Stockholms universitet

Discrimination-free pricing in motor insurance

Qi Lin

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2023:19
Försäkringsmatematik
September 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2023:19**
http://www.math.su.se

Discrimination-free pricing in motor insurance

Qi Lin*

September 2023

Abstract

The main objective of this thesis is to analyze the effects of discriminationfree pricing (DFP) on avoiding direct and indirect discrimination in insurance with respect to gender. This is done by comparing the DFPs to standard insurance prices that both include and exclude gender as a covariate. Using a French motor thirdparty liability insurance dataset, we explore two methods: Generalized Additive Models (GAM) and Gradient Boosting Machines (GBM), to build a model for claim frequency. A grid search with 10-fold cross-validation selects the optimal parameters of GBMs. Generalized cross-validation is adapted to find smoothing parameters for the GAMs. We evaluate the predictive performance of the models using concentration curves, Root Mean Square Error (RMSE) and deviance loss. The DFPs are also compared to the standard insurance prices w.r.t partial dependence plots (PDPs) and a certain type of coefficient of determination. We investigate the impact of nondiscriminatory pricing based on the GAM and GBM models that include gender as a discriminatory variable. In the analyses, we find DFPs lie closer to unawareness prices for GBM than GAM. The best-estimate prices have the best predictive performance. Differences in DFPs compared to the best-estimate prices are less for GBM than GAM.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: levxy@hotmail.com. Supervisor: Mathias Lindholm.

Acknowledgements

I would like to express my deep gratitude to my supervisor Mathias Lindholm for his invaluable patience and advice throughout this process. I am also thankful for the support and encouragement from my friends. Lastly, I cannot overlook the immense contribution of my family, whose unwavering belief in me served as a constant source of motivation during this entire period.

Contents

Chapter 1

Introduction

In Europe, the law does not allow the use of policyholder information on gender in insurance pricing. Direct discrimination is defined as treatment differences for gender reasons [3]. It is not an appropriate way forward to delete discriminatory covariates. This is because sensitive information may be derived from other non-protected attributes, see [9]. Such inferring results in indirect discrimination, which differs from group fairness criteria adapted to restrict the influence of protected features in machine learning. The primary purpose of the thesis is to analyze the effects of using discrimination-free pricing (DFP) proposed by Lindholm et al. in [8] to avoid direct and indirect discrimination in insurance pricing w.r.t gender. Ensuring protected features remain in the predictive model but do not represent discriminatory characteristics.

We explore this using GAMs and GBMs on a French third-party motor liability insurance data set. With generalized additive models (GAM) and gradient boosting machines (GBM), we build a predictive model to capture the best-estimate price (including gender as a covariate), unawareness price (excluding gender as a covariate), and discrimination-free price, respectively. The optimal parameters in the GAM are obtained using generalized cross-validation, which includes finding the optimal smoothing parameter. The hyperparameters in the GBM are obtained using a grid search. The predictive performance of the models is evaluated using concentration curves, root mean square error, and deviance loss. Further, the partial dependence plots (PDPs) are analyzed together with a coefficient of determination type fidelity measure that tries to quantify the differences between DFPs and best-estimate prices/unawareness prices. Throughout the thesis, we will explore which covariates and interaction effects impact the output most. In general, a discrimination price will not be unbiased. Therefore, an adjusted marginal distribution is used for obtaining unbiased discrimination-free prices.

Structure of the thesis. The second section introduces methods adapted to model building and inspection. In the third section, each covariate is presented. We describe the discrimination-free price based on models from the dataset in the fourth section. In the following section, we show how best-estimate prices, unawareness prices, and discrimination-free prices vary for GBM and GAM with simulations. Conclusion and discussion in the last section.

Chapter 2

Theory

2.1 Discrimination-free insurance pricing

Our objective is to develop an insurance tariff that ensures non-discrimination. Let $\mathbf{D}$ denote the vector of discriminatory explanatory variables, and $\mathbf{X}$ denote the vector of non-discriminatory explanatory variables. Furthermore, Let $\mathbf{X} \sim \mathbb{P}(\mathbf{x})$, $\mathbf{D} \sim \mathbb{P}(\mathbf{d})$ and $(\mathbf{D} \mid \mathbf{X} = \mathbf{x}) \sim \mathbb{P}(\mathbf{d} \mid \mathbf{x})$ be the marginal and conditional distribution of covariate under physical probability $\mathbb{P}$. Let the number of claims Y_i be $\text{Pois}(\omega_i \mu_i)$ given $\mathbf{X} = x_i$. Based on definition 2 in [8], the best-estimate price for claim frequency μ is defined by

$$\mu(\mathbf{X}, \mathbf{D}) := \mathbb{E}[Y \mid \mathbf{X}, \mathbf{D}] \tag{2.1}$$

Generally, the best-estimate price is not discrimination-free because discriminatory covariate $\mathbf{D}$ is directly used in $\mu(\mathbf{X}, \mathbf{D})$. The best-estimate price is unbiased which means $\mu = \mathbb{E}[Y] = \mathbb{E}[\mu(\mathbf{X}, \mathbf{D})]$.

For the unawareness price for Y w.r.t $\mathbf{X}$, discriminatory covariate are not included in the modeling, so it is defined by

$$\mu(\mathbf{X}) := \mathbb{E}[Y \mid \mathbf{X}] = \int \mu(\mathbf{X}, \mathbf{d}) d\mathbb{P}(\mathbf{d} \mid \mathbf{x}) \tag{2.2}$$

The unawareness price $\mu(\mathbf{X})$ evades direct discrimination. However, it may cause proxy or indirect discrimination because it still has a possibility to deduce protected policyholder information from these associated variables features. There has a special case that indirect discrimination avoids when $\mathbf{D}$ and $\mathbf{X}$ are independent. Unawareness price is unbiased due to $\mu = \mathbb{E}[Y] = \mathbb{E}[\mu(\mathbf{X})]$.

Lindholm et al.[8] indicate that the goal is to create pricing formulas free from discrimination while allowing insurers to differentiate policyholders based on non-discriminatory factors. A discrimination-free price for Y with reference to $\mathbf{X}$ is defined by

$$\mu^*(\mathbf{X}) := \int_d \mu(\mathbf{X}, \mathbf{d}) d\mathbb{P}^*(\mathbf{d}) \tag{2.3}$$

where $\mathbf{d} \in \mathbf{D}$ and distribution $\mathbb{P}^*(\mathbf{d})$ is defined in the same range as the marginal distribution of discriminatory covariate $\mathbf{D} \sim \mathbb{P}(\mathbf{d})$. Potential indirect discrimination is avoided by using $\mathbb{P}^*(\mathbf{d})$ in (2.3) instead of $\mathbb{P}(\mathbf{d} \mid \mathbf{x})$ in (2.2) when constructing

prices [9]. In the thesis, we specify gender as a discriminatory covariate. So we have discrimination-free prices as

$$\mu^*(\mathbf{x}) := \mu(\mathbf{x}, \mathbf{D} = \text{male})\mathbb{P}(\mathbf{D} = \text{male}) + \mu(\mathbf{x}, \mathbf{D} = \text{female})\mathbb{P}(\mathbf{D} = \text{female}) \quad (2.4)$$

where $\mathbb{P}(\mathbf{D} = \text{female})$ is the proportion of female in the total portfolio, $\mathbb{P}(\mathbf{D} = \text{male})$ is the proportion of men in the total portfolio, $\mu(\mathbf{x}, \mathbf{D} = \text{male})$ obtained by setting gender as male, same as $\mu(\mathbf{x}, \mathbf{D} = \text{female})$. That is, we use $\mathbb{P}^*(\mathbf{D} = \mathbf{d}) = \mathbb{P}(\mathbf{D} = \mathbf{d})$.

We need to pay attention to the fact that the Discrimination-free price (2.3) is not unbiased. Generally, there is the possibility that a bias exists with a discrimination-free price since $\mathbb{P}^*(\mathbf{d}) \neq \mathbb{P}(\mathbf{d} \mid \mathbf{x})$. For the sake of the premium of the entire portfolio held at a suitable level, the bias requires to be corrected. The portfolio bias of the discrimination-free price is defined by

$$B^* = \mathbb{E}[Y] - \int_{x,d} \mu(\mathbf{x}, \mathbf{d}) d\mathbb{P}^*(\mathbf{d}) d\mathbb{P}(\mathbf{x}) \quad (2.5)$$

One way to allocate the bias B^* is to allocate the proportion of the total premium to $h^*(\mathbf{X})$, therefore the adjusted discrimination-free price can be written as

$$\pi^*(\mathbf{X}) = h^*(\mathbf{X}) \frac{\mu}{\mu - B^*} = \int_d \mu(\mathbf{X}, \mathbf{d}) d\mathbb{P}^*(\mathbf{d}|\mathbf{X}) \frac{\mu}{\mu - B^*} \quad (2.6)$$

2.2 Generalized linear models

We assume a Poisson distribution for the number of damaged material claims of an individual policy during the period. Let N_i denotes the number of claims with exposure ω_i and claim frequency $Y_i = \frac{N_i}{w_i}$. Then let μ_i be the expected value of a number of claims when $\omega_i = 1$. So, $N_i|X_i \sim \text{Pois}(\omega_i \mu_i)$ [10]. Function of claim frequency $Y_i = \frac{N_i}{w_i}$ is given by

$$f_{Y_i}(y_i, \mu_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i y_i}}{(\omega_i y_i)!} \quad (2.7)$$

In the generalized linear model there is an arbitrary function $g(\mu_i)$, which g is the so-called link function. For Poisson data the log-link is often used:

$$g(\mu_i) = \eta_i = \ln(\mu_i) = \sum_{j=0}^{J} \mathbf{x}'_{ij} \beta_j \Rightarrow \mu_i = e^{\sum_{j=0}^{J} \mathbf{x}'_{ij} \beta_j} \quad (2.8)$$

where $\mathbf{x}$ is a $n \times J$ design matrix, β is a $J \times 1$ vector.

2.2.1 Poisson deviance loss

Deviance loss can measure the goodness-of-fit of models, which can be defined as likelihood-ratio-test statistic [10]. Let $\log L(\hat{\mu})$ denote log-likelihood function of the estimated $\hat{\mu}$. The deviance loss can be:

$$\begin{aligned} D(\mathbf{y}, \hat{\mu}) &= 2[\log(L(\mathbf{y})) - \log L(\hat{\mu})] \\ &= 2\left[\log \prod_{i=1}^{n} \exp(-y_i) \frac{y_i^{y_i}}{y_i!} - \log \exp(-\mu_i) \frac{\mu_i^{y_i}}{y_i!} \right] \\ &= 2 \sum_{i=1}^{n} (y_i \log \frac{y_i}{\mu_i} - y_i + \mu_i) \end{aligned} \quad (2.9)$$

Claim frequency modeling relates to count data, assumed to follow a Poisson distribution. Wüthrich and Buser [11] propose Possion deviance as a loss function for the Poisson distribution. The Poisson deviance can be weighted by duration as:

$$\begin{aligned} D(\mathbf{y}, \omega\hat{\mu}) &= 2[\log(L(y,y)) - \log L(y, \omega\hat{\mu})] \\ &= 2\Bigg[\sum_{i=1}^{n} -\omega_i y_i + \omega_i y_i \log(\omega_i y_i) - \log((\omega_i y_i)!) - \\ &\quad (-\omega_i\hat{\mu}_i + \omega_i y_i \log(\omega_i\hat{\mu}_i) - \log((\omega_i y_i)!))\Bigg] \\ &= 2\Bigg[\sum_{i=1}^{n} -\omega_i y_i + \omega_i y_i \log(y_i) + \omega_i\hat{\mu}_i - \omega_i y_i \log(\hat{\mu}_i))\Bigg] \\ &= 2\Bigg[\sum_{i=1}^{n} \omega_i(\hat{\mu}_i - y_i + y_i \log\frac{y_i}{\hat{\mu}_i}))\Bigg] \\ &= \sum_{i=1}^{n} \omega_i \hat{D}^*(y_i, \hat{\mu}_i) \end{aligned} \tag{2.10}$$

where $\hat{D}^*$ is unit deviance.

2.3 Generalized additive models

Hastie and Tibshirani introduced Generalized additive models in the 1980s, see [6]. They mentioned that a continuous variable usually takes on a much smaller number of values because of rounding. So instead of $g(\mu_i) = \eta_i = \sum_{j=0}^{J} \mathbf{x}'_{ij}\beta_j$ they assumed that

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(x_{ij}), i = 1, 2, \cdots, n \tag{2.11}$$

where f_j is some suitable functions, n denotes the numbers of observations, x_{ij} is the value of the variable j for observation i. Until now, the functions $f_j(\cdot)$ are not identified. We suppose to model with all variables except the first one, and then the model will look like

$$\eta_i = \sum_{j=1}^{J} \beta_j \mathbf{x}'_{ij} + f(x_{i1}), \cdots, n \tag{2.12}$$

A requirement is that the function should be twice continuously differentiable and also not vary too much, that is $\int_a^b (f''(x))^2 dx = 0$, where $a \leq z_1$, $z_1 \leq b$ and $z_1, \cdots, z_m$ denote the possible value of x_{i1}. To find the function that gives the best performance of the effect of a continuous variable, we take penalized deviance to measure the goodness of estimated values of the data,

$$\Delta(f) = D(\mathbf{y}, \mu) + \lambda \int_a^b (f''(x))^2 dx \tag{2.13}$$

where $D(y, \mu)$ is the Poisson deviance loss, $\mathbf{y}$ is the vector with observations, $\lambda \int_a^b (f''(x))^2 dx$ is a measure of variability of $f(\cdot)$, λ is a smoothing parameter. By tuning the smoothing parameter λ, we can find the balance between a good fit to the data and the function's variability by minimizing the penalized deviance $\Delta(f)$.

2.3.1 Natural cubic splines

The natural cubic spline is widely used for the function $f_j(\cdot)$ in (2.11). For a set of m knots with $u_1, \cdots, u_m$, we define the function s on the interval $[u_1, u_m]$ has $s(x) = p_k(x)$ where $k = 1, \cdot, m-1$. By setting a twice continuously differentiable function $p_k = a_k + b_k x + c_k x^2 + d_k x^3$ as spline function, we say function s is *cubic spline* if it satisfies the conditions $p_{k-1}(u_k) = p_k(u_k)$, $p'_{k-1}(u_k) = p'_k(u_k)$, $p''_{k-1}(u_k) = p''_k(u_k)$ for internal knots $k = 2, \cdots, m-1$ [10]. If we extend a cubic spline s to an interval $[a, b]$ which includes $[u_1, u_m]$ and $f'' = 0$ for $x \in [a, u_1]$, $x \in [u_m, b]$, we call such cubic spline as *natural cubic spline*.

For a set of m knots $u_1 < \cdots < u_m$ with given values $f_1^*, \cdots, f_m^*$, there exits a unique natural cubic spline s on $[a, b]$, that is $s(u_k) = f_k^*$, $k = 1, \cdots, m$, see Theorem 5.1 in [10]. Ohlsson & Johansson [10] indicates that it is adequate to consider natural cubic spline when we search for twice continuously differentiable function that minimizes equation (2.13).

Since a *B-splines* can be expressed as a linear combination of simple basis splines, see [10], we shall use it to parametrize the set of cubic splines. For a set of m knots has dimension $m + j - 1$, a spline s may be expressed as $s(x) = \sum_{k=1}^{m+j-1} \beta_k B_{jk}(x)$, where $k = 1, \cdots, m + j - 1$.

2.3.2 Choosing the smoothing parameter

With different smoothing parameters λ, the fitted cubic spline changes between two extremes. One is a straight line, and the other is a natural cubic spline that fits perfectly with the data. Wüthrich and Buser, see [11], point out that the Generalized cross-validation criterion can be applied to find suitable tuning parameters λ. Hastie et al., see [6], indicate that GCV takes advantage of K-fold cross-validation with much faster computation. The GCV criterion with a scaled in-sample loss is given by

$$\text{GCV}(\lambda) = (1 - \frac{M(\lambda)}{n})^{-2} L_D^{\text{in-s}} = (\frac{n}{n - M(\lambda)})^2 L_D^{\text{in-s}} \tag{2.14}$$

where $M(\lambda)$ is the effective degrees of freedom of the model, $L_D^{\text{in-s}}$ is the Poisson loss in-sample function. The effective degrees of freedom is obtained from the sum of the diagonal elements of an influence matrix only depending on the input vectors x_i and λ, see section 5.4.1 in [6]. For more details, refer to Hastie et al.[6], sections 7.6 and 7.10.1.

2.3.3 Interactions

Assume a two-way interaction exists between continuous and categorical variables in the GAM model. Let x_{1i} denote the value of a categorical variable for ith observation with possible values $z_{11}, \cdots, z_{1m_1}$, and x_{2i} denotes the value of the continuous variable with possible values $z_{21}, \cdots, z_{2m_2}$. There exits a function $\phi_j(x) = 1$ if $x = z_{1j}$, and $\phi_j(x) = 0$ otherwise. Set $B_1(\cdot), \cdots, B_{m_2+2}(\cdot)$ as the cubic B-splines for the knots $z_{21}, \cdots, z_{2m_2}$, see [10].

In the thesis, we shall investigate age effects for males and females, there would exist one spline for males and the other for females. Therefore the expression of our

model can be

$$\eta_i = \eta(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \sum_{j=1}^{m_1} \sum_{k=1}^{m_2+2} \beta_{jk} \phi_j(\mathbf{x}_{1i}) B_k(\mathbf{x}_{2i}) \tag{2.15}$$

By setting $s_j(x) = \sum_{k=1}^{m_2+2} \beta_{jk} B_k(\mathbf{x}_{2i})$ as cubic spline, we will get $\eta_i = s_j(\mathbf{x}_{2j})$ since $\phi_j(z_{1j}) = 1$ when $x_{1i} = z_{1j}$.

For the Poisson distribution we have $\mu(x_{1i}, x_{2i}) = \exp\{\eta(x_{1i}, x_{2i})\}$. Hence the corresponding penalized deviance is expressed as:

$$\begin{aligned} \Delta(s) &= D(y, \mu) + \lambda \int_{\mu_1}^{\mu_m} (s''(x))^2 dx \\ &= 2 \sum_i \omega_i (y_i \log y_i - y_i \log \mu(\mathbf{x}_{1i}, \mathbf{x}_{2i}) - y_i + \mu(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \sum_{j=1}^{m_1} \lambda_j (s_j''(x))^2 dx \end{aligned} \tag{2.16}$$

The penalized deviance can be minimized with transformed data $\tilde{\omega}_{jk}$ and $\tilde{y}_{jk}$, where $\tilde{\omega}_{jk} = \sum_{i \in I_{jk}} \omega_i$, $\tilde{y}_{jk} = \frac{1}{\tilde{\omega}_{jk}} \sum_{i \in I_{jk}} \omega_i y_i$, I_{jk} the set of i.

2.4 Gradient Boosting Machines

In data mining, usually only a small subset of predictor covariates are relevant for prediction. During data preprocessing, it is necessary to filter out irrelevant covariates and create relevant special features. Moreover, data mining applications also require models to explain the relationship between input covariates and predicted outputs. An off-the-shelf method can be directly adapted to data without data preprocessing or adjusting the learning process. Decision trees satisfy the conditions of off-the-shelf procedure for data mining. Decision trees produce interpretable models quickly and naturally incorporating a mixture of numeric and categorical covariates and missing values. Executing internal covariates can be selected as a component. Boosting a decision tree improves accuracy significantly while preserving most of the properties required for data mining. Gradient boosting machines (GBM) can generate accurate and effective off-the-shelf methods for data mining, see [6]. In a predictive learning problem, we have a response variable y and explanatory variable $\mathbf{x}$. Our goal is to estimate a function $\hat{F}(x)$ that approximates an unknown function $F^*(x)$, which minimizes the expected value of a specified loss function $L(y, F(x))$ over the joint distribution of all $(y, \mathbf{x})$. For claim frequency, we use Poisson deviance as loss function, which will be discussed in the next section. Following Friedman in [4], we restrict $F(x)$ to be of the form

$$F(\mathbf{x}; \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{2.17}$$

where $h(\mathbf{x}, \mathbf{a})$ is a simple parameterized function of variables $\mathbf{x}$ with parameters $\mathbf{a}$, also called a "base learner".

By the optimization method we take the solution $F^*(\mathbf{x}) = \sum_{m=0}^{M} f_m(\mathbf{x})$ where $f_0(x)$ is an initial guess with incremental functions $\{f_m(x)\}_1^M$. For steepest-descent,

$f_m(\mathbf{x}) = -\rho_m g_m(\mathbf{x})$ where ρ_m is step length, $\rho_m = \arg\min_\rho E_{y,\mathbf{x}} L(y, F_{m-1}(x) - \rho g_m(\mathbf{x}))$,$-g_m(\mathbf{x})$ gives the best steepest decent step direction,

$$g_m(\mathbf{x}) = E_y \left[\frac{\partial L(y, F(x))}{\partial F(x)} \;\middle|\; x \right]_{F(x)=F_{m-1}(x)} \tag{2.18}$$

We try with a greedy stagewise approach for $m = 1, 2, \cdots, M$,

$$F_m(x) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m) \tag{2.19}$$

By construction, we rewrite the unconstrained negative gradient with data-based analog as

$$g_m(\mathbf{x_i}) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \;\middle|\; x \right]_{F(x)=F_{m-1}(x)} \tag{2.20}$$

which is defined only at the data points $\{\mathbf{x}_i\}_1^N$. We choose $\{h(\mathbf{x}; \mathbf{a}_m)\}_1^N$ most parallel to $-g_m \in R^N$ and the line search a step length ρ_m can be written as

$$\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) - \rho h_m(\mathbf{x_i}; \mathbf{a}_m)) \tag{2.21}$$

By considering each base learner $h(\mathbf{x}; \mathbf{a}_m)$ as a J-terminal node regression tree, each regression tree model has additive form

$$h(\mathbf{x}; \{b_j, R_j\}_1^J) = \sum_{j=1}^{J} b_j \mathbb{1}\{\mathbf{x} \in R_j\} \tag{2.22}$$

where $\{R_j\}_i^J$ are disjoint regions that cover the space of all points predictor variables values. The indicator function is 1 if its argument is true. Otherwise, it is zero, $\{b_j\}_1^J$ are the coefficients. If $\mathbf{x} \in R_j$ then $h(\mathbf{x}) = b_j$ due to disjoint regions. Update equation about $F_m(\mathbf{x})$ to:

$$F_m(x) = F_{m-1}(\mathbf{x}) + \rho_m \sum_{j=1}^{J} b_{jm} \mathbb{1}\{x \in R_{jm}\} \tag{2.23}$$

where $\{R_{jm}\}_1^J$ are the terminal node of the regression tree at mth iteration. Set $\gamma_{jm} = \rho_m b_{jm}$ the model updates as

$$F_m(x) = F_{m-1}(\mathbf{x}) + \sum_{j=1}^{J} \gamma_{jm} \mathbb{1}\{x \in \mathbf{R}_{jm}\} \tag{2.24}$$

Therefore, the quality of the fit can be further improved in this case by using the optimal coefficients for each region. The optimal coefficients are:

$$\gamma_{jm} = \arg\min_\gamma \sum_{\mathbf{x_i} \in \mathbf{R_{jm}}} L(y_i, F_{m-1}(\mathbf{x}_i) + \gamma) \tag{2.25}$$

Each iteration reduces the training risk, which can lead to overfitting. Thus, the shrinkage technique is introduced for controlling the number of iterations M. The simplest way of using shrinkage is to scale each tree by a factor $\nu \in [0, 1]$.

$$F_m(x) = F_{m-1}(\mathbf{x}) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}\{x \in \mathbf{R}_{jm}\} \tag{2.26}$$

The parameter ν is also called the learning rate of the booting procedure. A trade-off exists between the number of boosting iterations M, training risk $L(f_M)$, and ν. Lager training risk for the same number of iterations M with a smaller shrinkage value. It can lead to larger values of M for the same training risk with a smaller shrinkage value.

Friedman, see [4], introduced stochastic gradient boosting to improve performance and computational efficiency. The algorithm of stochastic gradient boosting is shown in Appendix A. At each iteration, we sample η fraction of the training observations without replacement, acting as a regularization, and the next tree uses the subsample. Friedman presented that $0.5 < \eta < 0.8$ can have good results for small and medium-sized training datasets.

2.5 Interpretation

A single decision tree is easy to interpret. The entire model can be fully represented in a simple two-dimensional graph that is easy to visualize. However, linear combinations of trees lose this important feature, therefore it must be interpreted in a different way.

2.5.1 Relative importance of predictor variables

Input covariates are rarely equally relevant. Usually, only a few substantially impact the response variable, and the vast majority are insignificant. Knowing each input variable's relative importance or contribution in predicting the response is often useful. For measuring of relevance of each predictor covariate X_l, Breiman et al [1] suggested

$$I_l^2(m) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 I(v(t) = l) \tag{2.27}$$

which take the sum of the improvement $\hat{\imath}_t^2$ in Poisson deviance loss function over the $J-1$ nodes of the tree. At each node t, input variable $X_{v(t)}$ divides the region into two subregions with the node t, which means indicator $I(v(t) = l) = 1$ if predictor $X_{v(t)}$ same as predictor variable X_l in the tth node, otherwise $I(v(t) = l) = 0$. The most important covariate is the one that accumulates the max improvement $\hat{\imath}_t^2$ in the Poisson deviance loss function. The above measure can easily expand to adapt to additive trees. Summing up the importance measure for each variable among all decision trees M, then taking the average of it, it will have a more stabilizing effect, which is more reliable than the above equation.

$$I_l^2 = \frac{1}{M} \sum_{m=1}^{M} I_l^2(m) \tag{2.28}$$

2.5.2 Partial Dependence Plots

After identifying the most relevant predictors, we attempt to understand the effect of variables on the response variable, claim frequency. Through plots, we can view the partial dependence of proximities on selected subsets of input variables. Consider X_S as subvector and $S \subset \{1, 2, \cdots, p\}$. Let C be the complement set, with $S \cup C = \{1, 2, \cdots, p\}$. A general function $f(\mathbf{X})$ has $f(\mathbf{X}) = f(\mathbf{X}_S, \mathbf{X}_C)$. Partial dependence functions can be estimated by

$$f_S(\mathbf{X}_S) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_S, \mathbf{x}_{iC}) \tag{2.29}$$

where $\{x_{1c}, x_{2c}, \cdots, x_{nc}\}$ are the values of complement set $\mathbf{X}_C$ in the training set. The partial dependence functions represent the effect of $\mathbf{X}_S$ on $f(\mathbf{X})$ after computing the average effects of the other predictors $\mathbf{X}_C$ on $f(\mathbf{X})$. In the thesis, we are interested in the effect of age and gender on claim frequency.

2.5.3 Feature Interaction: Friedman's H-statistic

In section 2.5.2, we focus on the marginal partial dependence of grouped covariates, however, the interaction of covariates plays an important role. By applying Friedman's H-statistic where the properties of the partial dependence function (2.29) will be adapted we can measure the effects of the interaction in the predictive GBM. If two covariates x_j and x_k are dependent, then the partial dependence of $f_s(\mathbf{X}_s)$ on $\mathbf{X}_s = (x_j, x_k)$ can be expressed as:

$$f_{jk}(x_j, x_k) = f_{jk}(x_j, x_k) - f_j(x_j) - f_k(x_k) \tag{2.30}$$

Friedman's H-statistic for an interaction between two covariates (x_j, x_k) is defined [5]

$$H_{jk}^2 = \frac{\sum_{i=1}^{n} [\hat{f}_{jk}(x_{ij}, x_{ik}) - \hat{f}_j(x_{ij}) - \hat{f}_k(x_{ik})]^2}{\sum_{i=1}^{n} \hat{f}_{jk}^2(x_{ij}, x_{ik})} \tag{2.31}$$

In section 4, we will plot a partial dependence plot and H-statistic for claims frequency vs. age by gender.

2.6 Model performance assessment

In regression, the predicted values are compared to values of observations, and measure model performance should be taken. Then, various model parameters are adjusted iteratively to obtain the optimal value of the performance index. The performance measure used is Concentration curves, root mean square error (RMSE) and deviance loss. The concentration curve is a good choice to validate two or more competing alternative models. Additionally, the discriminatory-free prices will be compared to best-prices and unawarness prices using a coefficient of determination based fidelity measure.

2.6.1 Concentration curves

Denuit et al., see [2], propose that concentration curves can be used to compare two competing alternatives

$$CC[Y, \pi(\mathbf{X}); \alpha] = \frac{E\big[Y \mathbb{1}[\pi(\mathbf{X}) \leq F_{\pi}^{-1}(\alpha)]\big]}{E[Y]} \tag{2.32}$$

Assuming data samples $(Y_i, \mathbf{X}_i)$, $i = 1, \cdots, n$, to be independent and identically distributed, the concentration curve can be estimated as:

$$\begin{aligned} \hat{CC}[Y, \pi(\mathbf{X}); \alpha] &= \frac{1}{n\bar{Y}} \sum_{i|\hat{\pi}(\mathbf{X}_i) \leq \hat{F}_{\pi}^{-1}(\alpha)} Y_i \\ &= \frac{\sum_{i|\hat{\pi}(\mathbf{X}_i) \leq \hat{F}_{\pi}^{-1}(\alpha)} Y_i}{\sum_{i=1}^{n} Y_i} \end{aligned} \tag{2.33}$$

where $\hat{\pi}$ denotes the estimated predictor. In this thesis, we will evaluate the performance of the estimated claim frequency that is obtained through GAM, GBM, and discrimination-free prices. For the out-of-sample dataset, we have $(\mathbf{X}_i, \hat{\mu}(\mathbf{x}_i))_{i=1}^{k}$ and sort the estimated claim frequency $\hat{\mu}(x_{(i)})_{i=1}^{k}$ in ascending order, that means $\hat{\mu}(x_{(1)}) \leq \hat{\mu}(x_{(2)}) \leq \cdots \hat{\mu}(x_{(k)})$.

$$\hat{CC}(\hat{\mu}(x); \alpha) = \frac{\sum_{i=1}^{k} n_i \mathbb{1}\{\hat{\mu}(x_i) \leq \hat{\mu}(x_{([\alpha \cdot k])})\}}{\sum_{i=1}^{k} n_i} \tag{2.34}$$

where $\alpha \in [0, 1]$, indicator $\mathbb{1}\{\hat{\mu}(x_i) \leq \hat{\mu}(x_{([\alpha \cdot k])})\}$ means that the ith number of claims add to the denominator if the $\hat{\mu}(x_i) \leq \hat{\mu}(x_{[\alpha \cdot k]})$. The concentration curve will be far from the 45-degree line if claim frequency explains much information about the number of injuries. If a prediction is bad, the concentration curves will coincide with the 45-degree diagonal line, while a good predictor should lie as far below the 45-degree diagonal line as possible.

2.6.2 Root Mean Square Error and CV-error

Root-mean-square error is usually used to measure the divergence between the predicted value by the target model and observed values

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (n_i - \hat{n}_i)^2} \tag{2.35}$$

where m is the number of out-sample datasets, n_i is observed value, $\hat{n}_i$ is predicted value. For selecting the best parameters to fit the gradient boosting model on the in-sample dataset, we apply grid search with k-fold cross-validation, $k = 10$. In addition to Poisson deviance, the root mean square error is also the criteria to validate the model's performance. For K-fold cross-validation, we split data into K equal-sized parts, then use $K - 1$ partition to fit the model and kth part to test data. Additionally, we calculate the prediction error of the fitted model when making predictions on the K-th fold of the data. The cross-validation estimate of prediction error is:

$$\text{CV}(f) = \frac{1}{n} \sum_{i=1}^{n} \omega_i L(y_i, \hat{f}^{-k(i)}(x_i)) \tag{2.36}$$

where $f^{-k(i)}$ is the fitted function calculated with data that without kth part data, L is the poisson deviance loss function, n is the number of in-sample data.

2.6.3 Quantifying differences in prices

In order to quantify differences in discriminatory-free prices (DFP) compared to best-estimate prices, and unawareness prices we use the following R^2 measure

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i^{\text{DFP}} - \hat{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \tag{2.37}$$

where $\hat{y}_i^{\text{DFP}}$ is the predicted value for ith observation using DFP, $\hat{y}_i$ the prediction of the reference (best-estimate or unawareness) model and $\bar{\hat{y}}$ is the mean value of the DFP model predictions. If the R-square value is close to 1, the DFP model captures the behavior of the reference model very well. Otherwise, the DFP model cannot approximate the reference model.

The R^2 measure from equation (2.37) has been used as a measure of fidelity when comparing a black-box model with a surrogate model, see [7]. Using R^2, we can see the proportion of the variation that the model for the best-estimate price is captured by the model for unawareness price and the model for the discrimination-free price.

Chapter 3

Data description

In this thesis, we use the data **pg15training** included in the R package **CASdatasets** for building the model to material claim frequency. Original data **pg15training** contains $100,021$ policies for private motor insurance and 14 covariate for each policy. Because the information is confidential, several categorical levels are unclear. Table 3.1 shows a detailed description of each covariate.

Covariate	Description
Gender	the gender of the car driver: Male, Female
Occupation	the occupation of the driver: Employed, Housewife, Retired, Self-employed, Unemployed
Age	age of the driver in years
Group2	the region of the drive home with ten classes
Density	the density of inhabitants (number of inhabitants per km2) in the city the driver of the car lives in
Type	the car type: A, B, C, D, E, F
Category	the car Category: Medium, Large, Small
Value	the car value in Euro
PolDur	Individual insurance policy duration in years
Adind	a dummy variable indicating a material cover
Group 1	the group of the car with twenty classes
Bonus	the bonus-malus
Exposure	total exposure in yearly units
Numtpbi	the number of third-party material claims

Table 3.1: Data description for French Motor Insurance.

To have a more comprehensive understanding of claims, we calculated the total number of policies and duration. Table 3.2 provides an overview of the total policies and the corresponding total duration of different claims from least 0 to max 7. We observe that $11,905$ claims are one-year long, and the smallest duration is 0.249 year with 7 claims.

Number of claims	0	1	2	3	4	5	6	7
Number of Policies	85,273	10,358	3,036	918	248	85	54	49
Exposures	564,248	62,244	8,423	1,589	337	57.8	34	29

Table 3.2: Number of claims and total durations

In the thesis, we will investigate how we treat the discriminatory covariate, w.r.t Gender impacts the claim frequency. In Figure 3.1, the covariate **Gender** has a different distribution over ages from 18 to 75.

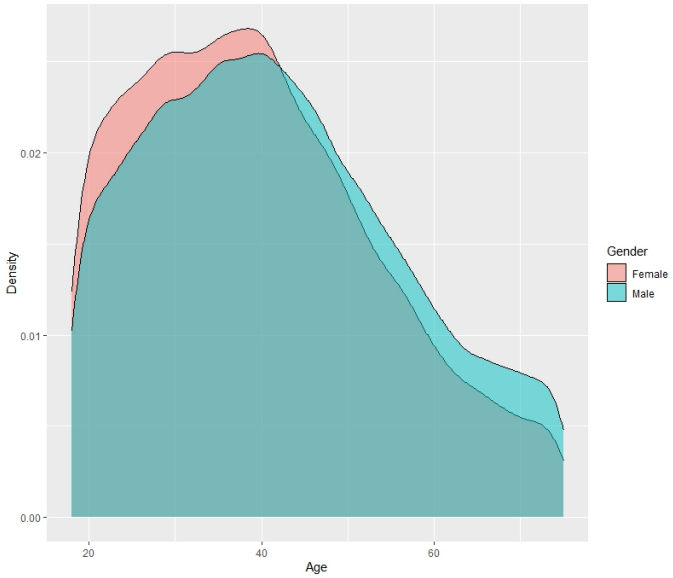


Figure 3.1: The Density of Age per Gender

It is not surprising to see that the driver occupation was employed with the greatest exposure compared to other occupational statuses. Because the description of the original data is not very comprehensive, it is not possible to explain the performance of the car in detail. According to the distribution of the type of car, the group of the car, and the cover material of the car, it shows that A type car belongs to the 11th classes group with 1 material cover has the longest duration. It is not the most expensive car with the greatest exposure, but it is worth about 70,000 euros. Living in an area with high residential density does not give you the greatest exposure. Exposure gradually rises to the top as residential densities start from a low of $14/km^2$ to $240/km^2$ before dropping again. In France, 18 years are allowed to start driving car. At age 40, the exposure gradually rises to the maximum value and then decreases with age. Compared with large cars, medium-sized cars are more exposed. Not surprisingly, the drivers with the fewest bonuses also had the most exposure. The longer contracts Those policyholders have with insurance companies, the lower exposure belong to they. Plotting of covariates is shown in Figure B.2 in Appendix.

Chapter 4

Model building

We build models for claim frequency $N \mid \mathbf{X}, \omega \sim \text{Pois}(\omega\mu(\mathbf{x}))$ with GAM and GBM for each of the $100,021$ individual motor insurance policies with 13-dimensional feature covariates, where $i = 1, \cdots, 100,021$.

$\mathbf{x}_i = (\text{Gender}_i, \text{Occupation}_i, \text{Age}_i, \text{Group2}_i, \text{Density}_i, \text{Type}_i, \text{Category}_i, \text{Poldur}_i, \text{Group1}_i, \text{Bonus}_i, \text{Value}_i, \text{Adind}_i, \text{YearDura}_i)$

Furthermore, we have a number of material claims as the response variable, $N_i \geq 0$, and yearly duration as weight $\omega_i \in (0.2491, 1)$. To compare the models' predictive performance, we partition our data set into two sets: the in-sample set denoted $\mathcal{D}$ for fitting models and the out-sample as $\mathcal{T}$ for comparing the predictive performance of models.

By randomly allocating 80% of data from the original dataset to $\mathcal{D}$ and holding 20% of the data to $\mathcal{T}$, we have

$$\mathcal{D} = \{(N_i, \mathbf{x}_i, \omega_i), i = 1, \cdots, 800,16\}$$

$$\mathcal{T} = \{(N_k, \mathbf{x}_k, \omega_k), k = 1, \cdots, 20,005\}$$

For fitting GAMs and GBMS, we will use the same data partitioning as shown above.

4.1 Generalized additive model (GAM)

4.1.1 GAM with Gender

Initially, we need to convert covariate **Bonus** to a factor in GAM. By adapting quantile as cut points, we divide the range of covariate **Bonus** into four different continuous classes, $x_{\text{Bonus}} \in [-50, -40] = 1$, $x_{\text{Bonus}} \in (-40, -30] = 2$, $x_{\text{Bonus}} \in (-30, 10] = 3$ and $x_{\text{Bonus}} \in (10, 150] = 4$.

There are four continuous covariates (**Age**, **Value**, **Density**, **Poldur**) and eight categorical covariates, **Gender**, **Type**, **Category**, **Occupation**, **Adind**, **Group2**, **Group1**, **Bonus**. We select the most extended duration as the reference level in GAM ($\text{Ref}_{\text{Gender}} = \text{Male}, \text{Ref}_{\text{Type}} = \text{A}, \text{Ref}_{\text{Category}} = \text{Medium}, \text{Ref}_{\text{Occupation}} = \text{Employed}, \text{Ref}_{\text{Adind}} = 1, \text{Ref}_{\text{Group2}} = \text{L}, \text{Ref}_{\text{Group1}} = 10, \text{Ref}_{\text{Bonus}} = 1$).

The thesis focuses on the two-way interaction between **Gender** and other covariates. We fit the GAM *Sat.GAM* shown in table 4.1 to the in-sample data set $\mathcal{D}$ and add all possible two-way interactions between **Gender** and other covariates. However, covariates **Age** and **Value** are treated as natural cubic splines.

Sat.GAM
SkadeFrekvens ∼ Gender + Type + Category + Occupation + Adind + Group2 + Group1 + s(Age, bs = "cr") + Density + Poldur + Bonus + s(Age, bs = "cr", by=Gender) + s(Value, bs="cr") + s(Value, bs="cr", by=Gender)+ Gender:Category + Gender:Occupation + Gender:Adind + Gender:Group2 + Gender:Group1 + Gender:Bonus, family=quasipoisson(link = "log"), data = train, weights = YearDura)

Table 4.1: GAM with all covariates and all possible two-way interactions between Gender and other covariates

After dropping covariates that are not significant at 5% level each time, we have the Model *Gam.AllSig* shown in the table 4.2, where the interaction between **Gender** and **Age**, the interaction between **Gender** and **Category** are significant at 5% level.

GAM.AllSig
SkadeFrekvens ∼ Gender + Type + Category + Occupation + Adind + Group2 + Group1 + Density +Poldur + s(Age, bs = "cr") + Bonus + s(Age, bs = "cr", by=Gender) + Gender:Category, family=quasipoisson(link = "log"), data = train, weights = YearDura)

Table 4.2: GAM model with covariates are all significant at 5% level

We find that the partial interaction between **Age** and **GenderMale** is not significant at 5% level. Therefore, we gradually increase the knot for the interaction by one. All covariates in the Model *GAM.Final* shown in the table 4.3 is significant at the 5% level. Generalized cross-validation(GCV) is applied in the algorithm to determine good smoothing parameter λ. For model*GAM.Final*, the optimal smoothing parameter (λ is $\lambda_{\text{Age}}, \lambda_{\text{Age:Female}}, \lambda_{\text{Age:Male}}$) = (10, 11, 772).

GAM.Final
SkadeFrekvens ∼ Gender + Type + Category + Occupation + Adind + Group2 + Group1 + Density +Poldur + s(Age, bs = "cr") + Bonus + s(Age, bs = "cr", by=Gender, k = 4) + Gender:Category, family=quasipoisson(link = "log"), data = train, weights = YearDura)

Table 4.3: GAM with each covariate is significant at 5% level, but the knot of interaction between age and gender increases to 4.

Additionally, analyzing poisson deviance and RMSE is shown in Table 4.4 for the Model *Sat.GAM*, Model *GAM.ALLSig* and Model *GAM.Final*. Finally, we determine that Model **Gam.Final** as our final generalized additive model with discriminatory Gender.

Model	In-sample loss	Out-of-sample loss	Out-of-sample RMSE
Saturated Model	0.76724	0.50525	0.40248
GAM.AllSig	0.49442	0.49941	0.40302
GAM.Final	0.49441	0.49935	0.403054

Table 4.4: Comparison of In-sample Poisson deviance, out-of-sample Poisson deviance, and RMSE of out-of-sample among models

Figure 4.1 shows the effect for cubic spline **Age** and the partial effect between **Age** and **Gender**. Effective degree freedom(EDF) for smooth terms, s(Age), s(age):-GenderFemale and s(age):GenderMale, are $7.033, 2.968, 1.035$, respectively. In the left upper panel, we can see the highest claim frequency is at the age allowed to

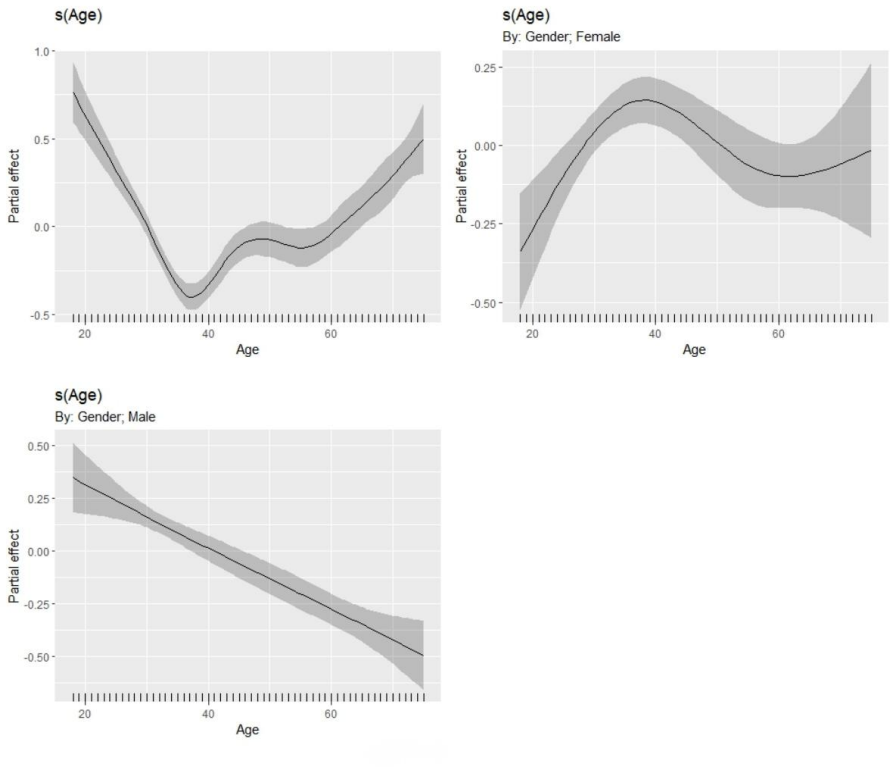


Figure 4.1: Partial effect of cubic spline for covariate Age, the interaction between Age and Gender

drive, 18 years in France. At about age forty, the lowest claim frequency occurs. After forty, however, claim frequency continued to increase, although there was a local minimum claim frequency near age sixty. The left upper panel shows the claim frequency of males decreased as age increased. The right upper panel shows that female has the highest claim frequency at about age thirty-eight from beginning to driving. Subsequently, the frequency declined until the age of sixty and then began to increase.

4.1.2 GAM without Gender

For determining how discriminatory covariate *Gender* affects the claim frequency. A model without *Gender* will be briefly presented here.

GAM.withoutGend
SkadeFrekvens ∼ Type + Category + Occupation + Adind + Group2 + Group1 + Density +Poldur + s(Age, bs = "cr") +Bonus, family=quasipoisson(link = "log"), data = train, weights = YearDura)

Table 4.5: GAM without Gender

More detail about the model will be introduced in the comparison among different models.

4.1.3 Discrimination-free analysis with GAM

After obtaining the suitable generalized additive model with/without Gender, we calculate discrimination-free price as mentioned in Section 2.1. For the computation

of the discrimination-free insurance price, we need the proportion of each class of gender of the total portfolio. $\mathbb{P}(\mathbf{D} = \text{female}) = 0.365$ and $\mathbb{P}(\mathbf{D} = \text{male}) = 0.634$. From Figure 4.2, we find bias in the discrimination-free price, which means bias needs to be considered in the discrimination-free analysis, and the price should be corrected. The average predicted price is $\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(\mathbf{x}_i, \mathbf{d}_i) = 0.1652$ and the average discrimination-free price is $\frac{1}{n}\sum_{i=1}^{n}\hat{h}(\mathbf{x}_i) = 0.1669$. We have a negative bias of 1% of μ. By using an adjusted marginal distribution $\mathbb{P}^*(\mathbf{D})$, we may obtain a bias-corrected price. Because discriminatory covariate gender has just two classes, we set $\frac{1}{n}\sum_{i=1}^{n}\hat{h}^*(\mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(\mathbf{x}_i, \mathbf{d}_i) = 0.1652$, adjusted $\mathbb{P}^*(\mathbf{D} = \text{female}) = 0.405$, which is higher than the empirical proportion $\mathbb{P}(\mathbf{D} = \text{female}) = 0.365$.

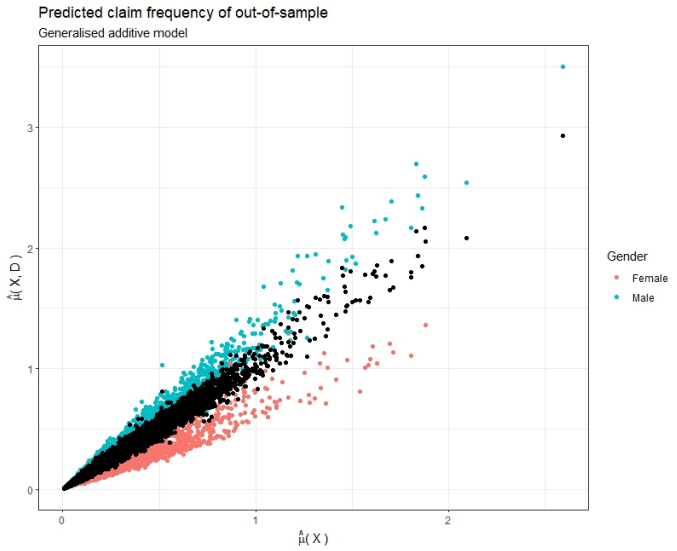


Figure 4.2: Predicted claim frequency obtained by GAM without Gender against predicted claim frequency through GAM with Gender. Red points are estimated claim frequency for females, blue points are claim frequency for males, and black points are discrimination-free prices

GAM	In-sample loss	Out-of-sample loss	Out-of-sample RMSE
Model with Gender	0.494418	0.499355	0.403054
Model without Gender	0.501764	0.505254	0.40594
Discrimination-free	0.494418	0.504469	0.40578

Table 4.6: In-sample Poisson deviance, out-of-sample Poisson deviance, and root mean square error of out-of-sample for GAM with Gender, GAM without Gender, and Discrimination-free, respectively.

GAM	Reference model	Out-of-sample R^2
Best-estimate	Unawareness	0.89463
Best-estimate	DFP	0.91076
DFP	Unawareness	0.98463

Table 4.7: Coefficient of determination to quantify the differences between best-estimate price and unawareness price, best-estimate price and discrimination-free pricing, unawareness price and discrimination-free pricing with GAM.

From Table 4.6, we can see that the performance of discrimination-free price is better than unawareness price regardless of training or testing data set. Furthermore, a discrimination-free price has a lower RMSE value of out-of-sample than the

unawareness price. The best-estimated price has the best-predicted performance. Additionally, Figure 4.2 shows $\hat{\mu}(\mathbf{x})$ against $\hat{\mu}(\mathbf{x}, \mathbf{D})$. The predicted $\hat{\mu}^*(\mathbf{x})$ is obtained from the discrimination-free analysis that lies between the estimated claim frequency for males and females. Table 4.7 shows a coefficient of determination type fidelity measure to quantify the differences between DFPs and BE prices/unawareness prices. The DFP model captures the behavior of the BE price model more than the unawareness price model captures the behavior of the BE price model. However, the R^2 is almost 1 between the DFP model and the unawareness price model, which quantifies the differences between these two models are quite small.

4.2 Gradient boosting model (GBM)

4.2.1 GBM with gender

Next, we consider a Poisson GBM; see Section [2.3]. First, we tune hyperparameters by grid search, which involves systematically iterating over each combination of hyperparameter values, the number of trees (J), subsample size, which is the percentage of training data for each tree, tree depth is restricted to 2, and learning rate ($\nu \in [0.01, 0.3]$). Hastie et al., see [6], point out that the number of terminal nodes of the trees, $4 \leq J \leq 8$ works well in the context of boosting. However, in the thesis, we will tune J from 1 to 8. Among the 699 different Gradient boosting Poisson models, we select the best ten based on RMSE and Poisson deviance criteria for the out-of-sample data set. In Table 4.8, the first GBM with 399 optimal trees, learning rate 0.14, and tree size $J = 4$ has the lowest out-sample poison deviance chosen as our GBM.

Model	J	ν	optimal trees	subsample	Out-of-sample Loss	out-of-sample RMSE
1	4	0.14	399	0.75	0.4850212	0.3968325
2	3	0.08	403	1	0.4851113	0.3964255
3	5	0.09	593	1	0.4852663	0.3958001
4	7	0.2	375	1	0.4853259	0.3964711
5	2	0.23	278	0.75	0.4858241	0.3969464
6	1	0.16	399	0.75	0.4860617	0.872603
7	6	0.17	599	1	0.4862308	0.3952606
8	4	0.07	581	0.5	0.4865334	0.3958239
9	3	0.12	581	0.5	0.4869393	0.3970489
10	2	0.14	393	0.5	0.4870474	0.3972603

Table 4.8: Performance indicators of GBMs and their respective parameters

Figure 4.3 shows which covariate impacts claim frequency most. The order of influence strength from strong to weak for covariate is *Bonus*, *Age*, *Density*, *Group1*, *Gender*, *occupation*, *Group2*, *Type of car*, *Policy duration*, *Value*. Covariate *Gender* has the fifth strongest effect.

Next, we analyze which covariate has interaction with any other covariate and two-way interaction strength with Friedman's H-statistic. Figure 4.4a shows that *Gender* and *Age* give the strongest interaction with others. Overall, interaction effects between the covariate are not strong about around 10% of prediction variance. Additionally, Figure 4.4b displays the two-way interaction with H-statistic between *Gender* and each other covariate. Interaction between *Age* and *Gender* significantly affects claim frequency.

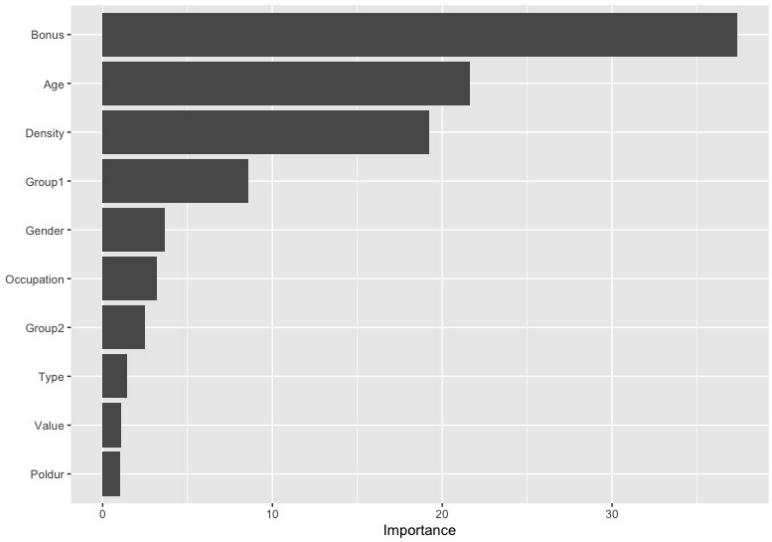
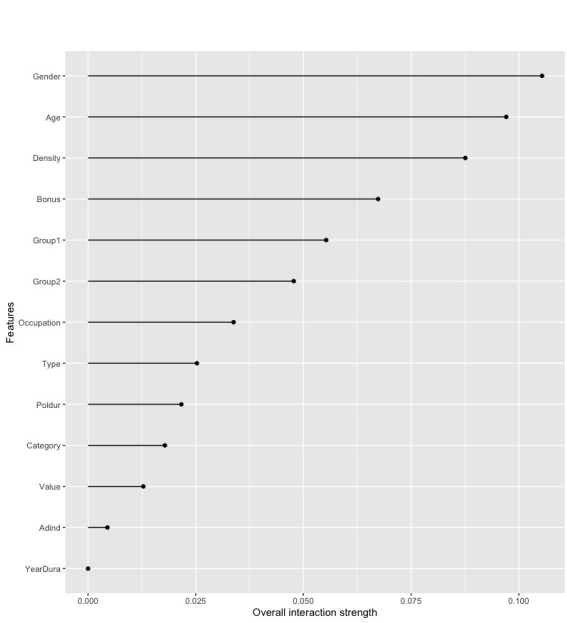
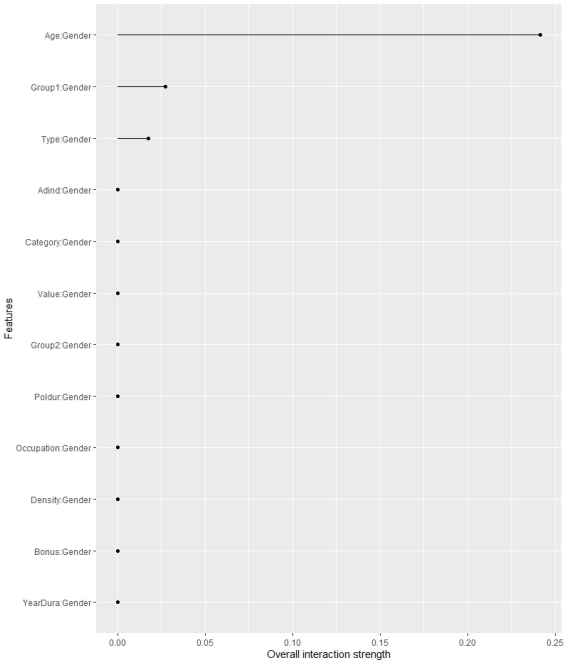


Figure 4.3: Relative importance for the motor insurance.



(a) The interaction strength for each covariate with all other covariates.



(b) The two-way interaction between gender and any other covariate strength

Figure 4.4: Interaction strength

4.2.2 GBM without gender

How does the GBM change if *Gender* is excluded from our model? Same as the previous step, first, we tune the hyperparameter by grid search. We find that model with optimal 474 trees, learning rate $\nu = 0.18$, tree size $J = 3$, and 80% of subsample has the best-predicted performance.

4.2.3 GBM with non-discriminatory Pricing

As in the Section 4.1.3, we calculate the discrimination-free price, $\hat{\mu}^*_{\text{GBM}}(\mathbf{X})$. The number of claims, $\mathbf{N}^*|(\mathbf{X}, \omega) \sim \text{Pois}(\omega\hat{\mu}^*_{\text{GBM}})$.

From Table 4.9, we can see that the performance of discrimination-free price is better than unawareness price with the testing data set. Furthermore, a discrimination-free price has a lower RMSE value than an unawareness price. Same as in the GAM, the best-estimated price has the best-predicted performance. As shown in the table 4.10 DFP captures better behavior of best-estimate prices than unawareness prices capture BE prices. Compared to R^2 shown in Table 4.7, DFP and Unawareness prices have smaller differences between best-estimate prices. Figure 4.5 shows $\hat{\mu}(\mathbf{x})$ against $\hat{\mu}(\mathbf{x}, \mathbf{D})$. The predicted discrimination-free prices $\hat{\mu^*}(\mathbf{x})$ lies between the estimated claim frequency for males and females. In contrast to Figure 4.2, it is obvious that the range of predicted claim frequency by the GBM is broader and more stable.

GBM	Out-of-sample loss	Out-of-sample RMSE
Best-estimate price	0.4863853	0.3960513
Unawareness price	0.5083443	0.4028289
Discrimination-free price	0.491571	0.4020983

Table 4.9: out-of-sample Poisson deviance, and root mean square error for GBM with Gender, GBM without Gender, and Discrimination-free, respectively.

GBM	Reference model	Out-of-sample R^2
Best-estimate	Unawareness	0.92172
Best-estimate	DFP	0.92824
DFP	Unawareness	0.98201

Table 4.10: Coefficient of determination to quantify the differences between best-estimate price and unawareness price, best-estimate price and discrimination-free pricing, unawareness price and discrimination-free pricing with GBM.

4.3 Comparison Between GAMs and GBMs

To compare the predicted performance of models, we plot concentration curves. Figure 4.6 shows that the GAM model performs better than GBM for small claims. As the number of claims increases, the predictive ability of the GBM gradually improves. GBM performs better than GAM as the number of claims increases. The figure shows clearly that the best-estimate price with GBM is better than the other models. DFP with GBM also has better performance than prices obtained from GAM as claims number increases. Especially at the end of the CC plot, GBM performs better than GAM because GAM almost lost much information at a large number of claims.

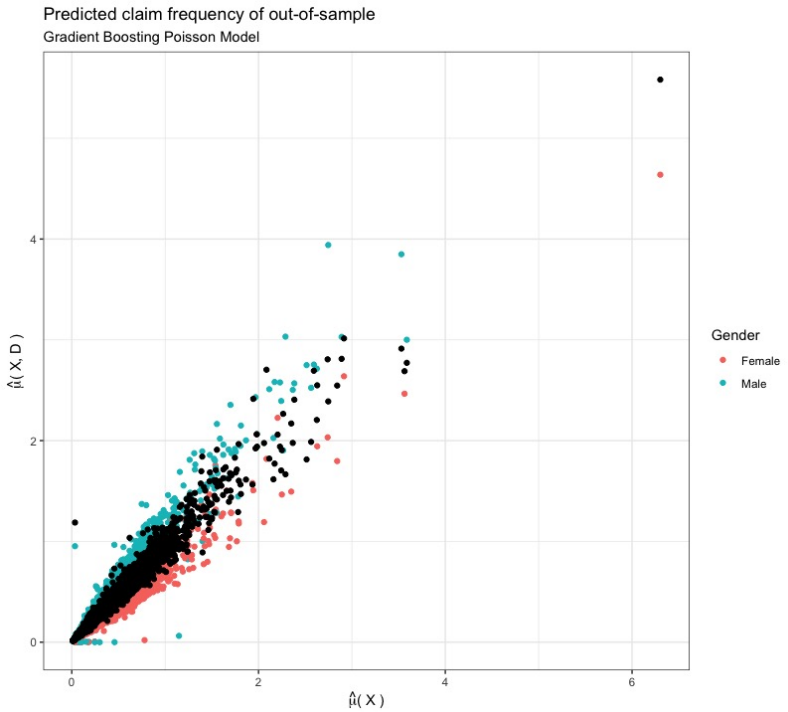


Figure 4.5: Predicted claim frequency obtained by GBM without Gender against predicted claim frequency through GBM with Gender. Red points are estimated claim frequency for females, blue points are claim frequency for males, and black points are discrimination-free prices

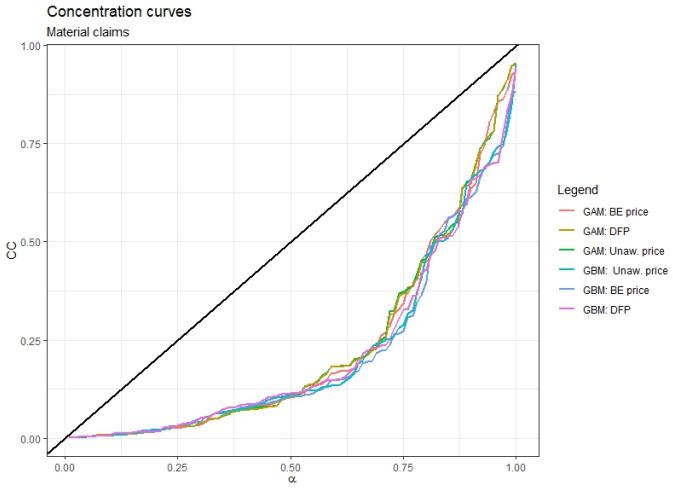
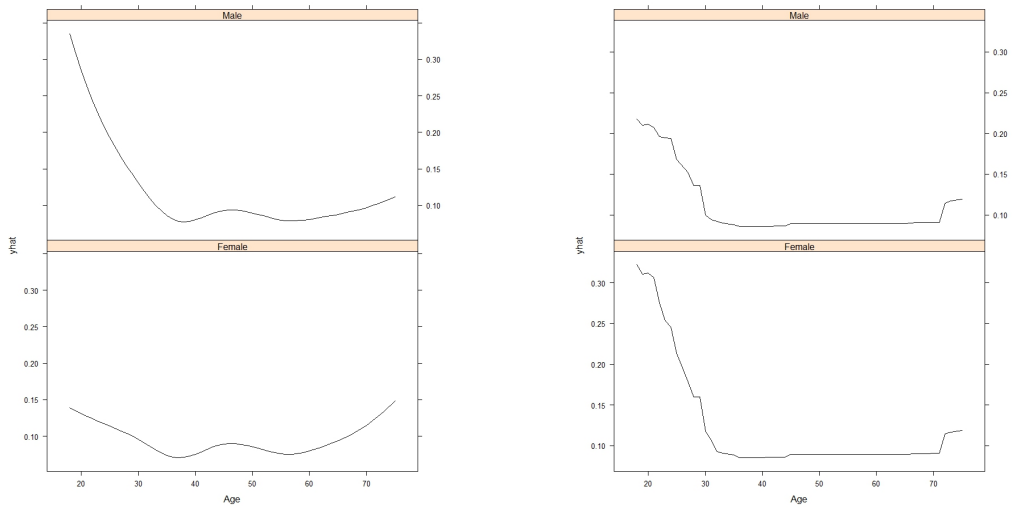


Figure 4.6: Concentration curves among best-estimate price, unawareness price, and discrimination-free prices based on GAM and GBM.

Figure 4.7 shows the partial dependence plot for the best-estimate price obtained from GAM and GBM. Regardless of sex, the highest value in GAM and GBM occurs when driving is allowed. The difference in claim frequency for males and females is less for GBM than for GAM. Males have a higher claim frequency than females in the GAM. However, GBM is exactly the opposite. Females reach a higher claim frequency than males.



(a) PDP for best-estimate price with GAM (b) PDP for best-estimate price with GBM

Figure 4.7: Comparison of the partial dependence plot for best-estimate price with GAM and GBM

Chapter 5

Simulation

In this part, with the optimal GAM and GBM, using the cross-validation method, we simulate 100 times to predict the number of claims, $N^{(k)}|(X_{ij}^{(k)}, \omega_i^{(k)}) \sim \text{Pois}(\omega_i \mu_i^{(k)}), k = 1, \cdots, 100$. Figure 5.1 shows that the GBM has a lower RMSE value than the GAM over 100 simulations, regardless of whether the model includes gender or discrimination-free analysis in the out-sample data set.

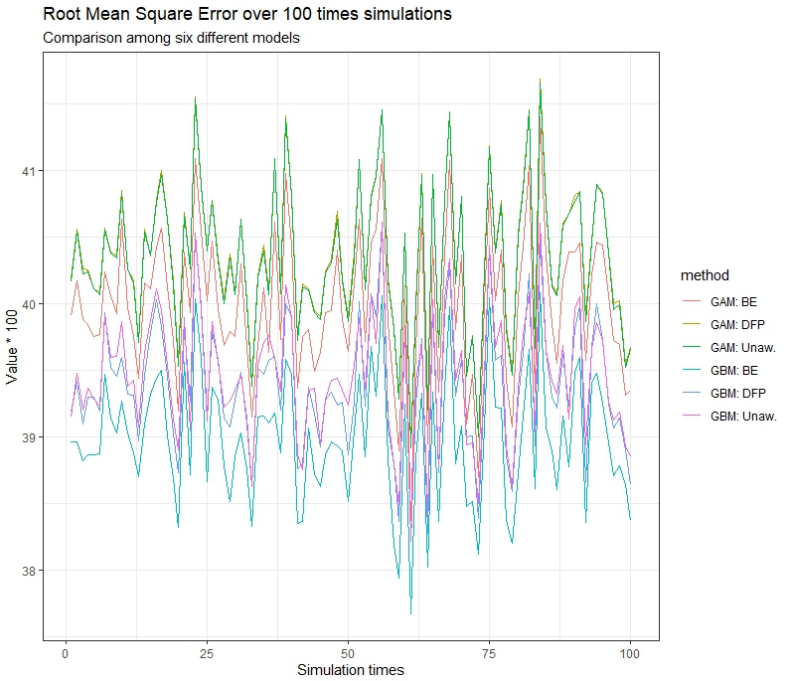


Figure 5.1: RMSE

Additionally, to examine how the prices vary with GAM and GBM, we compare R^2 between the reference models and DFPs, see equation 2.37. Figure 5.2 shows the differences among best-estimated price, unawareness price, and discrimination-free price for GAMs and GBMs. DFPs have smaller differences with BE prices obtained from GBM than DFPs compared to BE prices obtained from GAM. For GBM, the differences between unawareness prices and DFPs are fairly small. It seems R^2 between DFPs and BE prices for GBM overlay R^2 between DFP and unawareness prices for GAM. We assume that insufficient data about large claim numbers causes an unstable situation with the tail.

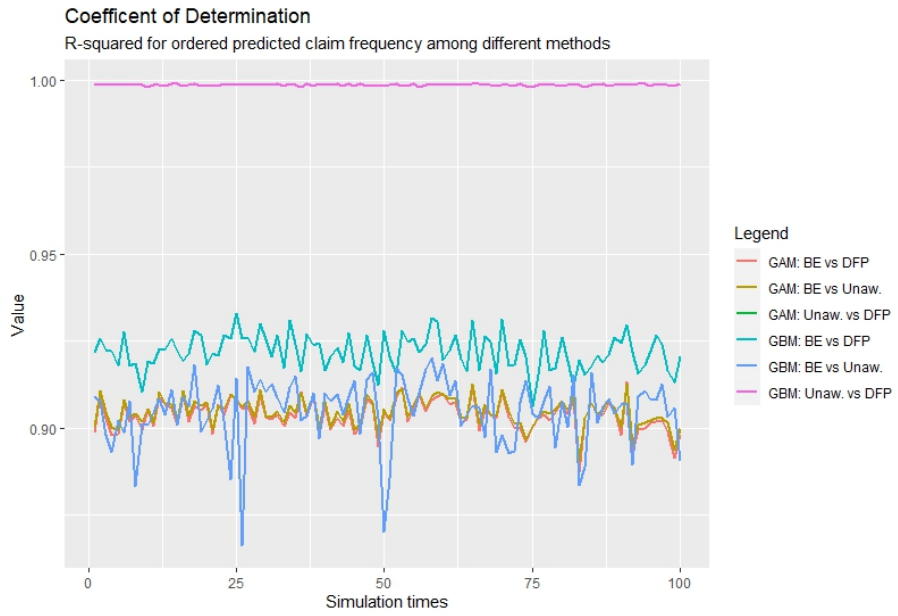


Figure 5.2: Coefficient of determination

Moreover, we plot the local coefficient of determination for more details about local prediction performance. First, we split $20,005$ estimated claim frequency $\hat{\mu}_i, i = 1 \cdots, 20,005$ into 400 groups and set it on the ascending sort. In Figure 5.3, we find that most regions are similar, the closer to the tail the greater the difference. The trend for R^2 between $\hat{\mu}(\mathbf{x})$ and $\hat{\mu}^*(\mathbf{x})$ for GBM shows that the biggest differences occur compared to others. Unawareness prices have the lowest differences with DFP for GBM.

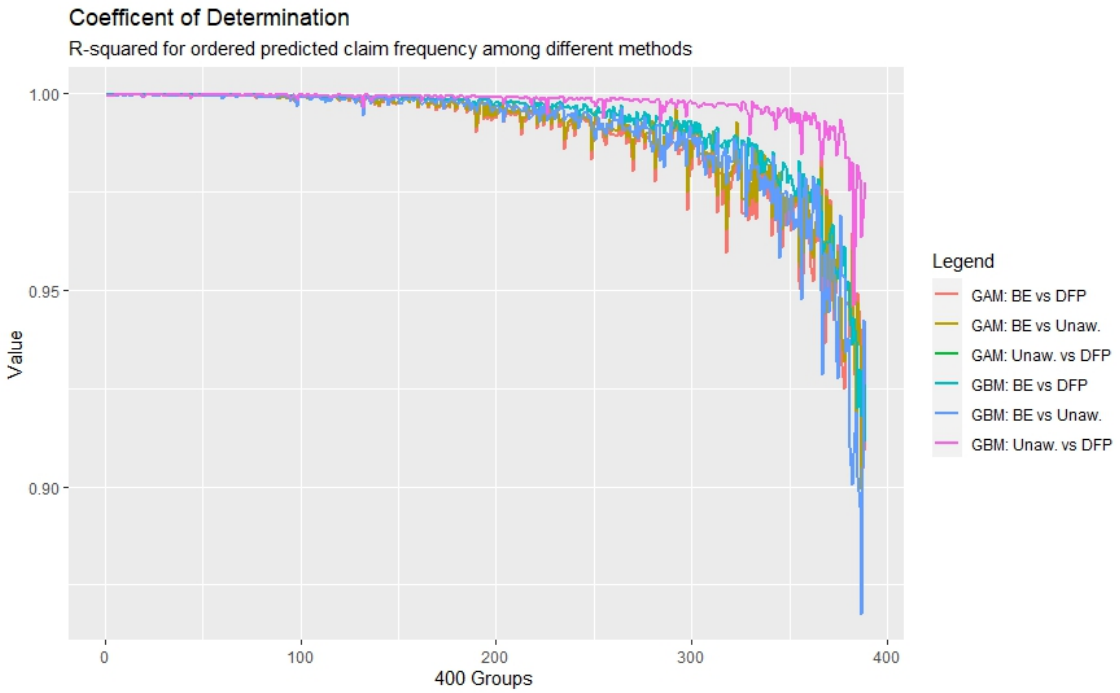
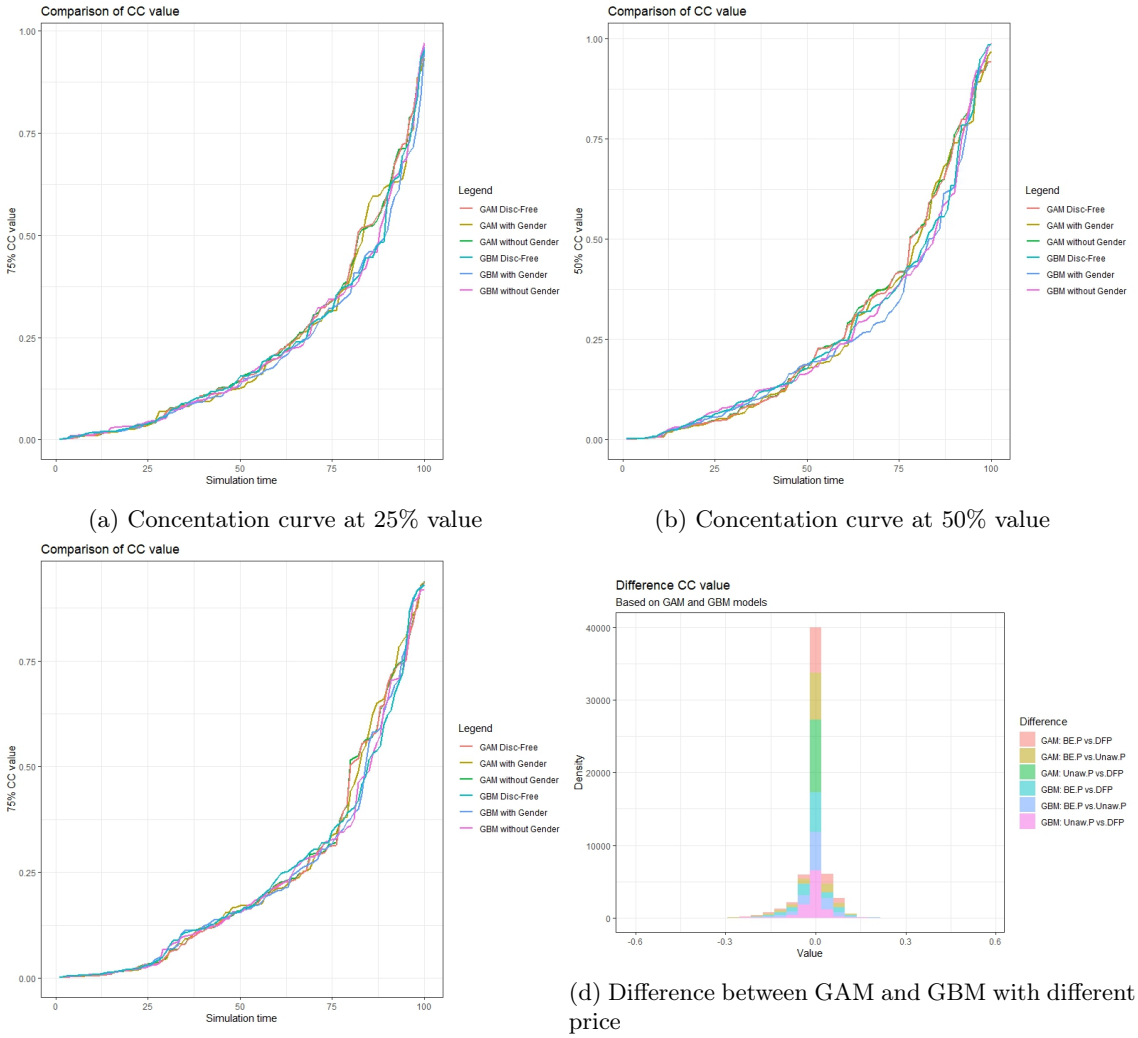


Figure 5.3: Grouped R-square for the best-estimate price, unawareness price and discrimination-free price for GAMs and GBMs

Partial Figure 5.4a, 5.4b, 5.4c in Figure 5.4 show the concentration curves value over 100 simulations for best-estimate prices, unawareness prices, and discrimination-free prices for GAM and GBM via 25%, 50% and 75%, which show prices obtained from GBM perform better than GAM in most situations. It reveals that GBM captures more information about covariates relationships than GAM and provides a wider estimated range. Moreover, differences between DFPs and best-estimate or unawareness price are shown in histogram 5.4d. It is obvious that DFPs get the nearest to the unawareness prices for GBM.



(a) Concentation curve at 25% value

(b) Concentation curve at 50% value

(c) Concentation curve at 75% value

(d) Difference between GAM and GBM with different price

Figure 5.4: Comparion of concentration curves over 100 simulations

As shown in the partial dependence plot 4.7, we are interested in the effect of incorporating those age ranges with significant claim frequency. So, we consider merging individuals over 70 years old. Using GAM and GBM, we get the proportion between the estimated number of claims and total observed data is 0.16. We also merged the individuals' age under 30 in the data, and the proportion obtained is slightly lower.

Chapter 6

Conclusion and Discussion

In the thesis, we have studied discrimination-free insurance pricing built on Gradient boosting machines and Generalized additive models. According to European law, sensitive information w.r.t the gender and ethnicity of policyholders should not be used in insurance pricing. It will still cause indirect discrimination because we can derive the protected features from associated information although this type of information is not included in insurance pricing.

By grid search, we find the optimal parameters for learning rate, bag fraction, tree size, and optimal trees of GBM. Using partial dependence plot and Friedman's H-statistic test show the most important individual feature and two-way interaction. Moreover, comparing local performance through concentration curves and R-squared shows that GAMs perform better than GBMs when small insurance claims occur. In contrast, GBMs perform better than GAMs with large insurance claims.

Unsurprisingly, the best-estimate price has the best predictive performance for GBMs and GAMs. Under investigation in the thesis, we find that the discrimination-free prices lie closer to the unawareness prices for GBM than GAM. Compared to the differences between the best-estimate and discrimination-free prices, the discrimination-free prices differ less from the best-estimate prices for GBM. Even though discrimination-free insurance pricing is not unbiased, it is easy to correct this bias by adapting the empirical marginal proportion.

In the thesis, we have shown that GBMs have better performance than GAMs. One is that GBM can produce interpretable models and naturally incorporate a mixture of numeric, categorical covariates and missing values. The other is that it may indicate that GAMs do not capture the dependence on gender well enough. The relative importance plot for GBM reveals that discriminatory covariate **Gender** does not have the greatest impact. Although our data set contains $100,021$ policies, most are distributed in 0 claim numbers. We believe that a bigger and more widely distributed dataset can enhance the impact of each covariate, even two-way interaction. Due to time limitations, we restrict grid search for training GBM when tuning the hyperparameters. To some extent, this limits the predictive power of the GBM models.

Several important changes need to be made for future works. Firstly, a more complex simulation method should be implemented in simulation for building models and exploring the effect of covariates. Secondly, model interpretation methods should be more proficiently adopted in the investigation.

Appendix A

Stochastic Gradient Boosting Algorithm

Algorithm 1. Stochastic Gradient Boosting
- Initialize $F_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
  - **For** $m = 1, \cdots, M$:
  Randomly choose subsample $\eta \cdot n$ from $\{y, \mathbf{x}_i\}_{i=1}^{N}$ entire training data sample, without replacement.
  - **For** $i = 1, \cdots, \eta \cdot n$, compute

$$g_{im}(\mathbf{x_i}) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

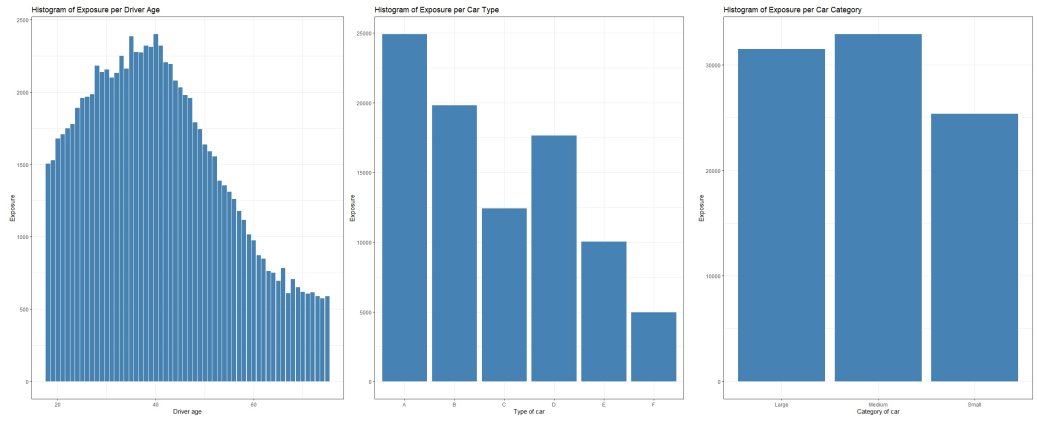
fit a regression tree to the targets g_{im} giving terminal regions $R_{jm}, j = 1, \cdots, J_m$
  - **For** $j = 1, \cdots, J_m$ compute

$$\gamma_{jm} = \arg\min_\gamma \sum_{\mathbf{x_i} \in \mathbf{R_{jm}}} L(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$$

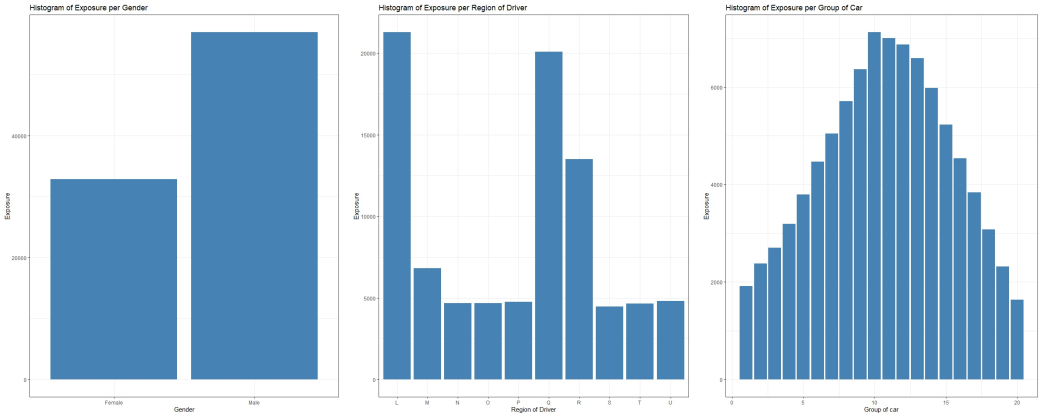
Update $F_m(x) = F_{m-1}(\mathbf{x}) + \nu \sum_{j=1}^{J} \gamma_{jm} 1(x \in R_{jm})$
  - Output $F_{GBM}(x) = F_M(x)$

Appendix B

Supplementary Figures

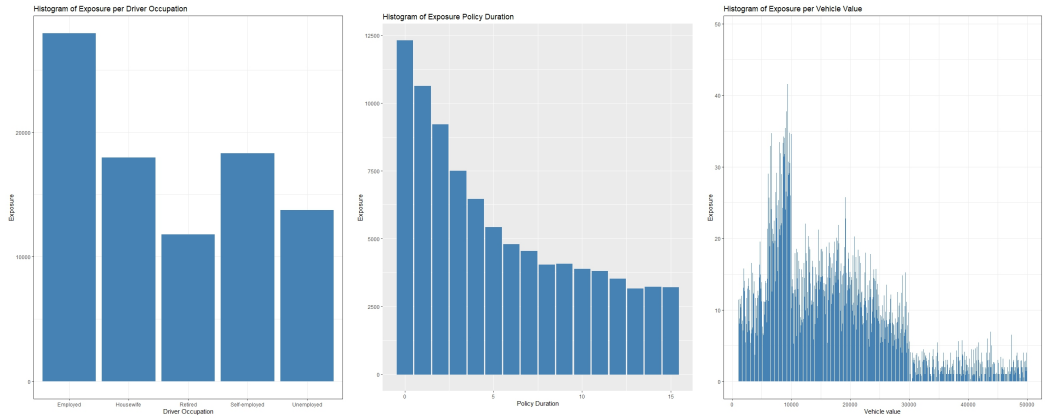


(a) Histogram of Exposure per Driver Age (b) Histogram of Exposure per Car Type (c) Histogram of Exposure per Car Category



(d) Histogram of Exposure per Gender (e) Histogram of Exposure per Region of Driver (f) Histogram of Exposure per Group of Car

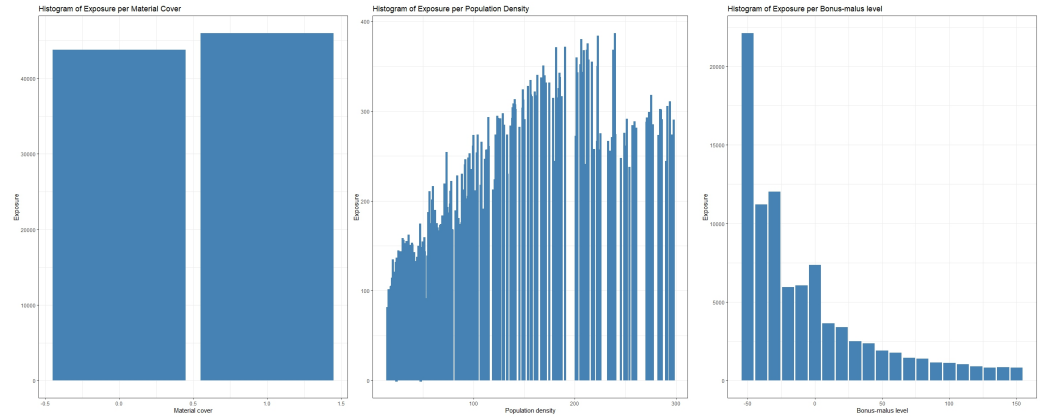
Figure B.1: Histogram of covariates



(a) Histogram of Exposure per Driver Occupation

(b) Histogram of Exposure Policy Duration

(c) Histogram of Exposure per Vehicle Value



(d) Histogram of Exposure per Material Cover

(e) Histogram of Exposure per Population Density

(f) Histogram of Exposure per Bonus-malus level

Figure B.2: Histogram of covariates

Bibliography

[1] Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*, 1 ed. Chapman and Hall, 1984.

[2] Denuit, M., Sznajder, D., and Trufin, J. Model selection based on lorenz and concentration curves, gini indices and convex order. *Insurance: Mathematics and Economics vol.89* (2019), 128–139.

[3] European Council. Council directive 2004/113/ec - implementing the principle of equal treatment between men and women in the access to and supply of goods and services. *Official Journal of the European Union L 373* (2004), 37–43.

[4] Friedman, J. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics vol. 29, No.5* (2001), 1189–1232.

[5] Friedman, J. H., and Popescu, B. E. Predictive learning via rule ensembles. *The Annals of Applied Statistics vol.2, No.3* (2008), 916–954.

[6] Hastie, T., Tibshirani, R., and Freidman, J. *The Elements of Statistical Learning: Data Mining, inference and prediction*, 2 ed. Springer, 2017.

[7] Henckaerts, R., Antonio, K., and Côté, M.-P. When stakes are high: Balancing accuracy and transparency with model-agnostic interpretable data-driven surrogates. *Expert System with Applications vol.202* (2022).

[8] Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. Discrimination-free insurance pricing. *ASTIN Bulletin 52* (2022), 55–89.

[9] Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. What is fair? proxy discrimination vs. demographic disparities in insurance pricing, (2023). Available at SSRN: https://ssrn.com/abstract=4436409.

[10] Ohlsson, E., and Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, 2010.

[11] Wüthrich, M. V., and Buser, C. Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper* (2023), 16–68.