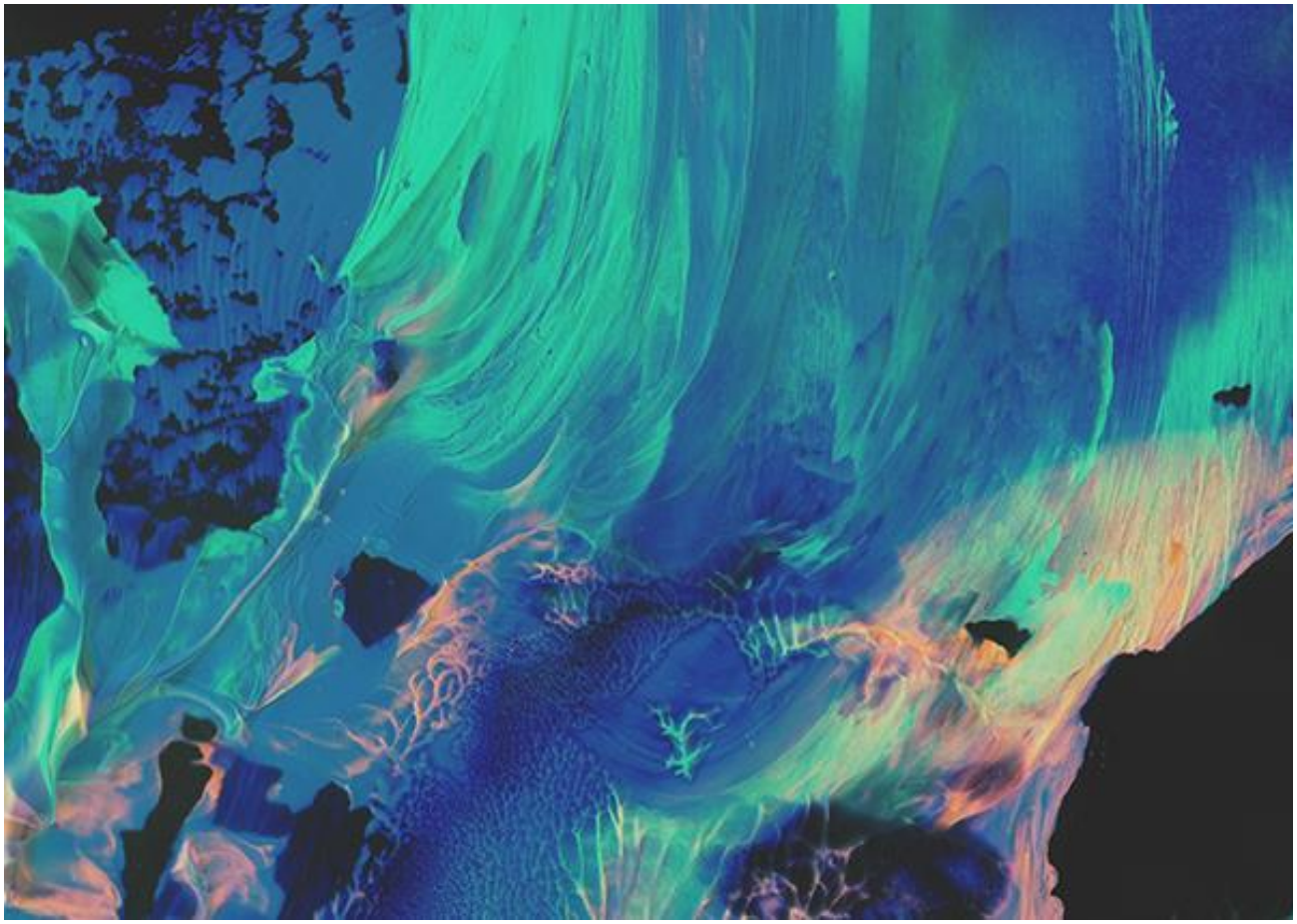
UMEÅ UNIVERSITY



Machine Learning Methods in Non-Life Insurance Pricing: A Comparative Study of Predictive Performance and Interpretability

Sanna Lockner

Master thesis, 30 ECTS
M.Sc in Industrial Engineering and Management, 300 ECTS
Spring term 2026

Abstract

This thesis investigates the use of machine learning methods in non-life insurance pricing, with a particular focus on claim frequency and claim severity modeling. The study compares the traditional Generalized Linear Model (GLM) with three machine learning methods: Random Forest (RF), Gradient Boosting Machine (GBM) and Neural Networks (NN). Three different insurance datasets were analyzed, consisting of insurance data from different countries.

The models were evaluated using Poisson deviance for claim frequency and Gamma deviance for claim severity. In addition, SHAP (SHapley Additive exPlanations) values were applied to investigate model interpretability.

The results show that GBM achieved the best or near-best predictive performance across the datasets. The ML methods performed the best, as expected, when the data includes stronger interaction effects and nonlinear patterns. The SHAP analyses showed that RF and GBM identified relatively consistent and interpretable relationships between risk factors and predicted claims, while the neural network produced more dispersed and less interpretable patterns.

To sum up, the machine learning methods can improve predictive performance in non-life insurance pricing but the traditional GLM remains relevant thanks to its interpretability and competitive performance. The ML methods performed better on two of three tested datasets.

Sammanfattning

Denna uppsats undersöker användningen av maskininlärningsmetoder inom prissättning av sakförsäkring, med särskilt fokus på modellering av skadefrekvens och skadekostnad. Studien jämför den traditionella Generalized Linear Model (GLM) med tre maskininlärningsmetoder: Random Forest (RF), Gradient Boosting Machine (GBM) och Neurala Nätverk (NN). Tre olika försäkringsdataset analyserades, bestående av försäkringsdata från olika länder.

Modellerna utvärderades med hjälp av Poissondevians för skadefrekvens och Gammadevians för skadekostnad. Därutöver användes SHAP-värden (SHapley Additive exPlanations) för att analysera modellernas tolkningsbarhet.

Resultaten visar att GBM uppnådde bäst eller nära bäst prediktiv prestanda över samtliga dataset. Resultaten visar också, som förväntat, att maskininlärningsmetoderna ger störst förbättringar i dataset som innehåller starkare icke-linjära samband och interaktionseffekter. SHAP-analyserna visade att RF och GBM identifierade relativt konsekventa och tolkningsbara samband mellan riskfaktorer och predikterade skador, medan neurala nätverk gav mer spridda och mindre tolkningsbara mönster.

Studien drar slutsatsen att maskininlärningsmetoder kan förbättra den prediktiva prestandan inom sakförsäkringsprissättning, men att den traditionella GLM-modellen fortfarande är relevant tack vare sin tolkningsbarhet och konkurrenskraftiga prestanda. Maskininlärningsmetoderna presterade bättre på två av de tre testade dataseten.

Acknowledgment

I would like to thank my supervisor Markus Ådahl at Umeå University for his valuable guidance, expertise and commitment throughout the Master thesis work. The support and feedback received from Markus has been very important for the writing of this thesis.

Sanna Lockner
May 2026

Contents

1 Introduction **7**
- 1.1 Background . . . 7
- 1.2 Project Overview . . . 7
- 1.3 Disposition . . . 8
- 1.4 Use of Artificial Intelligence . . . 8

2 Theory **9**
- 2.1 Non-Life Insurance Pricing . . . 9
- 2.2 Generalized Linear Models (GLM) . . . 10
- 2.3 Machine Learning . . . 13
- 2.4 Model Evaluation and Validation . . . 17
- 2.5 Model Interpretability . . . 19

3 Method **21**
- 3.1 Data . . . 21
- 3.2 Preprocessing . . . 24
- 3.3 Models . . . 24
- 3.4 Model Training . . . 25
- 3.5 Model Evaluation . . . 25

4 Result **27**
- 4.1 Research Question 1: Which machine learning method performs the best using deviance as a loss function? . . . 27
- 4.2 Research Question 2: How can these models be interpreted? . . . 29

5 Discussion **36**
- 5.1 Research Question 1: Which machine learning method performs the best using deviance as a loss function? . . . 36
- 5.2 Research Question 2: How can these models be interpreted? . . . 36
- 5.3 Previous Research . . . 37
- 5.4 Conclusions . . . 38

6 References **39**

7 Appendix **40**
- 7.1 Dataset 1: Motorcycle Case . . . 40
- 7.2 Dataset 2: Australian Motor Case . . . 40
- 7.3 Dataset 3: Belgian Motor Case . . . 40

Abbreviations

GLM	Generalized Linear Model
RF	Random Forest
GBM	Gradient Boosting Machine
NN	Neural Network
SHAP	SHapley Additive exPlanations
CV	Cross-Validation
EDM	Exponential Dispersion Model

1 Introduction

This section will provide the reader with some background knowledge about the subject, but also introduce the project and its statements, motivation, purposes and limitations.

1.1 Background

Non-life insurance pricing aims to determine premiums that reflect the underlying risk associated with an insurance policy while remaining competitive and financially sustainable. A claim is described as an event reported by the policyholder to seek economic compensation. Premiums are typically based on the expected cost of future claims, which can be explained by the key ratios claim frequency and claim severity [1, pp. 1-4]. Accurate modeling of these two components is therefore central to pricing and risk assessment in non-life insurance. Generalized linear models (GLM) were introduced in actuarial science during the 1990s and have since become the standard approach for modeling claim frequency and severity in many countries [1, p. 15].

There are many advantages with GLM; it is a part of general statistical theory, which means, for instance, that formulas to calculate confidence interval exists. Additionally, software for fitting GLM is available in various programs, such as in R. Another advantage is its interpretability, which facilitates communication of results to stakeholders [1, pp. 15–16].

In recent years, advances in machine learning have generated increasing interest in more flexible modeling approaches for insurance pricing. Machine learning methods have demonstrated the ability to learn complex patterns and interactions from data and have shown promising results in the non-life insurance domain. This is more difficult and time consuming when using the GLM. For instance, the user then has to create interactions between the variables themselves, while in machine learning, the model finds the interactions while training. These developments motivate an investigation of machine learning methods for claim frequency and severity modeling. Machine learning is showing great promise in non-life insurance pricing because it can automatically capture nonlinear relationships and interaction effects that are often difficult and time-consuming to specify in a traditional GLM framework [2, pp. 1–2].

It is already established that the model needs to be able to predict the premium but pricing models also need to justify premiums to customers under Swedish regulations [3]. While complex models such as deep neural networks are often difficult to interpret, simpler machine learning models such as Logistic Regression and decision-tree-based methods are often easier to explain. It can be described as a trade-off between the performance of the method and how explainable it is to stakeholders [4, pp. 1–2].

1.2 Project Overview

This section presents an overview of the thesis, including project description, research questions, motivation, purpose, limitations, and structure.

Project Description and Purpose

This thesis investigates the use of the three supervised machine learning methods Gradient Boosting Machine (GBM), Random Forest (RF), and Neural Networks (NN) for modeling claim frequency and claim severity in non-life insurance. The predictive performance of these methods is compared with that of a traditional Generalized Linear Model (GLM), which is used as the baseline model. In addition, the interpretability of the models is evaluated to assess the trade-off between predictive performance and model transparency in an insurance pricing context.

Problem Statement

The following aspects will be studied:

- Which machine learning method performs the best using deviance as a loss function?
- How can these models be interpreted?

Motivation

Accurate pricing is of central importance in non-life insurance as it affects risk management, profitability and competitiveness. Improving predictive accuracy for claim frequency and severity is therefore of significant practical relevance for insurance companies. From an academic perspective, the increasing availability of insurance data makes it easier to apply and compare machine learning methods as alternatives to traditional actuarial models. Since these methods have the potential to capture more complex patterns in data, it is relevant to investigate whether they can improve predictive performance in non-life insurance pricing.

Limitations

This study is limited to a small number of non-life insurance datasets that differ with respect to characteristics such as company, country, and type of insurance product. As a result, the findings may not be directly generalizable to all non-life insurance contexts. Furthermore, the analysis is restricted to the three machine learning models: Random Forest (RF), Gradient Boosting Machine (GBM) and Neural Networks (NN). While these models represent commonly used approaches in the literature, other machine learning methods not included in this study may yield different results and performance.

Finally, it should be noted that the models in this study do not directly estimate insurance premiums. Instead, claim frequency and claim severity are modeled separately, and the premium can in principle be derived as a combination of these components. However, the focus of this study is not on pricing itself, but on comparing the predictive performance of different models for frequency and severity.

1.3 Disposition

The remainder of this thesis is structured as follows. Section 2 presents the theoretical framework relevant to this study. Section 3 describes the methodology and datasets, including preprocessing and model implementation. Section 4 presents the results, while Section 5 discusses the findings, compares them with earlier research and presents the conclusions.

1.4 Use of Artificial Intelligence

Artificial intelligence tools have been used in this study as a supportive aid during the research process. Specifically, AI has been utilized for tasks such as code debugging and structuring of text. All methodological choices, model implementations, and interpretations of results have been made independently by the author. The use of AI has not replaced critical thinking or academic judgment, but has instead served as a tool to improve efficiency and support understanding. All final content has been reviewed and validated by the author.

2 Theory

This chapter provides the theoretical foundation for the thesis and is divided into subsections covering the relevant models and concepts.

2.1 Non-Life Insurance Pricing

This section presents the theoretical framework for the price of non-life insurance, focusing on the decomposition of premiums into key ratios (Y_i) claim frequency and claim severity. A tariff structure defines how different risk factors influence the insurance premium and how policyholders are grouped into different risk classes. Therefore a tariff cell refers to a specific combination of risk factors within a tariff structure [1, p. 3].

2.1.1 Basic Model Assumptions

The theory relies on several basic assumptions. These assumptions ensure that the model provides consistent and interpretable estimates. The following three assumptions form the foundation for the statistical model used in this thesis [1, pp. 6–7].

1. **Policy Independence.** Consider n different policies. For any response variable in Table 1, let X_i denote the response for policy i. Then X_1, ..., X_n are independent.
2. **Time Independence.** Consider n disjoint time intervals. For any response variable in Table 1, let X_i denote the response in time interval i. Then X_1, ..., X_n are independent.
3. **Homogeneity.** Consider any two policies in the same tariff cell, having the same exposure. For any response variable in Table 1, let X_i denote the response for policy i. Then X_1 and X_2 have the same probability distribution.

2.1.2 Key ratios

The key ratios can be described as a ratio between the outcome of a random variable and an exposure, which provides an important part, as the more exposure is available, the less variation. They are presented in the form

$$Y_i = \frac{X_i}{w_i}$$

where, Y_i is the chosen key ratio, X_i is the response variable, and w_i is the exposure [1, pp. 4–6]. Table 1 illustrates how the different components of the premium are constructed from observable quantities. As mentioned, the pure premium can also be calculated from multiplying claim frequency and claim severity. This decomposition forms the basis for separate modeling of the two [1].

Key Ratio Y_i	Response X_i	Exposure w_i
Claim Frequency	Number of claims	Duration
Claim Severity	Claim cost	Number of claims
Pure Premium	Claim cost	Duration

Table 1: Relationship between exposure, response variables and key ratio.

By the assumptions 1, 2 and 3, one can write the expectation of a key ratio Y_i as in Definition 2.1 [1, p. 8].

Definition 2.1 (Expectation of key ratio and response variable)**.** *Consider the situation where w is the number of claims, so that we can write X as a sum of w individual responses $Z_1, \ldots, Z_w$. Assumptions*

3 and 1 imply that Z_k are independent. Assumption 2 implies equal distribution. Now elementary rules for expectation and variance yield [1, p. 8].

$$E[Y_i] = \mu_i$$

$$Var[Y_i] = \frac{\sigma_i^2}{w_i}$$

$$E[X_i] = \mu_i w_i$$

$$Var[X_i] = \sigma_i^2 w_i$$

2.1.3 Tariff Analysis

To illustrate a simplified tariff structure, Table 2 presents four tariff cells corresponding to all possible combinations of two rating factors, each with two levels. In tariff analysis, one cell is selected as the reference. In this example, Cell 1 (Rural, Old) is chosen as the reference cell, and all other tariff cells are evaluated relative to that. Relativity is defined as a multiplicative factor that measures the risk level of a given tariff cell relative to the reference cell. From Table 2, it can be observed that Cell 4 (Urban, Young) has a pure premium equal to four times that of the reference cell. Consequently, the relativity for this cell is 4, indicating a substantially higher expected risk [1, pp. 9–10].

	Rating Factors		Claim	Claim	Pure	Relativity
Cell	Zone	Age	Frequency	Severity	Premium	
1	Rural	Old	0.03	10,000	300	1.00
2	Rural	Young	0.06	10,000	600	2.00
3	Urban	Old	0.06	10,000	600	2.00
4	Urban	Young	0.12	10,000	1,200	4.00

Table 2: Example of tariff cells with claim frequency, severity and pure premium. Cell 1 is the reference tariff cell.

To clarify, the goal of the tariff analysis is to determine how the chosen key ratio Y_i varies with the rating factors. This problem is conceptually similar to regression analysis, where the relationship between a response variable and explanatory variables is modeled. However, ordinary linear regression is generally not suitable for non-life insurance pricing, as the response variables often exhibit non-negative values and non-constant variance. Therefore, the generalized linear model (GLM) framework is introduced [1, p. 15].

2.2 Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are an extension of ordinary linear regression models that allow the response variable to follow distributions other than the normal distribution. A GLM consists of three main components: a random component specifying the probability distribution of the response variable, a systematic component represented by a linear predictor, and a link function connecting the expected response to the linear predictor. GLMs are commonly formulated using exponential dispersion models (EDMs) [1, pp. 15–30].

2.2.1 Multiplicative model and the log-link

In non-life insurance pricing, key ratios are commonly modeled using a multiplicative structure, where each rating factor contributes a multiplicative effect to the expected value. Consider a tariff cell (i, j) defined by two rating factors with levels i and j. For each cell (i, j), let w_{ij} denote the exposure, X_{ij}

the response variable, and Y_{ij} the corresponding key ratio. The expected value of the key ratio in tariff cell (i, j) can then be expressed as a multiplicative model, as shown in Definition 2.2 [1, pp. 9–10].

Definition 2.2 (Multiplicative model with two rating factors)**.**

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}$$

where μ_{ij} is the expected value of the key ratio in cell (i, j), γ_0 is the base expectation corresponding to the reference cell, and γ_{1i} and γ_{2j} are the relativities associated with rating factors 1 and 2, respectively.

The multiplicative structure in Definition 2.2 can be illustrated using the tariff analysis presented in Table 2. In the table, the reference tariff cell (1,1) has a pure premium of 300. The relativities associated with the rating factors adjust this baseline multiplicatively. According to the multiplicative model, the expected pure premium for the tariff cell (2,2) is therefore

$$\mu_{2,2} = 300 \times 2 \times 2 = 1200,$$

which corresponds exactly to the value observed in Cell 4 of Table 2.

$$\log(\mu_{ij}) = \log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j}),$$

which transforms the multiplicative structure into an additive model. This property motivates the use of a log-link function in generalized linear models [1, p. 27].

2.2.2 Exponential Dispersion Models

The probability distribution of an exponential dispersion models is given by the following Definition 2.3 [1, pp. 16–23].

Definition 2.3 (Exponential Dispersion Model (EDM))**.** *A random variable Y is said to follow an exponential dispersion model if its probability density (or mass) function can be written in the form*

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi / w_i} + c(y_i, \phi, w_i) \right),$$

where θ is a parameter that is allowed to depend on i, $\phi > 0$ is the dispersion parameter and is the same for all i, and $b(\cdot)$ and $c(\cdot)$ are known functions.
For an exponential dispersion model, the mean and variance are given by

$$E[Y] = \mu = b'(\theta_i), \qquad \mathrm{Var}(Y) = b''(\theta)\phi \,/ w_i.$$

Many distributions are an EDM, for example Poisson and Gamma, the Normal distribution is also an EDM. See Johansson and Ohlsson [1, pp. 18, 19, 21] for examples of distributions belonging to the exponential dispersion model family.

Within the EDM framework, maximum likelihood estimation is used to estimate the model parameters by selecting the values that maximize the likelihood of observing the data under the assumed distribution [1, pp. 30–31].

2.2.3 Poisson GLM

Claim frequency represents the number of claims during a given exposure period. Since the response variable is a count variable taking non-negative integer values, it is commonly modeled using the Poisson distribution in non-life insurance pricing. The Poisson distribution for claim counts is given in Definition 2.4 [1, pp. 18–19].

Definition 2.4 (Poisson distribution for claim counts)**.** *Let X_i denote the number of claims observed for policy i during the exposure period w_i. The claim count is assumed to follow a Poisson distribution*

$$X_i \sim Poisson(w_i\mu_i),$$

where $E[X_i] = w_i\mu_i$, denotes the expected number of claims for policy i. $Y_i = \frac{X_i}{w_i}$ follows a relative Poisson distribution with frequency function,

$$f_{Y_i}(y_i;\mu_i) = \exp(-w_i\mu_i)\,\frac{(w_i\mu_i)^{w_i y_i}}{w_i y_i!}$$

The claim frequency will later be modeled as a function of the rating factors within the generalized linear model framework.

2.2.4 Gamma GLM

Claim severity represents the average cost per claim. Since claim amounts are positive and typically right-skewed, the Gamma distribution is commonly used to model claim severity in non-life insurance pricing. See Definition 2.5 [1, p. 20]

Definition 2.5 (Gamma distribution for claim severity)**.** *Let X_i denote claim cost observed for policy i and let the exposure w_i denote number of claims. The claim severity is assumed to follow a Gamma distribution*

$$X_i \sim Gamma(w_i, \alpha, \beta),$$

Let Y_i denote the claim severity observed for policy i and let w_i denote the number of claims.

$$f_{Y_i}(y_i;\mu_i,\phi) = \frac{w\beta^{w\alpha}}{\Gamma(w\alpha)}y^{w\alpha-1}\exp(-w\beta y)\,, \qquad y > 0.$$

where $\alpha > 0$ is the index parameter and $\beta > 0$ is the scale parameter.

2.2.5 Model Formulation

Generalized linear models relate the expected value of a response variable to a set of explanatory variables through a link function and a linear predictor. Let Y_i denote the response variable and let

$$\mu_i = E[Y_i].$$

The generalized linear model is defined by

$$g(\mu_i) = \eta_i,$$

where $g(\cdot)$ denotes the link function and η_i is the linear predictor

$$\eta_i = x_i^\top \beta.$$

Here x_i represents the vector of rating factors and β the corresponding parameter vector.
In non-life insurance pricing, the log-link function is commonly used. The model can therefore be written as

$$\log(\mu_i) = x_i^\top \beta,$$

which implies

$$\mu_i = \exp(x_i^\top \beta).$$

This formulation corresponds to the multiplicative tariff structure introduced earlier in the chapter.

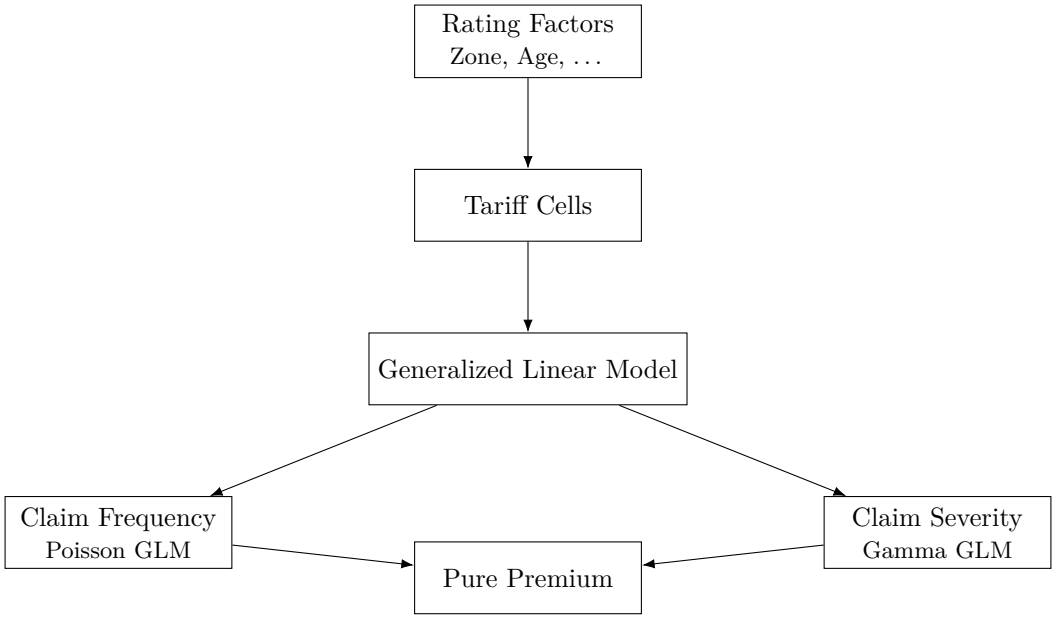


Figure 1: Conceptual relationship between rating factors, tariff cells, generalized linear models, and pure premium in non-life insurance pricing.

2.3 Machine Learning

Machine learning methods have gained increasing attention in non-life insurance pricing due to their ability to capture complex nonlinear relationships and interactions between variables [2]. In this thesis, several methods are considered and compared to the generalized linear model. In this thesis, only supervised machine learning methods are considered.

2.3.1 Supervised Learning

Supervised learning refers to a class of machine learning methods in which a model is trained using labeled data to learn the relationship between a set of explanatory variables, often referred to as rating factors, and a response variable. In non-life insurance pricing, historical data are used to estimate premiums that reflect the expected cost of future claims. The training data consist of observations $(\boldsymbol{x_i}, y_i)$ for $i = 1, \ldots, n$, where $\boldsymbol{x_i}$ denotes a vector of input variables (features) and y_i represents the observed response variable. For each observation, there is also an exposure variable w_i as mentioned in Table 1 [5, p. 5].

2.3.2 Overfitting and Regularization

Overfitting occurs when a model fits the training data too closely and captures noise, leading to poor generalization to new data. This issue is particularly relevant for flexible models, which have the capacity to model complex patterns but are also more prone to fitting random variation in the data. This is closely related to the bias–variance trade-off, where more flexible models have lower bias but higher variance, while simpler models exhibit the opposite behavior. The optimal model balances these two components to minimize the expected test error. To address overfitting, regularization techniques are applied to control model complexity and improve generalization performance. Regularization can be implemented in different ways depending on the model [6, pp. 22–24, 33–36]. In practice, the

choice of regularization technique depends on the specific model, and the corresponding methods are discussed in connection with each model in Chapter 3.

2.3.3 Decision trees

Decision trees can be used for both regression and classification problems. In the context of regression, the method aims to approximate the relationship between a response variable and a set of explanatory variables by partitioning the feature space into a set of disjoint regions [6, p. 327].

More specifically, the predictor space is divided into J non-overlapping regions $R_1, R_2, \ldots, R_J$. For each observation that falls into a region R_j, the prediction is given by a constant value, typically the mean of the response variable within that region. The partitioning is performed using a recursive binary splitting procedure, where at each step a predictor and a split point are chosen in order to best separate the data into more homogeneous groups, typically by minimizing a loss function [6, p. 330–331]. See James et al [6, Ch. 8] for a detailed introduction to decision trees.

Figure 2 illustrates how a decision tree can be interpreted in the context of non-life insurance pricing. In this simplified example, the feature space is partitioned using the rating factors Zone and Age, resulting in four terminal nodes corresponding to the tariff cells introduced in Table 2. Each terminal node is associated with a constant expected value μ, corresponding to the pure premium within that tariff cell.

Although decision trees are easy to interpret and align naturally with the concept of tariff cells, they tend to have high variance and may be sensitive to small changes in the data. This motivates the use of ensemble methods such as Random Forest, where multiple trees are combined to improve predictive performance and stability. [6, pp. 339–340].

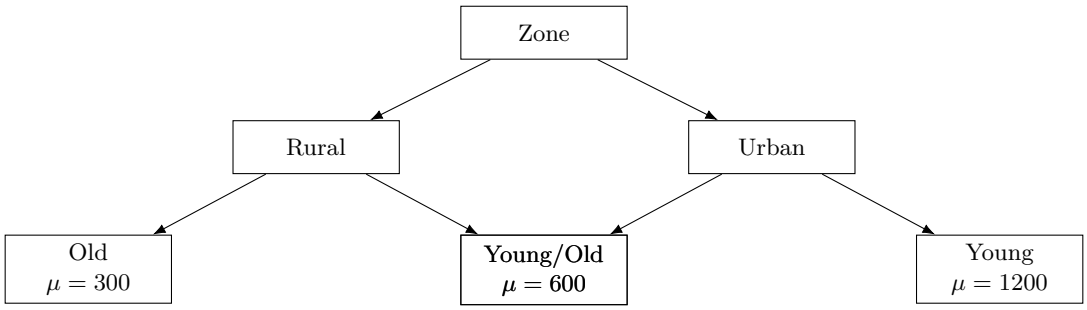


Figure 2: Decision tree representation of the tariff structure in Table 2.

2.3.4 Bagging

Bagging is an ensemble method that aims to reduce the variance of prediction models by combining multiple decision trees. The method is particularly effective for decision trees, since single trees tend to suffer from high variance, meaning that small changes in the training data can lead to substantially different tree structures and predictions. The main idea behind bagging is that averaging many high-variance models can produce a more stable and accurate prediction model [6, pp. 340–341].

2.3.5 Random Forest

Random Forest extends bagging by introducing additional randomness during tree construction. As in bagging, multiple decision trees are trained on bootstrap samples of the training data, and their predictions are combined for a final prediction. For clarification, bootstrap sampling refers to drawing

observations from the dataset with replacement to create multiple new training datasets. [6, pp. 340–341].

However, in contrast to bagging, Random Forest modifies the tree-building process by considering only a random subset of the available predictors at each split. More specifically, when a split is considered, only m randomly selected predictors are evaluated from the total of p, typically, $m \approx \sqrt{p}$. The motivation is to reduce the correlation between trees. In bagging, strong predictors are often repeatedly selected, leading to similar trees and highly correlated predictions. By considering only a random subset of predictors at each split, Random Forest increases tree diversity, which improves averaging and reduces variance [6, pp. 343–345]. See James et al [6, 6, Ch. 8] for a more detailed introduction to Random Forest.

2.3.6 Gradient Boosting Machines

Boosting is an ensemble method in which a large number of small decision trees are combined into a single predictor. In contrast to bagging and Random Forest, where trees are built independently, boosting grows the trees sequentially, such that each new tree is fitted using information from the previously fitted trees [5, pp. 21–23].
The basic idea is to start with an initial model and iteratively improve the predictor by adding new trees that focus on the remaining errors. In the standard regression setting, this corresponds to fitting each new tree to the residuals of the current model. More generally, boosting is performed by fitting each new tree to the pseudo-residuals, defined as the negative gradient of the loss function [5, pp. 25–29].

This results in a forward stagewise additive model, where the predictor is updated step by step by adding a new tree in each iteration. The method is therefore able to capture complex nonlinear relationships and interactions between rating factors, while still relying on simple trees as building blocks [5, pp. 21–22].

Boosting reduces overfitting through a combination of regularization mechanisms. First, the model is built in a forward stagewise manner, where small trees are added sequentially. By restricting each tree to have a limited number of terminal nodes, the complexity of the model is controlled. Additionally, the procedure can be improved by combining subsampling and shrinkage. Subsampling means that, in each iteration, only a fraction η of the observations is used, where $\eta \approx 0.5$ is a common choice. This introduces randomness and reduces variance. Shrinkage scales the contribution of each tree by a factor ν, typically chosen such that $0 < \nu < 0.1$, which slows down the learning process and improves generalization.Consequently, the main tuning parameters in boosting are the number of trees, the tree depth, the subsampling rate, and the shrinkage parameter [5, pp. 25–30].

Gradient Boosting algorithm

For a given number of trees M and number of terminal nodes J in each tree, initialize the model with a single tree $f_1(x)$. The following steps are then repeated for $m = 2, \ldots, M$.

1. For $m = 2, \ldots, M$:
   (a) Compute pseudo-residuals r_i.
   (b) Fit a regression tree with J terminal nodes to the weighted pseudo-residuals $w_i r_i$, using a quadratic loss function. Denote the resulting terminal regions by $R_{1m}, \ldots, R_{Jm}$.
   (c) For each region R_{jm}, compute the constant γ_{jm} as the minimizer of

$$\sum_{i: x_i \in R_{jm}} w_i \, L\big(y_i, g^{-1}(f_{m-1}(x_i) + \gamma_{jm})\big).$$

(d) Define the new tree

$$g_m(x) = \sum_{j=1}^{J} \gamma_{jm} \, \mathbf{1}_{\{x \in R_{jm}\}}.$$

(e) Update the model

$$f_m(x) = f_{m-1}(x) + g_m(x).$$

2. Output the final model:

$$\hat{f}(x) = \hat{f}_M(x).$$

See Johansson and Ohlsson [5, pp. 27–28] for a detailed derivation and discussion of the algorithm.

2.3.7 Neural Networks

Neural networks are flexible models used to approximate complex nonlinear relationships between input variables $X = (X_1, \ldots, X_p)$ and a response Y. A feed-forward neural network consists of an input layer, one or several hidden layers, and an output layer, where each layer applies a linear transformation followed by a nonlinear activation function.
They can be viewed as an extension of linear models, where the linear predictor is replaced by a sequence of transformations that allow for nonlinearity. A neural network consists of layers of interconnected units, referred to as neurons. Each hidden layer is composed of units that compute a weighted sum of the inputs, followed by a nonlinear activation function.
For a network with multiple layers, the activations in layer l can be written as

$$A_k^{(l)} = g\left(w_{k0}^{(l)} + \sum_{j=1}^{K_{l-1}} w_{kj}^{(l)} A_j^{(l-1)} \right),$$

where l denotes the layer index, k indexes the activations in layer l, j indexes the activations in the previous layer $(l-1)$, and K_{l-1} denotes the number of activations in layer $(l-1)$. The activation $A_j^{(l-1)}$ denotes the activations from the previous layer, $w_{kj}^{(l)}$ are weights, and $g(\cdot)$ is a nonlinear activation function.

The final output is obtained by combining the activations from the final hidden layer in a linear fashion,

$$f(X) = \beta_0 + \sum_{k=1}^{K_L} \beta_k A_k^{(L)}.$$

where L denotes the final hidden layer and K_L is the number of activations in that layer.

which can be interpreted as a linear model applied to nonlinear transformations of the input variables. Each neuron performs a two-step computation. First, it computes a weighted sum of the inputs, including a bias term. Second, this value is passed through a nonlinear activation function. The purpose of the activation function is to introduce nonlinearity, which enables the network to capture complex relationships in the data [6, pp. 403–407].

Modern neural networks typically consist of multiple hidden layers. Although a single hidden layer is theoretically sufficient to approximate many functions, deeper architectures often make learning more efficient and allow the model to capture more complex structures in the data [6, pp. 407–408]. Network parameters are estimated by minimizing a loss function, typically using gradient-based optimization methods such as stochastic gradient descent. In gradient descent, the model parameters are iteratively updated in the direction of the negative gradient of the loss function to minimize the prediction error. Gradients are efficiently computed using backpropagation, which updates the weights by propagating

the error backwards through the network [6, pp. 434–436]. The overall structure of a feed-forward neural network is illustrated in Figure 3 [6, p. 405].

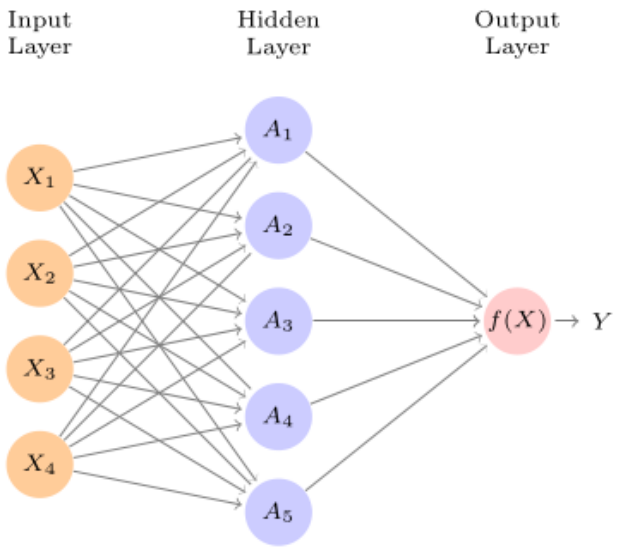


Figure 3: Feed-forward neural network illustrating how input variables are transformed through a hidden layer to produce a prediction [6, p. 405].

Due to their flexibility, neural networks can achieve high predictive performance, but are also prone to overfitting. Regularization techniques are therefore commonly used to improve generalization. One common approach is dropout, where a fraction of the units in a layer are randomly deactivated during training. This prevents individual neurons from becoming overly specialized to the training data and reduces the risk of overfitting, similar in spirit to subsampling methods. In addition to regularization, several hyperparameters need to be carefully tuned in practice. These include the number of hidden layers, the number of units per layer, the batch size, and the number of training epochs. The performance of the model is sensitive to these choices, and they are typically selected using cross-validation [6, pp. 436–438].

Despite their strong predictive capabilities, neural networks are often considered less interpretable than traditional models such as generalized linear models. This trade-off between predictive accuracy and interpretability is particularly relevant in insurance pricing, where regulatory requirements often demand transparent and explainable models. See James et al. [6, Ch. 10] for a broader review of neural networks.

2.4 Model Evaluation and Validation

Model evaluation is performed to assess predictive performance and the ability of the models to generalize to unseen data. To achieve this, cross-validation is applied, including nested cross-validation to separate model selection from evaluation. The models are assessed using loss functions consistent with the assumed distributions, where Poisson deviance is used for claim frequency and Gamma deviance for claim severity.

2.4.1 Cross Validation

Cross-validation (CV) is a resampling technique used to estimate the predictive performance of a model and to assess how well it generalizes to unseen data. The available dataset is repeatedly split into training and validation subsets, allowing the model to be trained and evaluated on different parts of the data. Cross-validation allows for efficient use of data, as all observations are used for both training and validation [6, pp. 198–200]. CV is commonly used for both model assessment and model selection. In particular, it provides a data-driven approach to choosing between competing models or tuning hyperparameters by comparing their predictive performance on validation data [6, pp. 232–237].

There are different ways to implement CV. In K-fold cross-validation, the dataset is partitioned into K mutually exclusive subsets (folds) of approximately equal size. The model is trained on $K-1$ folds and evaluated on the remaining fold. This procedure is repeated K times such that each fold is used once as validation data, and the results are averaged to obtain an estimate of the test error or loss. The choice of K in K-fold cross-validation involves a bias–variance trade-off. Empirical studies suggest that values such as $K = 5$ or $K = 10$ often provide a good balance between bias and variance [6, p. 206].

Another approach is the stratified K-fold cross-validation, where the folds are constructed such that the distribution of some variable is approximately preserved within each fold [7, pp. 99–103]. This is particularly important in imbalanced datasets, such as insurance data, where the proportion of observations with claims is typically low. Stratification ensures that each fold contains a representative share of claim and non-claim observations, leading to more stable and reliable performance estimates.

In addition to standard CV, nested cross-validation can be applied when model selection and hyperparameter tuning are involved. When using standard cross-validation for both tuning and evaluation, there is a risk of obtaining overly optimistic performance estimates, since the model is indirectly evaluated on data that has influenced the tuning process. Nested cross-validation addresses this by separating model selection and evaluation into two loops. The inner loop is used to select the optimal hyperparameters, while the outer loop is used to evaluate the model on unseen data. The performance is then averaged across the outer folds to obtain an estimate of the model's generalization performance [8].

2.4.2 Deviance and Loss Functions for EDM

Within the GLM framework, deviance is derived from the likelihood function and is used to measure the goodness-of-fit of a model. More specifically, the deviance compares the fitted model with the saturated model, where the saturated model represents a hypothetical model that fits the data perfectly [1, p. 39]. A smaller deviance therefore indicates a better fit to the observed data. This can be compared with the null model as well, where the null model is a baseline model that does not include any explanatory variables [6, p. 79].

In non-life insurance modeling, the choice of loss function is usually closely linked to the assumed distribution of the response variable. Within the framework of exponential dispersion models, the loss function is typically defined in terms of the unit deviance, which measures the discrepancy between observed values y and predicted values $\hat{y}$ [5, pp. 6–8].
The total loss is given by

$$L = \sum_{i=1}^{n} w_i \, d(y_i, \hat{y}_i),$$

where w_i denotes the exposure and $d(y, \hat{y})$ is the unit deviance.
For commonly used distributions in insurance, the unit deviations take the following forms:

Poisson deviance
The Poisson deviance is used to model the claim frequency. It is defined as

$$d(y, \hat{y}) = 2\left(y \log y - y - y \log \hat{y} + \hat{y}\right)$$

with the convention that $d(0, \hat{y}) = 2\hat{y}$. This loss function is appropriate when the response represents counts and the variance is proportional to the mean.

Gamma deviance
The Gamma deviance is commonly used for modeling claim severity. It is given by

$$d(y, \hat{y}) = 2\left(-1 - \log y + \frac{y}{\hat{y}} + \log \hat{y}\right)$$

and is suitable when the variance increases with the square of the mean, which is often realistic for monetary quantities.

Interpretation of deviance
The deviance can be interpreted as a distance measure between observed and predicted values, and minimizing the total deviance corresponds to maximizing the likelihood in a generalized linear modeling framework [5, pp. 6–7].

Furthermore, for EDMs, the deviance can be approximated by

$$d(y, \hat{y}) \approx \frac{(y - \hat{y})^2}{v(\hat{y})},$$

where $v(\hat{y})$ is the variance function. This highlights how the loss function adapts to the underlying distribution and variance structure [5, pp. 8–9].

2.5 Model Interpretability

Model Interpretability refers to the ability to understand how a model produces its predictions and how different input variables influence the output. While traditional models such as generalized linear models are inherently interpretable, more flexible machine learning models often act as black boxes [6, pp. 24–26]. In this study, interpretability is particularly important when comparing models with different levels of complexity. As more advanced models are introduced, additional methods are required to explain their predictions and ensure transparency.

2.5.1 Importance of Explainability in Insurance

In non-life insurance pricing, the primary objective is to obtain accurate predictions of future risk premiums. However, model interpretability also plays an important role in this process, as it supports decision-making and evaluation of the model. A well-interpreted model can help determine whether adjustments to the tariff structure are needed. Beyond predictive accuracy, interpretability serves several important purposes. First, it supports market considerations, where pricing decisions may need to balance risk-based pricing with customer acceptance. For example, although risk may increase with age, higher premiums for certain customer groups may negatively affect customer relations. Furthermore, interpretability is essential for customer communication. Insurers must be able to explain pricing decisions, and transparency is increasingly required by regulation. Finally, model interpretation can support claim prevention by identifying risk factors and encouraging safer behavior among policyholders. These considerations highlight that, in insurance applications, interpretability is not only a technical requirement but also a business and regulatory necessity. To address the interpretability challenges of complex models, methods such as SHAP values can be applied [5, p. 41].

2.5.2 SHAP Values

SHAP (SHapley Additive exPlanations) values are a model interpretability method based on Shapley values from cooperative game theory. The method explains individual predictions by decomposing them into additive contributions from each feature.

The SHAP value for feature j is defined as

$$\phi_j = \sum_{S \subseteq K \setminus \{j\}} \frac{|S|!(k - |S| - 1)!}{k!} \left[f_{S \cup \{j\}}(x_i) - f_S(x_i) \right],$$

where K denotes the set of all features, $k = |K|$ is the total number of features, and S is a subset of features excluding feature j. Here, $f_S(x_i)$ denotes the prediction obtained using only the subset of features S. The weighting term

$$\frac{|S|!(k - |S| - 1)!}{k!}$$

represents the probability of selecting the subset S from the set of all features, while

$$f_{S \cup \{j\}}(x_i) - f_S(x_i)$$

represents the marginal contribution of feature j when added to the subset S.

The SHAP value therefore represents the expected marginal contribution of a feature across all possible subsets of features. A positive SHAP value indicates that the variable increases the prediction relative to the baseline, while a negative value indicates a decrease. The baseline in this case, is the model's average prediction. SHAP values can be used both for local interpretation, where individual predictions are explained, and for global interpretation through aggregated feature importance measures. In this study, SHAP values are used to identify the most influential rating factors and to analyze how different feature values affect the predicted claim frequency and severity. The additive decomposition provided by SHAP values aligns naturally with gradient boosting models, where predictions are constructed as sums of contributions from multiple trees [5, pp. 45–50]. SHAP values are therefore particularly useful for improving transparency and interpretability in insurance pricing models. See Johansson & Ohlsson [5, Ch. 4.3] for a broader review of SHAP values.

A practical limitation of SHAP values is that categorical variables are typically transformed into numerical representations before the SHAP values are computed. Consequently, SHAP methods are generally more straightforward to apply and interpret for numerical variables. For categorical variables with many levels, interpretation may become more difficult since the contribution of one original rating factor can be distributed across several encoded variables.

3 Method

The approach is structured around the two main research questions, where the methodology for the first question is described first, followed by the approach applied to the second question.

Research Question 1: Which machine learning method performs the best using deviance as a loss function?

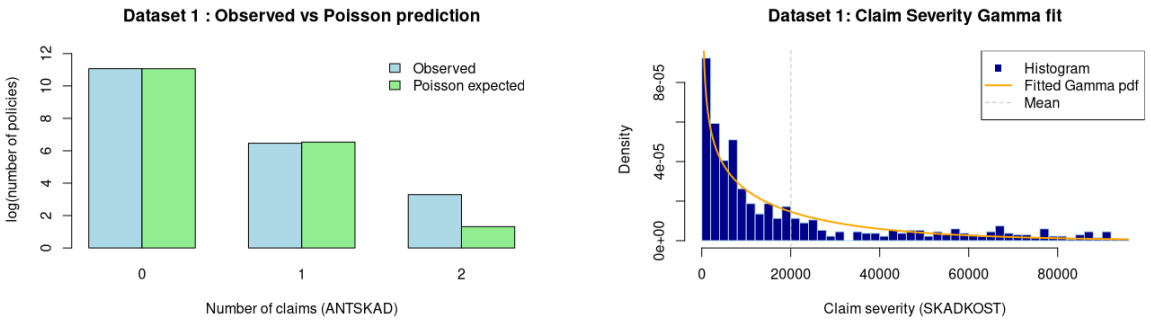
3.1 Data

In this study, three different datasets are analyzed. Each dataset has been preprocessed, including data cleaning and feature engineering. Furthermore, relevant statistical assumptions have been assessed. Three models are then applied to the data and evaluated based on their predictive performance. This section describes each dataset used. In particular, the assumptions of the Poisson and Gamma distributions are examined for each dataset. All data processing, analysis, and model implementation are conducted in the programming language R. More information about the rating factors are in the Appendix.

3.1.1 Dataset 1: Motorcycle Case

This dataset is provided by Länsförsäkringar Alliance and consists of motorcycle insurance data from the years 1994 to 1999. It is also used as an example dataset in the course book by Johansson and Ohlsson [1, pp. 3–4]. Descriptive statistics for the dataset are presented below. Furthermore, Figure 4 illustrates the empirical distributions of claim frequency and claim severity. These distributions provide empirical support for modeling claim frequency using a Poisson distribution and claim severity using a Gamma distribution. Although the data do not perfectly follow these distributions, the assumptions are considered reasonable for modeling purposes.

- The dataset contains 64,548 observations (policies).
- Of these, 62,474 observations have positive exposure (DURATION > 0).
- Approximately 11% of the policies have at least one claim. The dataset is imbalanced, with a large proportion of zero-claim observations.
- 17% of the policyholders are female.
- The dataset includes categorical rating factors: vehicle class, geographical zone, owner age, gender and vehicle age.
- There is imbalance between categories; for example, zone 7 (Gotland) contains only 367 policies, of which only one has a claim.



(a) Claim frequency

(b) Claim severity

Figure 4: Empirical distributions for claim frequency and severity.

Table 3: Extract from Dataset 1

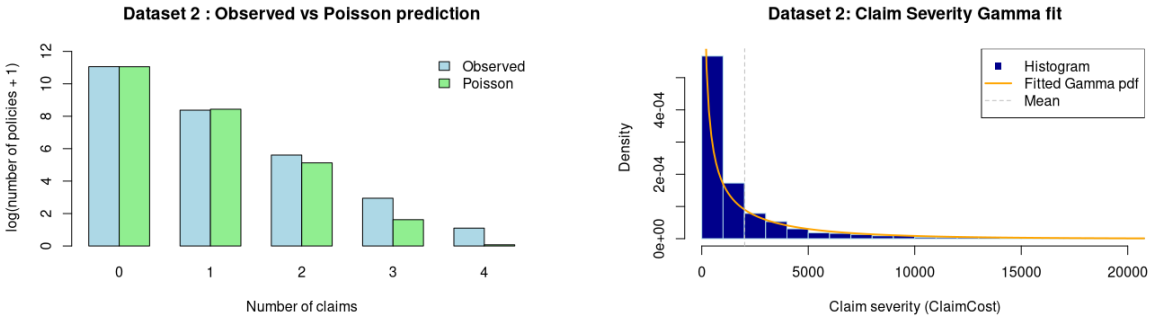
OwnerAge	Gender	Zone	VehAge	McClass	Bonus	Duration	NrofClaims	ClaimCost
33	M	2	2	3	5	0.748	0	0
26	M	3	17	4	7	0.118	0	0
23	M	4	0	2	6	0.384	0	0
50	M	2	3	4	6	0.164	0	0
64	M	3	23	1	2	1	0	0

3.1.2 Dataset 2: Australian Motor Case

This dataset consists of Australian motor third-party liability insurance data. It includes several related datasets, where the main dataset used in this study is `ausprivauto0405`, containing vehicle insurance policies from the years 2004 and 2005. This dataset is available in R after installing `CASdatasets`.

Descriptive statistics for the dataset are presented below. Furthermore, the empirical distributions of claim frequency and claim severity are examined in order to assess the suitability of standard actuarial assumptions. Similar to Dataset 1, although the data do not perfectly follow these distributions, they are considered reasonable approximations for modeling purposes. The rating factor `VehBody` originally consisted of several vehicle body categories as one can view in Table 4. To simplify the interpretation of the SHAP plots, similar vehicle types were grouped into four broader categories and encoded numerically from 1 to 4.

- The dataset contains 67,856 observations (policies) and all of them have positive exposure.
- Approximately 6.8% of the policies have at least one claim. The dataset is imbalanced, with a large proportion of zero-claim observations.
- 56.9% of the policyholders are female.
- The dataset includes categorical rating factors: vehicle value, vehicle type, vehicle age, gender and age of the driver.
- There is imbalance between categories: for example, there were a number of vehicle types that had few data points, for instance, there were 81 convertible while 22233 of Sedan. There are also imbalances in the category vehicle value, as there are few vehicles that has a high value.



(a) Claim frequency

(b) Claim severity

Figure 5: Empirical distributions for claim frequency and severity.

Table 4: Extract from Dataset 2

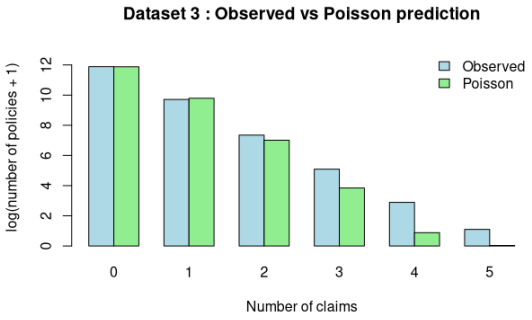
OwnerAge	VehValue	VehAge	VehBody	Gender	Duration	ClaimOcc	NrofClaims	ClaimCost
3	2.09	4	Station wagon	M	0.307	0	0	0
4	0.87	3	Utility	M	0.799	0	0	0
1	1.83	1	Sedan	F	0.602	0	0	0
4	1.27	4	Station wagon	M	0.345	1	1	1885
2	2.07	3	Utility	F	0.813	0	0	0

3.1.3 Dataset 3: Belgian Motor Case

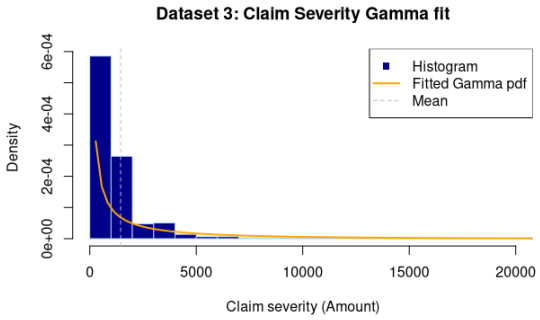
This dataset consists of motor third-party liability insurance data from a Belgian insurance portfolio. The dataset, `beMTPL97`, contains 163,212 observations, each associated with an exposure period representing the fraction of a year at risk. The dataset is available in R through the `CASdatasets` package. It includes information on claim frequency and severity, as well as several explanatory variables describing policyholders and vehicles.

Descriptive statistics are presented below, and the empirical distributions of claim frequency and severity are examined. Similar to the previous datasets, the data support the use of Poisson and Gamma models, although the assumptions are not perfectly satisfied.

- The dataset contains 163,212 observations, all with positive exposure.
- Approximately 11.2% of the observations have at least one claim. The dataset is imbalanced, with a large proportion of zero-claim observations.
- 26.5% of the policyholders are female.
- The dataset includes categorical rating factors: type of coverage, fuel, use, and fleet.
- There is imbalance between some categories; for example, few vehicles are older than 16 years and relatively few policyholders are aged 18–25. Additionally, only 4.8% of the observations correspond to non-private use.



(a) Claim frequency



(b) Claim severity

Figure 6: Empirical distributions for claim frequency and severity.

Table 5: Extract from Dataset 3

Coverage	OwnerAge	Gender	Bonus	Power	VehAge	Fuel	Use	Fleet	Expo	NrClaims	Amount
TPL	29	F	3	44	16	G	private	0	1	0	0
TPL	76	M	0	32	1	G	private	0	1	0	0
TPL	30	M	1	47	6	G	private	0	1	0	0
TPL++	38	F	4	33	8	G	private	0	1	0	0
TPL	24	M	9	66	3	D	private	0	1	1	145

3.2 Preprocessing

Before model training, the datasets are preprocessed to ensure consistency and suitability for modeling. Observations with zero exposure are removed, and when modeling claim severity, only observations with at least one claim are included. Categorical variables are encoded as factors for statistical models. For certain machine learning models and SHAP analyses, some categorical variables are transformed into numerical representations depending on the model requirements. Several variables, including bonus class and age variables, are also grouped into broader categories to reduce dimensionality. The datasets are highly imbalanced, with a large proportion of zero-claim observations. To account for this, stratification based on claim occurrence is applied when splitting the data into training and test sets using an 80/20 split. Sample weights are incorporated in the modeling process, where exposure is used for claim frequency and the number of claims is used for claim severity (see Table 1). Several datasets contain categorical rating factors with many levels. To improve interpretability and readability in the SHAP analyses, some categorical variables are aggregated into broader groups and subsequently encoded numerically. Further details regarding these transformations are provided in the Appendix.

3.3 Models

This study compares a traditional actuarial model with machine learning models for insurance pricing. The models are applied separately to claim frequency and claim severity. All models are trained using the same set of explanatory variables to ensure a fair comparison. All modeling and programming was conducted in the programming language R.
The following models are considered:

- Generalized Linear Models (GLM), implemented using the `stats` package
- Random Forest, implemented using `randomForest`

- Gradient Boosting Machines, implemented using `lightgbm`
- Neural Networks, using two layers, implemented using `keras` and `tensorflow`

3.4 Model Training

Models are trained separately for claim frequency and claim severity. Tree-based models and neural networks are trained using the same input variables as the GLM. Hyperparameters are tuned using nested cross-validation, where the inner loop is used for model selection and the outer loop is used to evaluate model performance. For gradient boosting, a Poisson objective is used for frequency and a Gamma objective for severity modeling.

3.4.1 Cross-Validation

A nested stratified K-fold cross-validation framework is applied for hyperparameter tuning and model evaluation. Before cross-validation, the data are divided into training and test sets using an 80/20 split, where the test set is held out for final evaluation. Since the insurance datasets are highly imbalanced, stratification is applied based on the presence of claims to preserve the proportion of claim and non-claim observations across the splits. Within the training set, K-fold cross-validation with $K = 5$ is used. The inner loop is applied for hyperparameter tuning and model selection, while the outer loop is used to estimate the generalization performance of the selected model. This separation reduces the risk of optimistic performance estimates, since the data used for tuning are not reused for evaluation.

3.4.2 Hyperparameter Tuning

Hyperparameters were tuned using nested cross-validation. The following hyperparameters were considered for each model:

- **GLM**: Different combinations of explanatory variables and interaction terms were evaluated using a grid search within the nested CV framework. The selected interaction terms and their contribution to model performance were later used as an indication of the importance of interaction effects in the datasets.
- **Random Forest**: Number of trees (`ntree`), number of variables considered at each split (`mtry`), and minimum node size (`nodesize`).
- **Gradient Boosting Machine**: Shrinkage (`learning_rate`), number of leaves (`num_leaves`), minimum number of observations in each leaf (`min_data_in_leaf`), proportion of observations randomly sampled (`subsample`), and proportion of explanatory variables randomly sampled (`colsample_bytree`).
- **Neural Network**: Number of neurons in the first and second hidden layers (`units1`, `units2`), dropout rate (`dropout`), learning rate (`lr`), number of training epochs (`epochs`), and batch size (`batch_size`).

3.5 Model Evaluation

Model performance is evaluated using deviance-based loss functions. The deviance is normalized by the sum of exposure for claim frequency and by the total number of claims for claim severity.
For claim frequency, Poisson deviance is used:

$$D = 2 \sum_i w_i \left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right)$$

For claim severity, Gamma deviance is applied, the formula can be found in Section 2.4.

To evaluate model performance relative to a baseline, the null model, i.e. a model containing only an intercept term, is used as a reference.

The improvement of a model is measured as the reduction in deviance compared to the null model, defined as:

$$\text{Relative improvement} = 1 - \frac{D_{\text{model}}}{D_{\text{null}}}$$

This measure can be interpreted as the proportion of deviance explained by the model and allows for a consistent comparison across different modeling approaches.

Research Question 2: How can these models be interpreted?

This section presents the methodology used to interpret the fitted models. The focus is on understanding how individual features influence the predicted claim frequency and how these effects differ across modeling approaches.

Generalized Linear Model

For the Generalized Linear Model, interpretation is based on the model structure. Since the model is additive on the log scale, the contribution of each feature is determined by its estimated coefficient. Global feature importance was measured using the mean absolute SHAP value. This measure represents the average magnitude of the contribution of each variable to the model prediction across all observations. So, this means which variables are affecting the prediction the most.

Random Forest

For the Random Forest model, feature importance is evaluated using permutation importance, which measures the increase in prediction error when a variable is randomly permuted. This is included in the RandomForest function in R. To further interpret the model, SHAP values are used to analyse how each feature contributes to individual predictions.

Gradient Boosting Machine and Neural Networks

For GBM and NN, interpretability is achieved using SHAP values which provide a local explanation of each prediction by decomposing it into contributions from each feature. Global feature importance is obtained by averaging the absolute SHAP values across all observations.

4 Result

In this section, the final results are presented. For each research question and for each dataset the model performance will be reported and the figures concerning interpretability will be presented.

4.1 Research Question 1: Which machine learning method performs the best using deviance as a loss function?

The results concerning model performance will be presented in this section, divided for each dataset and if it concerns claim frequency or severity.

4.1.1 Dataset 1: Motorcycle Case

Table 6: Model performance for claim frequency for Dataset 1, compared with the null model

Method	**Frequency (Poisson Dev.)**	**Relative performance**
Null Model	0.1004	-
GBM	0.08633	+13.99%
NN	0.08697	+13.35%
GLM	0.08779	+12.53%
RF	0.08916	+11.17%

Table 7: Model performance for claim severity for Dataset 1, compared with the null model

Method	**Frequency (Gamma Dev.)**	**Relative performance**
Null Model	1.7837	-
GBM	1.5203	+14.76%
RF	1.5437	+13.46%
GLM	1.5905	+10.83%
NN	1.7742	+0.53%

Looking at Table 6 all methods performs better than the null model. Gradient boosting achieves the lowest deviance for claim frequency, closely followed by the neural network. The GLM performs competitively, with only a small increase in deviance compared to the best-performing model. The random forest yields the highest deviance among the considered methods. When looking at claim severity in Table 7, GBM performs the best of the methods, while NN performs the worst, with a significant difference.

4.1.2 Dataset 2: Australian Motor Case

Table 8: Model performance for claim frequency for Dataset 2, compared with the null model

Method	**Frequency (Poisson Dev.)**	**Relative performance**
Null Model	0.7674	-
GLM	0.7645	+0.38%
RF	0.7652	+0.29%
GBM	0.7653	+0.27%
NN	0.7678	-0.052%

Table 9: Model performance for claim severity for Dataset 2, compared with the null model

Method	**Frequency (Gamma Dev.)**	**Relative performance**
Null Model	1.5348	-
GBM	1.5162	+1.21 %
RF	1.5195	+0.10 %
GLM	1.5334	+0.091 %
NN	1.5903	-3.62 %

The interaction analysis presented in Section 3.4.2 indicates fewer strong interaction effects in Dataset 2 than in Dataset 1. Considering claim frequency, in Table 8, no machine learning method outperforms the GLM, although the differences in performance are relatively small. The neural network performs slightly worse than the null model, indicating limited predictive improvement for this dataset. For claim severity, the GBM and Random Forest models achieve the best performance, while the GLM performs only marginally better than the null model. The neural network again performs worse than the null model. Overall, the results suggest that the additional flexibility of machine learning methods provides very limited improvement for Dataset 2.

4.1.3 Dataset 3: Belgian Motor Case

Table 10: Model performance for claim frequency for Dataset 3, compared with the null model

Method	**Frequency (Poisson Dev.)**	**Relative performance**
Null Model	0.6205	-
GBM	0.5997	+3.35 %
NN	0.5998	+3.34 %
GLM	0.6002	+3.27 %
RF	0.6043	+2.61 %

Table 11: Model performance for claim severity for Dataset 3, compared with the null model

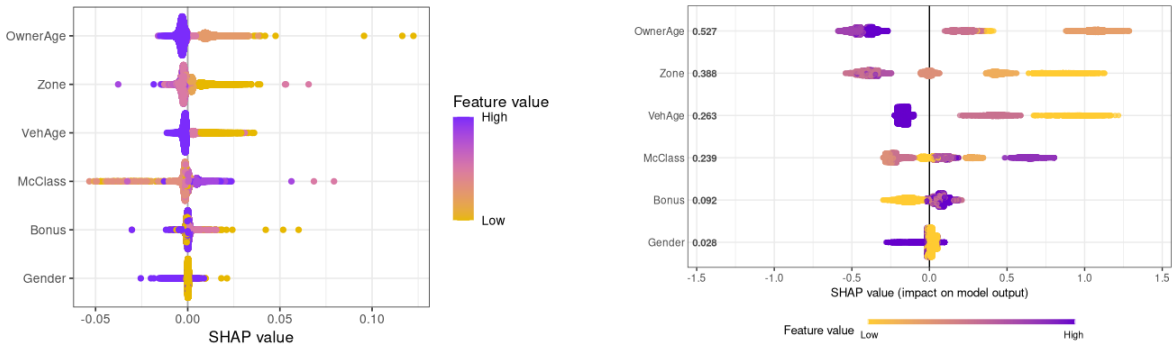
Method	Frequency (Gamma Dev.)	Relative performance
Null Model	2.001	-
GLM	1.9751	+1.29 %
GBM	1.9801	+1.05 %
RF	1.9829	+0.90 %
NN	2.1386	-6.88 %

Similar to Dataset 2, there seems to be few interactions in this dataset. Regarding claim frequency in Table 10, all methods performs better than the null model. When it comes to claim severity the GLM outperforms the ML methods, where NN stands out being - 6.88 % worse than the null model looking at Table 11.

4.2 Research Question 2: How can these models be interpreted?

This section presents SHAP analyses for the machine learning models across all datasets. The analyses are conducted for both claim frequency and claim severity to identify important rating factors and investigate how different variables influence the predictions.

4.2.1 Dataset 1: Motorcycle Case



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

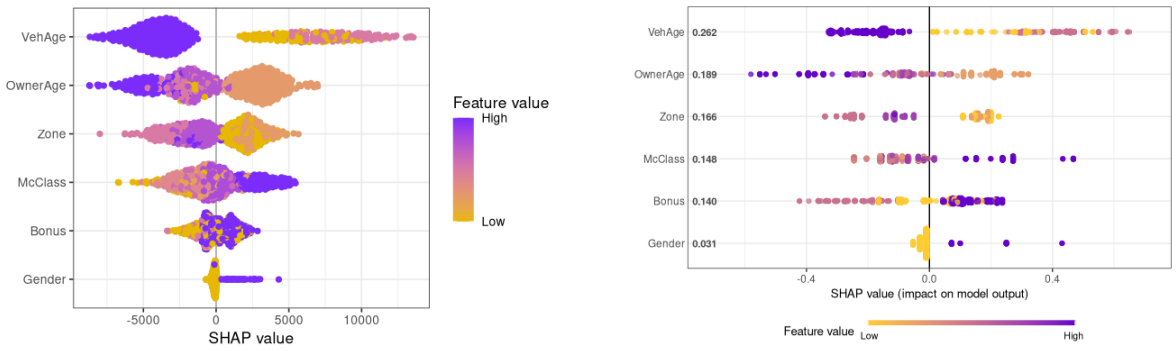
Figure 7: SHAP-analysis for RF and GBM-methods concerning claim frequency

Figure 8: SHAP-analysis for the NN-method concerning claim frequency

Table 12: Most important variables based on mean absolute SHAP values for claim frequency

Rank	GLM	RF	GBM	NN
1	Zone	OwnerAge	OwnerAge	OwnerAge
2	VehAge	Zone	Zone	Zone
3	McClass	VehAge	VehAge	VehAge
4	Gender:OwnerAge	McClass	McClass	McClass
5	Bonus	Bonus	Bonus	Bonus

The methods consider similar variables most important, looking at Table 12. Concerning the SHAP analysis in Figure 7 and 8, the machine learning models have most things in common. They all consider younger owners more likely to have higher claim frequency than older owners, also owners living in the bigger cities and has a higher motorcycle class are more likely to have higher claim frequency. Concerning bonus and gender, it is more of a cluster and more difficult to assess clear patterns.



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

Figure 9: SHAP-analysis for RF and GBM-methods concerning claim severity

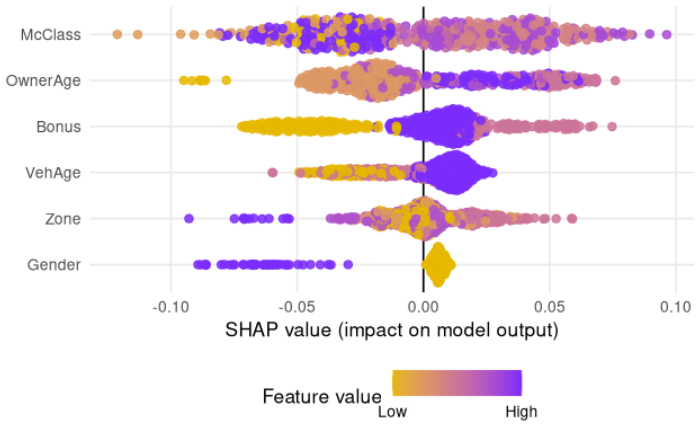


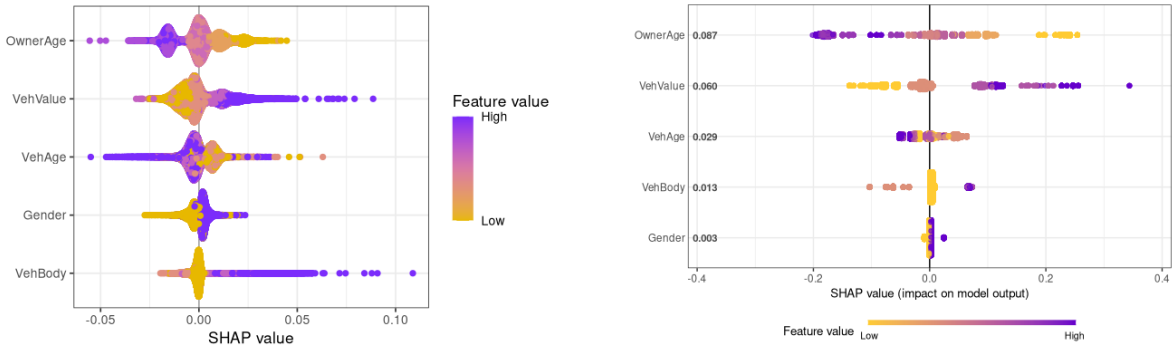
Figure 10: SHAP-analysis for the NN-method concerning claim severity

Table 13: Most important variables for claim severity in Dataset 1

Rank	GLM	RF	GBM	NN
1	VehAge	VehAge	VehAge	McClass
2	Gender:OwnerAge	OwnerAge	OwnerAge	OwnerAge
3	Zone	Zone	Zone	Bonus
4	McClass	McClass	McClass	VehAge
5	Bonus	Bonus	Bonus	Zone

Concerning claim severity, all methods expect NN has the similar rank in variable importance, which can be viewed in Table 13. Where NN considered motorcycle class the most important while the other methods considered vehicle age, owner age and geographical zone the most important. RF and GBM have the same variable rank and found similar patterns in the SHAP analysis, in Figure 9. For example, both methods found that newer vehicles and younger drivers are associated with higher claim severity. The SHAP-analysis for NN in Figure 10 showed more clustered patterns and occasionally produced relationships that differed from those identified by the other methods. For example, NN suggested that younger drivers are associated with lower claim severity.

4.2.2 Dataset 2: Australian Motor Case



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

Figure 11: SHAP-analysis for RF and GBM-methods concerning claim frequency

Figure 12: SHAP-analysis for the NN-method concerning claim frequency

Table 14: Most important variables based on mean absolute SHAP values for claim frequency

Rank	GLM	RF	GBM	NN
1	OwnerAge	OwnerAge	OwnerAge	OwnerAge
2	VehValue	VehValue	VehValue	VehValue
3	VehBody	VehAge	VehAge	VehAge
4	Gender	Gender	VehBody	VehBody
5	VehAge	VehBody	Gender	Gender

Looking at Table 14, all machine learning methods identify the same three variables as the most important rating factors. The GLM differs slightly by assigning lower importance to vehicle age compared to the machine learning methods. The SHAP analyses in Figures 11 and 12 reveal similar overall patterns across the models. However, the effect of gender appears relatively weak and inconsistent. For clarity, men are encoded as 0 and women as 1. Both RF and GBM associate female policyholders with slightly higher predicted claim frequencies, although the effect is limited.

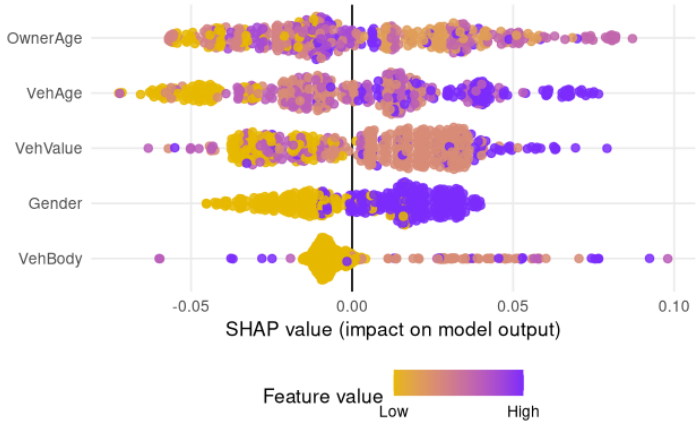
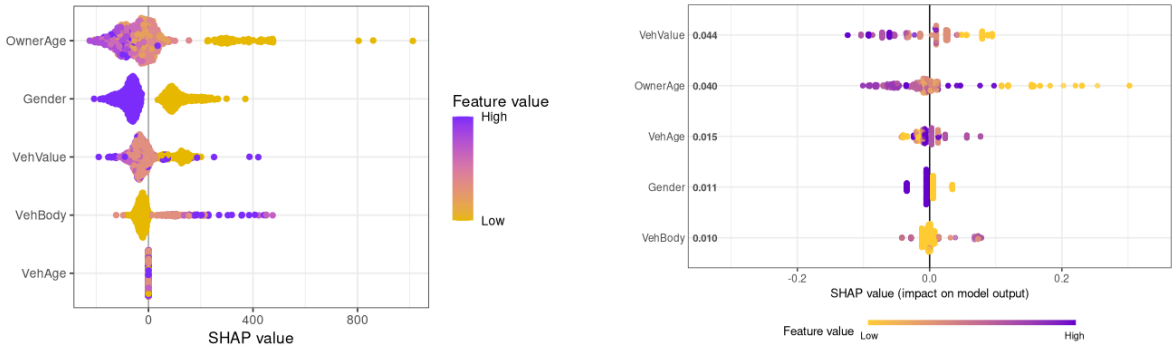


Figure 14: SHAP-analysis for the NN-method concerning claim severity



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

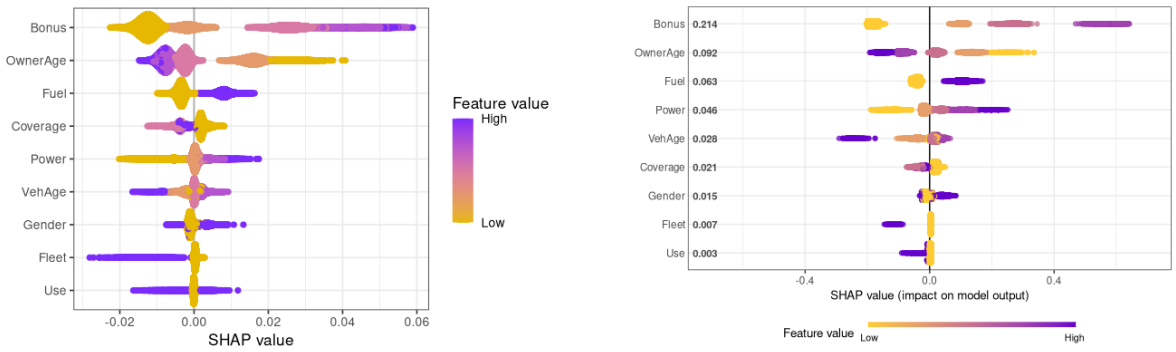
Figure 13: SHAP-analysis for RF and GBM-methods concerning claim severity

Table 15: Most important variables for claim severity in Dataset 2

Rank	GLM	RF	GBM	NN
1	OwnerAge	OwnerAge	VehValue	OwnerAge
2	Gender	Gender	OwnerAge	VehAge
3	-	VehValue	VehAge	VehValue
4	-	VehBody	Gender	Gender
5	-	VehAge	VehBody	VehBody

For claim severity, the variable importance rankings differ more between the models, as shown in Table 15. The SHAP analyses in Figures 13 and 14 show that RF and GBM identify similar overall patterns. GBM indicates a clearer relationship between vehicle value and predicted claim severity, while RF suggests a more nonlinear relationship. RF and GBM also identify similar patterns for owner age, whereas the neural network displays less consistent and less interpretable SHAP patterns across several variables.

4.2.3 Dataset 3: Belgian Motor Case



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

Figure 15: SHAP-analysis for RF and GBM-methods concerning claim frequency

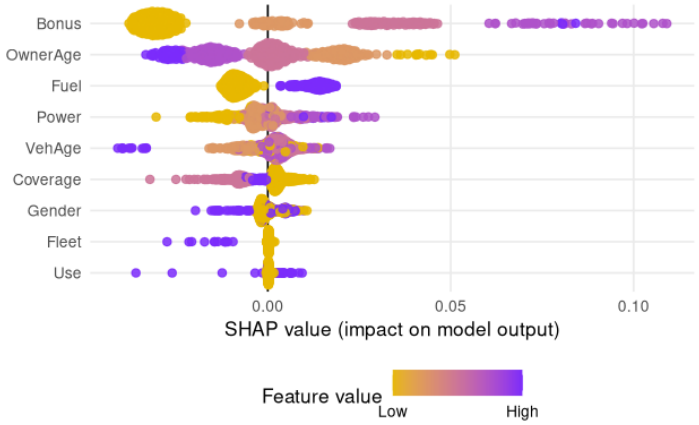


Figure 16: SHAP-analysis for the NN-method concerning claim frequency

Table 16: Most important variables based on mean absolute SHAP values for claim frequency

Rank	GLM	RF	GBM	NN
1	Bonus	Bonus	Bonus	Bonus
2	OwnerAge	OwnerAge	OwnerAge	OwnerAge
3	Power	Fuel	Fuel	Fuel
4	Fuel	Coverage	Power	Power
5	VehAge	Power	VehAge	VehAge

All methods considered similar variables important, which is seen in Table 16. The machine learning methods also found the same patterns, looking at the SHAP analysis in Figure 15 and 16.

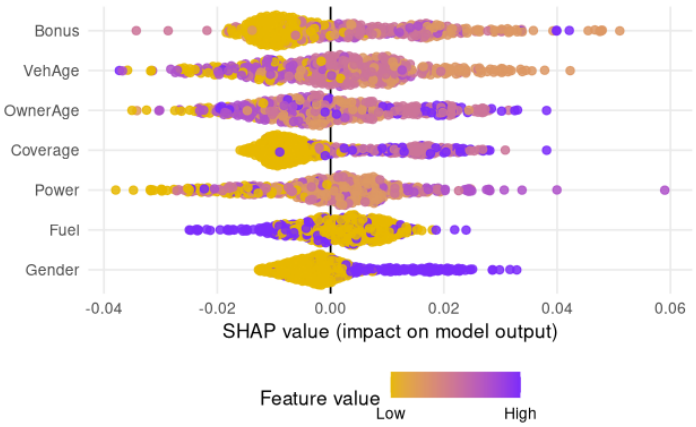
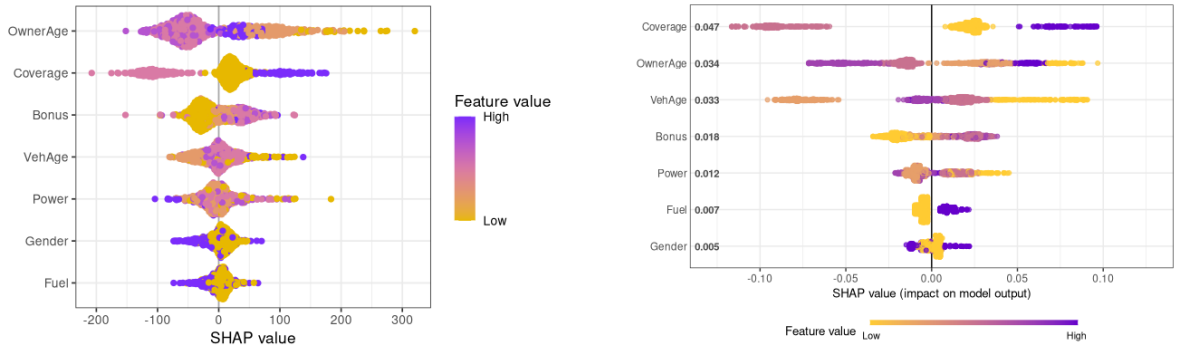


Figure 18: SHAP-analysis for the NN-method concerning claim severity



(a) Random Forest (RF)

(b) Gradient Boosting Machine (GBM)

Figure 17: SHAP-analysis for RF and GBM-methods concerning claim severity

Table 17: Most important variables for claim severity in Dataset 3

Rank	GLM	RF	GBM	NN
1	OwnerAge	OwnerAge	Coverage	Bonus
2	Coverage	Coverage	OwnerAge	VehAge
3	Power	Bonus	VehAge	OwnerAge
4	-	VehAge	Bonus	Coverage
5	-	Power	Gender	Power

When analyzing the severity, the GLM, RF and GBM were considered the similar variables important and similar rank, looking in Table 17. NN stood out in considering Bonus the most important variable. RF and GBM found the same patterns in the SHAP analysis except in the category Gender, which is notable in Figure 17. However, the NN exhibits more dispersed clusters and less distinct patterns in the SHAP analysis, making the relationships more difficult to interpret, as can be viewed in Figure 18.

5 Discussion

In this section the results will be evaluated, divided in research question.

5.1 Research Question 1: Which machine learning method performs the best using deviance as a loss function?

Considering the claim frequency results, there are notable differences between datasets with strong interaction effects and datasets where interactions appear to be less important. Dataset 1 exhibits the largest relative improvement over the null model. A likely explanation is that the machine learning methods are better able to capture nonlinear relationships and interactions between variables. GBM performs the best in two of three datasets while NN is generally the second-best performing method and RF performs the worst of the methods. Dataset 2 is the more unique in this case, where the GLM outperforms the machine learning methods, but with a small percentage. An explanation to this result, would be that this data lacks interactions all together which means that the machine learning methods do not contribute to more information. Dataset 3 has the same rank as dataset 1 looking at the methods in Tables 6 and 10.

Something that all datasets has in common concerning claim severity, is that NN performs worse than the null model. The most probable cause of that is that NN demands more observations to find patterns. As mentioned, there are imbalances in insurance data as there are few claims. Among the machine learning methods, GBM consistently achieves the best performance for claim severity. The GLM only outperforms the machine learning methods in Dataset 3, as shown in Table 11. A possible explanation is that the underlying structure in this dataset is more additive and therefore better aligned with the assumptions of the GLM. In contrast, Dataset 1, the motorcycle insurance dataset, shows the largest relative improvement from machine learning methods. Similar to the claim frequency results, this likely indicates the presence of stronger interaction effects and more complex nonlinear relationships which the ML methods capture easier.

5.2 Research Question 2: How can these models be interpreted?

The interpretability results for the three datasets are discussed separately below.

5.2.1 Dataset 1: Motorcycle Case

Looking at claim frequency, the machine learning methods revealed largely consistent patterns in the SHAP analyses and generally agreed on the most important risk factors. This makes sense when all methods are considered better than the null model (Table 6), with relatively little difference looking at relative performance. Even though, RF performed the worst of them all is was still much better the null model. Something that differed between the methods in the SHAP-analysis for GBM in Figure 7 was that is looks like GBM use `Bonus` more than the other variables as it is less of a cluster. It could be a possible explanation.

Concerning claim severity, NN performs only slightly better than the null model and the other models performs 10 % better. This is connected to the difference between NN and the other models in the SHAP-analysis and the difference in variable importance in Table 13, where the difference is significant. Additionally, looking at the SHAP-analysis in Figures 9 and 10 where GBM and RF has pretty much the same patterns where NN goes the opposite in some factors. For instance, NN consider younger drivers to have less severity while GBM and RF considers the opposite. NN's inferiority of the other methods is very likely to be caused by the method itself, as the dataset used when modeling severity is just a fraction of the training data as we only consider claims when modeling severity. Generally, NN needs more observations.

5.2.2 Dataset 2: Australian Motor Case

Regarding the claim frequency for dataset 2, all the methods considered ranked the same two variables top, noticeable in Table 14. All the models was almost the same relative performance as the null model. This leads to the conclusion the dataset contains relatively weak relationships between the rating factors and claim frequency, as well as limited interaction effects. Furthermore, GLM outperforms the other methods, but only with a small percentage. The ML methods agrees on the point that younger drivers, expensive, newer and sport cars had a higher claim frequency. The only category that was diffuse was if the driver's gender mattered for the frequency of the claim.

The machine learning methods nevertheless identify several similar patterns in the SHAP analyses. In particular, younger drivers, expensive vehicles, newer vehicles, and sports cars are generally associated with higher predicted claim frequency. The effect of gender is less consistent across the models, indicating that this variable has relatively weak predictive influence for claim frequency in Dataset 2. When it comes to claim severity, all methods except NN performed better than the null model. GBM and RF performed approximately 1 % improvement, while NN performed approximately 4 % worse, as shown in Table 9. The reason for NNs negative outcome is probably that NN demands more observations to find patterns. The SHAP analyses in Figure 13 show that RF and GBM identify several similar patterns. Both methods associate younger male drivers and sports cars with higher predicted claim severity. RF suggests that both low-value and high-value vehicles are associated with increased predicted severity, whereas GBM mainly associates lower vehicle values with higher severity. In addition, GBM identifies a clearer relationship between vehicle age and claim severity, where older vehicles are associated with higher predicted severity, while RF shows less distinct patterns for vehicle age. The neural network exhibits substantially more dispersed SHAP patterns in Figure 14, making the relationships more difficult to interpret. The only clearly identifiable pattern is that sports cars are associated with higher predicted severity, which is consistent with the RF and GBM models. Otherwise, the NN displays patterns that differ from the other ML methods.

5.2.3 Dataset 3: Belgian Motor Case

Looking at variable importance in Table 16, the methods consider roughly the same variables important and the SHAP analysis in Figures 7 and 8 agrees on overall the same patterns. All the methods outperform the null model and the difference is relatively small, as shown in Table 10 and the ranking was similar to that observed for Dataset 1.

Regarding claim severity, `Use` and `Fleet` provided limited predictive information and were therefore excluded from the SHAP analyses presented in Figures 17 and 18. GBM and RF found similar patterns, including younger drivers having more severe claims. he neural network, however, exhibits more dispersed SHAP patterns and less agreement with the other machine learning methods. Consistent with the previous datasets, the NN also performs worse than the null model, while the GLM achieves the best overall performance for claim severity, as shown in Table 11.

5.3 Previous Research

The findings of this study are broadly consistent with previous research in non-life insurance pricing. Generalized Linear Models remain the standard approach in insurance pricing due to their transparency and interpretability [1, pp. 15–16].

The strong performance of GBM observed in this thesis is consistent with earlier findings. Johansson and Ohlsson [5, p. 21] note that boosting methods are among the preferred machine learning approaches in non-life insurance pricing, and Henckaerts et al. [9, p. 255] found that boosted trees outperform traditional GLMs in motor insurance pricing. Similarly, GBM achieved the best overall predictive performance in most datasets considered in this study. The larger improvements observed in Dataset

1 are consistent with previous research suggesting that machine learning methods provide the greatest benefits when nonlinear relationships and interactions are present in the data [2].

5.4 Conclusions

Machine Learning methods can improve predicative performance, as seen in our results for dataset 1 and dataset 3; however, hardly no improvement were observed for dataset 2. GBM was the method that achieved the highest predictive performance across most of the datasets and modeling tasks, as shown in the results presented in Chapter 4. The traditional method GLM still is not outrun by the ML methods and performs well across all three datasets, despite its simplicity. Regarding claim severity, NN was not as suitable for any of the three datasets, a likely explanation for this is that NN needs more data in general.

6 References

[1] Björn Johansson and Esbjörn Ohlsson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, 2010. DOI: 10.1007/978-3-642-10791-7.

[2] Charles Blier-Wong et al. "Machine Learning in P&C Insurance: A Review for Pricing and Reserving". In: *Risks* 9.1 (2021), pp. 1–26. DOI: 10.3390/risks9010004.

[3] Svensk Försäkring. *Försäkringslagstiftning*. n.d. URL: https://www.svenskforsakring.se/omforsakring/foraskringslagstiftning/ (visited on 03/02/2026).

[4] Pantelis Linardatos, Vasileios Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2020), p. 18. DOI: 10.3390/e23010018.

[5] Björn Johansson and Esbjörn Ohlsson. *Gradient Boosting Machines and Non-Life Insurance Pricing: Lecture Notes*. SSRN Working Paper. 2022. DOI: 10.2139/ssrn.4294965.

[6] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. 2nd ed. New York: Springer, 2021. ISBN: 978-1071614174.

[7] Mario V. Wüthrich and Michael Merz. *Statistical Foundations of Actuarial Learning and its Applications*. Springer, 2023. DOI: 10.1007/978-3-031-12409-9.

[8] Eduardo Blancas. *Model selection done right: A gentle introduction to nested cross-validation*. 2022. URL: https://ploomber.io/blog/nested-cv/.

[9] Christophe Henckaerts et al. "Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods". In: *North American Actuarial Journal* 25.2 (2020), pp. 255–285. DOI: 10.1080/10920277.2020.1745656.

7 Appendix

7.1 Dataset 1: Motorcycle Case

Description of the variables in Table 3. The information is based on [1, p. 35].

- **OwnerAge**: The age of the policyholder, ranging from 0 to 99 years.
- **Gender**: The gender of the policyholder, with categories M (male) and F (female).
- **Zone**: Geographical zone (1–7) based on a standard classification of Swedish parishes, where Zone 1 represents central urban areas, Zones 2–4 correspond to suburban and smaller urban areas, Zones 5–6 represent northern regions, and Zone 7 corresponds to Gotland.
- **McClass**: Motorcycle class, defined based on the EV ratio:

$$\frac{\text{Engine power (kW)} \cdot 100}{\text{Vehicle weight (kg)} + 75},$$

  where 75 kg represents the average driver's weight.
- **VehAge**: The age of the vehicle, ranging from 0 to 99 years.
- **Bonus**: Bonus class, taking values from 1 to 7. A new driver starts in class 1, and the class increases by one for each claim-free year. After a claim, the bonus class is reduced by two levels. Returning to class 7 requires at least six consecutive claim-free years.

7.2 Dataset 2: Australian Motor Case

Description of the variables in Table 4. The information is based on the `CASpackage` in R.

- **OwnerAge**: The age of the policyholder, divided into five categories between youngest and oldest people.
- **VehValue**: The vehicle value in thousand of AUD.
- **VehBody**: The vehicle body group, originally divided into 13 categories. For the SHAP analysis, similar vehicle body types were aggregated into four broader categories (Standard, Utility, Passenger/Other, Sport) and encoded numerically from 1 to 4 to improve readability and interpretation.
- **VehAge**: The vehicle age group divided in four categories between youngest and oldest cars.
- **Gender**: The gender of the policyholder, in this case man (0) or women (1).

7.3 Dataset 3: Belgian Motor Case

Description of the variables in Table 5. The information is based on the `CASpackage` in R.

- **Coverage**: Describes the insurance level, where "TPL" means third party liability, "TPL+" stands for limited material damage and "TPL++" means comprehensive material damage. In the SHAP-analysis this is coded $1 - 3$, with three being the better coverage.
- **OwnerAge**: The age of the policyholder.
- **Gender**: The gender of the policyholder, in this case man (0) or woman (1).

- **Bonus**: Level in the former compulsory Belgian bonus-malus scale. From $0-22$, where a higher level indicates a worse claim history.
- **Power**: Horsepower of the vehicle in kilowatt.
- **VehAge**: The age of the vehicle.
- **Fuel**: Type of fuel for the vehicle, gasoline (0) or diesel (1).
- **Use**: The use of the vehicle, private (0) or work (1).
- **Fleet**: Integer 1 or 0 indicating if the vehicle is a part of a fleet, which means if a vehicle is managed by a company.
- **Expo**: Duration in years.
- **NrClaims**: The number of claims.
- **Amount**: Claim amount.