Hochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

Seminar Project

in the
Department of Mathematical Finance, Actuarial Science and
Risk Management
Faculty 4

XGBoost in Non-Life Insurance Pricing

by

Hans Alcahya Tjong

Supervisor: Dr. Alla Petukhina

Contents

1 Motivation: Fundamentals of Insurance Pricing

An essential component of non-life insurance is *pricing*. Pricing refers to the process of determining the insurance premium, which is the amount of money paid to the insurer to obtain insurance coverage for risks. The introduction to pricing in this subsection is based on the book *Non-Life Insurance Pricing with Generalized Linear Models* by Ohlsson and Johansson [OJ2010].

1.1 Rating Factors

The premium for each policy depends on rating factors. The rating factors can be categorized into three main characteristics:

- Characteristics of the policyholder: gender, age, marital status;
- Characteristics of the insured objects: car manufacturer, engine size, model year; and
- Characteristics of the geographical region: city and per capita income.

A rating factor can be categorical or continuous. Gender is an example of a categorical rating factor. An example of a continuous rating factor is age, before it is grouped into intervals. A rating factor typically consists of variables that are already classified into categories. If a rating factor is not yet categorical, it should be transformed; for example, age is grouped into age classes because a linear relationship rarely exists between them [cf. OJ2010, p. 3].

Example 1.1 (Automobile UK Collision Claims)**.** This example refers to the dataset `AutoCollision`, which is available in the `R` package `insuranceData`[1]. The rating factors in the dataset include, for instance, driver age and vehicle usage. In this example, the driver age is categorical, meaning that the rating factor is already classified into categories (from class `A` to class `H`). The other rating factor, vehicle usage, is divided into four categories: `Business`, `DriveLong`, `DriveShort`, and `Pleasure`. The class `Business` refers to business use, `Pleasure` refers to private use, `DriveLong` indicates commuting more than ten miles to work, and `DriveShort` means commuting less than ten miles to work.

Rating Factor	Class
Driver Age	`A`
	`B`
	`C`
	`D`
	`E`
	`F`
	`G`
	`H`
Vehicle Usage	`Business`
	`DriveLong`
	`DriveShort`
	`Pleasure`

Table 1.1: Rating factors in Automobile UK Collision Claims (own representation)

The following section explains key ratio used in a tariff (a detailed explanation is provided by Goelden et al. [Goe2016, pp. 120 - 126]).

Duration n_o refers to the total duration for which individual policies provide insurance coverage.

Claim number N represents the number of claims reported.

[1] `https://cran.r-project.org/web/packages/insuranceData/index.html`[31.01.2025]

Claim frequency CF is calculated as the number of claims divided by the duration.

Total claim amount S represents the sum of all claim amounts.

Claim severity CS is the total claim amount divided by the number of claims. It also reflects the average cost per claim.

Pure premium SB is defined as the total claim amount divided by the duration.

Duration and claim number are exposures. An exposure defines the unit in which the size of risk is measured [cf. Bus2015, p. 12]. Other examples of exposures include the number of policies, the number of risks, the cumulated sum insured, and the total premium amount [cf. Goe2016, p. 122]. The following table summarizes the relationships between key ratios:

Exposure $[\omega]$	Target Variable $[X]$	Normalized Target Variable $\left[Y = \frac{X}{\omega}\right]$
Duration n_o	Claim number N	Claim frequency $CF = \frac{N}{n_o}$
Claim number N	Total claim amount S	Claim severity $CS = \frac{S}{N}$
Duration n_o	Total claim amount S	Pure premium $SB = \frac{S}{n_o}$

Table 1.2: Key claim-related figures (own representation)

By multiplying the claim severity and claim frequency, the pure premium can also be derived. Thus, the following equation holds:

$$SB := \frac{S}{n_0} = \frac{S}{N} \cdot \frac{N}{n_0} = CS \cdot CF. \tag{1.1}$$

1.2 Multiplicative Model

The premium can be determined using either an additive model or a multiplicative model. However, this work considers only the multiplicative model. Let M be the

number of rating factors, m_k the number of classes of rating factor k, and i_k denote the class of rating factor k. In Example 1.1, it holds that $M = 2$, $m_1 = 8$, and $m_2 = 4$. The following table illustrates another example.

Age	Vehicle Usage	i_1	i_2
21-40	Business	1	1
21-40	Private	1	2
41-60	Business	2	1
41-60	Private	2	2
61-80	Business	3	1
61-80	Private	3	2

Table 1.3: Illustrative example of a tariff (own representation)

Two rating factors are given, namely *Age* and *Vehicle Usage*. This implies that $M = 2$. The first rating factor (Age) consists of three classes: 21-40, 41-60, and 61-80, meaning that $m_1 = 3$ and $i_1 \in \{1, 2, 3\}$. The second rating factor (Vehicle Usage) has two classes: business and private. Thus, $m_2 = 2$ and $i_2 \in \{1, 2\}$. The table shows that there are six tariff cells (i_1, i_2) formed from the combination of the two rating factor classes, e.g., tariff cell $(1, 1)$.

The multiplicative model for this example is given by:

$$\mu_{i_1,i_2} = \gamma_0 \gamma_{1,i_1} \gamma_{2,i_2}, \tag{1.2}$$

where γ_0 is the base value and γ_{k,i_k} for $k \in \{1, 2\}$ represents the price relativity. If a tariff cell (i_1, i_2) is chosen as the base, then the terms γ_{1,i_1} and γ_{2,i_2} are set to 1, so that γ_0 represents the base value. The base value is preferably chosen such that the corresponding tariff cell has a large exposure. The price relativity describes the relative difference of a tariff cell compared to the base.

Example 1.2 (Multiplicative Model)**.** Consider Table 1.3. Let the tariff cell $(1, 1)$ be the base cell, and for the normalized target variable, the pure premium is selected. Then, $\mu_{1,1} = \gamma_0$ holds. This means that the pure premium for the age group between 21 and 40 years with business vehicle usage corresponds to the base

value γ_0. If the price relativity $\gamma_{1,2} = 1.3$ is given, then according to equation (1.2):

$$\begin{aligned}\mu_{2,1} &= \gamma_0\gamma_{1,2}\gamma_{2,1} \\ &= \gamma_0 \cdot 1.3 \cdot 1 \\ &= 1.3\gamma_0.\end{aligned}$$

This implies that the pure premium for the age group between 41 and 60 years with the same vehicle usage is 30% higher than for the age group between 21 and 40 years. The following table summarizes the multiplicative model for all tariff cells.

Age	Vehicle Usage	i_1	i_2	μ_{i_1,i_2}
21-40	Business	1	1	γ_0
21-40	Private	1	2	$\gamma_0\gamma_{2,2}$
41-60	Business	2	1	$\gamma_0\gamma_{1,2}$
41-60	Private	2	2	$\gamma_0\gamma_{1,2}\gamma_{2,2}$
61-80	Business	3	1	$\gamma_0\gamma_{1,3}$
61-80	Private	3	2	$\gamma_0\gamma_{1,3}\gamma_{2,2}$

Table 1.4: Continuation of Table 1.3 with μ_{i_1,i_2} (own representation)

Equation (1.2) can be extended for any M. Thus, the following holds:

$$\mu_{i_1,i_2,\ldots,i_M} = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i_k}. \tag{1.3}$$

2 Methods

In this chapter, the methods applied in this study are explained. It is divided into two parts, starting with the generalized linear models in Chapter 2.1. Subsequently, the extreme gradient boosting are explained in Chapter 2.2. This chapter is based on Ohlsson and Johansson [OJ2010], Denuit et al. [DHT2019], and Tianqi et al. [CG2016].

2.1 Generalized Linear Models (GLM)

In linear regression models, it is assumed that the error terms $\boldsymbol{\epsilon}$ are normally distributed. Generalized Linear Models (GLM) differ from linear regression models in this assumption. In GLM, the response variable Y_i can follow a different distribution from the exponential family (Exponential Dispersion Family (EDF)), such as the Poisson, Binomial, or Gamma distribution.

Furthermore, Ohlsson and Johansson [cf. OJ2010, p. 15] point out that linear regression models are not entirely suitable for pricing. The reason is that in linear models, the expected value is a linear function of the explanatory variables $\mathbf{X}$, whereas in pricing, the multiplicative model is commonly used.

GLM were introduced by Nelder and Wedderburn [NW1972] in 1972. This method consists of three components:

- Random Component:
  The response variable Y_i follows a distribution from the exponential family,
- Systematic Component:

The explanatory variables result in a linear predictor η_i,

$$\eta_i = X_i^T \beta = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$,

- Link Function:
  The bijective and twice continuously differentiable link function $g(\cdot)$ connects the random and systematic component:

$$g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i) \text{ with } \mathbb{E}(Y_i|X) = \mu_i.$$

2.1.1 Random Component

The random component of GLM describes the probability function of the response variable Y_i, if discrete. If Y_i is continuous, then it refers to a density function. It is assumed that Y_i follows a distribution from the exponential family. The following definition is based on the definitions from Denuit et al. [DHT2019, pp. 47] and Ohlsson-Johansson [OJ2010, pp. 17].

Definition 2.1 (Exponential Family)**.** Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables. Y_i follows a distribution from the exponential family for $i = 1, \ldots, n$, if Y_i satisfies a probability function or a density function of the following form:

$$f_{Y_i}(y_i; \theta_i, \phi) = exp\left(\frac{y_i\theta_i - a(\theta_i)}{\phi/\nu_i}\right) c(y_i, \phi, \nu_i), \tag{2.1}$$

where:

y_i : Realization of the random variable Y_i

θ_i : Canonical parameter (real-valued)

ϕ : Dispersion parameter (positive)

ν_i : Weight (positive)

$a(\cdot)$: Measurable cumulant function (bijective and twice continuously differentiable)

$c(\cdot)$: Positive measurable normalizing function.

The expectation and variance of function (2.1) can be described as follows:

$$\mathbb{E}(Y_i) = a'(\theta_i) \tag{2.2}$$

$$\mathbb{V}ar(Y_i) = \frac{\phi}{\nu_i}\, a''(\theta_i). \tag{2.3}$$

Example 2.1 (Poisson Distribution)**.** This example demonstrates that the Poisson distribution is a member of the exponential family. Let $X_i \sim \mathcal{P}oi(\lambda_i)$ with $\lambda_i > 0$. The probability function is given for $x_i \in \{0, 1, 2, \dots\}$ as follows [Fah2016, pp. 243]:

$$\begin{aligned} f_{X_i}(x_i) &= exp(-\lambda_i)\, \frac{(\lambda_i)^{x_i}}{x_i!} \\ &= exp(x_i\, ln(\lambda_i) - \lambda_i)\, \frac{1}{x_i!}. \end{aligned}$$

Thus, the Poisson distribution is a member of the exponential family with:

$$\begin{aligned} \theta_i &= ln(\lambda_i) \\ \phi &= 1 \\ \nu_i &= 1 \\ a(\theta_i) &= \lambda_i = exp(\theta_i) \\ c(x_i, \phi, \nu_i) &= \frac{1}{x_i!}. \end{aligned}$$

In this work, claim frequency and average claim cost are modeled. For modeling

the claim frequency, duration must be considered. Ohlsson and Johansson define the relative Poisson distribution accordingly [cf. OJ2010, pp. 19].

Example 2.2 (Gamma Distribution)**.** Define $\mu_i := \alpha/\beta_i$, $\phi := 1/\alpha$, and $\theta_i = -1/\mu_i$. The density function of the gamma distribution can be rewritten as follows:

$$\begin{aligned}
f_{Y_i}(y_i) = f_{Y_i}(y_i;\mu_i,\phi) &= \frac{1}{\Gamma(\omega_i/\phi)} \left(\frac{\omega_i}{\phi\mu_i}\right)^{\omega_i/\phi} y_i^{\omega_i/\phi-1} exp\left(-\frac{\omega_i y_i}{\phi\mu_i}\right) \\
&= exp\left(-ln(\Gamma(\omega_i/\phi)) + (\omega_i/\phi)\left(ln\left(\frac{\omega_i}{\phi}\right) - ln(\mu_i)\right) + (\omega_i/\phi - 1)\ ln(y_i) - \frac{\omega_i y_i}{\phi\mu_i}\right) \\
&= exp\left(\frac{-y_i/\mu_i - ln(\mu_i)}{\phi/\omega_i} - ln(\Gamma(\omega_i/\phi)) + (\omega_i/\phi)\ ln\left(\frac{\omega_i}{\phi}\right) + (\omega_i/\phi - 1)\ ln(y_i)\right) \\
&= exp\left(\frac{-y_i/\mu_i - ln(\mu_i)}{\phi/\omega_i}\right) \frac{1}{\Gamma(\omega_i/\phi)} \left(\frac{\omega_i}{\phi}\right)^{\omega_i/\phi} y_i^{\omega_i/\phi-1}.
\end{aligned}$$

Substituting θ_i into the equation yields:

$$\begin{aligned}
f_{Y_i}(y_i) &= f_{Y_i}(y_i;\theta_i,\phi) \\
&= exp\left(\frac{y_i\theta_i + ln(-\theta_i)}{\phi/\omega_i}\right) \frac{(\omega_i/\phi)^{\omega_i/\phi}}{\Gamma(\omega_i/\phi)}\ y_i^{\omega_i/\phi-1}.
\end{aligned}$$

Thus, the gamma distribution is a member of the exponential family with:

$$\begin{aligned}
\theta_i &= -\frac{1}{\mu_i} = -\frac{\beta_i}{\alpha} \\
\phi &= \frac{1}{\alpha} \\
\nu_i &= \omega_i \\
a(\theta_i) &= -ln(-\theta_i) \\
c(y_i,\phi,\nu_i) &= \frac{(\omega_i/\phi)^{\omega_i/\phi}}{\Gamma(\omega_i/\phi)}\ y_i^{\omega_i/\phi-1}.
\end{aligned}$$

2.1.2 Systematic Component

In Chapter 2.1.1, it was explained that claim frequency follows a relative Poisson distribution and the average claim cost follows a Gamma distribution, both of which belong to the exponential family. This subsection focuses on the second

component of GLM, namely the systematic component. The systematic component describes how explanatory variables $X_1, X_2, \ldots, X_p$ are combined into a linear predictor η_i.

Consider a simple multiplicative model with two rating factors, namely i_1 and i_2. The factor i_1 consists of three classes, and the factor i_2 has two classes. Thus, there are a total of six tariff cells. Let tariff cell $(1, 1)$ be the base. Then, the multiplicative model is given by:

$$\mu_{i_1,i_2} = \gamma_0 \gamma_{1,i_1} \gamma_{2,i_2}.$$

The goal is to transform the expectation value μ_{i_1,i_2} so that it is no longer represented as a multiplication but as a linear combination. By taking the logarithm on both sides, we obtain:

$$log(\mu_{i_1,i_2}) = log(\gamma_0) + log(\gamma_{1,i_1}) + log(\gamma_{2,i_2}).$$

Table 2.1 summarizes the multiplicative model for individual tariff cells.

i_1	i_2	$log(\mu_{i_1,i_2})$
1	1	$log(\gamma_0)$
1	2	$log(\gamma_0) + log(\gamma_{2,2})$
2	1	$log(\gamma_0) + log(\gamma_{1,2})$
2	2	$log(\gamma_0) + log(\gamma_{1,2}) + log(\gamma_{2,2})$
3	1	$log(\gamma_0) + log(\gamma_{1,3})$
3	2	$log(\gamma_0) + log(\gamma_{1,3}) + log(\gamma_{2,2})$

Table 2.1: Multiplicative model with $\gamma_{1,1} = \gamma_{2,1} = 1$ (own representation)

Next, the logarithms in Table 2.1 are replaced with new parameters. Define $\beta_1 := log(\gamma_0),\ \beta_2 := log(\gamma_{1,2}),\ \beta_3 := log(\gamma_{1,3})$ and $\beta_4 := log(\gamma_{2,2})$.

Tariff Cell i	i_1	i_2	$log(\mu_i)$
1	1	1	β_1
2	1	2	$\beta_1 + \beta_4$
3	2	1	$\beta_1 + \beta_2$
4	2	2	$\beta_1 + \beta_2 + \beta_4$
5	3	1	$\beta_1 + \beta_3$
6	3	2	$\beta_1 + \beta_3 + \beta_4$

Table 2.2: Continuation of Table 2.1 (own representation)

Next, define a dummy variable x_{ij}, which satisfies the following condition:

$$x_{ij} = \begin{cases} 1, & \text{if } \beta_j \text{ is in } log(\mu_i), \\ 0, & \text{otherwise.} \end{cases}$$

By defining dummy variables, the logarithms can be rewritten as a linear combination of $x_{ij}\beta_j$. The dummy variables for this example are summarized in Table 2.3.

Tariff Cell i	i_1	i_2	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	1	1	1	0	0	0
2	1	2	1	0	0	1
3	2	1	1	1	0	0
4	2	2	1	1	0	1
5	3	1	1	0	1	0
6	3	2	1	0	1	1

Table 2.3: The corresponding dummy variables (own representation)

Thus, the linear structure of the multiplicative model is given by:

$$\eta_i := log(\mu_i) = \sum_{j=1}^{4} x_{ij}\beta_j, \quad i \in \{1, 2, 3, 4, 5, 6\}$$

or in matrix notation $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ with:

$$\boldsymbol{\eta} = \begin{pmatrix} log(\mu_1) \\ log(\mu_2) \\ log(\mu_3) \\ log(\mu_4) \\ log(\mu_5) \\ log(\mu_6) \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \\ x_{61} & x_{62} & x_{63} & x_{64} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}.$$

The matrix $\mathbf{X}$ is called the design matrix. If the model includes an intercept, then the first column of the design matrix consists of ones, and the first element of the vector $\boldsymbol{\beta}$ is β_0.

2.1.3 Link Function

From the example in Chapter 2.1.2, it can be shown that $\eta_i = log(\mu_i)$. In the general case, η_i may correspond to another function $g(\cdot)$ rather than the logarithm function. It is important that the function $g(\cdot)$ is bijective, monotonic, and continuously differentiable [cf. DHT2019, p. 103]. The function $g(\cdot)$ is called the link function because it links the parameter μ_i with the linear structure, or the systematic component through:

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

The logarithm function is a suitable link function for a multiplicative model [cf. OJ2010, p. 28].

2.2 Extreme Gradient Boosting (XGBoost)

Tree boosting is a powerful machine learning technique that has achieved state-of-the-art results in various applications, including ranking, classification, and regression tasks. XGBoost (Extreme Gradient Boosting) is an optimized implementation of gradient tree boosting that provides significant speed and performance improvements over existing approaches. This section provides an overview of the XGBoost algorithm, its regularized learning objective, and key optimizations. All formulas in Chapter 2.2 are based on Tianqi et al. [CG2016].

2.2.1 Simple example in pricing

In XGBoost, predictions are adjusted step-by-step based on errors from prior predictions. Three insured individuals are considered: one is 25 years old and resides in an urban city, another is 40 years old and lives in a small town, and the third is 32 years old and also resides in an urban city. An initial predicted premium $\hat{\mu}_i$ of 1000 is assigned equally to all.

i	Age	Location	$\hat{\mu}_i$	S_i	$Error$	$Error^2$
1	25	Urban	1000	1300	+300	90000
2	40	Small town	1000	500	-500	250000
3	32	Urban	1000	1000	0	0

Table 2.4: XGBoost in pricing (own representation)

After one year, the claim amount for each individual S_i is recorded. Errors between the predicted and actual claim amounts are computed, and decision trees are iteratively constructed to minimize these errors. Two decision trees are applied to adjust the predicted premium. The first tree modifies the premium based on age: if the person is younger than 35 years, 100 is added; otherwise, 200 is subtracted. The second tree considers the place of residence: if the person lives in an urban city, 150 is added to the predicted premium; otherwise, 300 is subtracted.

The predictions have been refined, resulting in a greater variation in the predicted premiums. Additionally, the sum of squared errors has significantly de-

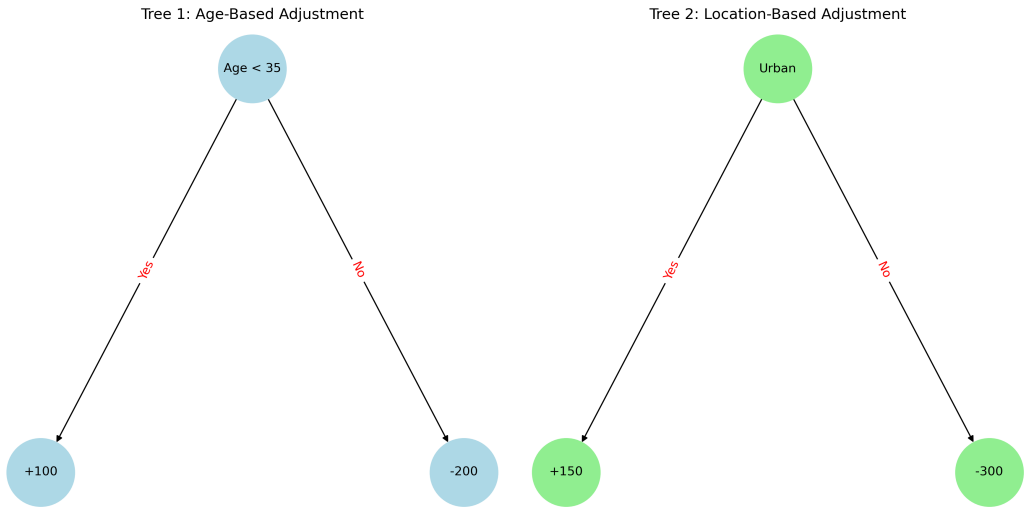

Figure 2.1: Decision trees for premium adjustment (own representation)

creased. XGBoost continues this iterative process until the stopping criterion is met.

i	Age	Location	$\hat{\mu}_i$	S_i	$Error$	$Error^2$
1	25	Urban	1250	1300	+50	2500
2	40	Small town	500	500	0	0
3	32	Urban	1150	1000	-150	22500

Table 2.5: Continuation of Table 2.4 (own representation)

2.2.2 Regularized Learning Objective

For a dataset with n examples and m features:

$$D = \{(x_i, y_i)\} \quad \text{where } |D| = n, \quad x_i \in \mathbb{R}^m, \quad y_i \in \mathbb{R}$$

XGBoost builds an ensemble of K trees to predict the output:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{2.4}$$

where $\mathcal{F}$ is the space of regression trees. Each tree f_k has a structure that maps an example x_i to a corresponding leaf weight. Unlike traditional decision trees, regression trees in XGBoost contain continuous scores instead of class labels.

The learning objective consists of a loss function l and a regularization term Ω:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{2.5}$$

where the regularization term controls the complexity of the trees:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.6}$$

Here, T represents the number of leaves in the tree, γ is the minimum loss reduction parameter, λ is the ridge regularization parameter and w_j represents the weight of the j-th leaf. This regularization term helps prevent overfitting and ensures smooth model generalization.

2.2.3 Gradient Tree Boosting

The tree ensemble model includes functions as parameters, making it unsuitable for direct optimization using traditional methods. Instead, XGBoost follows an additive approach, iteratively constructing new trees to minimize the objective function at step t:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.7}$$

By applying a second-order Taylor expansion to the loss function:

$$L^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.8}$$

where:

$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ represents the first-order gradient.

$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ corresponds to the second-order gradient (Hessian).

Let I_j denote the set of instances that belong to leaf j:

$$I_j = \{i | q(x_i) = j\} \tag{2.9}$$

Using this definition, the objective function can be expanded by incorporating the regularization term Ω:

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2. \tag{2.10}$$

Rewriting the objective function in terms of leaf-specific gradients and Hessians:

$$\tilde{L}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T. \tag{2.11}$$

For a given tree structure $q(x)$, the optimal leaf weight w_j^* can be computed as:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}. \tag{2.12}$$

By substituting w_j^* into the objective function, the optimal objective function is obtained:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{2.13}$$

2.3 Model Fit and Model Selection

Let $\hat{\boldsymbol{\mu}}$ be an estimated vector and $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ its log-likelihood function. If the number of independent parameters j matches the number of observations n, a perfect model

can arise, in which all $\hat{\mu}_i = y_i$ are set. This model is referred to as a saturated model. The saturated model is used as a reference to assess the goodness-of-fit of each GLM that is based on the same distribution.

Various measures exist to evaluate the goodness-of-fit of the model to the data, e.g. Deviance and Root Mean Square Error (RMSE). Deviance compares the estimated model to the saturated model. Deviance or residual deviance is defined as follows [DHT2019, p. 152]:

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2\phi\big[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\mu}};\mathbf{y})\big] \\ &= 2\sum_{i=1}^{n} \nu_i\big[y_i(\theta_i - \hat{\theta}_i) - a(\theta_i) + a(\hat{\theta}_i)\big], \end{aligned}$$

where ϕ is the dispersion parameter, and $l(\mathbf{y};\mathbf{y})$ corresponds to the maximum log-likelihood of the saturated model. The relationship between deviance and scaled deviance is given by:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \phi D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}).$$

Unscaled deviance has the advantage of being independent of the parameter ϕ.

Example 2.3 (Deviance of the Gamma Distribution)**.** The deviance for the Gamma distribution is described as follows:

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n} \omega_i \left[y_i \left(-\frac{1}{y_i} + \frac{1}{\hat{\mu}_i} \right) + ln\left(\frac{1}{y_i}\right) - ln\left(\frac{1}{\hat{\mu}_i}\right) \right] \\ &= 2\sum_{i=1}^{n} \omega_i \left[\frac{y_i}{\hat{\mu}_i} - 1 - ln\left(\frac{y_i}{\hat{\mu}_i}\right) \right]. \end{aligned}$$

Example 2.4 (Deviance of the Relative Poisson Distribution)**.** The deviance for the relative Poisson distribution is described as follows:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \omega_i [y_i ln(y_i) - y_i ln(\hat{\mu}_i) - y_i + \hat{\mu}_i].$$

Root Mean Squared Error is a commonly used metric for evaluating the accuracy of a predictive model. It measures the differences between predicted and observed

values. RMSE is particularly useful because it gives higher weight to large errors due to squaring each difference before averaging. The RMSE is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

3 Modelling using GLM and XGBoost

This chapter focuses on the implementation of the methods for tariff calculation, as described in Chapter 2. The modelling is carried out using the programming language **Python**. Several packages are required for modelling in this study. Therefore, it is recommended to first identify and install the necessary packages.

Before starting the modelling, the data is first analyzed and prepared in Chapter 3.1. Chapter 3.2 deals with data preprocessing. This includes, for example, checking whether the dataset contains missing values. After preprocessing, models for claim frequency are developed in Chapter 3.3, while Chapter 3.4 focuses on modeling claim severity.

The dataset `dataOhlsson`, which is used in this chapter, is publicly available in an **R** package called `insuranceData`. The data originates from the former Swedish insurance company *Wasa Insurance*, which merged with *Länsförsäkringar Alliance* in 1999. The dataset covers the period from 1994 to 1998 and contains information about comprehensive insurance for motorcycles.

The dataset consists of 64,548 observations, each corresponding to an insurance policy. Each policy contains nine variables, which can be categorized into rating variables and response variables. The following subsection provides a detailed description of these variables.

3.1 Data description

The explanatory variables included in the dataset are as follows. First, the information about the policyholder:

`agarald` Age of the policyholder, continuous variable, measured in years, ranging from 0 to 99.

`kon` Gender, categorical variable with two values: `K` (female) and `M` (male).

`zon` Policyholder's zone, categorical variable with seven levels, ranging from 1 to 7.

Next, information about the insured vehicle:

`mcklass` A classification based on the so-called EV ratio, defined as (engine power in kW $\times$ 100) / (vehicle weight in kg + 75), rounded down to the nearest integer. Categorical variable with seven levels, ranging from 1 to 7.

`fordald` Age of the vehicle, continuous variable, measured in years, ranging from 0 to 99.

`bonuskl` Bonus class, categorical variable with seven levels, ranging from 1 to 7. A new driver starts in bonus class 1; for each claim-free year, the bonus class increases by 1. After the first claim, the class decreases by 2. The driver cannot return to class 7 unless they have had at least six consecutive claim-free years.

Finally, the key ratios are described as follows:

`duration` Policy years, representing the total duration of all individual insurance policies. Continuous variable, measured in years.

`antskad` Number of claims, categorical variable with three levels: 0, 1, 2.

`skadkost` Claim costs, continuous variable.

3.2 Data preprocessing

This subsection explains the process of data preprocessing. The dataset consists of 64,548 records with nine features, each having different data types. The only preprocessing in this study is handling missing values. There are various methods for dealing with missing data. In this study, data containing missing values are removed.

3.3 Modelling claim frequency

A total of three GLMs and one XGBoost model are developed for modeling claim frequency. Before proceeding with model training, the dataset must be filtered accordingly. It is important to recall the relationships between claim-related variables, as summarized in Table 1.2. The claim frequency is calculated as follows:

$$CF = \frac{N}{n_0}.$$

In this case, the duration n_0 must not be zero, as this would result in an undefined claim frequency. As a result, only records where the duration is greater than zero are used in the analysis. The dataset is then split into training and test sets. The training dataset is a random sample comprising 70% of all observations, while the remaining 30% of the data is assigned to the test set.

The model GLM0 is the intercept-only model, meaning it includes only β_0. GLM1 is the model that incorporates all categorical rating factors. Thus, GLM1 has the following structure:

$$\log(\mu_i) = \log(\omega_i) + \sum_{j=1}^{7} x_{ij}\beta_j, \quad i \in \{1, 2, \ldots, 43731\}, \tag{3.1}$$

where $x_{i0} = 1$ for all $i \in \{1, 2, \ldots, 43731\}$, and ω_i represents the exposure for the i-th observation. GLM2 extends GLM1 by incorporating an interaction effect between `bonuskl` and `zon`. The XGBoost model, on the other hand, considers all available features.

Model	Features
GLM0	(Intercept model)
GLM1	`agarald` + `kon` + `zon` + `mcklass` + `fordald` + `bonuskl`
GLM2	`agarald` + `kon` + `zon` + `mcklass` + `fordald` + `bonuskl` + `bonuskl` * `zon`
XGB	`agarald` + `kon` + `zon` + `mcklass` + `fordald` + `bonuskl` + `skadkost`

Table 3.1: Summary of all models

The results of our models are presented in Table 3.2. To assess the model performance, the RMSE and Poisson Deviance are used. Lower values indicate better predictive accuracy. The table clearly shows that the XGBoost outperforms all GLM models, achieving the lowest Poisson Deviance of 0.12% and an RMSE of 0.0253. Compared to the GLM models, XGBoost provides significantly improved predictive accuracy, demonstrating its superior capability in modeling claim frequency.

Model	Claim Frequency (Test)		Poisson Deviance		RMSE
	Actual	Predicted	Train	Test	
GLM0	1.00%	1.09%	8.75%	7.50%	0.1066
GLM1	1.00%	1.10%	8.49%	6.83%	0.1055
GLM2	1.00%	1.10%	8.38%	6.76%	0.1055
XGB	1.00%	1.01%	0.00%	0.12%	0.0253

Table 3.2: Claim frequency, Poisson Deviance, and RMSE for different models.

3.4 Modelling claim severity

A total of two GLMs and one XGBoost model are developed for modeling claim severity. Before proceeding with model training, the dataset must be filtered accordingly. The claim severity is calculated as follows:

$$CS = \frac{S}{N}.$$

In this case, the claim number N must not be zero, as this would result in an undefined claim severity. As a result, only records where the claim number is greater than zero are used in the analysis. The dataset is then split into training and test sets. The training dataset is a random sample comprising 70% of all observations, while the remaining 30% of the data is assigned to the test set.

The model GLM0 is the intercept-only model, meaning it includes only β_0. GLM1 is the model that incorporates all categorical rating factors. The XGBoost

Model	Features
GLM0	(Intercept model)
GLM1	`agarald` + `kon` + `zon` + `mcklass` + `fordald` + `bonuskl`
XGB	`agarald` + `kon` + `zon` + `mcklass` + `fordald` + `bonuskl` + `duration`

Table 3.3: Summary of all models

model, on the other hand, considers all available features.

Table 3.4 presents the results for the claim severity models, comparing actual and predicted values across different modeling approaches. The models are evaluated based on their ability to predict total claim severity, with RMSE as the primary performance metric. A lower RMSE indicates better predictive accuracy.

Model	Claim Severity (Test)		RMSE
	Actual	Predicted	
GLM0	4,762,611.00	5,116,337.08	38,224.70
GLM1	4,762,611.00	4,861,973.03	38,418.00
XGB	4,762,611.00	4,984,814.42	42,570.22

Table 3.4: Claim severity and RMSE for different models.

The GLM0 model, which only includes an intercept, produces the highest predicted claim severity at $5,116,337.08$, resulting in an RMSE of $38,224.70$. The GLM1 model, which incorporates categorical rating factors, provides an improved prediction of $4,861,973.03$, maintaining a similar RMSE of $38,418.00$. In contrast, the XGBoost model, which leverages all available features, predicts claim severity at $4,984,814.42$, but with a slightly higher RMSE of $42,570.22$. These results suggest that while XGBoost provides a more flexible modeling approach, it does not necessarily outperform GLMs in terms of RMSE for claim severity prediction. The GLM1 model demonstrates the best overall performance by achieving the closest predicted value to the actual claim severity while maintaining a relatively lower RMSE.

4 Conclusion

This study aimed to evaluate different statistical and machine learning models for claim frequency and claim severity prediction. GLM and XGBoost were compared using key performance metrics, including Poisson Deviance and RMSE.

The results demonstrate that XGBoost significantly outperforms GLM in predicting claim frequency. With a Poisson Deviance of 0.12% and RMSE of 0.0253, XGBoost provided the most accurate predictions among all models. The GLM models, particularly GLM1 and GLM2, showed reasonable performance but were less precise than XGBoost.

However, in claim severity prediction, the findings indicate that GLM perform better than XGBoost in terms of RMSE. Among the models, GLM1 achieved the lowest deviation from the actual claim severity values, making it the best model for claim severity estimation. Although XGBoost remains a flexible and powerful machine learning approach, its slightly higher RMSE suggests that traditional GLM may still be preferred for modeling claim severity.

One major difference between the two approaches lies in how they handle interactions among rating variables. In GLM, interactions must be explicitly defined and included in the model to improve predictive accuracy. On the other hand, XGBoost automatically captures interactions between variables, making it more adaptable to complex datasets without the need for manual feature engineering.

Additionally, parameter tuning could further enhance XGBoost's performance, particularly in claim severity prediction. By optimizing hyperparameters such as learning rate, tree depth, and regularization terms, XGBoost could potentially reduce its RMSE and become more competitive with GLM in this context.

These results highlight the importance of selecting the appropriate modeling technique based on the target variable. While XGBoost is well-suited for predicting claim frequency, GLM remain competitive for modeling claim severity.

Bibliography

[Bus2015] M. Buse, K. Dräger, C. Dubowik, F. Ellgring, T. Franze, K. Geisler, P. Go, and D.G.V. Finanzmathematik. *Aktuarielle Methoden der Tarifgestaltung in der Schaden-/Unfallversicherung*. Versicherungs- und Finanzmathematik. Verlag Versicherungswirtschaft, 2015. ISBN: 978-3-862-98372-8.

[CG2016] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[DHT2019] Michel Denuit, Donatien Hainaut, and Julien Trufin. *Effective Statistical Learning Methods for Actuaries I*. Springer Cham, Sept. 3, 2019. ISBN: 978-3-030-25820-7.

[Fah2016] Ludwig Fahrmeir, Christian Heumann, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik*. Springer Spektrum Berlin, Heidelberg, Aug. 18, 2016. ISBN: 978-3-662-50372-0.

[Goe2016] Heinz-Willi Goelden, Klaus Th. Hess, Martin Morlock, Klaus Schmidt, and Klaus Schröter. *Schadenversicherungsmathematik*. Springer Spektrum Berlin, Heidelberg, Jan. 22, 2016. ISBN: 978-3-662-48860-7.

[HK2017] Andreas Handl and Torben Kuhlenkasper. *Multivariate Analysemethoden*. Springer-Verlag GmbH Deutschland, Aug. 3, 2017. ISBN: 978-3-662-54754-0.

[HT1986] Trevor Hastie and Robert Tibshirani. "Generalized Additive Models". In: *Statistical Science* 1.3 (1986), pp. 297–310. ISSN: 08834237. URL: http://www.jstor.org/stable/2245459 (visited on 06/16/2023).

[MBL2000] Karl P Murphy, Michael J Brockman, and Peter KW Lee. "Using generalized linear models to build dynamic pricing systems". In: *Casualty Actuarial Society Forum, Winter*. 2000, pp. 107–139.

[MN1989] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989. ISBN: 978-0-412-31760-6.

[NW1972] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pp. 370–384. ISSN: 00359238. URL: http://www.jstor.org/stable/2344614 (visited on 05/06/2023).

[OJ2010] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer Berlin, Heidelberg, Mar. 18, 2010. ISBN: 978-3-642-10791-7.

[WB2017] Mario Wuthrich and Christoph Buser. "Data Analytics for Non-Life Insurance Pricing". In: *SSRN Electronic Journal* (Jan. 2017). DOI: 10.2139/ssrn.2870308.

[WM2022] M.V. Wüthrich and M. Merz. *Statistical Foundations of Actuarial Learning and its Applications*. Springer Actuarial. Springer International Publishing, 2022. ISBN: 978-3-031-12411-2.