles behind the frequency-severity model, which can be utilized for understanding how often claims are made and the financial impact of each claim is essential for predicting the total cost that insurers may face.

2.2.1 Customer Premium Determination

Customer premium determination is an important aspect of the insurance industry. It involves assessing individual risk factors to set the price of an insurance policy. This process is not just about covering potential claims but also about accurately reflecting the likelihood of a claim being made. Insurers analyze various data points, such as the customer's history, the type of coverage sought, and industry-wide risk assessments, to establish a fair premium [28]. By doing so, they balance the need to be competitive with the necessity of maintaining financial solvency to support claims, ensuring that the premium aligns with the risk level and the value of the policy provided to the customer. A typical insurance premium for a customer is made up of three different components, such as the company premium[1], natural disaster premium[2], and government levies and taxes[3] [28].

Ideally, the premium should have a layered structure. By splitting the premium into various components, trends in customer claims behavior, costs/overheads, and commercial considerations can be determined and monitored independently of each other. In this thesis, one can assume the following premium structure, which is based on the expertise by the `the consultancy firm`:

1. *Pure premium* or *risk premium*. Outcome of the technical premium model per policy, differentiated by customer characteristics (postal code, age, etc.). Determined in the

[1]Company premiums fund expected claim costs of the insurance premium, administrative expenses such as staff and technology, and commissions for intermediaries like agents and brokers.

[2]Natural disaster premiums are made to secure reinsurance for insurance for large-scale natural catastrophes, as these events incur the highest costs for insurers.

[3]Government levies and taxes are taxes one has to pay to the government. In the Netherlands, for example, insurance premium tax (or insurance tax) is a tax on insurance premiums [29].

technical premium model (GLM, see Section 2.3) based on expected frequency and claims cost.
2. *Gross premium.* Includes the risk premium defined in the previous point plus profit margin, expenses, underwriting commissions, commissions, profit sharing, and any trends. The gross premium reflects the total costs per policy. These are the bottom-up determined premiums.
3. *Commercial premium.* Further differentiation of the gross premium within predefined ranges to optimize returns, inflow and outflow, and strategy. Commercial strategy could be differentiated by the respective distribution channel.

In this thesis, an emphasis is placed on the expected claim costs, commonly referred to as the *pure premium* or *risk premium*. The pure premium represents the cost that an insurance company anticipates will be paid out for claims, excluding any additional expenses such as overheads or profits [30]. This will be further handled in Section 2.2.2.

2.2.2 Pure Premium Determination

Actuaries are responsible for calculating the pure premium, or risk premium, which is the fundamental cost that insurance companies associate with covering potential losses for policyholders. It represents the core expense of compensating policyholders for their claims, excluding any overhead or profit margins for the insurer. Once determined, the pure premium serves as the starting point for insurers to set premium rates. Insurers add their desired profit margins and account for administrative expenses to ensure financial sustainability [30].

2.2.2.1 Frequency & Claim Severity

Many insurance datasets feature information about how often claims arise, the frequency. In addition to this claim size, the severity of a claim is featured. Observable responses include [31]:

- N, the number of claims (frequency). The claim frequency ("frequency" in Equation 2.1) is the number of claims divided by the duration[4], for some group of policies in force during a specific time period, i.e., the average number of claims per time period (usually one year). Claim frequency is assumed to be distributed according to a Poisson distribution or Negative Binomial distribution [32].
- $y_k = 1, ..., N,$, the amount of each claim (severity). The claim severity ("conditional severity" in Equation 2.1) is the total claim amount divided by the number of claims, i.e., the average cost per claim. The claim severity tends to follow a Gamma distribution. The Gamma distribution is a popular choice due to its parameter interpretability [31]. Looking into the reality, these distributions, which belong to a family of exponential distribution, are reasonable. Another example of a severity distribution is the Inverse Gaussian distribution [32].
- $S = y_1 + ... + y_N$, the aggregate claim amount.

The aggregate claim amount (S) is the key element for an insurer's balance sheet, as it represents the amount of money paid on claims. This would pose the question why it would be relevant for the insurance companies to track frequencies (N) and claim amounts (y_k) as well. A review by Frees et al. segments these reasons into four categories [31]:

[4] $Frequency = \#_{Claims}/Duration_{Policy}$

1. *Features of contracts.* In the insurance industry, deductibles and policy limits are commonly applied on a per occurrence and per contract basis.
2. *Policyholder behavior and risk mitigation.* The distinction between claim frequency (occurrence rate) and severity (financial impact) is significant. For instance, covariates influencing frequency might include personal characteristics, while those affecting severity could be related to external factors.
3. *Databases that insurers maintain.* Insurers often maintain separate databases for policyholders and claims, enabling distinct modeling of frequency and severity. Regulators require reporting of both claim numbers and amounts.
4. *Regulatory requirements.* Insurers may use different administrative systems for handling frequent, small losses versus rare, high-impact events, with claims frequency playing a crucial role in determining expenses.

Also, in practice, insurers create a compound (joint) model that integrates both frequency and severity models to predict the total claim cost, called the frequency-severity model. This approach involves first predicting the number of claims using a frequency model and then estimating the cost of each claim with a severity model. The overall expected cost is the aggregation of the severity estimates, weighted by the probability distribution of the frequency model. This compound approach allows for a more nuanced understanding of risk, accommodating the inherent variability and uncertainty in both the frequency and severity of the claims. The frequency-severity model is elaborated upon in the Section 2.2.2.2.

For simplicity, it is also common to use only the aggregate loss as a dependent variable in a regression [31]. Because the distribution of the aggregrate loss typically contains a positive mass at zero representing no claims, and a continuous component for positive values representing the amount of a claim, a widely used mixture is the Tweedie distribution [31] [32]. Tweedie distributions correspond to compound Poisson-Gamma distributions, that is, compound Poisson sums with Gamma-distributed summands [33]. However, in this thesis, focus lies on using separate distributions for both frequency and severity calculations, combining them according to the frequency-severity model principle ($Frequency_{Claim} * Severity_{Claim} = Loss_{Total}$). This principle is elaborated upon in Section 2.2.2.2.

2.2.2.2 Frequency-Severity Model Principle

As discussed in Section 2.2.2.1, in practice, insurers often adopt a compound (joint) model that integrates both frequency and severity models. For modeling the joint outcome (N, S) (or equivalently, $(N, \bar{S})$)[5], it is customary to initially condition on the frequency and subsequently model the severity [31]. The study breaks down the distribution of the dependent variables as follows (see Equation 2.1). Through this breakdown, the frequency and severity components don't need to be independent [31].

$$\begin{aligned} f(N,S) &= f(N) * f(S|N) \\ joint &= frequency\ *\ conditional\ severity \end{aligned} \tag{2.1}$$

Here, $f(N, S)$ denotes the joint distribution of (N, S).

The distribution of both frequency and severity models play a crucial role in insurance pricing, as these models help insurers predict the number of claims (frequency) and the

[5]The notation $\bar{S}$ stands for the average claim amount (defined to be 0 when $N = 0$).

cost of these claims (severity) more accurately [31] [32]. Understanding and selecting appropriate distributions for these models are important to accurately estimating risk and setting premiums.

2.3 Industry Model Standard in Insurance Pricing

GLMs are statistical models and were formulated by Nelder and Wederburn in 1972, and become standard industry practice for insurance pricing in 1990s and are still the standard in insurance pricing [34]. However, it was not until the second half of the 90's that the use of GLMs really started spreading, partly in response to the extended needs for tariff analysis due to the deregulation of the insurance markets in many countries [35].

2.3.1 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) is a class of statistical methods [35]. GLMs allow modeling a non-linear behavior and a non-Gaussian distribution of residuals [20]. This aspect is very useful for the analysis of non-life insurance, where claim frequency and claim cost follow an asymmetric density that is non-Gaussian. It generalizes the ordinary linear models (LMs) in two directions [35]:

1. *Probability distribution.* Instead of assuming the normal distribution, GLMs work with a general class of distributions, which contains a number of discrete and continuous distributions as special cases, in particular the Normal, Poisson and Gamma distributions.
2. *Model for the mean.* In linear models the mean is a linear function of the covariates x. In GLMs some monotone transformation of the mean is a linear function of the x's, with the linear and multiplicative models as special cases.

A key disadvantage of an ordinary LM is that its range is from $-\infty$ to ∞, which is not suitable for non-negative values only, such as claim severity and frequency. To address this issue, a GLM with an appropriate distribution, like the Gamma family, can be used. The development of GLMs has contributed to quality improvement of the risk prediction models and to the process of establishing a fair tariff or premium given the nature of the risk [20]. One strength of the GLM approach is its versatility, as the same set of routines can be applied seamlessly to model both continuous and discrete outcomes [31]. This adaptability makes GLMs a powerful tool for insurance modeling, where the outcomes can vary widely in nature. Moreover, the strength of the linear exponential family, which serves as the foundation for a GLM, lies in the fact that a sample average of outcomes follows the same distribution as the outcomes themselves [31]. This property enhances the reliability of predictions, ensuring that the modeling framework accurately captures the characteristics of both frequency and severity outcomes. The ability to handle diverse types of data and maintain consistency in distributional assumptions further solidifies the GLM approach as a robust and widely employed method in insurance analytics [31].

2.3.2 Limitations of GLMs

GLMs, while robust in their own right, have limitations when it comes to handling the complexities and nuances of modern insurance data [36]. Understanding these constraints is crucial for effectively applying GLMs and interpreting their results.

A fundamental limitation of GLMs is their assumption of complete data credibility [32]. In practice, this means the model fully trusts the data provided for each parameter, without accounting for potential data scarcity or unreliability. For instance, if a GLM is used to predict auto insurance claim severity based on territorial divisions with varying claim volumes, it might overly rely on sparse data from less common territories. This can lead to predictions that are overly specific to the small sample sizes available for these territories, potentially skewing the model's overall effectiveness. Another critical limitation within GLMs is that the randomness of outcomes across the dataset is uncorrelated [32]. This assumption does not preclude the existence of correlations driven by the model's predictors but asserts that the random errors, those variations not explained by the model, are uncorrelated. This can become problematic in scenarios where underlying, unmodeled factors lead to correlated outcomes.

However, the insurance landscape is evolving, with increasingly complex variables and interactions. ML techniques can better capture these interactions [18], as discussed in Section 2.4.1.

2.4 Adoption of Machine Learning (ML) in Insurance Pricing

In the ever-evolving landscape of the insurance industry, technological advancements have become catalysts for transformative changes. Among these innovations, ML stands out as a key player reshaping traditional practices and introducing unprecedented efficiencies [37]. These advanced technologies are reshaping the way insurers assess risk, determine premiums, and interact with policyholders. The application of AI in insurance pricing not only enhances accuracy but also promotes fairness, transparency, and a more customer-centric approach [38].

The adoption of AI in insurance underwriting, in particular ML, is proving to be important for improved risk management, fair pricing, and increased customer loyalty [39]. This technological evolution not only ensures profitable operations for insurers but also enhances efficiency, encourages upselling and cross-selling, and fortifies customer relationships [39]. The transitions to AI underscores the insurance industry's recognition of ML's potential to revolutionize pricing strategies, risk assessment, and customer service.

2.4.1 Transition from GLMs to ML

In response to the limitations of conventional frequency-severity models discussed in Section 2.2.2.2, the insurance industry has increasingly turned to ML algorithms for more flexible and accurate modeling [38]. Unlike GLMs, ML algorithms such as random forests, gradient boosting machines, and neural networks, offer the advantage of capturing complex and nonlinear relationships within data [18]. These algorithms excel in handling intricate interactions among feature variables, making them well-suited for scenarios where traditional models like GLMs and GAMs face challenges such as the capturing of intricate interaction among its feature variables, as discussed. ML, however, is capable of automatically learning patterns and relationships from the data without relying on predefined linear structures, making them better at capturing non-linear effects in data [36].

Furthermore, the use of ML allows for the exploration of diverse sets of predictors and features, enabling the incorporation of a wide range of variables that might exhibit nonlinear effects on the response variable, such as claim severity [36]. This flexibility is particularly beneficial when dealing with complex insurance datasets, where relationships between variables may not adhere to linear assumptions. Additionally, ML models can

adapt and evolve over time, continuously improving their predictive performance as they are exposed to new data [36]. Adaptability is crucial in dynamic insurance environments where risk factors and customer behaviors may change over time. By continuously learning from new data, ML models contribute to more informed decision-making, improved risk management, and thus an enhanced overall performance in insurance pricing.

2.4.2 Current State of ML Applications

The insurance sector, historically conservative in adopting technological innovations, is now embracing ML to address complex challenges in pricing policies, risk management, and fraud detection [40]. ML algorithms, including decision trees (DTs), neural networks (NNs), and ensemble methods like random forests (RFs) and gradient boosting machines (GBMs), are increasingly being applied to predict claims frequency and severity, thereby refining pricing accuracy and competitiveness. Grize et al. highlight the significant advantages of ML over traditional GLMs, noting the superior predictive power of ML models in capturing complex nonlinear relationships and interactions among variables [37].

The current state of ML applications in insurance pricing showcases a diverse range of techniques aimed at enhancing various aspects of the (non-life) insurance sector. Cunha & Bravo found that, in the investigation of claim prediction models, ML is applied [18]. Actuarial science benefits from ML by incorporating big data to enhance risk assessment, policy pricing, and claims management. Various algorithms such as k-nearest neighbors, k-means clustering, kernel regression, boosting machines, regression trees, neural networks, and many other algorithms are applied to analyze historical insurance data [41].

The application of telematics data for personalized insurance pricing illustrates ML's capability to transform the industry. Telematics, with its ability to capture real-time driving behavior and patterns, has become a game-changer, enabling insurers to move away from traditional, static premium calculations [42]. The incorporation of these innovative methodologies, such as data-driven binning methods, stands out as a pivotal aspect of this transformative journey [43]. By analyzing driving behavior, insurers can offer customized premiums, rewarding safe drivers with lower rates. This not only enhances risk assessment but also encourages safer driving habits among policyholders. Techniques like logistic regression (LR), support vector machines (SVM), RF, extreme gradient boosting (XGBoost), and artificial NNs are among the sophisticated tools leveraged to navigate the intricacies of telematics data [42]. Additionally, algorithmic advancements in variable selection for claims frequency modeling using telematics car driving data are introduced in a paper by Chan et al., focusing on handling overdispersion [44]. The approach involves generating sub-samples of data corresponding to each component of the Poisson mixture model (FLEXMIX). These models not only enhance the accuracy of risk evaluation but also contribute to the efficiency of pricing strategies. In essence, the combination of telematics-driven insights and cutting-edge modeling approaches marks a significant departure from traditional auto insurance practices [44]. These innovative approaches such as telematics data for more accurate insurance claims forecasting can also be altered by incorporating weather conditions and car sales as additional variables [38]. A practical example of the integration of telematics data in car insurance is "Fairzekering", which is a Dutch AI-enabled car insurance product [45]. The insurance company analyses their customers' driving behaviour based on the data collected by a smart module 'Chipin' that is installed in their customers' cars. Depending on your risk score, you get a 0-3% discount on your monthly rate. Customers also have access to a dashboard that provides them with additional insights in their driving behaviour [45].

In auto insurance, there is a growing emphasis on the development of ratemaking frame-

works tailored for usage-based insurance (UBI) products, particularly harnessing the power of information provided by telematics driving data [42]. By leveraging telematics data from connected vehicles, insurers can monitor and analyze driving behavior in real time. This granular level of insight enables more precise risk assessment and personalized pricing based on individual driving habits [42]. Safe drivers may benefit from lower premiums, while those with riskier behaviors may face higher costs. This strategic shift underscores a broader industry trend towards more dynamic and personalized pricing structures.

The application of ML extends to various other branches within insurance as well, like lapse risk management in motor insurance. Here, algorithms like LR, classification and regression trees (CART), SVM, and XGBoost predict policyholders' likelihood of lapsing [46]. These innovative approaches such as telematics data for more accurate insurance claims forecasting can also be altered by incorporating weather conditions and car sales as additional variables [38]. In the case of Poufinas et al., SVM, RF, and XGBoost algorithms are utilized to predict the number of insurance claims [38].

For handling over-dispersed insurance frequency data in the area of motor insurance, a sophisticated actuarial modeling approach has been introduced, according to Lee [21]. This strategic evolution reflects an effort within the non-life insurance sector to develop methodologies that can more effectively capture the nuances of risk inherent in the frequency of insurance events. The application of generalized linear models and generalized additive models (GAMs) forms a foundational element of this advanced actuarial modeling approach [21]. These techniques, rooted in statistical modeling, enable insurers to capture the complexity of over-dispersed data by accounting for non-linear relationships and variations in risk factors. The incorporation of gradient boosting and delta boosting further enhances the predictive capabilities of these models [21].

The ML techniques used within the study by Lee (various boosting algorithms) excel at iteratively refining predictions by learning from the errors of previous iterations, thereby improving the overall accuracy of the risk assessments. The adaptive nature of boosting algorithms aligns well with the dynamic nature of insurance data, allowing for a more responsive and agile risk modeling framework [21]. The utilization of negative binomial regression within the study underscores a keen awareness of the unique distributional characteristics associated with over-dispersed data. This regression technique, tailored for count data exhibiting over-dispersion, provides a more accurate reflection of the underlying risk structure compared to traditional Poisson regression [21].

2.4.3 Challenges for the Application of ML

The adoption of ML in insurance pricing is not without challenges. Issues of model interpretability, data privacy, and regulatory compliance pose significant hurdles as stakeholders may find it challenging to comprehend the decision-making processes of sophisticated "black-box" algorithms [47]. With regard to execution times, the greater complexity of the models imply longer execution times compared to the GLMs [48]. Striking a balance between model complexity and interpretability is a key consideration in ensuring that pricing strategies are not only accurate but also comprehensible to regulatory bodies, customers, and industry professionals.

To address the challenges of interpretability and transparency in insurance pricing models, an emerging solution is the integration of Explainable Artificial Intelligence (XAI) techniques. By incorporating XAI methods, insurance companies can offer stakeholders, including regulatory bodies, customers, and industry professionals, insights into how the model arrives at specific pricing decisions [47]. The possible solution of using explainable algorithms is addressed in Section 2.5.

However, the use of complex ML models and explainable algorithms in insurance can lead to several disadvantages. A significant hurdle is regulatory compliance. The insurance industry is heavily regulated, and as ML models evolve, so too do the regulations governing their use. Insurers must stay abreast of changes concerning algorithmic transparency and the ethical use of data, which can vary widely across different jurisdictions. Failure to comply with these evolving regulations can expose companies to legal and financial risks. Insurance analytics already uses personal data to optimise front- and back-end operations, risk modelling and risk pricing [49]. This is due to insurance analytics providing important value in fraud management, claims management and better managing risk pooling by creating more accurate behavioural profiles of insureds [50] [51]. An example of the usage of personal data in the insurance industry in the UK is provided in Section 2.6.

Moreover, the technical complexity of these systems cannot be underestimated. Managing and updating ML models to ensure their accuracy and effectiveness over time requires a robust IT infrastructure. Additionally, integrating XAI into these models not only increases the complexity but also demands continuous training and upskilling of staff. This ongoing need for technical proficiency puts further strain on insurers, potentially leading to operational inefficiencies.

2.5 Explainable Artificial Intelligence in Insurance Pricing

XAI is becoming increasingly crucial in the realm of insurance pricing as insurers seek to balance the benefits of advanced ML models with the need for transparency and accountability [47]. In the context of insurance pricing, where accurate risk assessment is essential, the ability to explain how ML models arrive at specific pricing decisions becomes essential for stakeholders. If insurers can gain more insight into how ML models behave in risk classification contexts, it would increase their ability to reassure regulators and the public that accepted rate making principles are met [47].

2.5.1 About Explainable Artificial Intelligence (XAI)

The need for XAI arises because AI models, such as deep neural networks, are often considered "black boxes" because they can be extremely complex and difficult to understand [52]. This lack of transparency can be a problem, particularly in scenarios where AI is making decisions that affect people's lives, such as insurance [53]. In the view of Fritz et al., fairness is a pre-condition to develop and run algorithms of which the decisions can be trusted [54].

XAI aims to provide clear and understandable insights into the decision-making processes of complex algorithms, addressing the "black-box" nature of many ML models [47]. It wants to create AI systems that can be easily understood and interpreted by humans. XAI proposes creating a suite of ML techniques that produce more explainable models while maintaining a high level of learning performance, and enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners [52]. This is important because as AI becomes more prevalent in our daily lives, it is crucial that one can trust and understand the decisions made by these systems. To cite a passage given by Arrietta et al. [52]:

> *"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."*

Thus, the term Explainable Artificial Intelligence refers to an explanatory agent revealing underlying causes to its or another agent's decision-making [55]. It is important to realize that the solution to explainable AI is not to just have more AI. Ultimately, it can be seen as a human-agent interaction problem. Human-agent interaction can be defined as the intersection of artificial intelligence, social science, and human-computer interaction in which XAI is just one problem within human-agent interaction, as can be seen in Figure 2.1.

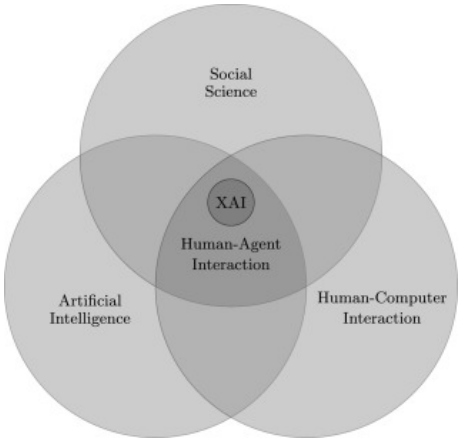


Figure 2.1: Scope of Explainable Artificial Intelligence (XAI) [55].

2.5.1.1 Establishing Common Understanding of XAI Terminology

One of the barriers to establishing common understanding within the domain of explainable artificial intelligence is the frequent confusion and incorrect use of terms such as explainability and interpretability in literature's terminology. These concepts have certain fundamental differences, which will be explored in this section.

In the area of XAI, there are several terms used to describe how people can understand and work with AI models, such as interpretability and explainabiliy, but also understandability, comprehensability, and transparency [52]. Interpretability is essentially a model's passive characteristic, indicating how comprehensible the model is to a human onlooker. Interpretability is about whether a person can understand what a model is doing and why, aiming for clear explanations of the model's decisions. On the other hand, explainability is making sure the reasons behind a model's decisions are not just clear, but also detailed and useful for humans. Explainability represents a model's active quality, signifying any effort or method it employs to make its internal functions more understandable or transparent [52]. It aims to answer questions like "Why did the AI system make this particular prediction?". Understandability means how easily a person can get how a model works without needing to look into its complex inner parts or how it deals with data. Comprehensibility is about how a learning algorithm can show its findings in a way that's easy for people to understand, suggesting that the insights from a computer should be similar to what a human expert would come up with, in both meaning and structure. Transparency refers to how naturally clear a model is. However, various proposals have been made for a unified theory of explanation, but none have successfully withstood critique [52]. All these various terminologies within XAI have been summarized in Table 2.2.

Table 2.2: Terms in Explainable Artificial Intelligence (XAI) [52].

Term	Definition
Interpretability	Ability to understand what a model is doing and why, aiming for clear explanations.
Explainability	Providing detailed and useful reasons behind a model's decisions.
Understandability	Ease with which a person can comprehend how a model works without delving into its complex inner parts.
Comprehensibility	Ability of a learning algorithm to present findings in a way understandable to people, akin to human expert insights.
Transparency	Clarity of a model's inner workings.

In the world of AI, there is no consensus on the definitions of interpretability and explainability [52]. Despite this lack of clarity, numerous efforts have been made to create interpretable models and techniques that aim to enhance explainability. It appears from the literature that there is not yet a common point of understanding on what interpretability or explainability are. However, many contributions claim the achievement of interpretable models and techniques that empower explainability [52].

2.5.1.2 The Trade-Off between Accuracy and Interpretability

The relationship between model accuracy and interpretability is a well-known trade-off in the field of ML and AI, particularly in the context of XAI. This trade-off is graphically depicted in Figure 2.2, illustrating a clear inverse relationship: as model complexity and accuracy increase, interpretability tends to decrease [52].

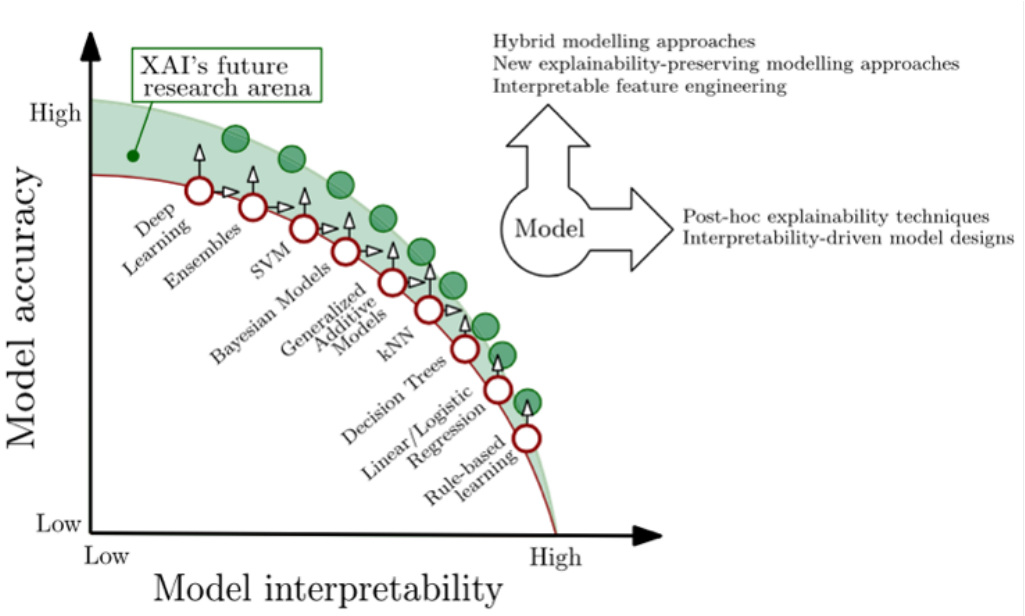


Figure 2.2: Visualizing the trade-off between model accuracy and model interpretability, and a representation of the area of improvement where the potential of XAI techniques and tools resides [52].

Models like LR and DTs are on the lower end of the complexity spectrum. These techniques offer a higher interpretability compared to other techniques due to their decision-

making process being straightforward and easy to understand [47]. For example, decision trees provide a clear, rule-based structure that can be followed logically to arrive at a decision [47]. Similarly, linear regression models offer transparent equations where the influence of each variable is directly observable [52].

Section 2.5.3 will provide some deeper understanding towards the multifaceted ways in which XAI is driving innovation within the actuarial sector, ultimately influencing the evolution of insurance pricing models.

2.5.2 Taxonomy of XAI Approaches

Diverse methods has been developed to make the decision-making processes clear of ML models. This subsection provides a structured overview of these methods. This subsection about the taxonomy of XAI approaches emphasizes the transparency of models, the distinctions between model-specific and model-agnostic techniques. Additionally, it considers the varying scope of explanations, from local to global. This subsection aims to guide the selection and application of XAI methods, ensuring that users can derive meaningful insights and understand how XAI approaches work.

2.5.2.1 Transparent Models & Post-Hoc Explainability

Transparent models are inherently interpretable, meaning their inner workings and decision processes are directly understandable without additional explanatory aids [52]. These models typically exhibit simplicity in their structure, making them easy to analyze and validate. Examples of transparent models include linear regression, decision trees, and logistic regression. These models are characterized by their direct interpretability through the examination of model parameters or decision rules. For instance, linear regression coefficients directly indicate the relationship between inputs and the prediction, allowing users to understand the impact of each feature on the output. Similarly, decision trees provide a clear, rule-based structure for decision making, where the path from the root to any leaf can be easily traced and understood [19].

While transparent models offer built-in interpretability, complex models such as deep neural networks or ensemble methods often act as black boxes, with their internal operations not easily accessible or understandable. Post-hoc explainability techniques are developed to address this gap, providing insights into the decision-making processes of these opaque models after they have been trained [54]. This differs to intrinsic model interpretation, draws conclusions from the structure of the fitted model and are typically associated with "interpretable" classes of models, such as the spare linear model and decision tree [47]. In the actuarial science literature, the GLM is probably the most mentioned example of an easily interpretable model [47].

Various approaches can be identified for post-hoc explainability techniques, as can be seen in Figure 2.3 [52].

1. *Local explanations.* These provide insights into the decision-making process of the model for individual predictions, which is crucial when the model's behavior needs to be justified in critical applications.
2. *Feature relevance.* Involves quantifying the importance and contribution of each input feature to the predictions of the model, which helps in understanding which variables play a pivotal role in the decision-making process.
3. *Explanations by example.* This approach uses specific data instances to show how the model behaves in particular cases, which can be particularly enlightening for end-users by providing relatable and concrete examples of the model's functionality.

4. *Text explanations.* These generate textual descriptions of why certain decisions or predictions were made, translating the model's operations into a more accessible and understandable format.
5. *Model simplification.* Creating a simpler model that approximates the behavior of the complex model, making it easier to understand and analyze while maintaining essential functionalities and predictive capabilities.
6. *Visualization.* Techniques in this category are used to visually represent data or model decisions, facilitating a better understanding of complex patterns or behaviors through graphical means, which can be more intuitive for users to comprehend.

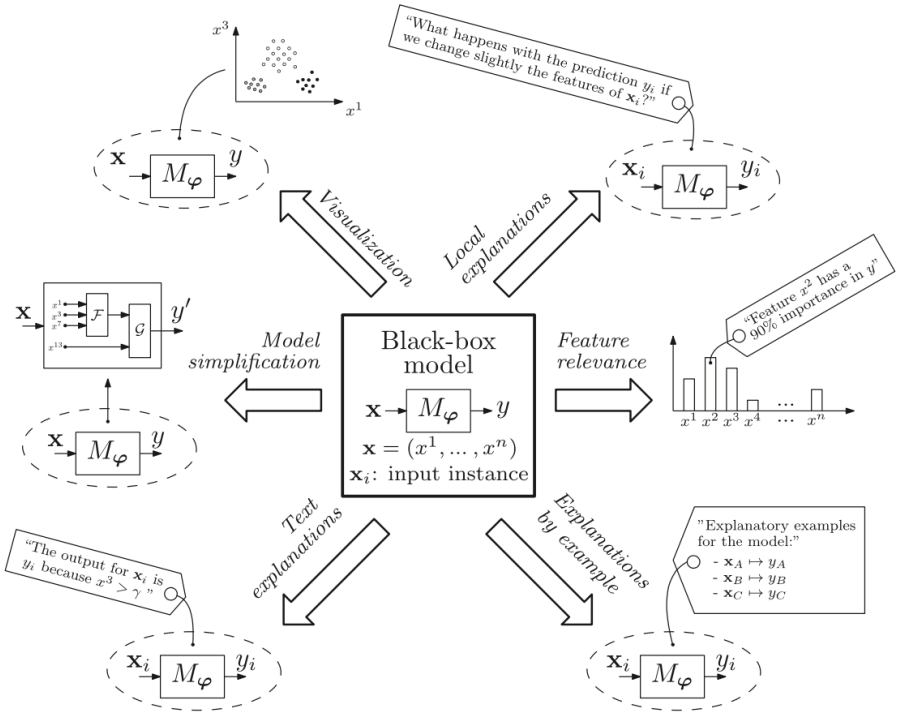


Figure 2.3: Conceptual diagram illustrating various post-hoc explainability approaches for a ML model M_φ [52].

Post-hoc techniques are essential for validating model behavior, ensuring fairness, and building trust, especially in domains where decisions have significant consequences, such as insurance [56]. These techniques bridge the gap between model performance and human interpretability (Figure 2.2), ensuring that users can understand, trust, and effectively manage AI systems.

2.5.2.2 Model-Specific & Model-Agnostic Explanation Methods

Model-specific explanation methods are designed to work with particular types of models, leveraging knowledge about the model's architecture and inner workings. These methods are based on the parameters of individual models and therefore can only be used on this specific model type [57]. For instance, for decision trees, explanations can often be derived directly from the paths and splits within the tree, offering a transparent view of the decision-making process [52]. This specificity allows for deep insights but at the cost of flexibility, as these methods are not generally applicable across different types of models.

Contrasting with model-specific techniques, model-agnostic methods are designed to be universally applicable across all types of models without any modifications or specific knowledge about the model's structure [52]. This universality is particularly advantageous in scenarios where the same explanation technique needs to be applied to different models or where the model's architecture is opaque or complex. Model-agnostic methods include techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which can provide insights into model behavior regardless of how the model processes its inputs [54]. These methods typically create interpretable approximations or use statistical techniques to estimate the impact of inputs on outputs across various models.

The choice between model-specific and model-agnostic explanation methods depends largely on the requirements and constraints of the specific application. Model-specific methods can provide deeper, more nuanced understandings of a model's behavior, which can be crucial for fine-tuning or diagnosing model performance in specialized applications. However, these methods can be complex to implement and understand, requiring in-depth knowledge of the model's architecture. On the other hand, model-agnostic methods offer flexibility and ease of use, making them suitable for a broader range of applications, especially when quick or general insights are needed across different models. They facilitate comparisons between models and are invaluable for applications involving multiple or changing model types.

2.5.2.3 Global & Local Explanation Methods

In XAI, a distinction can be made between local and global explanation techniques. Local explanation techniques focus on explaining model decisions for individual predictions, helping users and stakeholders understand the reasoning behind specific outcomes [52]. This is particularly important in scenarios where individual decisions have significant implications, such as loan approvals or medical diagnoses. Global explanation techniques aim to provide an overview of the model's overall behavior, offering insights into its general logic and decision patterns across a broad set of instances. This holistic view is crucial for evaluating the model's fairness, bias, and overall alignment with ethical guidelines [52].

To categorize explainability methods effectively, one must review the literature that suggests various classification frameworks. In Section 2.2.2.1, it is noted that insurance pricing involves a regression problem. Consequently, the research focuses on identifying XAI techniques suitable for this type of regression analysis. According to Arrieta et al., various XAI techniques could be identified [52]:

1. *LIME (Local Interpretable Model-agnostic Explanations).* This technique provides explanations for individual predictions by approximating the local decision boundary of any black-box model with a simpler, interpretable model such as a linear regression [58].
2. *SHAP (SHapley Additive exPlanations).* Leverages the concept of Shapley values from game theory to explain the contribution of each feature to a particular prediction. It provides a consistent and locally accurate attribution for each feature [58]. It can also provide an overview of the overall performance of a black-box model.
3. *Partial Dependence Plots (PDPs).* PDPs show the effect of a single feature or a pair of features on the predicted outcome of a model, averaged over a dataset, irrespective of the values of other features [21]. This can help understand the relationship between the feature(s) and the response variable.

4. *Individual Conditional Expectation (ICE) Plots.* ICE plots are a refinement of PDPs, showing the relationship between the feature and the response for individual observations, providing a more detailed view than PDPs [36].
5. *Feature Importance.* Involves ranking the features based on their importance to the model's predictions. Methods to determine feature importance can vary but often include techniques such as permutation feature importance, which assesses changes in model performance when a feature's values are randomly shuffled.

2.5.3 Integration of XAI in Insurance Pricing

It is important to highlight the growing importance of XAI in the decision-making processes of ML models within insurance pricing. The integration of XAI in insurance pricing marks a significant shift towards transparency and accountability within the industry. Notably, there are a few examples of XAI methodologies, such as feature importance analysis, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), being applied to insurance pricing [59]. However, the application of these methods is still limited. These XAI methods not only enhance transparency but also provide valuable insights into how ML algorithms arrive at specific pricing decisions, fostering trust and accountability in an industry where understanding the rationale behind predictions is important [53]. By incorporating XAI methods, insurance companies can offer stakeholders, including regulatory bodies, customers, and industry professionals, insights into how the model arrives at specific pricing decisions [60].

Through the usage of techniques such as SHAP and/or LIME, XAI methods strive to make AI decisions transparent and comprehensible [54]. McDonnell et al. underscores the importance of balancing model accuracy with interpretability, a critical aspect for regulatory compliance and operational transparency [61]. Currently, the integration of these techniques is not widespread in the insurance industry, signifying a promising area for future research and application. Fritz et al. emphasizes the necessity for AI systems in finance, including insurance, to not only be technically proficient but also ethically sound and understandable to humans [54]. This approach is particularly crucial in insurance, where decisions regarding risk assessment and pricing can have great implications for individuals and communities. Within telematics, the focus on creating explainable ML models for predicting claim frequency using telematics data underscores the importance of model explainability for regulatory compliance and stakeholder understanding [59].

The integration of XAI in insurance pricing represents a crucial step towards fostering trust, fairness, and accountability in the industry, ensuring that AI-driven decisions align with ethical standards and regulatory requirements while empowering stakeholders with transparent and interpretable insights. However, challenges persist, as highlighted by Fabris et al., which examines algorithmic decision-making's potential for unfair practices in insurance [60]. By employing statistical and ML methods to analyze patterns of discrimination and bias in insurance pricing, the research of Fabris et al. uncovers evidence of unfairness and discrepancies, prompting calls for regulatory oversight and the development of fair ML practices to mitigate bias in insurance algorithms [60]. Arrieta et al. also identifies several XAI challenges that must be addressed to ensure the responsible use of AI [52]. Beyond explainability, the guidelines for Responsible AI emphasize the importance of fairness, accountability, and privacy in the implementation of AI models in real-world environments [52]. An overview of the challenges is presented in Figure 2.4.

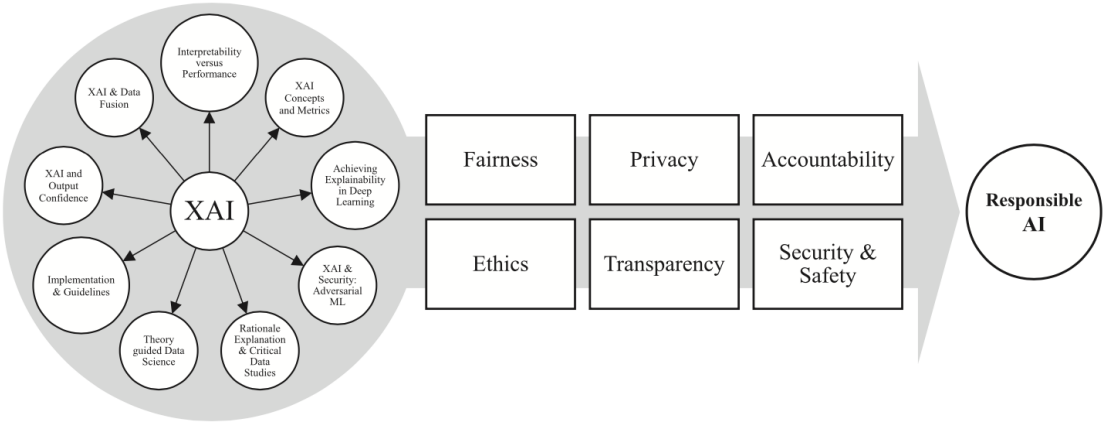


Figure 2.4: Summary of XAI challenges its impact on the principles for Responsible AI [52].

An example solution, according to Koster et al., to the problem of unfairness could be the development of a checklist to ensure the explainability and transparency of AI applications in the insurance industry [62]. This reinforces the importance of creating AI applications that are understandable to both experts and non-experts, addressing potential biases, and emphasizing continuous improvement based on feedback from the industry.

2.6 Regulatory Considerations in AI-Driven Insurance Practices

The integration of AI into insurance practices brings about a host of regulatory considerations that are essential for maintaining transparency and fairness [63]. Drawing insights from recent studies and papers, this section delves into the regulatory aspects of AI-driven insurance practices.

A key requirement in the insurance (and financial) industry is the need for interpretable pricing models which are easily explainable to all stakeholders (e.g., managers, customers, shareholders, regulators, auditors) [52]. The ethical guidelines outlined by Mullins et al. for AI and big data in the European insurance market stress the importance of transparency, fairness, and accountability [63]. These principles are essential for ensuring that AI-driven practices in insurance and financial services broadly adhere to standards that protect consumers while fostering innovation and efficiency. Insurance ratemaking models are highly regulated, and they must meet specific requirements[6] before being deployed in practice [63].

2.6.1 Regulatory Frameworks Governing AI in Insurance

As AI is becoming more embedded in industry sectors, regulatory frameworks play a pivotal role in ensuring that their deployment is both ethical and compliant with established standards of data protection and operational transparency. Two significant regulations in this context are the General Data Protection Regulation (GDPR)[7] and the Artificial

[6] see, e.g., the regime "algorithmic accountability" of decision-making machine algorithms imposed by the European Union's General Data Protection Regulation (GDPR) [63].

[7] https://gdpr-info.eu

Intelligence Act (AI Act)[8]. These regulations will be discussed separately in this section.

2.6.1.1 General Data Protection Regulation (GDPR)

Data privacy and security are at the forefront of regulatory concerns, especially given the vast amounts of personal and sensitive information handled by insurance companies [64]. The application of AI in areas like automobile insurance claim prediction and customer data analytics necessitates robust data protection measures. Regulations such as the GDPR in the European Union set stringent guidelines for data processing, requiring that insurance companies employ AI solutions that are compliant with data protection laws [63]. Although the GDPR aims to standardize data protection laws across Europe, the EU has granted member states the authority to establish specific regulations [65]. This includes determining the age at which children can consent to the processing of their personal data, handling criminal convictions data, and managing data processing deemed to be of substantial public interest.

Insurance companies often need to process sensitive personal data to underwrite risks, and provide claims handling and other insurance related services [65]. For example, in the UK, key stakeholders within the insurance sector collaborated closely with organizations[9] to establish specific legal grounds for processing personal data in insurance. Under this framework, special categories of personal data and information on criminal convictions can be processed without consent when necessary for legitimate insurance purposes, including underwriting, policy administration, and claims handling, and when it serves substantial public interests. According to Leonardi, the Parliament acknowledged the importance of insurance availability, risk-based pricing, fraud detection, and efficient claims administration as matters of significant public interest [65]. Thus, subject to certain exemptions, the UK insurance industry operates with a clear and defined legal basis for processing such data in underwriting and claims management.

2.6.1.2 Artificial Intelligence Act (AI Act)

There is necessity for transparency and explainability in AI applications within the insurance domain, as there's a growing demand for AI systems to be understandable to both practitioners and non-experts [62]. This involves clear communication regarding how AI models make decisions, the data used, and the reasoning behind specific outcomes. Regulatory frameworks are increasingly emphasizing the importance of explainable AI to ensure that stakeholders can trust and effectively scrutinize AI-driven decisions.

The European Union is making significant strides with its AI Act, aiming to create a comprehensive regulatory framework for AI applications within its member states [56]. This legislative effort is the world's first attempt to regulate AI comprehensively, focusing on ensuring that AI systems are safe, respect fundamental rights, and promote innovation and technological sovereignty in Europe. The AI Act introduces a risk-based classification system for AI applications, categorizing them into four levels: unacceptable, high, limited, and minimal risk, as can be seen in Figure 2.5.

This classification dictates the regulatory requirements each AI application must meet, from stringent compliance and testing for high-risk applications to minimal requirements for those deemed to pose little risk [56]. The legislation represents a pioneering step towards regulating the rapidly evolving field of AI, seeking to balance the need for innovation with

[8] https://artificialintelligenceact.eu

[9] Organisations such as the Association of British Insurers (ABI), the Lloyd's Market Association (LMA), and the government [65].

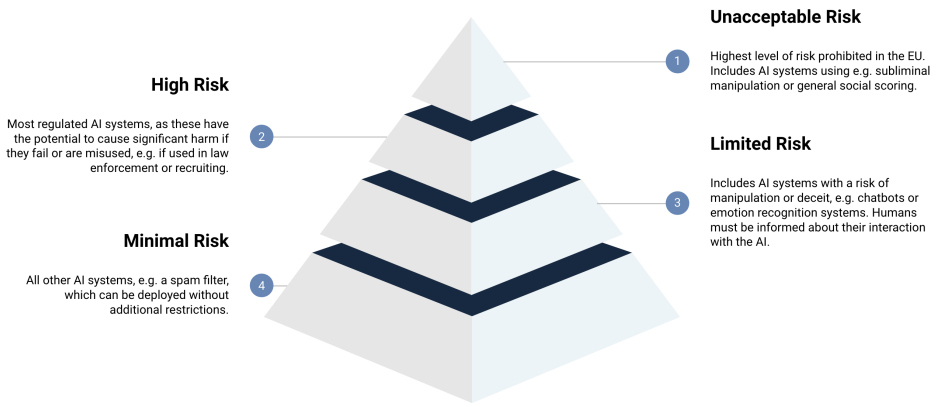


Figure 2.5: The AI Act defines four levels of risk for AI systems [66].

the imperative of safeguarding fundamental rights and societal values. It is noteworthy that Panigutti et al. argue that the AI Act neither mandates a requirement for XAI, which remains a focus of extensive scientific research and faces technical limitations, nor prohibits the use of black-box AI systems [56].

2.6.2 Regulatory Consequences for Insurance Companies & End Consumers

Ensuring compliance with existing regulations while also adapting to new requirements is essential for insurance companies utilizing AI. This includes continuous monitoring and updating of AI systems to align with evolving regulatory landscapes. Regulatory bodies are increasingly advocating for the establishment of oversight mechanisms that can assess and monitor the impact of AI applications over time, ensuring they continue to meet ethical, legal, and operational standards [56]. Traditional Model Risk Management (MRM) processes often struggle to address the unique risks associated with AI models. These models present distinct challenges, such as feature selection and hyperparameter tuning (the adjustment of technical parameters specific to AI models, like the number of layers in a NN). Additionally, inherent data biases can create further obstacles for validators [67].

There are several challenges in adapting traditional MRM processes to AI models. Firstly, the inherent opacity of AI models necessitates better coordination among developers, users, and validators [67]. Secondly, managing these models requires a different skill set than traditional MRM processes. Also, there is a critical need to avoid biases in models and ensure they uphold fairness and ethical standards. Unlike traditional models that assume static performance between reviews, AI models frequently evolve, challenging the agility of traditional MRM processes [67]. Additionally, AI models often require frequent recalibrations to remain effective, a demand that traditional MRM may not swiftly meet. Lastly, some specific AI applications, like resume screening or chatbots, do not conform easily to traditional MRM frameworks such as those outlined in SR 11-7[10], indicating a

[10]SR 11-7 is a regulatory guideline issued by the Federal Reserve that outlines standards for risk manage-

need for process adaptation to include these innovative use cases [67].

For insurance companies, regulations such as GDPR and the AI Act (Section 2.6.1.1 & 2.6.1.2) compel the adoption of advanced data governance and AI systems management practices. The GDPR mandates strict control over personal data, forcing insurers to refine their data collection and processing processes [69]. This includes ensuring that all personal data used in automated decision-making, such as claim adjustments or risk assessments, complies with privacy requirements, which adds layers of complexity. Moreover, the GDPR's right to explanation implies that any AI-driven decision must be transparent and justifiable to the customer, which can challenge the deployment of complex, non-transparent AI models such as deep learning algorithms [69]. The AI Act further intensifies this complexity by classifying AI systems according to risk, with high-risk applications in sensitive sectors like insurance subjected to more stringent regulatory scrutiny (see Section 2.6.1.2). For insurance companies, this means investing in robust AI audit and oversight mechanisms, potentially slowing innovation but also standardizing safer and more trustworthy AI deployments. In 2019, De Nederlandsche Bank (DNB) developed a framework known as "SAFEST", which stands for Soundness, Accountability, Fairness, Ethics, Skills, and Transparency, to guide the responsible use of AI in the financial sector [45]. This set of principles emphasizes the importance of AI applications being reliable, accurate, and compliant with existing laws (such as GDPR & AI Act), ensuring they operate predictably and mitigate any systematic risks.

From the consumer perspective, these regulatory frameworks are designed to enhance protection and empower individuals with more control over their personal data (see Section 2.6.1.1). The GDPR provides consumers the right to not only understand but also contest decisions made by AI, fostering greater transparency [69]. This empowers consumers to demand fairness in how their data is used in insurance premium calculation and claim processing, potentially leading to more consumer-friendly practices. The AI Act's emphasis on transparency and explainability can increase consumer trust in AI technologies. By understanding how and why certain AI-driven decisions are made, consumers can better appreciate the value and fairness of these technologies.

In addition to the impact on insurance companies and clients, new companies face challenges like navigating AI-related regulations (e.g., GDPR, AI Act) while leveraging AI for innovation. The relationship between insurers and regulators is evolving, requiring collaboration to ensure compliance without the discontinuation of innovation.

2.7 Summary

This chapter provides a foundational understanding of car insurance within Dutch law, explaining its classification under non-life insurance and covering mandatory requirements, policy structures, and various claim types relevant within insurance policies. The chapter also delves into insurance pricing, highlighting the role of risk management and actuarial assessments in determining premiums. Actuaries use frequency-severity models to estimate claim costs, focusing on the pure premium of an insured person.

The discussion extends to the evolution of modeling techniques in the insurance industry. GLMs have been standard since the 1990s but face various limitations, such as capturing correlations between features. To overcome these limitations, ML techniques are increasingly adopted for their flexibility and accuracy in handling complex data relationships. However, ML's adoption brings challenges like model interpretability, data privacy,

ment practices and policies concerning the use of models by financial institutions, ensuring their reliability, accuracy, and oversight [68].

and regulatory compliance. XAI methods address these challenges by making AI decisions transparent and comprehensible. The chapter also described the taxonomy of XAI, evaluating differences between model-specific and model-agnostic explanation methods, as well as global and local explanations. Additionally, it covers regulations such as GDPR and the AI Act, which aim to ensure the ethical and transparent use of AI in insurance, balancing technological advancement with consumer protection and trust.

Chapter 3

Methodology

3.1 CRISP-DM

This thesis handles a ML problem. A standard approach is needed which translates business problems into ML tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience. The CRISP-DM (*CR*oss *I*ndustry *S*tandard *P*rocess for *D*ata *M*ining) project defines a process model which provides a framework for carrying out data mining projects which is independent of both the industry sector and the technology used [22]. The CRISP-DM process model aims to make large data mining projects less costly, more reliable, repeatable, manageable, and faster. CRISP-DM is still a de-factor standard in data mining [23]. It can also be applied to ML problems.

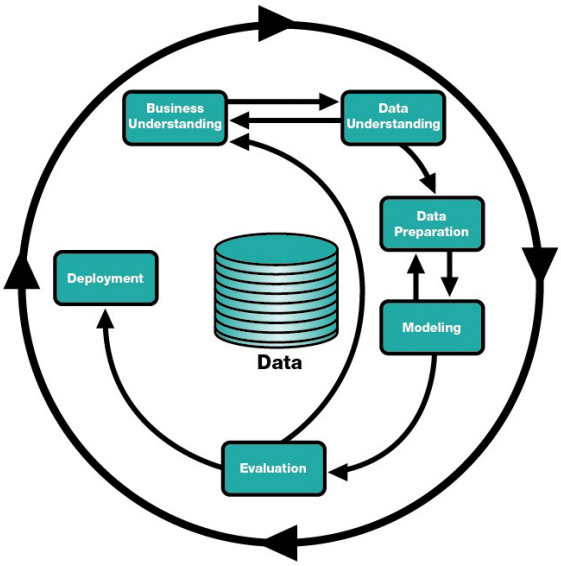


Figure 3.1: CRISP-DM, a framework for data mining projects [22].

The CRISP-DM framework is a widely recognized methodology that provides a structured approach to planning and executing data mining projects [23]. It comprises six major phases: *Business Understanding Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, and *Deployment*. Each phase is crucial for the success of a data mining project, and they are typically executed in sequence, although iterations and loops between phases are common to refine processes and improve outcomes [22]. The following is a general

description of each phase, based on the CRIPS-DM guide by Schröer et al [23]:

1. *Business Understanding.* Identify the business problem, define data mining goals aligned with business objectives, and establish success criteria. Develop a detailed project plan.
2. *Data Understanding.* Collect data from various sources, perform exploratory analysis to understand its attributes and quality, and identify any initial insights or issues.
3. *Data Preparation.* Cleanse the data, select relevant subsets, derive new variables as needed, and transform the data into a final dataset ready for modeling.
4. *Modeling.* Choose appropriate modeling techniques based on the business problem and data type, configure model parameters, build and assess model performance.
5. *Evaluation.* Assess the model against business objectives, interpret results, determine business impact, and decide on the next steps. Review the process for improvements.
6. *Deployment.* Implement the data mining solution in a business setting, which may range from generating reports to operationalizing models. Plan for monitoring and maintenance to ensure ongoing effectiveness.

This thesis focuses on the CRISP-DM methodology, which outlines the most essential steps in the ML process. However, it is crucial to highlight the existing literature on CRISP-ML, which includes an extra monitoring step as monitoring will play a significant role in maintaining ML models after deployment. Appendix B.1 provides some more information about both the CRISP-DM and CRISP-ML methodology.

3.2 Machine Learning Techniques

In this section, the thesis describes the theory that lays the groundwork for the (regression) ML models deployed. The selection and development of appropriate models are important, as these serve as the tools through which theoretical concepts are translated into measurable entities and examine the intricacies of the research questions. The ML models were chosen based on a comprehensive review of relevant studies discussed in Chapter 2, insights gained from an AutoML analysis in Section 4.5.1, and expert input from both `the consultancy firm` and `the insurance company`. This informed approach ensures that the models are well-suited to address the specific challenges and objectives of the study.

Regression models are models commonly used to predict the pure premium by modeling the expected claim frequency and claim severity separately. This section presents the various models utilized in this thesis, which will be compared to derive various insights of the selected models. The models are explained using a detailed description of the model's structure.

3.2.1 Generalized Linear Model (GLM)

GLMs have been generally covered in Section 2.3 of the thesis paper. However, a further description of the model will be provided in this section to introduce the technique to readers unfamiliar with it. Unlike standard linear regression, which presumes normally distributed errors, GLMs allow for response variables that have error distributions modeled by the exponential family (e.g., normal, binomial, Poisson, and gamma distributions)[31]. This adaptability makes GLMs particularly useful for modeling different types of data that do not meet the normality assumption. In this section, the gamma-distributed dependent and Poisson-distributed GLMs are handled. The gamma distribution is employed to model the severity of claims, whereas Poisson regression is utilized to model their frequency. An

example of the functioning of a GLM is provided in Figure 3.2, which demonstrates how a linear regression model predicts the mean of the response variable Y for given values of the predictor variable x, and how the actual values of Y are distributed around these predicted means.

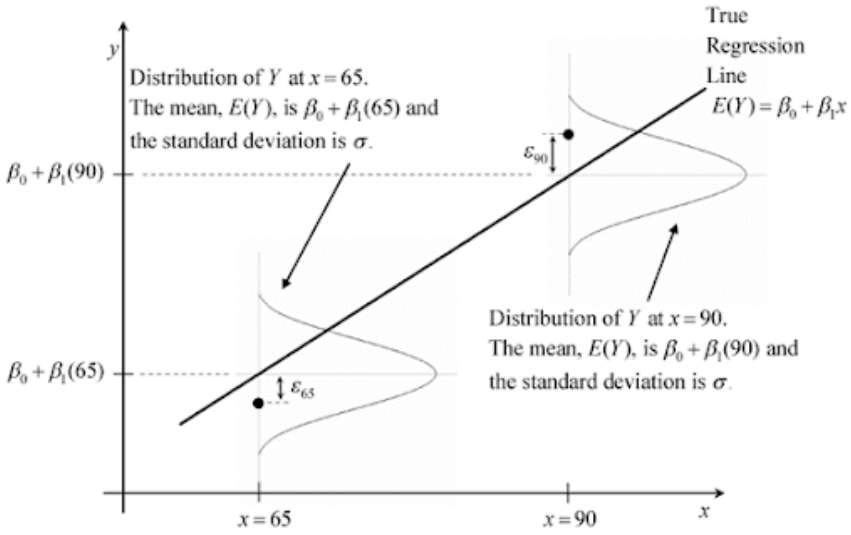


Figure 3.2: Illustration of a generalized linear regression model, showcasing the relationship between a predictor variable x and response variable y [70].

At the core of GLMs is the division of data into systematic and random components [32]. The systematic component accounts for the predictable aspects of the data through a combination of predictors or risk factors[1]. GLMs model the relationship between μ_i (the model prediction) and the predictors as follows. Equation 3.1 states that some specified transformation $g(.)$ (called a link function) of μ_i (denoted $g(\mu_i)$) is equal to the intercept (denoted β_0) plus a linear combination of the predictors and the coefficients, which are denoted $\beta_1 ... \beta_p$. x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. The values for the intercept (β_0) and the coefficients ($\beta_1 ... \beta_p$) are estimated by GLM software [32].

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{3.1}$$

The primary focus in Equation 3.1 is not on the value $g(\mu_i)$ but rather on μ_i itself. Therefore, after computing the linear predictor, one obtains the model prediction by applying the inverse of $g(.)$ to the result of Equation 3.1.

The random component, on the other hand, captures the inherent unpredictability of insurance claims [35]. This structure allows actuaries to model the expected loss or claim cost associated with an insurance policy more accurately. In a GLM, y (the target variable) is modeled as a random variable that follows a probability distribution that is assumed to be a member of the exponential family of distributions, as shown in Equation 3.2 [32].

$$y_i \sim Exponential(\mu_i, \phi) \tag{3.2}$$

[1] In insurance, one can think of the policyholder's age, vehicle type, and geographical location

"Exponential" in Equation 3.2 does not specify a particular distribution but serves as a stand-in for any distribution within the exponential family. These distributions are characterized by two parameters: μ, representing the mean, and ϕ, the dispersion parameter, which influences the variance.

Including GLMs in this thesis is a strategic choice, which is elaborated upon in Section 4.6.1. The implementation of GLMs serve as a benchmark against which the performance of various ML techniques can be compared. GLMs are still widely used across different studies, with various examples such as Wütrich & Buser who analyze the algorithm for non-life insurance pricing [71], Devriendt et al. use it for the prediction of insurance prices [43], and Hassani et al. mention GLMs in their paper meant for the actuarial science world [41]. There are many more examples, as GLMs are still considered industry standard for insurance pricing **Reil_A_2024]**[19].

3.2.2 Light Gradient Boosting (LightGBM)

LightGBM, a distributed and high-performance gradient boosting framework developed by Microsoft, is tailored for efficiency, scalability, and accuracy [72]. This framework operates on decision tree algorithms, enhancing model efficiency while minimizing memory consumption. It introduces innovative methodologies such as Gradient-based One-Side Sampling (GOSS), which preferentially maintains instances exhibiting larger gradients to optimize memory utilization and accelerate training processes [73]. Additionally, LightGBM employs a histogram-based algorithm to identify the optimal split points. This process involves converting continuous features into discrete bins, significantly accelerating the training process and thus efficiency [72]. The objective function of LightGBM can be written as follows:

$$g_i = \frac{\delta L(y_i, F(x_i))}{\delta F(x_i)} \tag{3.3}$$

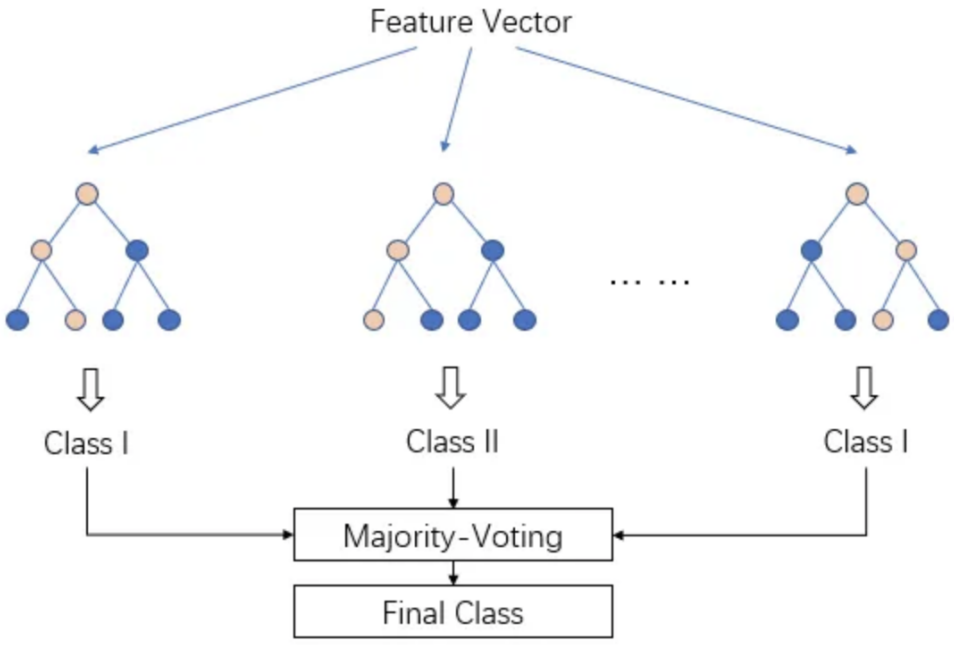


Figure 3.3: LightGBM, which stands for Light Gradient Boosting Machine, is a distributed, high-performance implementation of the gradient boosting framework. [72].

An overview of the LightGBM algorithm is provided in Figure 3.3. When it's time to make a prediction, each model casts its vote for the outcome, and the final prediction is determined by the majority's decision. For each instance i, the gradient g_i is calculated of the loss function with respect to the current prediction [74]. The data instances are then sorted based on the absolute values of their gradients, where the top $a\%$ of the data instances with the largest gradients are meant to be crucial for model improvement.

LightGBM can be a promising ML technique for this thesis to implement, as its design and performance capabilities offer several advantages for handling complex and large-scale datasets. Various studies have used LightGBM in similar contexts, where Wütrich & Buser use the algorithm for non-life insurance pricing [71], Grize et al. mention it as one of their ML applications in non-life insurance [37], and Maillart uses the LightGBM algorithm for predicing claim frequencies [59].

3.2.3 Extreme Gradient Boosting (XGBoost)

Unlike methods that depend on a single predictive model, ensemble techniques like boosting algorithms harness the combined strengths of multiple models, sometimes yielding superior performance compared to what any individual model could achieve alone [75]. Boosting is an ensemble technique that can assemble thousands of forecasting models with lower performance into a strong, high-performance model by repeatedly merging the models within permissible parameter values [76].

Extreme Gradient Boosting, also known as XGBoost, is an advanced implementation of the (gradient) boosting machines[2] designed by Tianqi Chen to optimize the speed and

[2]Gradient boosting machines (GBMs) are a type of ML boosting. It relies on the intuition that the

performance of gradient boosting frameworks [78]. XGBoost is renowned for its efficiency and scalability, making it particularly well-suited for handling large and complex datasets across various computing environments [79]. The objective function of XGBoost can be written as follows:

$$obj(\theta) = \sum L(\hat{y}_i, y_i) + \sum \Omega(f_k) \tag{3.4}$$

The objective function of XGBoost consists of two main components: a loss function (L), and a regularization term, denoted as $\Omega(f_k)$. This regularization term helps to reduce the output variation of each new tree added to the model [76]. In this context, $\hat{y}_i$ represents the predicted value, while y_i denotes the observed value.

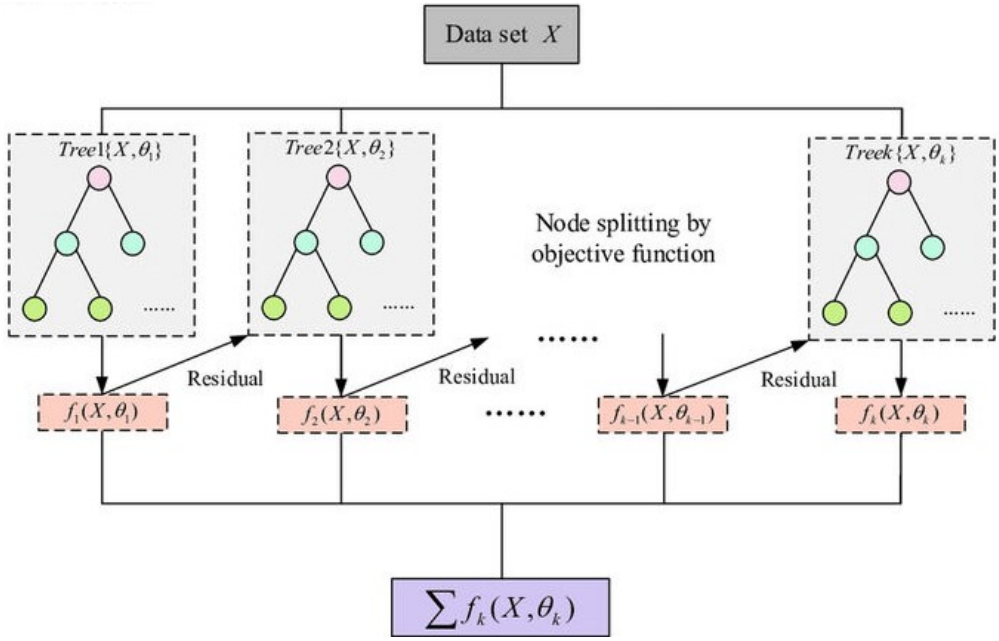


Figure 3.4: XGBoost, which stands for Extreme Gradient Boosting, is an advanced implementation of the gradient boosting machine (GBM) algorithm [78].

Figure 3.4 depicts the mechanism of XGBoost. This method involves the sequential use of multiple decision trees, each built to correct the errors of its predecessors. Initially, the dataset X serves as the input for the first decision tree, labeled as $Tree_1$ in the diagram [80]. Each tree in the sequence is characterized by parameters θ_j, which dictate how nodes within the tree are split. The objective function, explained in Equation 3.4, orchestrates these splits. As $Tree_1$ processes the dataset, it leaves behind residuals (differences between the predicted outcomes and the actual data). These residuals then become the target outputs for the next tree, which aims to predict and rectify these discrepancies [80]. This pattern continues, with each subsequent tree (up to $Tree_k$) focusing on the residuals left by its predecessor, thereby progressively refining the model's accuracy. The final prediction model, represented by the summation $\sum f_k(X, \theta_k)$ at the bottom of Figure 3.4, is the aggregate of the outputs from all individual trees. This cumulative approach leverages the collective strengths of each component model, resulting in a robust and precise ensemble predictor.

best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error [77].

XGBoost looks like a promising ML technique for this thesis due to its robust performance in both theoretical applications and practical scenarios. Its efficiency and scalability make it a good choice for handling large datasets, which is critical in this research. Various studies have used XGBoost in similar contexts, such as Wütrich & Buser, who use the algorithm for non-life insurance pricing [71], Poufinas et al. [38] use XGBoost for predicting motor insurance claims, as well as Clemente et al. [36]. Other mentions of XGBoost within the literature are in automobile insurance classification algortihms based on telematics driving by Huang & Meng [42], Maillart uses the XGBoost algorithm for predicing claim frequencies [59], and Henckaerts et al. use it for predicting tariff plans with tree-based ML methods [19].

3.2.4 Multi-Layer Perceptron Neural Network (MLP)

A Multi-Layer Perceptron, or MLP, falls under the category of feedforward algorithms. It is a type of artificial NN capable of handling both linearly separable and non-linearly separable data [81]. It is a Neural Network (NN) where the mapping between inputs and output is non-linear [82]. In this approach, inputs are combined with initial weights to form a weighted sum, adds a bias, which is then subjected to an activation function, similar to a perceptron. The difference between the predicted output and the actual output is measured using a loss function shown in Equation 3.5, such as mean squared error (MSE) for regression tasks [83].

$$\Delta_w(t) = -\varepsilon \frac{dE}{dw_{(t)} + \alpha \Delta_{w\,(t-1)}} \tag{3.5}$$

The error derived from this loss function is then propagated backward through the network during backpropagation, updating the weights and biases [82]. The network employs optimization algorithms like gradient descent to minimize the loss by adjusting the weights and biases accordingly. Training the MLP involves multiple iterations (epochs) of forward propagation and backpropagation to iteratively reduce the loss and enhance the network's accuracy.

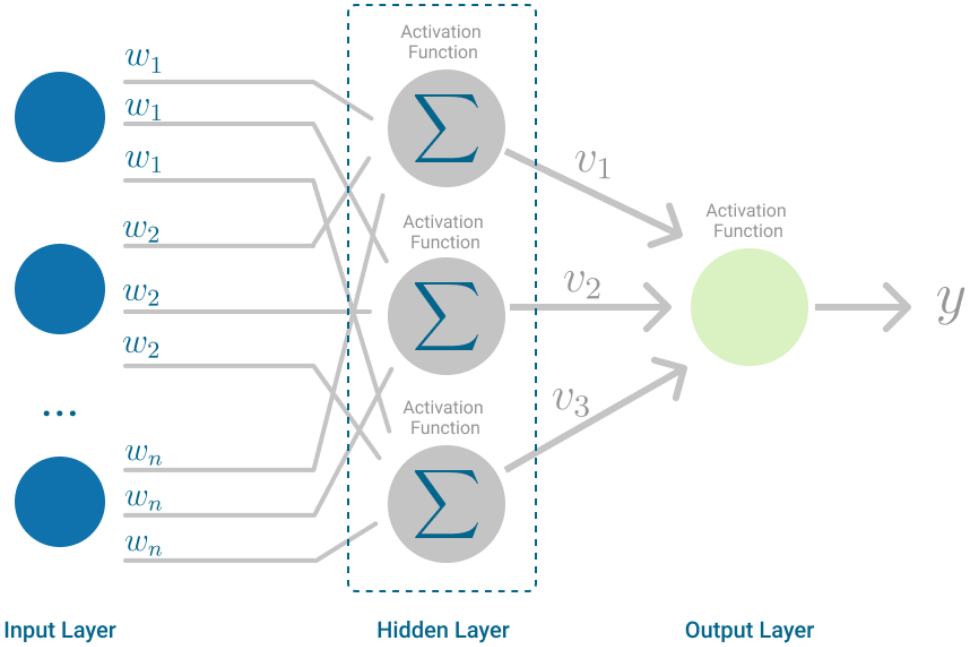


Figure 3.5: MLP, which stands for Multi-Layer Perceptron, falls under the category of feedforward algorithms, which is type of NN [82].

As discussed, MLP consists of input and output layers, along with one or more hidden layers containing multiple neurons, as shown in Figure 3.5. Unlike a simple perceptron, which requires its neurons to have an activation function imposing a threshold (such as ReLU[3] or sigmoid[4]), the neurons in a MLP can utilize any arbitrary activation function [82].

MLP looks like a promising ML technique due to its ability to learn and model complex non-linear relationships, its flexibility in handling various types of data, and its robust performance. However, in practice, MLPs take a lot of time for training [86]. However, various studies deploy MLPs in their studies, such as the paper by Hassani et al. talking about big data and actuarial science [41]. Also, NNs are mentioned various times in papers from Wütrich & Buser, who use the algorithm for non-life insurance pricing [71], Clemente et al., using it for modelling motor insurance claim frequency and severity [36], and in the tariff analysis of automobile insurance by Martinez et al. [48]. The aforementioned reasons make MLP a good choice for investigation within this study.

3.3 XAI Methods

In this section, the study delves into various explainability techniques, a crucial aspect of deploying ML models, particularly in domains requiring transparency and accountability, such as finance and insurance. Explainability in ML can be approached from two main perspectives: global and local explanations [87]. This has also been explained in Section 2.5.2.3. This section will mainly handle the XAI techniques SHAP and LIME, where

[3]The Rectified Linear Activation Function, commonly known as ReLU, is a piecewise linear function that outputs the input directly if it is positive; otherwise, it outputs zero. It has become the default activation function for many types of neural networks because models using ReLU are easier to train and often achieve better performance [84].

[4]The sigmoid function is one of the most commonly used activation functions [85]. Its mathematical representation is $\sigma(x) = \frac{1}{1+e^{-x}}$.

both share the idea to quantify how much a feature contributes to the prediction for a given example [58].

3.3.1 SHapely Additive exPlanations (SHAP)

SHAP (SHapely Additive exPlanations) is a XAI technique used for explaining the output of ML models [71]. It is grounded in the framework of game theory, particularly through the use of Shapley values, a method developed to fairly allocate the payout of a cooperative game to its players based on their contribution to the total payout. In the context of SHAP, features of a data instance are considered as "players" in a game where the "payout" is the prediction made by the model [88].

The SHAP explanation model has its own formula as shown in Equation 3.6, where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, which is a binary variable indicating whether the j-th feature is included in the subset [88]. M is the maximum coalition size (total number of features) and $\phi_j \in R$ is the feature attribution for a feature j, the Shapley values. ϕ_0 represents the model output when no features are present.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{3.6}$$

To determine the importance of a feature, one can intuitively consider the following process: for each iteration, feature values are drawn in random order for all features except the feature in question. Then, the difference in the prediction with and without this specific feature is calculated. The Shapley value is computed by averaging these differences across all possible combinations [88]. Essentially, the Shapley value represents the average marginal contribution of a feature, taking into account every possible combination of feature inclusions. This is visualized in Equation 3.7.

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)] \tag{3.7}$$

The SHAP value for a feature ϕ_j is calculated by considering all possible subsets S of the set of features N that do not include the feature j [71]. For each subset S, the model's output when only the features in subset S are present is denoted by $f(S)$, and the model's output when the feature j is added to subset S is denoted by $f(S \cup \{j\})$. The difference $f(S \cup \{j\}) - f(S)$ represents the marginal contribution of feature j to the model's prediction when added to subset S. The SHAP value ϕ_j is obtained by taking a weighted average of these marginal contributions across all possible subsets S [71]. The weighting factor $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ ensures that the contribution of feature j is averaged over all possible permutations of the features, thereby providing a fair attribution.

While the general SHAP framework provides a robust method for interpreting the outputs of ML models, different variations of SHAP have been developed to address the specific needs and structures of different types of models. These variants include:

1. *KernelSHAP*. Uses a weighted linear regression where the weights are determined by the distance between the instance being explained and the simplified input representations (coalitions). This method is particularly useful when the underlying model is a black-box or a non-linear model, such as SVMs or NNs [89].

2. *TreeSHAP*. Specialized SHAP variant for tree-based models like DTs, RFs, and GBMs. It efficiently computes exact Shapley values by leveraging the tree structure, making it significantly faster than KernelSHAP for these models [89].
3. *DeepSHAP*. Extension of SHAP tailored for deep learning (DL) models. DeepSHAP approximates Shapley values by backpropagating through the network to attribute the prediction to each input feature. Particularly effective in tasks where DL models excel [90].

Shapley values provide an estimate of the impact each feature has on the prediction relative to a baseline prediction, which can be done for both global (Appendix B.2.1) as well as local explanations (Appendix B.2.2) [88]. This method ensures that each feature's contribution is uniquely calculated, taking into account the interaction with other features, which provides a more detailed and nuanced interpretation of the model's behavior.

Global explanations using SHAP aim to provide insights into the overall behavior of the ML model across the entire dataset. By averaging SHAP values over many instances, one can identify which features are generally most important in influencing the model's predictions. This helps in understanding the model's decision-making process on a broad scale and identifying any potential biases or dependencies on particular features [91]. Examples of evaluating global SHAP explanations are provided in Appendix B.2.1.

Local explanations with SHAP focus on understanding the model's prediction for a specific instance. By calculating the SHAP values for an individual prediction, one can determine how each feature contributed to that particular outcome. This localized insight is critical for explaining specific predictions to end-users, diagnosing model errors, and improving transparency in decision-making processes. Figure B.3 shows the functionality of local Shapley explanations for house price predictions [92]. Examples of evaluating local SHAP explanations is provided in Appendix B.2.2.

3.3.2 Local Interpretable Model-agnostic Explanations (LIME)

LIME, or Local Interpretable Model-agnostic Explanations, is a technique used to provide explanations for predictions made by complex ML models, often referred to as "black-box" models [93]. The core idea of LIME is to approximate the complex model locally with a simpler, interpretable model around the prediction of interest. This is achieved by perturbing the input data and observing the corresponding changes in the output, thereby allowing an assessment of which features significantly influence the output. The overall goal of LIME is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier [94]. This method addresses the challenge of "local fidelity," meaning the explanation must accurately reflect the model's behavior near the specific instance, even though it may not generalize to the entire dataset [94]. This is crucial because features that are important on a global scale might not be as influential locally, and vice versa.

LIME operates by generating new samples around the instance being explained and observing how the predictions change when inputs are changed. This local dataset is then used to train a simpler model, such as a LM or a DT, which serves to approximate the complex model's behavior near the instance. This approach helps highlight which features significantly influence the output near the instance, providing insights into why the model made a particular decision [94]. The explanation produced by LIME at a local point x is obtained by the following generic formula, shown in Equation 3.8 [95].

$$\xi(x) = argmin\mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.8}$$

Based on an original input x, the algorithm aims to minimize the loss function $\mathcal{L}(f, g, \pi_x)$, which represents how unfaithful g is in approximating f in the locality defined by π_x. The loss function L lets us minimize the local mismatch between the original complex function f and approximating the simplified function g.

An analysis provided in the paper by Garreau & von Luxburg discusses LIME's effectiveness and limitations [58]. It demonstrates that when explaining LMs, LIME can effectively highlight important features by using coefficients that are proportional to the gradient of the function being explained. An example of the representation by LIME based on an example is provided in Section B.3 of the Appendix. The analysis with LIME also reveals vulnerabilities, such as the potential to miss important features due to poor parameter choices, which could lead to misleading interpretations [58]. Another limitation of such linear explanations is their locality. The explanations are specifically tailored to the instance they were generated for, making it uncertain whether these explanations can be generalized to other, unseen instances. Moreover, it is often not apparent which specific region of the input space these explanations are valid for, adding to their limitations [58].

3.4 Validation Methods

This section handles the methods employed to validate the models developed in this thesis. It covers the evaluation metrics used to assess model accuracy and the techniques implemented to ensure reliable results. Additionally, it describes the use of of an interview methodology which is used to validate the explainability of the models from a domain-specific perspective.

3.4.1 Evaluation Metrics

In the context of predicting insurance prices, it is crucial to evaluate the performance of the regression model(s) using a variety of evaluation metrics. This helps ensure that the model not only fits the data well but also works effectively when applied to unseen data. Various metrics will be used to assess model performance:

1. *Mean Squared Error (MSE).* The MSE is a metric that measures the average of the squares of the errors. Squaring the errors has the effect of amplifying large errors. The greater the difference between the predicted and actual values, the more significant the resulting squared error becomes. This approach penalizes models by increasing the average error score when used as a metric [96]. The formula of MSE is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{3.9}$$

2. *Root Mean Squared Error (RMSE).* RMSE is an extension of the MSE [96]. The square root of the error is calculated in RMSE, ensuring that its units match those of the target value being predicted. MSE squares each error to remove its sign and penalize large errors. Taking the square root in RMSE reverses this operation while keeping the result positive. RMSE's formula is given by:

$$RMSE = \sqrt{MSE} \tag{3.10}$$

3. *Mean Absolute Error (MAE).* This is the average of the absolute errors between the predicted values and the actual values. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE [96]. Unlike the RMSE, the changes in MAE are linear and therefore intuitive. The MAE formula is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{3.11}$$

4. *Root Mean Squared Logarithmic Error (RMSLE).* The RMSLE is calculated by first applying the logarithm to both the actual and predicted values, and then taking the difference between these logarithmic values [97]. This metric is robust to outliers, ensuring that small and large errors are treated more evenly. RMSLE penalizes the model more for underestimations than overestimations due to its logarithmic nature. This results in lower penalties for high errors and can be advantageous in scenarios where overestimations are acceptable, but underestimations are not. The formula of RMSLE is:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\log(1+y_i) - \log(1+\hat{y}_i)^2} \tag{3.12}$$

5. *Coefficient of Determination (R^2).* The R^2 (or R-Squared) is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s) [98]. It indicates how well the data points fit a statistical (regression) model. Its value ranges from 0 to 1, where 0 means that the model does not explain any of the variance in the dependent variable and 1 indicates that the model perfectly explains the variance. The formula for R^2 is [59]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum i = 1^n (y_i - \bar{y_i})^2} \tag{3.13}$$

6. *Explained Variance (EV).* EV is a metric that measures the proportion of the variance in the dependent variable that is "explained" (captured) by the model. It helps assess how well the model accounts for the variability in the data, where, just like R^2, 1 is the highest value [99]. A higher explained variance value indicates that the model is capturing a higher amount of the variance present in the target variable, typically suggesting a better fit of the model to the data. The EV can be expressed as:

$$EV = 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)} \tag{3.14}$$

7. *Tuning Time.* Tuning time refers to the total computational time required to optimize the model's hyper-parameters [100]. In real-world applications, especially with large datasets or complex models, the time taken to tune the model is a critical factor. It impacts the overall efficiency of the modeling process and the speed at which results can be delivered. Tuning time includes the duration of processes such as hyper-parameter optimization techniques (explained in Section 4.6.2) [100].

3.4.2 Cross-Validation (CV)

Cross-validation (CV) is a fundamental technique used in ML for hyperparameter tuning and model validation. K-fold CV is the most common approach to ascertaining the likelihood that a ML outcome is generated by chance. It is used to limit the problems like overfitting, underfitting and get an intuition how the model will generalize to an independent dataset [101]. CV involves partitioning the dataset into a specified number of subsets or "folds". In this thesis, 5-fold CV was used.

In 5-fold CV, the dataset is divided into five equal parts as visualized in Figure 3.6. For the 1^{st} iteration, the first fold is used for validation, and the remaining four folds are used for training. In the 2^{nd} iteration, the second fold is used for validation while the other four folds are used for training. This pattern continues through the 3^{rd}, 4^{th}, and 5^{th} iterations, with each iteration using a different fold for validation and the other folds for training.

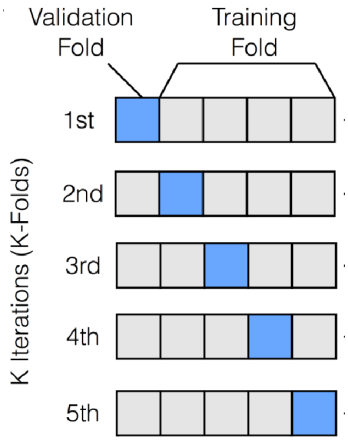


Figure 3.6: 5-Fold iteration CV [102].

The K-fold CV method reduces the variance associated with a single train-test split, providing a more comprehensive understanding of the model's performance on unseen data [103]. By leveraging multiple validation sets, CV ensures that the model generalizes well to new data, which improves the quality of predictions.

3.4.3 Semi-Structured Interviews

This thesis conducts interviews with experts from target groups to validate explainability results. Expert opinions provide unique insights, enhancing the understanding of the gathered results. This validation process involves conducting semi-structured expert interviews. Semi-structured expert interviews, originally from psychology and social sciences, are a well-established method that has been effectively applied in exploring requirements within both software engineering and ML [104]. This method stands between fully structured interviews, where predetermined questions are asked in a fixed order and rated using a standardized scoring system [105], and unstructured interviews, which are more open-ended and allow the conversation to flow naturally without a predefined set of questions (see Figure 3.7). The semi-structured interview approach offers some flexibility during the conversation, allowing experts to share their unique perspectives without being influenced by overly specific questions [104].

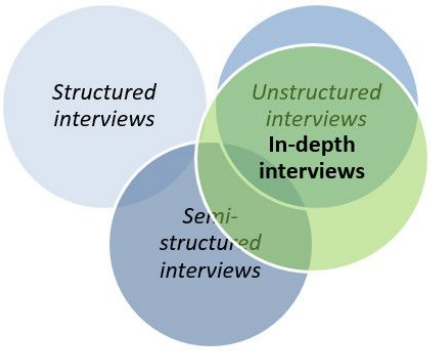


Figure 3.7: An overview of structured, semi-structured, and unstructured interviews [106]. As visualized, semi-structured interviews are a combination of both structured and unstructured interview concepts.

In the context of this thesis, semi-structured expert interviews serve several purposes. Firstly, these interviews try to provide validation for the explainability results derived from ML models. By engaging with experts who have domain-specific knowledge, the research can assess whether the model's explanations are meaningful and accurate from a practical standpoint. This process helps bridge the gap between technical model outputs and real-world applicability. Secondly, expert interviews play an important role in the understanding of the research findings. Experts provide valuable insights by interpreting the data, explaining trends, and pointing out potential implications that may not be immediately clear. Their feedback can help refine the research by identifying limitations, offering suggestions for improving the models, and proposing new avenues for future exploration. Moreover, engaging with experts ensures that the research aligns with current industry standards and needs, making it more relevant and practical. So, the input of experts strengthens the credibility and applicability of the findings.

3.5 Summary

This thesis employs the CRISP-DM framework to address a machine learning problem, focusing on a structured approach to translate business issues into ML tasks. The methodology handles six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These phases guide the project from problem definition to solution deployment, ensuring alignment with business objectives.

Furthermore, the study explores various ML techniques, including GLMs, LightGBM, XGBoost, and MLP. To ensure model interpretability, the thesis investigates SHAP and LIME as the main XAI techniques. Validation of the models is carried out using performance metrics such as MAE, RMSE, and Tuning Time, along with 5-fold CV to ensure generalizability. Additionally, semi-structured expert interviews are conducted to qualitatively assess the explainability of the models, allowing experts to offer insights and suggest improvements. This combined approach ensures that the models are both effective and understandable, trying to bridge the gap between technical outputs and practical application.

Chapter 4

Experimental Setup

4.1 Approach

To address the defined problems and research questions outlined in Section 1.2 and Section 1.4 respectively, it is necessary to establish a framework defining an approach for the project. This approach should facilitate the most effective way of retrieving the necessary insights and ensure that the data is thoroughly analyzed and interpreted. The framework ensures that the project is conducted systematically and follows the CRISP-DM methodology defined in Section 3.1. Within this methodology, a proper roadmap is provided to ensure retrieving necessary insights and data exploration prior to modeling. This will be handled per section, where every section handles a new part of the framework. Additionally, this section handles the coding environment used for this thesis.

4.1.1 Experimental Framework

As mentioned, a (structured) experimental framework is essential. The framework offers a roadmap for the project, ensuring that each phase is carefully planned and executed. The final framework and the steps taken are shown in Figure C.1, which can be found in Appendix C. As shown on the left in Figure C.1, the various CRISP-DM phases are displayed, making it easy to identify which processes fall under each phase. Every phase of the framework is discussed in this chapter, except deployment. The deployment phase is not covered in this chapter, as it involves steps taken after the model has been created, making it irrelevant to the experimental setup.

4.1.2 Coding Environment

The coding environment used for this study is Jupyter Notebook[1]. This choice was made primarily due to the advantages Jupyter Notebook offers for Python programming, which is the preferred language for this study. Additionally, research conducted on Kaggle[2] for similar financial problems indicates that notebooks are widely used as a preferred tool for both coding and easy visualization. As mentioned, Jupyter Notebook allows for an easy and interactive way of coding, where code can be written and executed in small, manageable blocks. This is particularly beneficial for debugging, as it allows for real-time testing and iteration on individual pieces of code. Additionally, Jupyter Notebook

[1]The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience [107].

[2]Kaggle allows users to find datasets, publish their own, collaborate with other data scientists and machine learning engineers, and participate in competitions to solve data science challenges [108].

allows for comprehensive documentation alongside the code, which enhances readability and understanding.

4.2 Business Understanding

This section handles the strategic importance of understanding the business environment of the thesis, the role of various departments within the organization, and the governance frameworks that guide the adoption of new models. The insights derived from a comprehensive business understanding will help the company's decision-making processes.

4.2.1 The Three Lines of Defence (3LoD) Model

Solvency II requires all insurers to have an effective governance system. Good governance in insurance companies is essential for controlled business operations and aims to ensure that the insurer is adequately managed, with risks being identified and controlled in a timely and appropriate manner [109]. In an insurance company, the adoption process for pricing models involves multiple departments, each with unique roles and perspectives on the necessity and depth of model explanations. This process is framed by the three lines of defence model (3LoD), which ensures comprehensive risk management and accountability throughout the organization. This method describes three lines within an organization: operational management (1st line), a controlling/coordinating/advisory function (2nd line), and internal audit (3rd line) [110]. An overview of the 3LoD model is presented in Figure 4.1.

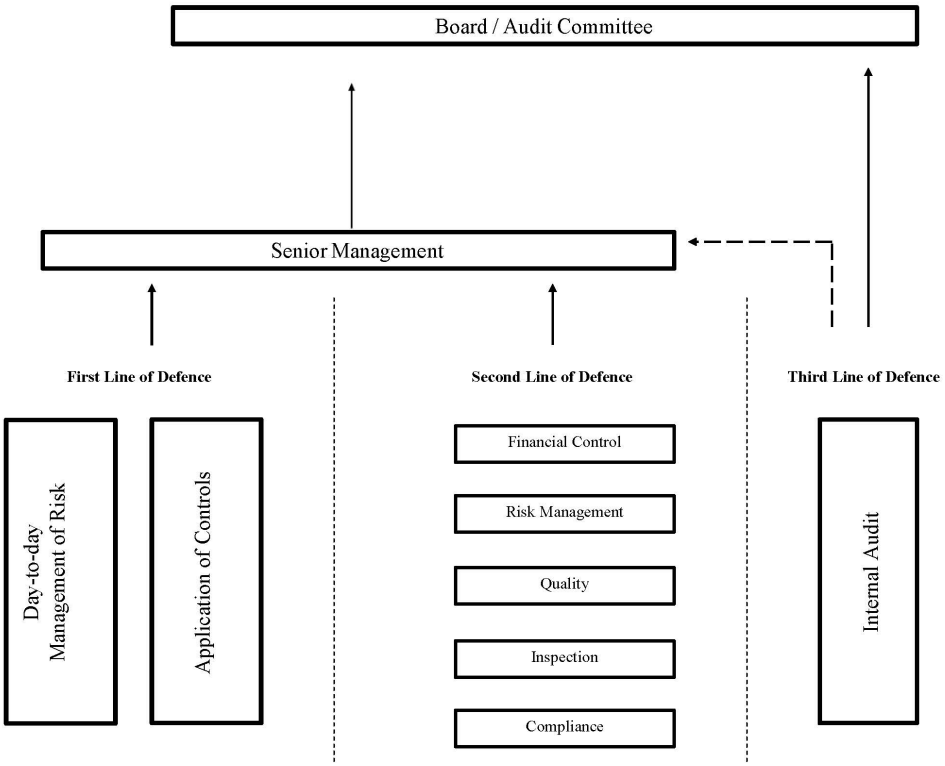


Figure 4.1: The three lines of defence (3LoD) model visualized [111].

4.2.2 Process Stakeholders

A way to identify different explainability expectations of the designed models is to identify relevant departments within each line of the 3LoD model discussed in Section 4.2.1. By interviewing these departments (stakeholders), a comprehensive business understanding of their views and concerns regarding the explainability of the newly proposed models can be gained. This helps ensure that the proposed models are tested on their ability to be transparent, trusted, and effectively utilized across the organization. An overview of the identified departments is based on the organisational structure of `the insurance company`, which has been visualized in Figure 4.2 and has been identified through an exploratory interview conducted (see Appendix E.1).

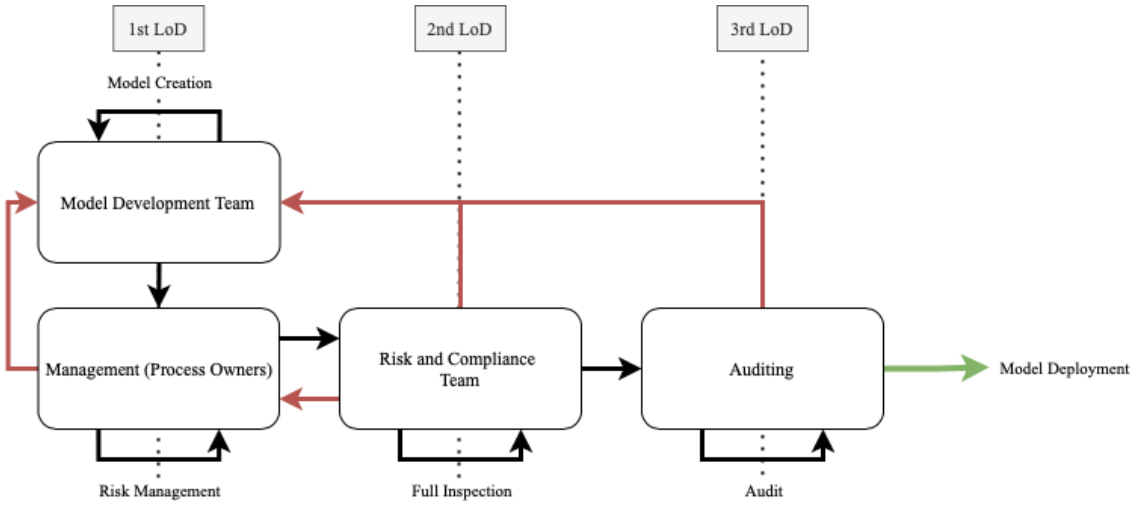


Figure 4.2: An overview of the process for a model based on the 3LoD model and information provided by `the insurance company`. For this thesis, one could look at this visualization for a premium pricing model deployment process. This is a conceptual representation, making this potentially differ compared to other companies.

The 3LoD model has three different stages of risk management, where the first line of defence (1st LoD) focus on day-to-day risks and the application of controls (see Figure 4.1) [111]. One could map two different departments for these roles, namely the *Model Development Team* and *Management*, who both need to look at the day-to-day risks accompanied with the creation of a model. At `the insurance company`, these respective departments are called *Data Management* (responsible for model development) and *Commercial Risk Management Committee* (a committee consisting of the company's management and commercial pricing analysts). The *Data Management* team focuses on the technical aspects of the model, including algorithm selection, data preprocessing, feature engineering, and performance metrics. This team understands the model's workings and ensure it performs optimally. Transparency in the model's development process is crucial for debugging, optimizing, and improving the model. The *Commercial Risk Management Committee* needs explanations that bridge technical details with business implications. The department requires clear insights into how risk premiums (or pure premiums) are evaluated and priced, ensuring that the model's outputs align with the company's risk profile and profitability goals. They look for a balance between technical accuracy and practical application in risk assessment.

The second line of defence (2nd LoD) is responsible for conducting a thorough re-

view of financial control, risk management, quality, inspection, and compliance (see Figure 4.1) [111]. For `the insurance company`, this function is carried out by external consultants provided by `the consultancy firm`. As shown in Figure 4.2, the *Risk and Compliance Team* at `the insurance company` is composed of an external team, though this structure is not a requirement for all companies. The *Risk and Compliance Team* looks into the entire model development process to ensure compliance with industry standards and regulations. This department requires comprehensive documentation and transparency in the coding and algorithmic decisions to validate that the model is built correctly and ethically. These actuaries also check for biases and regulatory adherence. `The consultancy firm` has the in-house expertise to perform these checks and does so for `the insurance company`.

Finally, the third line of defence (3rd LoD) contains internal actuaries, according to Figure 4.1 and Figure 4.2 [111]. *Internal Actuaries* perform yearly reviews to ensure that the models comply with internal policies and actuarial standards. The department requires detailed explanations that allow them to understand the model's assumptions, methodologies, and any changes made over the year. They focus on long-term consistency and reliability of the models. This way of working is the same for `the insurance company`.

Based on the identified stakeholders according to the 3LoD model, stakeholders can be categorized based on their placement within the three lines of defence. This division allows for specific questions regarding the explainability of both GLMs and ML models. Each line of defence has distinct responsibilities and perspectives, potentially necessitating specific types of explanations.

4.3 Data Understanding

This section outlines the process of acquiring and exploring of two different kind of datasets used in this study, one dataset focusing on the severity of claims and the other on the frequency of claims. Each dataset serves a distinct purpose in understanding insurance claims and their implications, and are explained below.

1. *Severity of claims.* This dataset records detailed information on the severity of insurance claims. It only includes records of individuals who have filed claims, resulting in a relatively smaller dataset. The severity data helps us understand the magnitude and cost associated with claims. See Section 4.3.3.1 for more information.
2. *Frequency of claims based on insurance policies.* This dataset encompasses the frequency of insurance claims and includes records for every person that has an insurance policy, regardless of whether they have filed a claim. This dataset is significantly larger than the severity dataset. The frequency data aids in analyzing the occurrence patterns of claims across the insured population. See Section 4.3.3.2 for more information.

Car insurance data is often split in multiple (main) causes that resulted in a claim. The main claim causes that are often registered are reparation, window damage, nature (wind, hail), theft, fire and injury. `The insurance company` has provided a part of their car insurance data (2016-2024) for the research of this paper. The year 2024, however, is not fully available and will be removed from the analysis. `The insurance company` supplied the datasets containing claim data for all policy claim types. When a client files a claim, a new entry is created in the claim dataset, capturing all important details of the incident. Concurrently, an update is made to the frequency dataset, incrementing the count of claims associated with the specific policy. This dual recording ensures a comprehensive

view of both individual claim details and the overall claim frequency linked to each policy. Section 4.3.3.1 and Section 4.3.3.2 discuss the chosen datasets for this thesis.

The primary goal of the premium pricing modeling process described in Section 4.2 is to accurately predict both the severity and frequency of claims. An employee of `the insurance company` emphasizes the importance of focusing on accurately predicting the majority of cases rather than the few outliers. This focus should be taken into account when validating the developed models.

4.3.1 Naming Conventions

Section 2.1.3 mentioned a few main causes for claims, such as damage to others, theft and break-ins, window damage, weather or fire damage, damage caused by yourself, and vandalism. `The insurance company` employs a different naming strategy for the same kind of cause types. An example: `the insurance company` categorizes "damages caused by yourself" as *"reparation"* (in Dutch: *"reparatie"*). So, from now on, damages caused by yourself will be referred to as reparation data. An important mention is that the reparation data is only a subset of all policy data. The data per policy claim type consists of two datasets: the claim frequency data and the claim severity data, following the frequency-severity method described in Section 2.2.2.2. Table 4.1 shows an overview of all naming conventions used for the claim cause types within `the insurance company`, while also describing whether certain datasets of cause types will be handled within the thesis. `The insurance company` does not use an individual model for vandalism.

Table 4.1: Mapping the taxonomies of claim causes (identified by Section 2.1.2) with the naming conventions used by `the insurance company`. Multiple "..."'s within a column mean that this cause type was divided into multiple datasets.

Type of Cause	Naming Convention of the insurance company [EN]	Naming Convention of the insurance company [NL]	Handled in Thesis [Yes/No]
Damage to others	"Damage to others - Material damage" & "Damage to others - Injury damage"	"WA - Materieel" & "WA - Letsel"	No
Theft and break-ins	"Theft" & "Break-ins"	"Diefstal" & "Inbaak"	No
Window damage	"Window"	"Ruit"	No
Weather or fire damage	"Weather" & "Fire"	"Weer" & "Brand"	No
Damage caused by yourself	"Reparation"	"Reparatie"	Yes
Vandalism	-	-	No

For modeling the claim severity and frequency for these claim types, `the insurance company` utilizes a GLM model built on the respective data linked to the cause. The GLMs and their underlying error distribution differ for claim frequency and claim severity data. After one of the datasets has been chosen, the data needs to be prepared for the steps to follow. For some of the datasets identified in Table 4.1, GLMs are not utilized due to insufficient data. This thesis only uses datasets with a reasonable amount of data.

4.3.2 Data Confidentiality

Due to confidentiality and non-disclosure agreements not all descriptive statistics can be shown. Features in both datasets have been anonymized (highlighted with the following font indication: `anonymized`) as well based on their respective column number within the

dataset. This would mean that a certain `Feature X`, which takes the 21st column within the dataset, would be `Feature 21`. Also, informative information about the feature is removed as well. This means that any personally identifiable information or sensitive data has been removed or obscured to protect the privacy of individuals. Despite this anonymization, the integrity and utility of the data for analysis are maintained, allowing us to extract meaningful insights while ensuring compliance with privacy and data protection standards as mentioned in Section 2.6. In addition to the anonymized results presented in the thesis, the respective companies are receiving the original results directly.

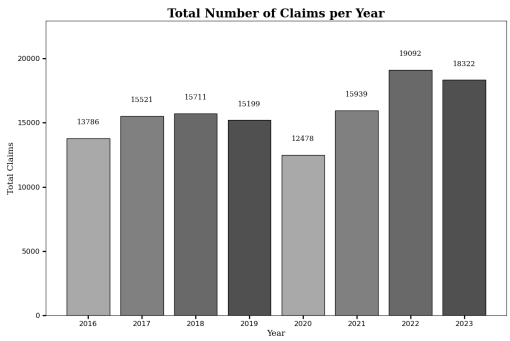
4.3.3 Data Exploration

This section focuses on the Reparation data provided in the datasets from `the insurance company`. The dataset is divided into two main categories: severity data and frequency data. The following subsections will explore these datasets in detail, providing a comprehensive analysis of their characteristics, distributions, and any notable trends or patterns. This exploration establishes a solid foundation for the subsequent analysis and modeling work in the thesis.

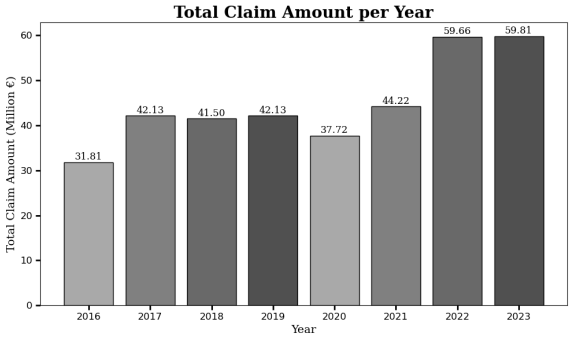
4.3.3.1 Severity Data

The full dataset for claim severity analysis consists of 128,212 rows and 162 columns. This dataset eventually contains 116,918 entries due to the removal of data gathered in the year 2024 due to incompleteness. Each row represents an individual claim, while the 162 columns represent various features relevant to the claims. The features include policyholder demographics, vehicle characteristics, claim data, and other factors that might influence the severity of the claims. The primary target feature in this dataset is *"claim burden"* (in Dutch: *"schadelast"*), which represents the cost associated with each claim. This target feature is critical for modeling and predicting the financial impact of claims.

Figure 4.3a shows that the number of claims fluctuates over the years, with a noticeable drop in 2020, likely due to the impact of the COVID-19 pandemic. However, the claim numbers significantly increase in 2022 and 2023, indicating a potential rise in incidents or changes in reporting behavior. The total claim amount, shown in Figure 4.3b, shows variability between 2016 (€31.81 million) and 2023 (€59.81 million), with a significant increase shown in 2022 and 2023. This rise in claim amounts could be due to the higher claim frequency identified, increased claim severity, inflation, or changes in claim processing and payouts. It could also be due to increased policy numbers at `the insurance company`.



(a) The total number of claims over the years.



(b) The total claim amount over the years.

Figure 4.3: An overview of the claim (severity) data over the years at `the insurance company`.

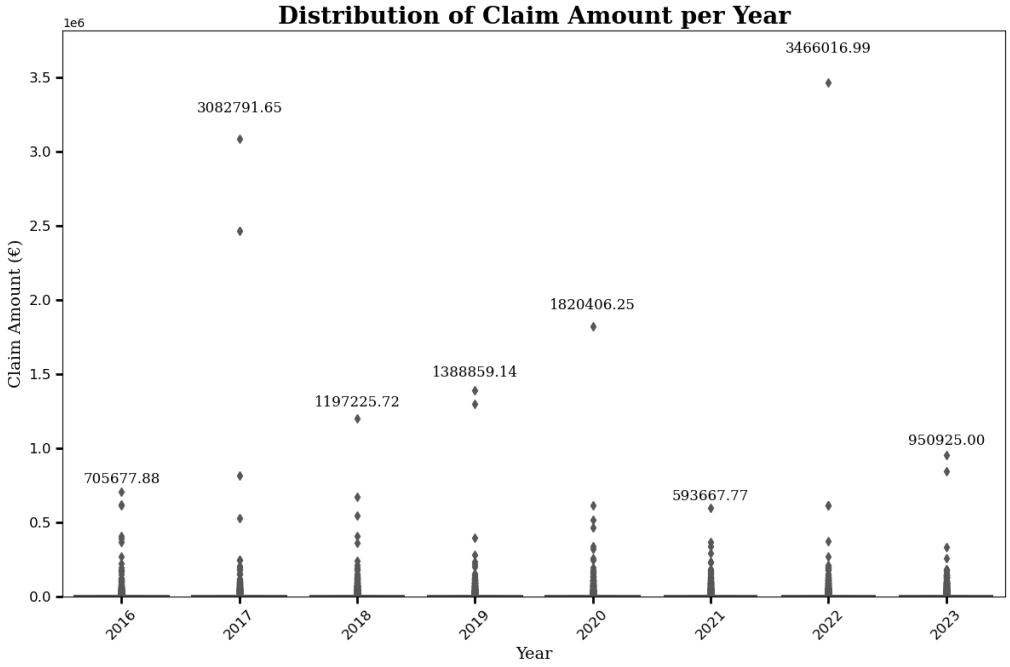


Figure 4.4: An overview of the distribution of claim amounts over the years at `the insurance company`.

Car insurance data can have heavy outliers, as shown in Figure 4.4 with the numerical values displayed for all types of claims. These outliers can be actual outliers, but can also be mistakes. It is no simple task to automate the checking of outliers for validity. Therefore, most outlier values are checked by hand. In case of doubt, `the insurance company`'s data analysts provided additional information regarding the outlier and whether to include them or not. According to `the insurance company`, the focus of this thesis should be on accurately predicting the gross number of claims. Consequently, Figure 4.5 provides a better representation of the data points that are the focus of this thesis. It is also interesting to know that most outliers are caused by physical damage towards a person, rather than car damages themselves. According to `the insurance company`, these physical damages

can indeed be extremely high. An example would be the (reimbursed) claim of €3.47 million in 2022.

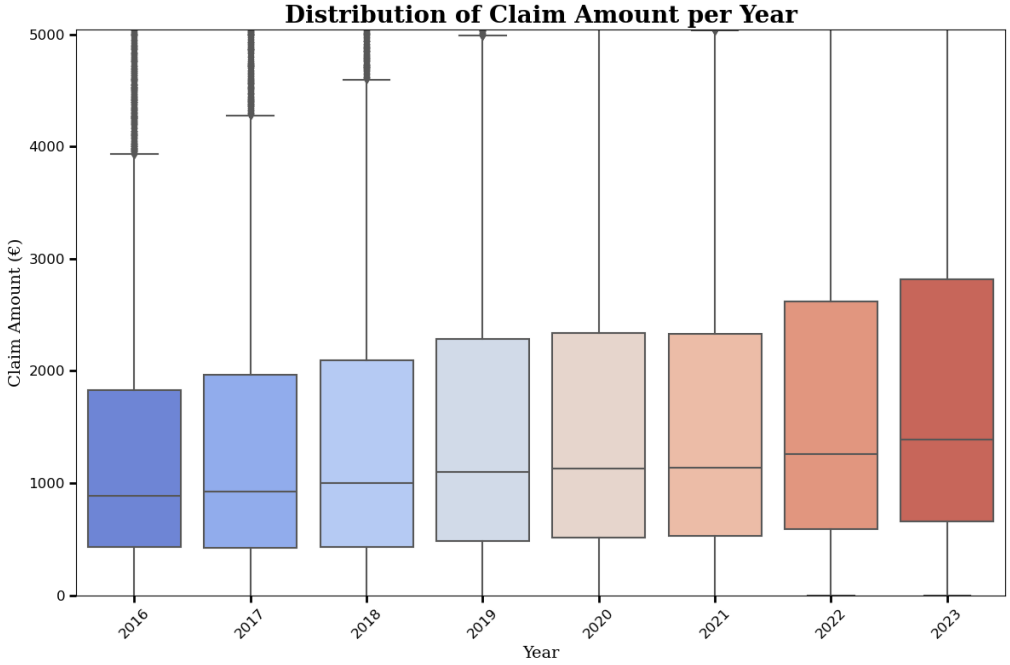


Figure 4.5: An overview of the distribution of claim amounts over the years at `the insurance company`, focused on the gross number of claims.

Each box in the plot represents the interquartile range (IQR) of the claim amounts for a given year, depicting the middle 50% of the data, with the horizontal line inside each box indicating the median claim amount. Starting from 2022 to 2023, there is a noticeable increase in both the median claim amount and the IQR, indicating higher and more variable claim amounts. A reason for this is that `the insurance company` sees a trend in new technological advances in cars which tend to have higher claim damages. This trend was previously identified in the analysis of Figure 4.3, which also showed an increase in both the total number of claims and the total claim amount during these years.

(a) Word cloud of car brands.

(b) Pie chart of biggest car brands.

Figure 4.6: An overview of the various car brands within the `the insurance company`'s portfolio.

One could also look at the various car brands `the insurance company` manages. The word cloud shown in Figure 4.6 illustrates the diverse range of car brands within the portfolio, with the size of each brand's name reflecting its level of representation. The largest names in the word cloud are "BMW", "AUDI", "MERCEDES-BENZ", "PORSCHE", and "TESLA", signifying that these brands hold the highest representation within the insurer's portfolio. Their prominent size highlights their significance in `the insurance company`'s portfolio management.

Reparation Severity Data

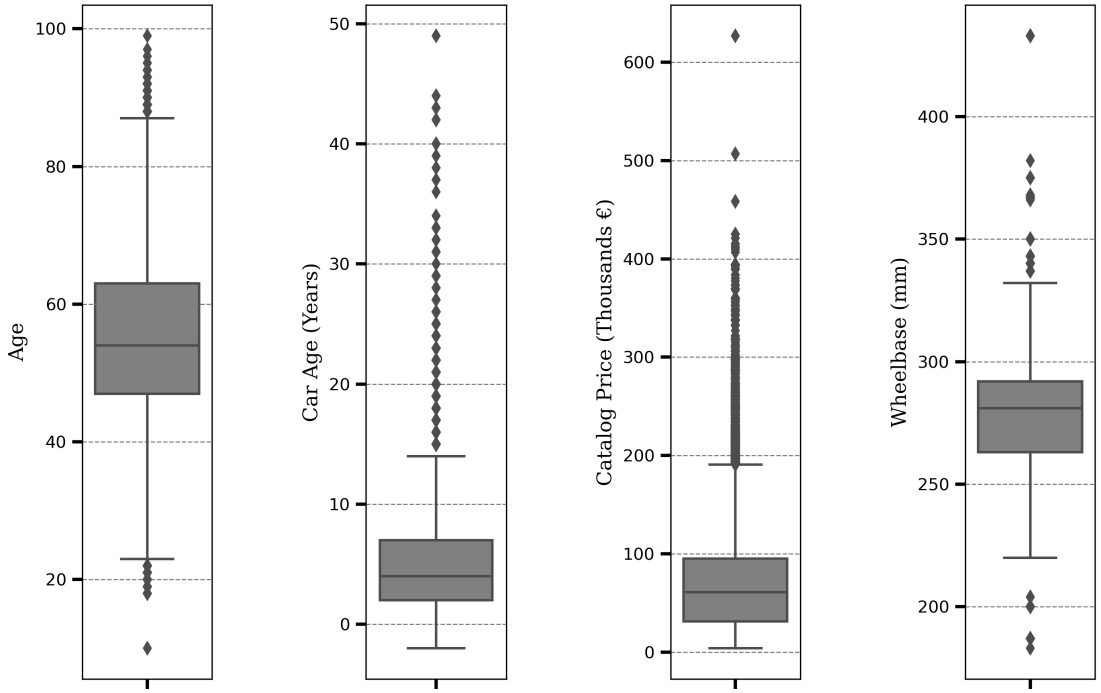
The reparation severity dataset is the largest (severity) dataset available at `the insurance company`, providing an good foundation for exploring the potential of ML. The reparation severity dataset, initially comprising 52,353 rows and 162 columns, included several additional features created through manual inclusion compared to the initial dataset. These new features enhanced the dataset's analytical potential. However, for privacy and compliance reasons, all these features will be anonymized in the final results. It is important to note the use of the logarithmic value for the insured amount. For instance, a car valued at €200,000 does not necessarily carry twice the risk of a car valued at €100,000. By applying a logarithmic function, `the insurance company` aims to account for this non-linear relationship, ensuring a more accurate assessment of risk.

An overview is given by some features found in the dataset, which is done in Table 4.2. These descriptive features are just a few of the 162 features existing in the reparation severity dataset. Boxplots of the few selected features are provided in Figure 4.7. Due to confidentiality and non-disclosure agreements not all descriptive statistics can be shown, but the chosen features show a good overview of the types of data `the insurance company` has access to.

1. *"Age insured"* (in Dutch: *"Leeftijd verzekerd persoon"*). See Figure 4.7a.
2. *"Car age"* (in Dutch: *"Leeftijd auto"*). See Figure 4.7b.
3. *"Catalog value vehicle"* (in Dutch: *"Cataloguswaarde auto"*). See Figure 4.7c.
4. *"Wheelbase vehicle"* (in Dutch: *"Wielbasis auto"*). See Figure 4.7d.

Table 4.2: Descriptive statistics for some of the features within the reparation severity dataset at `the insurance company`.

Feature	Quantile 1	Median	Quantile 3
Age insured	47	54	63
Age vehicle	2	4	7
Catalog value vehicle	€31,145	€60,812	€94,996
Wheelbase vehicle	263.0	281.0	292.0
And 158 more features...			



(a) Age insured. **(b)** Age vehicle. **(c)** Catalog value vehicle. **(d)** Wheelbase vehicle.

Figure 4.7: Boxplots of selected features for the reparation severity dataset at `the insurance company`.

The boxplots in Figure 4.7 and the descriptive statistics in Table 4.2 reveal various insights into the reparation severity dataset: the insured population predominantly consists of middle-aged to older individuals. Also, most vehicles are relatively new with some significantly older ones. Vehicle catalog values vary widely, where most cars (approximately 75%) are below €100,000, with some very high-value cars worth over €100,000. The wheelbase of vehicles shows considerable diversity, with a median wheelbase of 281 mm.

It is important to note that the features discussed are just 4 out of the 162 features in the reparation severity dataset. The final dataframe used for determining reparation outcomes will involve some preparation steps before being handled. This is handled in Section 4.4.

4.3.3.2 Frequency Data

The dataset for claim frequency analysis consists of 7,801,559 rows and 144 columns. Each row represents a policy, while the 144 columns represent various features relevant to the policies. The features include policyholder demographics, vehicle characteristics, claim data, and other factors that might influence the frequency of the claims. The primary target feature in this dataset is *"Number of Claims {Type}"* (in Dutch: *"Aantal Schades {Type}"*), which represents the number (frequency) of claims associated with each policy at `the insurance company`. This target feature is critical for modeling and predicting the estimated frequency of the claims. {Type} can differ for the policy claim type identified, such as "Reparation" (see Table 4.1 in Section 4.3.1 for the type of cause and its respective naming convention).

As discussed, the total claim frequency dataset includes a variety of features. Some of these features can be explored further for analytical purposes.Figure 4.8 presents the percentage trends of different fuel types used in vehicles over time, from the year 2000 to 2023, focused on the main three fuel types *"Petrol"* (in Dutch: *"Benzine"*), *"Diesel"* (in Dutch: *"Diesel"*), and *"Electricity"* (in Dutch: *"Electriciteit"*). The y-axis represents the percentage of total policies, while the x-axis represents the year. The legend on the right indicates the fuel types, which are color-coded for the main three fuel types for easy identification.

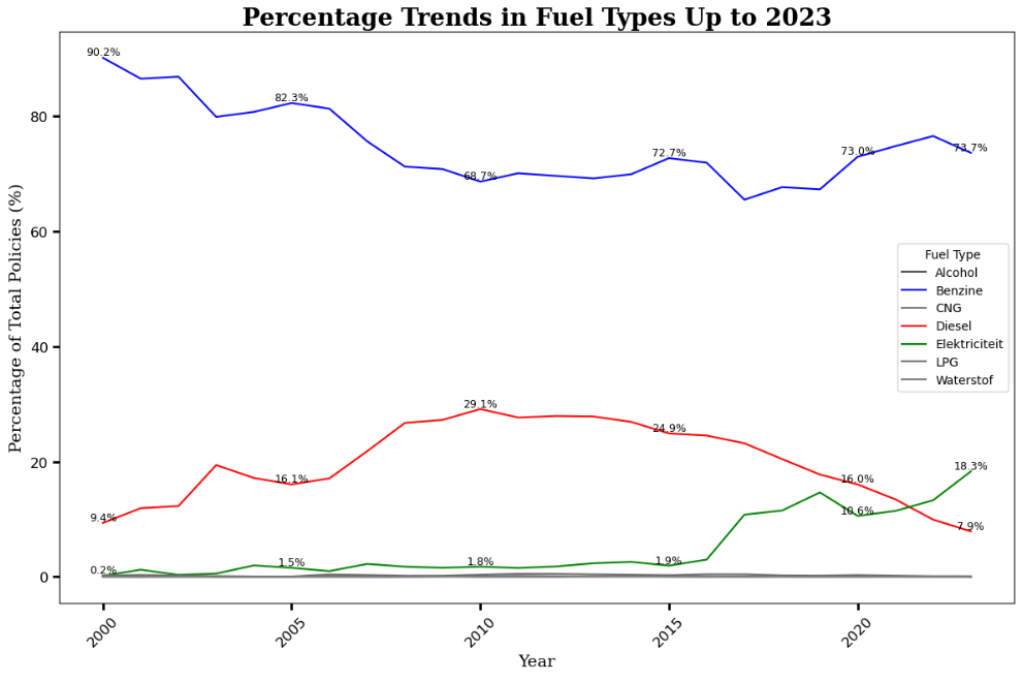


Figure 4.8: An overview of the distribution of fuel type of cars over the years at `the insurance company`.

Petrol remains the most prevalent fuel type, though its usage has declined over time, from 90.2% at the start of the 2000's and 73.7% at the end of 2023. Diesel saw significant growth until around 2010 (29.1%) but has since decreased, while electricity has seen rapid growth, especially in recent years, overtaking diesel fueled cars. The total percentage of electric type cars within `the insurance company`'s portfolio now stands at 18.3% at the end of 2023. An article by Osinga stated that the main reason for higher claim amounts at insurance companies could be that there are more and more electric cars on the road. These cars are more expensive to repair than gasoline cars [112]. This could be another indicator for the higher claim amount in the recent years which is shown in Figure 4.3a. Additionally, car parts are becoming increasingly advanced and therefore more expensive to replace, which is validated by `the insurance company` as well [112].

Reparation Frequency Data

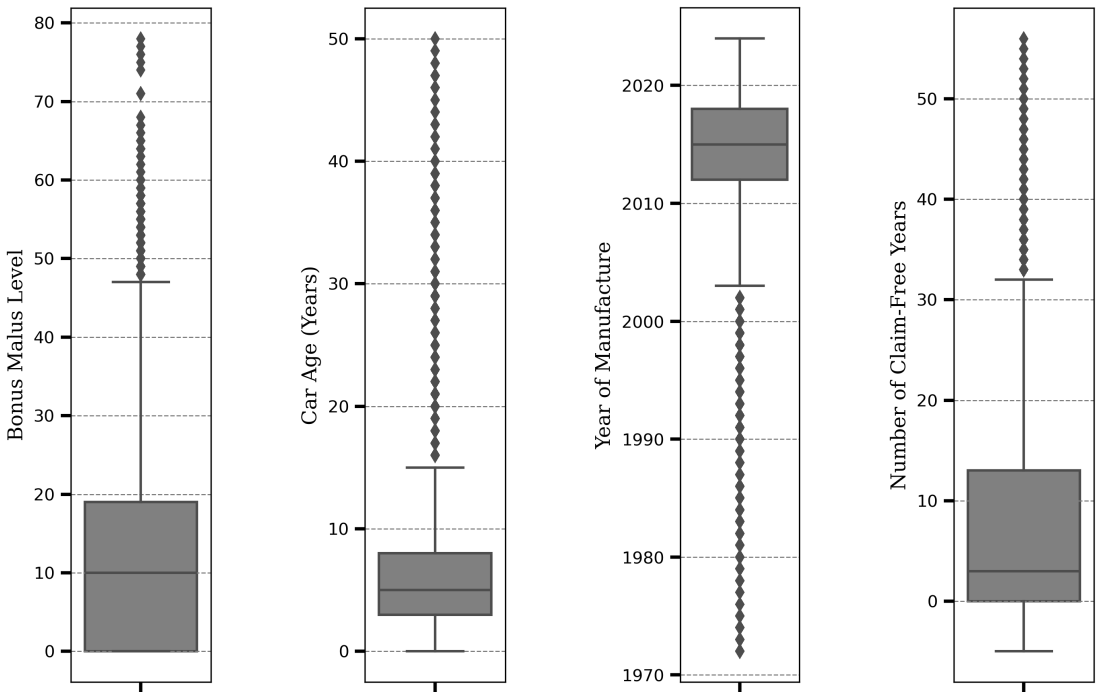
The reparation frequency dataset is one of the largest (frequency) datasets available at `the insurance company`, providing an excellent foundation for exploring the potential of ML, just like the severity dataset. The reparation frequency dataset, initially comprising 7,801,559 rows and 144 columns, included several additional features created through manual inclusion compared to the initial dataset based on a feature engineering strategy by the

insurance company. An overview can be given by some features found in the dataset, which is done in Table 4.3. These descriptive features are just a few of the 144 features existing in the reparation severity dataset. Boxplots of the few selected features are provided in Figure 4.9. Due to confidentiality and non-disclosure agreements not all descriptive statistics can be shown, but the chosen features show a good overview of the types of data `the insurance company` has access to.

1. *"Bonus-malus Level"* (in Dutch: *"Bonus-malus Level"*). See Figure 4.9a.
2. *"Age vehicle"* (in Dutch: *"Leeftijd Auto"*). See Figure 4.9b.
3. *"Year of manufacture"* (in Dutch: *"Bouwjaar"*). See Figure 4.9c.
4. *"Number of claim-free years"* (in Dutch: *"Aantal schadevrije jaren"*). See Figure 4.9d.

Table 4.3: Descriptive statistics for some of the features within the reparation frequency dataset at `the insurance company`.

Feature	Quantile 1	Median	Quantile 3
Bonus-malus level	0	10	19
Age vehicle	3	5	8
Year of manufacture	2012	2015	2018
Number of claim-free years	0	3	13
And 140 more features...			



(a) Bonus-malus level.

(b) Age vehicle.

(c) Year of manufacture.

(d) Number of claim-free years.

Figure 4.9: Boxplots of selected features for the reparation frequency dataset at `the insurance company`.

The boxplots visualized in Figure 4.9 provide a visual summary of the distribution of four important features: bonus-malus level, age of vehicle, year of manufacture, and the number of claim-free years. These boxplots reveal the spread of the data within the reparation frequency dataset. The distribution of the bonus-malus level shows a median value around 10, with a range extending from 0 to nearly 80. The first and third quartiles are at 0 and 19, respectively, indicating that while many individuals have a low bonus-malus level, there is a significant spread among those with higher values. The presence of several outliers above 50 suggests that some drivers are considered very high risk, which could have a significant impact on reparation frequency.

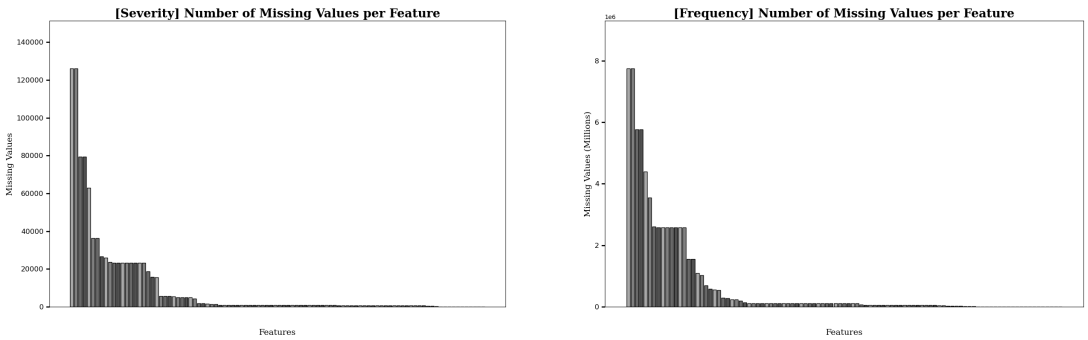
It is important to note that the features discussed are just 4 out of the 144 features in the reparation frequency dataset. The final dataframe used for determining reparation outcomes will involve some preparation steps before being handled. This is handled in Section 4.4.

4.4 Data Preparation

Data preparation is essential for ensuring the quality and reliability of the ML model results. The process begins with data cleaning, which is done to address any inconsistencies or errors in the data. Following this, a train-test split is performed to divide the data into subsets for training and testing. Encoding and scaling techniques are then applied to convert categorical and numerical variables respectively into a suitable format for ML algorithms.

4.4.1 Cleaning

It is important to understand the initial data cleaning process that was undertaken. Data cleaning is an essential step in the data preprocessing pipeline, as it ensures the integrity and quality of the data used for the modeling process. The initial assessment of the datasets revealed a significant presence of missing values within the total dataset structure for both severity and frequency, as can be seen in Figure 4.10. Addressing these missing values was a priority, as they can negatively affect the performance of ML models.



(a) The total number of NAs in the severity dataset.

(b) The total number of NAs in the frequency dataset.

Figure 4.10: An overview of the NAs with the provided datasets at `the insurance company`.

The presence of a significant number of missing values, shown in Figure 4.10, highlights the need for a proper data preprocessing strategy. However, it's important to note

that the features with the highest number of missing values are not incorporated into the final feature selection, ensuring that these missing values do not affect the model's performance. The chosen features that are based on a preprocessing strategy predefined by `the insurance company`[3] actually have no missing values, which eliminates the concern of missing data impacting the accuracy and reliability of subsequent data analysis or modeling efforts. Therefore, one could say that 100% of the chosen feature data was usable in this thesis. Table 4.4 and Table 4.5.

Table 4.4 shows the process from the initial dataset (ID) to the final number of rows of the dataset per claim type (SD). The removal of 2024 (YR) is also considered, as this year is incomplete within the dataset. Therefore, this data is not used within the thesis. Table 4.5 illustrates the process of refining the dataset based on claim type, leading to a reduction in the number of columns and potential removal of NA values. Notably, no rows are eliminated during this process. However, there is a substantial reduction in the number of columns due to the feature selection methodology employed by `the insurance company`. This results in the usage of the most important variables only, thus giving a low percentage of used columns in Table 4.5. However, due to this methodology, only columns with rows containing 100% information were used, thus making sure no cleaning was necessary for the rows (100% selection). The current methodology opens some possibilities for the incorporation of additional features in the future.

Table 4.4: Row information of the reparation dataset for severity and frequency after determining the selected data (SD) gathered from the initial full dataset (ID), where also the size of the dataset after the removal of the year 2024 is indicated (YR).

	Severity	**Frequency**
Number of Rows (ID)	128,212	7,801,559
Number of Rows (YR)	116,918	7,270,276
Number of Rows (SD)	52,353	2,633,546
Selected Data (%)	**44.78**	**36.22**

Table 4.5: Final row and column information of the datasets per claim type for severity and frequency before cleaning (BC) and after cleaning (AC).

	Severity	**Frequency**
Number of Rows (BC)	52,353	2,633,546
Number of Rows (AC)	52,353	2,633,546
Used Rows (%)	**100**	**100**
Number of Cols (BC)	162	144
Number of Cols (AC)	6	7
Used Cols (%)	**4.32**	**4.86**

[3]In this thesis, the specific choice has been made to replicate the preprocessing strategy (for feature selection used by `the insurance company`. This approach allows for the validation of potential improvements made by ML models, as it provides a known baseline of GLM performance from the company. Consequently, it enables a proper comparison between the performance of the GLM model and that of the ML models. The feature selection methodology is based on experts and can not be disclosed.

4.4.2 Train-Test Split

To ensure robust model evaluation and to prevent bias, a 70-30 train-test split was employed on the data before the modeling process [42]. This means that 70% of the available (used) data (see Table 4.5) was used to train the ML models, while the remaining 30% was reserved for testing and validating the models' performance. The 70-30 split is a commonly used ratio in ML and data science, just like the 80-20 split, balancing the need for a sufficiently large training set to accurately learn the underlying patterns in the data, while also retaining a substantial portion of data to evaluate the model's performance on unseen data [44]. The chosen train-test method helps in assessing how well the model generalizes to new, independent data. During the training phase, the model learns from the training data (70%) by identifying patterns and relationships. Once training is complete, the model's performance is evaluated on the test data (30%) based on the metrics defined in Section 3.4.1, with a specified focus on the metrics defined in Section 4.6.3.1.

There was specifically opted for a 70-30 split, rather than the other commonly used 80-20 split, due to the variable nature of insurance claims. This decision ensures that the test set contains a relatively higher proportion of new variables and claim scenarios, which helps to challenge the model's generalization capabilities. By increasing the size of the test set, the study aims to reduce the risk of overfitting on the initial training data, thereby providing a more robust evaluation of the model's performance in predicting unseen data. This approach is particularly important in the insurance domain, where the diversity and unpredictability of claims can significantly impact the model's accuracy and reliability.

Additionally, stratified sampling was utilized during the train-test split to ensure that the distribution of classes in both the training and test datasets reflects the original dataset's distribution. Stratified sampling is particularly important in cases where the data might be imbalanced, such as when certain classes or categories are underrepresented [43]. This technique enhances the reliability of the model's performance metrics by preventing scenarios where the model might perform well on the test set simply due to an unbalanced or unrepresentative split.

4.4.3 Encoding

In the dataset, one can identify numerical and categorical variables. A categorical variable refers to a variable that is divided into distinct groups or categories, such as gender ("Male", "Female" or "Other"). In contrast, a numeric variable is one that can be quantified and measured, such as height, weight, or speed. Categorical variables in the dataset need to be converted into a numerical format to be used in ML models [71]. There are two primary methods for this transformation: One-Hot Encoding (OHE) and Label Encoding (LE). OHE involves creating a new binary column for each category in the original variable. Each binary column will have a value of "1" or "0", indicating the presence or absence of the category in the observation [71]. Each label for a categorical factor is assigned a (typically) $1 \times n$ vector representation that can be trained with the model [113]. LE converts each category in the variable into a numerical label. Each unique category is assigned a unique integer. This approach is useful when there is an ordinal relationship[4] between categories.

In this study, OHE is chosen as the encoding method because some of the categorical features in the dataset do not have a natural ordinal relationship. Using OHE ensures

[4]Ordinal data can be classified into categories that are ranked in a natural order. Ordinal data often include ratings about opinions or feelings or demographic factors like social status or income that are categorized into levels [114].

that these non-ordinal categorical variables are represented in a way that avoids imposing any artificial ranking, allowing the model to treat each category as distinct and equally important.

4.4.4 Scaling

Numerical variables in the dataset need to be scaled to ensure that each feature contributes equally to the model's performance. This thesis identifies common methods for scaling numerical variables: Standard Scaling (SS)[5], Min-Max Scaling (MMS)[6], and Robust Scaling (RS)[7]. SS transforms the data to have a mean (μ) of 0 and a standard deviation (σ) of 1 [115]. MMS rescales the data to a fixed range, typically [0, 1]. This method is sensitive to outliers, as it uses the minimum and maximum values for scaling [115]. RS uses the median and the interquartile range (IQR) for scaling, making it robust to outliers [115]. This method is particularly useful when the data contains outliers that might skew the scaling process. By choosing the appropriate scaling method, numerical variables can be effectively normalized.

For this thesis, the choice is made to use RS, as Section 4.3.3.1 highlights the importance of accurately predicting the gross number of claims, making outliers less important. By utilizing the median and IQR for scaling, RS ensures that the influence of outliers is minimized, providing a more accurate representation of the data. This approach allows the model to focus on the gross of the data without being affected too much by extreme values, leading to more reliable model performance for this specific case.

4.5 Model Exploration

In the initial phase of model exploration, H2O AutoML[8] is utilized to evaluate a variety of ML models and determine which ones perform the best. This automated process facilitates a thorough examination of multiple models, enabling the identification of the most promising candidates for further development and tuning. During this exploratory analysis, most variables are considered without variables accounting for potential issues such as bias, data leakage, or ethical considerations. This approach ensures that no variables that might contribute to the model's predictive power are prematurely excluded. It's important to emphasize that the outcomes of this approach are not intended for practical solutions. Instead, the focus is on gaining insights into the potential of ML and identifying possible techniques to explore.

4.5.1 H2O AutoML: Automatic Machine Learning

H2O AutoML is a highly scalable, automated ML platform designed to simplify the process of building and deploying ML models. It provides a comprehensive solution that includes robust data preprocessing, diverse model training, and powerful ensemble techniques, all accessible through a simple interface in multiple programming languages [116]. In terms of performance, H2O AutoML demonstrates competitive accuracy.

H2O AutoML automatically handles data preprocessing tasks such as imputation, normalization, and OHE, ensuring that the data is prepared effectively for model training [116].

[5] SS: $z = \frac{(x-\mu)}{\sigma}$
[6] MMS: $x' = \frac{(x-x_{min})}{(x_{max}-x_{min})}$
[7] RS: $x' = \frac{(x-\text{median})}{\text{IQR}}$
[8] https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

It also plans to support advanced techniques like target encoding for high-cardinality features and text encoding in future releases, which will further enhance its preprocessing capabilities [116]. For model training, H2O AutoML uses various algorithms[9]. The system employs random search for tuning of hyperparameters and creates stacked ensembles. The AutoML workflow involves several stages including data preparation, feature engineering, model selection and hyperparameter optimization, as described by Salehin et al. [117].

In the end, H2O AutoML generates a leaderboard as a final result that ranks all the models based on their performance metrics [116]. This feature allows users to easily identify the best-performing models. The leaderboard is customizable, enabling users to sort models by different metrics, such as accuracy, training time, or prediction speed, providing comprehensive insights into model performance. This is what is used in Section 4.5.2 and Section 4.5.3 for investigating the outcomes of the exploratory modeling phase.

4.5.2 AutoML: Severity Model

This section presents the detailed results of the AutoML process for the reparation severity data of `the insurance company`, highlighting the performance metrics of each model, including RMSE, MSE, MAE, RMSLE, and Mean Residual Deviance. The primary objective was to identify models capable of making reliable predictions, so that these models can be potentially used within the thesis domain for severity modeling.

Table 4.6: Model performances of the reparation severity dataset from `the insurance company` found via AutoML. *Model ID* is abbreviated due to space issues. Values of interest are highlighted in blue.

Model ID	RMSE	MSE	MAE	RMSLE	Mean Residual Deviance
GBM_grid_model_7	5843.04	3.41411e+07	2551.16	8.964702	3.41411e+07
GBM_grid_model_11	5844.59	3.41651e+07	2541.07	NaN	3.41651e+07
GBM_grid_model_9	5846.84	3.41762e+07	2536.89	8.951579	341759949
GBM_grid_model_2	5854.40	3.4274e+07	2541.74	8.956145	3.4274e+07
GBM_grid_model_3	5866.16	3.44119e+07	2548.71	0.953743	3.44119e+07
GBM_grid_model_17	5876.75	3.45362e+07	2547.59	8.95398	3.45362e+07
GLM_1_223801	5893.30	3.4731e+07	2545.61	8.952905	3.4731e+07
GBM_3_223801	5903.30	3.48489e+07	2548.61	8.950843	3.48489e+07
XRT_1_223801	5911.36	3.49442e+07	2622.18	8.962985	3.49442e+07
GBM_2_223801	5921.52	3.50644e+07	2549.68	NaN	3.50644e+07
GBM_grid_model_4	5931.99	3.51885e+07	2585.45	NaN	3.51885e+07
GBM_grid_model_12	5937.98	3.52596e+07	2548.65	8.958086	3.52596e+07
GBM_4_223801	5939.13	3.52733e+07	NaN	NaN	NaN
DeepLearning_1_223801	5947.52	3.5373e+07	2467.48	NaN	3.5373e+07
GBM_grid_model_14	5951.61	3.54216e+07	2570.74	0.95559	3.54216e+07
GBM_grid_model_6	5954.64	3.54577e+07	2572.50	8.953687	3.54577e+07
GBM_grid_model_16	5964.49	3.55751e+07	2577.82	NaN	3.55751e+07
GBM_grid_model_1	5970.50	3.56468e+07	2616.43	NaN	3.56468e+07
GBM_1_223801	5974.52	3.56949e+07	2580.12	NaN	3.56949e+07
GBM_grid_model_10	5980.43	3.57656e+07	2596.67	NaN	3.57656e+07
GBM_grid_model_15	6021.18	3.62546e+07	2619.87	8.965174	3.62546e+07
GBM_grid_model_5	6828.68	3.63449e+07	2596.48	NaN	NaN
DRF_1_223801	6846.62	3.65616e+07	2654.96	8.965077	NaN
GBM_grid_model_8	6139.53	3.76938e+07	2664.07	NaN	3.76938e+07
GBM_grid_model_13	6152.94	3.78587e+07	2682.62	NaN	3.78587e+07

[9] Algorithms included: Gradient Boosting Machines (GBMs), Distributed Random Forests (DRFs), Extremely Randomized Trees (XRT), Deep Neural Networks (DNNs), and Generalized Linear Models (GLMs) [116].

The DL algorithm achieved the lowest MAE of 2467.48, which is significantly better than other models, indicating that it has the smallest average error between the predicted and actual values. It suggests that these models can be useful for this dataset, particularly in minimizing prediction errors. The GBMs achieved the lowest RMSE, which was 5843.04, which implies that this model has the lowest average squared error among the evaluated models. The fact that nearly every model in the table is a GBM highlights the robustness of GBM as a potential ML method. Only one GLM is shown among the 25 models evaluated. The GLM produced an RMSE of 5893.30 and an MAE of 2545.61, which are respectable but not the best compared to the top-performing ML models. This indicates that while GLM is a solid baseline model, the more advanced ML models, particularly tree-based models like GBMs and DL, have the potential to outperform GLMs.

These results provide insights into selecting ML models for further experimentation. The performance of tree-based models (such as LightGBM, XGBoost) and DL models (could be related to NNs) in various metrics suggests that these should be prioritized in the ML algorithm choice in the experimental setup. The GLM model, while not the top performer, still offers a useful comparison and can serve as a baseline model to benchmark the ML algorithms, which will elaborated in Section 4.6.1.

4.5.3 AutoML: Frequency Model

This section presents the detailed results of the AutoML process for the reparation severity data of `the insurance company`, highlighting the performance metrics of each model, including RMSE, MSE, MAE, RMSLE, and Mean Residual Deviance. The primary objective was to identify models capable of making reliable predictions, so that these models can be potentially used within the thesis domain for frequency modeling. There are less models in the final result presented in Table 4.7, as these ML models took much longer to process due to the dataset size differences between the severity data and frequency data.

Table 4.7: Model performances of the reparation frequency dataset from `the insurance company` found via AutoML. *Model ID* is abbreviated due to space issues. Values of interest are highlighted in blue.

Model ID	RMSE	MSE	MAE	RMSLE	Mean Residual Deviance
DRF_1_208639	0.203278	0.0413221	0.0745038	0.136162	0.0413221
GBM_1_208639	0.203493	0.0414896	0.0747609	0.136227	0.0414096
GBM_4_208639	0.203500	0.0414121	0.0747684	0.136227	0.0414121
XRT_1_208639	0.203512	0.0414173	0.0739201	0.136075	0.0414173
GBM_3_208639	0.203537	0.0414274	0.0747789	0.136252	0.0414274
GBM_2_209639	0.203557	0.0414356	0.0748350	0.136263	0.0414356
GBM_S_208639	0.203574	0.0414425	0.0748291	0.136279	0.0414425
GBM_grid_mode1_2	0.203609	0.0414568	0.0748523	0.136303	0.0414568
GBM_grid_model_1	0.203711	0.0414982	0.0749618	0.136297	0.0414982
GLM_1_208639	0.203733	0.0415872	0.0749323	0.136410	0.0415072
DeepLearning_1_200639	0.203748	0.0415133	0.0741946	0.136419	0.0415133

The DRF achieved the lowest RMSE of 0.203278 and the lowest MSE of 0.0413221, showing potential in predicting the reparation frequency, suggesting an effectiveness in minimizing prediction errors. XRT demonstrated the lowest MAE of 0.0739201, making

it a strong candidate for scenarios where minimizing the MAE is important, which is the case for the thesis. GBMs, just like the severity data in Section 4.5.2, showed competitive performance, with GBMs being particularly notable for its stable performances. This consistency across GBM models, which could also be seen in the severity data, provides reasoning for also using it for the frequency data. The DL algorithm exhibited strong performances again, particularly in terms of MAE (0.0741946), highlighting its ability to capture complex patterns in the data that other models might miss.

The analysis shows that the DRF model stands out as the top performer in terms of RMSE and MSE, making it a strong candidate for prediction of reparation frequency. However, the XRT model excels in minimizing absolute errors, which may be advantageous depending on the specific needs of the prediction task. The consistency in performance across various GBM models further suggests that gradient boosting is a reliable method for this type of dataset. The DL model also shows promise, particularly in minimizing MAE.

4.6 Modeling

This section introduces the various modeling techniques and processes used in this thesis, focusing on both traditional and modern ML approaches. The goal is to compare the performance of baseline GLMs with ML models, highlighting the strengths and weaknesses of each method. The tuning process of the ML techniques used is also described, making sure an optimal model is found. The model results are validated through various metrics and methods, which are detailed in this section.

4.6.1 Baseline GLM & ML Models

`The insurance company` provided valuable insights into their GLM used in practice. Additionally, non-life insurance consultants at `the consultancy firm` offered further insights into the variables included in a practical GLM. By using insights from the modeling process at `the insurance company`, it is possible to reproduce their GLM model and achieve performances closely matching theirs. The process and results of this baseline GLM are validated with `the insurance company`. This provides a solid baseline for this thesis to compare results obtained through ML methods with the same data, which is provided by `the insurance company`. Combining this information results in two baseline GLMs, similar to those used in practice and known to perform well. One GLM is designed for claim frequency, while the other is for claim severity. These baseline models are based on practical experience and refined through expert insights from both `the consultancy firm` and `the insurance company`, ensuring their relevance and applicability.

The three additional ML methods selected for the analysis are LightGBM, XGBoost, and MLP, as discussed in Section 3.2. These models were chosen due to their strong performance in the model exploration phase of Section 4.5 in various predictive modeling tasks via AutoML, and these models have a significant presence in literature. However, these advanced models inherently function as "black boxes," making it challenging to directly interpret their predictions. This lack of transparency necessitates the use of post-hoc analysis techniques to make the models' decision-making processes more understandable and interpretable. To address these challenges, XAI methods such as SHAP and LIME (both post-hoc analysis techniques) could be used, explained in Section 3.3.

4.6.2 Hyper-Parameter Tuning

To develop an optimal ML model, a variety of options must be explored. The process of designing the best model architecture with the most effective hyper-parameter configuration is known as hyper-parameter tuning. Hyper-parameters are the parameters set prior to the training process and can significantly impact the model's performance. The tuning process involves selecting the optimal set of hyper-parameters that yield the best performance on a validation dataset [118].

When tuning hyper-parameters, popular methods are Grid Search (GS) and Random Search (RS). GS is a decision-theoretic approach that exhaustively searches the optimal configuration in a fixed domain of hyper-parameters [118]. It creates a grid of all possible combinations of hyper-parameters and evaluates each one. While it performs an extensive search, this approach can be computationally expensive, especially with a large number of hyper-parameters. RS is another decision-theoretic method that randomly selects hyper-parameter combinations in the search space, given limited execution time and resources [118]. RS is often more efficient, particularly when dealing with high-dimensional spaces where certain hyper-parameters have more impact on performance than others.

A different approach could be Bayesian Optimization (BO). BO determines the next hyper-parameter value based on the previous results of tested hyper-parameter values, which avoids many unnecessary evaluations. This can also be a limitation because, in BO, the cost function is assumed to vary ideally within a predicted, preferred, and restricted bound around the observed value. However, the true solution might realistically lie far outside this limited region of search and inspection. As a result, the optimal solution found using Bayesian Optimization could be quite unsatisfactory and have a significant error [119]. The three hyper-parameter tuning methods are shown in Figure 4.11.

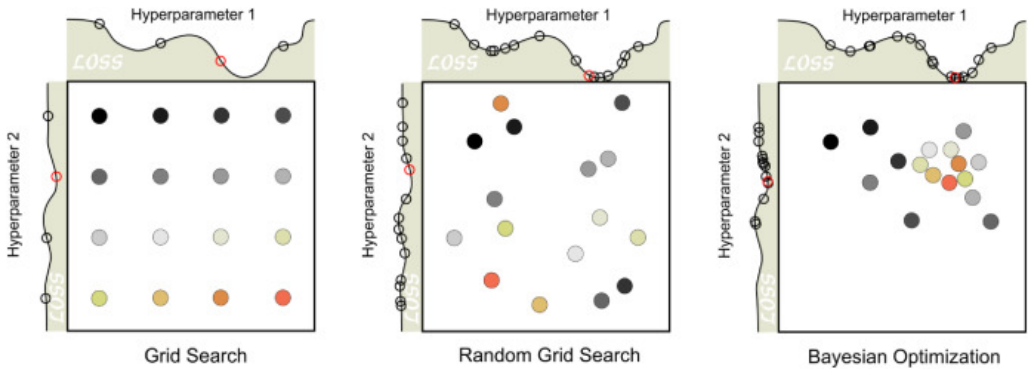


Figure 4.11: The different hyper-parameter tuning methods visualized (GS, RS, and BO) [118].

For this thesis, RS was chosen for hyper-parameter tuning due to its computational efficiency. RS allowed for a broad exploration of possible combinations within a reasonable time frame. This approach facilitated solid improvements on the initial (not tuned) developed ML model.

4.6.3 Model Validation

Validation is an important step in the development of ML models as it ensures that the models perform well on unseen data and can generalize beyond the training dataset. This section covers the validation process for the developed ML models, focusing on the chosen validation metrics and validation plots.

4.6.3.1 Metrics

For ML model outcome validation, five primary metrics have been chosen out of the possibile metrics identified (see Section 3.4.1): *MAE*, *RMSE*, R^2, *EV*, and *Tuning Time*. These metrics are chosen based on their relevance to the business understanding described in Section 4.2, which emphasizes the importance of accurately predicting the majority of cases rather than focusing on the few outliers.

MAE is chosen to be the most critical metric for the model validation process. One of the main reasons MAE is chosen as the key metric is its robustness to extreme outliers. Unlike other metrics, MAE is less sensitive to outliers, which ensures that the evaluation remains reliable even when the data contains extreme values. This robustness is essential in this business context, where understanding the average error magnitude directly translates to assessing the financial impact. Consequently, MAE serves as the primary metric for evaluating the developed model's performance. RMSE is the second most important metric in the validation process. RMSE can be more informative when comparing models, as it highlights differences in performance that may not be as apparent with MAE. However, its sensitivity to outliers means it is less robust than MAE, which is why it is considered secondary in the evaluation process. R^2 and Explained Variance (EV) are also included as important metrics in the validation process to look how well the model's predictions capture the variance in the actual data. Tuning time is considered as the last metric for validation. A model that requires extensive tuning time may delay deployment and increase costs, whereas a model with a shorter tuning time can be implemented more quickly, providing faster insights and value. Therefore, tuning time is a practical consideration that balances model accuracy with operational efficiency.

4.6.3.2 Actual Predicted Performance (APP) Plots

To evaluate the performance of the created ML models, APP[10] plots were generated comparing the predicted (mean) values against the actual (mean) values for each segment. These plots are a standard validation method at `the insurance company`. and provide a visual representation of how well the model's predictions align with the true values across different features and their respective segments [120]. In this case, the segments could represent the predicted mean values for car brands such as "BMW", "AUDI", "MERCEDES-BENZ", "PORSCHE", and "TESLA", as discussed and shown in Figure 4.6 of Section 4.3.3.1.

By plotting the mean values on the y-axis and the various segments on the x-axis, it is possible to observe how closely the predictions match the actual data points. Ideally, the points should align, indicating that the predicted values are very close to the actual values. Each segment is represented by a set of data points, allowing for the identification of any patterns or discrepancies within specific groups. Key observations from the plots include alignment with the line created for the mean of the actual values, systematic deviations, and segment-specific trends. Points that lie close to the actual data mean suggest accurate predictions, while systematic deviations reveal biases or errors. The spread of points within each segment provide insights into segment-specific model performance.

[10]APP stands for Actual Predicted Performance, which refers to the plots used to validate the mean values of actual data points against the mean values of predicted data points.

4.7 Evaluation

In this section, the XAI techniques used are presented. This evaluation aims to look into the findings gathered from the ML models and assess the applicability and effectiveness (for explainability and transparency) of the proposed ML techniques in a real-world context through interviews. The section also delves into the process overview of the interviews with stakeholders in the insurance industry.

4.7.1 XAI Techniques

In Section 3.3, various techniques for XAI were outlined, showing the importance of transparency and explainability in ML models. Among the various available methods mentioned, SHAP and LIME were selected for providing both global and local explanations of model behavior. It excels in providing both global explainability plots, which depict overall feature importance, and local explanations, which clarify individual predictions. This dual capability makes SHAP an interesting tool for understanding ML models, as discussed in Section 3.3.1. On the other hand, LIME provides a different approach by locally approximating the model around the prediction of interest with a simpler, interpretable model. This method could be useful for explaining the reasons behind specific predictions, making it easier to debug and trust the model on a case-by-case basis.

For the global explanations, the beeswarm and bar plots from SHAP were chosen, while for the local explanations, the waterfall plot (SHAP), force plot (SHAP), and LIME predictions were used.

4.7.2 Interviews

As discussed in Section 3.4.3, a semi-structured interview methodology is used to validate the gathered outcomes. In this interview, the aim is to explore the potential integration of ML algorithms into the insurance premium pricing process, replacing the currently used GLMs. The focus is to validate the various departments' perception (see Section 4.2) of the ML model results and to understand their views on the potential performance improvements and level of explainability offered by these models.

4.7.2.1 Approach

For the interview approach, the paper by Jentzch & Hochgeschwender was used as inspiration for dividing the interview in various question blocks [104]. Using this methodology, an semi-structured interview guideline was established to help the interviewer maintain an overview of essential topics and to later align the experts' insights with the validation of results. This approach also ensures both comparability and a methodological approach. The interview was divided into various main topics, inspired by the qualitative research study about ML by Jentzsch & Hochgeschwender [104]. In addition, various additional topics were included by the author to cover all relevant subjects. A more detailed description of the individual interviews conducted and the various questions asked are provided in Appendix E.2.

1. Technical Background.
2. Model Validation.
3. Importance of Model Explainability.
4. XAI Outcome Analysis.
5. Comparison GLMs and ML Explainability.
6. Views on AI Adoption.
7. Feedback and Suggestions.

4.7.2.2 Sample Size

Determining the appropriate sample size for validating ML explainability problems through interviews is a critical aspect of ensuring reliable and usable results. Also, data saturation was considered during the research process. Bryant & Charmaz define theoretical saturation as the point when all important issues or insights are exhausted from the data, making the emerging theory comprehensive and well-grounded [121]. This occurs when additional data collection reveals no new properties or theoretical insights. This is directly related to the sample size, ensuring that sufficient interviews were conducted to capture comprehensive insights.

Traditionally, it has been suggested by Nielsen et al. that observing five users can identify approximately 80% of usability problems, which could be related to explainabilty validation for this thesis [122]. The concept is built on the notion that the initial few users will uncover the majority of unique issues, while subsequent users tend to encounter repetitive problems [123]. However, this number has often been a discussion point, especially for studies that are more complex or diverse in nature. Small sample sizes may miss critical issues due to the limited variability and diversity in user feedback [122]. A paper by Faulkner indicates that increasing the sample size to 10 users can identify around 95% of usability problems [124]. Based on these guidelines provided by Nielsen et al. and Faulkner, validating the explainability outcomes through interviews should involve a sample size of 5 to 10 users. This sample size would yield an initial accuracy between 85.55% and 94.69% [124]. For greater statistical relevance, the sample size should be increased.

4.7.2.3 Process Overview

A final overview of the various interviews conducted is provided in Table 4.8. The interview process was planned in such a way to ensure thoroughness and reliability. This included obtaining consent for recording interviews, providing transcripts post-interview for verification, and maintaining a structured yet flexible interview environment to gather comprehensive insights.

Table 4.8: An overview of the various interviews conducted for this thesis, including essential background information about each interviewee and the context of the interviews.

	Consultancy Firm			**Insurance Company**	
Interviewee	*Individual A*	*Individual B*	*Individual C*	*Individual D*	*Individual E*
Department	Data Analytics	Non-Life Insurance	Non-Life Insurance	Data Management	Innovation & Product Management
# Interviews	1	1	1	2	2
Line of Defence (LoD)	1st	2nd	2nd	1st	1st & 3rd
Reference(s)	Appendix E.2.1 [125]	Appendix E.2.2 [126]	Appendix E.2.3 [127]	Appendix E.1 [128] & E.2.4 [120]	Appendix E.1 [128] & E.2.5 [129]
Location	Microsoft Teams	Microsoft Teams	Microsoft Teams	Physical; Physical	Physical; Microsoft Teams
Date	24th of July, 2024	26th of July, 2024	31st of July, 2024	24th of June, 2024[10] & 15th of July, 2024	24th of June, 2024[10] & 12th of August, 2024
Duration	40m 46s	1h 4m 14s	1h 11m 59s	1h 02m 17s & 59m 47s	1h 02m 17s & 51m 13s
Language	Dutch	Dutch	Dutch	Dutch	Dutch
Audio/Video	Yes/Yes	Yes/Yes	Yes/Yes	Yes/No; Yes/No	Yes/No; Yes/Yes

Table 4.8 provides various other notable insights. There are diverse expertise backgrounds within the interview group. The interviewees come from a variety of departments, including Data Analytics, Non-Life Insurance, Data Management, and Innovation & Product Management. This diversity ensures a comprehensive understanding of the different perspectives and requirements of model explainability across an organization. The duration of the interviews varied, with some lasting over an hour. Dutch was used consistently as the primary language in these discussions. The order of interviews was determined based on the experts' availability. For example, `Individual D` and `Individual E` were interviewed first in a combined interview because their availability was limited between June 2024 and August 2024 due to vacation. This approach was applied to all other interviewees as well. Some interviewees participated in multiple interviews. This indicates a deeper exploration of opinions of the stakeholders in the adoption and validation of premium pricing models.

[10]This is a combined interview, discussing the same type of questions with `Individual D` representing *LoD-1* and `Individual E` representing both *LoD-1* and *LoD-3*.

Effective time scheduling and management were crucial in ensuring that the required number of interviews was conducted with all departments. A more detailed description of the individual interviews conducted is provided in Appendix E.2.

4.8 Summary

In this chapter, the experimental setup for the thesis was outlined, covering the approach, data handling, and modeling techniques used. The methodology follows a structured framework based on the CRISP-DM process. The experimental framework was detailed, including the business context, data understanding, preparation, modeling, and evaluation approaches. Two datasets were explored, focusing on claim severity and frequency from the reparation dataset of `the insurance company`. Data confidentiality was noted and will be taken into account for the thesis. The modeling phase compares traditional GLMs with modern ML techniques through H2O AutoML, where performances were compared using metrics such as MAE and RMSE. The results showed that GBMs and DL models performed particularly well, highlighting their potential over GLMs for `the insurance company`'s case. A baseline GLM was developed to provide a comparison point for evaluating the potential performance improvements offered by ML models. Hyper-parameter tuning and validation processes, such as APP plots, were also discussed in the process, emphasizing the importance of robust model evaluation. The interview approach was introduced as well and is used to validate model explainability and effectiveness, providing insights from various stakeholders in the insurance industry, which could be identified through the 3LoD model discussed in the business understanding phase.

Chapter 5

Results & Discussion

5.1 Model

This section presents a detailed discussion on the performance and validation of the GLM alongside the selected ML models. As outlined in Section 3.2, the chosen ML models for this study include the GLM, applied to both severity (using the Gamma distribution) and frequency (using the Poisson distribution), as well as LightGBM, XGBoost, and a MLP (type of NN). The results for each model are presented, a comparative analysis follows, highlighting the strengths and limitations of each approach based on various performance metrics. The validation process is described in detail as well, emphasizing how each model was evaluated to ensure reliability and robustness in predictions. The severity model will focus on the following features, where the features are vaguely described to provide some more understanding towards the outcomes:

- ◇ **Target**. Claim burden.
- ◇ `Feature 163`. Policy details.
- ◇ `Feature 171`. Vehicle information.
- ◇ `Feature 6`. Claim details.
- ◇ `Feature 88`. Vehicle information.
- ◇ `Feature 127`. Vehicle information.

The frequency model will focus on the following features, where the features are vaguely described to provide some more understanding towards the outcomes:

- ◇ **Target**. Number of claims {Type}[1].
- ◇ `Feature 145`. Policy details.
- ◇ `Feature 76`. Vehicle information.
- ◇ `Feature 167`. Vehicle information.
- ◇ `Feature 110`. Vehicle information.
- ◇ `Feature 169`. Vehicle information.
- ◇ `Feature 18`. Policy details.

5.1.1 Performance

The model performance of both the severity and frequency datasets is critical to understanding the effectiveness of the predictive models used. In this section, a comprehensive overview of the performance metrics and outcomes is provided for both the severity

[1]For this section, the type is "Reparation", as the results section focuses on the reparation data retrieved from `the insurance company`.

and frequency datasets, specifically focusing on the reparation dataset of `the insurance company`.

5.1.1.1 Severity

Table 5.1 presents a comparative analysis of several ML models such as the GLM (Gamma), LightGBM, XGBoost, and the MLP for the severity data. The models are evaluated based on their performance on the reparation severity dataset from `the insurance company`. The metrics used to compare the models include MAE, RMSE, and Tuning Time (in seconds). Additionally, R^2 and EV are mentioned. The best values in each metric are highlighted in green to emphasize their significance.

Table 5.1: An overview of the results of the various ML techniques for the reparation severity dataset.

Model Type (Reparation Severity)	MAE	RMSE	R^2	EV	Tuning Time (s)
GLM (Gamma)	2603.22	6243.74	0.0374	0.0376	2.47
LightGBM	2596.05	6190.87	0.1025	0.1026	8.07
XGBoost	2597.86	6205.70	0.0982	0.0983	0.10
MLP	2580.19	6121.43	0.1225	0.1226	697.32

General. It is important to note that the R^2 and EV values across all models are relatively low, especially considering that these metrics can theoretically reach a maximum of 1. This could be due to the complexity and variability in the severity of reparations, which may be influenced by factors not captured in the model, which makes it challenging for models to predict every case perfectly. This inherent unpredictability in the data likely contributes to the lower R^2 and EV values observed in this analysis. The results also show a clear trade-off between prediction power and computational efficiency. While the MLP model provides slightly better accuracy, the significant increase in tuning time may not be justified in all situations, especially if quick model deployment is necessary. On the other hand, models like XGBoost and LightGBM offer a good balance between accuracy and efficiency, making them strong candidates when both factors are considered.

MLP. The MLP achieves the lowest MAE and RMSE, with values of 2580.19 and 6121.43 respectively. It also has a significant difference compared to the GLM, showing the potential of ML. The LightGBM and XGBoost models also outperform the GLM. The MLP model, with the highest R^2 value (0.1225), explains the most variance among the models compared, suggesting it has the best fit for this dataset. Higher EV values indicate a better model, as it suggests that the model is capturing more of the variance in the data. The EV values closely mirror the R^2 values, with the MLP model again performing the best (0.1226). While the MLP model is more accurate, it requires substantially more computational time (697.32s), which could be a limiting factor in some applications. The differences in tuning time highlight the potential trade-off between model complexity and computational efficiency that should be considered in a future stage.

GLM. The GLM based on the Gamma distribution for severity data does not excel in any of the evaluated metrics. Its performance is inferior to that of the other models. Additionally, with its respective tuning time, it is neither the fastest (2.47s) nor the one with the lowest absolute error (2603.22). The model's performance indicates that the

current industry standard models like GLMs may struggle to compete with more complex ML algorithms when it comes to predictive power, which is shown by the GLM having both the lowest MAE, RMSE, R^2, and EV values.

GBMs[2]. LightGBM presents a balanced performance with the MAE (2596.05) and RMSE (6190.87) metrics, both of which are closest to the best-performing MLP model. Although it does not lead in any single metric, it provides a good balance between accuracy and tuning time. Considering this, LightGBM could be an interesting option to explore. Additionally, since LightGBM is tree-based, it allows the use of TreeSHAP, which is significantly more computationally efficient than KernelSHAP, as discussed in Section 3.3.1. While the MLP provides the lowest error, XGBoost stands out for its efficiency in terms of tuning time (0.10s), but relatively does not differ much to the GLM (2.47s) and LightGBM (8.07s) models. The minor difference in tuning time would not make the difference in choosing XGBoost over LightGBM. However, when severity data sizes increase, tuning time could become a more important factor for determining the best performing model.

Discussion. The models show close performance in terms of prediction error metrics, with the MLP model slightly ahead. As already established, there is a preference in explaining the gross of the data correctly. This would make MLP the best model due to its lowest MAE (and RMSE) value. A potential reason for the performance of MLPs is that these algorithms are capable of learning interactions between features that may not be explicitly captured by other models. MLPs utilize non-linear activation functions, such as ReLU or sigmoid, in their hidden layers, enabling them to capture these complex, non-linear relationships. Given that the severity data is likely influenced by interactions between features (e.g., combining different aspects of the claim or repair process), MLP's hidden layers can learn and model these interactions, potentially improving its performance compared to models that may not capture these interactions as effectively. However, the choice of the model should take into account the balance between accuracy and tuning time, particularly when dealing with large datasets or when rapid model deployment is required. When the dataset scales or the company needs to retrain models more frequently (e.g., more than once a month), faster-tuning models like XGBoost or LightGBM may become more attractive because of their lower computational overhead. Unlike MLPs, which use NNs with multiple layers of interconnected nodes, XGBoost and LightGBM use DTs. DTs are generally faster to train because they involve simple, recursive splitting of data based on feature values. This simplicity makes tree-based models less computationally demanding compared to the matrix operations and backpropagation required by MLP. The significant computational cost (697.32s tuning time) of MLP may not always be justifiable.

5.1.1.2 Frequency

Table 5.2 presents a comparative analysis of several ML models such as the GLM (Gamma), LightGBM, XGBoost, and the MLP for the frequency data of `the insurance company`.

Table 5.2: An overview of the results of the various ML techniques for the reparation frequency dataset, indicating the best values with a green color.

Model Type (Reparation Frequency)	MAE	RMSE	R^2	EV	Tuning Time (s)
GLM (Poisson)	0.074679	0.203362	0.0064	0.0064	43.69

[2] Assigned combined naming for the LightGBM and XGBoost methods, which are both based on GBM principles.

LightGBM	0.074533	0.203172	0.0079	0.0079	429.25
XGBoost	0.074533	0.203172	0.0082	0.0082	632.76
MLP	0.074237	0.203444	0.0056	0.0056	28836.50

General. Some remarks that one could identify from Table 5.2 is that the models show close performance in terms of MAE and RMSE, with differences being minimal. This indicates that all the models are generally performing similarly well on this dataset, though slight differences are noted. Also, the R^2 and EV values are generally low across all models. This suggest that the models are not explaining much of the variability in the claim frequency data. This could be due to the stochastic nature of claim frequency, where the occurrence of claims may be influenced by random events (natural disasters, economic conditions, random individual behavior) or factors not captured in the dataset. These outliers are difficult to model accurately, hence the reason why the mean of the predicted values is checked in Section 5.1.2.2 rather than relying to much on the R^2 and EV values.

MLP. The MLP has the lowest MAE (0.074237), indicating it has the most accurate predictions for the gross of predictions. This model minimizes the mean error, making it the best performer in this area across the models tested. However, the RMSE for the MLP is slightly higher than that of LightGBM and XGBoost, suggesting that while it performs well on average, it may be more susceptible to larger errors compared to the tree-based ML models built. The lower R^2 (0.0056) and EV (0.0056) suggest that the explained variance with the MLP is worse than the LightGBM and XGBoost models. Also, the MLP requires over 28,800 seconds (approximately 7.5h) for tuning, which could give issues in the future as this could be deemed extremely long when compared to the other models.

GLM. The GLM (Poisson) model outperforms the other models in terms of tuning time, taking only 43.69 seconds. This is substantially faster than the more complex models, such as the MLP. However, it has to be noted that the other metrics, such as MAE, RMSE, R^2, and EV, do not provide good results for the GLM compared to the ML algorithms.

GBMs. Both LightGBM and XGBoost strike a balance between the best predictive performance and computational efficiency, making them an interesting option for insurance pricing due to improved results, but taking less computational power (also for SHAP in the explanation stage, as discussed in Section 3.3.1). These models provide excellent performance in MAE and RMSE with reasonable tuning times, making them versatile options for applications requiring the best of both worlds.

Discussion. When considering a balance between prediction and computational efficiency, models like LightGBM and XGBoost could be preferable, as they provide a good compromise between error metrics and tuning time. LightGBM and XGBoost outperform simpler models like GLM due to their gradient boosting architecture approach, regularization techniques, and scalability, allowing them to manage large, complex, and imbalanced datasets effectively. However, the model choice could also depend on the specific requirements of the task, such as the need for faster processing or higher prediction power. This could differ per insurance company.

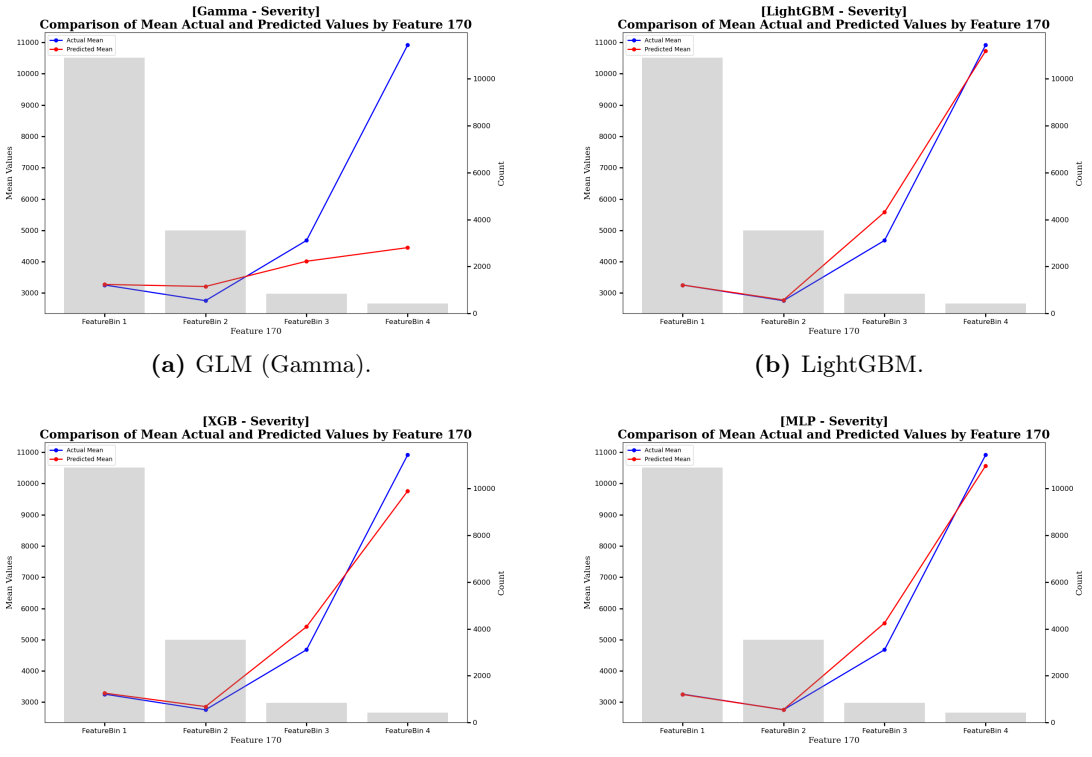
5.1.2 Validation

Validation of the ML models will be performed using APP plots, discussed in Section 4.6.3.2. These plots are important for assessing the model's accuracy by comparing the predicted values against the actual values across different segments of the data and is one of the main validation methods for GLMs at `the insurance company`. The APP plots allow us to visually inspect whether the model's predictions align with the observed data trends.

In practice, the APP plot is constructed by segmenting the data into groups, based on certain criteria (e.g., predefined categorical groups). For each segment, the mean of the actual values is plotted against the mean of the predicted values. Ideally, if the model is performing well, the predicted points on the APP plot should lie close to the actual data points, indicating that the predicted means are in line with the actual means. This approach provides a more nuanced validation method than simply relying on aggregate metrics such as MAE or RMSE. By examining the APP plots, one can identify whether the model consistently underestimates or overestimates in specific segments, which might not be apparent from overall error metrics.

5.1.2.1 Severity

An overview of a comparison for the models for the severity data is provided in Figure 5.1. The validation process centers around a specific anonymized feature, `Feature 170`, which is part of the final model features. This feature is divided into four distinct bins, ranging from the most frequent (left) to the least frequent (right), with the APP plot used to compare the actual versus predicted values across these bins. The method provides a clear visual representation of the model's performance for the particular feature, ensuring that the predicted values accurately reflect the trends observed in the actual data across all feature bins.



(a) GLM (Gamma). **(b)** LightGBM.

(c) XGBoost. **(d)** MLP.

Figure 5.1: APP plots for the four models for `Feature 170` for the reparation severity dataset at `the insurance company`, where the distribution of the feature bins is provided as well.

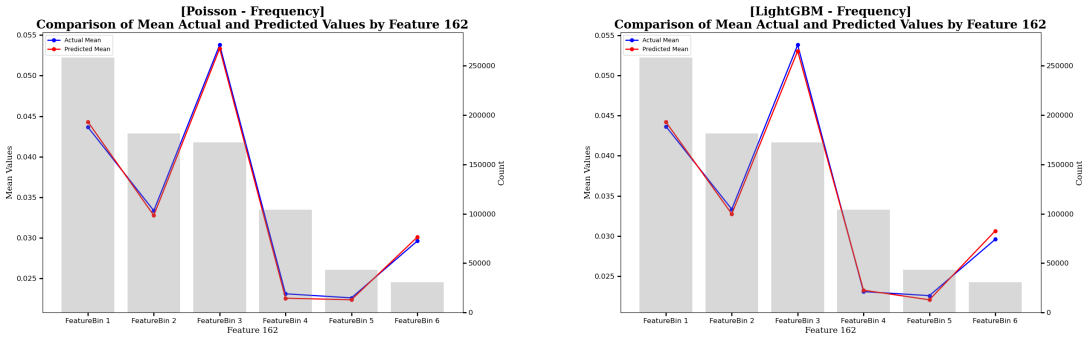
Figure 5.1a shows that the predicted means deviate significantly from the actual means, where this is the most prominent in the last bin (`FeatureBin 4`), which is the least frequent

bin. The GLM seems to underestimate the actual mean value here, which indicates that the GLM model is not capturing the trend accurately in this segment of the data when there is not much information available. As one could see as well, the actual mean value is the highest here as well, meaning that this is a bin feature that could have some high claim potential, which makes it harder to predict such a value when taking the variability into account. The other ML models, however, show a better alignment between the actual and predicted means, especially in the last two bins. Both LightGBM (Figure 5.1b) and XGBoost (Figure 5.1c) demonstrate good consistency, with predicted means closely aligning with actual means across all bins, especially in the challenging last `FeatureBin 4` where GLM struggles. The MLP model (Figure 5.1d) shows some good performances in predicting the mean values across different bins of `Feature 170`, showing the ability of capturing complex, non-linear relationships in the data.

Some interesting comparisons could be derived from Appendix D.1.1, which handles the APP plots for the other features from the final (severity) model. When analyzing `Feature 62` in Figure D.1, one can observe high variability in actual severity, with numerous peaks in certain bins (mostly happening at the side with less frequent occurences). These peaks suggest that specific bins see dramatic increases in severity, yet the models struggle to capture these spikes, as the predicted mean remains relatively flat compared to the actual mean. `Feature 88` displays a similar pattern of peaks, though less often than in `Feature 62`, and once again, the models find it challenging to predict these sudden increases in severity. In contrast, `Feature 6` presents a more consistent trend across bins, with noticeable fluctuations. Here, models like LightGBM and XGBoost closely follow the actual mean values, indicating better performance for this feature. `Feature 171`, with its simpler pattern and fewer bins, shows that all models predict the peak in the third bin relatively well, except the GLM (Gamma) model that has the highes differences. This is a trend that occurs throughout the severity validation process, and could potentially advocate for the usage of ML models.

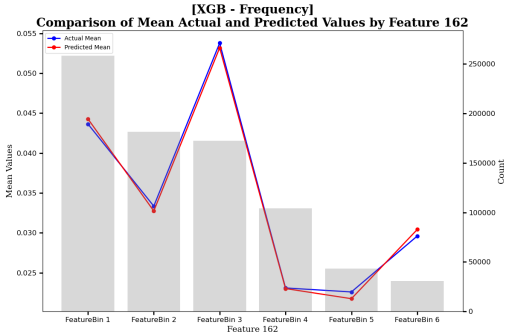
5.1.2.2 Frequency

An overview of a comparison for the models for the severity data is provided in Figure 5.2. The validation process centers around a specific anonymized feature, `Feature 162`, which is part of the final model features. Just like the severity validation, feature are divided into distinct bins, ranging from the most frequent (left) to the least frequent (right).

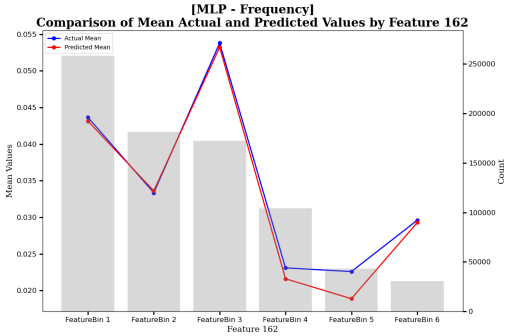


(a) GLM (Poisson).

(b) LightGBM.



(c) XGBoost.



(d) MLP.

Figure 5.2: APP plots for the four models for `Feature 162` for the reparation frequency dataset at `the insurance company`, where the distribution of the feature bins is provided as well.

When comparing the models, the MLP plot (Figure 5.2d) reveals some discrepancies between the actual and predicted mean values, most in `FeatureBin 4` and `FeatureBin 5`. In these bins, it fails to respond as accurately to the individual feature bin predictions compared to the GLM, LightGBM, and XGBoost model. In contrast, the plots for GLM, LightGBM, and XGBoost show a much closer alignment between the actual and predicted values in these specific bins, indicating better model performance. The LightGBM and XGBoost models, in particular, demonstrate a strong ability to capture the relationship between `Feature 162` and the target variable (*number of claims*), as evidenced by the near overlap of the actual and predicted mean lines. The MLP model also performs well in the other bins, though there are slight deviations in the bins as mentioned (`FeatureBin 4` and `FeatureBin 5`).

Some interesting comparisons could be derived from Appendix D.1.2, which handles the APP plots for the other features from the final (frequency) model. For `Feature 165`, `Feature 145`, `Feature 54`, `Feature 76`, `Feature 168`, and `Feature 172`, a similar pattern of these models (LightGBM, XGBoost, and MLP) performing better than GLM is observed, just like the `Feature 162` described in this section. The GLM and MLP show more irregular behavior across all features by somehow averaging some features (For both GLM and MLP: `Feature 145`, for GLM only: `Feature 76`), which might indicate that these respective features have a more complex relationship with the target variable that the GLM and MLP struggle to capture, which would advocate for the usage of ML algorithms such as LightGBM or XGBoost. Across the different features, the trend of LightGBM and XGBoost performing better than the GLM is consistent. This analysis suggests that while

certain models (XGBoost, LightGBM, and in some way MLP) generally outperform others (GLM) in predicting the actual values, some features might give challenges that require further investigation or potentially other modeling techniques.

5.2 Explanations

In this section, the explanations for the severity and frequency predictions within the reparation dataset are explored, focusing on both global and local explanations. The discussion will highlight the most intriguing insights drawn from the data and provide a detailed guide on interpreting SHAP's global/local explanations and LIME local explanation plots, which are essential tools for understanding the ML model's predictions. By utilizing SHAP and LIME, one can get both global (SHAP) and local (SHAP & LIME) insights into the model's behavior. For the model explanations, the GLM model will be investigated for global explanation, but not for local explanation. Reasoning is the GLM already uses factor explanations, which already provide accurate reasoning for the models. So, for local explanation, only the three ML models (LightGBM, XGBoost, and MLP) will be investigated. But, it is interesting to see how SHAP performs on the GLM, hence the incorporation in the global evaluation.

5.2.1 Severity

The discussion will delve into the global and local explanations provided by the best-performing model for severity, which is used to explain the various explanation plots generated based on its predictions. Additionally, some interesting insights obtained from the other models will be highlighted to offer a broader perspective.

Some features, such as `Feature 163` and `Feature 171`, have a feature name that appears multiple times in the shown explanations. This duplication occurs because the features have been converted with OHE, which is often applied to convert categorical variables into a form suitable for ML models. As a result, each different bin of the original feature becomes a separate feature after encoding, leading to the repeated appearance of the same feature name.

5.2.1.1 Global Explanations: SHAP Beeswarm & Bar Plots

This section handles the global (SHAP) plot explanations of the models described in Section 5.1.1.1. In this section, the focus is on the best-performing model in the analysis of the reparation severity dataset: the MLP. Both the beeswarm plot and the bar plot have been examined in Section 4.7.1 and briefly handled and discussed through an example dataset in Section B.2.1. Figure 5.3 provides the beeswarm plot for the MLP model, which shows the impact of the various features incorporated within the final model.

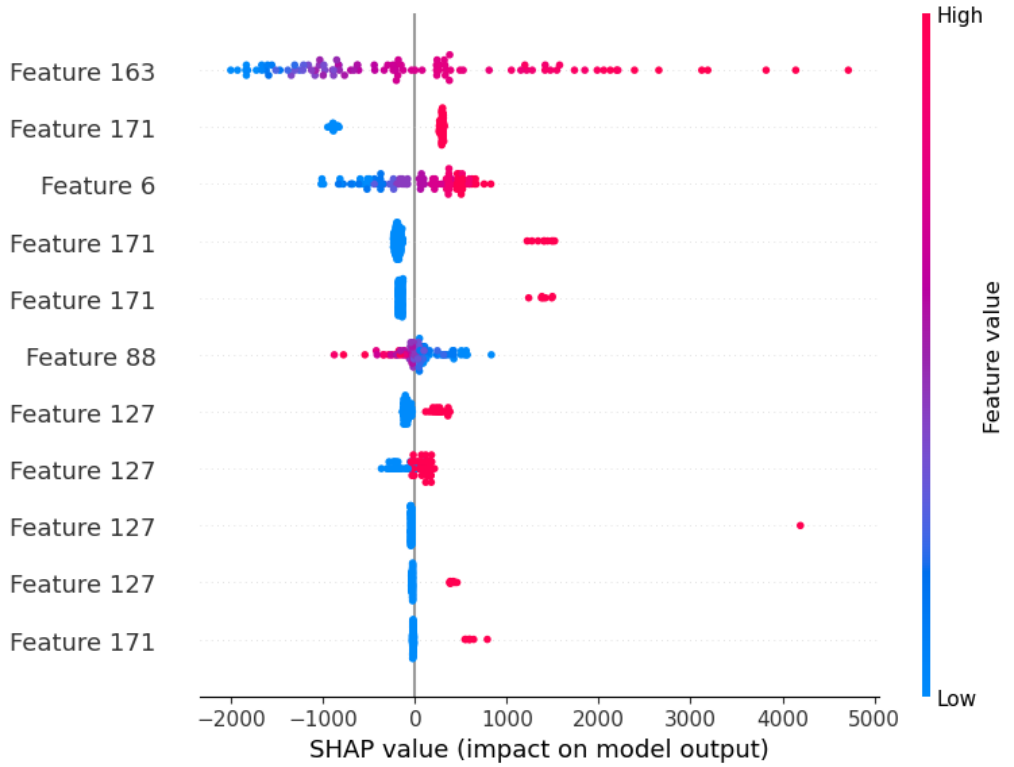


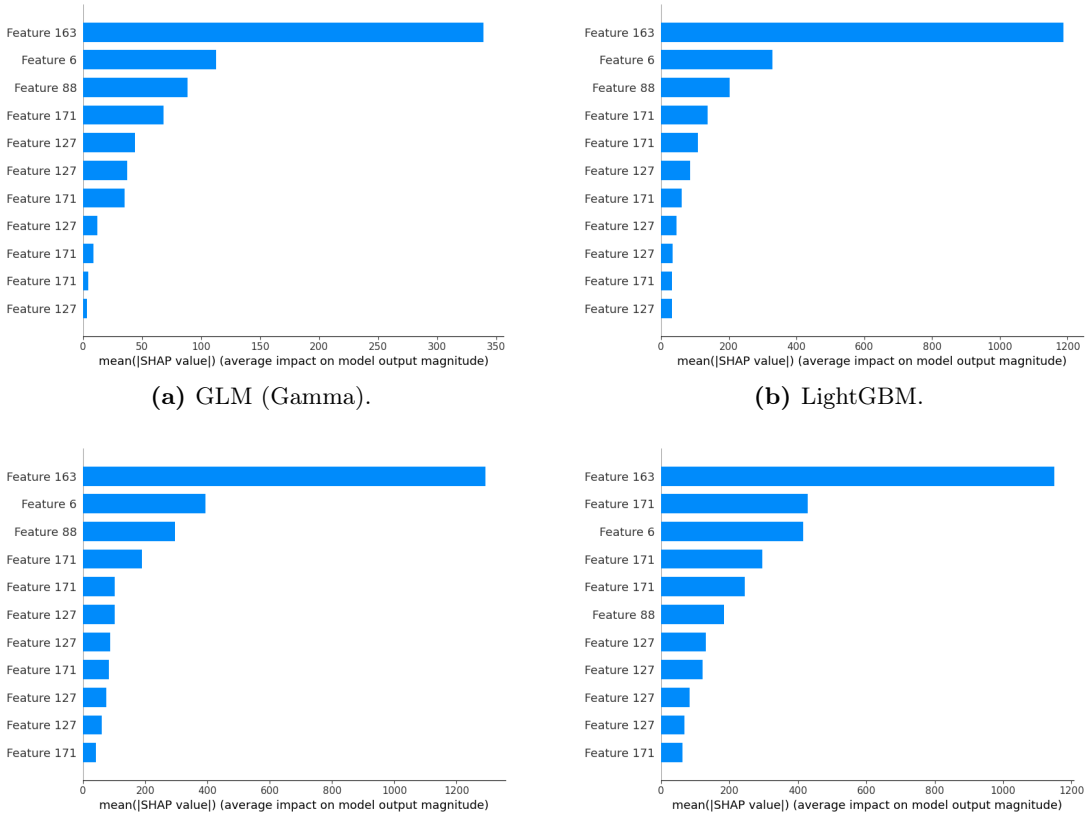
Figure 5.3: Beeswarm plot made for the (best-performing) MLP model based on the severity data of `the insurance company`, showing global feature contributions for the respective features within the model.

The x-axis represents the SHAP value, indicating the impact of each feature on the model's output. Positive SHAP values, which appear to the right of the centerline, show feature values that contribute to increasing the prediction (increasing the target *"claim burden"*), while negative SHAP values, on the left, indicate that the respective feature value contributes to decreasing the prediction (in this case, decreasing the target). Each feature can have either a positive or negative impact, depending on its specific value. The color gradient from blue to red represents the feature value itself, where blue signifies low values and red represents high values. An example for interpreting this output could be as follows: one could deduce from this model that when `Feature 163` decreases in its feature value (i.e., if the color gradient shifts from red to blue), the impact of this feature on the final model output will decrease. This specific instance (and the other features as well) has also been validated with `the insurance company` to strengthen the claims made in the discussion.

Also looking at the beeswarm plots created for the other models, which are described in Appendix D.2.1.1, one can see that the plots share several similarities, such as the prominence of `Feature 163` and `Feature 6`, which are describe in the top 3 features for all the different models. `Feature 171` and `Feature 88` are also found high up within the SHAP explanations. However, also some notable differences exist, including the scale of SHAP values, with XGBoost and LightGBM showing much larger ranges compared to the GLM (Figure D.11a) and MLP (Figure D.11d). An important note to make here is the fact that KernelSHAP had to be used for both the GLM and MLP, which was limited by the computing power of the device which was used for the calculations. Therefore, a subset had

to be taken for the SHAP outcomes for GLM and MLP models[3]. While using a subset of 100 samples for SHAP analysis can provide insights, the results might not fully represent the model's behavior on the full dataset, potentially leading to different interpretations compared to using the entire data, which should be considered when interpreting the results. This highlights a reason for choosing tree-based models, which makes it possible to accelerate the retrieval of explanations from the XAI techniques used.

Figure 5.4 shows bar charts for the four final models made, where the mean SHAP value measures the average impact of that feature on the model's predictions. The provided SHAP value-based feature importance plots highlight the global contributions of various features across the four different models.



(a) GLM (Gamma). **(b)** LightGBM.

(c) XGBoost. **(d)** MLP.

Figure 5.4: Bar plots made for the four models based on the severity data of `the insurance company`, showing global feature contributions based on SHAP values for the respective features within the model.

A consistent observation across all models is the prominence of `Feature 163`, which emerges as the most influential feature in each case. Higher mean SHAP values indicate that the feature has a more significant influence on the model's output. Its dominance is particularly noticeable in the LightGBM (Figure 5.4b) and XGBoost (Figure 5.4c) models and MLP model, where the SHAP values for `Feature 163` are significantly higher compared to other features, surpassing 1200, showing the great influence of this feature towards the final target output *"claim burden"*. Comparing this to the GLM model (Figure 5.4a),

[3]**Note:** This is the case for all the other SHAP and LIME models to come. A subset of 100 has been used for both GLM and MLP, giving these models the same subset distribution.

where the SHAP value is under 350, one can see a big difference in impact, which could explain why the GLM does not follow individual feature bins as well in the validation stage of Section 5.1.2.1. Furthermore, `Feature 6`, `Feature 88`, `Feature 127`, and `Feature 171` also consistently rank among the top contributors across all four models, indicating that they play a significant role in the prediction process. However, the importance of these features varies across models.

Looking at the individual chart for the best-performing severity model identified in Section 5.1.1.1, Figure 5.4d shows a bar chart representing the mean SHAP values for different features within the MLP model. `Feature 163`, `Feature 171`, and `Feature 6` are shown to be having the highest influence within the MLP model, as could be derived. `Feature 163` has the highest mean SHAP value, indicating it is the most important feature in the model. It has the greatest average impact on the model's output. `Feature 171`, appearing multiple times due to OHE, is second highest. This feature appears multiple times on the vertical axis, but the order is validated with `the insurance company`.

5.2.1.2 Local Explanations: SHAP Waterfall & Force Plots, LIME

A waterfall plot effectively breaks down the prediction process, showing how the individual contributions of different features lead to the final model output. This allows for a deeper understanding of the model's behavior, particularly in understanding how specific feature values drive predictions in either direction, which could be helpful to insurers when having to investigate a client and look at reasons for a specific pure premium.

The waterfall plot depicted in Figure 5.5 provides a local explanation for the predictions of the MLP model based on the severity data from `the insurance company`. In this waterfall plot, the base value, which is the average model output across all predictions for the target *"claim burden"* (in €), is denoted as $E[f(X)] = 3085.60$. The contributions of the various features towards the target *"claim burden"* are then added or subtracted from this base value to arrive at the final model prediction, a *"claim burden"* (in €) of $f(x) = 3181.65$. This means that this particular person has an expected claim burden higher than the average.

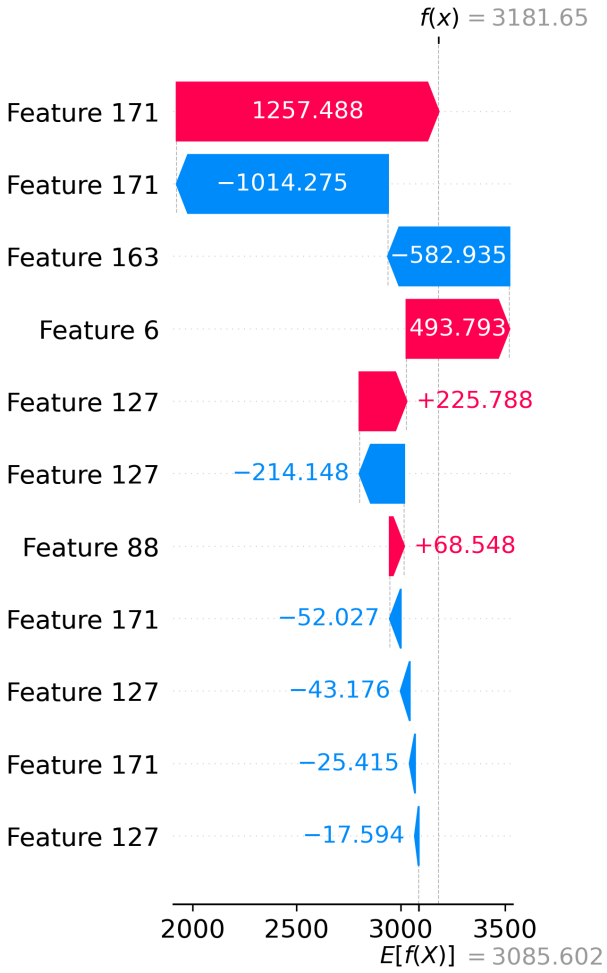


Figure 5.5: Waterfall plot made for the MLP model based on the severity data of `the insurance company`, showing local feature contributions based on SHAP values for the respective features within the model.

For this particular instance, `Feature 171` has the most significant positive impact, increasing the prediction by approximately €1257.49, as indicated by the red bar. On the other hand, another bin of `Feature 171` decreases the prediction by about €1014.28, shown by the blue bar. This highlights the varying impact of different categories within the same feature due to OHE. `Feature 163` also has a substantial negative impact, reducing the prediction by approximately €582.93. `Feature 6` and a specific bin of `Feature 127` contribute positively, with increases of €493.79 and €225.79, respectively. Other features such as `Feature 88` and additional bins of `Feature 127` and `Feature 171` have smaller but still notable contributions, either increasing or decreasing the final prediction.

Looking at the created waterfall plots for both LightGBM and XGBoost (Figure D.12a & Figure D.12b in Appendix D.2.1.2 respectively), one can identify some common trends and differences. The provided SHAP waterfall plots for the LightGBM, XGBoost, and MLP models highlight differences in how each model predicts outcomes based on feature contributions. LightGBM and XGBoost produce similar prediction outputs (€3248.33 and €3328.78, respectively), while the MLP model predicts a slightly lower value (€3181.65). `Feature 171` emerges as the most critical factor for the MLP, but LightGBM and XGBoost differ, with `Feature 163` contributing the highest. `Feature 163` is also consistently con-

tributing negatively (which makes sense as the same as this is the investigation of the same instance) and `Feature 6` , `Feature 88` and `Feature 127` (the same bin for this respective feature) adding positive value across all models as well. This demonstrates that all three models can recognize the same patterns within the given data and draw similar conclusions. The main differences among the models are in the degree of feature contribution (measured in €) attributed to specific feature values.

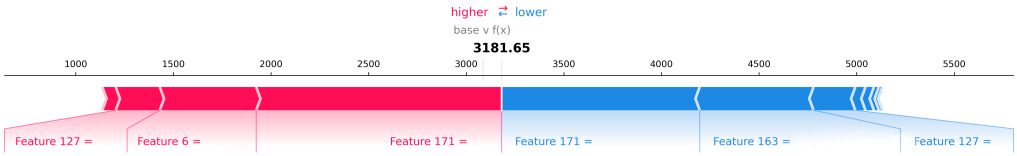


Figure 5.6: Force plot made for the MLP model based on the severity data of `the insurance company`, showing local feature contributions based on SHAP values for the respective features within the model.

Another way of visualizing the impact of features on an individual prediction is using SHAP force plots. The force plot depicted in Figure 5.6 provides a visual representation of how different features push the prediction either higher or lower relative to the base value for a particular instance in the dataset. This plot is particularly useful for understanding the local explanation of a model's prediction at the individual instance level.

Looking at the force plot in Figure 5.6, the base value is denoted, just like in the waterfall plot, as $f(x) = 3181.65$. The length and direction of the arrows in the force plot represent the magnitude and impact of each feature on the prediction, with longer arrows indicating a greater impact. The final prediction is the result of summing the base value with all the individual feature contributions. This visualization effectively illustrates how each feature pushes and pulls the model's prediction for a single instance, providing a clear understanding of which features are most influential in determining the outcome. Appendix D.2.1.3 show the force plots for both the LightGBM and XGBoost model, which have the same contributions as explained for the waterfall plot. Hence, no further discussion is needed.

Next to the SHAP explanations, LIME plot explanations have been made to explain local instances. An example of a LIME explanation based on the severity data has been provided in Figure 5.7. The final plots are visualized in Figure D.14 of Appendix D.2.1.4 for the ML models, where the values on the horizontal axis represent the contribution or weight of each feature to the prediction for a specific instance. The minimal predicted value was negative, namely a value of €-794.93. The maximum value approximated was €14,189.12. The values on the horizontal bars indicate how much each feature, with a particular value, pushes the prediction in a positive or negative direction. For all three ML models, `Feature 127` (same bin) and `Feature 163` show significant contributions. This was also recognized as the most impactful feature according to the global model evaluation with SHAP bar plots conducted in Section 5.2.1.1. However, the impact of the different features within the LIME explanation varies across models in magnitude. `Feature 6` and `Feature 88` also show explanations in the same direction, showing some form of robustness for the perturbation-based LIME model.

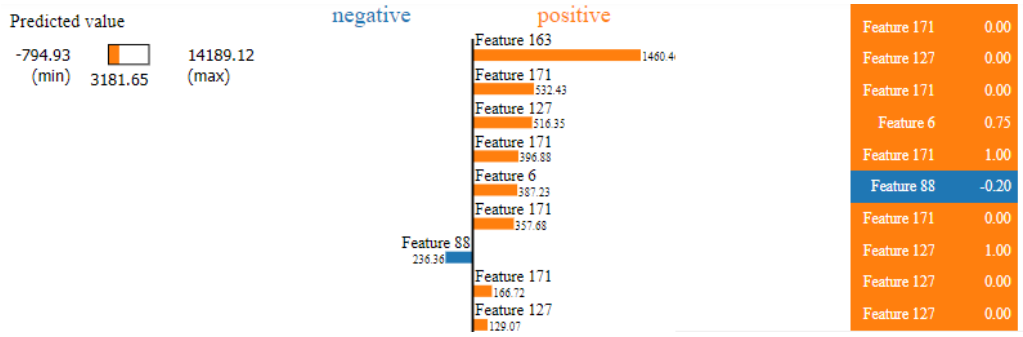


Figure 5.7: Example of a LIME explanation for the MLP model based on the severity data of `the insurance company`, showing local feature contributions for the respective features within the model.

The LIME outputs, while useful for understanding local model behavior, have a key limitation in that they are not particularly robust during modeling. The results produced by LIME are not deterministic, meaning that the explanations generated for the same instance can vary with each run. This variability arises because LIME uses a perturbation-based approach to approximate the local decision boundary of the model, which introduces randomness into the explanation process. Consequently, different perturbations or different samples used in the process may lead to different results, making the explanations less reliable compared to more stable methods like SHAP. The retrieved results in Figure D.14 of Appendix D.2.1.4 could therefore, according to the author, not be used in the financial sector, where robust explanations are key. Therefore, while LIME can provide insights into how features influence individual predictions, its outputs (when used) should be interpreted with caution, especially in scenarios where consistency and robustness of the explanation are critical such as the insurance sector. It is therefore moved to Appendix D.2.1.4.

5.2.2 Frequency

The discussion will delve into the global and local explanations provided by the best-performing model for frequency, which is used to explain the various explanation plots generated based on its predictions. Additionally, some interesting insights obtained from the other models will be highlighted to offer a broader perspective. Some features, such as `Feature 110` or `Feature 167`, have a feature name that appears multiple times in the shown explanations. This duplication occurs because the features have been converted with OHE.

5.2.2.1 Global Explanations: SHAP Beeswarm & Bar Plots

This section handles the global (SHAP) plot explanations of the models described in Section 5.1.1.2. In this section, the focus is on the best-performing model in the analysis of the reparation frequency dataset: XGBoost. Both the beeswarm plot and the bar plot have been examined in Section 4.7.1 and briefly handled and discussed through an example dataset in Section B.2.1. Figure 5.8 provides the beeswarm plot for the XGBoost model, which shows the impact of the various features incorporated within the final model. An example for interpreting the output provided could be as follows: one could deduce from this model that when `Feature 145` decreases in its feature value (i.e., if the color gradient shifts from red to blue), the impact of this feature on the final model output will increase (this means the target *"number of claims"* will increase). An alternative example is `Feature`

76, where an increase in the feature value (represented by a gradient color transitioning from blue to red) leads to a higher SHAP value. This, in turn, results in a greater impact on the model's output, ultimately increasing the target value of *"number of claims"*.

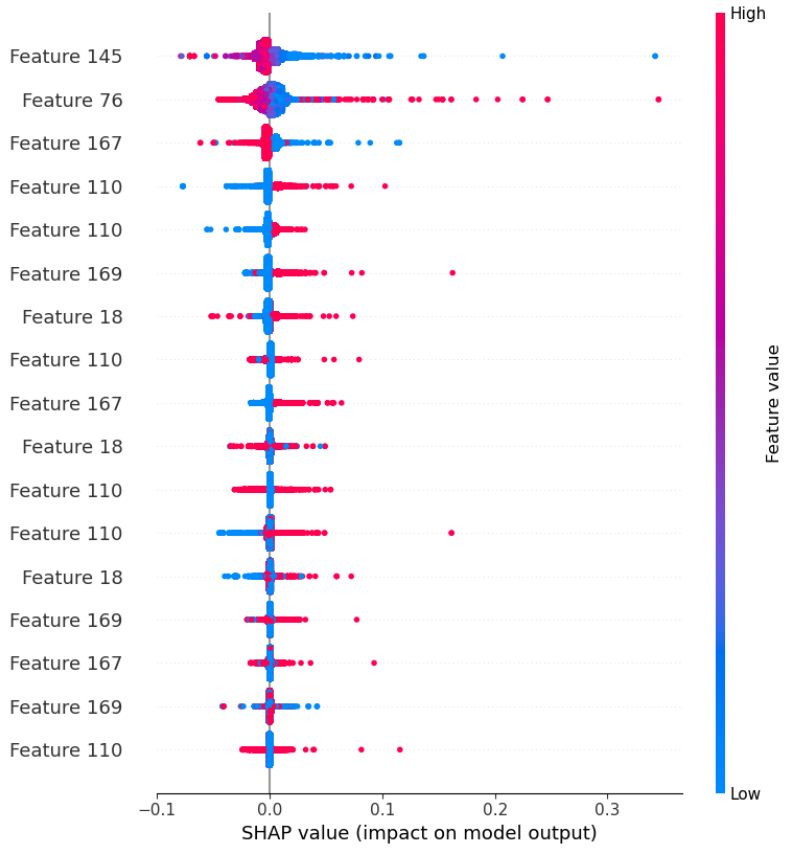
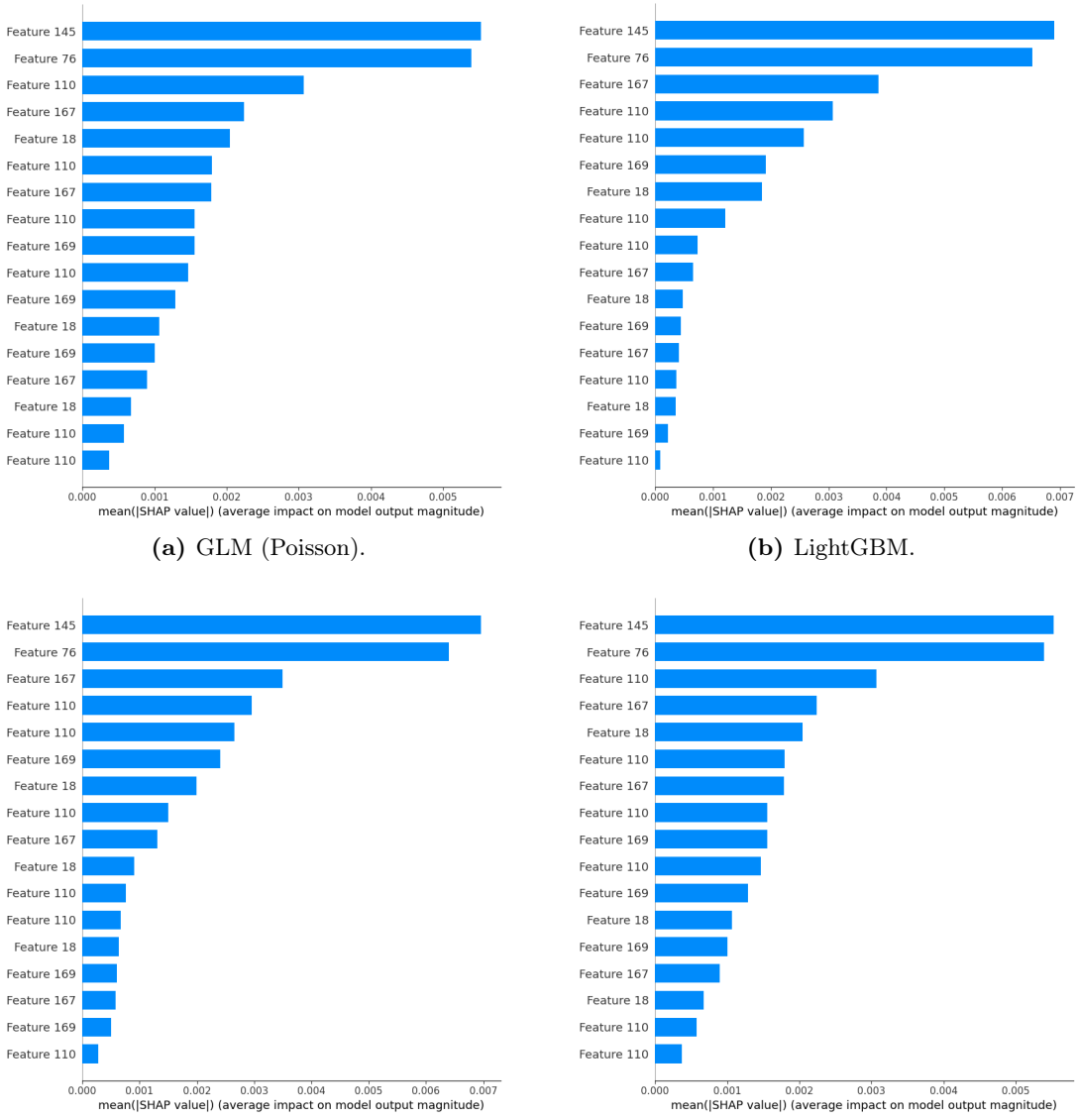


Figure 5.8: Beeswarm plot made for the XGBoost model based on the frequency data of `the insurance company`, showing global explanations for the respective features within the model.

The Beeswarm plots for the four models (GLM, LightGBM, XGBoost, and MLP) are visualized together in Figure D.15 of Appendix D.2.2.1. Across all these four explanations, certain features, such as `Feature 145`, `Feature 76`, `Feature 110`, and `Feature 167` consistently emerge as important, indicating their significant impact on the models' predictions. This is the case despite the differences in model architecture and underlying assumptions. The order of top features differs slightly between models, such as `Feature 167` and `Feature 169` being more prominent in XGBoost than in LightGBM and MLP, highlighting how different models prioritize features based on their structure and how they capture relationships within the data. All models display a mix of positive and negative SHAP values for these features, suggesting that their influence on predictions can vary depending on their specific values. The range of SHAP values varies significantly across models. XGBoost shows a broader range of SHAP values (from -0.1 to 0.3), suggesting it might be assigning more weight to certain features or capturing more complex interactions compared to LightGBM and MLP, which have SHAP values in a narrower range. This could indicate potentially more conservative predictions by these models. Based on this information, it might make sense that XGBoost was the best-performing model in Section 5.1.1.2.

The analysis of the beeswarm plots for the models show that all models identify key features as important, they differ in the weight they assign to these features and in their ability to capture complex interactions, with XGBoost appearing to handle more complex relationships and GLM reflecting a simpler, linear approach.

The importance of individual features can be potentially better captured through the SHAP bar plots, which are visualized for all four models in Figure 5.9. In all four models analyzed, `Feature 145` is consistently identified as the most important feature, as it appears at the top of each plot with the highest mean SHAP value. `Feature 76` and `Feature 110` are also highly influential across all models, usually ranking within the top three features. `Feature 18`, `Feature 167`, and `Feature 169`'s importance varies slightly depending on the model.



(a) GLM (Poisson).

(b) LightGBM.

(c) XGBoost.

(d) MLP.

Figure 5.9: Bar plots made for the four models based on the frequency data of `the insurance company`, showing global feature contributions based on SHAP values for the respective features within the model.

5.2.2.2 Local Explanations: SHAP Waterfall & Force Plots, LIME

The waterfall plot depicted in Figure 5.5 provides a local explanation for the predictions of the XGBoost model based on the severity data from `the insurance company`. In this waterfall plot, the base value, which is the average model output across all predictions for the target *"number of claims"* (in €), is denoted as $E[f(X)] = 0.040123$. The contributions of the various features towards the target *"claim burden"* are then added or subtracted from this base value to arrive at the final model prediction, a *"claim burden"* (in €) of $f(x) = 0.031244$. This means that this particular person has an expected number of claims lower than the average.

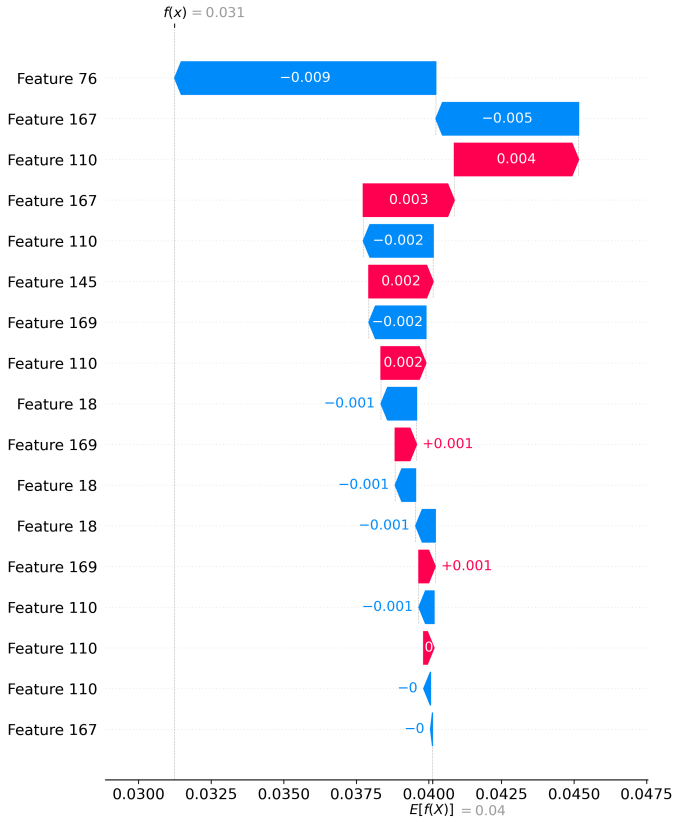


Figure 5.10: Waterfall plot made for the XGBoost model based on the frequency data of `the insurance company`, showing local feature contributions based on SHAP values for the respective features within the model.

The waterfall plot visualized in Figure 5.10 starts with a base value at the bottom, which represents the average prediction made by the model across all observations (in this case, around 0.04013). As you move from the bottom to the top of the waterfall, the effects of individual features are cumulatively added or subtracted from the base value, leading to the final prediction $f(x)$. The features are ordered by the magnitude of their contributions, with the features having the most impact based on their contribution towards the target appearing first (at the top). In this case, `Feature 76`'s value has the highest impact on the prediction, which could be derived by looking at the plot and seeing the largest horizontal bar within the model (-0.008998). In this instance, `Feature 76`'s feature value made the model decrease the expected number of claims. `Feature 167` also is a significant feature in lowering the predicted value. If one would predict the claim frequency (number of

claims), these might be indicators of lower-risk cases, as the predicted value is lower than the average. Features that have both positive and negative contributions (e.g., `Feature 110`) are involved in OHE, where some categorical classes impact the prediction positively, while others impact the prediction negatively. `Feature 18` is for all its bins in the lower ranges for the XGBoost model, indicating that this feature was not driving the claim frequency predictions for this respective individual. If such features consistently show small contributions across many instances, they might be candidates for removal to simplify the model, unless they provide critical interpretability.

Looking at Appendix D.2.2.2, where all three ML models have their respective waterfall plot, one can identify various similarities and differences. In all these three models, `Feature 76` has a significant negative contribution to the prediction, while `Feature 145` shows a consistent positive contribution across the different models. The magnitude of `Feature 76`'s impact is relatively similar across the models (strongest for LightGBM model in Figure D.16a with a value of -0.009790), indicating that the feature consistently decreases the prediction value for the given instance.

Despite these similarities, there are notable differences in how each model weighs and interacts with the features. The MLP model, for example, appears to focus on a narrower set of features, with only a few, such as `Feature 145` and `Feature 76`, having significant impacts. Interestingly, while `Feature 145` has a strong negative impact in the MLP model, this effect is not as prominent in the other two models, highlighting how MLP might weigh certain features differently. Understanding these differences between models can eventually help in choosing the most appropriate model based on the specific characteristics and requirements of the use-case at hand.

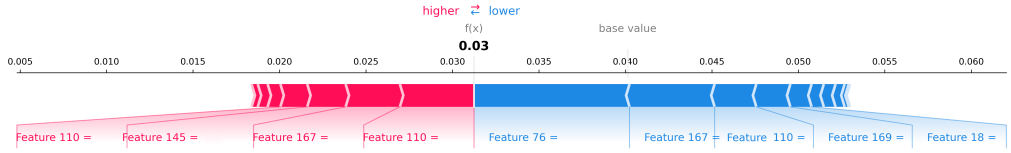


Figure 5.11: Force plot made for the XGBoost model based on the frequency data of `the insurance company`, showing local feature contributions based on SHAP values for the respective features within the model.

As discussed earlier in Section 5.2.1.1, force plots can also be utilized to derive similar insights as those provided by the waterfall plots. Consequently, the results for the XGBoost model are briefly illustrated in Figure 5.11, with additional results presented in Appendix D.2.2.3. Similarly as for severity, LIME explanations were considered. However, due to their non-deterministic nature, they were not found to be sufficiently reliable for extracting meaningful insights from the frequency results. Therefore, insights are not explained in this section. One could find the LIME results in Appendix D.2.2.4.

5.3 Interview Evaluation

The interviews, which are individually discussed in Appendix E, collectively try to answer questions related to current insurance practices, model designs of GLMs and ML techniques, model validation, and the evaluation techniques based on XAI outcomes that

could be deployed. An overview of the interview process is provided in Table 4.8 of Section 4.7.2.3.

Within the five interviews conducted, GLMs are widely regarded as the industry standard in insurance pricing due to their explainability, regulatory acceptance, and ease of validation. All interviewees recognize the robustness and reliability of GLMs in delivering transparent and interpretable results, which is crucial for compliance and stakeholder trust. While ML models are acknowledged for their potential to enhance predictive accuracy, especially in handling complex datasets and capturing intricate patterns, they are currently seen as supplementary rather than replacement tools. The main challenges associated with ML models include their "black box" nature and lack of transparency.

"Explainability is a critical factor in the usage of insurance pricing models."

Explainability is another critical factor in the usage of pricing models. GLMs are favored by all interviewees for their straightforward interpretation, which is essential for justifying premium calculations to customers and meeting regulatory requirements. While tools like SHAP and LIME are being explored to explain ML model outputs, their current effectiveness compared to GLM factor explanations is still debated by the experts, with concerns about their ability to fully replicate the transparency offered by GLMs.

"Various challenges that need to be addressed with ML."

The transition from GLMs to ML models comes with several challenges, including the need for specialized knowledge, computational power, and infrastructure changes. The complexity of ML models demands significant training and resources, which are currently seen as barriers to widespread adoption in the industry. The regulatory environment heavily influences model adoption in insurance. The interviews consistently emphasize that any transition to ML models must address concerns about regulatory compliance and maintain the transparency necessary for both internal validation and customer-facing applications. The interviews reveal varying views on the level of explainability necessary for insurance pricing models. Some individuals, particularly those with extensive experience in traditional actuarial roles, emphasize the critical importance of explainability, favoring GLMs. Others take a more balanced view, acknowledging the importance of explainability while also recognizing the potential benefits of ML models. They are open to using tools like SHAP and LIME to mitigate the transparency issues of ML models.

"A hybrid approach would allow for insurers to benefit from the strengths of both models while mitigating the risks associated with ML's complexity and lack of transparency."

Looking ahead, Interview C suggested a hybrid approach, where ML models could be used alongside GLMs, particularly in enhancing feature selection and fine-tuning predictions. This idea was supported by the other stakeholders with an informal talk. The hybrid approach would allow insurers to benefit from the strengths of both models while mitigating the risks associated with ML's complexity and lack of transparency. There is also a consensus that for ML models to gain wider acceptance, significant improvements in explainability tools are needed. Developing more robust and deterministic methods for explaining ML model outputs could help bridge the gap between the precision of ML and the interpretability of GLMs. The interviews suggest that a gradual, well-validated adoption of ML models could occur in the future, provided that challenges around explainability, validation, and resource allocation are adequately addressed. There is optimism about the potential of ML models to revolutionize insurance pricing, but this will require careful planning and stakeholder acceptance.

"There is still significant variance in the level of adoption of pricing models across the insurance industry."

The interviews reveal a significant variance in the level of adoption of pricing models across the insurance industry, with larger companies using a dual approach that combines traditional methods with GLMs and sometimes ML, while smaller insurers may still rely on these simpler models alone. The integration of ML models might progress differently across companies depending on their size, resources, and regulatory environment.

"Interest in ML adoption is growing, but is dependent on overcoming significant challenges related to explainability, validation, and resource requirements."

The interviews collectively show that there is a growing interest in ML models due to their potential to enhance predictive power. However, the widespread adoption of ML models is dependent on overcoming significant challenges related to explainability, validation, and resource requirements. The (near) future likely holds a hybrid approach, where ML models in combination with the XAI tools such as SHAP and LIME complement rather than replace GLMs, with a gradual integration shaped by ongoing advancements in technology and explainability tools.

The key points from the interviews conducted with the respective individuals can be summarized as follows:

- ◇ *GLMs.* GLMs are widely regarded as the industry standard in insurance pricing due to their explainability, regulatory acceptance, and ease of validation.
- ◇ *ML.* While ML models are recognized for their potential to enhance predictive accuracy, they are currently seen as supplementary tools rather than replacements for GLMs. The main challenges associated with ML models include their "black box" nature and lack of transparency. A hybrid approach is suggested
- ◇ *Explainability.* Explainability is crucial in the adoption of pricing models. GLMs are favored for their straightforward interpretation, which is essential for justifying premium calculations and meeting regulatory requirements. However, various interviewees see the potential of ML for the future, as there are differing views on the level of explainability necessary. Some emphasize the critical importance of explainability, while others are open to using ML models with appropriate tools to mitigate transparency issues.
- ◇ *XAI.* Tools like SHAP and LIME are being explored to improve the explainability of ML models, but their effectiveness compared to GLMs is still debated. However, it was recognized that these tools could provide additional insights.
- ◇ *Transitioning.* The transition from GLMs to ML models is challenged by the need for specialized knowledge, computational power, and infrastructure upgrades, as well as the complexity of ML models, which demands significant training and resources. There is significant variance in the adoption of pricing models across the insurance industry, with larger companies using a dual approach combining traditional methods with GLMs and sometimes ML, while smaller insurers may still rely solely on simpler models.
- ◇ *Gradual Adoption.* Adoption of ML models is anticipated in the future, but should be taking with caution due to challenges such as explainability, validation, and resource allocation are addressed.

5.4 Summary

This chapter analyzed various models for the prediction of insurance claim severity and frequency, such as the GLM (Gamma & Poisson), LightGBM, XGBoost, and MLP models. The MLP model showed the best predictions for severity, though it required longer tuning times, while GLM lagged behind the ML models (both the GBMs and MLP). For frequency predictions, LightGBM and XGBoost offered balanced performance, with XGBoost slightly ahead in predictive power and efficiency. Therefore, the MLP model was chosen as the main model for severity and XGBoost as the main model for frequency. Validation of the models using APP plots revealed that ML models, particularly LightGBM and XGBoost, aligned more closely with actual data trends compared to GLM, especially in complex data segments.

The chapter also explored model explainability using SHAP and LIME. SHAP's global and local explanations highlighted key features influencing predictions, with some features being particularly impactful. SHAP's results (both global and local) were deemed useful, and could provide various insights into the inner workings of the more complex ML algorithms. However, LIME's inconsistent results made it less reliable and thus less useful, particularly in a regulated financial sector.

Interviews with industry professionals underscored the continued reliance on GLMs due to their explainability and regulatory compliance, though there is interest in ML models for their potential to improve predictions, where the incorporation of XAI techniques helped explaining model reasoning. Fully transitioning to ML-based pricing models may be premature. Challenges such as model complexity, transparency, and the need for specialized resources must first be addressed.

ML models offer significant benefits in predictive accuracy, their integration into insurance pricing will likely be gradual and slow, requiring improvements in explainability and regulatory acceptance. A hybrid model combining the strengths of both GLMs and ML models appears to be the most feasible path forward. This hybrid approach, combining GLMs with ML models and supported by robust explanation tools like SHAP, could be a potential solution to improve insurer's insights in their data.

Chapter 6

Conclusion

This thesis explored the integration of ML techniques into insurance pricing models, specifically as a replacement for the industry-standard GLMs. The central research question guiding this study was:

How can ML be applied or integrated into insurance pricing forecasting to replace traditional statistical models (GLMs) in such a way that interpretability and explainability are taken into account?

To address this question, three sub-questions were investigated, where **RQ1** was defined as: *How does a GLM work, and how is explainability ensured in this context?* The answer to this question is that GLMs, particularly within the context of insurance pricing, function by establishing a linear relationship between predictors and the target variable, allowing for easy interpretation of the impact of each variable. This inherent transparency is achieved by the simplicity of the model's structure, making GLMs the preferred choice when explainability is crucial. The explainability in GLMs is largely due to their predictable and understandable nature, which aligns with the regulatory requirements and stakeholder needs for transparent pricing models.

For the second sub-question, **RQ2**, the question investigated: *How do ML models compare to GLM approaches, and how can ML models be integrated into the insurance pricing process?* The answer to this, based on investigation in this thesis, is that ML models (techniques like LightGBM, XGBoost, and a MLP) could offer better prediction power compared to GLMs, particularly in complex datasets, which could be derived from the model exploration done with more variables. However, integration of these models into insurance pricing is challenging due to the higher complexity and lower transparency. The thesis found that while ML models can capture non-linear relationships and intricate patterns that GLMs might miss, their explainability remains a significant hurdle. The study suggests that the best way of integrating ML models into the pricing process would be via a hybrid approach, where ML models are supplemented with explainability tools like SHAP and LIME to maintain a level of transparency comparable to GLMs. This would allow insurers to focus on challenges such as explainability, validation, and resource requirements.

RQ3, the last sub-question, looks at: *How can one validate the explainability outcomes through the usage of XAI methods in ML-based insurance pricing models?* The thesis validated the explainability of ML models using SHAP and LIME. SHAP provided consistent global and local explanations that were useful for understanding the impact of individual features on predictions, making it a reliable tool for explaining (complex) ML models. However, LIME was found to be less consistent, particularly in regulated environments where

stable and predictable explanations are necessary. The validation interviews with industry experts underscored the importance of explainability, with stakeholders expressing a preference for models that balance predictive power with explainability and transparency of these models, as this would be required for compliance as well. The findings suggest that while SHAP can be effectively used to explain ML models, a full transition to ML-based pricing may require either further advancements in explainability techniques or the adoption of less complex, inherently transparent ML models that reduce the need for extensive post-hoc analysis.

Final conclusion. Answering the main research question of the thesis: the integration of ML models into insurance pricing holds promise for improving predictive accuracy but poses significant challenges in terms of explainability and regulatory acceptance. The study recommends a gradual adoption of ML, potentially through a hybrid model that combines the strengths of both GLMs and ML techniques explored within the thesis, supported by robust explainability tools like SHAP. This approach could provide the necessary balance between boosting performance and the level of explainability and transparency required by the insurance industry.

Limitations & Future Research

While the thesis offered valuable insights into the application of ML and XAI in insurance pricing, several limitations must be acknowledged. These limitations point to potential areas for future research, particularly concerning data constraints, model limitations, and the sample size of the conducted interviews. The key limitations identified are as follows:

- ◇ **Data constraints.** Like practically every study, this research was constrained by the availability of data, mainly for severity, which could potentially have influenced the outcomes of the model performance and validation.

- ◇ **Model limitations.** The study primarily focused on specific ML models like LightGBM, XGBoost, and MLP, which were chosen via an AutoML approach which incorporated a different feature set compared to the final ML models. Also, AutoML does not incorporate all models, which made the scope smaller in the process. Other potentially suitable models were not explored in depth in this thesis. Alternative ML models, particularly those with inherent transparency, could have been more appropriate for the research problem, offering a balance between performance and explainability that better satisfies stakeholder requirements.

- ◇ **Interview sample size.** The sample size for the interviews was relatively small (5), which may limit the generalizability of the findings. While the insights obtained are valuable, it could potentially not capture the broader insurance industry's perspective.

- ◇ **Post-hoc explainability reliance.** The study relied heavily on post-hoc explainability tools such as SHAP and LIME to interpret the complex ML models. While these XAI techniques proved to be able to provide valuable insights, their application is not without limitations such as robustness and consistency, which is crucial in regulated environments such as the insurance sector.

- ◇ **Practical implementation.** The integration of ML models into insurance pricing was explored theoretically and through model validation, but a full-scale practical implementation within an operational environment was beyond the scope of this

thesis. This was, however, also never really an option, as the interviews showed that a full-size practical implementation of ML models could prove to be too soon.

Building upon the findings and limitations discussed in the previous sections, several opportunities for future research in the application of ML in insurance pricing emerge. By exploring these possibilities, future studies can contribute to advancing this research field and work towards the development of more effective, transparent, and adaptable XAI models in insurance pricing. The following areas for future exploration have been identified:

- ◇ **Other preprocessing methods**: Implementing techniques such as BO or GS for hyper-parameter tuning could enhance model performance and robustness by systematically exploring and optimizing model parameters. This could also be the case for other scaling and encoding methods potentially.
- ◇ **Other ML techniques**: Expanding (or changing) the range of ML techniques evaluated for insurance pricing could give further insights in the potential of ML. It is recommended to begin with simpler models before progressing to more complex, NN-based models, as the regulatory landscape in the financial sector is likely to favor more interpretable ML techniques in the close term. These models, therefore, have greater immediate applicability. Techniques such as DTs, Bayesian Networks, or K-Nearest Neighbors (KNN) would be best suitable in this context [52].
- ◇ **Evaluation metrics**: Future research should also consider using a broader set of evaluation metrics beyond those used in this study. Metrics that assess fairness, robustness, and long-term stability of the models would provide a more comprehensive understanding of their performance in the insurance sector. This could also be done for assessing explainability results of the various interviews.
- ◇ **Other XAI techniques**: While SHAP and LIME were the main XAI techniques used in this study to explain the ML models created, other XAI techniques or ML models with more inherent interpretability should be investigated.
- ◇ **Variable inclusion**: Expanding the dataset to include more variables, particularly those related to behavioral data and external economic factors, could improve the predictive power and relevance of the ML models. The current incorporation of features is low, due to the specific choice of taking the same feature selection process as `the insurance company`. The integration of more might lead to pricing models with more predictive power.
- ◇ **Validation on different data**: Using additional datasets, such as the window data, to validate claims and assess model performance could provide a more robust evaluation and increase the generalizability of the findings.
- ◇ **Increasing interviewee sample size**: As discussed in the limitations of this research, enhancing the reliability and generalizability of the findings could by achieved by involving a larger (and more diverse) sample of industry experts. This would provide a more comprehensive view of the industry's readiness to adopt ML models and the challenges they foresee.
- ◇ **Ontology-based ML**: Future studies should explore the integration of ontology-based ML techniques in insurance pricing. Ontology-based approaches can provide a more structured representation of the data, which might enhance the interpretability and explainability of ML models.

Bibliography

[1] B. Beers, *A brief overview of the insurance sector*, May 2023. [Online]. Available: https://www.investopedia.com/ask/answers/051915/how-does-insurance-sector-work.asp.

[2] A. A. Turgaeva, L. V. Kashirskaya, Y. A. Zurnadzhyants, O. A. Latysheva, I. V. Pustokhina, and A. V. Sevbitov, "Assessment of the financial security of insurance companies in the organization of internal control," *Entrepreneurship and Sustainability Issues*, vol. 7, no. 3, pp. 2243–2254, Mar. 2020. DOI: 10.9770/jesi.2020.7.3(52).

[3] M. Rosanes, *Insurance premium: What is it and how does it work?* Dec. 2022. [Online]. Available: https://www.insurancebusinessmag.com/us/guides/insurance-premium-what-is-it-and-how-does-it-work-430049.aspx.

[4] EY, 2023. [Online]. Available: https://assets.ey.com/content/dam/ey-sites/ey-com/es_es/topics/insurance/ey-non-life-insurance-pricing.pdf.

[5] Indeed, *Actuarial pricing vs valuation*, Jun. 2022. [Online]. Available: https://www.indeed.com/career-advice/finding-a-job/actuarial-pricing-vs-valuation.

[6] E. Erguen, *Why machine learning for insurance pricing is a game-changer*, Jul. 2023. [Online]. Available: https://www.sapfioneer.com/blog/blogpost/why-machine-learning-for-insurance-pricing-is-a-game-changer/.

[7] Origin, *The changing landscape of actuarial services in the insurance industry*, Jul. 2023. [Online]. Available: https://www.originaffinity.com/the-changing-landscape-of-actuarial-services-in-the-insurance-industry/.

[8] J. Kagan, *What is actuarial science? definition and examples of application*, Sep. 2023. [Online]. Available: https://www.investopedia.com/terms/a/actuarial-science.asp.

[9] Google, *Ai vs. machine learning: How do they differ? | google cloud*, 2024. [Online]. Available: https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning#.

[10] V. Yakymiv, *The impact and applications of machine learning in insurance*, Feb. 2024. [Online]. Available: https://forbytes.com/blog/machine-learning-in-insurance/.

[11] S. Toms, D. A. Simon, E.-C. Vermynck, and J. A. Kamyar, *Ai insights: How regulators worldwide are addressing the adoption of ai in financial services: Insights: Skadden, arps, slate, meagher flom llp*, Dec. 2023. [Online]. Available: https://www.skadden.com/insights/publications/2023/12/how-regulators-worldwide-are-addressing-the-adoption-of-ai-in-financial-services.

[12] Hslade, *The role of machine learning explainability in insurance*, Jan. 2024. [Online]. Available: https://fintech.global/2024/01/31/the-role-of-machine-learning-explainability-in-insurance/.

[13] A. Riskfinance, *Triple a risk finance - consultancy in risk management and actuarial*, Aug. 2019. [Online]. Available: https://www.aaa-riskfinance.nl/en/about-us/.

[14] A. Riskfinance, *Triple a risk finance - consultancy in risk management and actuarial*, Sep. 2022.

[15] T. Co., *Over turien co.*. 2024. [Online]. Available: https://turien.nl/over-ons/over-turien-co.

[16] TurienHolding, 2023. [Online]. Available: https://www.turienholding.nl/assets/JaarverslagTurienenCoHolding2021.pdf.

[17] Turien, *Linkedin turien co*, Apr. 2024. [Online]. Available: https://www.linkedin.com/company/turien-&-co-assuradeuren/posts/?feedView=all.

[18] L. Cunha and J. M. Bravo, "Automobile usage-based-insurance: Improving risk management using telematics data," *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, Jun. 2022. DOI: 10.23919/cisti54924.2022.9820146.

[19] R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen, "Boosting insights in insurance tariff plans with tree-based machine learning methods," *North American Actuarial Journal*, vol. 25, no. 2, pp. 255–285, Jul. 2020. DOI: 10.1080/10920277.2020.1745656.

[20] M. David, "Auto insurance premium calculation using generalized linear models," *Procedia Economics and Finance*, vol. 20, pp. 147–156, 2015. DOI: 10.1016/s2212-5671(15)00059-3.

[21] S. C. Lee, "Delta boosting implementation of negative binomial regression in actuarial pricing," *Risks*, vol. 8, no. 1, p. 19, Feb. 2020. DOI: 10.3390/risks8010019.

[22] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–39, Apr. 2000. DOI: 10.1.1.198.5133.

[23] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. DOI: 10.1016/j.procs.2021.01.199.

[24] F. Rodrigues, *Key differences between life and general insurance – hdfc life*, Nov. 2023. [Online]. Available: https://www.hdfclife.com/insurance-knowledge-centre/about-life-insurance/differences-between-life-and-general-insurance.

[25] Rijksoverheid, *Third-party liability insurance for motor vehicles*, 2024. [Online]. Available: https://business.gov.nl/regulation/vehicle-insurance/.

[26] Expat, *Car insurance in the netherlands*, 2024. [Online]. Available: https://www.iamexpat.nl/expat-info/driving-netherlands/car-insurance-netherlands.

[27] Mar. 2024. [Online]. Available: https://www.expatica.com/nl/finance/insurance/car-insurance-in-the-netherlands-115185/.

[28] ICNZ, Oct. 2019. [Online]. Available: https://www.icnz.org.nz/wp-content/uploads/2023/01/ICNZ_Guide_to_Insurance_Pricing_Oct19_Updated.pdf.

[29] R. Netherlands Enterprise Agency, *Insurance premium tax*, 2024. [Online]. Available: https://business.gov.nl/regulation/insurance-premium-tax/.

[30] A. Sajumon, *Pure premium*, Oct. 2023. [Online]. Available: https://www.fisdom.com/glossary/pure-premium/#:~:text=Pure%20Premium%20is%20the%20fundamental,profit%20margins%20for%20the%20insurer..

[31] E. Frees, G. Lee, and L. Yang, "Multivariate frequency-severity regression models in insurance," *Risks*, vol. 4, no. 1, p. 4, Feb. 2016. DOI: 10.3390/risks4010004.

[32] M. Goldburd, A. Khare, D. Tevet, and D. Guller, *Generalized Linear Models For Insurance Rating*, 2nd ed. Casualty Actuarial SocietyCasualty Actuarial Society, 2020, vol. 2.

[33] M. Denuit, A. Charpentier, and J. Trufin, "Autocalibration and tweedie-dominance for insurance pricing with machine learning," *Insurance: Mathematics and Economics*, vol. 101, pp. 485–497, Nov. 2021. DOI: 10.1016/j.insmatheco.2021.09.001.

[34] W. Yue, *Non-life insurance pricing models*, Apr. 2021. [Online]. Available: https://wenyue2021.medium.com/non-life-insurance-pricing-models-ace87fe63e9.

[35] E. Ohlsson and B. Johansson, "The basics of pricing with glms," *EAA Lecture Notes*, pp. 15–38, 2010. DOI: 10.1007/978-3-642-10791-7_2.

[36] C. Clemente, G. R. Guerreiro, and J. M. Bravo, "Modelling motor insurance claim frequency and severity using gradient boosting," *Risks*, vol. 11, no. 9, p. 163, Sep. 2023. DOI: 10.3390/risks11090163.

[37] Y.-L. Grize, W. Fischer, and C. Lützelschwab, "Machine learning applications in nonlife insurance," *Applied Stochastic Models in Business and Industry*, vol. 36, no. 4, pp. 523–537, May 2020. DOI: 10.1002/asmb.2543.

[38] T. Poufinas, P. Gogas, T. Papadimitriou, and E. Zaganidis, "Machine learning in forecasting motor insurance claims," *Risks*, vol. 11, no. 9, p. 164, Sep. 2023. DOI: 10.3390/risks11090164.

[39] Avcontentteam, *Applications of machine learning and ai in insurance in 2024*, Feb. 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/03/applications-of-machine-learning-and-ai-in-insurance/.

[40] B. R. E. T. |. M. 14, R. E. Team, B. P. I. C. |. A. 3, and P. I. Companies, *Insurance industry increasingly adopting ai technologies, study shows*, Mar. 2024. [Online]. Available: https://riskandinsurance.com/insurance-industry-increasingly-adopting-ai-technologies-study-shows/.

[41] H. Hassani, S. Unger, and C. Beneki, "Big data and actuarial science," *Big Data and Cognitive Computing*, vol. 4, no. 4, p. 40, Dec. 2020. DOI: 10.3390/bdcc4040040.

[42] Y. Huang and S. Meng, "Automobile insurance classification ratemaking based on telematics driving data," *Decision Support Systems*, vol. 127, p. 113 156, Dec. 2019. DOI: 10.1016/j.dss.2019.113156.

[43] S. Devriendt, K. Antonio, T. Reynkens, and R. Verbelen, "Sparse regression with multi-type regularized feature modeling," *Insurance: Mathematics and Economics*, vol. 96, pp. 248–261, Jan. 2021. DOI: 10.1016/j.insmatheco.2020.11.010.

[44] J. S. Chan, S. T. Choy, U. Makov, A. Shamir, and V. Shapovalov, "Variable selection algorithm for a mixture of poisson regression for handling overdispersion in claims frequency modeling using telematics car driving data," *Risks*, vol. 10, no. 4, p. 83, Apr. 2022. DOI: 10.3390/risks10040083.

[45] DNB, *General principles for the use of artificial intelligence in the ...* 2019. [Online]. Available: https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf.

[46] S. Loisel, P. Piette, and C.-H. J. Tsai, "Applying economic measures to lapse risk management with machine learning approaches," *ASTIN Bulletin*, vol. 51, no. 3, pp. 839–871, Jun. 2021. DOI: 10.1017/asb.2021.10.

[47] K. Kuo, "Towards explainability of machine learning models in insurance pricing," *Variance*, Mar. 2020. DOI: 10.48550/arXiv.2003.10674.

[48] Z. Díaz Martínez, J. Fernández Menéndez, and L. J. García Villalba, "Tariff analysis in automobile insurance: Is it time to switch from generalized linear models to generalized additive models?" *Mathematics*, vol. 11, no. 18, p. 3906, Sep. 2023. DOI: 10.3390/math11183906.

[49] E. Owens, B. Sheehan, M. Mullins, M. Cunneen, and J. Ressel, "Explainable artificial intelligence (xai) in insurance: A systematic review," *SSRN Electronic Journal*, 2022. DOI: 10.2139/ssrn.4088029.

[50] L. Barry and A. Charpentier, "Personalization as a promise: Can big data change the practice of insurance?" *Big Data amp; Society*, vol. 7, no. 1, p. 205 395 172 093 514, Jan. 2020. DOI: 10.1177/2053951720935143.

[51] A. Cevolini and E. Esposito, "From pool to profile: Social consequences of algorithmic prediction in insurance," *Big Data amp; Society*, vol. 7, no. 2, p. 205 395 172 093 922, Jul. 2020. DOI: 10.1177/2053951720939228.

[52] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020. DOI: 10.1016/j.inffus.2019.12.012.

[53] E. Aizenberg and J. van den Hoven, "Designing for human rights in ai," *Big Data amp; Society*, vol. 7, no. 2, p. 205 395 172 094 956, Jul. 2020. DOI: 10.1177/2053951720949566.

[54] S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy, responsible ai," *Frontiers in Artificial Intelligence*, vol. 5, Feb. 2022. DOI: 10.3389/frai.2022.779799.

[55] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019. DOI: 10.1016/j.artint.2018.07.007.

[56] C. Panigutti, R. Hamon, I. Hupont, *et al.*, "The role of explainable ai in the context of the ai act," *2023 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2023. DOI: 10.1145/3593013.3594069.

[57] Btd, *Explainable ai (xai): Model-specific interpretability methods*, Jan. 2024. [Online]. Available: https://baotramduong.medium.com/explainable-ai-model-specific-interpretability-methods-02e23ebceac1#:~:text=Model%2Dspecific%20interpretability%20methods%20are,underlying%20model%20to%20generate%20explanations..

[58] D. Garreau and U. von Luxburg, "Theoretical analysis of lime," *Explainable Deep Learning AI*, pp. 293–316, 2023. DOI: 10.1016/b978-0-32-396098-4.00020-x.

[59] A. Maillart, "Toward an explainable machine learning model for claim frequency: A use case in car insurance pricing with telematics data," *European Actuarial Journal*, vol. 11, no. 2, pp. 579–617, Mar. 2021. DOI: 10.1007/s13385-021-00270-5.

[60] A. Fabris, A. Mishler, S. Gottardi, *et al.*, "Algorithmic audit of italian car insurance: Evidence of unfairness in access and pricing," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Jul. 2021. DOI: 10.1145/3461702.3462569.

[61] K. McDonnell, F. Murphy, B. Sheehan, L. Masello, and G. Castignani, "Deep learning in insurance: Accuracy and model interpretability using tabnet," *Expert Systems with Applications*, vol. 217, p. 119 543, May 2023. DOI: 10.1016/j.eswa.2023.119543.

[62] O. Koster, R. Kosman, and J. Visser, "A checklist for explainable ai in the insurance domain," *Communications in Computer and Information Science*, pp. 446–456, 2021. DOI: 10.1007/978-3-030-85347-1_32.

[63] M. Mullins, C. P. Holland, and M. Cunneen, "Creating ethics guidelines for artificial intelligence (ai) and big data analytics customers: The case of the consumer european insurance market," *SSRN Electronic Journal*, 2021. DOI: 10.2139/ssrn.3858056.

[64] N. Mehndiratta, G. Arora, and R. Bathla, "The use of artificial intelligence in the banking industry," *2023 International Conference on Recent Advances in Electrical, Electronics amp; Digital Healthcare Technologies (REEDCON)*, May 2023. DOI: 10.1109/reedcon57544.2023.10150681.

[65] T. Leonardi, *What the gdpr means for the insurance industry*, Sep. 2018. [Online]. Available: https://www.linkedin.com/pulse/what-gdpr-means-insurance-industry-thomas-leonardi/.

[66] Trail, *Eu ai act: Risk-classifications of the ai regulation*, Apr. 2023. [Online]. Available: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified.

[67] KPMG, 2022. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2022/07/modern-risk-management-for-ai-models.pdf.

[68] Fed, *Board of governors of the federal reserve system*, 2011. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

[69] KPMG, 2018. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/ie/pdf/2018/03/ie-gdpr-for-insurance-industry.pdf.

[70] A. Kumar, *Generalized linear models explained with examples*, Oct. 2022. [Online]. Available: https://vitalflux.com/generalized-linear-models-explained-with-examples/.

[71] M. V. Wuthrich and C. Buser, "Data analytics for non-life insurance pricing," *SSRN Electronic Journal*, 2017. DOI: 10.2139/ssrn.2870308.

[72] İ. Kılıç, *Light gbm: A powerful gradient boosting algorithm*, Sep. 2023. [Online]. Available: https://medium.com/@ilyurek/light-gbm-a-powerful-gradient-boosting-algorithm-fe145a1cd8a6.

[73] ArcGIS, *How lightgbm algorithm works*, 2017. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-lightgbm-works.htm.

[74] N. S. Chauhan, *Lightgbm demystified: Understanding the math behind the algorithm*, Jun. 2024. [Online]. Available: https://www.theaidream.com/post/lightgbm-demystified-understanding-the-math-behind-the-algorithm.

[75] D. Thakran, *Boosting model accuracy with ensemble learning*, Aug. 2023. [Online]. Available: https://medium.com/@thakrandisharth/boosting-model-accuracy-with-ensemble-learning-2742f360ae0.

[76] M. Noorunnahar, A. H. Chowdhury, and F. A. Mila, "A tree based extreme gradient boosting (xgboost) machine learning model to forecast the annual rice production in bangladesh," *PLOS ONE*, vol. 18, no. 3, Mar. 2023. DOI: 10.1371/journal.pone.0283452.

[77] J. Hoare, *Gradient boosting explained - the coolest kid on the machine learning block*, Jan. 2024. [Online]. Available: https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/.

[78] M. Dancho, *Xgboost: A must have data science tool*, Jan. 2024. [Online]. Available: https://www.linkedin.com/posts/mattdancho_xgboost-is-now-my-go-to-number-1-must-have-activity-7158152248865275904-6GDo.

[79] T. Chen and T. He, *Xgboost: Extreme gradient boosting*. [Online]. Available: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.

[80] C. Leo, *The math behind xgboost*, Jan. 2024. [Online]. Available: https://medium.com/@cristianleo120/the-math-behind-xgboost-3068c78aad9d.

[81] baeldung, *Multi-layer perceptron vs. deep neural network*, Jun. 2023. [Online]. Available: https://www.baeldung.com/cs/mlp-vs-dnn.

[82] C. Bento, *Multilayer perceptron explained with a real-life example and python code: Sentiment analysis*, Sep. 2021. [Online]. Available: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141.

[83] Shankar, *Understanding loss function in deep learning*, Apr. 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/06/understanding-loss-function-in-deep-learning/.

[84] J. Brownlee, *A gentle introduction to the rectified linear unit (relu)*, Aug. 2020. [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/.

[85] Y. Chen, L. Li, W. Li, Q. Guo, Z. Du, and Z. Xu, *Ai Computing Systems: An application driven perspective*. Morgan Kaufmann, an impritn of Elsevier, 2024.

[86] J. d. D. Nyandwi, *Why is it so hard to train neural networks?* Nov. 2021. [Online]. Available: https://jeande.medium.com/why-is-it-so-hard-to-train-neural-networks-433af273ce5f.

[87] C. Molnar, *Interpretable machine learning*, Aug. 2023. [Online]. Available: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html.

[88] C. Molnar, *Interpretable machine learning*, May 2024. [Online]. Available: https://christophm.github.io/interpretable-ml-book/shap.html.

[89] C. O'Sullivan, *Kernelshap vs treeshap*, Mar. 2023. [Online]. Available: https://towardsdatascience.com/kernelshap-vs-treeshap-e00f3b3a27db.

[90] ArizeAI, *Deep explainer (deep shap)*, Apr. 2023. [Online]. Available: https://arize.com/glossary/deep-explainer-deep-shap/.

[91] B. John, *How to use shap values to optimize and debug ml models*, Aug. 2023. [Online]. Available: https://neptune.ai/blog/shap-values.

[92] A. Cooper, *Explaining machine learning models: A non-technical guide to interpreting shap analyses*, Apr. 2024. [Online]. Available: https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/.

[93] Mar. 2022. [Online]. Available: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/.

[94] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?"," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016. DOI: 10.1145/2939672.2939778.

[95] D. Nguyen, *Explain your ml model predictions with local interpretable model-agnostic explanations (lime)*. Mar. 2020. [Online]. Available: https://medium.com/publicis-sapient-france/explain-your-ml-model-predictions-with-local-interpretable-model-agnostic-explanations-lime-82343c5689db.

[96] J. Brownlee, *Regression metrics for machine learning*, Feb. 2021. [Online]. Available: https://machinelearningmastery.com/regression-metrics-for-machine-learning/.

[97] P. M, *A comprehensive introduction to evaluating regression models*, Jun. 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#.

[98] J. Fernando, *R-squared: Definition, formula, uses, and limitations*, Jul. 2024. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp.

[99] F. Aust, *Explained variance score*, Oct. 2023. [Online]. Available: https://medium.com/@floaust/explained-variance-score-962a9d8ba778#:~:text=The%20Explained%20Variance%20Score%20is,is%20the%20best%20possible%20result..

[100] J. Schneppat, *Hyperparameter tuning in ml*, 2024. [Online]. Available: https://schneppat.com/hyperparameter-tuning-in-ml.html.

[101] N. Darapureddy, D. N. Karatapu, and D. T. Battula, "Research of machine learning algorithms using k-fold cross validation," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6s, pp. 215–218, Sep. 2019. DOI: 10.35940/ijeat.f1043.0886s19.

[102] R. Budiarto and K. Buana, "Performance comparison of feature extraction and machine learning classification algorithms for face recognition," *The IJICS (International Journal of Informatics and Computer Science)*, vol. 5, no. 3, p. 250, Nov. 2021. DOI: 10.30865/ijics.v5i3.3333.

[103] A. Jain, *Understanding cross-validation: Enhancing model evaluation*, Feb. 2024. [Online]. Available: https://medium.com/@abhishekjainindore24/understanding-cross-validation-enhancing-model-evaluation-ccad3e19cde0.

[104] S. Jentzsch and N. Hochgeschwender, "A qualitative study of machine learning practices and engineering challenges in earth observation," *it - Information Technology*, vol. 63, no. 4, pp. 235–247, Jul. 2021. DOI: 10.1515/itit-2020-0045.

[105] N. Bika, *How to conduct a structured interview*, Sep. 2023. [Online]. Available: https://resources.workable.com/tutorial/conduct-structured-interview#:~:text=A%20structured%20interview%20is%20a,likelihood%20of%20a%20bad%20hire..

[106] C. Gibbons, *In-depth interviews in qualitative research: Not "just a chat"*, Jun. 2024. [Online]. Available: https://www.quirkos.com/blog/post/in-depth-interviews-in-qualitative-research/.

[107] Jupyter, *Project jupyter*, 2024. [Online]. Available: https://jupyter.org/.

[108] Kaggle, *What is a kaggle?* 2022. [Online]. Available: https://www.kaggle.com/discussions/general/328265#:~:text=A%20subsidiary%20of%20Google%2C%20it,to%20solve%20data%20science%20challenges..

[109] DNB, *Pilaar 2: Governance*, Feb. 2016. [Online]. Available: https://www.dnb.nl/voor-de-sector/open-boek-toezicht/sectoren/verzekeraars/risicomanagement-en-governance-pilaar-2/pilaar-2-governance/.

[110] H. Stijnen, *Risicomanagement*, 2024. [Online]. Available: https://kpmg.com/nl/nl/home/sectoren/verzekeringen/risicomanagement.html.

[111] C. I. o. I. Auditors, *The three lines of defence*, Mar. 2015. [Online]. Available: https://www.iia.org.uk/policy-and-research/position-papers/the-three-lines-of-defence/.

[112] H. Ozinga, *Autoverzekering flink duurder door toename elektrische auto's*, Jul. 2024. [Online]. Available: https://www.nu.nl/economie/6321722/autoverzekering-flink-duurder-door-toename-elektrische-autos.html.

[113] J. Poon, "Penalising unexplainability in neural networks for predicting payments per claim incurred," *Risks*, vol. 7, no. 3, p. 95, Sep. 2019. DOI: 10.3390/risks7030095.

[114] Simplilearn, *What is ordinal data? definition, examples, variables analysis: Simplilearn*, Mar. 2023. [Online]. Available: https://www.simplilearn.com/what-is-ordinal-data-article.

[115] L. B. de Amorim, G. D. Cavalcanti, and R. M. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109 924, Jan. 2023. DOI: 10.1016/j.asoc.2022.109924.

[116] E. LeDell and S. Poirier, "H2o automl: Scalable automatic machine learning," *7th ICML Workshop on Automated Machine Learning*, 2020.

[117] I. Salehin, M. S. Islam, P. Saha, *et al.*, "Automl: A systematic review on automated machine learning with neural architecture search," *Journal of Information and Intelligence*, vol. 2, no. 1, pp. 52–81, Jan. 2024. DOI: 10.1016/j.jiixd.2023.10.002.

[118] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020. DOI: 10.1016/j.neucom.2020.07.061.

[119] S. Chomachar, *What are the limitations of bayesian and convex ...* Feb. 2022. [Online]. Available: https://www.researchgate.net/post/What_are_the_limitations_of_Bayesian_and_Convex_optimization_techniques.

[120] J. Reil and I. D, *Insurance pricing interview individual d*, Jul. 2024.

[121] A. Bryant and K. Charmaz, "The sage handbook of grounded theory," *The sage handbook of grounded theory*, 2007. DOI: 10.4135/9781848607941.

[122] J. Nielsen, J. Lewis, and C. Turner, "Determining usability test sample size," *International Encyclopedia of Ergonomics and Human Factors, Second Edition - 3 Volume Set*, Mar. 2006. DOI: 10.1201/9780849375477.ch597.

[123] R. Alroobaea and P. J. Mayhew, "How many participants are really enough for usability studies?" *2014 Science and Information Conference*, Aug. 2014. DOI: 10.1109/sai.2014.6918171.

[124] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Behavior Research Methods, Instruments, amp; Computers*, vol. 35, no. 3, pp. 379–383, Aug. 2003. DOI: 10.3758/bf03195514.

[125] J. Reil and I. A, *Insurance pricing interview individual a*, Jul. 2024.

[126] J. Reil and I. B, *Insurance pricing interview individual b*, Jul. 2024.

[127] J. Reil and I. C, *Insurance pricing interview individual c*, Jul. 2024.

[128] J. Reil, I. D, and I. E, *Insurance pricing interview individual d and e*, Jul. 2024.

[129] J. Reil and I. E, *Insurance pricing interview individual e*, Aug. 2024.

[130] C. Andrade, "The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives," *Indian Journal of Psychological Medicine*, vol. 41, no. 3, pp. 210–215, May 2019. DOI: 10.4103/ijpsym.ijpsym_193_19.

[131] S. Studer, T. B. Bui, C. Drescher, *et al.*, "Towards crisp-ml(q): A machine learning process model with quality assurance methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, Apr. 2021. DOI: 10.3390/make3020020.

[132] D. Soken, *Clisp-ml(q)ml*, 2022. [Online]. Available: https://speakerdeck.com/isidaitc/clisp-ml-q-wohazimetositamlsisutemufalsepin-zhi-que-bao-niguan-surudiao-cha.

[133] J. Dieber and S. Kirrane, "Why model why? assessing the strengths and limitations of lime," *Institute for Information Systems and New Media*, Nov. 2020.

[134] Achmea, 2024. [Online]. Available: https://www.achmea.nl/reinsurance/achmea-reinsurance/history.

Appendix A

Additional Context Description

A.1 Application of GLMs

The application of GLMs in insurance pricing represents a sophisticated approach to understanding and quantifying risk, ultimately informing more accurate and fair pricing strategies for insurance policies. This process, as outlined by Goldburd et al., highlights the impact of GLMs in the property/casualty insurance sector [32]. In this section, a short case study is provided to encapsulate the application process of insurance pricing using GLMs combined with the frequency-severity model approach discussed in Section 2.2.2.2. Based on the information provided from the paper by David, this section focuses on the application, methodology, and results of using GLMs for calculating auto insurance premiums. The case study by David highlights the transition from traditional linear regression models to GLMs, underscoring the latter's capacity to handle non-linear relationships and non-Gaussian distributions of residuals, which are common in insurance data.

The core of this case study is the application of GLMs to estimate the pure premium, which involves two main components: claim frequency and claim cost. The analysis utilizes a Poisson distribution to model claim frequency and a Gamma distribution for claim cost, reflecting the actual data distribution more accurately than normal distribution assumptions [20]. The methodology section details the models used for the analysis, such as the calculation model for the pure premium. The pure premium is the mathematical expectation of the annual cost of claims declared by the policyholders and is obtained by multiplying the two components, the estimated claim frequency and cost, as visualized in Equation A.1 [20].

$$E[\sum_{i=1}^{N} C_i] = E[Y] \times E[C_i] \tag{A.1}$$

for the claims amount $(C_1, C_2, ...)$ independent of their number (Y).

Equation A.1 is in accordance with the frequency-severity methodology discussed in Section 2.2.2. This method, according to David, is particularly relevant because the risk factors, which influence the two components of the pure premium, are usually different [20]. Essentially, the separate analysis of the two phenomena provides a clearer perspective on how the risk factors are influencing the premium.

Using data from a French auto insurance portfolio containing 50,000 policies registered during the year 2009, the case study presents a numerical illustration that identifies risk

factors and divides the insurance portfolio into tariff classes [20]. It highlights the significant predictors of claim frequency and cost, such as the age of the insured, the type of vehicle, and the bonus-malus coefficient. Through this illustration, visualized in Figure A.1, the study demonstrates the stages of establishing the insurance premium through a Poisson and Gamma model.

Source	Chi-Square*	Pr > ChiSq*	Chi-Square**	Pr > ChiSq**
Age	87.75	<.0001	89.87	<.0001
Occupation	63.76	<.0001	63.71	<.0001
Type	46.02	<.0001	46.22	<.0001
Category	4.13	0.1268	-	-
Power	1.60	0.4499	-	-
Use	83.61	<.0001	83.58	<.0001
Bonus-Malus	451.76	<.0001	451.69	<.0001
Age of insurance contract	35.14	<.0001	35.23	<.0001

(*) Poisson regression including all the explanatory variables
(**) Poisson regression including only the significant explanatory variables

(a) Case study statistics of Poisson model (claim frequency).

Source	Chi-Square*	Pr > ChiSq*	Chi-Square**	Pr > ChiSq**
Age	46.13	<.0001	47.75	<.0001
Occupation	60.50	<.0001	61.00	<.0001
Type	39.81	<.0001	39.57	<.0001
Category	2.62	0.2701	-	-
Power	2.11	0.3491	-	-
Use	0.87	0.3510	-	-
Bonus-Malus	0.97	0.3253	-	-

(*) Gamma regression including all the explanatory variables
(**) Gamma regression including only the significant explanatory variables

(b) Case study statistics of Gamma model (claim severity)

Figure A.1: Overview of both Poisson and Gamma statistics of the case study [20].

In the case study of David, certain variables have been noted to lack statistical significance, which can be derived by looking at the p-values exceeding the threshold of 0.05 [20]. The standard error is a measure of uncertainty of the Poisson and Gamma regression coefficient. Andrade considers why 5% may be set as a reasonable cut-off for statistical significance [130]. Consequently, variables exceeding the threshold of 0.05 are omitted from the model, which retains only those explanatory variables that demonstrate statistical significance.

The explained variable is the product between the estimated frequency and the estimated cost of claims. Equation A.2 shows the calculated value representing the insurance pure premium established for insured i, characterized by the variables vector x_i.

$$E[\sum_{i=1}^{N} C_i] = E[N_i]E[C_{i1}] = exp((\beta_{freq} + \beta_{cost})^t x_i \tag{A.2}$$

In Figure A.2, two different regression analyses are shown for both the Poisson regression (claim frequency) and Gamma regression (claim severity) respectively. The presence of a dash $(-)$ for 'Use (private)', 'Bonus-Malus', and 'Age of insurance contract' in the Gamma regression estimates and their corresponding standard errors mean that these variables were not included in the claim severity model, because they were not significant predictors of claim severity (exceeding p-value threshold).

A negative estimate of a feature means a negative relation to the prediction, which in this case is the pure premium. One would read the Estimate of 'Age' as follows: "Age has a pure premium value of -0.0314 (as $\beta_{freq} + \beta_{cost} = -0.0189 + -0.0124$) meaning that every year of aging, starting from $x = 0$, causes a $e^{-0.0314} \approx 0.97 = 3\%$ reduction in the pure premium compared to the intercept (β_0)" (use Equation A.2). The intercept parameter is the mean of the responses at $x = 0$. In cases where it does not make sense to set all the predictors equal to zero, one should interpret the intercept at some arbitrary value of the predictors, for example at the mean of the data.

Source	Estimate *	Std Error *	Estimate **	Std Error **	Pure premium
Intercept	-2.1541	0.1193	8.4556	0.1076	6.3029
Age	-0.0189	0.0020	-0.0124	0.0018	-0.0314
Occupation (employed)	-0.2611	0.0617	-0.1674	0.0638	-0.4276
Occupation (housewife)	-0.4189	0.0743	0.0238	0.0765	-0.3945
Occupation (retired)	-0.2348	0.1099	0.0231	0.1113	-0.2101
Occupation (self-employed)	0.0489	0.0657	0.2972	0.0693	0.3471
Type (A)	-0.3760	0.0905	-0.4205	0.0932	-0.7952
Type (B)	-0.4431	0.0940	-0.4457	0.0964	-0.8888
Type (C)	-0.3873	0.1003	-0.3000	0.1033	-0.6863
Type (D)	-0.2068	0.0929	-0.1718	0.0958	-0.3780
Type (E)	-0.0601	0.0989	-0.2026	0.1021	-0.2628
Use (private)	0.4282	0.0481	-	-	0.4282
Bonus-Malus	0.0082	0.0004	-	-	0.0082
Age of insurance contract	-0.0286	0.0049	-	-	-0.0286

(*) Poisson regression results

(**) Gamma regression results

Figure A.2: An analysis of parameter estimates of the case study [20].

This case study concludes with reflections on the potential of GLMs in non-life insurance pricing. It acknowledges the precision and flexibility GLMs offer in understanding and pricing risk. The study also acknowledges limitations, such as data specificity and model assumptions, suggesting areas for future research (further discussed in Section 2.3.2). Ultimately, it posits GLMs as a well-functioning tool in the actuarial science domain, capable of handling premium calculation processes to reflect the risk profile of policyholders.

Appendix B

Additional Methodology

B.1 CRISP-DM & CRISP-ML

Additional supplementary information and supporting documents are provided in this appendix section to enhance the comprehensiveness of the CRISP-DM methodology. These materials aim to offer readers further insight into specific details of CRISP-DM.

Table B.1 provides a comprehensive overview of all phases within the CRISP-DM methodology. It also references back to sections of the thesis that handle certain phases of the framework for easy navigation through the report. Every phase has its own outputs, which are provided in the CRISP-DM documentation[1]. Table B.1 describes the main idea, tasks, and outputs of these phases briefly, based on the user guide of CRISP-DM [23]. It also provides an index where every phase of the framework is found within this thesis document.

Table B.1: CRISP-DM process model descriptions [22].

Phase	Short Description	Section # of Thesis
Business Understanding	The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e.g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.	Section 4.2
Data Understanding	Collecting data from data sources, exploring and describing it and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describe the data description task with using statistical analysis and determining attributes and their collations.	Section 4.3
Data Preparation	Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent.	Section 4.4
Modeling	The data modelling phase consists of selecting the modeling technique, building the test case and the model. All data mining techniques can be used. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set.	Section 4.6
Evaluation	In the evaluation phase the results are checked against the defined business objectives. Therefore, the results have to be interpreted and further actions have to be defined. Another point is, that the process should be reviewed in general.	Section 4.7

[1] https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf

Deployment	The deployment phase is described generally in the user guide. It could be a final report or a software component. The user guide describes that the deployment phase consists of planning the deployment, monitoring and maintenance.	Not handled

Building on the CRISP-DM methodology, the CRISP-ML (CRISP for Machine Learning) methodology could also be applied to the ML insurance pricing problem. With each task of the process, this methodology, just like CRISP-DM, proposes quality assurance methodology that is suitable to address challenges in ML development that are identified in the form of risks [131]. CRISP-ML includes an extra monitoring step, which is particularly useful for continuously assessing model performance in real-world settings, identifying potential issues, and making necessary adjustments [132]. This continuous monitoring ensures the model remains effective over time, which is crucial for dynamic environments such as insurance pricing, where insurance prices are re-evaluated every year [120]. An overview of the CRISP-ML methodology is provided in Figure B.1.

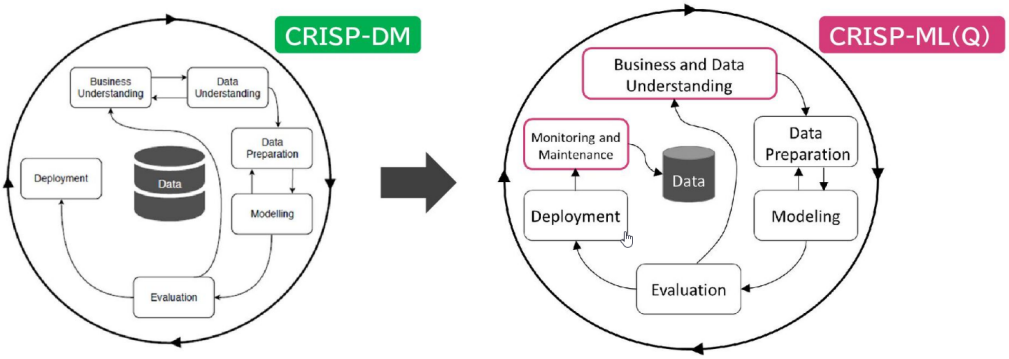


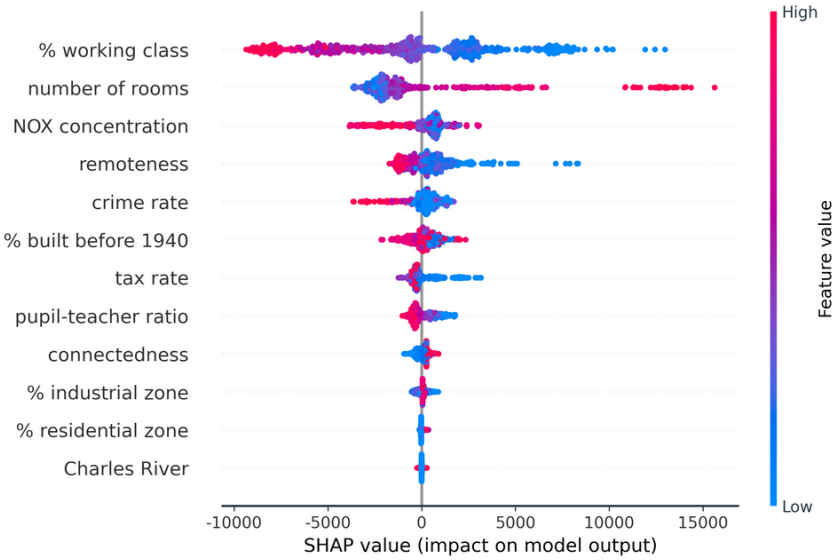
Figure B.1: An overview of the CRISP-DM and CRISP-ML methodology respectively, showing the added *Monitoring and Maintenance* step [132].

B.2 Example SHAP: Global & Local Explanations

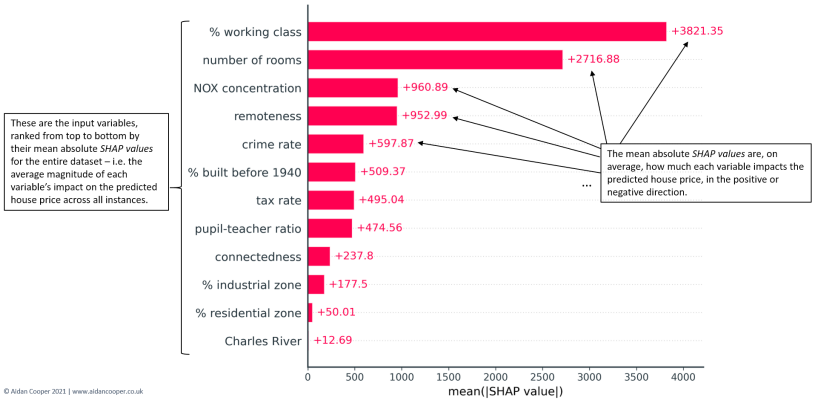
B.2.1 Example Beeswarm & Bar Plots

The functionality of global SHAP explanations is depicted in Figure B.2 through an example for house price predictions based on *The Boston Housing Dataset*[2], where the two figures in question provide a comprehensive visualization of SHAP, showing the impact of various features on the model's predictions [92]. This is done through either a beeswarm plot (shows the distribution of SHAP values for each feature across all instances in the dataset) and a bar plot (summarizes the average impact of each feature on the model's predictions across all instances).

[2] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html?xgtab=&ref=aidancooper.co.uk



(a) Global Shapley explanation through a beeswarm plot.



(b) Global Shapley explanation through a bar plot.

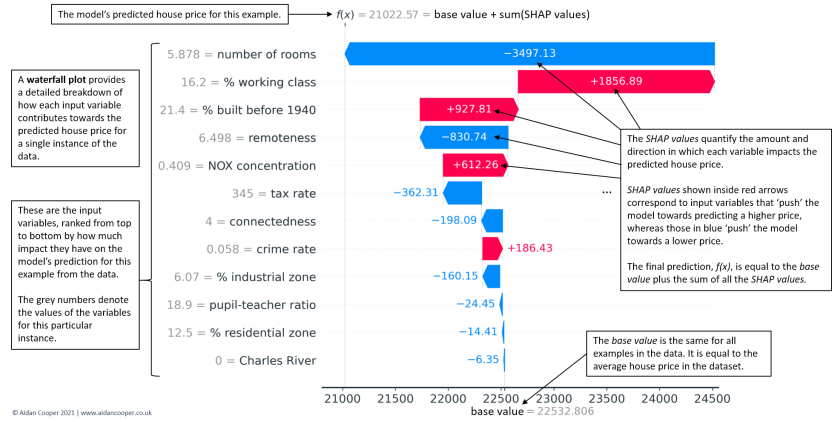
Figure B.2: Example: Global explanation plots with SHAP based on *The Boston Housing Dataset* [92].

The SHAP beeswarm plot, also known as the summary plot, provides a detailed view of how each feature's value affects individual predictions, showcasing the variability and distribution of SHAP values [88]. In the case of Figure B.2a, one sees that lower values of % working class have positive SHAP values (the points extending towards the right are increasingly blue) and higher values of % working class have negative SHAP values (the points extending towards the left are increasingly red). This indicates that houses in more working-class areas have lower predicted prices. The reverse is seen for number of rooms, where higher room counts lead to higher house price predictions [88].

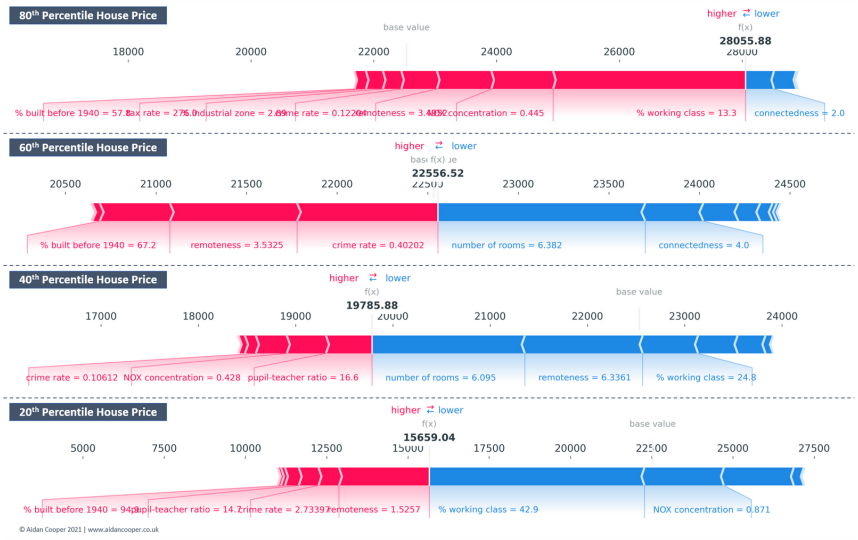
The SHAP bar plot provides an aggregate view, ranking features by their average importance across all predictions, making it easier to identify the most influential features in the model [88]. In the case of Figure B.2b, one can see that % working class is the most influential variable, contributing on average \$3,821 to each predicted house price. By contrast, the least informative variable contributes approximately \$13 [88].

B.2.2 Example Waterfall & Force Plots

The functionality of local Shapley explanations is depicted in Figure B.3 through an example for house price predictions based on *The Boston Housing Dataset* as well. This is done through either a waterfall plot (breaks down the prediction of a model by showing how each feature contributes to the final prediction) and a bar force plot (shows how different features push the prediction from a base value towards the final prediction).



(a) Local Shapley explanation through a waterfall plot.



(b) Local Shapley explanation through force plots.

Figure B.3: Example: Local explanation plots with SHAP based on *The Boston Housing Dataset* [92].

The waterfall plot of Figure B.3a provides a detailed breakdown of how each input variable contributes to the predicted house price for a specific instance, where features pushing the prediction higher are shown in red bars, while the features pulling the prediction lower are shown in blue bars. This can be related to the functioning of LIME, which will be explained in Section 3.3.2.

Force plots are useful for examining explanations for multiple instances of the data at once, as their compact construction allows for outputs to be stacked vertically for ease of comparison. These force plots in Figure B.3b illustrate the contributions of different

features to the predicted house prices for instances at different percentiles (80^{th}, 60^{th}, 40^{th}, and 20^{th}) within the dataset.

B.3 Example LIME: Local Explanation

The functionality of LIME is depicted in Figure B.4 through an example of the *Rain in Australia Dataset*[3], where the model in question is tasked with predicting one of two classes: class 0 or class 1. The respective model predicts class 0 with a probability of 0.79 and class 1 with a probability of 0.21. Therefore, the final prediction for this instance is class 0, as it has a higher probability. The bar charts on the right side illustrate how various features contribute to this prediction. Each feature's contribution is visualized in terms of its support for class 0 (in blue) and class 1 (in orange). It illustrates how LIME is used to assess the significance of different features in a manner that tries to be straightforward and interpretable.

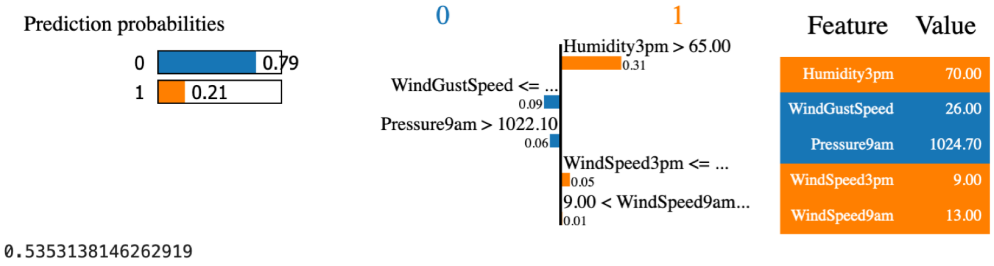


Figure B.4: Example: LIME output of a local observation. For this example, the *Rain in Australia Dataset* from Kaggle was used [133].

One can derive from Figure B.4 that the variables Humidity3pm (representing the humididy at 3PM) and WindGustSpeed contribute positively towards class 1, indicated by the orange bars. This means that higher humidity at 3PM increases the likelihood of the model predicting class 1, as well as higher wind speeds, which might be associated with certain weather events like storms that could influence the model towards predicting these outcomes. On the other hand, Pressure9am (representing the atmospheric pressure at 9AM) and WindSpeed9am (measured wind speed at 9AM) have a negative influence, pushing the prediction towards class 0, as shown by the blue bars. High atmospheric pressure often correlates with stable weather conditions, such as clear skies, which might reduce the likelihood of an event. The same can be said for the wind speed.

[3] https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Appendix C

Additional Experimental Setup

C.1 Framework (Roadmap)

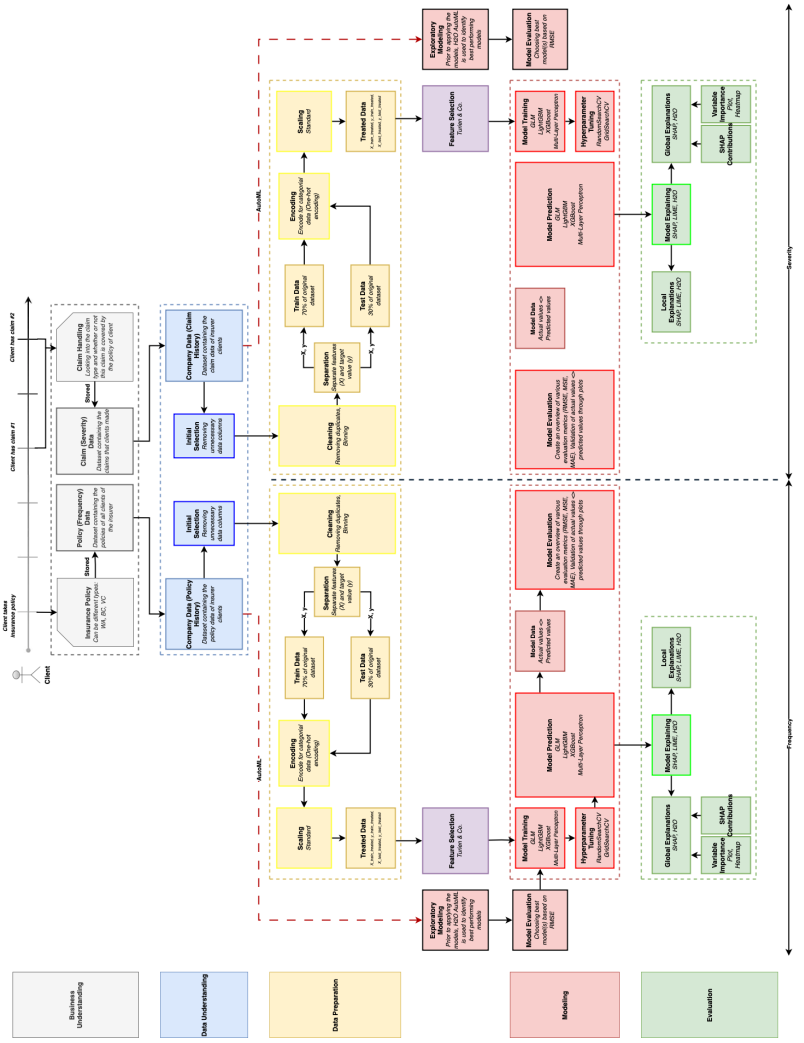


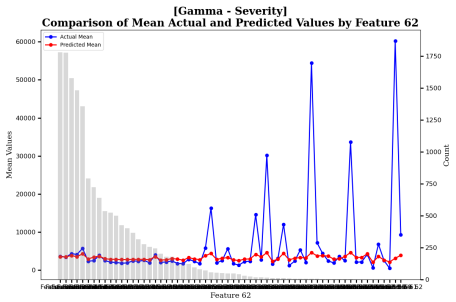
Figure C.1: A roadmap for the project based on CRISP-DM, detailing all the phases conducted within the thesis.

Appendix D

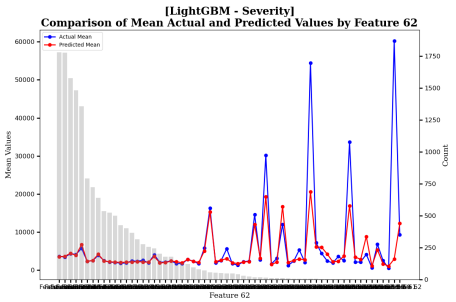
Additional Results & Discussion

D.1 Model Validation

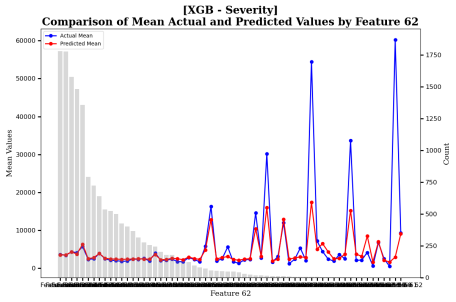
D.1.1 APP Plots Severity



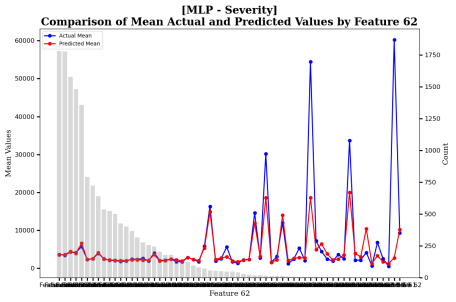
(a) GLM (Gamma).



(b) LightGBM.

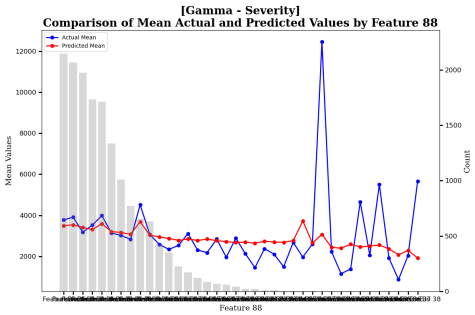


(c) XGBoost.

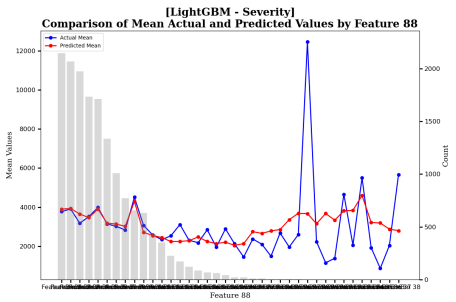


(d) MLP.

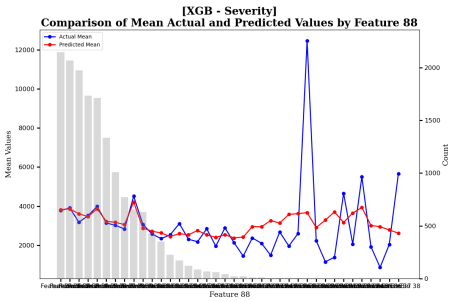
Figure D.1: APP plots for the four models for `Feature 62` for the reparation severity dataset at `the insurance company`.



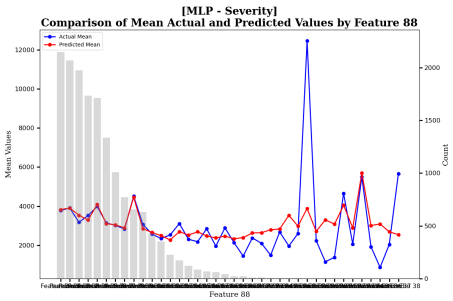
(a) GLM (Gamma).



(b) LightGBM.

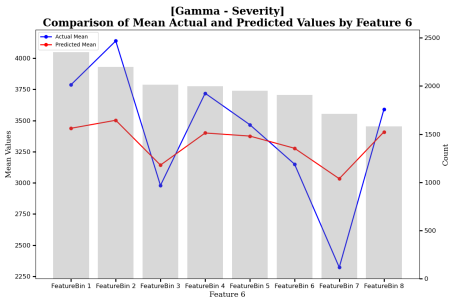


(c) XGBoost.

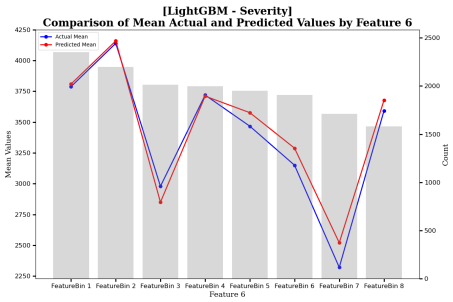


(d) MLP.

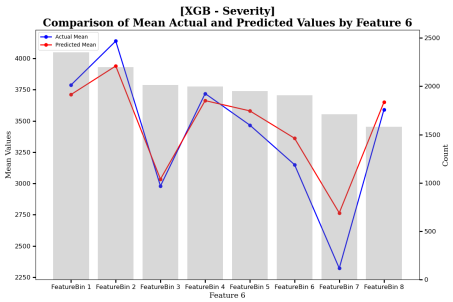
Figure D.2: APP plots for the four models for `Feature 88` for the reparation severity dataset at `the insurance company`.



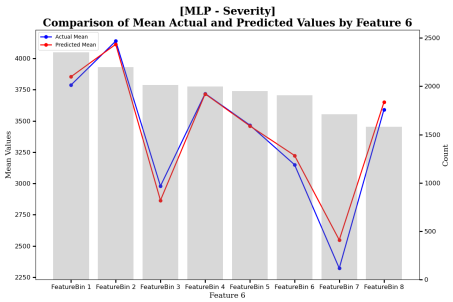
(a) GLM (Gamma).



(b) LightGBM.

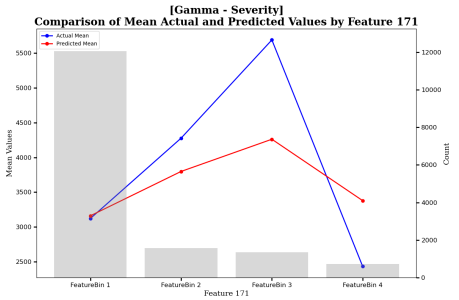


(c) XGBoost.

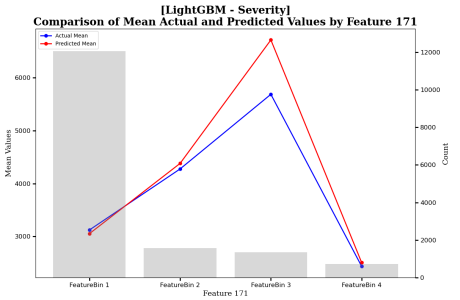


(d) MLP.

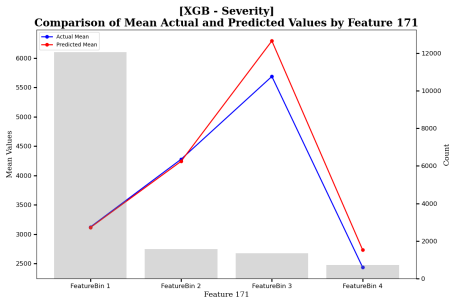
Figure D.3: APP plots for the four models for `Feature 6` for the reparation severity dataset at `the insurance company`.



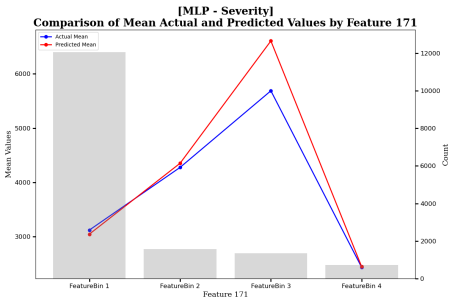
(a) GLM (Gamma).



(b) LightGBM.



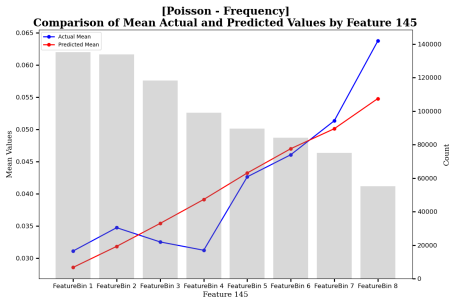
(c) XGBoost.



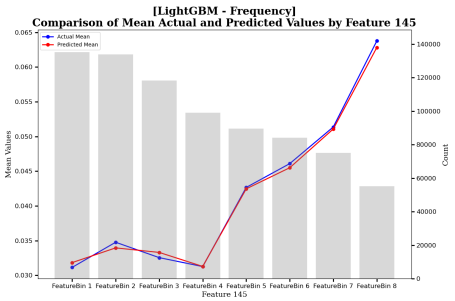
(d) MLP.

Figure D.4: APP plots for the four models for `Feature 171` for the reparation severity dataset at `the insurance company`.

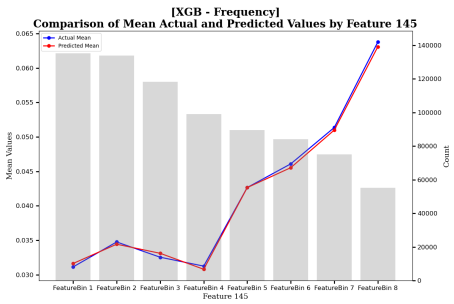
D.1.2 APP Plots Frequency



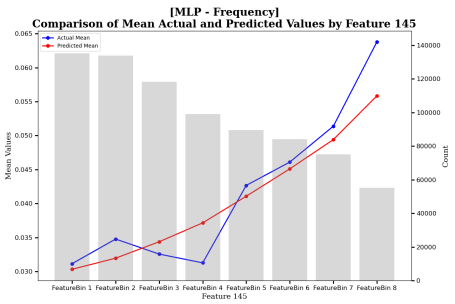
(a) GLM (Gamma).



(b) LightGBM.

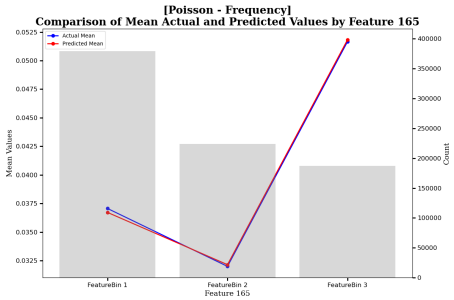


(c) XGBoost.

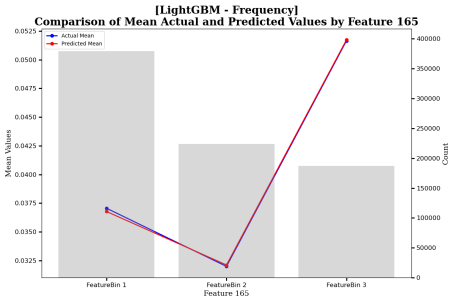


(d) MLP.

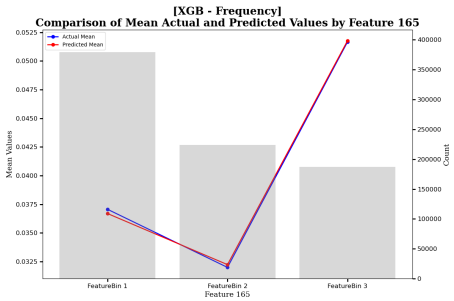
Figure D.5: APP plots for the four models for `Feature 145` for the reparation frequency dataset at `the insurance company`.



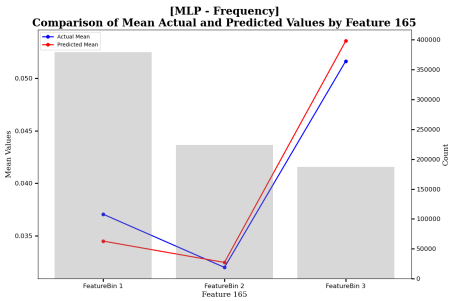
(a) GLM (Gamma).



(b) LightGBM.

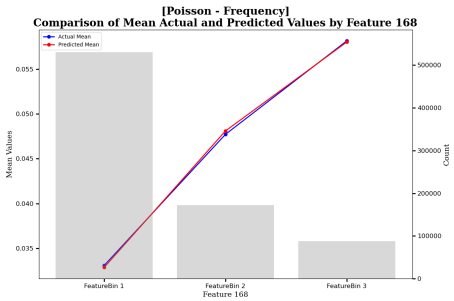


(c) XGBoost.

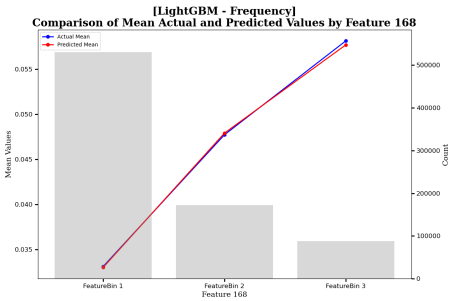


(d) MLP.

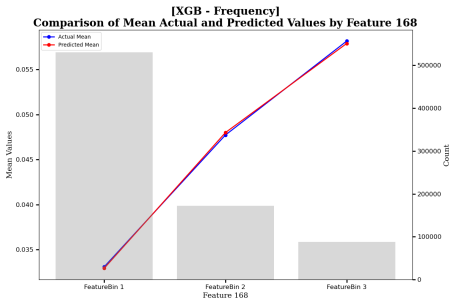
Figure D.6: APP plots for the four models for `Feature 165` for the reparation frequency dataset at `the insurance company`.



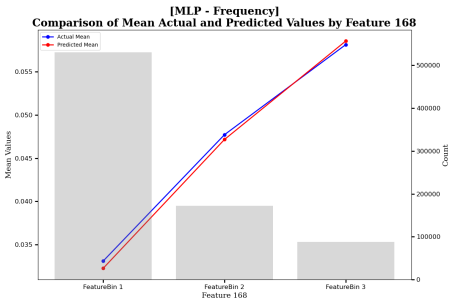
(a) GLM (Gamma).



(b) LightGBM.

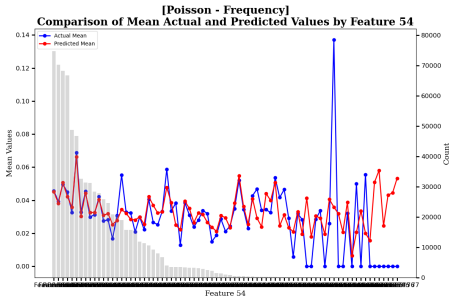


(c) XGBoost.

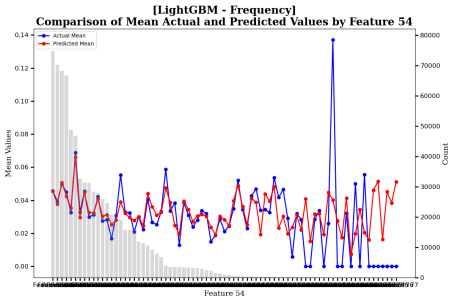


(d) MLP.

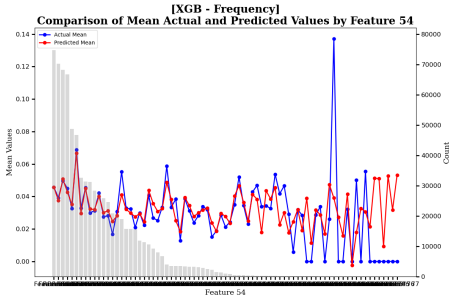
Figure D.7: APP plots for the four models for `Feature 168` for the reparation frequency dataset at `the insurance company`.



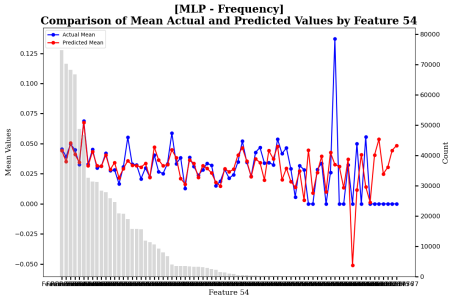
(a) GLM (Gamma).



(b) LightGBM.

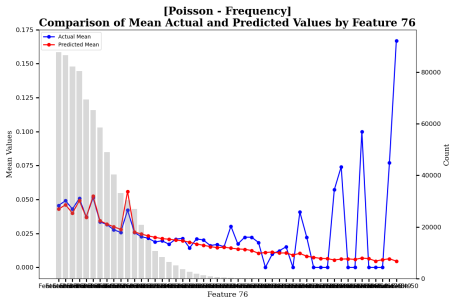


(c) XGBoost.

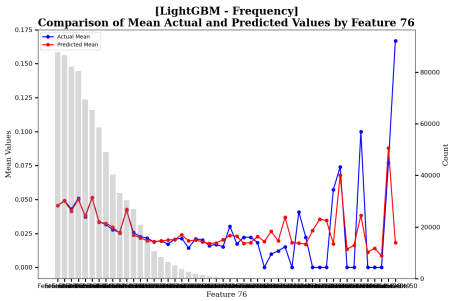


(d) MLP.

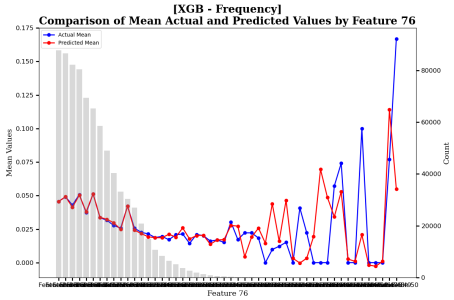
Figure D.8: APP plots for the four models for `Feature 54` for the reparation frequency dataset at `the insurance company`.



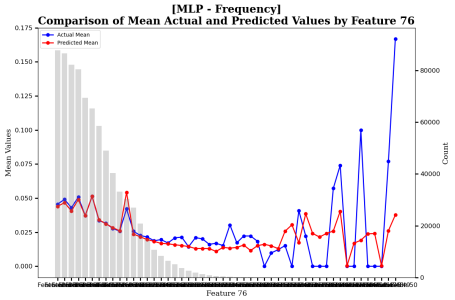
(a) GLM (Gamma).



(b) LightGBM.

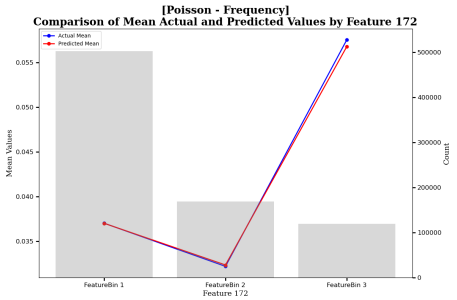


(c) XGBoost.

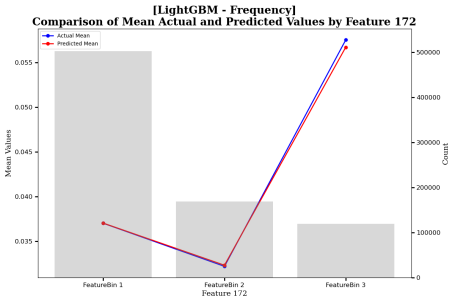


(d) MLP.

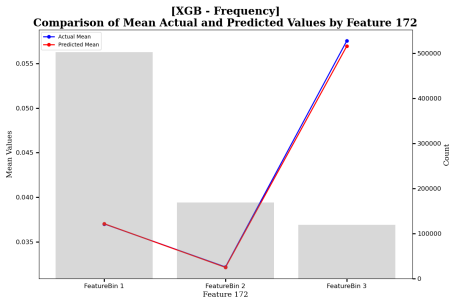
Figure D.9: APP plots for the four models for `Feature 76` for the reparation frequency dataset at `the insurance company`.



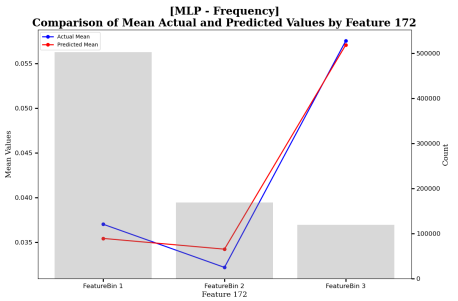
(a) GLM (Gamma).



(b) LightGBM.



(c) XGBoost.



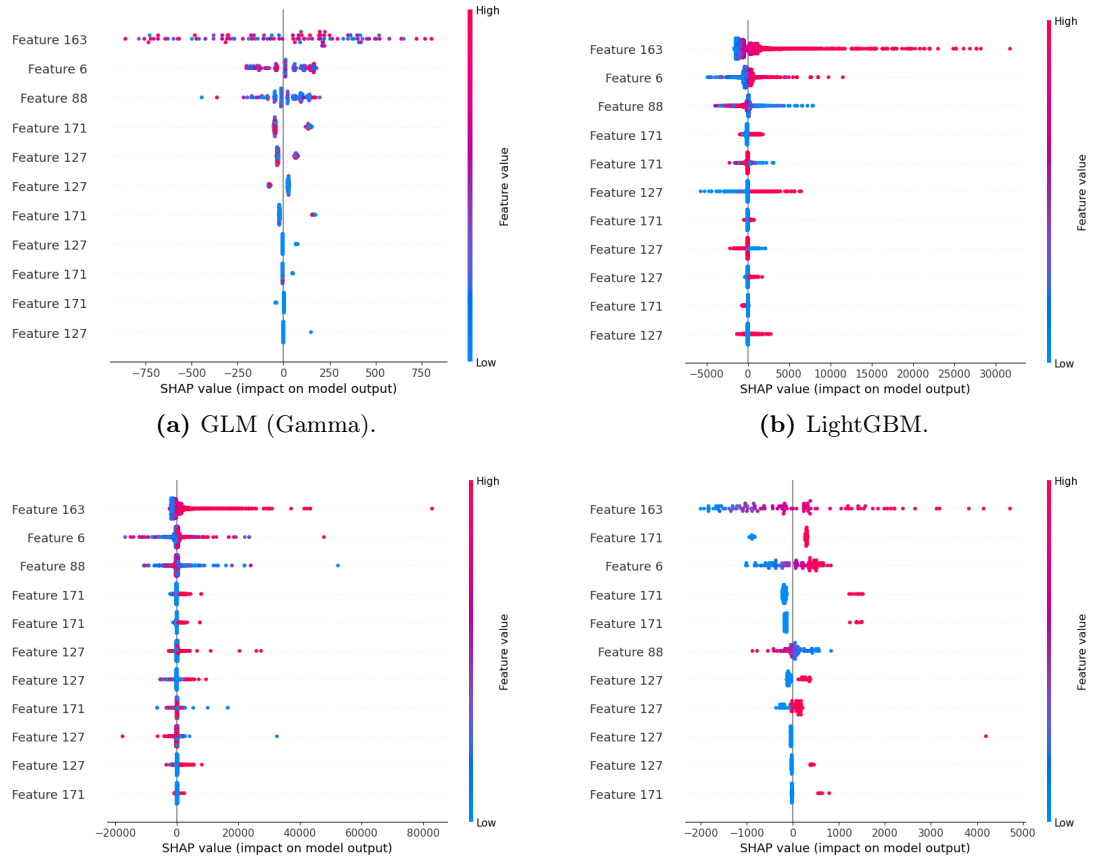
(d) MLP.

Figure D.10: APP plots for the four models for `Feature 172` for the reparation frequency dataset at `the insurance company`.

D.2 Model Evaluation

D.2.1 Severity Evaluation

D.2.1.1 Beeswarm Plots



(a) GLM (Gamma).

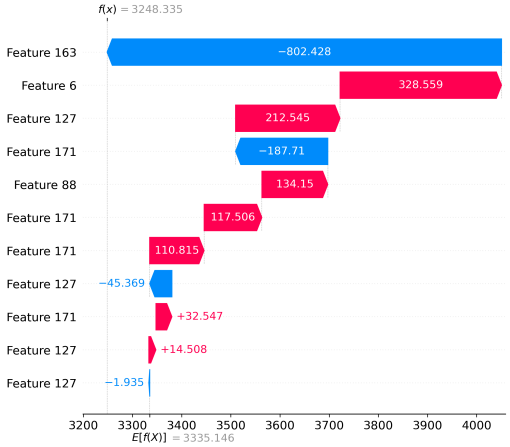
(b) LightGBM.

(c) XGBoost.

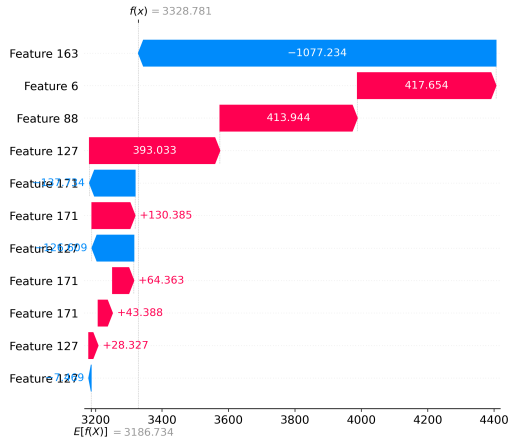
(d) MLP.

Figure D.11: Beeswarm plots for the four models for the reparation severity dataset at `the insurance company`.

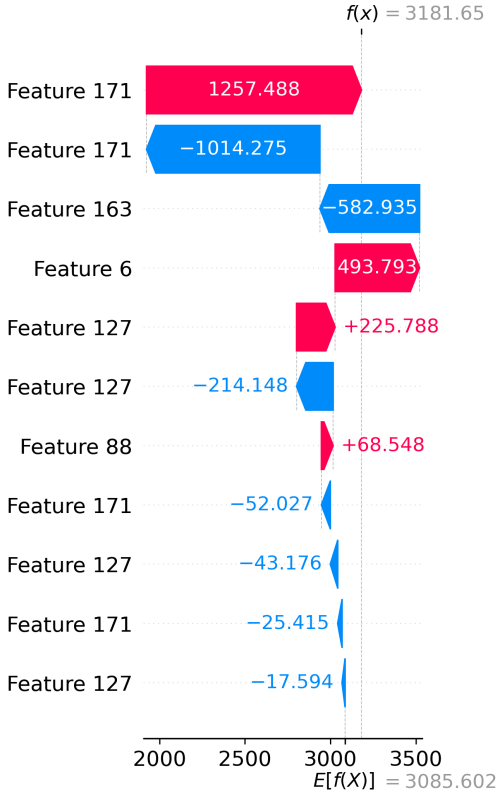
D.2.1.2 Waterfall Plots



(a) LightGBM.



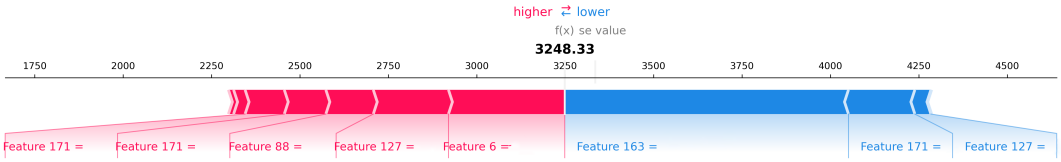
(b) XGBoost.



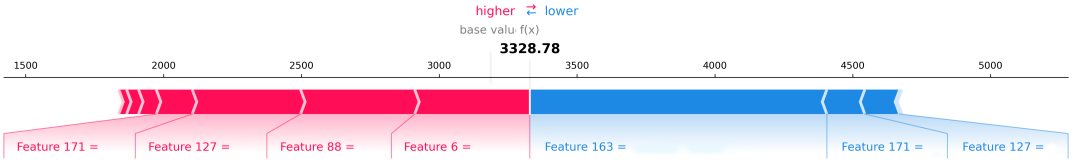
(c) MLP.

Figure D.12: Force plots for the ML models for the reparation severity dataset at `the insurance company`.

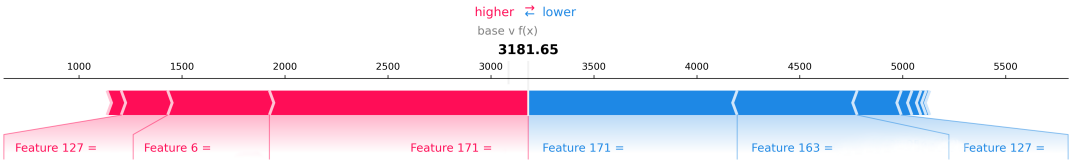
D.2.1.3 Force Plots



(a) LightGBM.



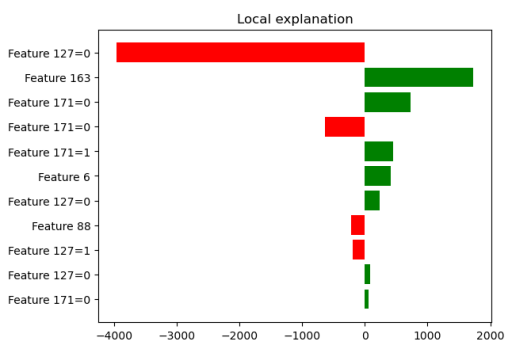
(b) XGBoost.



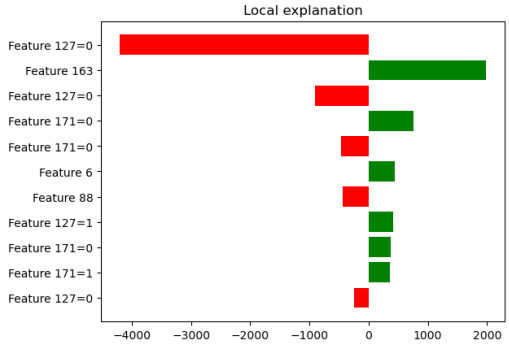
(c) MLP.

Figure D.13: Force plots for the ML models for the reparation severity dataset at `the insurance company`.

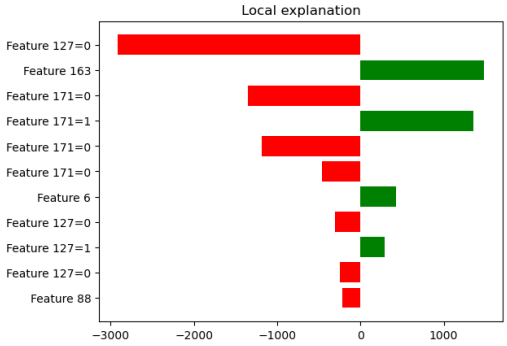
D.2.1.4 LIME Plots



(a) LightGBM.



(b) XGBoost.

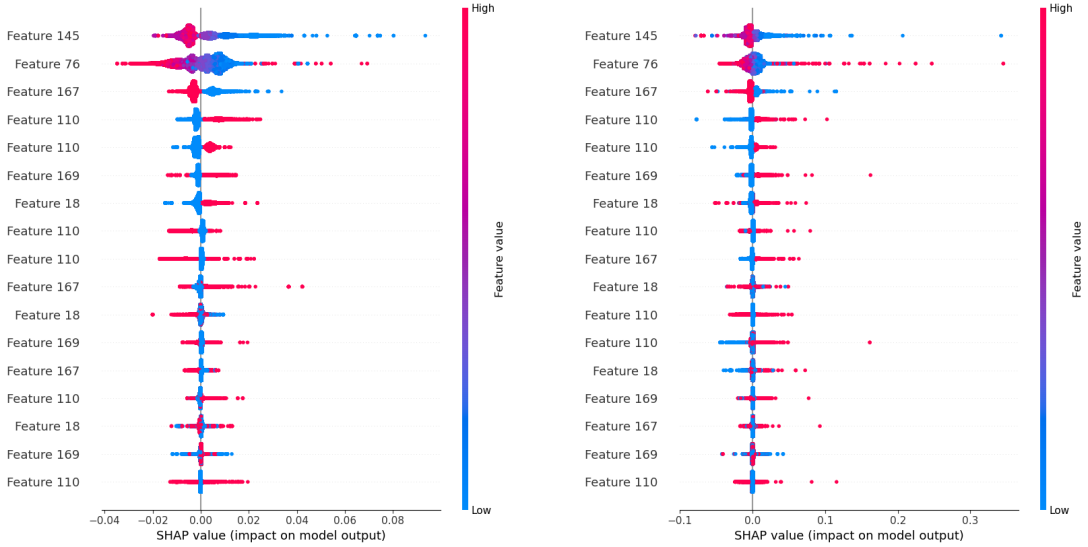


(c) MLP.

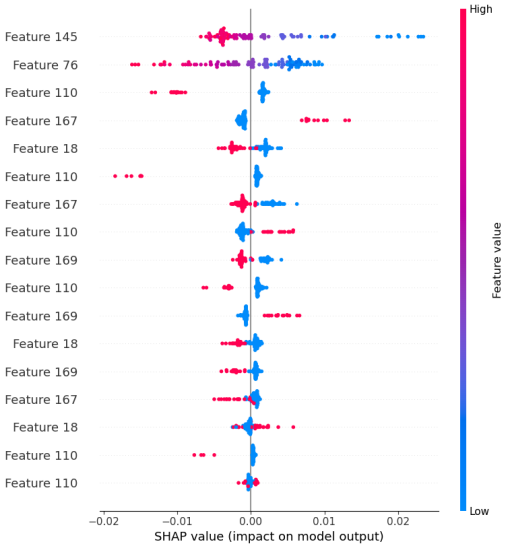
Figure D.14: LIME local explanation plots for the ML models for the reparation severity dataset at `the insurance company`.

D.2.2 Frequency Evaluation

D.2.2.1 Beeswarm Plots



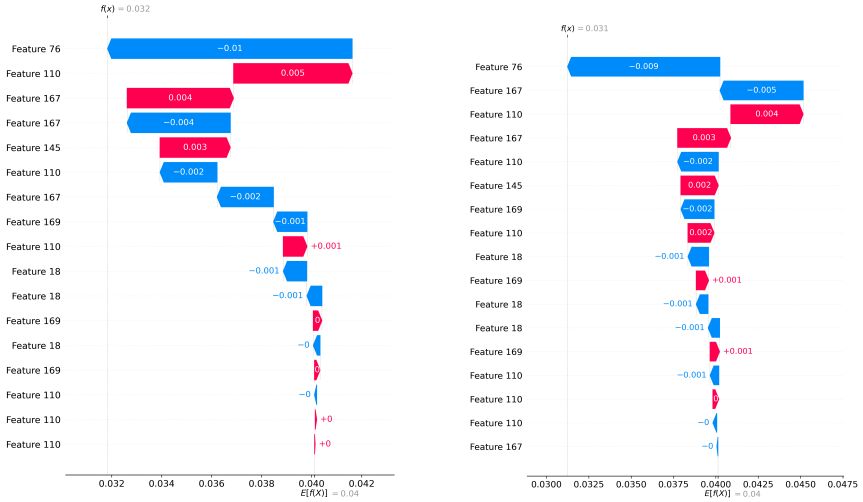
(a) LightGBM. **(b)** XGBoost.



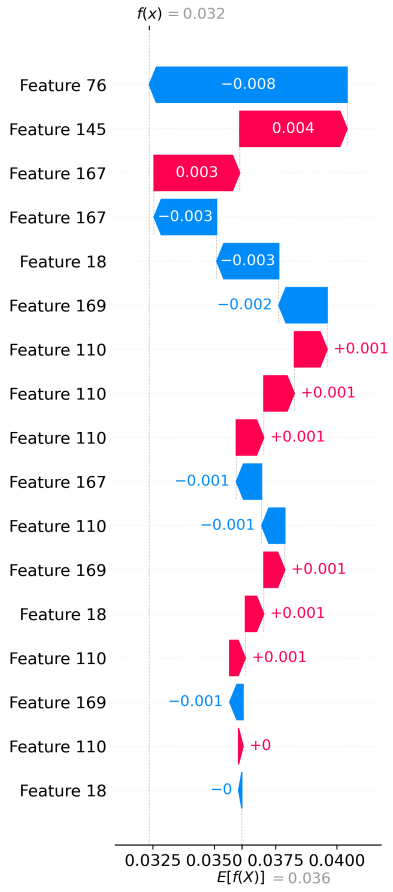
(c) MLP.

Figure D.15: Beeswarm plots for the four models for the reparation frequency dataset at `the insurance company`.

D.2.2.2 Waterfall Plots



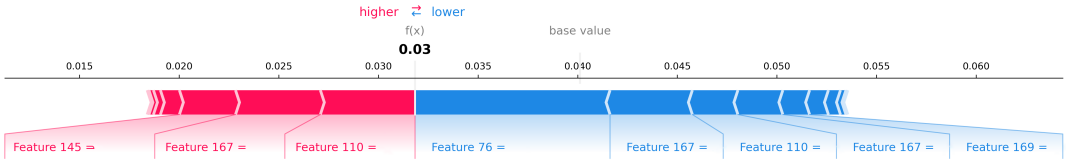
(a) LightGBM. **(b)** XGBoost.



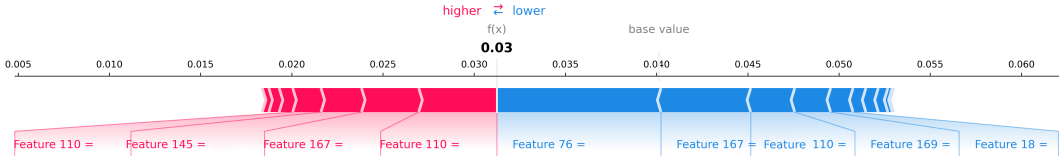
(c) MLP.

Figure D.16: Waterfall plots for the ML models for the reparation frequency dataset at `the insurance company`.

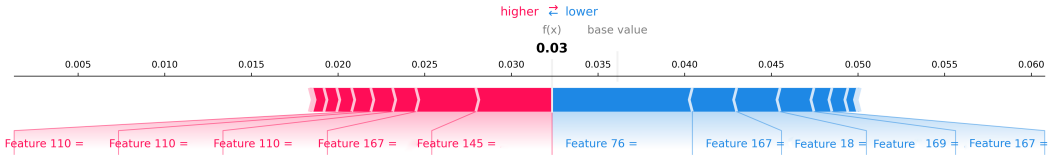
D.2.2.3 Force Plots



(a) LightGBM.



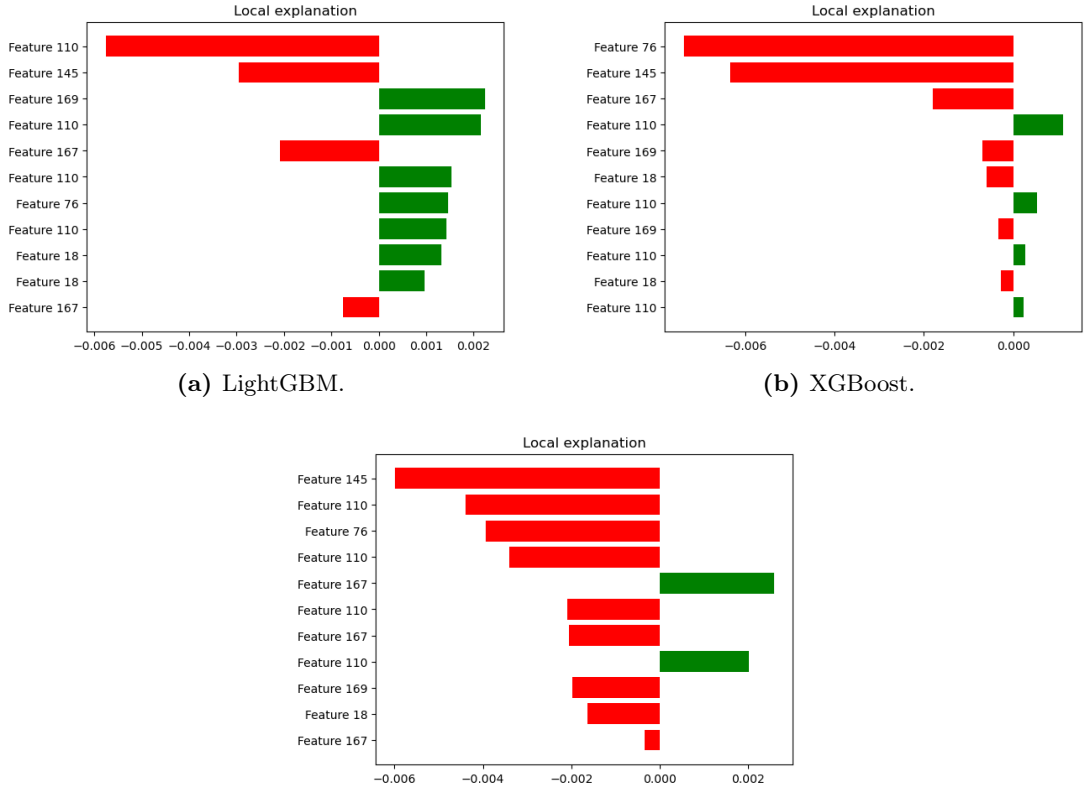
(b) XGBoost.



(c) MLP.

Figure D.17: Force plots for the ML models for the reparation frequency dataset at `the insurance company`.

D.2.2.4 LIME Plots



(a) LightGBM.

(b) XGBoost.

(c) MLP.

Figure D.18: LIME local explanation plots for the ML models for the reparation frequency dataset at `the insurance company`.

Appendix E

Interviews

During this thesis, an exploration is underway to potentially apply ML to risk pricing, a process traditionally handled with GLMs. The objective is to determine whether ML can improve predictive accuracy while preserving the transparency provided by GLMs. `The consultancy firm` is collaborating with `the insurance company` to explore this question. The following sections will cover interviews conducted at both companies, featuring insights from consulting field experts. These interviews will address questions related to GLM and ML models, their explainability and transparency, and the potential transition from GLMs and other traditional statistical methods to ML approaches. Also, an exploration interview in conducted to understand how pricing works within a business context.

E.1 Exploration Interview

This is the first exploratory interview between the author, a scriptant at `the consultancy firm`, and `Individual E`, who works in innovation and product management at `the insurance company`. `Individual D`, a member of the Data Management department of `the insurance company`, also joined the conversation. The aim of this exploratory interview is to examine the following:

1. *Understanding Pure Premium Pricing & Commercial Pricing*
    - ◇ Explore the components of pure premium pricing and commercial pricing.
    - ◇ Examine the factors and aspects considered in these pricing processes.
2. *Application of GLMs in Insurance Pricing*
    - ◇ Analyze the advantages and disadvantages of using GLMs in insurance pricing.
    - ◇ Investigate why GLMs are still preferred over ML techniques in current practices.
3. *Application of ML in Insurance Pricing*
    - ◇ Analyze the advantages and disadvantages of using ML in insurance pricing.
    - ◇ How explainability plays a part in insurance pricing.

During the interview, the components of pure premium pricing and commercial pricing were explored extensively. Pure premium pricing involves calculating the base premium based on risk factors such as the type of coverage, building type, province, and occupational category. This pure premium is crucial as it directly influences the final commercial pricing by incorporating additional factors. In commercial premium pricing, several additional costs are added to the base risk (pure) premium, including commissions, underwriting fees,

and operational expenses. This also includes analyzing loss ratios, which are influenced by factors like the size of the company and the frequency and severity of the registered claims. The importance of understanding the reasons behind high loss ratios was stressed, such as whether claims are due to large individual claims or a higher frequency of smaller claims. This analysis helps in setting appropriate premiums that reflect the actual risk.

`The insurance company` offers a variety of insurance policies, most of which pertain to cars. These car insurance policies are primarily managed using GLMs. For other products, pricing is more frequently based on loss data and different influencing factors. One of the primary benefits of GLMs is their explainability. GLMs allow for a clear understanding of how each factor affects the outcome, which is crucial for explaining model decisions to stakeholders, including regulatory bodies. It is also mentioned by `Individual D` that it is important that the gross of data is predicted correctly, rather than shifting too much modeling fitting towards the few outliers that exist within a claim portfolio.

`Individual E` explained that for auto insurance, `the insurance company` uses GLMs to evaluate numerous variables, such as the car's brand, the driver's age, and regional risk factors. They assess deviations in these variables to determine their impact on the overall risk. For instance, they have identified that certain high-value cars do not present twice the risk of cars valued at two million euros. This explanatory insight helps in setting premiums that are fair and competitive, avoiding disproportionate increases for higher-valued vehicles.

GLMs can precisely identify and explain the contributing factors. The ability to explain model predictions ensures that the models comply with legal and ethical standards. `The consultancy firm`, an external auditing body, requires models to be explainable to ensure they meet these standards. Model validation at `the insurance company` involves professional judgment, monitoring loss experiences, and examining specific characteristics of insured items. The validation process relies heavily on practical knowledge and market experience rather than formal backtesting. GLMs allow for straightforward validation of the significance and relevance of each factor included in the model.

A significant part of the discussion revolves around the application of ML models, balancing predictive accuracy and explainability. The benefits of ML, such as the possibility of having lower prediction errors, are acknowledged. However, there is a concern about the black-box nature of these models and its variability, which could hinder clear explanations to clients and advisors. Customers often prefer stable premiums over potentially lower, variable premiums due to predictability in expenses, similar to fixed mortgage rates. Stability is especially important in high-value items like car insurance, where customers value the assurance that their premiums won't fluctuate significantly year-to-year. Price comparison tools like Independer[1] play a significant role in how customers choose their insurance. When faced with slightly varying premiums, many customers prefer the stability of a higher, but more predictable premium. This lack of explainability is a significant drawback in insurance pricing with ML, where understanding the rationale behind premium calculations can be essential.

`The insurance company` emphasizes the importance of explaining premium calculations to customers, especially when premiums increase significantly. Tools and dashboards can help illustrate the factors contributing to premium changes. However, complex models like those generated by ML pose challenges in explainability compared to GLMs. These models allow insurers to explain price changes based on specific, understandable factors. However, there is interest in exploring ML for its potential to improve pricing accuracy.

[1] Independer.nl N.V. (operating under the name Independer) is a Dutch comparison site by DPG Media for financial products, healthcare institutions, and energy.

While `the insurance company` is open to innovation, there is also a cautious approach to implementation of ML in insurance pricing, considering customer and advisor expectations. Transitioning to ML models involves significant initial costs and complexities. These include training the models, validating their results, and ensuring they are robust enough for real-world application. Additionally, models must be reproducible and explainable to satisfy both regulatory bodies and customers. `The insurance company` highlights specific cases like Porsche insurance, where the premium might increase significantly due to higher risks associated with certain car models. These decisions need to be explained clearly to customers to maintain trust. The idea of a pilot project for a new product using ML is proposed, allowing for monitoring results and integrating lessons learned before broader implementation. Both parties agree on the necessity of innovation in risk pricing. However, the implementation must be careful and transparent to keep clients and advisors well-informed and engaged. A pilot project for a new product seems the best way to test ML techniques and evaluate their effectiveness, providing a pathway to potentially broader application if successful.

E.2 Validation Interviews

This section handles the various semi-structured interviews[2] conducted with respective stakeholders to identify possibilties of integrating ML while trying to preserve explainability and transparancy of these models. Here, all the interviews are summarized. The interview was divided into the following main topics, as discussed in Section 4.7.2.1, inspired by the qualitative research study about ML by Jentzsch & Hochgeschwender [104]:

1. *Technical Background.* Basic information as education, previous work experience, and the current job description of the interviewee are explored in this part of the interview. Understanding the interviewee's technical background provides valuable context for their responses and insights. It allows to assess the level of expertise and experience they bring to the discussion, which is crucial for evaluating their perspective. Questions involve:
   - ◇ Could you share details about your educational background?
   - ◇ What professional experience do you have?
   - ◇ Can you describe your role and responsibilities within the team?
2. *Current Practices.* The current practices and methodologies employed by the specific department in the insurance company's pricing process are explored in this phase of the interview. This includes the tools, models, and strategies they currently use. Understanding the current practices within an insurance company is essential to evaluate the potential impact and feasibility of integrating ML algorithms.
   - ◇ How many years of experience do you have working with GLMs?
   - ◇ Could you walk me through the entire model development process?
3. *Model Validation.* This process is crucial for ensuring the accuracy and reliability of predictive models. Model validation involves a variety of techniques and metrics to assess how well a model performs on unseen data. Each method provides different insights into the model's performance and error distribution. Evaluating the performance of ML models compared to traditional methods such as GLMs is essential for understanding the potential benefits and drawbacks of each approach. By comparing these models, one can determine whether the results are compelling enough to make

[2]For a detailed understanding of semi-structured interviews, please refer to Section 3.4.3.

a transition to or inclusion of ML techniques in the modeling process. Proper validation ensures that models are robust, generalizable, and capable of making accurate predictions in real-world scenarios.

- ◇ What methods do you use to validate the current models (e.g., MAE, RMSE, MSE, others)?
- ◇ How do you interpret the results of your validation metrics in a business context?
- ◇ How do you evaluate the performance of ML compared to traditional methods such as GLMs?
- ◇ Are the ML results compelling enough to consider using this technique?
- ◇ What can you say about the APP plots for ML compared to GLMs? Are they comparable?
- ◇ In your opinion, what could be a potential drawback of applying ML techniques?

4. *Importance of Model Explainability.* This topic seeks to uncover how crucial it is for the team to be able to explain and understand the models they use. It explores the role of explainability in building trust and confidence in the models, how it influences the adoption of new methodologies, and any specific instances where the ability to thoroughly explain a model's output was particularly beneficial.
   - ◇ Can you provide examples where the explainability of GLMs was advantageous? What did you need at that time?
   - ◇ How do you view the importance of explainability within the insurance sector? Is there room to reduce this for better pure premium pricing precision?
5. *XAI Outcome Analysis.* Focusing on evaluating the glocbal and local outcomes provided by XAI techniques used. The aim is to gather detailed feedback so these explanations can be compared to the traditional factor-based explanations provided by GLMs, assessing the usefulness of these explanations for ML techniques. The goal is to understand how well these XAI techniques meet the department's needs for transparency and whether they facilitate a better understanding of ML models.
   - ◇ (Global) How can the global XAI outcomes be compared to traditional factor-based explanations of GLMs?
   - ◇ (Global) Do the outcomes align with your experience and expertise?
   - ◇ (Global) What is your perspective on tools like SHAP, and do you trust the results they generate? Can you use them to achieve what you would typically do with GLMs to some extent?
   - ◇ (Local) How can the local XAI outcomes be compared to traditional factor-based explanations of GLMs?
   - ◇ (Local) Do the outcomes align with your experience and expertise?
6. *Comparison GLMs and ML Explainability.* Centers around identifying the strengths and weaknesses of each approach and understanding the interviewee's perspective on the feasibility of adopting ML models in light of their explainability.
   - ◇ Do you see specific advantages or disadvantages in using SHAP and LIME compared to GLM explanations?
   - ◇ Based on the evaluation, do you see a future where ML models could replace GLMs in your pricing process?
   - ◇ What additional information or results would you need to make an informed decision about the adoption of ML models?
7. *Views on AI Adoption.* Looking into the overall potential and readiness for AI adoption within the current infrastructure and workflow.

- ◇ What potential challenges do you foresee in integrating ML algorithms into the current pricing process?
- ◇ Are there infrastructural or resource limitations that could impact the adoption of ML models?
- ◇ How well do you think your team is prepared to transition from GLM to ML models in terms of skills and knowledge?
- ◇ What kind of training or resources would be necessary to facilitate this transition?

8. *Feedback and Suggestions.* This ensures that no steps are overlooked in the process and that any critical points vital to the project's outcome are included.
   - ◇ Do you have any feedback on the ML models and their explanation that can help improve their integration and acceptance?
   - ◇ Are there any other aspects or factors that you think should be considered when evaluating the transition from GLMs to ML models?

E.2.1 Interview #1: Individual A

`Individual A` holds a Bachelor's and Master's degree in Econometrics from Vrije Universiteit (VU) Amsterdam, specializing in Data Science. This academic background made him pursue a professional journey into the topics of data analytics and predictive modeling. With over 4.5 years of experience as a Data Analytics consultant at `the consultancy firm`, `Individual A` has worked on a variety of projects involving dashboarding and predictive modeling, particularly for insurance products. Their role has primarily focused on developing ML models for premium pricing, although `Individual A` acknowledges that GLMs remain the industry standard today.

The model development process at an insurer made was validated, looking into a general insurance company involving several stakeholders, including a commercial risk committee that evaluates the models for sufficient risk margins, external actuaries who validate the models against regulatory standards, and a product management team that deploys the models. This process ensures that the models are robust, compliant, and commercially viable. In some larger organizations, according to `Individual A`, this is done by internal actuaries. They develop the models, which are then reviewed by commercial departments. Therefore, the model could use some level of revision to make it more generalized.

To validate the models, various metrics such as MAE, RMSE, MSE, and tuning time are employed. MAE was highlighted by Individual D as particularly important because it focuses on the bulk of the data around the mean, reducing the impact of outliers. `Individual A` agrees that if this emphasis aligns with some insurers' priorities, which value the accuracy of the majority of predictions over the extremes. `Individual A` interprets validation results by comparing predicted values against actual values, focusing on the mean and distribution across different segments. They note that ML models, such as neural networks and tree-based models, show varying performance, with neural networks excelling in certain areas despite longer tuning times. The performance of ML models is evaluated through metrics and plots that compare actual versus predicted values, called APP plots. ML models tend to show more accurate predictions for segments with larger variances compared to GLMs. This detailed analysis helps in understanding the strengths and weaknesses of different models. When comparing APP plots, `Individual A` finds that ML models generally offer more precise predictions, especially in categories with significant variances. This increased accuracy is crucial for fine-tuning insurance premiums. Potential drawbacks of ML include the longer tuning times required for some models and the need for

extensive validation to ensure accuracy. These factors must be weighed against the benefits when considering the adoption of ML models. Explainability is crucial for ensuring models are free from errors and for validating the importance of features. Traditional GLMs allow for easy back-calculation of factors, which some actuaries prefer. In the insurance sector, explainability helps in validating models internally and can be vital for customer-facing applications like fraud detection. However, for premium determination, `Individual A` believes that current explainability tools for ML, such as SHAP, provide sufficient insights.

The usage of SHAP and LIME values to explain feature importance in ML models is also discussed. The interview finds that these outcomes align well with traditional factor-based explanations from GLMs, offering a robust and deterministic method for model validation. Local XAI outcomes, which provide insights on individual predictions, are particularly useful for explaining differences in premiums for similar cases. SHAP is preferred for its robustness and deterministic nature, as LIME's variability makes it less reliable. `Individual A` believes that the visual and interpretative clarity provided by SHAP and other XAI tools can match the explanatory power of GLMs, making them viable for use in insurance pricing.

Looking ahead, `Individual A` sees the potential for ML models to replace GLMs in the pricing process, provided that explainability and validation concerns are adequately addressed. He emphasizes the importance of continued validation and stakeholder acceptance for a successful transition. To make an informed decision about adopting ML models, additional comprehensive validation results and detailed insights into feature impacts are necessary. This information helps build confidence in the accuracy and reliability of ML models. Integrating ML algorithms into existing pricing processes poses several challenges, including meeting regulatory standards and ensuring model transparency. Infrastructural and resource limitations, such as the need for specialized knowledge and computational power, can also impact adoption. `Individual A` believes that teams need to be adequately trained in ML techniques and tools like SHAP to effectively transition from GLMs to ML models. Ongoing training and access to necessary resources are crucial for facilitating this transition. For better integration of ML models, `Individual A` suggests enhancing explainability tools and streamlining validation processes. Developing a more generic model to accommodate different types of insurers, from small to large companies, is also recommended ensuring wider applicability and acceptance. Overall, `Individual A`'s insights highlight the potential for ML models to revolutionize insurance pricing, provided that challenges around explainability and validation are addressed adequately.

E.2.2 Interview #2: Individual B

`Individual B`, currently working as a principal in non-life insurance, has a robust background in econometrics and actuarial studies from the University of Groningen. With over ten years of experience in the industry, `Individual B` has spent significant time working with major insurers like Achmea[3] and smaller insurers, providing him with a comprehensive understanding of the practical applications and challenges in the field of insurance. Throughout his career, `Individual B` has been involved in various facets of non-life insurance, including pricing. His experience has equipped him with a elaborate understanding of both traditional actuarial models and modern ML techniques. In his current role, `Individual B` focuses on model validation and advice, ensuring that the models used within his organization are robust, compliant, and commercially viable. He empha-

[3] Achmea is a large insurance company that provides a wide range of financial services and products. The company serves about 10 million people with Health, Life and Non-life insurances [134].

sizes the importance of understanding the practical implications of model outputs and the necessity of aligning them with regulatory standards and business objectives.

`Individual B` acknowledges the prevalence of GLMs in the industry, noting that while ML models offer promising results, GLMs remain the standard due to their explainability and regulatory acceptance. He acknowledges the model development process in various organizations (`the consultancy firm` involves multiple stakeholders, which in case of `the insurance company` a commercial risk committee and external actuaries, ensuring that models are thoroughly vetted and validated before deployment.

The organizational structure and stakeholder involvement was also discussed, comparing it to various companies such as Achmea and other insurance companies `Individual B` worked at. `Individual B` noted that Achmea operates with a dual approach: one following some more traditional methods such as fixed formulas and another incorporating more advanced (GLM) models. This dual approach allows Achmea to benefit from both worlds, with traditional methods for regulatory compliance and newer models for internal analysis and customer-specific adjustments. `Individual B` highlights the importance of focusing on validation metrics, aligning with insurers' priorities for accurate predictions around the mean. When interpreting these validation results, Jens showed comparisons between the predicted values against actual values, focusing on the mean and distribution across different segments. From the findings presented, `Individual B` concludes that ML models offer more precise predictions. This increased accuracy could be crucial for fine-tuning insurance premiums. However, he also acknowledges potential drawbacks of ML, including longer tuning times and the need for extensive validation. Explainability is the key concern for `Individual B`, particularly in the context of risk management. He believes that understanding model outputs and their implications is crucial for making informed decisions. Traditional GLMs offer ease of back-calculating factors, which some actuaries prefer. The author used the retrieved SHAP values to explain feature importance in ML models. `Individual B` confirms that these outcomes align well with traditional factor-based explanations from GLMs, although GLMs provide more detailed explanations. Additionally, it is challenging to identify relationships between variables using SHAP.

Looking ahead, `Individual B` sees the potential for ML models to replace GLMs in the pricing process, provided that explainability and validation concerns are adequately addressed. He emphasizes the importance of continued validation and stakeholder acceptance for a successful transition. To make an informed decision about adopting ML models, additional comprehensive validation results and detailed insights into feature impacts are necessary. Integrating ML algorithms into existing pricing processes poses several challenges, including meeting regulatory standards and ensuring model transparency. `Individual B` highlights infrastructural and resource limitations, such as the need for specialized knowledge and computational power, as potential barriers to adoption. At a lot of insurers, this is deemed insufficient for now. For better integration of ML models, `Individual B` suggests enhancing explainability tools and streamlining validation processes. He also recommends developing a more generic model to accommodate different types of insurers, ensuring wider applicability and acceptance.

E.2.3 Interview #3: Individual C

The interview between `Individual C` and the author delves into the insurance pricing models, focusing on the comparison between GLMs and ML, and the practical challenges and potential benefits of adopting new technologies in the industry. `Individual C` studied at the University of Amsterdam (UvA) and has been with `the consultancy firm` for about twelve years. His work primarily involves Solvency II, balance and capital management,

and actuarial functions.

The conversation started discussing the organisational structure of the actuarial modeling process at `the insurance company`, where `Individual C` explains that the Data Management team creates models as the first line of defence. The team combines inputs from various departments, including the commercial pricing department, to develop the pricing models used by the company. These models are then reviewed by the actuarial function, which he is a part of, representing the second line. The role of the second line is to review and audit the models created by the first line whether these meet regulatory and internal standards, ensuring an unbiased review process. The third line of defence is typically the internal and external audit functions. In the case of `the insurance company`, the internal actuarial team performs annual norm validation checks on the models used. However, this internal team is limited in size and scope. The third line focuses on compliance and high-level review to ensure that the first and second lines are functioning correctly.

The interview continued by delving into the comparison between GLM and ML models. `Individual C` emphasizes that while GLM models are currently more explainable and easier to validate, ML models, despite their potential for higher accuracy, present significant challenges due to their "black box" nature. This lack of transparency and explainability makes it difficult for insurers to fully trust and adopt ML models without robust validation methods in place. The validation process for GLM models at `the insurance company` involves deep reviews of code and data, a thoroughness that `Individual C` notes might be specific to `the insurance company` and not as common in other insurance companies. The importance of feature importance plots and XAI methods in making ML models more explainable is also discussed, where `Individual C` likes the factor-based GLMs due to its thoroughness. However, he acknowledges that while these methods can partially explain ML models, they do not offer an in-depth analysis of the relationships between individual variables. He also addresses the difficulties in adopting ML models due to these transparency issues.

The interview continues discussing the potential issues of data limitations and the risk of mispricing in segments with insufficient data. `Individual C` suggests that ML could be used alongside GLMs rather than replacing them entirely. This hybrid approach could combine the strengths of both models, leveraging ML's accuracy while maintaining the transparency and explainability of GLMs. `Individual C` acknowledges that while ML can offer potential in more accurate pricing algorithms, the transparency and ease of validation inherent in GLM models currently make them more suitable for insurance pricing. He points out that some insurers, especially smaller ones, are still not even using GLMs, indicating a significant variance in the adoption of pricing models across the industry.

Afterwards, a discussion started talking about the possibility of using ML models to enhance the feature selection process for GLMs. This would allow insurers to benefit from the strengths of both approaches, facilitating a gradual integration of ML into existing pricing processes. He emphasizes the need for additional information and results to make ML models adoptable and suggests involving ML experts for proper validation, as this is not `Individual C`'s primary field of work.

Towards the end of the interview, the conversation touches on the future adoption of ML in insurance. `Individual C` notes that while ML could become more prevalent, it would face diminishing returns after a certain point compared to the significant leap from judgment-based methods to GLMs. He discusses the competitive aspect, where insurers with better models can achieve better pricing, but this also raises the risk of anti-selection, where less advanced insurers might end up with worse risks. Advancements in other markets were discussed, particularly in the UK, where dynamic pricing and ML adoption might

be more advanced due to different market dynamics and competitive pressures. This global perspective provides additional context to the ongoing discussion about model adoption and implementation.

`Individual C`'s interview underscores the current state and future potential of ML in insurance pricing. While ML offers improved accuracy, its lack of explainability, even with XAI techniques, and the robust validation processes required pose significant challenges. The discussion highlights and brings up the possibility of combining ML with GLMs to leverage both their respective strengths. The necessity for a gradual, well-validated adoption of new models in the insurance industry is also discussed.

E.2.4 Interview #4: Individual D

`Individual D` from the Data Management department at `the insurance company`, has a background in business economics from the Vrije Universiteit (VU) in Amsterdam, where he also completed a master's degree. He began his career at `the insurance company` as an information analyst, focusing initially on software and functional descriptions. This role soon expanded into financial reporting for healthcare institutions, necessitating the creation of detailed and accurate financial reports. Following this, `Individual D` advanced to become the team leader of the Business Intelligence department. This team was responsible for significant developments, including the establishment of the current data warehouse. Over time, this team evolved into the Data Management department, where its current size of total employees shows the growing importance of data at insurance companies. `Individual D` is deeply involved in product development, portfolio management, and pricing strategies, focusing on creating and validating pricing models that align with the company's financial goals and regulatory requirements.

`The insurance company` transitioned to GLMs around 2018, replacing simpler, older methods that relied on fixed formulas. The shift to GLMs was driven by the need for more accurate risk assessments, particularly for pricing models with variables like postcode factors. Back then, GLMs were considered as they offered several advantages, including enhanced explainability, which allows for clear, interpretable relationships between inputs and outputs. This transparency is crucial in the insurance industry, where stakeholders need to understand the rationale behind premium calculations. The transition involved refining broad regional groupings into more precise estimations using ML techniques. The implementation of GLMs required extensive validation, in which the Commercial Risk Management Committee played a role as well. This committee, comprising product management, directorial staff, and other key stakeholders, evaluates the models' impact on the company's profitability and market competitiveness. The department ensures that the models are aligned with company goals, while providing accurate and actionable insights. The committee's feedback was important to the models' refinement and validation. If necessary, the models undergo further adjustments before being finalized.

Explainability plays a crucial role in the acceptance and implementation of these models. `Individual D` provided examples where the explainability of GLMs was particularly advantageous. For instance, when discrepancies arise between expected and actual premium calculations, GLMs allow the team to trace back and clearly explain these differences. This capability is essential for internal validation and addressing customer inquiries about their premiums. The ability to provide detailed explanations builds trust with stakeholders and ensures compliance with regulatory standards.

The interview explored the performance of various ML models, including the tree-based models LightGBM and XGBoost, and neural networks. Validation methods such as cross-validation and APP plots were used to ensure models were correctly handling the available

data. This process helped maintain model accuracy and reliability, which is crucial for their acceptance and practical application within the organization. While `Individual D` recognizes that ML models can capture more complex patterns and potentially improve prediction accuracy, they often lack the transparency required for regulatory compliance and stakeholder trust. `Individual D` noted that ML models function as "black boxes", making it difficult to justify their predictions. Despite this, the potential benefits of ML models, particularly in terms of performance improvements, are recognized. However, any transition to more complex models would require significant advancements in explainability and integration with existing systems. While ML models showed potential, GLMs remained preferred due to their explainability and reliability. This preference is critical in contexts where the consequences of model predictions can have significant financial implications. Significant change will only occur when organizations such as `the consultancy firm` advise them to do so.

Comparing ML models with GLMs, `Individual D` discussed the use of tools like SHAP to interpret ML model predictions. SHAP values provide insights into the importance of each feature, aiding in the explainability of ML models. However, there are still concerns about fully replacing GLMs with ML models due to the latter's inherent complexity. The current preference for GLMs stems from their balance of accuracy and transparency, which ML models have yet to consistently achieve.

The discussion also covered the challenges of integrating ML algorithms into the current pricing process. This integration faces several obstacles, including the need for infrastructural upgrades and potential resource limitations. Additionally, according to `Individual D`, transitioning from GLMs to ML models would require extensive training and resources to equip the team with the necessary skills and knowledge. This includes understanding new validation techniques and interpreting complex model outputs. When evaluating the transition from GLMs to ML models, it is essential to consider the broader impact on operational efficiency, regulatory compliance, and stakeholder trust. The decision should be based on a thorough cost-benefit analysis, weighing the potential gains in predictive accuracy against the practical challenges of implementation.

E.2.5 Interview #5: Individual E

The interviews primarily involve a discussion with a professional named `Individual E`, who is involved in product and portfolio managementat `the insurance company`, with a specific focus on car insurance. Throughout the conversation, there is a significant emphasis on the comparison between traditional GLMs and more advanced ML techniques for determining insurance premiums. The core of the discussion revolves around the potential benefits and challenges associated with transitioning from GLMs to ML models within the insurance industry, particularly in pricing strategies.

One of the central themes of the interviews is the performance and potential of ML models in improving the accuracy of insurance pricing. `Individual E` discusses how ML models, with their capacity to handle more extensive and complex datasets, have the potential to offer more precise predictions compared to GLMs. However, it is noted that the current GLMs in use still perform adequately, and the margin of improvement with ML models, while present, is relatively small. This leads to a nuanced consideration of whether the transition to ML is warranted at this stage, especially given the operational and interpretative challenges it presents.

A significant concern raised during the interviews is the issue of explainability. GLMs, while more straightforward and less sophisticated than ML models, offer a level of transparency that is crucial in the insurance sector. The ability to explain how premiums are

calculated is vital, particularly when dealing with stakeholders such as customers and insurance advisors. `Individual E` highlights that the predictability and consistency offered by GLMs are valued by these stakeholders, who often need to justify premium differences between similar cases. The challenge with ML models is that, while they might offer slight improvements in accuracy, they can be seen as "black boxes," making it harder to interpret and explain the results in a clear and understandable way.

The interview also delves into the importance of certain features within the models used for predicting insurance claims costs. `Individual E` points out that variables such as the insured amount, the year of damage (referred to in Dutch as "schadejaar"), and the age of the vehicle could play a significant role in both GLM and ML models. These features can help in determining the likelihood and cost of claims. The discussion suggests that while ML models can potentially incorporate more variables and provide a more nuanced analysis, the traditional features still hold substantial predictive power in GLMs.

`Individual E` also touches on the practical challenges of implementing ML models in an operational setting. One of the key challenges is maintaining predictability and consistency in pricing, which is especially important when dealing with insurance advisors who expect similar premiums for similar risks. The consistency provided by GLMs ensures that advisors can offer reliable quotes to clients without facing significant price fluctuations, which could undermine client trust. `Individual E` emphasizes the need for robustness in pricing models, advocating for models that not only provide accurate predictions but also maintain a level of transparency and predictability over time.

The conversation briefly explores the future of car insurance pricing, with a mention of telematics as a potential area for development. Telematics, which involves using data on driving behavior to set premiums, could allow for more personalized and potentially more accurate insurance pricing. `Individual E` acknowledges that while current models focus heavily on the characteristics of the car (such as its age and value), there is a growing need to consider factors related to the driver and their behavior. This shift could lead to more sophisticated pricing models that better reflect the actual risk. However, questions arise regarding the practicality of this approach. Mobile phones are often considered insufficient, and the challenge of asking clients to pay for their telematics tracking device remains a complex issue.

Additionally, the interviews handled the internal processes related to model validation and approval within the company. `Individual E` discusses the roles of various teams, including data management and internal audit, in ensuring that pricing models are robust, compliant with regulations, and effectively integrated into business operations. This process involves multiple layers of validation, where models are reviewed and adjusted before being implemented. The discussion suggests a well-structured approach to model governance, ensuring that any pricing model, whether GLM or ML, meets the company's standards for accuracy and reliability.

In summary, the interviews provide a detailed exploration of the current state and potential future of pricing models in car insurance. `Individual E` offers insights into the ongoing debate between sticking with GLMs, which are transparent and well-understood, versus moving towards more complex ML models that offer higher accuracy but come with challenges in explainability and operational consistency. The interviews underscore the importance of balancing innovation with practicality in the insurance industry, ensuring that any new models adopted not only improve accuracy but also maintain the trust and understanding of all stakeholders involved.