RISK PREMIUM PRICING METHODS IN NON-LIFE INSURANCE FOR ACCURATE ESTIMATION AND UNCERTAINTY QUANTIFICATION

Olivier de Groot

A thesis submitted for the degree of MSc. of Applied Mathematics

December, 2017

"Prediction is very difficult, especially if it's about the future."
–Nils Bohr, Nobel laureate in Physics

Abstract

Free competition in the insurance markets increases the competitiveness and lowers the premiums. If insurers lower their premiums without having a model that accurately quantifies the expected claim size, they can be in serious trouble. This research aims to accurately model the premiums and quantify the uncertainty involved using historic claims data from an insurer.
The current approach, Generalized Linear Models (GLMs), is compared to some Machine Learning techniques: Random Forests (RFs) and Gradient Boosting Machines (GBMs). Insights gained from these models and other methods (MARS) are then used to improve the GLMs.
Bayesian Additive Regression Trees (BART) and Hierarchical Models (HMs) are then used to quantify the uncertainty. HMs provides the insurer with the means to make proper credible intervals for the total expected claim size of the active portfolio. The HMs also allow the use of risk premium principles that include measures of uncertainty in the pricing of premiums.
All relevant models are applied on the dataset of the holdout year. It is apparent that the GLMs and HMs provide too low estimations of the premiums when the profit is tracked. It is therefore prudent to either use the HM with a risk premium principle that incorporates a percentage of the standard deviation in the estimation of the premiums or apply the RF model. The model lift shows that the Machine Learning techniques are better at recognising the risky policies from the non-risky.
We recommend the insurer to use RFs to price the premiums and HMs to measure the uncertainty of the active portfolio.
It is recommended for further study to either apply other techniques to further improve the predictive performance, to improve the structure of the Hierarchical Model or to include left truncation and right censoring into the model.

Keywords: non-life insurance, number of claims, claim severity, risk premiums, Machine learning techniques, predictive models, uncertainty quantification, Generalized Linear Models, Random Forests, Gradient Boosting Machines, Hierarchical Modeling, MCMC, Bayesian Additive Regression Trees, Multivariate Adaptive Regression Splines, Generalized Additive Models.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms

AIC	Akaike Information Criterion
AIC	Deviance Information Criterion
BART	Bayesian Additive Regression Trees
GAM	Generalized Additive Model
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
HM	Hierarchical Modelling.
JAGS	Just Another Gibbs Sampler.
MARS	Multivariate Adaptive Regression Splines
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
RF	Random Forest
RMSE	Root-mean-square error

Symbols

ρ_i	The risk premium of policy i.
c_i	The number of claims of policy i.
e_i	Duration of policy i.
l_{ij}	Loss paid out for a given claim j and policy i.
$n.sims$	The number of kept iterations of the Gibbs sampler.
dev	Short for deviance residuals.

Terminology

Burn-in iterations	The number of iterations that are discarded from the Gibbs sampling procedure.
Claim	A reported event where the insured asks for compensation.
Claim arrival	A claim j for a policy i is notified to the insurer at the time or claim arrival t_{ij}.
Claim severity	The average loss for a policy i, given a number of claims c_i.

Cook's distance	Measures the influence of an outcome by combining residuals and leverage. There is no clear rule when to delete an outlier.
Crunched residuals	Deviance residuals that are summed together within equally sized bins of fitted values.
Deductible	The amount of money an insured has to pay himself for a certain claim.
Deviance residuals	the signed square roots of the ith observation to the overall deviance. This are the residuals that are taken as a default in GLM analyses.
Duration	The lifetime of a policy. The duration is expressed in years.
Exposure	It is the state of being subject to loss because of some hazard or contingency. This can be the duration in policy years, the number of claims, the number of hours a policyholder has driven with an insured car, etc. The duration is the exposure of choice in this research.
Leverage	An observation that has unusual predictor (X) values, so that it is far from the mean of predictors, has leverage on (i.e., the potential to influence) the regression line.
Number of claims	The number of times a policy i resulted in a claim during the duration e_i of this policy.
Policy	It is a contract between the insurer and the insured, known as the policyholder, which determines the claims which the insurer is legally required to pay. This policy starts at a point in time and has a duration.
Risk premium	This is the expected total claim amount of a policy i that has duration e_i.
Thinning	The number of iterations that are discarded between kept iterations.

Chapter 1

Introduction

In a lot of countries the insurance market is deregulated. The pricing is not uniform across the different insurance entities, but free competition is allowed to ensure that fair premiums are attained. The fair or risk premium allows an insurer to cover the claims of a certain risk category (such as young drivers) during a period of coverage. In this manner the insurer will not lose money by setting this premium. We do not consider extra costs an insurer faces here. The strategy to pricing a premium can best be illustrated by an example:

> Young drivers are known to be more reckless and result in higher claim amounts than middle aged drivers. This results in a higher risk premium for the young drivers than the middle aged drivers. Suppose that the premiums for young drivers are set at an amount higher than the risk premium by other insurers in the market, this means it is possible to cut the price of the premium and still remain profitable for the young drivers. If an insurance company cuts on the premium while keeping the premium at a higher price than the risk premium, this will result in attracting more young drivers away from competitors. This increases the volume of profitable accounts and the insurer in question thus gains an edge.

Suppose that in the previous example the risk premium has been overestimated by the insurer in question, then the insurer still acquires a larger volume of young drivers but now the accounts might not be profitable anymore. This illustrates why it is of utmost importance to an insurer to make an accurate estimation of the risk premium, for a category such as young drivers, before making a strategic decision. Another important aspect here is the uncertainty of the estimation. Suppose we have two policies which are estimated at exactly the same amount, but the uncertainty of one estimation is higher than the other. It will therefore be riskier to offer a premium to the more uncertain one and therefore requires a different treatment.

1.1 Problem statement

In the competitive environment insurers face today, where prices are pressured to a lower rate to keep up with the competition, the need for prudent and accurate ratemaking is high. The way the premiums are priced have not changed a lot, if at all, in the past thirty years. The recent developments in machine learning has inspired some actuaries to seek alternative methods to price their premiums. But the knowledge or time is not always present to conduct such a research in-house, this is where this research steps in. This research seeks to answer one main question by answering its mutually exclusive and collectively exhaustive subquestions.

- How do the present techniques applied in premium pricing of non-life insurance perform compare to suitable alternative techniques in accuracy and in uncertainty quantification?
    i. What are the current methods utilized by actuaries?
    ii. What are the advantages and disadvantages of the current methods?
    iii. What are alternative or more advanced approaches that are suitable in an insurance application?

iv. How do we measure what method performs the best?

v. Is it possible to quantify uncertainty surrounding an estimate?

vi. How can we implement the uncertainty quantification in the premium calculation?

vii. Which method performs best and why?

Each of these subquestions will be answered in this research. Subquestions i,ii and iii will be answered in section 3, the theory section. Subquestions iv and vii are answered in sections 5 and 7. The remaining subquestions v and vi are addressed in section 6.

1.2 Significance of research

The actuarial environment is fairly traditional, it is not focused on innovating. This research provides actuaries alternative and more advanced (recent) methods which could more accurately estimate the total claim amounts for the different risk categories while also accounting the uncertainty of this estimate.

1.3 Objective of research

The focus of this research is to create a model that accurately estimates the total claim amounts an insured will accumulate based on the characteristics of the insured item and its owner. This model allows the insurer to price a risk premium for new customers based on their characteristics. Besides providing an accurate estimate of the total claim amount, the model should provide meaningful insights into the number of claims and claim severity responses.

1.4 Scope and limitation

The scope of this research is limited by risk premium pricing only. Competitive pricing, additional costs a company faces and price elasticity are left out of the equation. We only take into account the ultimate claim amounts, so all considered claims have been settled and the policies have expired.

Chapter 2

Literature Review

In the early 1900's there was no widespread used pricing method in non-life insurance. The easiest method was to not differentiate between its policyholders and simply use the average claim size over the insurer's whole portfolio to assign a premium per policyholder.
The work done by Lundberg [25] and Cramer [13] in 1903 and 1930, by many considered the founders of mathematical theory of risks, made actuaries think about approaching risks from the insurance companies' perspective. This was when one- and two-way analyses were born. In these analyses, the effect of a single or two covariates are investigated on either the number of claims or average claim size. The problem here is that correlation between covariates are not accounted for. Suppose you have a one-way analysis and we have two covariates: policyholder age and splitting of premium (yes/no). Now let's assume that young drivers split their premium payments more often than older drivers because of budgetary reasons. A result from the one-way analysis could be that being a younger driver increases the number of claims and splitting the premium does this as well. That means we will double-count the effect of age, because of the aforementioned unnoticed strong correlation between the covariates. There was thus a need for a multivariate approach.
In 1959 Wittick wrote a paper [34] on a merit system applied in Canada to differentiate between the group of insured and the individually insured by being in one of four risk categories: from 3 to no years without a claim. Bailey and Simon [3] followed up on Wittick's work and coined the term *credibility*, while creating an applicable formula for credibility instead of a constant discount for the four different risk categories. Credibility is thus the accumulated trust of the insurance company in an individual insurer. A first breakthrough of a multivariate approach was provided by Bailey and Simon in 1960 [4] in the form of the minimum bias procedure. This procedure sets the sum of all individual differences of estimated costs and observed costs to zero. This imposes a set of equations, which are solved after an iterative process. It is essentially a fixed point iteration technique. The issue with this method is that it does not provide a statistical framework, so the significance of a covariate can not be quantified. This was however one of the first widely used methods by actuaries. There were some follow-up papers that applied the minimum bias procedure on auto insurance data, such as the influential one by Jung in 1968 [23].
The next very influential paper for risk premium pricing methods in non-life insurance was provided by Nelder and Wedderburn in 1972 [28] on generalized linear models (GLM). It was a mathematical paper that was not focused on non-life insurance, but it introduced a very interesting new method. It was an improvement on the previously used minimum bias procedure, since it operates within a statistical framework while achieving very similar results. GLM is a generalization of linear regression, where the probability distribution of a response variable is extended from the Normal distribution to any distribution from the exponential family of distributions. GLMs also generalize the relation between the expected value of the response and the covariates by the link function. Nelder and Mcculagh later wrote a book on how to deal with many different aspects of GLMs in 1983 [27], including how to quantify the goodness of fit and how to validate GLMs. This is still one of the widely used methods to date to determine the price of a non-life insurance product.
One of the earliest papers where they implemented the GLIM computer package from the Royal Statistical Society for modelling the number of claims and average claim size was provided by Baxter et al. in 1980 [5]. There appears to be quite the time gap before the next few influential papers on the application of

GLMs in a non-life environment arrived. But this did not happen immediately after the paper by Nelder and Wedderburn, since GLMs were not easy to fit without proper software and computers. It therefore took until the 1990's, when computing power rapidly increased and it therefore became a lot easier to fit models such as the GLMs, that it gained more traction in the insurance community.
Papers encouraging the use of GLMs for insurance premium pricing followed in quick succession. A few examples are papers by Brockman and Wright in 1992 [9], Renshaw in 1994 [30] and Haberman and Renshaw in 1996 [20]. In all of these last three papers, they use a multiplicative GLM to model the number of claims and severity. The distributions used for the response is Poisson in case of the number of claims and Gamma for claim severity.

Chapter 3

Theory

Contents

Before we model the data, it is necessary to understand both the process we wish to model and the models we will use. We therefore start this chapter with some basic non-life insurance terminology (section 3.1). This is followed by the objective of pricing and an explainer of the two main approaches to premium pricing (section 3.2). We continue with the theory of the premium and its basic components: the number of claims and the claim severity (section 3.3). After discussing the basic response components, we proceed to the current approach (GLM) applied by actuaries in section 3.4. The chapter is concluded with the theory of some Machine Learning techniques (section 3.5) that are applied in this thesis.

3.1 Properties of interest and key responses

The level of premiums are determined by properties which influence the size of the claim amount. These properties can be put in the following overarching categories, the *rating factors*:

- **Policyholders**: age, gender, line of business, etcetera.
- **Insured item**: age, model type, size, etcetera.
- **Geographic region**: per capita income, population density, etcetera.

These overarching categories are not fixed, but are determined by the available data. Each dataset may lead to a very different kind of categorisation.
Furthermore, a claim contains certain information as well, such as the date of the incident and notification, but also what type of incident it was. You can imagine that the size and frequency of a claim where a car was scratched is quite different to a claim where a serious collision took place, it could be that a certain type of incident is affected by other rating factors.

This brings us to the basic concepts and key responses of interest in determining a fair or risk premium.

- Basic concepts:
  - A **policy** is a contract between the insurer and the insured, known as the policyholder, which determines the claims which the insurer is legally required to pay. This policy starts at a point in time and has a duration.
  - **Exposure** is the state of being subject to loss because of some hazard or contingency. This can be the duration in policy years, the number of claims, the number of hours a policyholder has driven with an insured car, etc. The duration is the exposure of choice in this research.
  - The **duration** e of a policy is the lifetime of a policy. The duration is expressed in years.
  - A **claim** is a reported event where the insured asks for compensation. This claim is notified to the insurer at time t, called the **claim arrival**.
  - **Loss** l is the paid out amount for a given claim.
- Key responses:
  - **Claim severity** s is the average loss for a number of claims.
  - the **number of claims** c is the number of times a policyholder reports a claim at the insurer during a period of exposure e.
  - **Risk premium** is the expected total claim amount during a period of exposure. This is the product of the expected number of claims and the expected claim severity conditioned on the number of claims: $\rho = E[c] \cdot E[s|c]$.

The key responses are the main items of interest in the pricing analysis.

3.2 Objective of premium pricing

The main objective is to accurately determine the fair or risk premium. We do this by pricing the premium at the expected loss or claim amount of a policy during its lifetime. This is a fair or risk premium, such that the expected loss of an insurer is zero. We are interested in discovering the relation between rating factors and the total claim amount. There are two ways to do this: directly model the risk premium or model the number of claims and claim severity separately. The former requires less modelling and is thus faster, while the latter could be more accurate if the relation between rating factors for the number of claims and claim severity differs a lot. That is the reason for modelling the separate building blocks of the pure premium. It allows a more layered explanation of the determination of a risk premium and can increase the accuracy of the estimation, which is the main objective. It is also very valuable to determine a fair premium within a

statistical framework such that we can express how sure we are of a certain result and its significance.
In general, the pricing of premiums are categorized in two classes: a priori and a posteriori pricing [1, 14]. We now explain these into more detail.

3.2.1 A priori pricing

In the a priori analysis the policies are priced based on information that is always known at the start of a policy. This is information that is not policyholder specific. So here the characteristics of a policy solely determine the eventual pricing and we do not further distinguish per policyholder based on a policyholder's specific claim history. This results in the iteration of policies i. This is different from the a posteriori analysis.
In a priori pricing, we tackle the insurance data in a cross-sectional manner. This approach is done because we do not differentiate between policyholders, so the history of a policyholder does not matter. The policies are then analyzed from either the same point in time or from multiple time periods. The latter is mostly used in practice if the differences per policy year are not very large, which could be the case if a catastrophe occurred.

3.2.2 A posteriori pricing

The a posteriori pricing methods are based on credibility theory [10]. Here we differentiate on the policyholders as well. So the premium of a specific policyholder changes when more information about the individual's claim history is known. Thus making the estimation more credible, hence its name. Most insurers simply use the a priori approach and add a covariate that describes the policyholder's specific claim history, such as the number of claim-free years.
A different approach is to impose a hierarchy and differentiate between characteristics of a policyholder and a policy (see figure 3.1). Suppose we have a dataset consisting of N policyholders. For each policyholder i ($1 \leq i \leq N$) we have T_i observations available. Suppose that for instance policyholder i has five policies, then we would have $T_i = 5$, so we have repeated measurements $Y_{11}, ..., Y_{15}$.

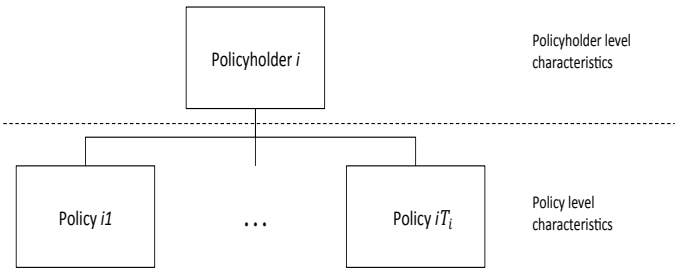


Figure 3.1: Policyholder to policy hierarchical structure.

This approach is attractive since observations on the same subject over time often are substantively correlated. But since policyholder characteristics are unmeasurable (without Telematics[1]), random effects are introduced per policyholder that determine the correlation structure between observations on the same subject, but also take heterogeneity among subjects into account. The value of this approach is in that it further distinguishes between persons and their unobserved characteristics at play such as reckless driving. The implementation of credibility theory is in practice quite limited due to its mathematical complexity. Bayesian statistics can play a big role here.

[1]Telematics allows the insurer to put a chip in the policyholder's car. This chip measures the driving properties of the policyholders. Examples of properties are: brake behaviour, speeding, miles driven, etcetera.

3.3 The premium and its basic components

The insurance ratemaking consists of pricing the premium by modelling the number of claims and claim severity separately as stated in (3.2). The assumptions that will follow are put in place to allow the construction of a basic model for the number of claims and claim severity.
For each policy i we have

- c_i, the number of claims during the lifetime e_i (duration) of a contract.
- l_{ij}, the loss corresponding to each claim made at claim arrival t_{ij}, with $j = 1, 2, ..., c_i$.

3.3.1 Main assumptions

There are three main assumptions in non-life insurance pricing [26], namely

1. *Partial ordering of Claim arrivals*:
   claim arrivals happen at times t_{ij}, satisfying $0 \leq t_{i1} \leq t_{i2} \leq ... \leq t_{ic_i}$. Thus enforcing a partial ordering on the set of claim arrivals per policy.
2. *Homogeneity of claim sizes and number of claims*:
   the j^{th} claim of policy i at time t_{ij} causes claim size l_{ij}. (l_{ij}), $j \in \{1, 2, ..., c_i\}$ is an identically and independently distributed (i.i.d.) sequence of non-negative random variables. Analogous for the number of claims.
3. *Mutual independence*:
   The claim size process (l_{ij}) and the claim arrival process (t_{ij}) are mutually independent.

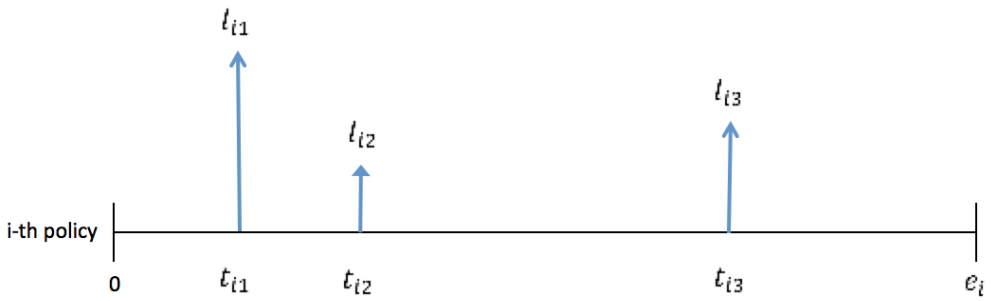


Figure 3.2: Visualisation of the claims arrival and size process for policy i. Each claim j has a specific claim arrival time t_{ij} and size l_{ij}. These can occur within the duration, e_i, of the policy.

3.3.2 The number of claims

Let $c_t, t \geq 0$ denote the number of claims during a period of exposure e_i. Also $c_0 = 0$ and $c_t \in \{0, 1, 2, ...\}$, so $c_t = \sum_{i=1}^{\infty} I_{[0,t]}(t_i), \quad t \geq 0$. I is the indicator function. This counting or claims process is said to be a *Homogeneous Poisson Process* (HPP) with rate $\lambda > 0$ if

i c_t has independent increments. So for $t_1, t_2, ..., t_n \in t$ with $t_1 < t_2 < .... < t_n$. The increments $c_{t_1}, c_{t_2} - c_{t_1}, ..., c_{t_n} - c_{t_{n-1}}$ are independent.

ii $c_t - c_s \sim$ Poisson$(\lambda(t - s))$ for $s < t$, stationary increments.

The claims process is now defined as a Homogeneous Poisson Process [6]. The total number of claims during an exposure period e_i for a policy i is thus

$$c_i|\lambda_i, e_i \sim \text{Poisson}(e_i\lambda_i),$$

where λ_i is the claim rate of a policy. The number of claims c_i with mean $E[c_i] = \mu_i = e_i\lambda_i$ are distributed according to the Poisson density function, see appendix (A.1.1). This distribution belongs to the exponential family.

3.3.3 Claim severity

The claim severity of a policy i is

$$s_i = L_i/c_i = \sum_{j=1}^{c_i} l_{ij}/c_i \mid c_i > 0.$$

The claim severity is often skewed to the right (many low values, few very large values) and non-negative on the real line. To model the claim severity different probability distributions can be used, although the Gamma distribution is most commonly used [9]. The gamma distributed s_i is parametrised as

$$s_i|c_i, \zeta_i, \nu \sim \text{Gamma}(\nu, \nu/\zeta_i),$$

such that $E[s_i|c_i] = \zeta_i$. But the Pareto, Log-Normal, Log-Gamma and Weibull distributions can be used as well. The Gamma distribution is preferred because it is a Tweedie distribution, this is a subfamily of the exponential family of distributions which is scale invariant. First we explain scale invariance, followed by its importance to severity modelling in an actuarial application. Suppose we have a positive constant $k > 0$ and a random variable Z from a certain family of distributions. This family is scale invariant if kZ follows a distribution in the same family. This property is attractive for insurance purposes since Z is measured in monetary units. Converting from one currency to another should not affect the analysis.

3.3.4 Premium

We want to price a premium for a policy that has a duration e_i. In actuarial practice the premium is set by using a combination of the number of claims and the claim severity by making use of the following relation

$$\rho_i = E\left[\sum_{j=1}^{c_i} l_{ij}\right] = E[c_i] \cdot E[s_i|c_i] = e_i \lambda_i \zeta_i. \tag{3.1}$$

The implementation of this is done with GLMs.

3.4 Generalized Linear Models

Definition 3.4.1. Linear regression with r predictors or covariates is defined as:

$$Y_i|\{X_{ij} = x_{ij}\}_j \sim \mathcal{N}(\beta_0 + \sum_{j=1}^{r} \beta_j x_{ij}, \sigma^2), \tag{3.2}$$

Linear regression is not suitable for modelling in non-life insurance because of the following two properties. Standard linear regression assumes (i) the response Y_i to be Normal distributed and (ii) its expected value to be a linear function of the covariates (see equation (3.2)). Property (i) is not very flexible. A typical model of the number of claims has a discrete probability distribution on non-negative integers, while the claim costs are often skewed to the right and non-negative. Besides that, suppose that we want to model the covariates in a different way than the usual linear function, for example a multiplicative model, then this will not be possible. Multiplicative models and others cannot be used in the standard linear regression setting because of property (ii). This calls for a more general form of linear regression, namely Generalized Linear Models (GLMs).

Definition 3.4.2. Generalized Linear Models assume

(i) the response Y_i to be distributed according to a probability distribution of the exponential family of distributions.

(ii) the expected value to be related to the linear predictor with a (link) function (equation 3.3).

In the next subsections we will discuss the key ingredients (i) and (ii) of a GLM in more detail.

3.4.1 Link function

Definition 3.4.3. A Link Function $g : \mathbb{R} \to \mathbb{R}$, which is monotone and differentiable, specifies the relation between the expected value of the response variable and the linear predictor of explanatory variables, such that

$$g(E[Y_i]) = \sum_{j=1}^{r} \beta_j x_{ij}. \tag{3.3}$$

The link function can take on a diverse number of forms. The different types of link functions for corresponding identities are listed in table (3.1). The identity link results in linear regression, while the logit regression is for example interesting for key responses that are proportions, such as for example the proportion of large claims, since the values are restricted to the closed set $[0, 1]$. In practice, the log link is used for both the number of claims and claim severity. Although the reciprocal link is the canonical link of the Gamma distribution. The canonical link is the link function that reduces the distribution of the exponential family to its easiest form. The reason why the log link is used, is because of the ease of use of the multiplicative model.

	Identity	Log	Logit	Reciprocal	Inverse squared
$g(\mu)$	μ	$\log(\mu)$	$\log(\frac{\mu}{1-\mu})$	$1/\mu$	$1/\mu^2$
$g^{-1}(\mu)$	μ	e^{μ}	$\frac{e^{\mu}}{1+e^{\mu}}$	$1/\mu$	$1/\mu^2$

Table 3.1: Different types of link functions

3.4.2 Exponential family of distributions

Definition 3.4.4. A probability distribution f_{Y_i} is from the exponential family [22],[30] if

$$f_{Y_i}(y_i|\theta_i, \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, \omega_i)\right]. \tag{3.4}$$

The scale function $a(\phi)$, where dispersion parameter $\phi > 0$ is fixed for all i, is commonly of the type $a(\phi_i) = \phi/\omega_i$. The cumulant function $b(\theta_i)$ is assumed to be twice differentiable and has an invertible first derivative.

The choice of the name for $b(\cdot)$ will become clear in section (3.4.3). It can be shown that the Poisson, Normal, Gamma distributions and more belong to this family of distributions.

3.4.3 Variance functions

Definition 3.4.5. A variance function is a smooth function $v : \mathbb{R} \to \mathbb{R}$ which depicts the variance of a random quantity as a function of its mean $v(\mu)$.

In this section we want to acquire a general expression for the variance and expectation of a distribution from the exponential family. To achieve this, we use the cumulant-generating function. This function is the logarithm of the moment-generating function, which is defined as

$$M(t) = E\left[e^{tY}\right].$$

If you work through some algebraic manipulations and use the assumptions of the exponential family, the following cumulant-generating function is achieved

$$\Psi(t) = \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}.$$

Now it should be clear why $b(\cdot)$ was named the cumulant function previously. From properties learned during an intermediate course in probability [19], it follows that

$$E[Y] = \Psi'(0) = b'(\theta) = \mu; \qquad Var[Y] = \Psi''(0) = b''(\theta)\phi/\omega = v(\mu)\phi/\omega.$$

Where the last step was achieved by using the relationship $\theta = b'^{-1}(\mu)$ and substitute it into the variance to achieve an expression for the variance function of a distribution from the exponential family

$$v(\mu) = b''(b'^{-1}(\mu)).$$

For clarity and simple use later on, we sum up some variance functions in table (3.2) for probability distributions which are of interest.

Distribution	Normal	Poisson	Gamma	Binomial	Inverse Gaussian
$v(\mu)$	1	μ	μ^2	$\mu(1-\mu)$	μ^3

Table 3.2: Variance functions for several distributions of the exponential family

3.4.4 Multiplicative models and Premium pricing

Definition 3.4.6. A multiplicative model is a GLM which uses the log as the link function.

The multiplicative models in general will be discussed and applied for the number of claims and claim severity in the following sections. The multiplicative model is used a lot in practice because of its ease of use. We motivate this further in the coming subsections.

3.4.4.1 Motivation and method

In the analysis we want to relate the key responses, the number of claims and claim severity, to the rating factors (see section 3.1). And we are especially interested in how we should alter the basic premium because of a certain combination of covariates. The choice of multiplicative models is motivated by its ease of use in altering an expected response due to a change in the set of covariates. We now show the ease of altering an expected outcome by changing a level of a covariate. Suppose we have a basic set of covariates, or the intercept, such that

$$E[Y|X = \boldsymbol{x}] = e^{\boldsymbol{x}\boldsymbol{\beta}}.$$

Suppose now we alter a level a level of a covariate x_j, such that

$$\boldsymbol{x}^{(j)} = (x_1, ..., x_j + \alpha_j, ..., x_r).$$

We can then calculate the altered expected value as follows

$$\frac{E[Y|X = \boldsymbol{x}^{(j)}]}{E[Y|X = \boldsymbol{x}]} = e^{(x^{(j)} - x)\boldsymbol{\beta}} \Rightarrow E[Y|X = \boldsymbol{x}^{(j)}] = e^{\beta_j \alpha_j} E[Y|X = \boldsymbol{x}].$$

The utilized approach in setting the risk premium for a policy i (see equation 3.1), is to first compute the $\eta_j = e^{\beta_j \alpha_j}$ for all different covariate levels for both the number of claims and claim severity models. The basic premium is determined by the intercept, then the basic premium level is adjusted by multiplying it with $\eta_j{}'s$ of both the number of claims and claim severity models. This illustrates the ease of use of a multiplicative model and it is the reason why it is so widely used by actuaries.

3.4.4.2 Fitting GLMs with logarithmic link (multiplicative model)

There are multiple methods to estimate the model parameters (model fit), for example least squares fit, variance stabilized responses, Bayesian approaches and Maximum Likelihood Estimation (MLE). Among these methods, MLE is the most widely used in non-life insurance. If the ML equations do not have an explicit solution, we resort to Newton-Raphson's method and Fisher's scoring method.
So let's have a look at MLE in this environment. Take the log-likelihood of the exponential family equation (3.4), such that

$$l(\boldsymbol{\theta}|\phi,\mathbf{y}) = \log\left[\prod_i f_{Y_i}(y_i|\theta_i,\phi)\right] = \frac{1}{\phi}\sum_i \omega_i(y_i\theta_i - b(\theta_i)) + \sum_i c(y_i,\phi,\omega_i)$$

Now the score equations can be derived. This is done by combining the already determined relations(see sections 3.4.3, 3.4.1) : $\mu_i = b'(\theta_i)$ and link function $g(\mu_i) = \sum_j x_{ij}\beta_j$. We take the derivative to β_j by applying the chain rule $\partial l/\partial\beta_j = \sum_i \partial l/\partial\theta_i \cdot \partial\theta_i/\partial\beta_j$ $(*)$ such that

$$(*) = \sum_i(\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial\theta_i}{\partial\beta_j} = \frac{1}{\phi}\sum_i(\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}.$$

Now we use that $\partial\mu_i/\partial\theta_i = b''(\theta_i) = v(\mu_i)$, so $\partial\theta_i/\partial\mu_i = 1/v(\mu_i)$. Furthermore $\partial\mu_i/\partial\eta_i = (\partial\eta_i/\partial\mu_i)^{-1} = 1/g'(\mu_i)$ and lastly that $\partial\eta_i/\partial\beta_j = x_{ij}$ because $\eta_i = \sum_j x_{ij}\beta_j$. Combine these expressions to acquire

$$\frac{\partial l}{\partial\beta_j} = \frac{1}{\phi}\sum_i \omega_i\frac{y_i - \mu_i}{v(\mu_i g'(\mu_i))}x_{ij}.$$

The dispersion parameter is ignored in this analysis. It will be estimated later on. So we set the r partial derivatives to zero and multiply by ϕ, so we acquire the general ML equations:

$$\sum_i \omega_i\frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)}x_{ij} = 0, \quad j = 1,2,...,r \tag{3.5}$$

As you may notice, the nominator lends itself quite well for Tweedie distributions. Because Tweedie distributions have expected value μ_i and variance $\phi\mu_i^p$, where p is an integer ≥ 0. So equation 3.5 becomes analytically solvable. This is the case if we model the risk premium, the number of claims and claim severity, since these use the Compound Poisson Process, the Poisson and gamma distributions respectively.

3.4.5 Model building

A widely used method to measure the goodness of fit is to use the Deviance Information Criterion (DIC)[33], where

$$D^*(\mathbf{y},\hat{\boldsymbol{\mu}}) = 2[l(\mathbf{y}) - l(\hat{\boldsymbol{\mu}})] \tag{3.6}$$

The Akaike Information Criterion (AIC), see equation 3.7, is also used in the actuarial environment to evaluate the goodness of the model. The AIC is a means to help in the choice between model complexity and goodness-of-fit. In equation 3.7, k is the number of fitted parameters.

$$AIC = 2k - 2ln(l(\hat{\boldsymbol{\mu}})) \tag{3.7}$$

3.4.5.1 Covariate selection

In common practice, a backward selection method is used to determine an optimized GLM model. This is an iterative process, where first all covariates are included to fit a GLM. Then either a p-test (Wald) or the AIC is used to quantify the contribution of a covariate. If the contribution is less than a specified level, the covariate is dropped. We use the AIC for covariate selection of a basic GLM model.

3.4.5.2 Residual Analyses

For residual analyses, some important measures are

i. Raw residuals: difference between observation and estimated value. Variations on residuals are standardized residuals, studentised residuals and Pearson residuals. These all have a different way of combining raw residuals with their standard deviations.

ii. Deviance residuals: are the signed square roots of the ith observation to the overall deviance. This are the residuals that are taken as a default in GLM analyses.

iii. Leverage: An observation that has unusual predictor (X) values, so that it is far from the mean of predictors, has leverage on (i.e., the potential to influence) the regression line.

iv. Cook's distance: measures the influence of an outcome by combining residuals and leverage. There is no clear rule when to delete an outlier. Some suggest deleting an outlier when the Cook's D > 1 [12], others say when Cook's D $> 4/n$ [7] and even $> 4/(n-k-1)$ where k is the number of explanatory variables fitted. The differences between the first and the next two rules are enormous. The last two are approximately the same when we handle very large datasets.

For a more complete description on residual analyses, we refer to [27].

3.5 Machine Learning techniques

In the methodology section we will apply some machine learning techniques of interest: Gradient Boosting Machines (GBM), Random Forests (RF), Multivariate Adaptive Regression Splines (MARS) and Bayesian Additive Regression Trees (BART). GBM, RF and BART are all tree-based ensemble methods. The main difference between these methods is the underlying ensemble method, which is boosting for GBM and BART and bootstrap aggregation or bagging for RF. The motivation for using these methods in the insurance environment is two-fold. These methods are generally known to provide good results in predictive performance, while allowing insight into variable importance and interaction effects. This makes it possible to enhance GLMs, which are more transparent and work within a statistical framework.
This section introduces how to build a regression tree first (section 3.5.1), since both RF, GBM and BART utilize this. Afterwards the further workings of RF (section 3.5.2), GBM (section 3.5.3), MARS (section 3.5.5) and BART (section 3.5.6) are discussed.

3.5.1 Regression Trees

Suppose we have our data $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ are the explanatory variables and y_i the response variable for each observation $i \in \{1, ..., N\}$. Let us say we have p explanatory variables such that $\boldsymbol{x}_i = (x_i^1, ..., x_i^p)$. Regression trees partition the space of these explanatory variable into M regions $R_1, ..., R_M$ and assign a constant c_m to each of these regions. This constant estimates the response in each region, such that

$$\hat{y}_i = T(\boldsymbol{x}_i; \Theta) = f(\boldsymbol{x}_i) = \sum_{m=1}^{M} c_m I(\boldsymbol{x}_i \in R_m), \qquad i \in \{1, ..., N\},\ \Theta = \{R_m, c_m\}_1^M. \tag{3.8}$$

In figure 3.3 we show how a regression tree splits the explanatory variable space into different regions. We shall now discuss how to determine these splits.
The splitting variable j and split point s divide an explanatory variable in a pair of half-planes R_m and R_{m+1}, as follows

$$R_m(j, s) = \{x | x^j \leq s\} \text{ and } R_{m+1}(j, s) = \{x | x^j > s\}.$$

At each node the splitting variable j and point s are chosen in a manner that minimizes the Euclidean distance of the constant c_m to all y_i given $\boldsymbol{x}_i$ is in region R_m and analogously for $m + 1$. This minimization

is done in the following manner,

$$\min_{j,t}\left[\min_{c_m}\sum_{\boldsymbol{x}_i\in R_m(j,t)}(y_i-c_m)^2+\min_{c_{m+1}}\sum_{\boldsymbol{x}_i\in R_{m+1}(j,t)}(y_i-c_{m+1})^2\right].$$

This minimization is solved by assigning the constants c_m , c_{m+1} to the average of all y_i given $\boldsymbol{x}_i$ is in each respective region R_m and R_{m+1}, no matter the choice of j and t. This process is repeated on all of the resulting regions. The more this process is repeated, the more likely the algorithm will overfit the available data. So the algorithm should stop at a certain tree size.
The strategy is to grow a full tree T_0, such that we have N terminal nodes and perfectly fit the data. Then a weakest link pruning strategy is applied using the cost-complexity criterion (equation 3.9) to find a tree $T_\alpha \subset T_0$. Where to prune means to collapse a non-terminal node.

$$\begin{aligned} C_\alpha(T) &= \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|, \\ Q_m(T) &= \frac{1}{N_m}\sum_{\boldsymbol{x}_i\in R_m}(y_i-\hat{c}_m)^2, \\ \hat{c}_m &= \frac{1}{N_m}\sum_{\boldsymbol{x}_i\in R_m} y_i. \end{aligned} \tag{3.9}$$

In equation 3.9, the indices m are the terminal nodes and $|T|$ is the number of terminal nodes or the tree size. The *tuning parameter* $\alpha \geq 0$ governs the trade-off between tree size and goodness of fit. It can be shown that for each α there is a unique smallest subtree T_α that minimizes $C_\alpha(T)$. As stated earlier, we get to this unique smallest subtree by a weakest link pruning strategy by successively collapsing the internal node with the smallest per-node increase of $\sum_m N_m Q_m(T)$. This process is repeated until we obtain a single node tree. It can be shown that this sequence of trees must contain the optimal tree T_α.

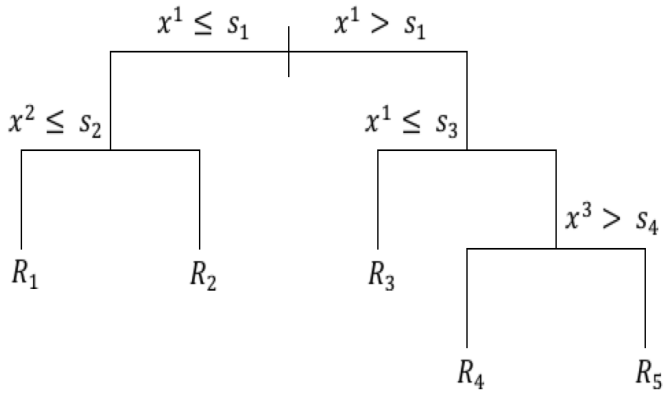


Figure 3.3: Building a regression tree by splitting a explanatory variable x^j at a splitting point s_m into a pair of half-planes.

3.5.2 Random Forests

Bootstrap aggregation plays a large role in Random Forests [8]. The key functionality of the bootstrap method that is used by Random Forests, works as follows. Suppose we have a similarly constructed dataset as in section 3.5.1. Bootstrap draws $b = 1$ to B samples with replacement from the full dataset, such that we have datasets $X^{*1}, ..., X^{*B}$ that have the same size as the original. Each regression tree generated in bagging is identically distributed (i.d.), the expectation of an average of B trees is the same as the expectation of any

of them [21]. In other words, we have

$$E[T_b(x)] = \mu, \quad Var[T_b(x)] = \sigma^2 \quad \text{and} \quad Cov[T_b(x), T_{b*}(x)] = \rho\sigma^2.$$

So the improvement sought after through use of bagging is variance reduction. An average of B i.d. with positive pairwise correlation ρ, the variance of the average is

$$Var\left[\frac{1}{B}\sum_{b=1}^{B} T_b(x)\right] = \left(\frac{1}{B}\right)^2 \left(B\sigma^2 + n(n-1)\rho\sigma^2\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

As B increases, the second term disappears but the first remains. Random forests attempt to reduce the pairwise correlation ρ without increasing the variance much by a random element in the tree-growing process. We now explain this tree-growing process. For each of these bootstrapped datasets $(X^{*1}, ..., X^{*B})$ a RF-tree T_b is constructed in the following manner. The tuning parameter m determines the number of explanatory variables that are randomly drawn from the possible p variables at each split. For the m drawn variables, the best splitting variable j and splitting point s are determined in a similar manner as described in section 3.5.1. The node is then split into two daughter nodes. This process is repeated until the minimum node size n_{min} is reached and we thus have a fully grown tree. The ensemble of trees $T_1, ..., T_B$ cast their vote on the prediction of new data, by

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Reducing m reduces the correlation ρ between any pair of trees, although it increases the variance as well. This is why m is in general quite low $(1-6)$. The algorithm can be found in the appendix, see A.2.1.

3.5.3 Gradient Boosting Machines

An important ingredient for GBMs [15] is the idea of boosting [32]. Boosting consists of making an initial guess of the target function $f : X \to Y$ where at first each observation $(\boldsymbol{x}_i, y_i)$ is given the same weight $\frac{1}{N}$ in a dataset of similar construction as in section 3.5.1. Each iteration the weights are reapplied in such a manner that more weight is given to an observation that was badly fitted. The target function is updated and this is repeated for some iterations until a final target function is achieved because of some stopping criterion. Boosted trees use a summation of regression trees as the target function. We will now go into more detail about the workings of GBMs.The algorithm can be found in appendix A.2.2.
A regression tree $T(x;\Theta)$, see equation 3.8, is described by its terminal nodes m. These nodes approximate response variables by a constant c_m for a partition R_m in the explanatory variable space. A Loss function $L(y_i, c_m)$ describes the goodness of this approximation for each response variable. A regression tree is optimal if the following parameter is found

$$\hat{\Theta} = arg\ \min_{\Theta} \sum_{m=1}^{M} \sum_{\boldsymbol{x}_i \in R_j} L(y_i, c_m). \tag{3.10}$$

This is quite the optimization problem, to reduce the complexity the following more convenient optimization criterion is used

$$\tilde{\Theta} = arg\ \min_{\Theta} \sum_{i=1}^{N} L(y_i, T(\boldsymbol{x}_i;\Theta)). \tag{3.11}$$

GBMs are a sum of sequential trees. So after each iteration a tree is added to the existing sum or ensemble of trees. We can thus express the target function up to iteration K of GBMs as

$$f_K(x) = \sum_{k=1}^{K} T(x;\Theta_k).$$

This leads to rewriting equation 3.11 to the following optimization problem at each iteration

$$\tilde{\Theta}_k = arg \min_{\Theta_k} \sum_{i=1}^{N} L(y_i, f_{k-1}(\boldsymbol{x}_i) + T(\boldsymbol{x}_i; \Theta_k)). \tag{3.12}$$

The GBM follows a very greedy procedure to reach an optimal target function. It follows the procedure of steepest descent, see equation 3.13. It chooses the path where the biggest reduction of the Loss function is achieved. You can imagine that this approach heavily relies on the chosen initial target function. Because of this initial target function it might lead to a local optimum and not the global optimum.

$$\begin{aligned} f_{ik} &= f_{ik-1} - g_{ik}, \\ g_{ik} &= \left[\frac{\partial L(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)} \right]_{f(\boldsymbol{x}_i)=f_{k-1}(\boldsymbol{x}_i)} \end{aligned} \tag{3.13}$$

The idea here is to fit a regression tree whose predictions t_k are as close as possible to the negative gradient. A regression tree is grown as explained in section 3.5.1. This reduces the possibility of overfit one would possibly achieve by just taking the negative gradient. Different versions of the GBM algorithm are obtained by applying a different Loss function. Then there remains the question of the tuning parameters. One of these parameters is the single *tree depth* γ; commonly a decision stump is used as the fitted regression tree to the negative gradient. A decision stump is a single node with two terminal nodes, so $\gamma = 1$. If we increase the individual tree depth, we allow more interaction between the explanatory variables; this can lead to interesting insights into variable interaction. In practice the tree depth tends to between depth 4 and 8.
The next tuning parameter is closely related to the tree depth, it is the *minimum number of observations in the terminal nodes* m_{min}. A fully grown tree uses all data and thus has one observation per terminal node. Suppose that the tree depth is chosen at a fairly low level, as suggested, then with a lot of data a low number of minimum observations will never be crossed. In that case this tuning parameter does not influence the process, but will only result in longer computing time.
Another tuning parameter for GBMs is the *number of trees K* used. Increasing the number of trees better approximates the target function f, but will likely lead to overfitting the data.
The last tuning parameter is the *shrinkage* ν. The shrinkage scales the contribution of each sequentially fitted tree in the following manner

$$f_k(x) = f_{k-1}(x) + \nu \cdot \sum_{m=1}^{M} c_{mk} I(x \in R_{mk}).$$

The shrinkage term is often called the learning rate of the boosting procedure. The shrinkage parameter interacts with the number of trees parameter. If the shrinkage is lower, the number of trees should be larger since the adjustments are done at a lower rate, such that

$$K \propto \frac{1}{\nu}.$$

3.5.4 Variable importance and partial dependence plots

Tree-based methods such as RFs and GBMs have some interesting features that allow us to interpret the results. As mentioned in the introduction of section 3.5, the features in question are variable importance and partial dependence plots. Variable importance is inherent to tree-based methods, since at each node of a tree we choose between covariates to make a good split. If a variable is more often used in splitting a node, its variable importance increases. Furthermore we can visualise the effects of one or two covariates on the response, in other words the marginal effects on the response. This allows us to discover non-linearities in a covariate and the interaction between two covariates.

3.5.5 Multivariate Adaptive Regression Splines (MARS)

A method that is especially good at extracting interaction terms is MARS [16]. MARS is a non-parametric method that employs a recursive partitioning regression procedure, which is of the form

$$\text{if } \boldsymbol{x} \in \mathcal{R}_m, \text{then } \hat{f}(\boldsymbol{x}) = g_m(\boldsymbol{x}|\{a_j\}_1^p) = a_0 + \sum_{m=1}^{M} a_m B_m(\boldsymbol{x}).$$

Here $\{\mathcal{R}_m\}_1^M$ are disjoint subregions of the domain D. Furthermore we have intercept a_0 and fitted coefficients $\{a_m\}_1^M$. The basis functions B_m take the form

$$B_m(\boldsymbol{x}) = \prod_{k=1}^{K_m} \left[s_{km} \cdot (x^v(k,m) - c_{km})\right]_+^q,$$

where s is the sign (positive or negative), x^v the specific covariate with $v \in \{1, ..., r\}$, c the knot location, q the order of the spline and the subscript $+$ indicates the positive part of the argument. Basis functions of this type are called hinge functions because of their shape (see figure 3.4).
The algorithm has two parts. It starts with forward - and is followed by backward recursive partitioning. In the forward phase, MARS guesses a suitable value for M_{max}, which is more than twice its optimal value. Then it starts with one constant basis function, the intercept. For each covariate and choice of knot positions, the space of basis functions is minimised with respect to the squared prediction error. This is recursively repeated until $M + 1$ terms are derived, where 1 is the intercept. Then the next phase of the algorithm commences: the backward phase. The basis functions are pruned that contribute least to the goodness of fit, which is in this case the generalized cross-validation (GCV) criterion,

$$GCV(M) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(\boldsymbol{x}_i)\right]^2 \Bigg/ \left[1 - \frac{C(M)}{N}\right]^2.$$

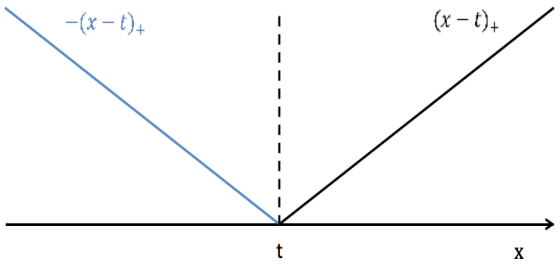


Figure 3.4: Visualisation of the Hinge functions used for MARS.

3.5.6 Bayesian Additive Regression Trees (BART)

An additional method of interest aims to combine the predictive power of tree-based methods, while attempting to incorporate the quantification of uncertainty. This method is called Bayesian Additive Regression Trees (BART) [11].
Suppose we have a response y_i and corresponding r covariates $\boldsymbol{x}_i = (x_i^1, ..., x_i^r)$ for policies $i = 1, ..., N$. The outcome is modeled as a sum-of-trees model

$$y_i = \sum_{k=1}^{K} g(\boldsymbol{x}_i, T_k, M_k) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{3.14}$$

where T_k is the k^{th} binary regression tree and $M_k = (\mu_{1k}, ..., \mu_{m_k k})$ is the set of m_k terminal node values of tree T_k. In BART priors are imposed over all parameters of the sum-of-trees model, namely

(T_1, M_1),...,(T_K, M_K) and σ. The joint prior distribution for equation 3.14 is $P[(T_1, M_1), ..., (T_K, M_K), \sigma]$. Because of independence of ϵ_i and (T_k, M_k) as well as independence between the K the structures and terminal node parameters, the joint prior distribution can be decomposed as

$$\begin{aligned} P[(T_1, M_1), ..., (T_K, M_K), \sigma] &= \left[\prod_{k=1}^{K} P[(T_k, M_k)]\right] P[\sigma] \\ &= \left[\prod_{k=1}^{K} P[M_k|T_k]P[T_k]\right] P[\sigma] \\ &= \left[\prod_{k=1}^{K} \left\{\prod_{j=1}^{m_k} P[\mu_{jk}|T_k]\right\} P[T_k]\right] P[\sigma]. \end{aligned}$$

So priors should be assigned to T_k, $\mu_{lk}|T_k$ and σ to obtain posterior distributions. For $P[T_k]$ there are three aspects to it: (i.) the probability a node at depth l is non-terminal, given by $\alpha(1+l)^{-\beta}$, where $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$. Here an increasing α increases the possibility a terminal node in the tree splits and an increasing β reduces the number of terminal nodes. The second aspect (ii.) is the distribution used to choose which covariate to select for the decision rule at an internal node, while the last aspect (iii.) is the distribution for the value of the selected covariate for the decision rule in an interior node. Chipman et al. [11] use a discrete uniform prior for both (ii.) and (iii.).
Chipman uses the following priors for $\mu_{jk}|T_k$ and σ:

$$\begin{aligned} \mu_{jk}|T_k &\sim \mathcal{N}(\mu_\mu, \sigma_\mu^2), \\ \sigma^2 &\sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}). \end{aligned}$$

IG is the inverse-gamma distribution with shape parameter $\nu/2$ and rate parameter $\nu\lambda/2$. Chipman makes an estimate of $\hat{\sigma}$ using the data [11], then picks a certain value for ν and a value of λ follows so that the q^{th} quantile of the prior on σ is located at $\hat{\sigma}$, $P[\sigma < \hat{\sigma}] = q$. Here ν and q are tunable parameters.
Furthermore Chipman et al. use $\alpha = 0.95$ and $\beta = 2$ and these are not recommended to be tuned. But the values for ν and λ can be tuned, although the recommended values are 3 and 0.9 respectively. Lastly they set $\mu_\mu = 0$ and $\sigma_\mu = 0.5/(\kappa\sqrt{K})$, where shrinkage factor $\kappa = 2$ is recommended although it is a tunable parameter. The rationale behind theses choices are mentioned in the paper [11].
Information is extracted from the posterior with a Bayesian backfitting MCMC algorithm. Given the observed data y, the Bayesian setup induces a posterior distribution $P[(T_1, M_1), ..., (T_K, M_K), \sigma|y_i]$. A Gibbs sampler entails K successive draws of $(T_k, M_k)|T_{-k}, M_{-k}, \sigma, y_i$, followed by a draw $\sigma|T_1, ..., T_K, M_1, ..., M_K, y_i$. The draw of σ conditioned on the rest is simply a draw from an inverse gamma distribution. The implementation of the K successive draws is a bit more tricky however. The key insight here is that the conditional $(T_k, M_k)|T_{-k}, M_{-k}, \sigma, y_i$ depends on (T_{-k}, M_{-k}, y_i) only through

$$R_{ik} = y_i - \sum_{w \neq k} g(X_i, T_w, M_w), \tag{3.15}$$

the N-vector of partial residuals that excludes the k^{th} tree. So the K draws of $(T_k, M_k)|T_{-k}, M_{-k}, \sigma, y_i$ are equivalent to draws from

$$(T_k, M_k)|R_{ik}, \sigma. \tag{3.16}$$

Because a conjugate prior is used for M_k we can obtain

$$P[T_k|Rik, \sigma] \propto P[T_k] \int P[R_{ik}|M_k, T_k, \sigma]P[M_k|T_k, \sigma]dM_k \tag{3.17}$$

in closed form up to a norming constant. Therefor we can draw from 3.16 by the following successive draws

$$T_k|R_{ik}, \sigma, \tag{3.18}$$

$$M_k|T_k, R_{ik}, \sigma. \tag{3.19}$$

Draw 3.18 is obtained using a Metropolis-Hastings (MH) algorithm. This MH algorithm proposes a new tree based on the current tee using one of four moves: (i.) grow a terminal node into two new child nodes ($P[i.] = 0.25$), (ii.) prune a pair of terminal nodes such that their parent node is the new terminal node ($P[ii.] = 0.25$), (iii.) change the splitting criterion of a single non-terminal node ($P[iii.] = 0.40$) and (iv.) swap the splitting criterion of a parent and child node ($P[iv.] = 0.10$). The grow and prune moves change the number of terminal nodes, but by integrating out M_k in 3.17, the complexities associated with reversible jumps between continuous spaces of varying dimensions are avoided.
And as last the draws 3.19 are a set of independent draws of terminal nodes $\mu_{jk}|...$ from a Normal distribution.
The backfitting algorithm we described generates a sequence of draws which is converging to the posterior $P[(T_1, M_1), ..., (T_K, M_K), \sigma|y_i]$. So the sequence of sum-of-trees functions $f^*(\cdot) = \sum_{k=1}^{K} g(\cdot, T_k^*, M_k^*)$, are converging to $P[f|y_i]$. Now we use this for inference by averaging the after burn-in sample $f_1^*, ..., f_B^*$,

$$\frac{1}{B}\sum_{b=1}^{B} f_b^*(\boldsymbol{x}_i), \tag{3.20}$$

so we predict with the posterior mean $E[f(\boldsymbol{x}_i)|y_i]$. Posterior uncertainty is quantified by the $(1-u)\%$ posterior intervals. Partial dependence functions can also be extracted by the marginal effects of one predictor on the response, see [11] for specifics.

Chapter 4

Exploratory Data Analysis

Contents

The data analysis is done in two parts. First we describe the available and relevant dataset that will be used for the risk premium pricing (section 4.1) and secondly we will present descriptive statistics about the dataset in question (section 4.2).

4.1 Available data

The available data is split into two sets of data that are linked by the policyholder ID. We have the portfolio dataset on the one hand and the claims dataset on the other. To discover the relations between the rating factors and the claim processes, we need a combination of these two. Each of these sets cover different insurance products. An insurance product might only cover theft for example. For analysis, we filtered a commonly used insurance product: Omnium insurance. Data preparation consisted of discovering inconsistencies in the claims and portfolio data and rectifying them. An example is renaming an identical but alternatively written brand name. The rectifications done are not interesting for this thesis. Furthermore we only regard claims that are finalized.

4.1.1 Portfolio

As mentioned in section (3.1), the rating factors are properties that describe the policyholders and the insured object. These are in other words the explanatory variables or covariates. Before we proceed with some histograms that describe our non-life insurance data, we sum up the available rating factors in table (4.1). The dataset of an insurance company is used. The type of this non-life insurance is *car insurance*. The dataset contains $443,287$ policies after data preparation. A policy starts in a year for a certain policyholder and has a duration. Two different policies can be for example of the same policyholder but in a different year. This makes sense because a policyholder's characteristics change over course of time. One year a policyholder may have a cheap car and does not drive a lot, while the next year he may have upgraded his car and drive more frequently. During the lifetime of a policy of a year this could happen as well. The insurance company tracks these changes by making different versions of this policy. Each time some covariate changes, a new version of the policy is made. The durations of these versions added up is the

lifetime of the whole policy, which is one year in this example. Each version is therefore regarded as a different policy that has a set of characteristics and a duration. This process is visualised in figure 4.1. Policies also have some additional information that is presented in table 4.1.

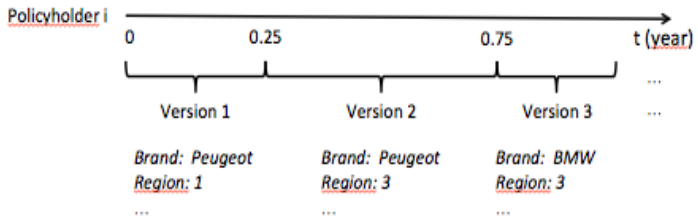


Figure 4.1: Process example of version changes for a policyholder.

Rating Factor	Type	Description
Car age	Integer	Lifetime of the car.
Credit Score	Integer	Credit score of the policyholder. The score ranges from 1, the best, to 4, the worst.
Fuel Types	Categorical	The car can use petrol, diesel, gas, electricity or hydrogen.
Experience years	Integer	The number of years a policyholder has owned a driver's license.
Yearly Mileage	Categorical	The range of miles the policyholder's car has driven.
Home Owner	Boolean	The policyholder owns a home or not.
Region type	Categorical	The policyholder lives in one of 19 anonymised regions.
Vehicle Power	Real	The policyholder owns a car with power expressed in cc.
Vehicle Weight	Real	The policyholder owns a car with weight expressed in kg.
Age	Integer	The policyholder's age.
Number of cars	Integer	The number of cars owned by the policyholder.
Brand of car	Categorical	The brand of the insured car.
Catalogue value	Real	The catalogue value of the insured car.
Accessories	Integer	The value of the add-ons of the insured car.
Number of licenses	Integer	The number of different types of licenses the policyholder owns.
Claim-free years	Integer	The number of years the policyholder did not have a claim.
Language	Categorical	The maternal language of the policyholder.

Table 4.1: Rating factors of each policy.

4.1.2 Claims

The claims dataset contains information about the claim reported to the insurer. The useful information is listed in table (4.3). The claim category might be especially interesting for modelling. The reason for this is that the rate at which a claim happens and the average claim amount might be quite different for an incident where a car was scratched or where two cars had a serious head-to-head collision. Therefore it could be interesting to model these categories separately to achieve better accuracy.
The claims dataset is filtered in such a way that only settled claims are considered.

4.1.3 Combination

The goal is to accurately model the number of claims and the claim severity of a policy by its characterizations. We are thus not interested in the number of payments, but the number of claims. So each incident contains at most one claim, no matter how many payments. To combine the portfolio and claims dataset, we

Information	Type	Description
Policy ID	Integer	The ID of a policy.
Policy Year	Integer	The year the policy was in effect.
Duration	Integer	The duration of a policy.
Deductible	Integer	The amount of money an insured has to pay himself.
Premium	Real	The actual premium paid by the policyholder.

Table 4.2: Additional information of each policy.

Information	Type	Description
Policy ID	Integer	The ID of a policy.
Incident ID	Integer	The ID of a particular claim.
Incident date	Date	Date when the incident occured.
Notification date	Date	Date when the incident was reported to the insurer.
Closing date	Date	Date when the claim is settled.
Reporting delay	Integer	Difference in months between date of incident and notification.
Closing delay	Integer	Difference in months between date of incident and closing.
Category	Categorical	The type of incident, e.g. scratches or serious collision.
Number of payments	Integer	The number of payments paid out by the insurer.
Total payment	Real	The total amount paid out by the insurer.

Table 4.3: Claims information about an incident.
[1]

link the policy ID of both datasets and place the claim within the policy where the incident happened. Multiple claims can happen within a policy's lifetime but we only keep one incident and notification date within a row. The date of the first claim of the policy is kept. Each incident within a policy iteratively increases the number of claims, while the total payments of each incident are summed within a policy which results in the claim size. There are plenty of claims that could not be linked to the proper policy because the policy ID was not found. About 97% of the claims was allocated, this is a minor data issue. Furthermore 12.5% of these allocated claims lead to an ultimate claim of size 0. We therefore do not regard these as claims and the number of claims is set to zero in these cases as well. There are furthermore a very small number of negative claims, these are similarly set to zero. There are $68,024$ claims in the combined dataset.

4.2 Descriptive statistics

To get a comprehensive view of the available covariates, we want to illustrate the exposure of the insurer to the different values of the covariates. This is different to the general approach where you simply illustrate the number of occurrences of the different aforementioned values. The exposure or the duration of a policy is used to weight the contributions of each policy. When this weighting is not applied in a figure, we name the y-axis frequency instead of cumulative exposure.

4.2.1 Portfolio

The portfolio rating factors can be divided in two categories, namely policyholder - and car specific. We also have some additional information about each policy. The summary statistics are taken of the whole portfolio from policy year 2007 to 2015.

4.2.1.1 Policyholder

The policyholder specific rating factors are represented in figures 4.2 and 4.3. In the former figure we notice a small amount of very bad credit (level 4), if we use this as a categorical variable this may indicate an opportunity to join the worst two levels. But in this case we use it as a numeric predictor, since the credit score's categorisation is ordinal. As can be expected, the exposure to the different values of experience years looks pretty similar to that of the policyholder's age. We notice low amounts of very inexperienced drivers, a large part of mid-experienced drivers and this decreases again over time. Correlation between experience years and age should not be disregarded, since we assume independent covariates when we use a GLM.
The exposure to the different ages of the insured seems what you would expect: a small amount of young and old drivers and a large portion of middle aged insured drivers, it is a bit skewed to the right.
The idea of multiple licenses seems quite odd, but it means the number of different types of licenses a policyholder holds, e.g. a car and motorcycle license. Some time ago people received two licenses, since you received a motorcycle license as well. Now you only receive one. So this covariate may be a driving experience covariate in disguise, so correlation investigation could again be worthwhile.

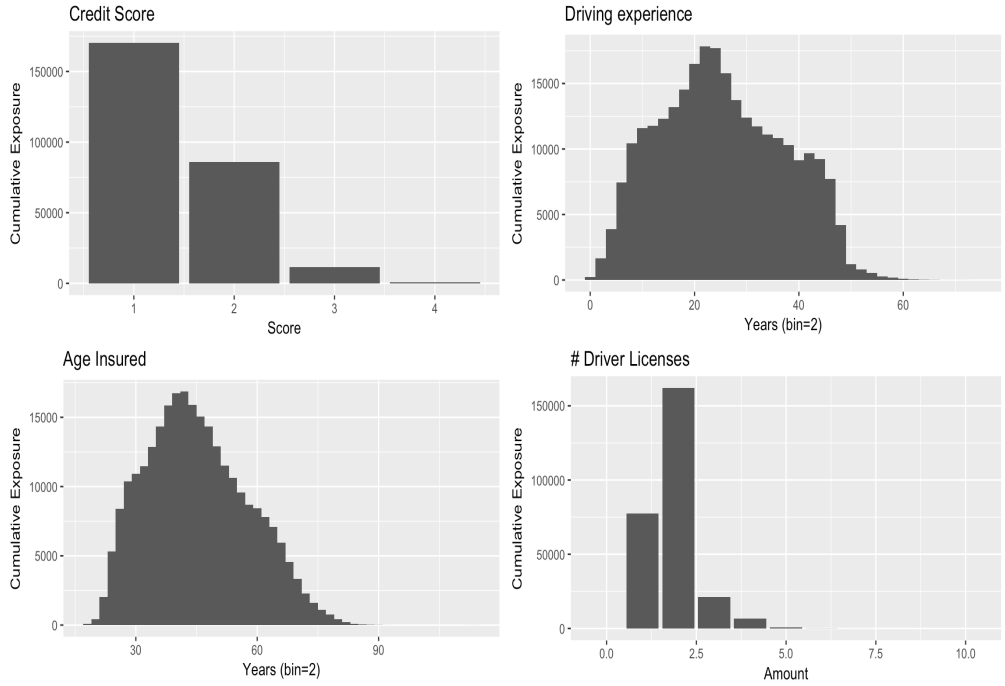


Figure 4.2: The first set of policyholder specific rating factors.

The exposures of the claim-free years in figure 4.3 appears a bit odd, since we have repeating patterns for $5-10$ and $20-25$ claim-free years. If this is very suspicious and not totally reliable, a solution might be to create a covariate that keeps track of claim-free years since the start of the portfolio. You can also notice that the claim-free years can be negative, this is a method to penalize irregular (a lot of claims) behaviour by individual policyholders. In the next subfigure, there are not many German (de) and English (en) speaking policyholders. This may indicate a need to merge these levels to reduce the effect of insignificant variables on prediction. There are 18 regions where the policyholders are located, with an additional U (unknown) location. Also here we should be wary for the effect of insignificant categories on prediction, f.e. regions $1D, 2B, ..., U$.

4.2.1.2 Car

The car specific rating factors are represented in figures 4.4 and 4.5. The former contains three histograms which are cut off at a certain level. A long tail with very low cumulative exposure would otherwise distort the representation. Both the data cumulative exposures of the car ages and car values are skewed to the right. Before using covariates for regression, we center and standardise continuous variables. But some explana-

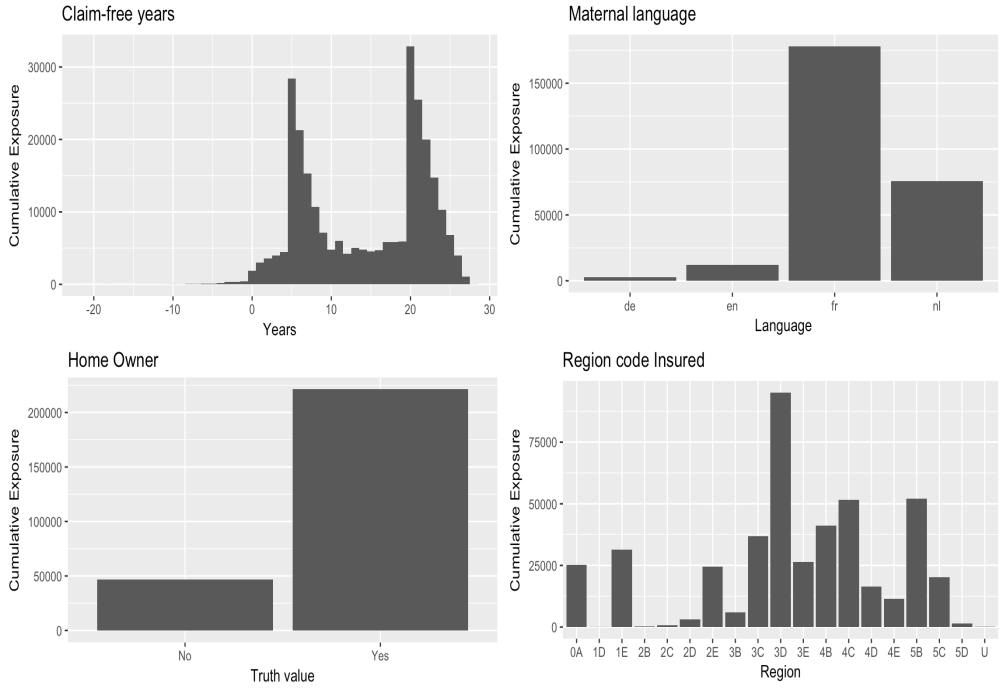


Figure 4.3: The second set of policyholder specific rating factors.

tory variables like vehicle power and catalogue value are very skewed. After centering and standardising these, we probably still have very skewed variables and these values can have a huge effect on regression. These therefore might require transforming by the logarithm function.
There were also some extreme (impossible) values for the age of the car, which transformed back to its production year was supposed to be 1360. This is impossible, an investigation on the type of car showed that the production years was supposed to be 1960. The vehicle weight in kg's seems to be more heavily massed in the middle than the power of the car in cc's and both are a bit skewed to the right. In practice the Power-to-Weight ratio is often used to replace the previous two by one covariate. We will consider using this as well.

The accessories in figure 4.5 have six distinct values, where four have a negligibly small cumulative exposure. Since this categorisation is again ordinal, we use this as a numeric predictor. Car Mileage has a heavy mass of exposure in the $10,000 - 30,000$ category, this variable was of origin a categorical variable. Mileage is an ordinal categorical variable as well. So we can transform it to the values 1,2 and 3 with increasing number of kilometers driven in a year. The next subfigure shows that most insured cars drive on diesel and petrol, while very few drive with electric, gas or hydrogen. We can transform this to Diesel drivers and non-diesel drivers. The final car specific rating factor is the number of cars owned, here we notice that the insurer has large exposure to policyholders who own one or two cars. The exposure rapidly decreases with increasing number of cars owned by the policyholders. This is yet again an ordinal categorical variables and therefore is used as a numeric predictor.
A car specific rating variable that is missing are the brands of the cars. There are 32 brands of cars after processing. There were a lot of duplicate but differently spelled brands, which were corrected. Furthermore we joined brands with an exposure less than 500 in the brand named *other*.

4.2.1.3 Additional

Visual representation of the addition information about the portfolio is present in figure 4.6. For the deductibles an actuary is faced with an important question: Do you segment the data on the different deductibles and model the data separately? It makes sense if you think about it; if a policyholder has to fund a larger amount of an incident himself before he can receive compensation from the insurer, then it should be less likely for this policyholder to report a claim, in both number and average size, in a given year. Most commonly the insurer just introduces the deductible as another predictor.

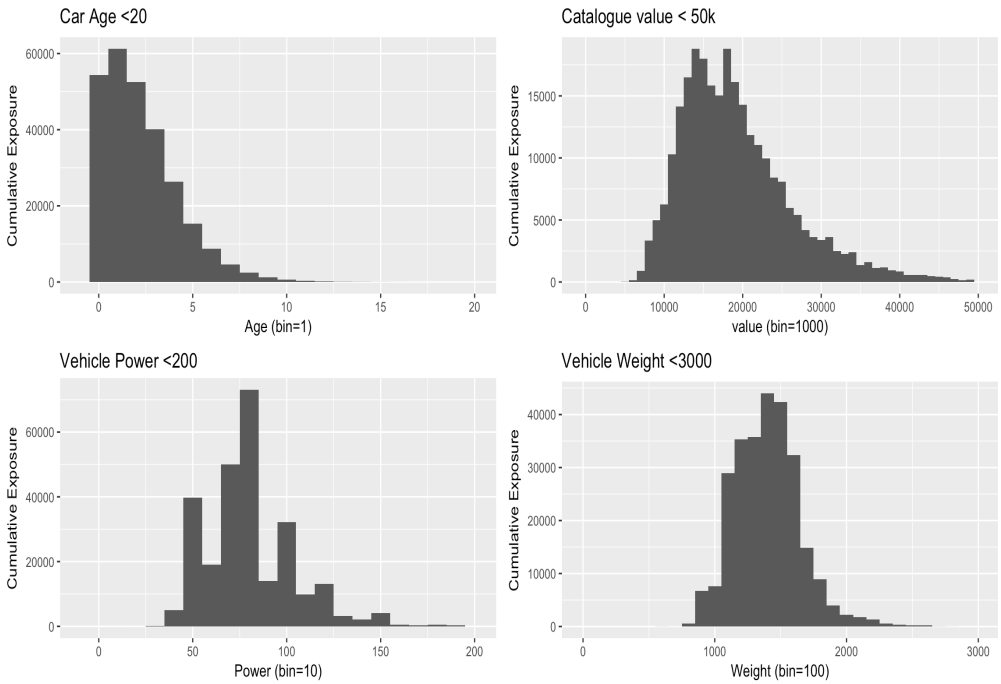


Figure 4.4: The first set of car specific rating factors.

The next subfigure shows the total duration of the policyholders and indicates how long a policyholder, up until the date of reference, has held a contract with the insurer. The final subfigure of 4.6 illustrates the premium paid by the accounts. It looks like a gamma distribution which is skewed to the right. Also here we choose a cut off, in this case of $5,000$, because it would otherwise elongate the figure a lot. There are only 63 policies with a premium higher than the cut off point. The highest premium is of $84,241.89$ for a Ferrari California with a catalogue calue of $650,000$.

4.2.2 Claims

The figures of portfolio characteristics are followed by some descriptive data about the claims. First some figures (4.7 and 4.8) about the information summarized in table 4.3, then an analysis of the average claim size for each claim category.

4.2.2.1 Claims information

The damage and notification dates in figure 4.7 give insight into the evolution of the portfolio (in size) of the insurer over the years and the thus increasing exposure. There is a slight decline in claims at the end of the portfolio, this is because we do not consider claims that have not been closed yet, which are of course mostly located at the end of the claims portfolio. It is also noticeable that some months show some peaks in claim activity.

How the categorisation of the different claim types was done, can be seen in the appendix A.3. There are 7 (A-G) distinct categories: Collisions, Traffic rules, Scratches, Parking and maneuvering, Weather/Nature, Unknown and Theft/Vandalism/Loss. In figure 4.8, we notice the difference in exposure to the different claim categories. The two largest categories by a distance are the Scratches (C) and Unknown (F). This suggests that if these two claim categories show quite different behaviour for the number of claims and claim size, we could model the claims differently. We could then join the remaining smaller categories to one of these two on the basis of largest similarity.

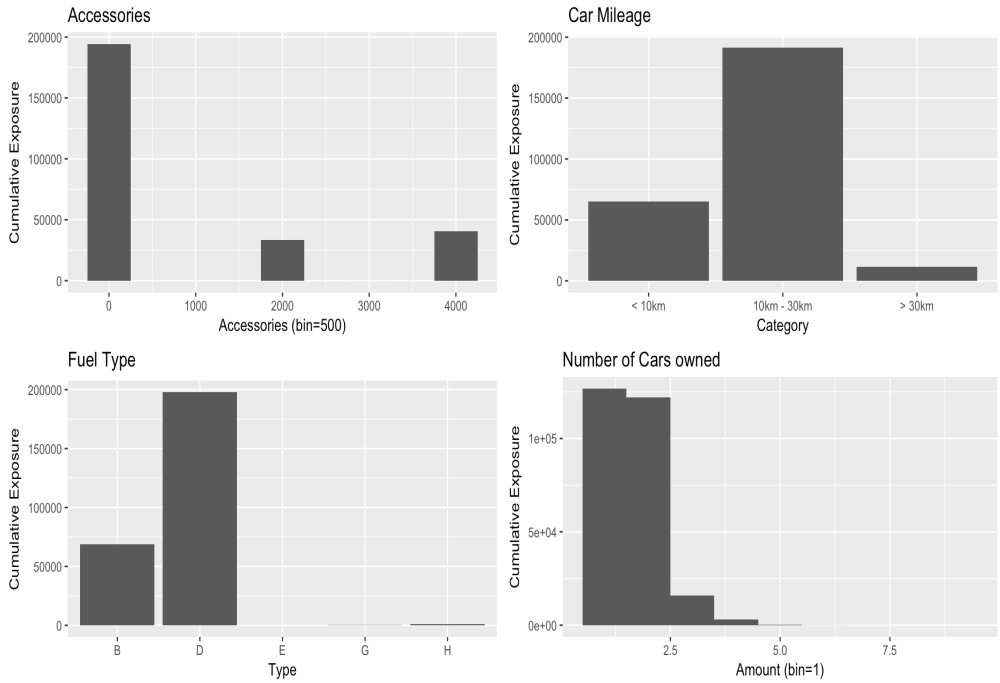


Figure 4.5: The second set of car specific rating factors.

4.2.2.2 Claim categories

The claims belong to different categories. Each signifies a type of damage that has occurred to the car. We expect different distributions of the claim size for the types. In each of the figures (4.8,4.9 and 4.10) the dashed red line signifies the mean value of all claims in a specific category. All categories combined are represented in figure 4.8, the claims combined have a mean value of $1,675.82$, you can also notice that it is skewed to the right. We noticed in the previous paragraph that the insurer is most exposed to claim categories C and F by a distance, so we start with those subfigures in figure 4.10. The mean value of claim in category C is 437.97, while the mean for claims in category F is $2,205.43$. The means for collisions (A), Weather/Nature (E) are pretty high as well, see table 4.4.

Claim type	A	B	C	D	E	F	G
Mean claim size	2,613.29	2,577.20	437.97	1,478.29	3,484.90	2,205.43	1,120.26

Table 4.4: Mean claim size for the different claim types.

4.2.2.3 Number of claims

To get a sense of the empirical observations for the number of claims, we provided table 4.5. There are a lot of zero claims in the portfolio, 86.91% of the whole portfolio to be precise. This is typical for an insurance portfolio. In the next table (4.6) we see some fluctuations throughout the years for the number of claims and the claim rate. The claim rate is the number of claims divided by the total amount of exposure.

Number of claims	0	1	2	3	4	5	6	7
Total	385,267	49,485	7,337	984	172	32	5	5
Percentage	86.91	11.16	1.66	0.22	0.039	0.0072	0.0011	0.0011

Table 4.5: Number of claims

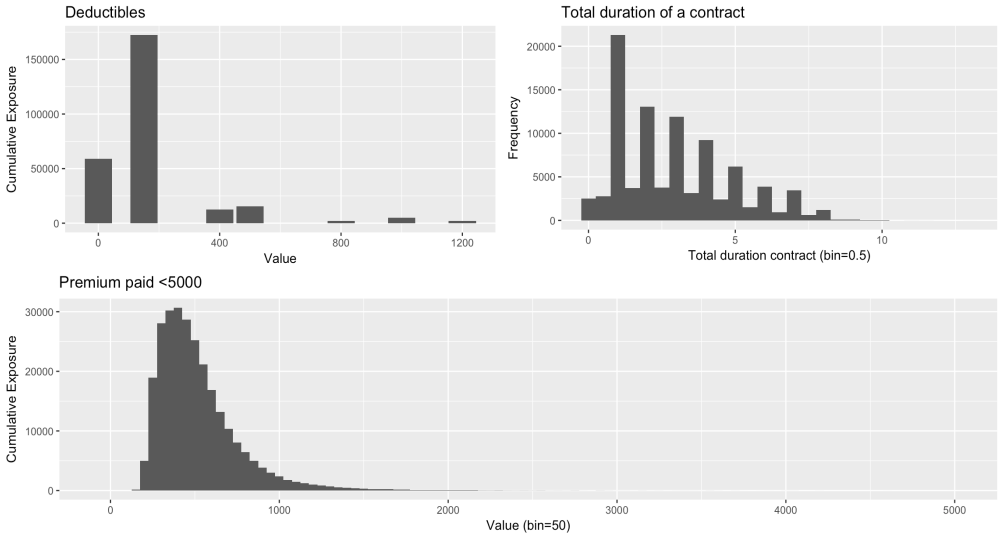


Figure 4.6: The additional information of each portfolio.

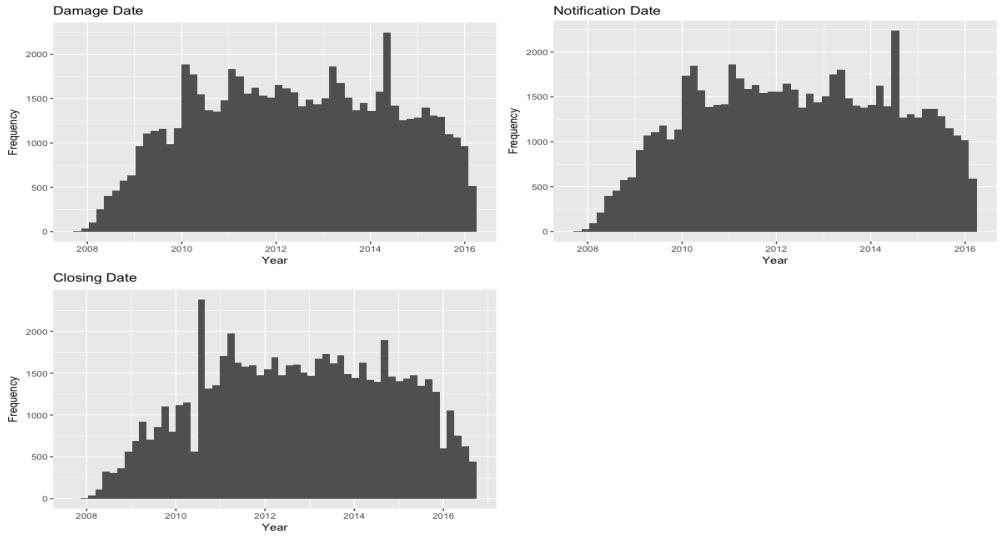


Figure 4.7: Reporting information of claims.

Policy Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Number of claims	340	4,745	8,524	9,973	10,345	9,813	9,750	9,626	4,833
Claims per exposure	0.16	0.29	0.33	0.30	0.25	0.22	0.22	0.22	0.28

Table 4.6: Number of Claims and claim rate per policy year

4.2.2.4 Claim severity

We use a similar approach for the claim severity and investigate these through the years. The findings are in table (4.7). It appears the the claim severity is gradually increasing through the years.

Policy Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Average claim severity	1,435	1,614	1,505	1,527	1,716	1,725	1,754	1,743	1,984

Table 4.7: Average claim severity per policy year

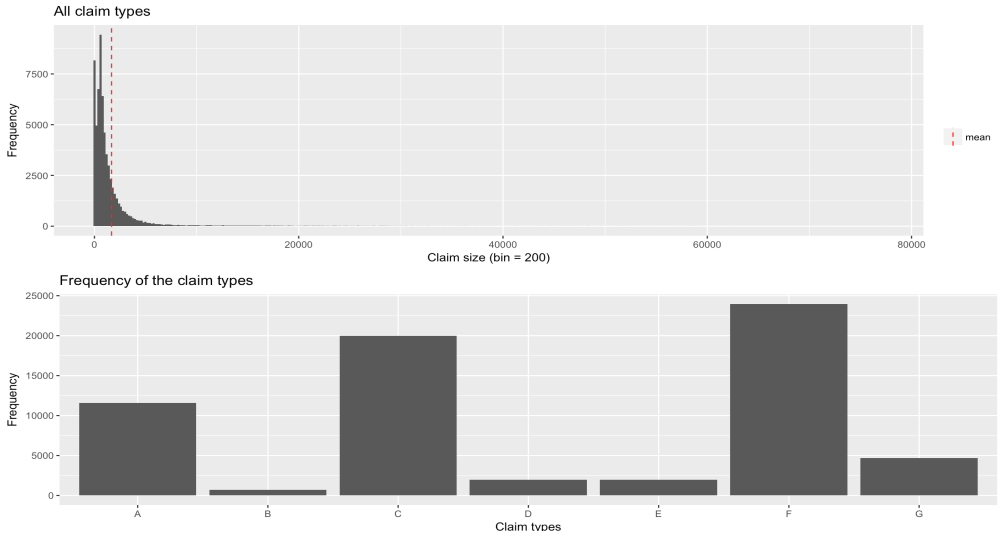


Figure 4.8: Additional information of the claims.

4.2.3 Conclusion

All the continuous covariates are centered and standardised for use by GLM. We do this enhance interpretability of the fitted coefficients. The Machine Learning techniques do not require centering and standardisation of the covariates since they are not sensitive to skewed covariates with extreme values because of the regression tree approach: each plane is assigned a constant. Covariates such as catalogue value and vehicle power are very skewed, therefore additional predictors are created that transform these predictors by the logarithm function.

In case of some methods such as the GLM, it is also important to consider the effect on prediction due to insignificant variables. This is an issue when a categorical variable is used where an instance has a very low count. This is solved by joining two levels that are the most homogeneously distributed. Other methods such as the Random Forest model are not very susceptible to category instances with low counts, since in that case they will be used far less in the ensemble of trees.

If the models do not perform very well, it can be of help to segment the data on the deductibles or the claim types. As a consequence the data will be more homogeneous.

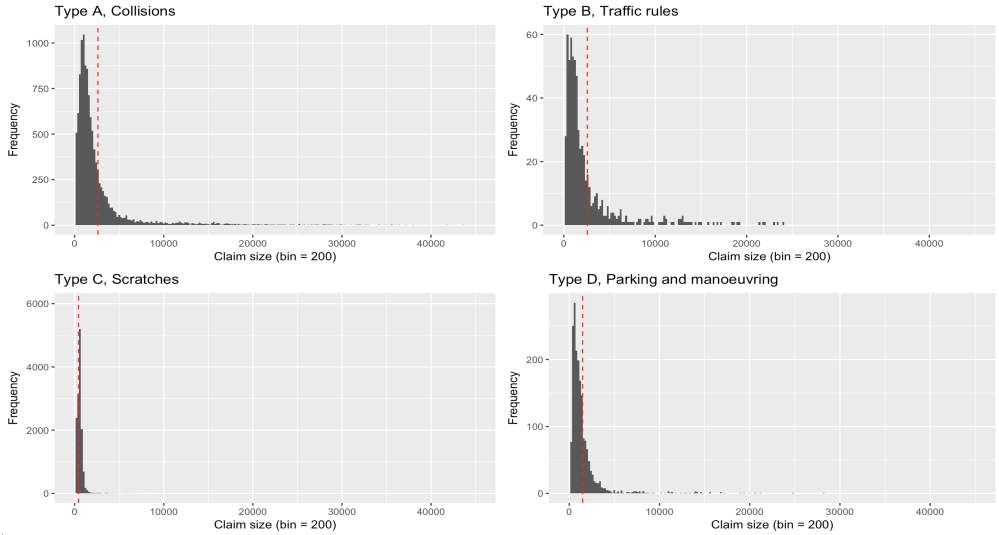


Figure 4.9: Average claim size for claim categories A to D.

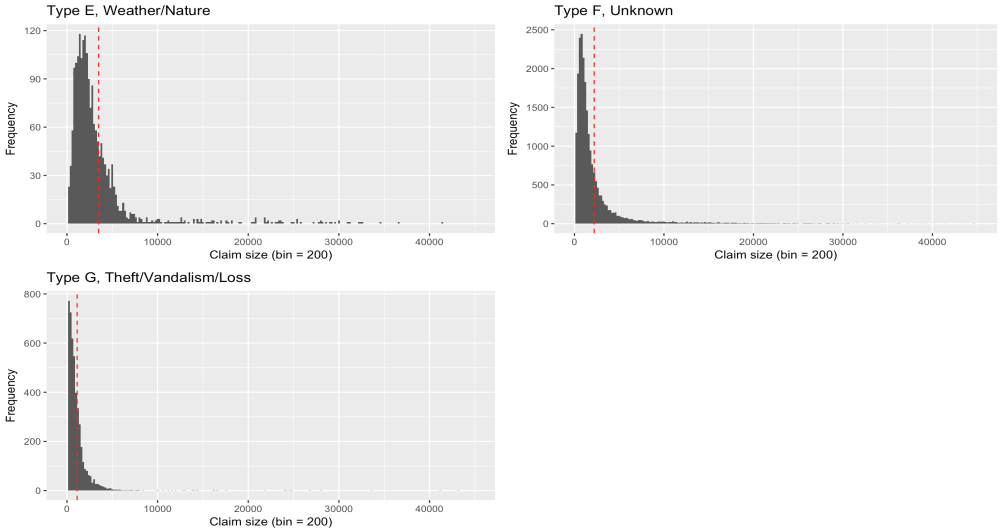


Figure 4.10: Average claim size for claim categories E to G.

Chapter 5

Predictive performance of various methods when applied to the car insurance dataset

Contents

In this chapter we commence by providing a framework to select a model (section 5.1). This is followed by a testing strategy to prevent overfitting the data (section 5.2). Then we compare the predictive performances of the current approach in the actuary field (GLM) to the tree-based Machine Learning techniques (RF and GBM) on a subset of the data, since training these machine learning techniques is quite time consuming (section 5.3). The fitted GLMs are analysed with residual analyses and the training of the Machine learning techniques is investigated. The Machine learning techniques contribute to insights into variable importance and interaction, which leads to the implementation of interaction effects (section 5.4) and the implementation of the non-linear effects (section 5.5). The deviance residual analyses of the fitted GLMs showed that some covariates, due to very low exposure, resulted in very high leverage; these are handled in section 5.6. Finally we predict the premiums and compare the separate models to some direct models (section 5.7.1). The programming domain of R is used throughout the whole methodology of this thesis.

5.1 Model selection

The objective of risk premium pricing is stated in section 3.2. An insurer would thus want to set the prices each year. If a new policyholder joins or a policyholder renews, a premium is proposed to the policyholder. This premium is determined by the acquired information about the policyholder. The smaller the deviation between the assigned premium and the total claim amount the policyholder will report during the lifetime of the contract, the better the model is. This distance is therefore the performance criterion of a model. The root-mean-square error (RMSE) is used as the model performance criterion.

5.2 Testing strategy

A testing strategy is implemented to simulate evaluating a model on new policyholders by leaving out part of the policyholder and claims data. A good testing strategy prevents overfitting of a model. There are two phases when building a model: the training and testing phase. In the training phase, each different algorithm is treated separately and this phase is used to improve the model. For example 10-fold cross-validation can be used to train and validate the different models. After this phase, a final model per algorithm is chosen. These different algorithms are finally evaluated on the test set. The best performing algorithm is then chosen based on the model selection criterion, the root-mean-square error (RMSE). But first, we select a method to separate the data into a training and testing set. The following splits are popular approaches.

i. Train model on all policies except those in the most recent year, which are left for testing purposes.

ii. Train model on a random subset (f.e. 75% of the data) and test on a random subset (f.e. 25% of the data).

We will do a combination of both. The policy year 2015 is held out for performance tracking, while the previous years are used the build the models. On the years before 2015, we use testing strategy ii, because it allows us to train models on all available years, so also the most recent one.
A basic GLM model is created by performing backward elimination by the Akaike Information Criterion (AIC). If new iterations of the GLM model are proposed, when f.e. introducing squared predictors, the AIC is used again to evaluate the model.
The Machine Learning models are trained by tuning the parameters. This is achieved by setting up a grid for the tunable parameters. Each combination of tuning parameters is then used to create sum-of-trees models and are then evaluated with 10-fold cross-validation (see figure 5.1). The grid values of the tuning parameters are chosen in a manner such that we consider three to six different values for each parameter. These values should be within a reasonable range. The range of values are recommended by numerous studies and books [21]. Use in practice has shown that when we deviate from this range of values, we will likely underfit or overfit the model. In case of RFs, the sole tuning parameter is the number of covariates to randomly use at splits m (see section 3.5.2). GBMs have four tuning parameters, namely the number of trees K, the tree depth γ, the shrinkage ν and the minimum number of observations in the terminal nodes m_{min} (see section 3.5.3). The selected grids are the following:

$$\text{RF:} \quad m \in \{1, 2, ..., 5, 6\} \qquad \text{GBM:} \quad \begin{cases} K \in \{250, 500, 1000, 2000\} \\ \nu \in \{0.1, 0.01, 0.001\} \\ \gamma \in \{2, 4, 6, 10, 15\} \\ m_{min} \in \{10, 20, 100\} \end{cases}$$

5.3 Basic GLM versus Machine Learning techniques

In this section (5.3) we separate the modelling of the number of claims response (5.3.1) and claim severity response (5.3.2). We use the data before policy year 2015. The resulting dataset is still quite large ($443,287$ rows), and since training RFs and GBMs is very slow, the models are trained on a 20% subset of the data and tested on a 10% subset of all data. As a result the training set has $80,176$ rows and the test set has

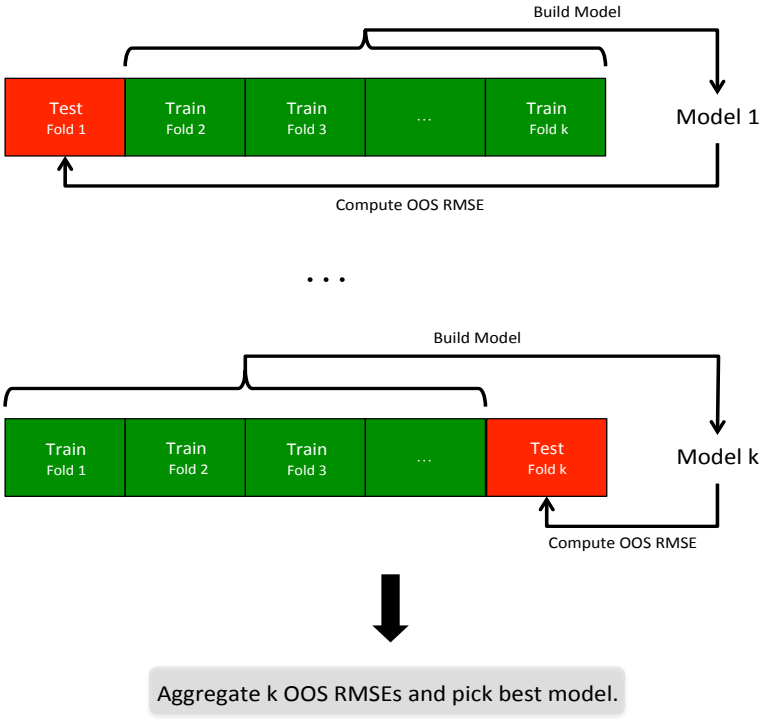


Figure 5.1: Illustration of k-fold cross validation process to select tuning parameters of the Machine Learning models.

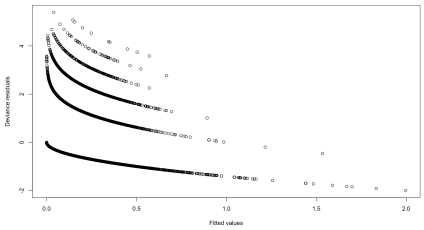
$40,089$ rows. The number of claims and claim severity models are created independently. The following models are tested on held-out data: the "mean model" for the number of claims, the "mean model" for the claim severity, a GLM, a GBM and a RF. The approach to creating these models is very similar for both the claim severity and number of claims models.
The GLMs are fitted using centered and standardised variables. Some of the variables (see section 4.1) are then still very skewed. This leads to very extreme predictions when using GLMs. Therefore the explanatory variables *Catalogue Value*, *Vehicle Power* and *Vehicle Weight* are log-transformed. There is also a new covariate (for GLM) named *Reference Year*, which is simply *Policy Year* subtracted by the year 2011. The range of integer values of *Reference Year* is therefore $[-4, 4]$ for the before policy year 2015 data.
The models are separately applied for the number of claims and claim severity.

5.3.1 Methods to model the number of claims response

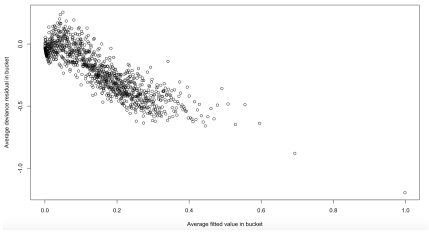
The results of the models applied to the number of claims c_i per policy i that have a duration e_i are first summarised. Each fitted model is then analysed in depth. The GLM is checked with a residual analysis and the training procedure of the Machine Learning techniques are investigated.

5.3.1.1 Summary of results

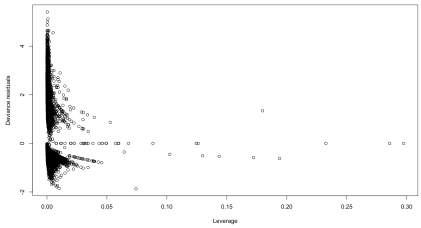
We first make a benchmark model for the number of claims. Just using the average number of claims out of the train set to predict on the test set is not a good model. This does not take into account the lifetime of each policy. This is why we do the following. The benchmark mean "model" is made by first taking the weighted average of the number of claims in the train set $\bar{f} = (\sum_i^{\text{train}} e_i c_i)/(\sum_i^{\text{train}} e_i)$, then we multiply this value with the exposures of the test set to get your predictions $c_i^{\text{test}} = \bar{f} * e_i^{\text{test}}$. We then move on to the GLM, this is created by backward elimination of some covariates using AIC. The result is summarised in table A.1 and figure A.2. RF achieves its best tune for $m = 3$, while the best tune for GBM is $K = 1000$, $l = 10$, $\nu = 0.01$ and $m_{min} = 100$. The RF performs best on the 10% out-of-sample dataset. Therefore we investigate the variable importance and partial dependence plots of the RF and not the GBM.



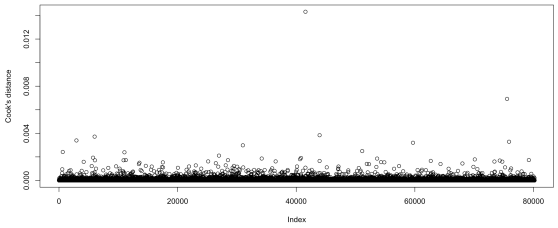
(a) Deviance residuals versus the fitted values of the Poisson GLM for the number of claims.



(b) Summed deviance residuals per bucket of fitted values.



(c) Deviance residuals vs leverage for the number of claims response.



(d) Cook's distance for the number of claims response.

Figure 5.2: Residual plots of the Poisson GLM for the number of claims.

	Mean	GLM	GBM	RF
RMSE	0.4515	0.4495	0.4364	0.410

Table 5.1: RMSE Out-of-sample for each number of claims model on a subset of the data.

5.3.1.2 Residual Analysis of the GLM

The residual plots of the fitted number of claims by the GLM are presented in figure 5.2a. Residual versus fitted plots of count data such as these are especially difficult to interpret since you will always see lines in the output. This is because we have a finite number of possible outcomes (0 to 7) and our fitted values are all quite low because of the high mass of no claims in the portfolio. There are some methods that do allow interpretation in this case. You either use LOWESS (locally weighted scatterplot smoothing) or you bin the fitted values from smallest to largest and add the residuals together for each bin. These binned values are visible in figure 5.2b. There is a downward linear trend visible in these 'crunched' residuals. So for increasing fitted values, the overall weight of negative deviance residuals increases. This indicates that the model does not do a very good job at fitting the data. There are two feasible reasons why: (i) the Poisson GLM model is not adequate or (ii) important predictors are missing. The latter is easy to try and was not an issue, so it seems the reason might be the former. The higher number of zeros in the number of claims data than the Poisson GLM fits is possibly the root of the issue here. Improvements can be made by fitting *Zero Inflation* or *Hurdle* models.
Subfigure 5.2c plots the deviance residuals against the leverage (hat-values). There are ten policies with leverages greater than 0.10. These policies are highly leverage because of a combination of very infrequent instances of categorical variables that these policies possess. Especially because of infrequent *Regions* such as $1D$ and *Fuel types* such as E and H. Only one of these ten policies has a absolute deviance residual larger than one. This is evidently the account with the largest cook's distance as well. The fitted value for this policy is 0.16, while the actual value was 1, so we do not see this as a problem. However, if we hold to the second threshold rule of the Cook's distance (see section 3.4.5.2), then we should remove this outlier. But then again about 83% of the data should be removed according to this threshold rule, since $4/n \approx 5e{-}5$. Removing this outlier does not visibly decrease the downward trend of the 'crunched' residuals. These are however policies that should be followed with care.

5.3.1.3 Training of Machine Learning techniques

The Machine Learning techniques are tuned using 10-fold cross-validation on the 20% train set. The performance is kept as the RMSE of the combined ten hold-out sets. The applied grids are stated at the end of section 5.2. We first start with the training of the GBM for the number of claims, followed by the training of the RF.
The results of training the GBM are summarised in figure 5.3. It appears that when the shrinkage is quite large ($\nu = 0.1$) and the iterative "improvements" or changes are thus bigger, increasing the number of trees or the interaction depth (complexity) has a negative effect on the performance. With shrinkage 0.001 the performance does gradually increase with an increased interaction depth although it does stagnate at the end. Increasing the number of trees with this shrinkage does seem to improve the model at every step. The iterative improvements on the hold-out sets during training with shrinkage 0.001 while increasing the number of trees seem to indicate that when we keep on increasing the number of trees that the performance will increase as well. This has been tested outside of this training environment and the result did not perform better than the best train of this grid that is achieved with shrinkage $\nu = 0.01$, $K = 1000$ trees, $m_{min} = 100$ observations in the terminal nodes and interaction depth $\gamma = 10$.
The training of the RF is a lot easier to interpret. The RF increases performance at first when increasing the randomly selected predictors per tree m. This process reverses with more variables drawn than 3. The best performance on the hold-out sets is therefore reached when $m = 3$.

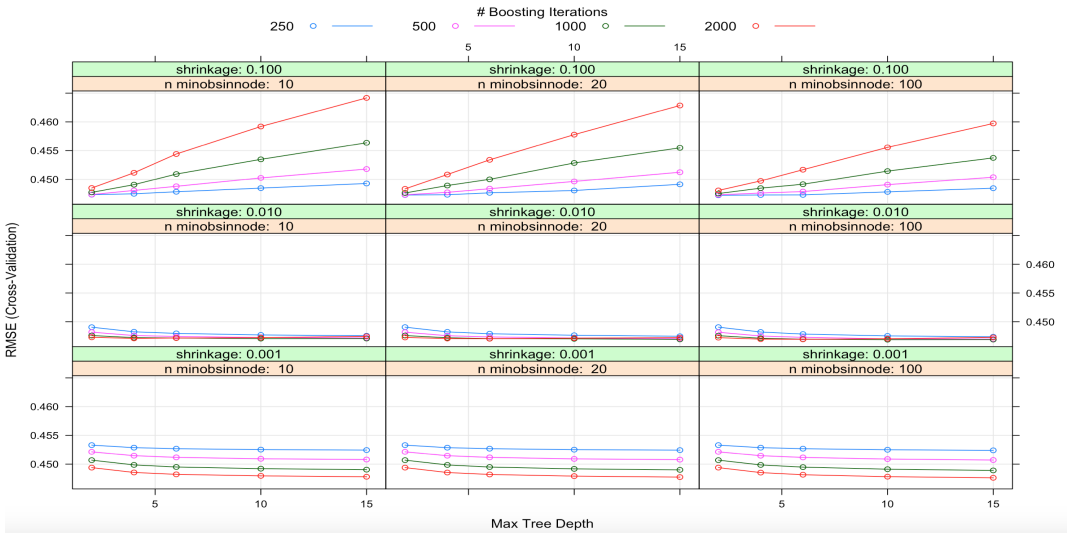


Figure 5.3: RMSE on the hold-out folds with the different tuning parameters of the GBM for the number of claims response.

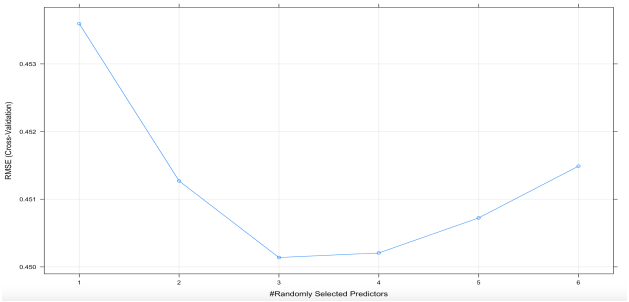


Figure 5.4: RMSE on the hold-out folds with the tuning parameter of the RF for the number of claims response.

5.3.1.4 Variable importance and Partial Dependence of the RF model

It is in line of expectation that the duration of a policy is one of the most important covariates (see figure 5.5) when a RF model is built to predict the number of claims. Owning a home or having a car with a different fuel type does not greatly influence the number of claims an insured reports. There seems to be a big drop in importance from the variable *Policy year* to *Number of Licenses*, this therefore seems a good cut-off point to separate between the important and unimportant variables.

The marginal effect of the duration on the number of claims is not as linear as one would expect. The issue with the marginal here is that it does not monotonously increase, this is something an insurer would require however. An alternative is to use alternative software that allows RFs to be implemented with an offset term.

There are some partial dependence plots or marginals that show similar nonlinear or quadratic effects on the number of claims. The covariates that induce these effects are the *Policyholder Age*, *Experience Years*, *Vehicle Weight*, *Claim-free Years*, *Catalogue Value*, *Car Age* and the *Deductible* plots; where the two latter plots are in the appendix figure A.4.

If these non-linear effects have a big contribution to the predictive performance, then this could be a reason the GLMs perform less well since the fitted GLMs do not include quadratic terms yet.

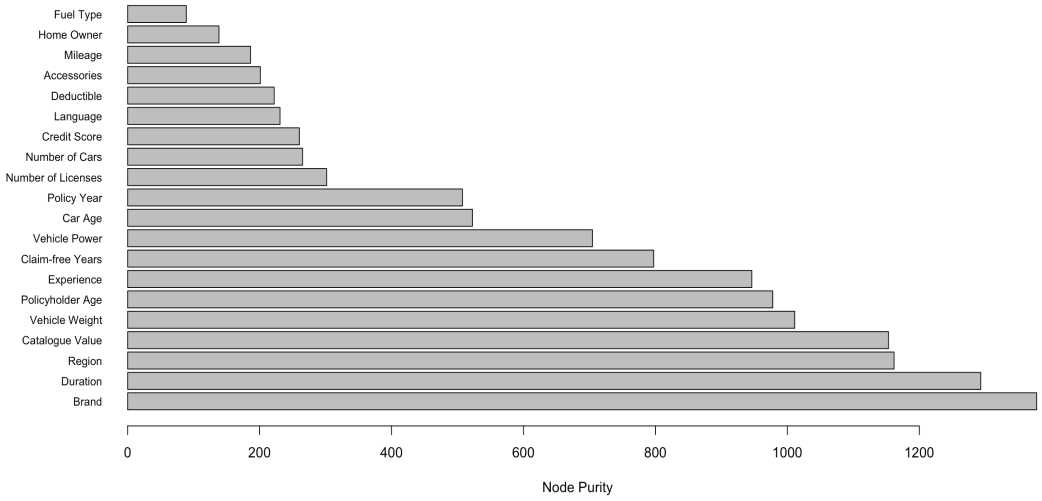


Figure 5.5: Variable importance according to the Random Forest number of claims model on a subset of the data.

5.3.2 Methods to model the claim severity response

To model the claim severity response s_i, we apply the same approach as used when modelling the number of claims response: first a summary of the results, followed by an in depth analysis of each model.

5.3.2.1 Summary of results

The dataset for severity models is a lot smaller, since we model the average claim size given the number of claims is larger than zero. As a result we have a training set of size $10,624 \times 20$ and a testing set of size $5,523 \times 20$. The mean "model" is this time simply projecting the mean value of the claim severity in the train set on the test set. We perform backward AIC and eliminate some covariates, the fitted GLM severity model is summarised in table A.2 and figure A.3. RF achieves it best tune with $m = 2$, while the best tune for GBM is $K = 250$, $\gamma = 6$, $\nu = 0.01$ and $m_{min} = 100$. A residual analysis of the GLM for the claim severity is done, followed by an inspectation of the trained models.

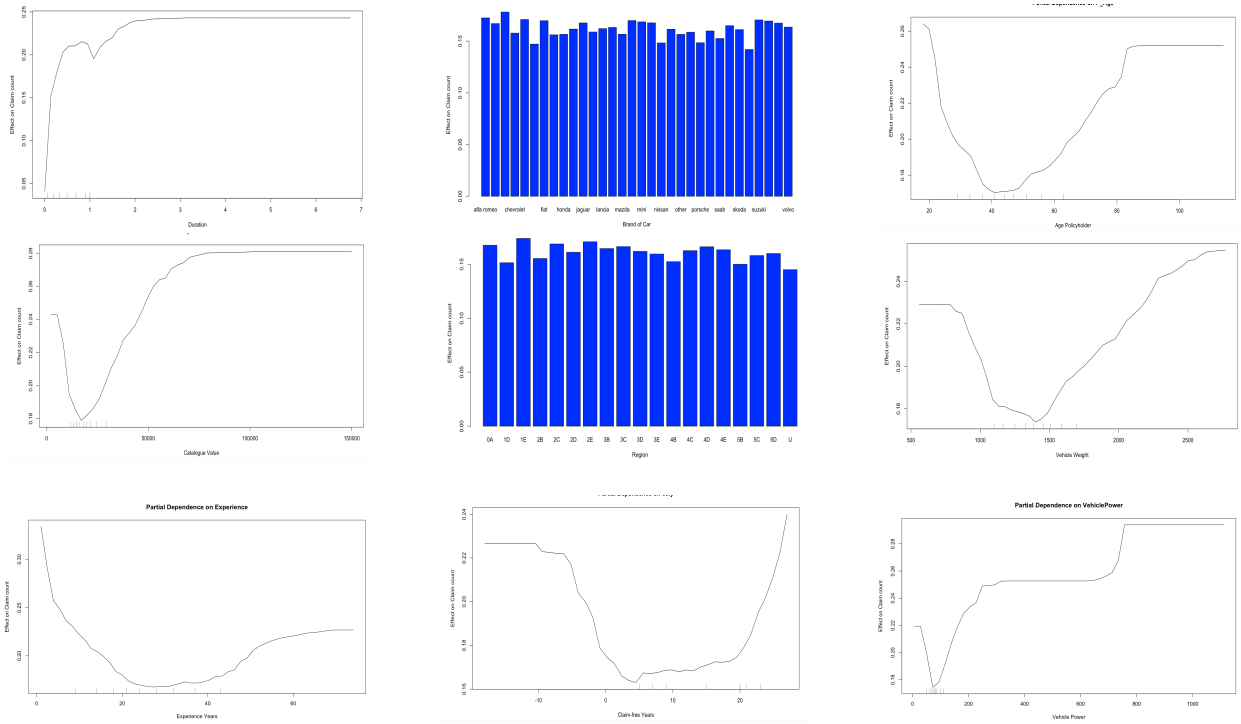


Figure 5.6: Marginal effects of the different covariates on the number of claims. The nine most important variables of the Random Forest model are illustrated. The excluded plots can be found in the appendix (figure A.4).

	Mean	GLM	GBM	RF
RMSE	3159.1	3147.5	3144.1	2964.9

Table 5.2: RMSE Out-of-sample for each severity model on a subset of the data.

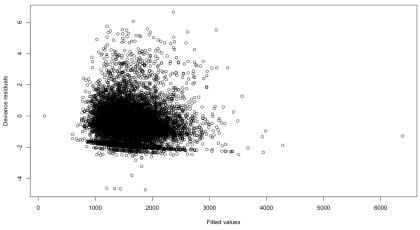
5.3.2.2 Residual Analysis of the GLM

For deviance residuals (dev) versus fitted values plots we ideally want the deviance residuals to be unstructurally dispersed in a symmetrical fashion around zero. If we look at figure 5.7a we notice two concerning patterns.

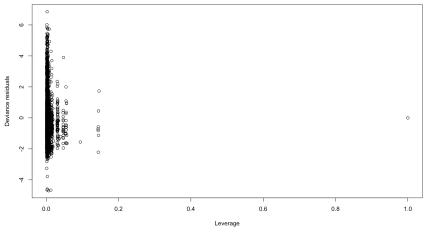
i. The variance is substantially greater for deviance residuals larger than zero.

ii. There are strange line patterns below zero.

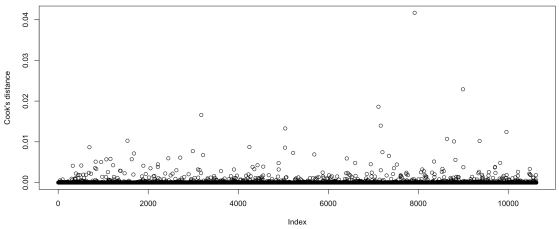
Let's first focus on (i.) the difference in variance of the deviance residuals above and below zero. There are two possibilities in solving this issue. We either fit a more heavy-tailed distribution such as the Lognormal distribution or we put a ceiling on the average claim sizes and leave these values for a seperate analysis. We apply both methods to the data. The first approach is applied by taking the claim severity dataset and transform the average claim sizes s_i by $f : \mathbb{R}_{(0,\infty)} \to \mathbb{R}_{(0,\infty)}$, where $f(s_i) = \log(1 + s_i)$. Then we apply a linear regression to this dataset.
The second approach is sensitive to the choice of the threshold $\mathcal{T}$ for the severity dataset. For simplicity we set $\mathcal{T}$ approximately at 97.5% of $s_i \leq \mathcal{T}$. As a result $\mathcal{T} = 8000$.
If we look at subfigures 5.8a and 5.8b, you notice that the variance of the deviance residuals (dev) above and below zero are more equally dispersed. In fact for the newly applied methods we have $\sigma(\text{dev}_{logN}|\text{dev}_{logN} > 0) = 0.756$, $\sigma(\text{dev}_{logN}|\text{dev}_{logN} < 0) = 0.814$, $\sigma(\text{dev}_{cap}|\text{dev}_{cap} > 0) = 0.637$ and $\sigma(\text{dev}_{cap}|\text{dev}_{cap} < 0) = 0.594$; while we had $\sigma(\text{dev}_{gam}|\text{dev}_{gam} > 0) = 1.031$ and $\sigma(\text{dev}_{gam}|\text{dev}_{gam} < 0) = 0.609$. Both methods reduce the absolute difference $\left|\sigma(\text{dev}|\text{dev} > 0) - \sigma(\text{dev}|\text{dev} < 0)\right|$, although the *capped-gamma*



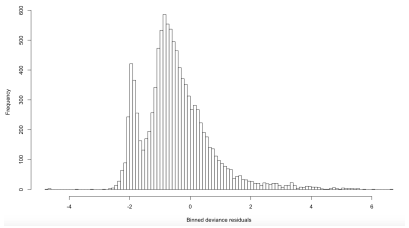
(a) Deviance residuals vs fitted.



(b) Deviance residuals vs leverage.



(c) Cook's distance.



(d) Histogram deviance residuals.

Figure 5.7: Deviance residual plots for claim severity.

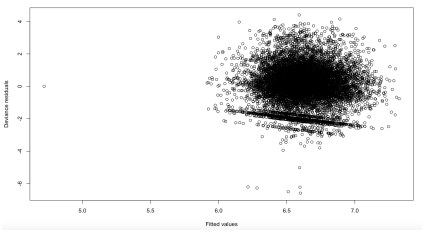
model reduces the absolute difference the most. But a downside is that it disposes about 2.5% of the data and requires an additional model for the remaining large-claims dataset.
Now let's revisit the issue (ii.) of the line patterns. We notice that these occur for deviance residuals between -1.5 to -2.2, notice the spike on histogram 5.7d as further evidence. When we extract these data points, we see that 74% of claims were of claim category C, while this category occurs for 31% of the claims in the 20% train set. Furthermore we recognise a lot of recurring claim sizes, which indicates this might be some fixed handling fee of claims from category C (Scratches).
Now we refit the original GLM model as well as the methods mentioned in (i.) to the dataset excluding claim category C, as a result we acquire the residual plots 5.8c, 5.8d and 5.8e. It is visible the line patterns below zero reduce but they do not completely disappear. The portfolio and claims datasets were combined (see section 4.1) such that a policy can hold multiple claims in a row. But only claim information such as category and date are kept of the first of these claims within a single policy. This setup is therefore not fit to seperate on the claim categories fully.

5.3.2.3 Training of Machine Learning techniques

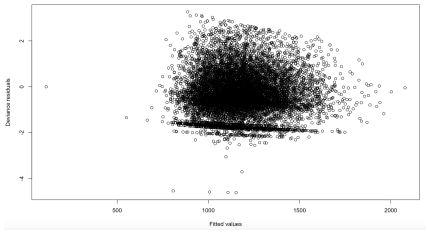
The approach here is exactly the same as it was for the number of claims models. We first start with the training of the GBM for the claim severity process, followed by the training of the RF.
The results of training the GBM are summarised in figure 5.9. It appears that when the shrinkage is $\nu = 0.1$ or $\nu = 0.01$ that increasing the number of trees or the interaction depth (complexity) has a negative effect on the performance. Increasing the observations held in the terminal nodes does have a positive effect on the performance on the hold-out sets. With a low shrinkage $\nu = 0.001$, the performance does not improve at all and falls flat with changes to the interaction depth or the number of trees used. The predictive performance of the model with shrinkage $\nu = 0.01$ does however achieve the best performance with $K = 250$ trees, $m_{min} = 100$ observations in the terminal nodes and interaction depth $\gamma = 6$.
The training of the RF is again more straightforward to interpret, see figure 5.10. The RF increases performance at first when increasing the randomly selected predictors per tree m. This process reverses when more variables are drawn than 2, as was the case for the number of claims response. Therefore $m = 2$ is the best tune for RF in case of the claim severity.

5.3.2.4 Variable importance and Partial Dependence of the RF model

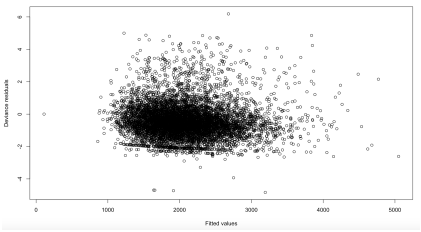
One of the main motivations to split modelling of the total claim amounts into the number of claims and claim severity was because of the difference in importance of different variables to each of these different



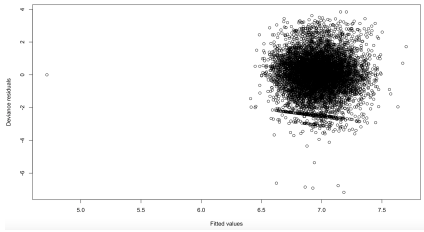
(a) Lognormal distribution.



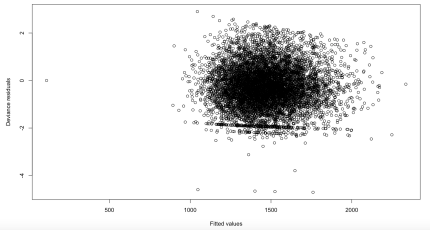
(b) Capped.



(c) Not claim type C.



(d) Not claim type C and Lognormal distribution.



(e) Not claim type C and Capped.

Figure 5.8: Capped versus Lognormal approach for claim severity, followed by excluding C and applying these methods on the excluded C dataset.

models. But when we compare figure 5.11 to figure 5.5, we see that the same variables are in the top ten of most important variables (excluding the duration), where some have shifted one place up or down.

If we have a look at the partial dependence plots in case of the claim severity, we again notice some non-linear or in some cases quadratic effects here. This is most pronounced for the covariates *Policyholder Age* and *Experience*. The correlation between the two covariates is more pronounced when modelling the claim severity. The marginals are very similar. Correlation between independently assumed explanatory variables can be harmful but that does not necessarily have to be the case.

5.3.2.5 Multicollinearity

Especially the marginals of the RF severity model suggested that both the *Policyholder Age* and the *experience* are likely correlated. Pair plots are an easy way to visualize the relation between pairs of explanatory variables. We make pair plots for the variables and illustrate a subset of these in figure 5.13. It seems that most pairs of explanatory variables do not show clear (linear or non-linear) correlation patterns, except for the aforementioned pair and the pair *Catalogue value* and *Vehicle power*. For these pairs we see they are directly proportional in most cases. Furthermore we notice that $Age \geq Experience$ holds for all data. A way to better fit this, is to incorporate the correlation between these variables in the models.

5.3.3 Conclusion

Both the GLMs for the number of claims and claim severity have some flaws. The Poisson-GLM for the number of claims response exhibits a downward trend in the 'crunched residuals' for increasing fitted values. This indicates the model is flawed in some manner. An approach that can resolve this issue is to better model the large proportion of zeros present in the number of claims dataset. Approaches that allow this are *Zero Inflation* and *Hurdle* models.
The residual analysis of the claim severity Gamma-GLM reveals that it has some issues modelling the heavy-tailed aggregate claim size data and that there are some strange line patterns that predominantly occur for claim category *Scratches* (C). When the Lognormal distribution is used with linear regression or when the aggregate claim sizes are capped at 8000 and the Gamma-GLM is still employed, the deviance residuals are more similarly dispersed above and below zero. But when the Lognormal distribution is used with linear regression, the RMSE on the hold-out set is 3293.2. The RMSE on the hold-out set of the capped model is

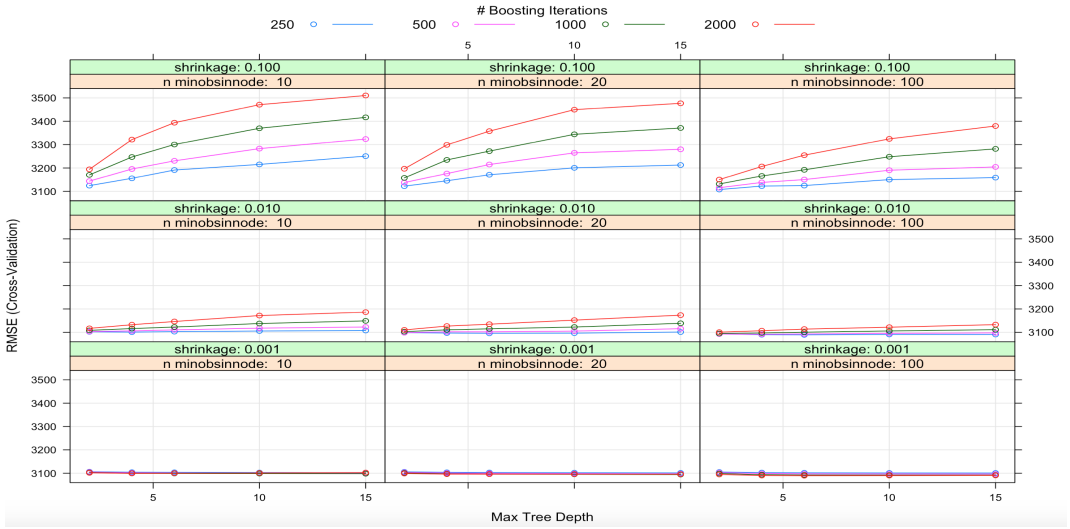


Figure 5.9: RMSE on the hold-out folds with the different tuning parameters of the GBM for claim severity.

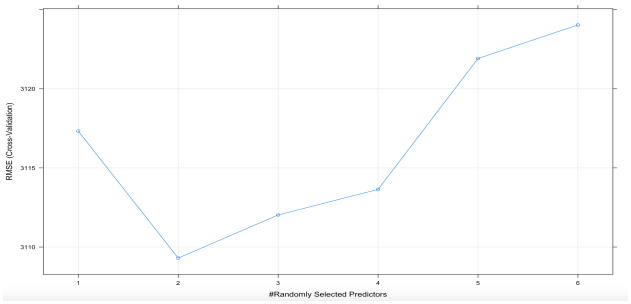


Figure 5.10: RMSE on the hold-out folds with the tuning parameter of the RF for claim severity.

3199.2. Both have a predictive performance worse than simply projecting the average claim severity of the train set on the hold-out set, see table 5.2. We therefore either trying another more heavy-tailed distribution than the Gamma distribution or simply improving the model at hand. We continue with the latter.
There are different approaches on how to improve the GLMs. The RF model has a better performance for both the number of claims and claim severity models than the GLM, so we can learn from the insights gained from the RF to improve the GLMs. The RF allows (i) non-linearities and (ii) interactions. These seem like areas where the GLMs can be improved.
The marginals of the RF for the number of claims response suggest (i) *Policyholder Age*, *Experience Years*, *Vehicle Weight* and others (see section 5.3.1.4) can be introduced as non-linear predictors. A similar set of covariates exhibit these non-linear effects for claim severity, see section 5.3.2.4. We will implement these in GLM and GAM. GAM allows the use of functions of predictors, which are most commonly fit as cubic regression splines.
Recognising (ii) the interaction effects is a bit problematic however. The RFs allow interpretation through $3D$ partial dependence plots, where the effect of two covariates on the response is shown. If these $3D$ plots deviate substantially from the seperate marginals, then interaction among the covariates occurs. This is hard to derive from the $3D$, see figure A.6. We therefore resort to a model that is known to be good at this, namely MARS. In the following sections we will first use MARS to detect possible interaction effects. This is followed by the implementation of the non-linear effects in GLMs and GAMs.

5.4 Interaction detection

We concluded that although the RFs provided meaningful insights into the (non-linear) effects of a single predictor on the response via the 2D partial dependence plots. The interaction effects, which are 3D effects

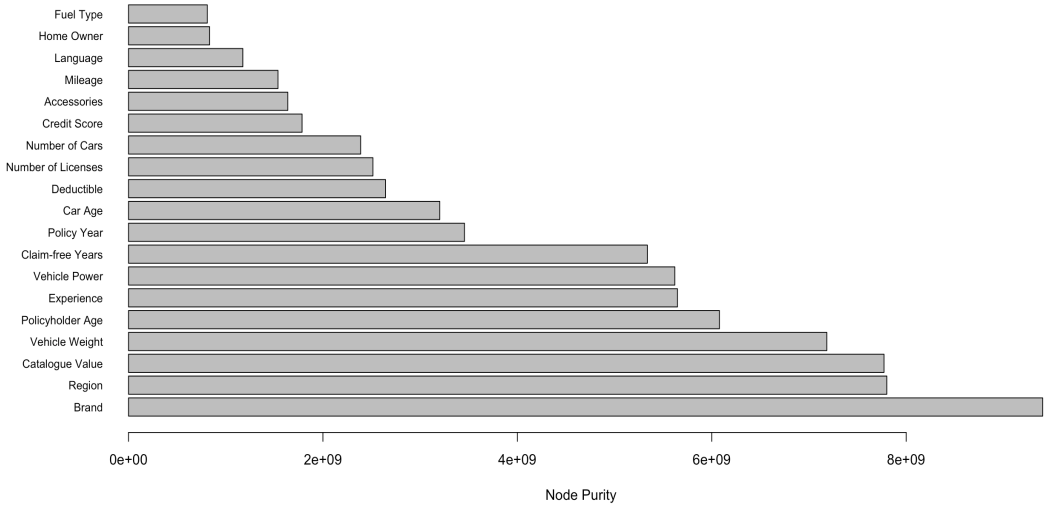


Figure 5.11: Variable importance according to the Random Forest claim severity model on a subset of the data.

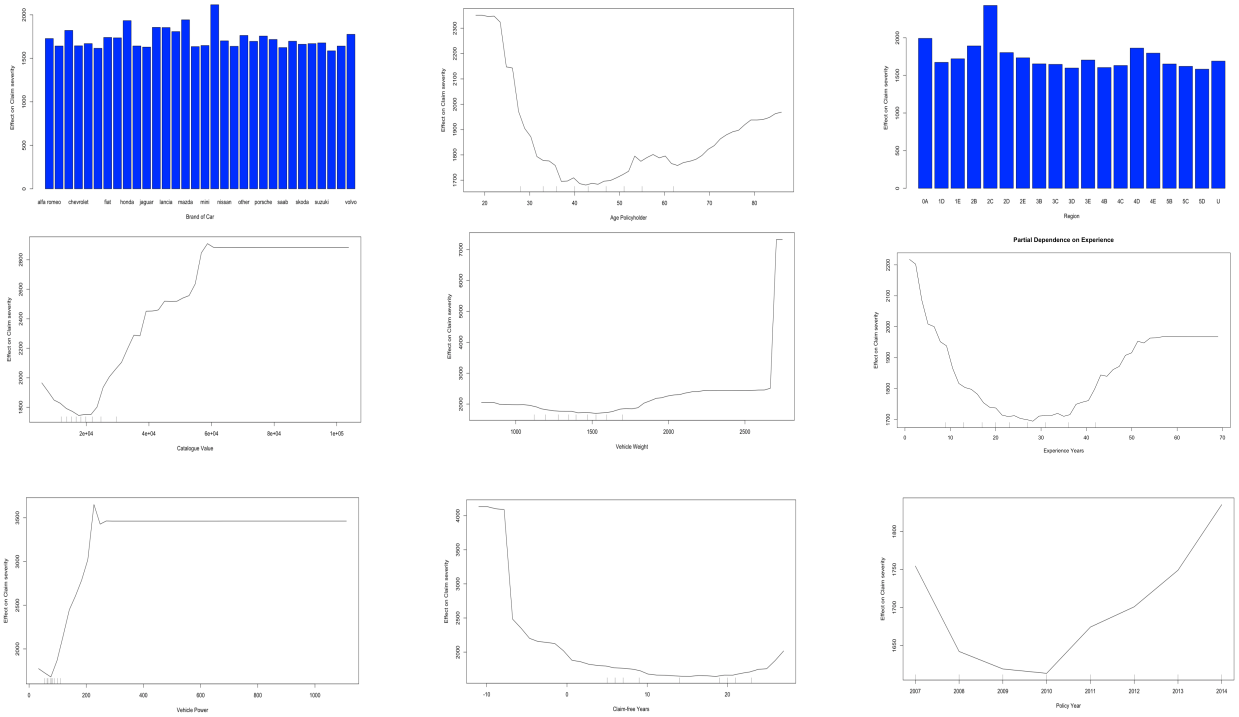


Figure 5.12: Marginal effects of the different covariates on the claim severity. The nine most important variables of the Random Forest model are illustrated. The excluded plots can be found in the appendix (figure A.5).

since we investigate the effects of two predictors on the response, were too difficult to interpret (see figure A.6). We therefore need an alternative method to investigate this. A method that automatically searches for these interaction effects and is widely used in practice are Multivariate Adaptive Regression Splines (MARS).

5.4.1 Multivariate Adaptive Regression Splines

This method is again applied for both the number of claims and the claim severity process on the $20\%/10\%$ train-to-test set split. The standard forward building and backward pruning algorithm is used (see section

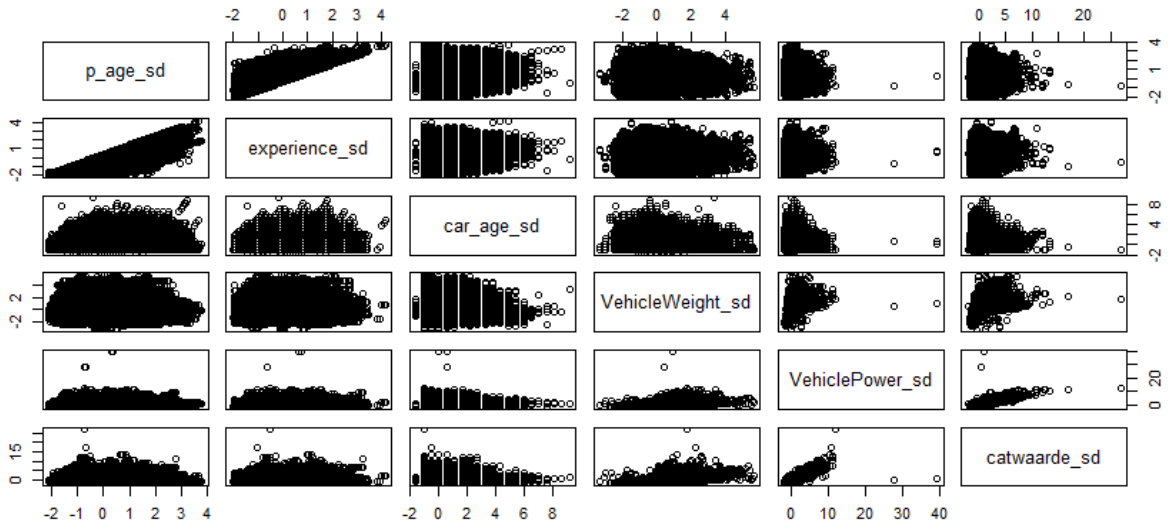


Figure 5.13: Pairs plots for a selection of explanatory variables.

3.5.5). The cumulative exposure for each (product) of hinge function(s) is plotted as well. This can show if a fitted coefficient is sensitive to outliers and thus overfitting, because a coefficient of MARS is attached to a single or a product of *hinge* functions. The general notation of a hinge function is $h[\pm(x_i - c)] = [\pm(x_i - c)]_+ = max\{0, \pm(x_i - c)\}$, where x_i is an explanatory variable and c is the knot. So in case of a single hinge function the coefficient only applies for values greater (or less) than the knot. The size of the set of values for which the coefficient is applicable decreases when more hinge function are combined in product. For more information on MARS, see section 3.5.5 or paper [16].
The coefficient fitting process uses a combination of the known MARS method from [16] and uses either the GLM Gamma or Poisson fitting process in the cases of claim severity and the number of claims respectively. The specifics of this coefficient fitting process are described in paper [24]. This method is used so that f.e. in case of the claim severity only strictly positive values are fitted.

5.4.1.1 Claim severity

MARS is fitted for the claim severity data and the results are presented visually in figure 5.14. The fitted coefficients can be found in the appendix (A.3). The RMSE on the hold-out 10% set is 3178.9, which is a worse performance than all other models (see table 5.2). The number of coefficients fitted is 15, which is considerably less than the 39 used by the GLM severity model. Some products of hinge functions have a very low cumulative exposure, so the model could be quite sensitive to outliers for these products. In all but one of these cases the fitted coefficients are not of a large absolute size (< 2.5) so the overfitting does not seem to be an issue here. But the last fitted interaction coefficient between *Reference Year* (ref_year) and *Vehicle Weight* (VehicleWeight_log) ≥ 7.844 is however an extreme case. The products of hinge functions that do not have an extreme coefficients fitted but do have a very low exposure are the following (from left to right): *Brand* is Mazda and *Credit score* (Creditsc) ≥ 3, *Brand* is Mitsubishi and *Language* is French, *Brand* is Renault and *Credit score* ≥ 3, and lastly *Region* is $2C$ and *Language* is French. We deem these hinge function combinations (interactions) less valid because of the very low exposure.
The combinations that do apply for larger sets of data are (from left to right): *Brand* is BMW and *Language* is French, *Car Age* (Car_Age_sd) ≤ 1.42 and *Language* is French, *Claim-free Years* (ccfy_sd) ≤ 1.43 and *Language* is French, *Reference Year* ≤ 2 and *Vehicle Weight* ≤ 7.84, *Number of Cars* (Ncars) ≥ 2 and *Language* is French, and finally *Vehicle Power* (VehiclePower_log) ≥ 4.36 and *Vehicle Weight* ≤ 7.84.
Some combinations of explanatory variables appear in the products of hinge functions that have a very low exposure and of a higher exposure. The combinations *Brand* $\times$ *Language*, *Reference Year* and *Vehicle weight* are examples of these.

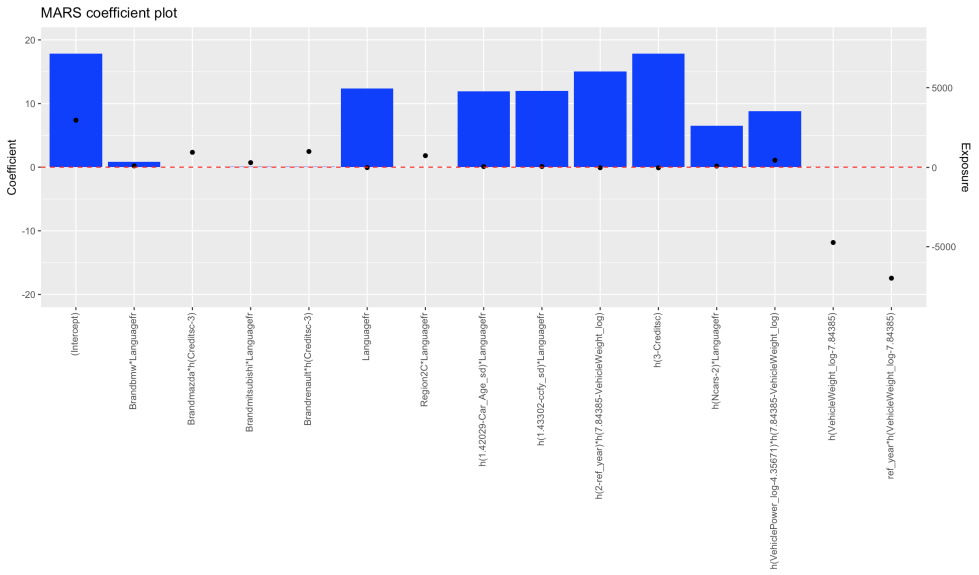


Figure 5.14: Coefficient plot for MARS for claim severity with exposures. The blue bars (right y-axis) are the cumulative exposures and the dots are the fitted coefficients (left y-axis).

Furthermore it should be said for clarification that the subscript log is used to indicate that this variable has been log-transformed and the subscript sd to indicate that these variables are centered and standardised.

5.4.1.2 The number of claims

We do the same for the number of claims response and this time we notice no products of hinge functions in figure 5.15. The fitted coefficients are stated in the appendix (A.4). No new information is revealed about the potential interactions at work using the MARS model for the number of claims. The RMSE on the hold-out 10% set is 0.4444, which is a better performance than the mean value and the basic GLM (see table 5.1).

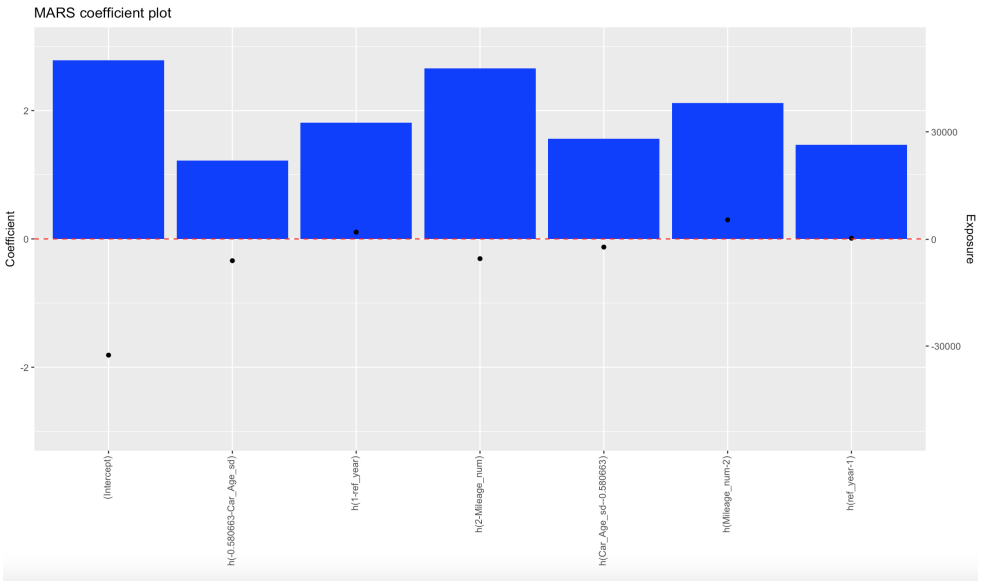


Figure 5.15: Coefficient plot for MARS for the number of claims response with exposures. The blue bars (right y-axis) are the cumulative exposures and the dots are the fitted coefficients (left y-axis).

5.4.1.3 GLM application

The only applicable interactions extracted from the MARS model were *Brand* × *Language*, *Car Age* × *Language*, *Claim-free Years* × *Language*, *Reference Year* × *Vehicle Weight*, *Number of Cars* × *Language*, and *Vehicle Power* × *Vehicle Weight* for claim severity. The MARS model did not find any interactions in case of the number of claims response.
Each interaction term is added to the model separately in a 'forward' manner and checked for the new AIC value. Only terms that improve the AIC are included. The interactions that are included because of forward AIC improvement are all interaction terms except *Brand* × *Language* and *Vehicle Weight* × *Vehicle Power*. The AIC is eventually slightly improved from $177,677$ to $177,647$. The performance on the hold-out set also barely improves from RMSE $3,147.5$ to $3,147.1$.

5.4.2 Conclusion

The MARS model performs worse on the subset than the GLM for claim severity, but performs better for the number of claims response however. The MARS model does not find any interactions in case of the number of claims response. In case of the claim severity we deem six interactions as worth pursuing, see section 5.4.1.3. The other cases have a very low exposure and are thus less credible. The interactions are applied and the improvements are minimal, this makes us question the improvements possible through adding interaction effects.

5.5 Non-linearities implementation

Random Forests performed the best on a hold-out set, this could indicate that the basic GLM model is not able to capture non-linearities that are at work in the data. These Random Forests provided meaningful insights into the non-linear relationships of the explanatory variables with respect to the response. Here we recapture these relationships through a GLM with added quadratic terms and a fitted Generalized Additive Model (GAM) for both the number of claims and claim severity responses. The GAM is defined by the following relationship

$$g(E[Y_i]) = \beta_0 + \sum_{j=1}^{r_l} \beta_j x_{ij} + \sum_{k=1}^{r_{nl}} f_k(x_{ik}), \tag{5.1}$$

where Y_i is distributed according to a distribution from the exponential family and $g(\cdot)$ is the link function as was the case for the GLMs. The used link function remains the log-transformation. The first summation term collects the number r_l of linear explanatory variables, while the second summation the number r_{nl} of non-linear explanatory variables. The explanatory variables are chosen to be included in the non-linear implementation by evaluation of the partial dependence plots of the Random Forest model, see figures 5.12 and 5.6. The f_j's are assumed to be relatively complicated functions and are therefore most commonly modeled using cubic regression splines. Cubic regression splines will be used to fit the GAMs in the following subsections. The explanatory variables that are chosen to be implemented in the non-linear implementation are *Policyholder Age*, *Experience*, *Car Age*, *Vehicle Power*, *Vehicle Weight* and *Catalogue Value*. The last two explanatory variables are not used in modelling the number of claims, since they are not included in the basic Poisson GLM.

5.5.1 Claim severity

The GAM is first fitted for the claim severity, this is followed by introducing explanatory variables that showed quadratic effects according the the RF model to the basic GLM.
The functions $f_k(x_{ik})$ (equation 5.1) fitted by cubic regression splines are plotted in figure 5.17. The pronounced non-linear effects of some variables on the response in the RF model are not easily recognisable in these figures. The *Policyholder Age* and *Experience* variables for example showed strong quadratic effects on the response, while the GAM functionals of these show either a slightly oscillating linear effect upwards

or a linear effect downwards respectively. These effects are thus quite conflicting with the effects found in the RF model. But if you compare the trends of the functionals to the fitted coefficients of the basic glm model (see table A.2 and figure A.3), you'll find that all but the *Catalogue value* and *Vehicle Weight* follow the same overall trend. This is because of the effect that the outliers of these explanatory variables have on their functionals. If you focus only on the more densely populated values of these variables, then the trends are still the same. The performance in RMSE of this GAM is 3146.1, which is slightly better than the GLM (see table 5.2).
Now the explanatory variables which showed strong non-linear effects according to the RF model are introduced to the GLM by continually adding higher order terms. These variables are the following: *Policyholder Age*, *Experience*, *Car Age* and *Reference Year*. For each variable the higher order terms are iteratively included until the effect of the RF is emulated and the AIC of the model has improved. This is only achieved for the *Policyholder Age*, by adding higher terms until the 4^{th} order. The other variables did both not emulate the effect found in the RF model and did not improve the AIC. The new GLM model improved the AIC slightly from $177,677.1$ to $177,673.3$, while the predictive performance on the hold-out set improved from RMSE 3147.5 to 3144.4. The fitted coefficients of the new model are illustrated in table A.5 of the appendix. The new effect because of the higher order polynomial is illustrated in figure 5.16, this is more similar to the RF model effect (see figure 5.12) than the linearly increasing coefficients with increasing values of *Policyholder Age* of the original GLM.

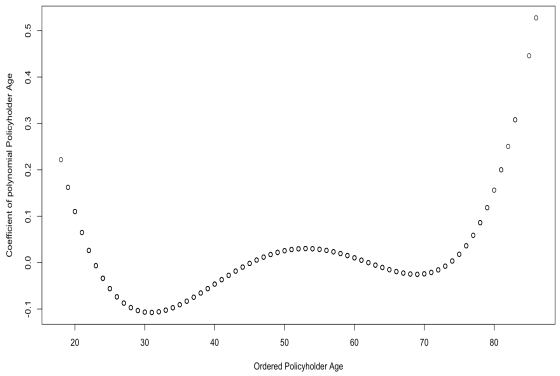


Figure 5.16: The estimated coefficients of the combined polynomials of the Policyholder Age covariate for the ordered Policyholder Ages of the severity train set.

5.5.2 The number of claims

The same procedure is applied here as conducted for the claim severity. If you consider figure 5.19, it is yet again apparent that the fitted functions $f_k(x_{ik})$ do not resemble the partial dependence plots (see figure 5.6) of the RF model for the number of claims. These cubic regression splines seem mostly linear. These effects do follow the same trends as the GLMs as was the case for the claim severity. The only function that deviates a bit in the outliers is the one for *Policyholder Age*. But the trend still holds in the more densely populated values of the variables. The performance (RMSE) of this fitted GAM is 0.4817, which is worse than all other models.
The explanatory variables which showed strong non-linear effects according to the RF model are introduced to the GLM by continually adding higher order terms. So the same procedure is followed as previously conducted for the claim severity. The chosen variables are the following: *Policyholder Age*, *Experience*, *Car Age*, *Catalogue Value* and *Vehicle Weight*. For each variable the higher order terms are iteratively included until the effect of the RF is emulated and the AIC of the model has improved. Two variables improved the AIC, while leaving the RMSE unchanged on the hold-out set. This is achieved by including *Policyholder Age* from order one until order 4. The *Car Age* variable was included until the third order. With this new model the AIC improves from $75,703.73$ to 75688.95, while the RMSE remains unchanged. The fitted coefficients of the new model are illustrated in table A.6 of the appendix. The new effects because of the higher order polynomials are illustrated in figure 5.18. The effects are not similar to the corresponding

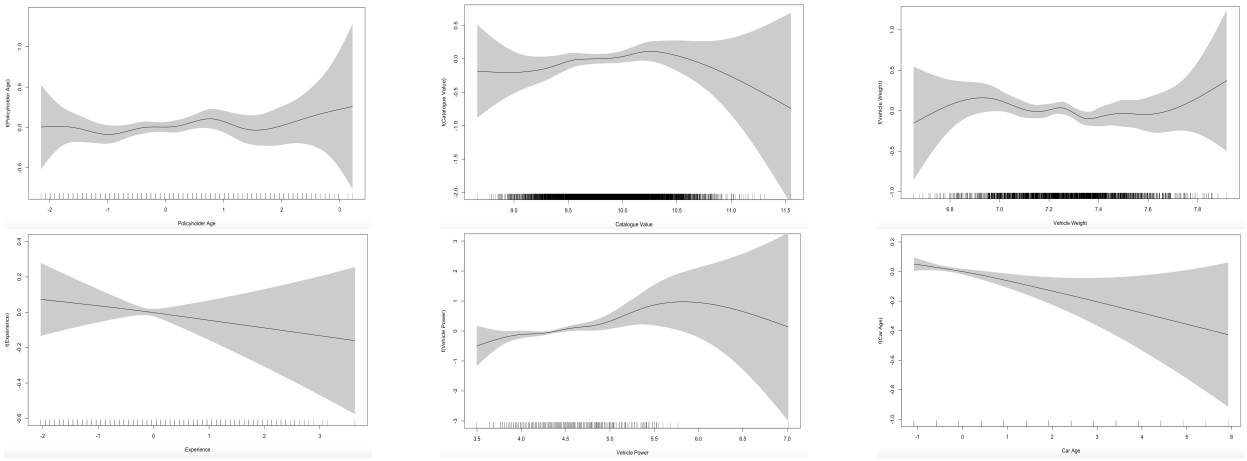


Figure 5.17: GAM cubic regression spline plots for claim severity. Describes $f_k(x_{ik})$ and thus its effect on the response s_i.

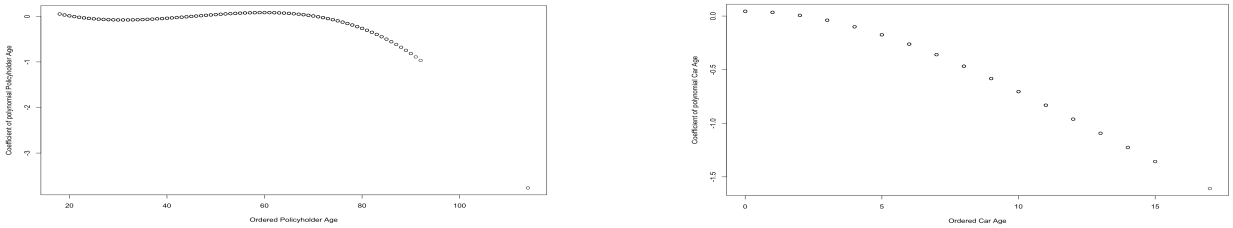


Figure 5.18: The estimated coefficients of the combined polynomials of the the covariates for the ordered covariates of the number of claims train set.

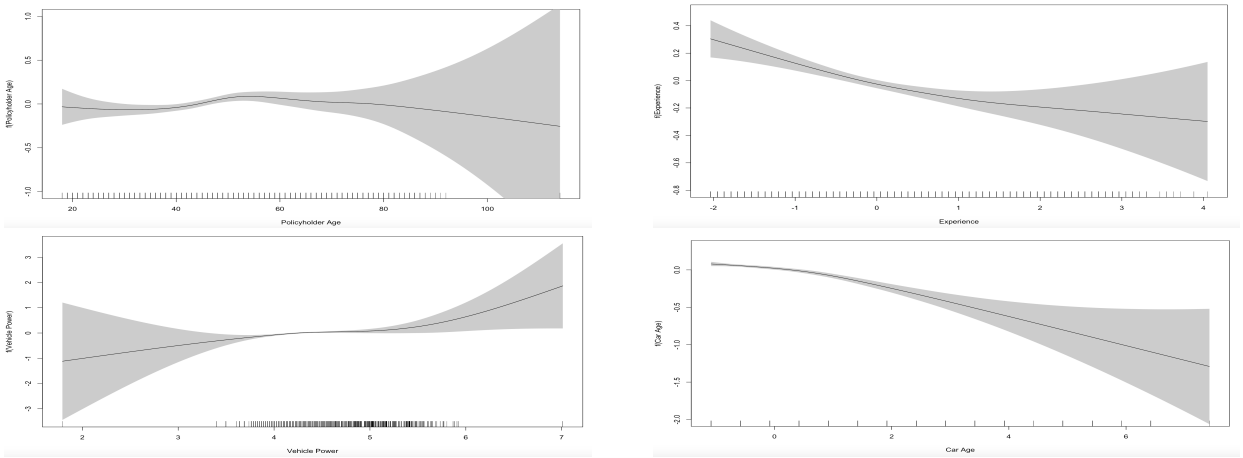
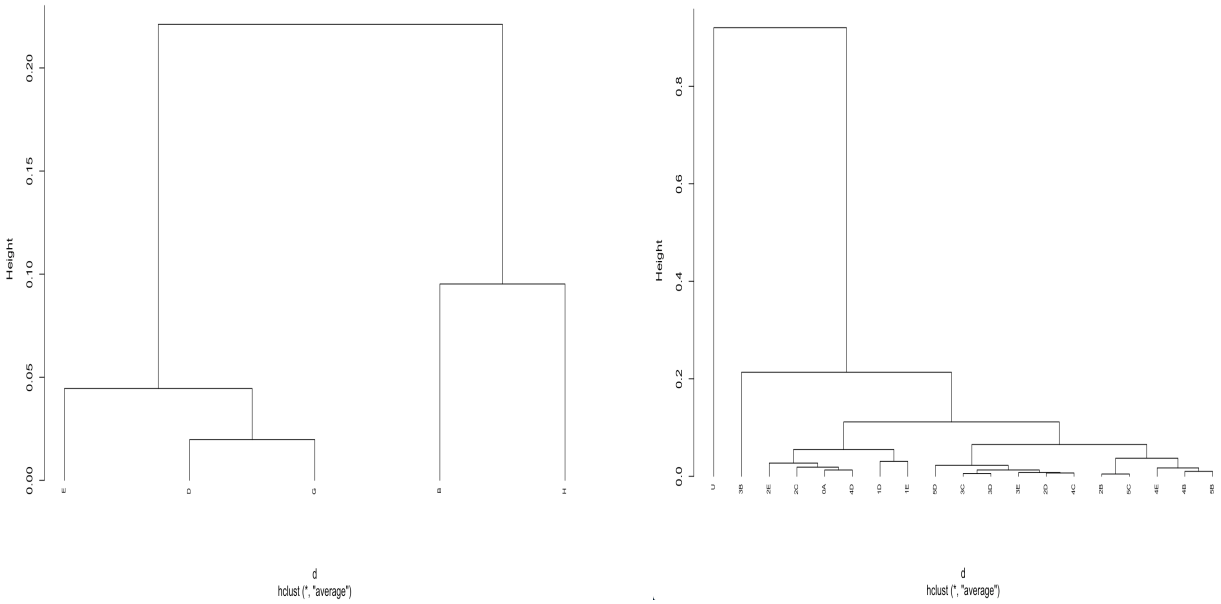


Figure 5.19: GAM cubic regression spline plots for claim severity. Describes $f_k(x_{ik})$ and thus its effect on the response c_i.

partial dependence plots in figure 5.6.

5.5.3 Conclusion

The introduction of non-linearities by using either a GAM or augmenting the GLM does not have significant results in predictive performance for either the claim severity or the number of claims, although the AIC did improve somewhat. These models still do not perform well in comparison to the RF model.
In the next section another issue concerning the fitted GLM is targeted. This issue was raised by the high



(a) Fuel types. (b) Regions.

Figure 5.20: The results of Hierarchical Clustering of the single kernels achieved by the K-means Clustering based on the number of claims and the log-transformed claim sizes in the train set.

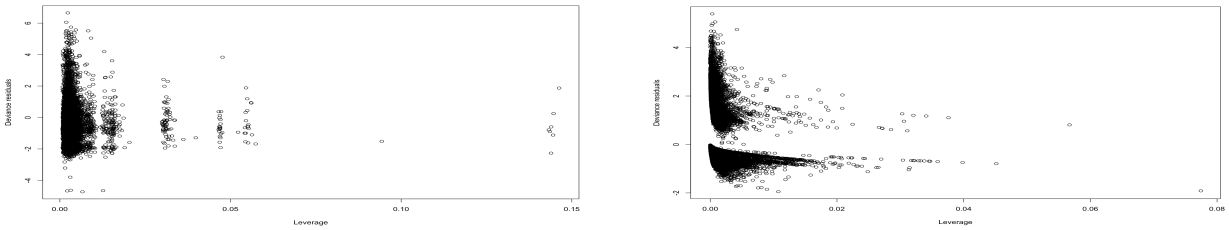
leverage for both the number of claims and claim severity GLMs in the residual analyses.

5.6 GLM variables with low exposures

There are some categorical variables that are used in the basic GLM approach that due to a very low exposure lead to an extremely (small or large) coefficient. Some examples are the instances U (Unknown) of variable *Region* and *Fuel Type* E (Electric) for the number of claims; as well as *Region* $1D$ for the claim severity.
There are a number of different ways to handle this issue, one that seems promising is using Hierarchical Clustering based on the single kernels determined by the K-Means algorithm to see which instances of categorical variables could be joined. Only information from the train set is used here. The analyses use the number of claims and the log-transformed claim sizes.
The results of the analyses for both explanatory variables are presented in figure 5.20. *Fuel type* E is closest to types D and G. The policies of type E are therefore randomly divided between these other two types.
Region $1D$ is most similar to those of *Region* $1E$, so these are combined into a new *Region* $1DE$. The policies with an unknown *Region* are not similar to any of the others, so these are divided randomly among the others.
The GLMs fitted with these new categorical variables do not improve on the previous models in terms of AIC or predictive performance (RMSE), but they do however improve in robustness. The RMSE is unchanged and the AIC worsens from $177,677.1$ to $177,678$ for claim severity and from $75,703.7$ to $75,710.1$ for the number of claims. The robustness improves, this is visible by comparing figure 5.21 with 5.2c and 5.7b. The maximum leverages for both the number of claims and claim severity have dropped considerably. The maximum leverage for the number of claims dropped from 0.3 to 0.08 and for claim severity it dropped from 1 to 0.15.

5.7 Combination or direct models: Premium calculation

Separate analyses have been conducted for the claim severity and the number of claims responses and now these analyses are combined to provide the risk premiums. This is easily achieved by multiplying the expected value of the number of claims and the conditional expectation of the claim severity severity for



(a) claim severity.

(b) Number of claims.

Figure 5.21: The updated Leverages for claim severity and the number of claims.

each policy i, see equation 3.1. This separation of expectations makes it possible to combine the different type of models and create hybrid models, since each model estimates one term of equation 3.1. To challenge the idea of separately modelling the number of claims and severity processes, a direct-to-premium RF model is implemented. First the direct-to-premium RF is tuned on the sole parameter, afterwards we will compare it to the other combinations and even hybrid models.

5.7.1 Fitting the direct-to-premium RF

The direct-to-premium RF is trained on the subset of data using 10-fold cross-validation and the best performance in rmse is achieved with the number of covariates to randomly use at splits $m = 2$, see figure 5.22. The variable importance is computed and can be found in the appendix, see figure A.7. The marginal effects of the ten most important variables are displayed in the appendix as well, see figure A.8. The marginal effects of some variables resemble the non-linear effects of the RF for the number of claims model better, while others that of the RF claim severity model. So some variables are more important for the number of claims, while others more for the claim severity. Examples are *Catalogue Value*, *Vehicle Power* and *Vehicle Weight* for claim severity, while *Claim-free Years* and *Policy Year* for the number of claims.

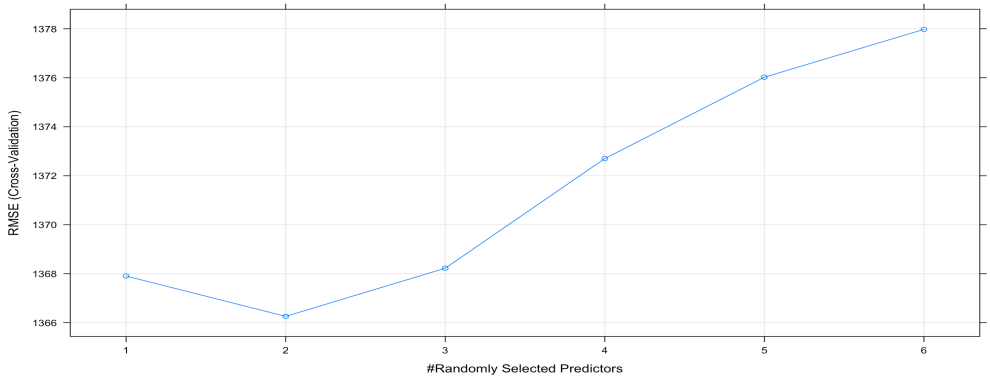


Figure 5.22: Training of a direct-to-premium Random Forest model.

5.7.2 Model comparisons

Now the models are compared on the calculation of the risk premiums. This is done for the GLM and RF models, since it is the current methodology with the best performing machine learning algorithm. The used benchmark is the yet again the mean model. The duration is again an important factor here, so the duration weighted mean of claim sizes is used, this is done in the same manner as done for the number of claims. So $\bar{\rho} = (\sum_i^{\text{train}} e_i \rho_i)/(\sum_i^{\text{train}} e_i)$, then $\rho_i^{\text{test}} = \bar{\rho} * e_i^{\text{test}}$ for each policy in the test set. The direct-to-premium RF performs best, just narrowly better than using the RFs separately for the number of claims and claim severity responses, see table 5.3. These are followed by the hybrid models, where using the RF for the number of claims results in the biggest improvement in predictive performance. It is surprising however to see that the combined GLMs perform worse than the benchmark, since the GLMs performed better than the benchmarks for both the number of claims and claim severity.

	Mean	$\text{GLM}_c \times \text{GLM}_s$	$\text{GLM}_c \times \text{RF}_s$	$\text{RF}_c \times \text{GLM}_s$	$\text{RF}_c \times \text{RF}_s$	$\text{RF}_{\text{direct}}$
RMSE	$1,410.7$	$1,412.2$	$1,405.0$	$1,376.7$	$1,335.0$	$1,323.8$

Table 5.3: Predictive performance for the different model combinations applied to the risk premium.

5.8 Conclusion

In this chapter we started by applying the basic GLM and Machine Learning techniques on the separate number of claims and claim severity processes on a subset of the whole dataset (20% train - 10% test) since the dataset is very large. The RF improved on the predictive performance (RMSE) of the basic GLM by approximately 6% and 9% for the claim severity and the number of claims respectively.
The GLM showed some issues in modelling both the number of claims and claim severity. In case of the Poisson GLM, there was a downward trend visible in the crunched deviance residuals (see figure 5.2b), indicating that the model does not do a very good job at fitting the data. The residual analysis of the GLM fitted for the claim severity process showed that the Gamma distribution doesn't do the best job at capturing the heavy tail of the claim severity. The Lognormal distribution and a gamma distribution on a capped dataset were fitted. Both improved the deviance residuals but the predictive performance of the Lognormal claim severity was a lot worse than the benchmark *mean model*, while the capped claim severity requires additional analysis for the withheld data. There were furthermore some pronounced line patterns found in the residual analysis. These seemed to occur more for claim type C (Scratches). So it seems this is a fixed payout that occurs a lot for these types of damages.
The Machine Learning techniques showed some non-linear marginal effects of the different covariates on both claim severity and the number of claims. This caused us to investigate the possibilities of improving the GLMs sufficiently so that the performances are more comparable to the Machine Learning techniques. The most notable marginal effects by covariates were introduced in a new GLM and a GAM was fitted as well. Both did not compare well to the predictive performance of the Machine Learning techniques.
An additional strength of tree-based methods is the natural inclusion of interaction effects which are absent in the basic GLM. To discover the interactions there are two options: investigate the 3D marginal effects or apply a method that is good at extracting interaction effects. The former is very hard, so the latter was applied in the form of MARS. This method on its own does not improve the predictive performance but it showed some possible interactions. The predictive performance of the GLM that includes these interactions does not improve on the basic GLM, neither does the AIC.
The residual analyses pointed out some covariates with very low exposures, by investigating the merging possibilities by sequential K-means and Hierarchical Clustering provided options to increase the robustness of the basic GLM. this did not affect the predictive performance.
The last step was applying these methods to estimate the premiums and evaluate the predictive performance here. Hybrid methods between GLM and RF improve on the basic separated GLMs. The best improvement was achieved in modelling the number of claims by the RF. Although the separate RF and direct RF perform best.

RMSE	Number of claims	Claim severity	Premium
Mean	0.4515	3159.1	$1,410.7$
GLM	0.4495	3147.5	-
GBM	0.4364	3144.1	-
RF	0.410	2964.9	-
MARS	0.4444	3178.9	-
GLM interaction	-	3147.1	-
GLM non-linear	0.4495	3144.4	-
GAM	0.4817	3146.1	-
GLM robust	0.4495	3147.5	
$GLM_c \times GLM_s$	-	-	$1,412.2$
$GLM_c \times RF_s$	-	-	$1,405.0$
$RF_c \times GLM_s$	-	-	$1,376.7$
$RF_c \times RF_s$	-	-	$1,335.7$
RF_{direct}	-	-	$1,323.8$

Table 5.4: Summary of the predictive performances of the various models for the number of claims, claim severity and premium. The best performances are coloured green.

Chapter 6

Uncertainty quantification

Contents

Up until now we have provided techniques that are the most interesting when we only consider predictive performance and thus place priority on the predictive part. But another important component in this setting is measuring the uncertainty surrounding our predictions and used explanatory variables. This can be of interest to a single policy or the whole portfolio. Since the RF model performed best for accurate prediction, a new model named BART is included (section 6.1). BART aims to combine the predictive strength of RFs and provides uncertainty quantification through a MCMC backfitting algorithm.
This model is followed by Hierarchical modelling (section 6.2), which is a Bayesian extension of regression models. These models are interesting since they provide uncertainty quantification through the Bayesian method, they are also very flexible so it is easy to further increase the complexity of the model or use different distributions (section 6.3). The hierarchical structure can also be of interest when the model is applied to multiple insurance products for example or when enforcing a policyholder-policy structure.
Both models use the same $20\%/10\%$ train-to-test split used in the previous chapter, see section 5.3.
In this chapter we furthermore provide a metric (DIC) to compare HMs (section 6.4). Since we have quantified the uncertainty properly, we present ways to implement this into the pricing of premiums (section 6.5). We end this chapter with two sections that discuss the approach of an individual policy and for the whole (active) portfolio (sections 6.6 and 6.7).

6.1 Bayesian Additive Regression Trees

This method runs into some memory problems quite easily when applied in the R environment. The issue is that the memory used by R is RAM (Random Access Memory), which runs out pretty quickly on most computers (around $16GB$). BART requires a lot of memory for the MCMC backfitting algorithm and therefore fails on the subset of data chosen in section 5.3. We therefore set up a virtual machine of 50GB RAM with *Google Cloud Computing*. How to setup such a VM (Virtual Machine) is described in the appendix A.7.
BART is directly applied to the premium. It is not split into two components because of the uncertainty quantification. The predictive intervals would not make sense anymore if we would merely make a product of the two components. We would need to simulate this and we do not have access to these parts of the algorithm, so a custom-made algorithm would then have to be made. The good predictive performance of the direct-to-premium RF also contributes to this choice.
BART can be tuned with 5-fold cross-validation on the following tuning parameters: number of trees K, shrinkage parameter κ and the σ-shape parameters ν and q. The grid setup is chosen in accordance with the recommendations by the paper of Chipman et al. [11]. The selected grid setup is the following:

$$\text{BART:} \quad \begin{cases} K \in \{50, 200\} \\ \kappa \in \{1, 2, 3, 5\} \\ (\nu, q) \in \{(3, 0.9), (3, 0.99), (10, 0.75)\} \end{cases}$$

6.1.1 Premium model

BART is trained using 5-fold cross-validation on the subset of data mentioned (data used for the number of claims models) at the start of section 5.3. The best performance (RMSE) is achieved using: $K = 200$, $\kappa = 5$ and $(\nu, q) = (3, 0.99)$. The results of the cross-validation are summed up in figure A.9 of the appendix. We now refit BART with these parameters and use a number of burn-in iterations which is many orders of magnitude higher than proposed by Chipman et al. We use $10,000$ burn-in iterations and let it run for another $10,000$ iterations, while Chipman et al. propose 250 and 1000 iterations respectively.
Now the model is ready to use and first the predictive performance is checked in comparison with the previous models used for the risk premium at the end of previous chapter, see section 5.7. The RMSE on the hold-out set is 1391.8, which is better than the basic GLM. We continue with the uncertainty quantification and create 95% credible intervals for the hold-out set. Now we notice that merely 43% can be found within the 95% credible intervals when we create the credible intervals for the test set. So the model does not do a good job at all at quantifying the uncertainty.

6.2 Hierarchical modelling

The goal of this research is twofold: accurate prediction and meaningful insights. As up until now we achieved greater accuracy in prediction on a hold-out set by employing Random forests on the data. Although we were able to extract insights from the Random Forests such as variable importance and partial dependence plots (marginal effects), we are not able to express our beliefs about new data in probabilistic terms. This is why Hierarchical modelling (HM) is an interesting method. Because of its Bayesian design, it allows us to easily access credible intervals, posterior predictive (joint) distributions of the number of claims, claim severity and premiums of new data.
The hierarchical models are built using a combination of R and *JAGS*, where JAGS is an acronym for Just Another Gibbs Sampler. The data is handled using R, which feeds it to *JAGS* along with the *JAGS* model written in a text file.

6.2.1 Design

For every policy

$$
\begin{aligned}
& c_i \sim \text{Poisson}(e_i \lambda_i) \\
& \lambda_i = exp(\alpha^F + \sum_{k=1}^{K} x_{ik}^F \cdot \beta_k^F) \\
& \alpha^F \sim \mathcal{N}(\mu_\alpha^F, (\sigma_\alpha^F)^2) \\
\text{Each ind.} \quad & \beta_k^F \sim \mathcal{N}(0, 10^4) \qquad \text{k = 1 ... K} \\
& \mu_\alpha^F \sim \mathcal{N}(0, 10^4) \\
& \sigma_\alpha^F \sim \mathcal{U}(0, 100)
\end{aligned}
$$

When $c_i > 0$

$$
\begin{aligned}
& s_i | c_i > 0 \sim \text{Gamma}(\nu, b_i) \\
& b_i = \nu / \zeta_i \\
& \zeta_i = exp(\alpha^S + \sum_{k=1}^{K} x_{ik}^S \cdot \beta_k^S) \\
& \nu \sim \mathcal{U}(0, 10) \\
& \alpha^S \sim \mathcal{N}(\mu_\alpha^S, (\sigma_\alpha^S)^2) \\
\text{Each ind.} \quad & \beta_k^S \sim \mathcal{N}(0, 10^4) \qquad \text{k = 1 ... K} \\
& \mu_\alpha^S \sim \mathcal{N}(0, 10^4) \\
& \sigma_\alpha^S \sim \mathcal{U}(0, 100)
\end{aligned}
\tag{6.1}
$$

For this HM $5,000$ burn-in iterations are used, followed by another $5,000$ iterations. A thinning of 10 is used, so the final model keeps 500 iterations in the memory. This setup has two main reasons: (i.) memory and (ii.) extracting useful information. In the presence of autocorrelation, however, we may obtain for example 5000 samples from the posterior, but those samples contain less (possibly much less) information about the 2.5^{th} and 97.5^{th} percentiles than $5,000$ independent draws would. The lower the autocorrelation, the greater the amount of information contained in a given number of draws from the posterior; this is referred to as the efficiency or mixing of the chain. The thinning of the iterations reduces the autocorrelation. We expect a high amount of autocorrelation so that's the reason thinning is used in this setup.
Each policy i has a recorded number of claims c_i and a claim severity s_i. The risk premium is acquired as

$$\rho_i = e_i \lambda_i \zeta_i, \qquad \text{(see equation 3.1)}. \tag{6.2}$$

This is a result of the chosen parametrisation of the Poisson and Gamma distributions for the count and severity components respectively (see sections 3.3.2 and 3.3.3). So the risk free premium consists of a number of claims (F) and a claim severity (S) component, we therefore separate the predictors and coefficients of these components with the appropriate superscript since the effects differ. The number of claims component consists of an exposure e_i and λ_i, where

$$\lambda_i = exp(\alpha^F + \sum_{k=1}^{K} x_{ik}^F \cdot \beta_k^F). \tag{6.3}$$

The severity component is similarly defined as

$$\zeta_i = exp(\alpha^S + \sum_{k=1}^{K} x_{ik}^S \cdot \beta_k^S). \tag{6.4}$$

The parameter α serves as an intercept in both models. The covariates x_{ik} and belonging parameter β_k use K selected number of covariates. Initially we used the ten covariates that are the most important according to the Random Forest model and added mileage to this group as well because of expert judgment. But since the Brand and Region parameters were very hard to estimate (the procedure would not converge after 500k iterations), they were not considered. These are the same group of covariates for both the number of claims as the claim severity response. The total number of covariates is $K = 9$. For a full list of the used covariates see tables A.7.
The claim count is analysed for every component of the 20% subset and claim severity conditional on the number of claims being larger than 0.
The hierarchical model 6.1 can be interpreted graphically as well by a parameter design model (figure 6.1). This illustrates which parameters serve as either an input for a distribution (solid line) of a higher level parameter or a direct connection by either passing the exact value or an equation to the specified parameter (dashed line). The number of data iterations is not the same (see figure 6.1) since we have a s_i conditioned on $c_i > 0$.

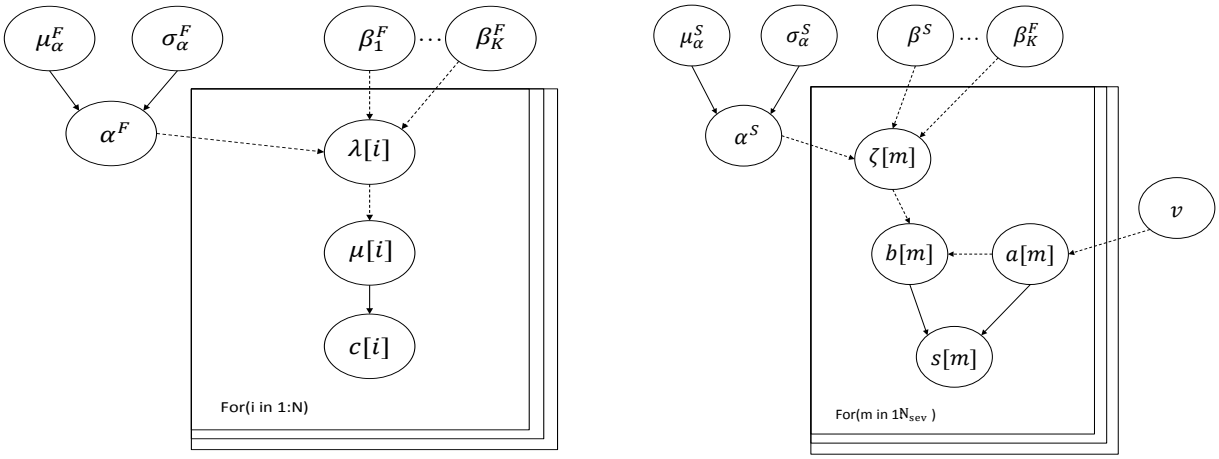


Figure 6.1: Graphical parameter design of the HM.

6.2.2 Validate on fake data

A quick check to show that a statistical method works as advertised, is to impose true values and check how the estimated parameters compare to the "true" values. This procedure goes as follows:

1. Specify reasonable *true* values for each parameter in the model.
2. Simulate a fake dataset of count - and severity data (c^{fake}, s^{fake}) using the model and the assumed *true* parameters.
3. Fit the model to the fake data and check if parameters are consistent with the *true* parameters.

To provide a good test we apply the exact model (we do not exclude any parameters) 6.1. We judge the model to be adequate if 95% of the parameters fall within the 95% credible intervals for each parameter. We randomly choose $5,000$ entries from the whole datset (before 2015). The true values of the parameters are stated in table 6.1.

i	1	2	3	4	5	6	7	8	9
β_i^F	0.1	0.07	0.07	0.1	-0.3	-0.3	0.1	0.2	-0.2
α^F	-2	-	-	-	-	-	-	-	-0.5
β_i^S	0.3	0.1	0.02	0.1	-0.1	0.01	-0.1	-0.1	0.3
α^S	6	-	-	-	-	-	-	-	-
ν	0.25	-	-	-	-	-	-	-	-

Table 6.1: *True* parameters choice to create fake count- and severity data.

We run the simplified model with merely 1000 burn-in iterations, followed by another 1000 iterations where we save the history of the selected parameters. This is usually a small amount of burn-in iterations, but we justify it by the fast convergence of the parameters. We then acquire the derived parameter values for the number of claims and claim severity component stated in tables 6.2 and 6.3. Of the 21 estimated parameters in the simplified model, 20 of the "true" parameters are within the 95% credible intervals of the estimated parameters. Only parameter β_6^S is not properly quantified, its 95% credible interval is too narrow by a small margin. So in other words, 95.2% of the derived parameters comply with the assigned *true* parameters. It should be noted that some parameters have a wide credible interval. Using more prior information helps to narrow these intervals.

6.2.3 Convergence diagnostics of the MCMC chain

Iterative simulation methods such as MCMC are tricky because after a certain number of iterations the distribution used to draw from lies between the starting and target distribution. The difficulty herein lies that a

	Mean	sd	2.5%	97.5%	True
α^S	6.155	0.384	5.400	6.897	6
β_1^S	0.543	0.268	0.077	1.062	0.3
β_2^S	-0.030	0.229	-0.475	0.418	0.1
β_3^S	0.081	0.153	-0.200	0.367	0.02
β_4^S	-0.185	0.277	-0.676	0.302	0.1
β_5^S	-0.072	0.104	-0.261	0.130	-0.1
β_6^S	0.563	0.283	0.015	1.064	0.01
β_7^S	-0.071	0.098	-0.255	0.131	-0.1
β_8^S	-0.117	0.204	-0.487	0.293	-0.1
β_9^S	0.258	0.053	0.147	0.356	0.3
ν	0.234	0.010	0.215	0.255	0.25

Table 6.2: Derived parameters for the fake claim severity response.

	Mean	sd	2.5%	97.5%	True
α^F	-1.832	0.159	-2.121	-1.541	-2
β_1^F	0.219	0.096	0.036	0.392	0.1
β_2^F	-0.096	0.112	-0.310	0.138	0.07
β_3^F	0.113	0.077	-0.053	0.256	0.07
β_4^F	-0.054	0.098	-0.228	0.144	0.1
β_5^F	-0.283	0.045	-0.371	-0.195	-0.3
β_6^F	-0.131	0.131	-0.392	0.121	-0.3
β_7^F	0.122	0.041	0.043	0.198	0.1
β_8^F	0.132	0.083	-0.019	0.283	0.2
β_9^F	-0.190	0.021	-0.231	-0.149	-0.2

Table 6.3: Derived parameters for the fake number of claims response.

random walk can remain within a certain region that is heavily influenced by the starting distribution. There is some disagreement among scientists on how to handle this issue.
Rubin and Gelman [18] advocate multiple starts (or chains) with dispersed starting values. They argue that one single long run cannot find all maxima and therefore multiple iterations are required. Rafferty and Lewis on the other hand [29] propose a single sufficiently long run. They argue that the main reason for doing a multi-start is not of significance to standard statistical models with a realistically large number of MCMC iterations. They do acknowledge the importance of the chosen starting values, since poorly chosen ones can lead to very slow convergence. They recommend to handle this issue by trial-and-error.
In this analysis we stick to the approach recommended by Rafferty and Lewis.
By a general rule-of-thumb the number of burn-out iterations is taken as double of the burn-in iterations [2]. The number of burn-in iterations is $5,000$. The analysis of the convergence of this MCMC chain requires a number of helpful tools. The most common tools are the (i.) trace, (ii.) density and (iii.) autocorrelation function plots of each parameter. The trace plots (i.) should appear stationary: relatively constant mean and variance. If the process does show some form of drift, then the parameter has not converged yet and the number of burn-in iterations should be increased. A chain that mixes well traverses its posterior space rapidly, and it can jump from one remote region of the posterior to another in relatively few steps. The trace plot will appear very dense since the jumps can be quite large per iteration. This indicates low autocorrelation. The density plots help with the evaluation of the trace plots.
The convergence diagnostics contain a lot of figures and are therefore placed in the appendix (figures A.11 and A.12). Both figures look quite proper. This also is apparent from their running means that converge to a point quite rapidly for most parameters. The autocorrelation plots indicate the efficiency of mixing the distributions and it decreases to zero quite rapidly for most parameters (see model A.7 for the proper covariate names).

6.2.4 Posterior predictive check

Since our model passes the *fake data test*, we move on to another test applied by Bayesian statisticians, namely the posterior predictive check. This test is performed on the train set. We run our hierarchical model (see equation 6.1) with $5,000$ burn-in iterations and let it go on for another $5,000$ iterations. We choose this amount of iterations so we have comfortably reached convergence of the parameters. Therefore we have $n.sims = 500$ draws from the posterior distribution of the parameters. For each of these $j = 1 \dots n.sims$ draws, we create a replicated claims c_j^{rep} set and a replicated severity s_j^{rep} set. Naturally the risk premium ρ_j^{rep} is the inner product of the two vectors. These replicated sets are $N \times 1$ vectors of length $N = 80,176$, since we use the explanatory variables matrix X of the 20% train set, mentioned in section 5.3, to create them. These $n.sims$ replicated vectors represent the posterior predictive distribution.

$$\begin{aligned} f(\rho_{\text{new}}|\boldsymbol{\rho}) &= \int f(\rho_{\text{new}}|\theta) f(\theta|\boldsymbol{\rho}) d\theta \\ &= \int f(c_{\text{new}}|\theta) f(s_{\text{new}}|\theta, c>0) f(\theta|\boldsymbol{c}) f(\theta|\boldsymbol{s}, c>0) d\theta \end{aligned} \tag{6.5}$$

The bold-faced $\boldsymbol{\rho}, \boldsymbol{c}$ and $\boldsymbol{s}$ are the sets of the respective response with the covariate matrix, as such $\{\rho, X\}$. So to summarise, we simulate equation 6.5 in the following manner

- For $j = 1, ..., n.sims$ sample
  1. θ_j^F from $f(\theta^F|\boldsymbol{c})$,
  2. c_j^* from $f(c_{\text{new}}|\theta_j^F)$,
  3. If $c_j^* > 0 : \theta_j^S$ from $f(\theta^S|\boldsymbol{s})$,
  4. If $c_j^* > 0 : s_j^*$ from $f(s_{\text{new}}|\theta_j^S)$, else $s_j^* = 0$,
  5. ρ_j^* = $c_j^* \cdot s_j^*$
- Then ρ_1^*, ... , $\rho_{n.sims}^*$ are samples from the posterior $f(\rho_{\text{new}}|\boldsymbol{\rho})$.

Once we have the posterior predictive distribution, we use some test statistics $T_i(\rho_l^{rep}, \theta_l), i \in \{1, ..., n.sims\}$ to check if different models are consistent with the data (train set). Test statistics show in what areas a model fails to recreate the data. The test statistics are shown in table 6.4. The minimum value is not interesting since it is highly likely the minimum value will be zero in almost every set. We compare these statistics of each l replicated response to the observed response. We can extract the posterior predictive p-value, which is the proportion $T_i(\rho_l^{rep}, \theta_l) \geq T_i(\rho, \theta_l)$, where ρ is the observed premium vector. The closer the predictive p-value is to 0.5, the better the fit is to the data. From the L replicated sets we also create a 95% credible set for each test statistic.

Test statistic	95% credible interval	observed value	p-value
% of zero claims	$[0.857\ ,\ 0.868]]$	0.867	0.11
% of one claim	$[0.110\ ,\ 0.128]$	0.112	0.91
Mean value severity (s)	$[1\,593\ ,\ 1\,812]$	$1\,633$	0.706
Mean value premium (ρ)	$[237\ ,\ 270]$	243	0.84
Maximum value premium (ρ)	$[2.63 \cdot 10^4\ ,\ 2.95 \cdot 10^5]$	$6.24 \cdot 10^4$	0.24

Table 6.4: The posterior predictive check for the HM. Each test statistic accompanied with a 95% credible interval in the predictive distribution, a true observed value in the train set and the predictive p-value.

If you evaluate the predictive p-values of table 6.4, you notice there is definitely room for improvement. It seems that the Poisson process fitted for the number of claims has some difficulties. It underestimates the number of zeros present in the data, while it overestimates the occurrence of one or more (than 4) claims. An improvement that might tackle this is introducing a process that separates the two by introducing a Bernoulli variable (u_i) in the number of claims model, so that the number of zeros is controlled in the following manner:

$$\begin{aligned} c_i &\sim \text{Poisson}(u_i e_i \lambda_i), \\ u_i &\sim \text{Bernoulli}(p). \end{aligned}$$

Another visible issue is the overestimates of the size of the claims, for both claim severity and the premiums. The claim severity is actually close to 0.5, so we leave this model untouched in the improvement step. The premiums are consistently higher estimated than observed follows from too high estimated claim severity and number of claims. Although the number claims suffer more from this effect. The maximum value of the premiums is also a bit off, but we think the biggest room for improvement can be found in modelling the

number of claims differently.
The issues we just discussed were also visible in the residual analyses of the GLMs of the previous chapter, see sections 5.3.2.2 and 5.3.1.2. The conclusion of the previous chapter also showed that the number of claims model (predictive performance) was best improved by an alternative such as the RF model.

6.2.5 Evaluation uncertainty quantification

The BART model did not prove to be effective at quantifying the uncertainty involved in modelling the premium. This led to a coverage of only 42.88% of the observed claim sizes in the test set by the 95% credible interval created by the BART model. This is repeated for the HM model and the results are more in line with the expectation.
When the HM model is applied on the data from the test set, $95,60\%$ of the claim sizes is within its 95% credible interval. So although the model does not do the greatest job in some areas pointed out in the posterior predictive check, it has a pretty accurate coverage of the 95% credible interval.

6.2.6 Model improvement possibilities

There are three ways to (possibly) improve the existing model (see equation 6.1), namely:

i. Complexity of predictors:

- Adding non-linear predictors based on the insights gained from the partial dependence plots of the best performing model (RF).
- Introducing interactions between predictors based on insights from MARS and RF.

ii. Alternative distributions:

- Number of claims response:
  - Negative binomial (NB) is a possibility but not a viable option. Residual plots of a NB GLM in a Frequentist setting on the count data, does not visibly improve the fit.
  - Opting for a model that either increases the amount of zero claims or separates the $N = 0$ and $N > 0$ process. These are Zero-Inflation and Hurdle models respectively.
- Claim severity response:
  - Introducing a heavier-tailed distribution than the gamma distribution, such as the Lognormal, Pareto or Weibull distribution.

iii. Hierarchical structure:

- Policyholder hierarchy: Use a hierarchical structure where each policyholders i has a number of policies t_i. Therefore a regression is possible on the level of policyholders and the level of the policies.
- Insurance product hierarchy: Use a hierarchical structure where each insurance product is the upper level and the lower levels are the data contained by each insurance product. A third (middle) level could then be the policyholders i as well.

It does not seem worth pursuing the increase of complexity in the predictors (i.), since the findings in previous chapter showed that there was not much to be gained by doing so. Approach (iii.) seems worth pursuing, but there is not a lot to be gained since the characteristics are all policy specific and not policyholder specific. So the only predictor that can be used for the policyholders are random effects. That leaves us with the seemingly most impactful option, namely changing the distributions (iii.). This seems especially convincing since the main fallacies revealed by the posterior predictive check point in this direction.
The most attractive options for (ii.) are introducing hurdle models in case of the number of claims, the claim severity we leave untouched.

6.3 Alternative model

For every policy			When $c_i > 0$		
	$c_i \sim \text{Poisson}(u_i e_i \lambda_i)$			$s_i \mid c_i > 0 \sim \text{Gamma}(\nu, b_i)$	
	$u_i \sim \text{Bernoulli}(p)$			$b_i = \nu / \zeta_i$	
	$\lambda_i = exp(\alpha^F + \sum_{k=1}^{K} x_{ik}^F \cdot \beta_k^F)$			$\zeta_i = exp(\alpha^S + \sum_{k=1}^{K} x_{ik}^S \cdot \beta_k^S)$	
	$p \sim \mathcal{U}(0,1)$			$\nu \sim \mathcal{U}(0,10)$	
	$\alpha^F \sim \mathcal{N}(\mu_\alpha^F, (\sigma_\alpha^F)^2)$			$\alpha^S \sim \mathcal{N}(\mu_\alpha^S, (\sigma_\alpha^S)^2)$	
Each ind.	$\beta_k^F \sim \mathcal{N}(0, 10^4)$	k = 1 ... K	Each ind.	$\beta_k^S \sim \mathcal{N}(0, 10^4)$	k = 1 ... K
	$\mu_\alpha^F \sim \mathcal{N}(0, 10^4)$			$\mu_\alpha^S \sim \mathcal{N}(0, 10^4)$	
	$\sigma_\alpha^F \sim \mathcal{U}(0, 100)$			$\sigma_\alpha^S \sim \mathcal{U}(0, 100)$	

(6.6)

With this model we attempt to fix the main issues described in the posterior predictive check by adapting the number of claims distribution. The issues are addressed in the manner described at the end of the previous section. Model 6.3 is implemented. A summary of the fitted parameters can be found in the appendix (figures A.9 and A.15). The convergence diagnostics are in the appendix as well, see figures A.13 and A.14. We will now continue with another posterior predictive check.

6.3.1 Posterior predictive check

The test statistics that check the severity component are excluded, since the alternative model does not use a different approach to model the claim severity. This model still definitely has some flaws. The posterior

Test statistic	95% credible interval	observed value	p-value
% of zero claims	[0.859 , 0.870]]	0.867	0.18
% of one claim	[0.114 , 0.139]	0.112	0.97
Mean value premium (ρ)	[244 , 293]	243	0.97
Maximum value premium (ρ)	$[2.94 \cdot 10^4 , 3.45 \cdot 10^5]$	$6.24 \cdot 10^4$	0.29

Table 6.5: The posterior predictive check for the alternative model. Each test statistic accompanied with a 95% credible interval in the predictive distribution, a true observed value in the train set and the predictive p-value.

predictive check is however not a manner to compare models. We therefore use a criterion in the next section.

6.4 Comparing models: Deviance

The Frequentist approach uses the AIC (see chapter 5), while the Bayesian approach uses a similar criterion, namely the Deviance Information Criterion (DIC). The DIC combines model fit and model complexity, as does the AIC. The model fit can be summarized with deviance, which is defined as -2 times the log-likelihood [17], such as

$$D(y, \theta) = -2 \log[f(y|\theta)]. \tag{6.7}$$

The model complexity is measured by p_D, also the effective number of parameters of a Bayesian model. The sum of the differences between the posterior mean of the model-level deviance and the deviance at each draw i of θ_i [17],

$$p_D = \bar{D}(y) - D_{\theta_i}(y). \tag{6.8}$$

Combine 6.7 with the model complexity measure pD and we get the DIC,

$$DIC = 2\bar{D} - D_{\theta_i}(y) = \bar{D}(y) + p_D. \tag{6.9}$$

So now the models can be compared by this measure, see table 6.6 for the results. Although the alternative model has a lower deviance, the complexity of the model has increased a lot. Therefore the DIC of the original model is better.

	Model 6.1	Model 6.3
$\bar{D}$	253 633	231 933
p_D	22	108 041
DIC	253 655	339 974

Table 6.6: Comparison of the DICs of the implemented models.

6.5 Risk premium principles

The access to the uncertainty quantification granted by the HM model, allows us to implement different risk premium principles. The most common risk premium principle used by insurers is the expected value principle

$$\rho_i = (1 + \lambda)E[Z_i], \tag{6.10}$$

where $Z_i = \sum_{j=1}^{c_i} l_{ij}$, with c_i the number of claims reported for policy i and l_{ij} the loss per claim. If λ is 0, this is the net risk premium principle. In case $\lambda > 0$ there is a loading margin that increases with the expected value. This risk premium principle does not need access to the uncertainty quantification, it is therefore the most common.
Risk premium principles that do make use of the uncertainty are the following

$$\rho_i = E[Z_i] + \lambda Var[Z_i], \quad \lambda > 0, \tag{6.11}$$

$$\rho_i = E[Z_i] + \lambda\sqrt{Var[Z_i]}, \quad \lambda > 0, \tag{6.12}$$

$$\tag{6.13}$$

the variance and the standard deviation principles respectively [31].

6.6 Individual policies

By simulating replicated sets for both claim severity and the number of claims of the new policies, we can acquire some nice visualisations of the estimation of the premium per policy with its uncertainty. Model 6.1 is applied in this setting. The sets are simulated and some different policies are extracted, namely new policies 1, 2423, 33614 and 39351. These policies are extracted since the eventual size of the premium of these policies is driven by a different process (either number of claims or claim severity). Policy 1 has an above average estimated number of claims and claim severity, policy 33614 has a very high estimated number of claims and claim severity (caused by the high Catalogue value, $227k$, of this car), policy 39351 has a low estimated number of claims and a high estimated claim severity and lastly policy 2423 has a high estimated number of claims and a low estimated claim severity. The estimated values of these components can be found in table A.10 of the appendix.

6.7 Whole portfolio

Measuring uncertainty is especially interesting for setting received premium targets of the whole portfolio and reserving. The insurance industry should set prudent premiums, such that they reduce the possibility of

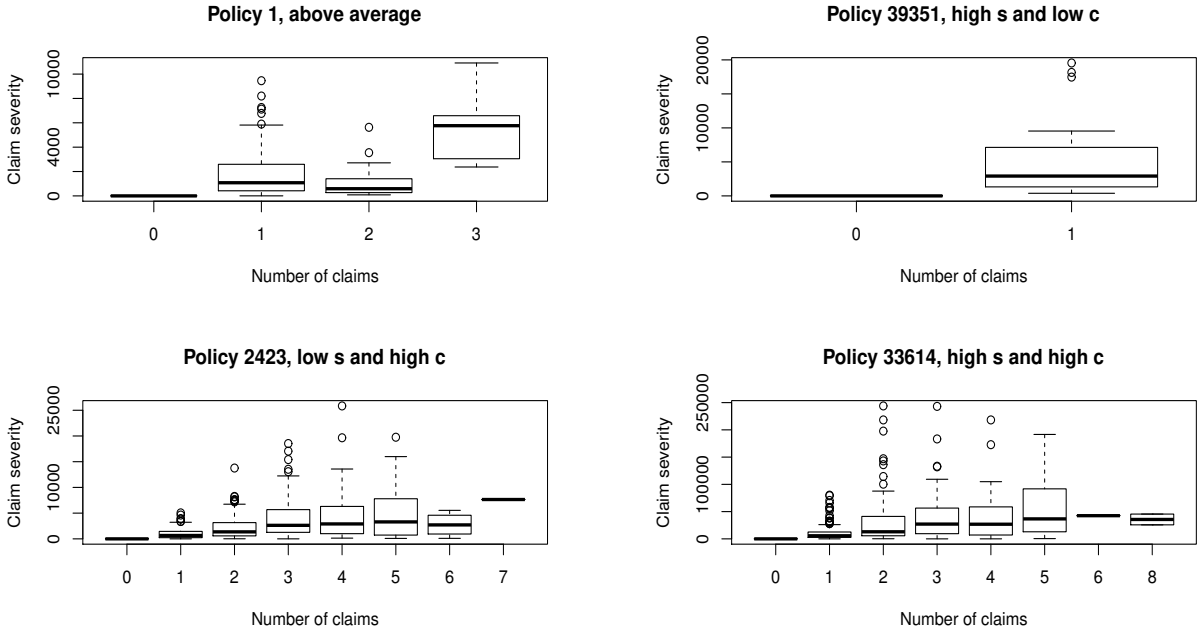


Figure 6.2: Simulated number of claims and claim severity responses for different policies of the test set.

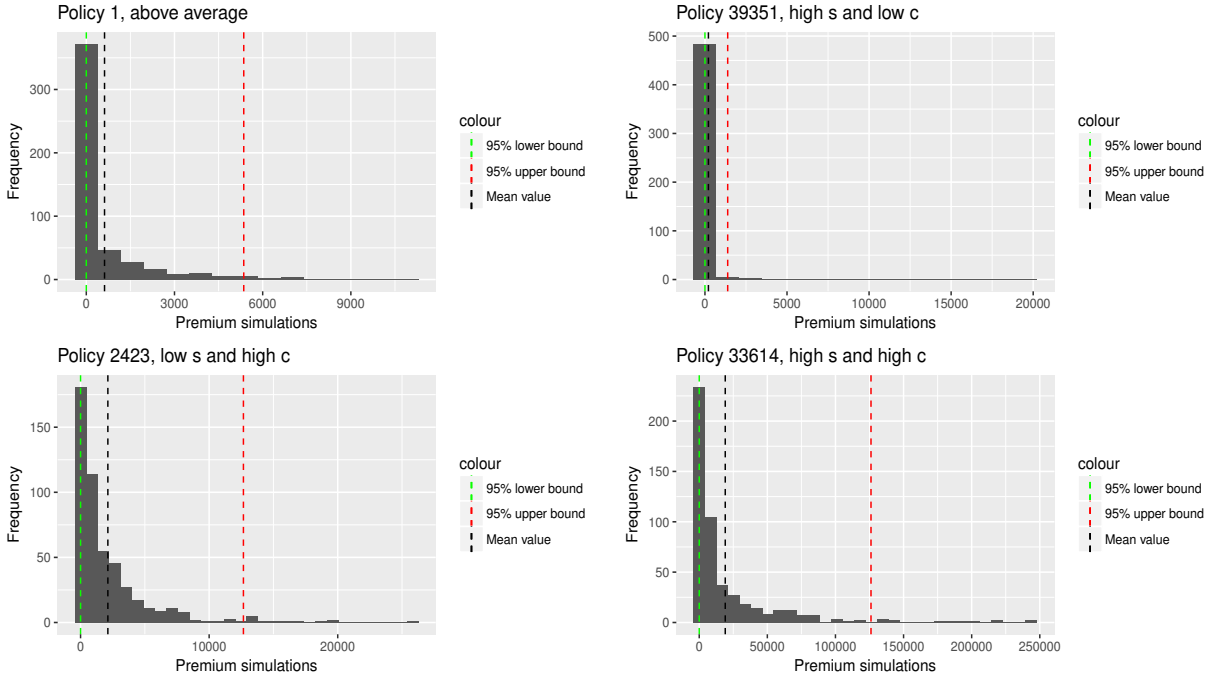


Figure 6.3: Histogram of simulated premiums for different policies with 95% credible intervals of the test set.

not being profitable or in the very least reserve money so they can pay out the claims during the business year. The built hierarchical model covered the 95% credible interval of the hold-out set quite well (see section 6.2.5). The predictive distribution for the whole portfolio is represented by the $n.sims$ draws for each of the $n_{rows}(testset)$ policies in the test set. Suppose we want to know the 95% credible interval of this portfolio as a whole, we can then sum over each replicated set and calculate the proper quantiles for the combined (500) replicated summations. This can be of interest for the active portfolio of an insurer. In this manner an insurer knows the amount of money they should have at all times for the active portfolio within a certain degree of confidence. Model 6.1 is applied in this setting.
All the credible intervals actually give a good indication for the expected paid out claims for the whole (active) portfolio, see table 6.7. This method allows the insurer to set prudent buffers for his active portfolio.

95%	75%	50%	Observed
$[9.73 \cdot 10^6\ ,\ 11.48 \cdot 10^6]$	$[9.91 \cdot 10^6\ ,\ 10.55 \cdot 10^6]$	$[10.03 \cdot 10^6\ ,\ 10.42 \cdot 10^6]$	$10.41 \cdot 10^6$

Table 6.7: Credible intervals for the size of the claims for the whole test portfolio compared to the observed total claim size.

6.8 Conclusion

The first method used to quantify uncertainty is BART. This method combines a tree-based approach with a MCMC backfitting algorithm to capture the uncertainty. This method is computationally very expensive and did not prove good at quantifying uncertainty even though way more iterations were used than recommended by the original paper by Chipman. Only 43% of observed values were found within the 95% credible intervals. Increasing the number of iterations did not improve this significantly.
We then continued with Hierarchical modelling and applied some conventional methods to validate the model such as checking the model's capabilities to retrieve the set parameters of the fake response data and the posterior predictive check. The first model did not perform well in the posterior predictive check although it does a good job at quantifying the uncertainty: 95.4% of the observed values were found within the 95% credible intervals. Proper quantification of uncertainty allows us to make use of different risk premium principles where we not only use the expected value but implement a part of the uncertainty per policy in the premium calculation as well.
The first model was used to simulate the posterior predictive distribution per policy and for the whole portfolio. This allows us to visualise the effects of both the number of claims and claim severity responses for a single portfolio and quantify the credible interval for the whole portfolio. The latter is especially interesting for reserving use. In this way the insurer know approximately how much money to reserve for his active portfolio within the chosen credible intervals.

Chapter 7

Tracking models in recent hold-out year

The models from both previous chapters 5 and 6 are applied on the holdout set that contains policies starting from January 2015. Policies that started before this month but are still open during the year 2015 are ignored in this analysis, so these claims are not taken into account. The contract starting and ending dates are visualised in figure 7.1. This figure shows that most contracts that have been opened at the beginning of the year are also closed before the end of the time period of this analysis. It is important to realise however that some claims have likely not been reported yet in this set. Therefore although risk premiums are sharply set if the loss across the whole portfolio is as close to zero as possible, the model is a lot more prudent if the overall loss is negative since some claims have been incurred but not reported yet. The earned premiums versus the reported claims are compared in this chapter for the different models.

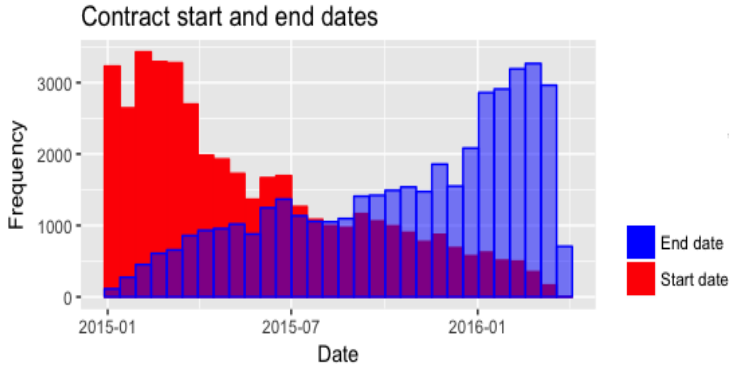


Figure 7.1: Contract start and end dates during the holdout year (from start 2015).

7.1 Predictive performance

Five models are used on the holdout data to set premiums for the new policies. The models that are employed to set the premiums are the following: GLM, GBM, RF (separate + direct) and the HM. The other models are not included because of either worse performance in accuracy or in quantifying uncertainty. For the HM, we use multiple type of predictors. The mean for each policy, the upper bounds of different credible intervals and we apply the standard deviation principle (section 6.5) with $\lambda = 0.1$ (msd).
The predictive performance is quantified for these models. The RMSE is used as the measure for predictive performance.
The GBM surprisingly has the best predictive performance, while both RFs perform slightly better than the basic GLM model on the holdout year. The mean value of the HM has a similar predictive performance to the GLM, which is to be expected since its main advantage is quantifying uncertainty and not increasing the predictive performance. All other versions of the HM have a worse predictive performance than applying the expected premium of the HM.

	GLM	GBM	RF sep	RF dir	HM mean	HM 75%	HM 80%	HM msd
RMSE	1398	1386	1393	1396	1399	1422	1449	1404

Table 7.1: RMSE for each premium model on the holdout year.

7.2 Tracking earned versus paid

The predictive performance in RMSE quantifies the error in absolute terms, so if the model consistently estimates the premium higher or lower than the observed with some margin, the predictive performance will not reveal this. To give insight into this, the earned premiums minus the paid claims are tracked for the different models in the most recent year (see figure 7.2). The final values of the earned premiums minus the paid out claims are registered in tables A.11 and A.12 of the appendix.

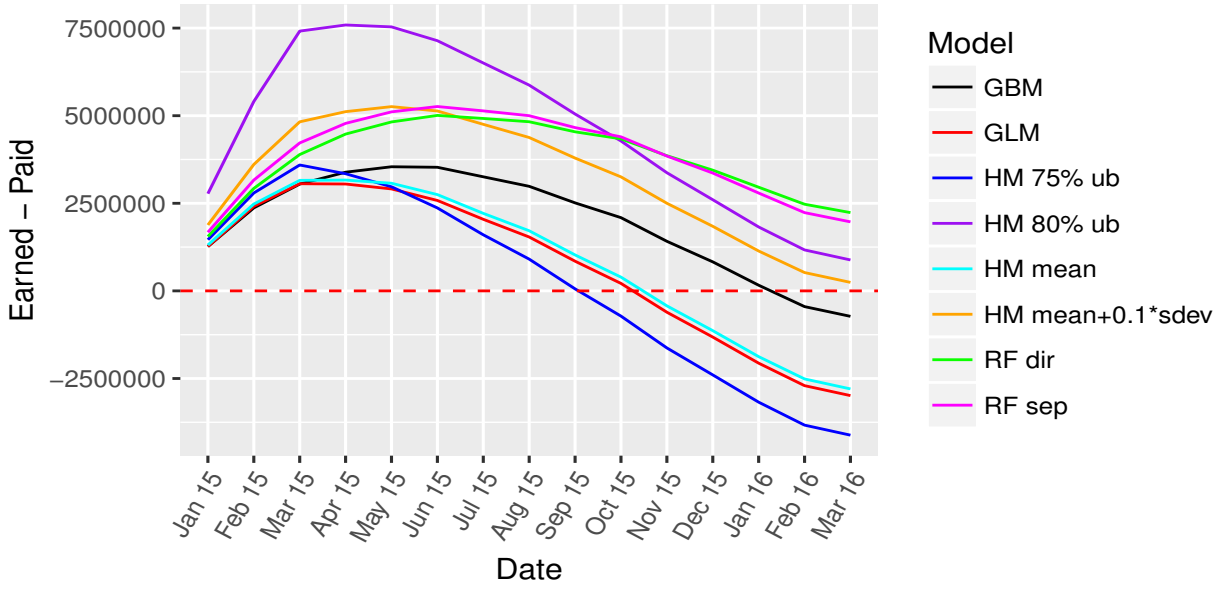


Figure 7.2: Earned premiums minus the paid claims for the different models in the holdout set (from January 2015).

Evaluating the earned premiums minus the paid claims is not very straightforward. If the insurer attaches more value to the overall absolute outcome, then the *HM msd* model performs best, closely followed by the *GBM* model. Suppose we assume that there are quite some incurred but not reported (IBNR) claims out there and therefore more value is attached to ending up with a surplus, then *HM msd*, *RF sep*, *HM 80% upper bound* and *RF dir* might be more attractive. If stability is an additional requirement, the *HM 80% upper bound* is less attractive since it swings more and therefore creates the misconception of having a huge surplus early in the year.
Even though the predictive performance of the GLM was quite good compared to the other models in the previous section, the notion of consistently too low estimation by the GLM and HM is shown by figure 7.2. The *HM msd* model performs better than its mean or the GLM because of this consistent lower estimation and it is therefore prudent to incorporate a percentage (λ) of the uncertainty for each policy.

7.3 Model lift

Another valuable tool to measure model performance is plotting the model lift. We plot the lift of the models by sorting the predicted values (by each model) for the policies of the holdout set from small to large and plot the belonging cumulative observed exposure (duration of a policy) and cumulative observed loss. The cumulative observed exposure (or loss) are the summed ratios of the exposures (or loss) of the reordered

policies (from small to large predicted values) to the total summed exposure (or loss) of the whole holdout set. This means that if we indeed recognise the risky accounts properly, we should find a low amount of the cumulative observed loss for the policies we did not deem risky (low predicted values). The model lift therefore shows if the model is good at separating the risky policies from the non-risky. The perfect model would be exactly predicting the observed claim sizes, this is also plotted in 7.3. We notice that the regression models do a poor job at recognising the risky accounts.

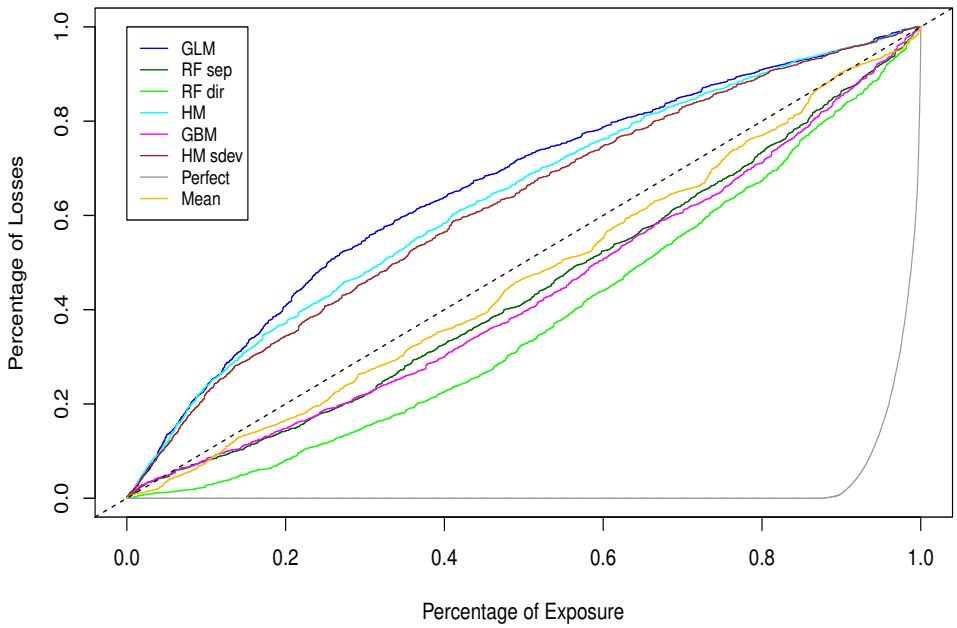


Figure 7.3: Model lift.

7.4 Conclusion

Various models were tested on the recent hold-out year by measuring predictive performance (RMSE), by tracking the difference of earned premiums and paid claims and by plotting the model lift.
The predictive performance on the holdout set revealed the GBM as the best model, closely followed by the two RF models and then by the GLM. The various HMs perform the worst.
Evaluating the difference between the earned premiums and paid out claims in the holdout year is not very straightforward. Ideally we want the absolute difference to be as close to zero as possible, but since there might be some incurred but not reported claims out there it is more prudent to have a small buffer. The HM that uses the 80% upper bound does not seem very stable, it moves from a large surplus early in the year to a lot less by the end of the duration of the policies. The HM that applies the standard deviation principle with $\lambda = 10\%$ performs performs best. Other methods that end with a surplus and fairly close to zero are the RFs. The GBM ends with a deficit but its absolute difference is fairly low as well. We see here that the basic GLM and HM does cause the insurer to have a deficit in revenue by the end of the duration of the policies.
Finally we used the model lift to investigate which model does effectively recognises the risky policies from the non-risky. We see that both the GLM and HMs do not perform well here. All Machine Learning techniques are able to recognise the risky policies. The direct RF is the most successful in this respect.

Chapter 8

Discussion

The answers to the main - and subquestions of this research are summarised in this chapter.

(i) Frequentist GLMs and GAMs are still the most widely used methods for actuaries in the current environment. If these are not used, then some actuaries use variations of the Frequentist GLM. The most common are the Negative Binomial, Zero-inflation and Hurdle models instead of the standard Poisson GLM for the number of claims or alternatives to the Gamma distributed claim severity are used such as the Log-normal, Inverse-Gamma, Pareto etcetera.

(ii) The insurance market is a very competitive one and the risk premium prices are therefore more and more sharply set. As a consequence some insurers in the Netherlands suffer from overall too low estimation of the total claim sizes and therefore think that the risk premiums do not match the paid claims in some manner. The current methods are however very easy to apply in the market. The premiums are easy to set with the readouts of a GLM and the model does not have to be consulted during this process.

(iii) There are many machine learning techniques that can be applied to an insurance dataset to set the premiums such as Random Forests, Gradient Boosting Machines and Neural Networks. RFs and GBMs filter better on the importance of variables and allow interpretation by partial dependence plots. Furthermore we have Hierarchical modeling which is a Bayesian approach to GLMs. HMs are very time consuming to fit but are flexible in further increasing the complexity of the models by either changing the distributions used or changing the hierarchical structure. They also provide a natural way of quantifying the uncertainty in the model. Bayesian Additive Regression Trees combine machine learning techniques and attempts to quantify the uncertainty as well.

(iv) There are different methods to measure if a model performs better than another. Predictive performance on a random subset or on a holdout year by the root-mean-square error is a widely used technique. Tracking the earned premiums versus the paid claims of different models on a holdout year and checking the trajectory and final position is also recommendable. A method that is often used in data science is to plot the model lift. this shows which models are better at recognising the risky policies from the non-risky.

(v) Methods that attempt to combine a machine learning technique with uncertainty quantification do not prove very useful. But a Bayesian approach such as Hierarchical Modeling that incorporates confidence (or credible) intervals in a natural way allow uncertainty quantification.

(vi) There are a number of different premium pricing principles that allow the incorporation of uncertainty into the estimation such as the standard deviation principle, where a percentage λ of the standard deviation per policy is included in the pricing of a premium. Credible intervals can also be used in estimation by either using one of the bounds for estimation. Uncertainty quantification can also be used to evaluate the credible interval of an entire (active) portfolio.

(vii) The machine learning techniques RF and GBM do consistently outperform the GLM in predictive performance, tracking of the models and model lift. But using a HM that incorporates a percentage of the standard deviation works well to counter the consistent too low estimation by both GLMs and HMs, this is visible when tracking the models. The model lift also shows the improvements made by the machine learning techniques in comparison to GLM and HM.

Chapter 9

Recommendations for Future Research

In this final chapter of the thesis, the results of the discussion are used to suggest possible ideas for further research. I believe there are three key areas where the built models in this thesis can be improved. These three areas are the following:

i. Increase predictive performance.

ii. Improve hierarchical structure.

iii. Include left truncation and right censoring.

Each of these key areas of improvement are discussed in separate sections.

9.1 Potential increase of predictive performance

There are numerous alternative Machine Learning methods that could have a better predictive performance on the non-life insurance data such as artificial neural networks. Investigating more Machine Learning techniques is not a cumbersome task in R. You simply apply the additional technique with a training mechanism on a subset of the data and derive the tuning parameters. This is the same basic approach we used in section 5.2.

9.2 Improve hierarchical structure

In section 6.2.6 two hierarchical structure improvements were proposed. The policyholder structure improvement (see figure 3.1) is not very attractive since we only have policy level characteristics that are applicable in the regression formula since using information about sex, race or religion is not allowed. This structure becomes increasingly attractive when more data is acquired to describe policyholder characteristics through Telematics. Telematics registers the driving behavior of a certain policyholder by application of a chip in the car of the policyholder. We can therefore create policyholder characteristics such as *brake behaviour*, *acceleration behaviour*, etc.
A possible extension to this structure is an additional insurance product level as the highest hierarchy level. This will considerably increase the data size and this is not very attractive since HMs are quite time consuming to fit.

9.3 Include left truncation and right censoring

Two key aspects that can be included in a statistical model and especially in a flexible one, such as HMs, is left truncation and right censoring. First we shall indicate why the true claims process is actually left truncated. The possible right censoring is a more complicated issue.
Most insurers use the deductible level as a covariate in the predictive models. An alternative and possibly rewarding approach is the following. The true claims process is left truncated since an accident (and thus

claim) is not reported if the amount of the claim is less than the deductible of policy i, so when $l_{ij} - d_i \leq 0$. That means that when $d_i \geq 0$ there are some some unreported claims and the data is therefore *left truncated*. If we consider left truncation, the mass density of the probability distribution of the claim sizes changes for policies with different deductible levels. This moves the mass of the probability distribution to the right and could lead to less consistently too low estimated premiums by the HM and GLM.
In section 4.2.2.1 we stated that we only consider claims that are closed. So that means that some reported claims are not included in the models. This is not very strange since we do not know what the final claim size will be when an initial claim estimate is reported at the insurer. The difficulty here is how we could use this data. In an orthodox right censoring example we have the following: we know claim loss l_{ij} is at least c, so $l_{ij} \geq c$. But the issue with the early estimates is that the ultimate claim can be either lower or higher than the initial estimate (even 0), so we have a censoring interval from zero to infinity which does not provide us with additional information. This censoring issue is what the reserving branch of the insurer deals with. They estimate the ultimate claims based on the reported non-finalised claims. An additional predictive model that estimates the ultimate claims from the non-finalised claims could be interesting to incorporate in the existing predictive model.

Appendix A

Appendix

A.1 Density functions

A.1.1 Poisson

The Poisson distribution has the following density function

$$f_{N_i}(n_i|\lambda_i, \Delta_i) = e^{-\Delta_i\lambda_i}\frac{(\Delta_i\lambda_i)^{n_i}}{n_i!}, \quad n_i \in \mathbb{N}. \tag{A.1}$$

A.1.2 Gamma

The Gamma distribution has the following density function

$$f_{S_i}(s_i|\alpha_i, \beta_i, \omega_i) = \frac{(\omega_i\beta_i)^{\omega_i\alpha_i}}{\Gamma(\omega_i\alpha_i)}s_i^{\omega_i\alpha_i-1}e^{-\omega_i\beta_i s_i}, \quad s_i \in (0, \infty).$$

A.2 Algorithms

A.2.1 Random Forest

The following algorithm can be found in [21].

1. For $b = 1$ to B:
   a. Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   b. Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

A.2.2 Gradient Boosting Machines

As was the case for the Random Forest algorithm, the GBM algorithm can be found in [21].

1. Initialize $f_0(x) = arg\ \underset{c}{min}\sum_{i=1}^N L(y_i, c)$
2. For $k = 1$ to K:

a. For $i = 1, ..., N$ compute

$$r_{ik} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{k-1}}.$$

b. Fit a regression tree to the targets r_{ik} giving terminal regions $R_{mk}, m = 1, ..., M_k$.

c. For $m = 1, ..., M_k$ compute

$$c_{mk} = arg\ \min_{c} \sum_{x_i \in R_{mk}} L(y_i, f_{k-1}(x_i) + c).$$

d. Update $f_k(x) = f_{k-1}(x) + \sum_{m=1}^{M_k} c_{mk} I(x \in R_{mk})$

3. Output $\hat{f}(x) = f_K(x)$.

A.3 Claims categorisation

Here we made the categories for the claims as used in section 4.1.

Category	Type
Collisions	A
Traffic rules	B
Scratches	C
Parking and manoeuvring	D
Weather/Nature	E
Unknown	F
Theft/Vandalism/Loss	G

	Cause	Frequency	Type
1		9	F
2	Aanrijding Vast Object	8771	A
3	Aanrijding dier	1125	A
4	Aardbeving	2	E
5	Achteruit rijden	34	D
6	Beide over middellijn	2	A
7	Brand	90	E
8	Carjacking	8	G
9	Diefstal toebehoren	70	G
10	Dubbel maneuver	10	D
11	Geen contact tussen voertuigen	6	A
12	Grondverschuiving	1	E
13	Haakslaan	1	A
14	Hagelschade	1609	E
15	Homejacking	10	G
16	Inhalen	3	B
17	Inrijden of verlaten parking	4	D
18	Kanteling	3	A
19	Keren op de weg	2	D
20	Klapband	42	A
21	Kop-staart	423	A
22	Kortsluiting	15	E
23	Lawine	1	E
24	Maneuver	579	D
25	Neervallende stenen	7	E
26	Negeren verkeerslicht	16	B
27	Negeren voorrangsteken	111	B
28	Ontploffing	3	E
29	Openen deur	22	C
30	Opspringend steentje	17988	C
31	Over middenlijn	78	B
32	Overig / Onbekend	23960	F
33	Overstroming	35	E
34	Parkeerstand	1296	D
35	Poging tot diefstal	1984	C
36	Schade door lading	15	D
37	Slippen	1198	A
38	Sneeuw-of ijsdruk	35	E
39	Stormschade	171	E
40	Totaal Diefstal Voertuig	292	G
41	Uitzwenken voertuig	8	D
42	Vandalisme en Kwaad Opzet	4270	G
43	Verandering van rijstrook	133	B
44	Verboden richting	1	B
45	Verlies lading	22	G
46	Vervoer voertuig incl op/afladen	1	C
47	Voorrang van rechts	102	B
48	Y met voorrang nog niet op kruispunt	2	B
49	Zwakke weggebruiker	250	B

Figure A.1: The categorisations of the different kept claim causes.

A.4 Data Handling log

1. Reduced portfolio and claims data by removing all non-Omnium columns.
2. Analysed available columns and removed those that were severely corrupted or totally useless. Analysis in the Unexplained vars tab of the excel "categorieen+covariaten".
3. Reordering of portfolio in categories: ID , Contract info, Car info, insured info, claim info, least useful info.
4. Removed duplicates of brand names.
5. Transformed acquired license year to experience. Year of reference is 2016. Renamed it as well.
6. Handled 2 exceptions of Building year (1360), after research we noticed 1960 was meant.
7. Transformed Building year car to age of car. Year of reference is 2016 (last available info). Renamed it as well.
8. Transformed birth date to continuous age of policyholder. Renamed it as well.
9. Date transformations of gebdat_b, ingdatum and hproldat to the same date format. First checked for NA values.
10. Transformed birth date to policyholder age. Overwritten column.
11. Fixed NA values of the contract end date (veinddOM).
12. Removed overlap some consecutive policies had from the same policyholder.
13. Duration calculation: time difference between contract start date (vbgindOM) and contract start date (veinddOM). New column introduced.
14. Recurring series of accounts removed (marked by .[number]).
15. Transformed some numerical covariates (standardised and some by taking the logarithm).
16. Negative claims set to zero.
17. Claims that did not result into an amount set to zero.
18. Claims dataset and policy dataset are joined by placing the claims in the correct policies. The reported claim dates should be within the policy start and end dates. Each additional claim increases the number of claims within that policy and the total claim size.
19. Split the dataset in claims reported before 2015 and after 2015.
20. Test and train sets are made.
21. Severity datasets are created by taking the policies with at least one claim and taking the average of the claim sizes (claim severity).

A.5 Further covariates info and handling

Numerical (some partially categorical already):

- Number of cars, range = 1 to 9.
- Catalogue value, range = 2,331.4 to 650,000. We standardise this variable.
- Vehicle power, range = 30 to 1110cc.
- Vehicle weight, range = 560 to 2805.
- Power-to-Weight ratio in percentage, range = 2.42 to 68.52.
- Age of car, range = 1 to 54.
- Age of policyholder, range = 18.12 to 93.23.
- Years of driving experience, range = 1 to 74.
- Number of claim-free (ccfy), range = 0 to 26.
- Number of driver's licenses, range 1 to 10.
- Credit score, range from 1 to 4.
- Value of the Accessories in the car.
- Mileage of the car, range from 1 to 3.

Categorical (ordinal and nominal):

- Brand of car, 59 types (reduced to 32).
- Fuel type, 5 types.
- Regions, 19 types; 18 regions + 1NA: 0A, 1D-E, 2B-E ,3B-E ,4B-E ,5B-D.
- Home owner, 2 types: Yes/No.
- Maternal language, 4 types.

A.5.1 Grouping actions

Categorical

- Brand: We put the small groups together in "Other brands". So all brands with less than 1000 instances are pooled together

A.6 Results

Covariate	Estimate	Std. Error	z-value	Pr(> |z|)
(Intercept)	-2.8112	0.2254	-12.47	0.0000
Brand: audi	-0.0319	0.0792	-0.40	0.6874
Brand: bmw	0.0802	0.0777	1.03	0.3024
Brand: chevrolet	-0.0770	0.1276	-0.60	0.5462
Brand: citroen	0.0111	0.0767	0.14	0.8850
Brand: dacia	-0.2556	0.1065	-2.40	0.0164
Brand: fiat	0.0546	0.0889	0.61	0.5392
Brand: ford	-0.1307	0.0789	-1.66	0.0975
Brand: honda	0.0384	0.1169	0.33	0.7427
Brand: hyundai	-0.0522	0.0908	-0.57	0.5654
Brand: jaguar	0.1527	0.2025	0.75	0.4510
Brand: kia	-0.1215	0.0900	-1.35	0.1770
Brand: lancia	-0.0986	0.1579	-0.62	0.5322
Brand: land rover	0.0092	0.1385	0.07	0.9468
Brand: mazda	-0.0907	0.1169	-0.78	0.4378
Brand: mercedes	0.0657	0.0816	0.80	0.4209
Brand: mini	-0.0006	0.0946	-0.01	0.9950
Brand: mitsubishi	0.0084	0.1322	0.06	0.9494
Brand: nissan	-0.1827	0.0905	-2.02	0.0435
Brand: opel	-0.0574	0.0776	-0.74	0.4595
Brand: other	-0.1028	0.1225	-0.84	0.4014
Brand: peugeot	-0.0522	0.0774	-0.67	0.5001
Brand: porsche	-0.2956	0.2711	-1.09	0.2756
Brand: renault	-0.0835	0.0775	-1.08	0.2809
Brand: saab	-0.5750	0.2602	-2.21	0.0271
Brand: seat	-0.0384	0.0902	-0.43	0.6699
Brand: skoda	-0.0756	0.0884	-0.86	0.3922
Brand: smart	-0.4809	0.2776	-1.73	0.0831
Brand: suzuki	0.0663	0.1060	0.63	0.5318
Brand: toyota	0.0521	0.0845	0.62	0.5374
Brand: volkswagen	-0.0204	0.0748	-0.27	0.7854
Brand: volvo	-0.0287	0.0976	-0.29	0.7689
Fuel: D	0.1698	0.0251	6.75	0.0000
Fuel: E	-11.0940	197.0035	-0.06	0.9551
Fuel: G	0.3625	0.2353	1.54	0.1234
Fuel: H	-0.0946	0.1750	-0.54	0.5887
Ncars	-0.0633	0.0182	-3.48	0.0005
Deductible	-0.0985	0.0107	-9.24	0.0000
Reference Year	-0.0446	0.0050	-8.90	0.0000
Vehicle Power log	0.1774	0.0421	4.21	0.0000
Mileage	0.2694	0.0205	13.15	0.0000
Car Age	-0.1009	0.0106	-9.54	0.0000
Policyholder Age	0.0389	0.0249	1.56	0.1189
Region: 1D	-0.4353	1.0013	-0.43	0.6638
Region: 1E	0.1433	0.0487	2.94	0.0033
Region: 2B	-0.1960	0.3802	-0.52	0.6062
Region: 2C	0.0623	0.2039	0.31	0.7601
Region: 2D	-0.1004	0.1170	-0.86	0.3911
Region: 2E	0.0968	0.0519	1.87	0.0619

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
Region: 3B	0.0942	0.0888	1.06	0.2884
Region: 3C	-0.0191	0.0476	-0.40	0.6874
Region: 3D	-0.0363	0.0414	-0.88	0.3808
Region: 3E	-0.0712	0.0523	-1.36	0.1735
Region: 4B	-0.1793	0.0498	-3.60	0.0003
Region: 4C	-0.0507	0.0447	-1.13	0.2565
Region: 4D	-0.0335	0.0591	-0.57	0.5706
Region: 4E	-0.0266	0.0672	-0.40	0.6920
Region: 5B	-0.1906	0.0469	-4.07	0.0000
Region: 5C	-0.1035	0.0568	-1.82	0.0681
Region: 5D	-0.0661	0.1685	-0.39	0.6950
Region: U	-11.4472	78.6205	-0.15	0.8842
Experience	-0.1132	0.0259	-4.36	0.0000
Nlicenses	0.0612	0.0171	3.58	0.0003
Claim-free Years	-0.0491	0.0113	-4.33	0.0000
Credit score	0.0705	0.0157	4.50	0.0000
Accessories	0.0270	0.0105	2.56	0.0103
Language: en	-0.2314	0.1087	-2.13	0.0333
Language: fr	0.0074	0.0986	0.07	0.9405
Language: nl	0.0064	0.0990	0.06	0.9483

Table A.1: Summary of the GLM fit for the number of claims.

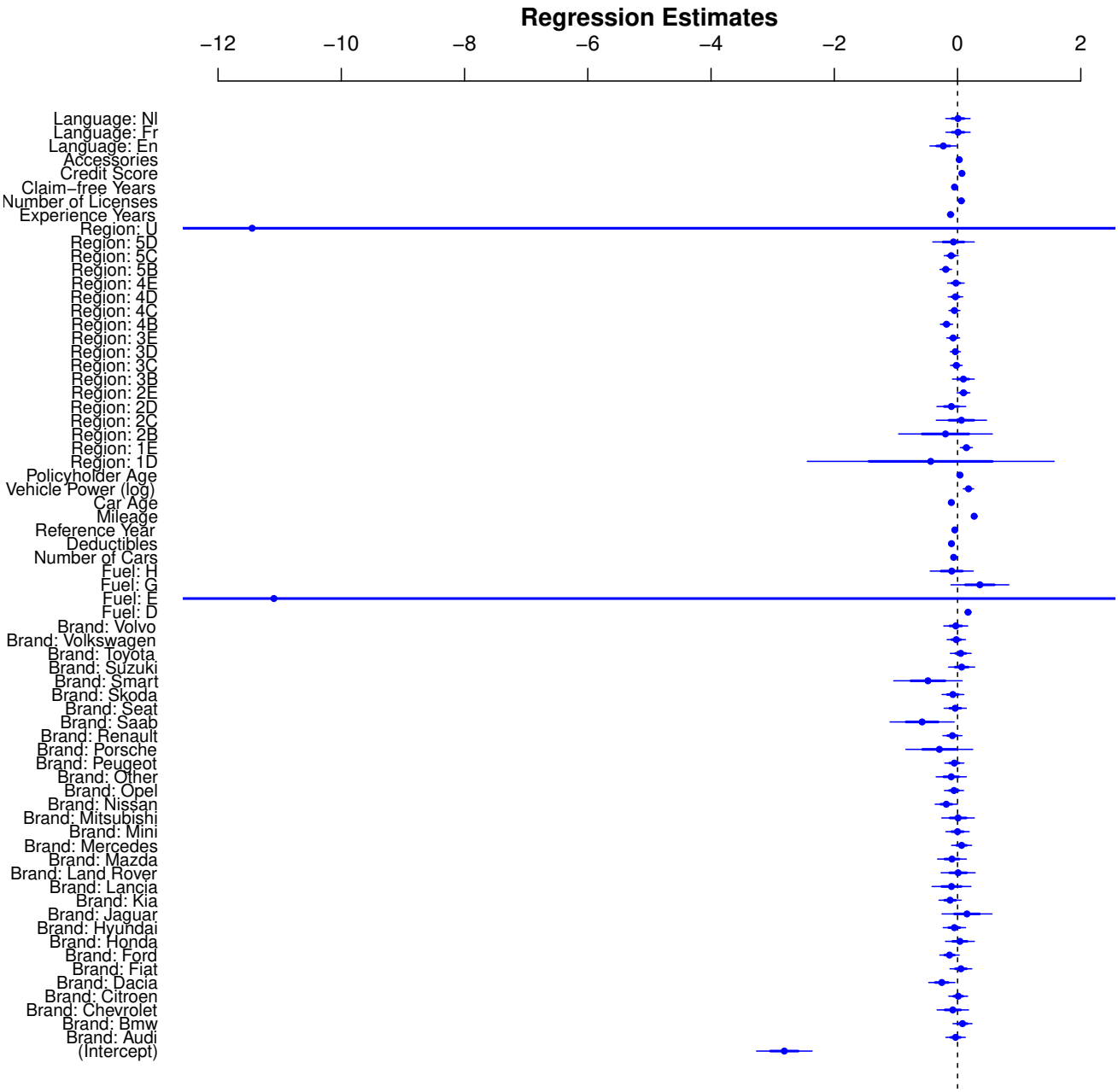


Figure A.2: Visual summary of the GLM fit for the number of claims.

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
(Intercept)	6.0074	0.9815	6.12	0.0000
Fuel: D	-0.0541	0.0518	-1.04	0.2963
Fuel: G	-0.6092	0.4046	-1.51	0.1321
Fuel: H	-0.0854	0.3059	-0.28	0.7800
Ncars	0.0579	0.0345	1.68	0.0930
Catalogue Value log	0.1477	0.1329	1.11	0.2666
Deductible	-0.0230	0.0205	-1.12	0.2607
Policy year	0.0429	0.0097	4.43	0.0000
Vehicle Power log	0.4331	0.1392	3.11	0.0019
Mileage	-0.0202	0.0385	-0.53	0.5994
Car Age	-0.0601	0.0202	-2.98	0.0029

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
Vehicle Weight log	-0.2682	0.1979	-1.36	0.1754
Policyholder Age sd	0.0353	0.0492	0.72	0.4730
Region: 1D	-3.0073	1.7494	-1.72	0.0856
Region: 1E	-0.2346	0.0941	-2.49	0.0127
Region: 2B	0.3514	0.6644	0.53	0.5969
Region: 2C	-0.0815	0.3812	-0.21	0.8308
Region: 2D	-0.1826	0.2188	-0.83	0.4042
Region: 2E	-0.2202	0.1002	-2.20	0.0281
Region: 3B	-0.2809	0.1698	-1.65	0.0981
Region: 3C	-0.2820	0.0909	-3.10	0.0019
Region: 3D	-0.3257	0.0792	-4.11	0.0000
Region: 3E	-0.2616	0.0997	-2.62	0.0087
Region: 4B	-0.2718	0.0945	-2.88	0.0040
Region: 4C	-0.2535	0.0855	-2.96	0.0030
Region: 4D	-0.0990	0.1121	-0.88	0.3771
Region: 4E	-0.1844	0.1290	-1.43	0.1528
Region: 5B	-0.2308	0.0891	-2.59	0.0096
Region: 5C	-0.3044	0.1077	-2.83	0.0047
Region: 5D	-0.6618	0.3079	-2.15	0.0316
Experience	-0.0435	0.0508	-0.86	0.3912
Nlicenses	0.0023	0.0324	0.07	0.9423
Claim-free Years	-0.0825	0.0218	-3.79	0.0002
Credit Score	0.0910	0.0314	2.90	0.0038
Home Owner: Yes	-0.0118	0.0491	-0.24	0.8106
Accessories	-0.0184	0.0188	-0.98	0.3280
Language: en	-0.2518	0.2123	-1.19	0.2357
Language: fr	0.1277	0.1928	0.66	0.5079
Language: nl	-0.1246	0.1937	-0.64	0.5202

Table A.2: Summary of the GLM fit for claim severity.

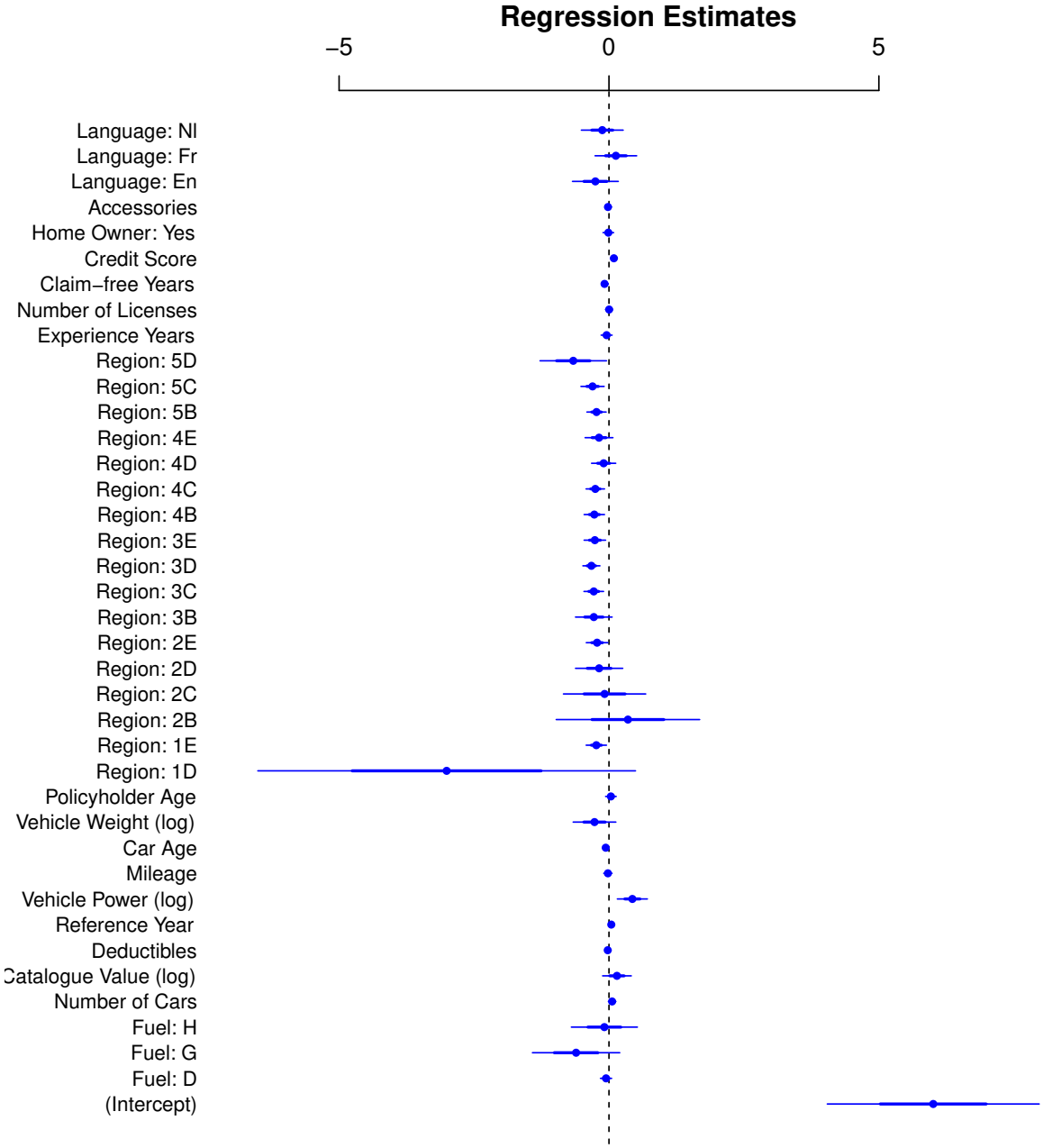


Figure A.3: Visual summary of the GLM fit for claim severity.

Covariate	Estimate
(Intercept)	7.3793
Language:fr	-0.0681
h(VehicleWeight_log-7.84385)	-11.8419
h(3-Creditsc)	-0.0984
Brand:bmw * Language:fr	0.2032
Brand:mitsubishi * Language:fr	0.7134
Region:2C * Language:fr	1.8072
Brand:mazda * h(Creditsc-3)	2.3359
Brand:renault * h(Creditsc-3)	2.4655
h(Ncars-2) * Language:fr	0.1727
ref_year * h(VehicleWeight_log-7.84385)	-17.4473

Covariate	Estimate
h(1.42029-Car_Age_sd) * Language:fr	0.0920
h(1.43302-ccfy_sd) * Language:fr	0.1040
h(2-ref_year) * h(7.84385-VehicleWeight_log)	-0.0892
h(VehiclePower_log-4.35671) * h(7.84385-VehicleWeight_log)	1.0913

Table A.3: Summary of the MARS fit for claim severity.

Covariate	Estimate
(Intercept)	-2.0662
h(ref_year-1)	0.0106
h(1-ref_year)	0.1059
h(Mileage_num-2)	0.2971
h(2-Mileage_num)	-0.3072
h(-0.580663-Car_Age_sd)	-0.3392
h(Car_Age_sd- -0.580663)	-0.1265

Table A.4: Summary of the MARS fit for the number of claims.

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
(Intercept)	5.8963	0.9889	5.96	0.0000
Fuel: D	-0.0591	0.0518	-1.14	0.2539
Fuel: G	-0.6440	0.4044	-1.59	0.1113
Fuel: H	-0.0777	0.3058	-0.25	0.7995
Number of cars	0.0524	0.0345	1.52	0.1291
Catalogue Value log	0.1483	0.1329	1.12	0.2642
Deductible	-0.0248	0.0205	-1.21	0.2268
Vehicle Power log	0.4217	0.1392	3.03	0.0025
Mileage	-0.0233	0.0385	-0.61	0.5449
Car Age	-0.0591	0.0201	-2.94	0.0033
Vehicle Weight log	-0.2387	0.1989	-1.20	0.2303
Region: 1D	-3.0089	1.7480	-1.72	0.0852
Region: 1E	-0.2312	0.0941	-2.46	0.0140
Region: 2B	0.3271	0.6639	0.49	0.6222
Region: 2C	-0.0666	0.3809	-0.17	0.8613
Region: 2D	-0.1821	0.2187	-0.83	0.4051
Region: 2E	-0.2174	0.1002	-2.17	0.0300
Region: 3B	-0.2751	0.1697	-1.62	0.1052
Region: 3C	-0.2813	0.0908	-3.10	0.0020
Region: 3D	-0.3240	0.0791	-4.10	0.0000
Region: 3E	-0.2548	0.0997	-2.56	0.0106
Region: 4B	-0.2685	0.0944	-2.84	0.0045
Region: 4C	-0.2499	0.0855	-2.92	0.0035
Region: 4D	-0.0965	0.1121	-0.86	0.3891
Region: 4E	-0.1765	0.1289	-1.37	0.1709
Region: 5B	-0.2267	0.0891	-2.55	0.0109
Region: 5C	-0.3060	0.1076	-2.84	0.0045
Region: 5D	-0.6740	0.3077	-2.19	0.0285
Experience	-0.0444	0.0516	-0.86	0.3899
Number of licenses	-0.0013	0.0325	-0.04	0.9681
Claim-free Years	-0.0862	0.0224	-3.85	0.0001
Credit score	0.0872	0.0315	2.77	0.0057

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
Home Owner: Yes	-0.0127	0.0492	-0.26	0.7961
Accessories	-0.0178	0.0188	-0.95	0.3440
Language: en	-0.2561	0.2123	-1.21	0.2276
Language: fr	0.1233	0.1927	0.64	0.5222
Language: nl	-0.1280	0.1935	-0.66	0.5084
Reference year	0.0439	0.0097	4.52	0.0000
Policyholder Age	0.0942	0.0631	1.49	0.1352
I[(Policyholder Age)^2]	-0.0598	0.0368	-1.62	0.1042
I[(Policyholder Age)^3]	-0.0321	0.0179	-1.80	0.0726
I[(Policyholder Age)^4]	0.0178	0.0099	1.81	0.0707

Table A.5: Summary of the GLM polynomial fit for claim severity.

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
(Intercept)	-2.8082	0.2261	-12.42	0.0000
Brand: audi	-0.0274	0.0792	-0.35	0.7298
Brand: bmw	0.0832	0.0777	1.07	0.2844
Brand: chevrolet	-0.0757	0.1276	-0.59	0.5531
Brand: citroen	0.0174	0.0768	0.23	0.8203
Brand: dacia	-0.2460	0.1065	-2.31	0.0209
Brand: fiat	0.0598	0.0889	0.67	0.5015
Brand: ford	-0.1281	0.0789	-1.62	0.1043
Brand: honda	0.0344	0.1170	0.29	0.7687
Brand: hyundai	-0.0497	0.0908	-0.55	0.5839
Brand: jaguar	0.1438	0.2026	0.71	0.4779
Brand: kia	-0.1203	0.0900	-1.34	0.1812
Brand: lancia	-0.0941	0.1579	-0.60	0.5511
Brand: land rover	0.0154	0.1385	0.11	0.9114
Brand: mazda	-0.0890	0.1169	-0.76	0.4463
Brand: mercedes	0.0702	0.0816	0.86	0.3899
Brand: mini	0.0041	0.0946	0.04	0.9653
Brand: mitsubishi	0.0103	0.1322	0.08	0.9381
Brand: nissan	-0.1776	0.0906	-1.96	0.0499
Brand: opel	-0.0551	0.0776	-0.71	0.4776
Brand: other	-0.0988	0.1225	-0.81	0.4198
Brand: peugeot	-0.0475	0.0774	-0.61	0.5397
Brand: porsche	-0.2905	0.2712	-1.07	0.2841
Brand: renault	-0.0781	0.0775	-1.01	0.3131
Brand: saab	-0.5644	0.2602	-2.17	0.0301
Brand: seat	-0.0406	0.0902	-0.45	0.6526
Brand: skoda	-0.0674	0.0884	-0.76	0.4456
Brand: smart	-0.4818	0.2776	-1.74	0.0826
Brand: suzuki	0.0635	0.1061	0.60	0.5495
Brand: toyota	0.0548	0.0845	0.65	0.5166
Brand: volkswagen	-0.0151	0.0748	-0.20	0.8403
Brand: volvo	-0.0245	0.0976	-0.25	0.8020
Fuel: D	0.1645	0.0252	6.53	0.0000
Fuel: E	-11.0936	196.9594	-0.06	0.9551
Fuel: G	0.3670	0.2353	1.56	0.1188
Fuel: H	-0.0924	0.1751	-0.53	0.5974
Number of cars	-0.0658	0.0182	-3.60	0.0003

Covariate	Estimate	Std. Error	z-value	Pr($>\|z\|$)
Deductible	-0.1001	0.0107	-9.36	0.0000
Reference Year	-0.0447	0.0051	-8.85	0.0000
Vehicle Power log	0.1900	0.0423	4.50	0.0000
Mileage	0.2665	0.0205	12.99	0.0000
Region: 1D	-0.4589	1.0013	-0.46	0.6468
Region: 1E	0.1442	0.0487	2.96	0.0031
Region: 2B	-0.2019	0.3802	-0.53	0.5953
Region: 2C	0.0611	0.2039	0.30	0.7644
Region: 2D	-0.1053	0.1170	-0.90	0.3682
Region: 2E	0.0961	0.0519	1.85	0.0640
Region: 3B	0.0960	0.0888	1.08	0.2796
Region: 3C	-0.0186	0.0476	-0.39	0.6962
Region: 3D	-0.0355	0.0414	-0.86	0.3913
Region: 3E	-0.0696	0.0523	-1.33	0.1831
Region: 4B	-0.1791	0.0498	-3.60	0.0003
Region: 4C	-0.0507	0.0447	-1.13	0.2567
Region: 4D	-0.0333	0.0591	-0.56	0.5737
Region: 4E	-0.0256	0.0672	-0.38	0.7030
Region: 5B	-0.1924	0.0469	-4.11	0.0000
Region: 5C	-0.1033	0.0568	-1.82	0.0689
Region: 5D	-0.0672	0.1685	-0.40	0.6900
Region: U	-11.4548	78.8018	-0.15	0.8844
Number of licenses	0.0555	0.0172	3.23	0.0012
Claim-free Years	-0.0460	0.0117	-3.94	0.0001
Credit score	0.0650	0.0158	4.12	0.0000
Accessories	0.0269	0.0105	2.56	0.0105
Language: en	-0.2298	0.1088	-2.11	0.0346
Language: fr	0.0020	0.0986	0.02	0.9839
Language: nl	0.0013	0.0990	0.01	0.9899
Policyholder Age	0.1058	0.0323	3.28	0.0011
I[$(\text{Policyholder Age})^2$]	0.0015	0.0139	0.11	0.9139
I[$(\text{Policyholder Age})^3$]	-0.0273	0.0088	-3.11	0.0018
I[$(\text{Policyholder Age})^4$]	0.0041	0.0033	1.24	0.2136
Experience	-0.1270	0.0264	-4.82	0.0000
Car Age	-0.0796	0.0134	-5.96	0.0000
I[$(\text{Car Age})^2$]	-0.0335	0.0136	-2.46	0.0138
I[$(\text{Car Age})^3$]	0.0020	0.0042	0.48	0.6311

Table A.6: Summary of the GLM polynomial fit for the number of claims.

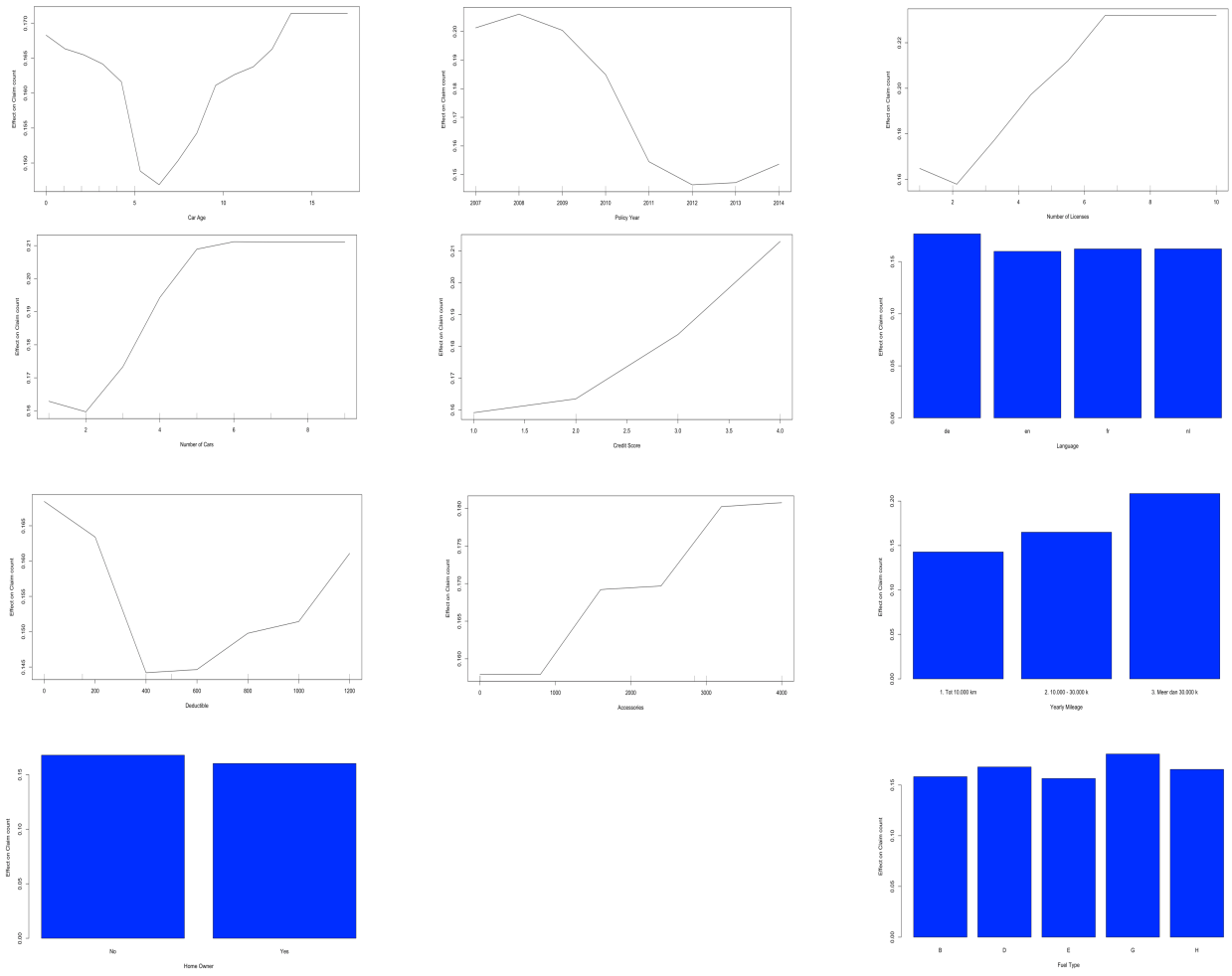


Figure A.4: Marginal effects of the different covariates on the number of claims. The remaining variables of the Random Forest model are illustrated.

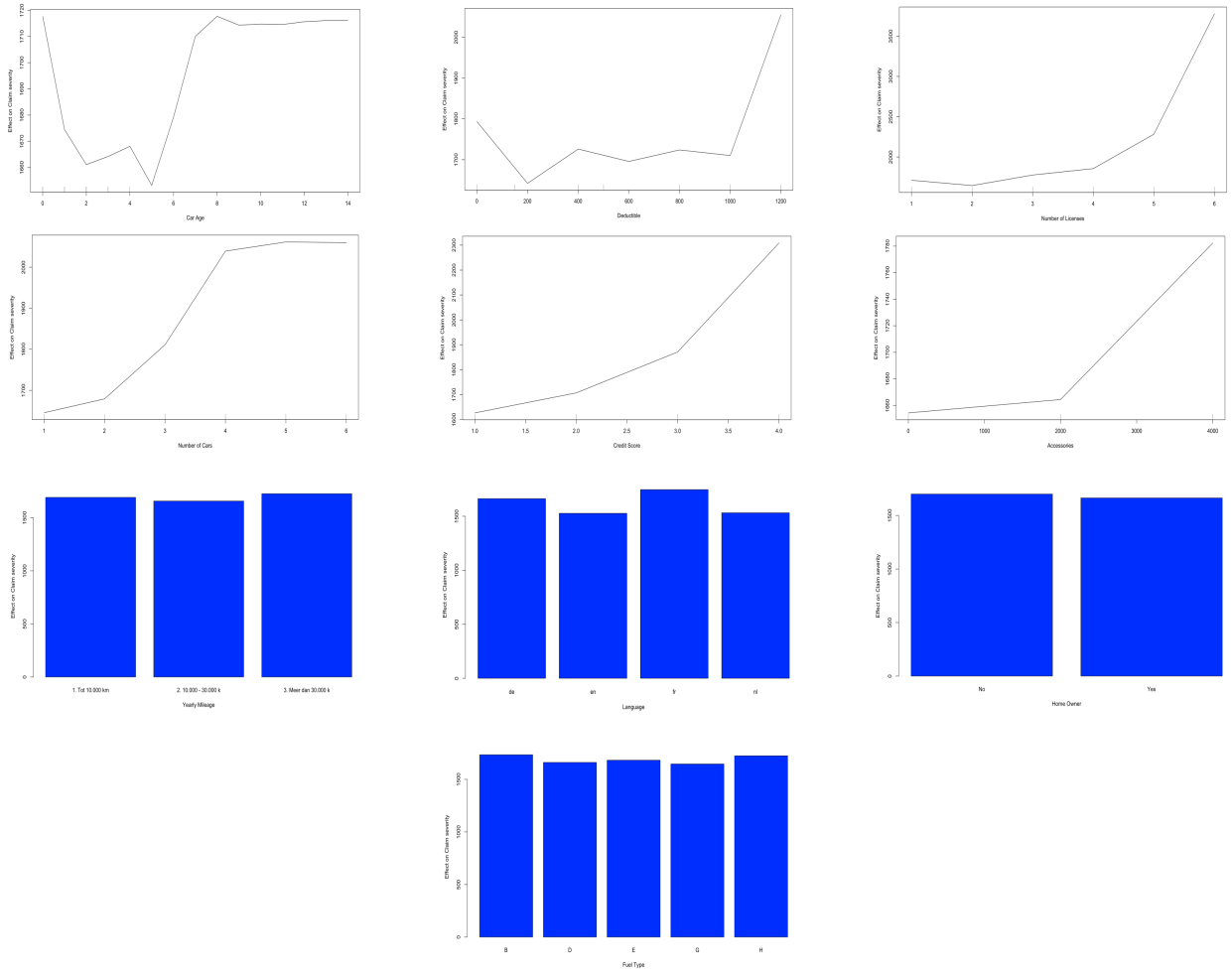
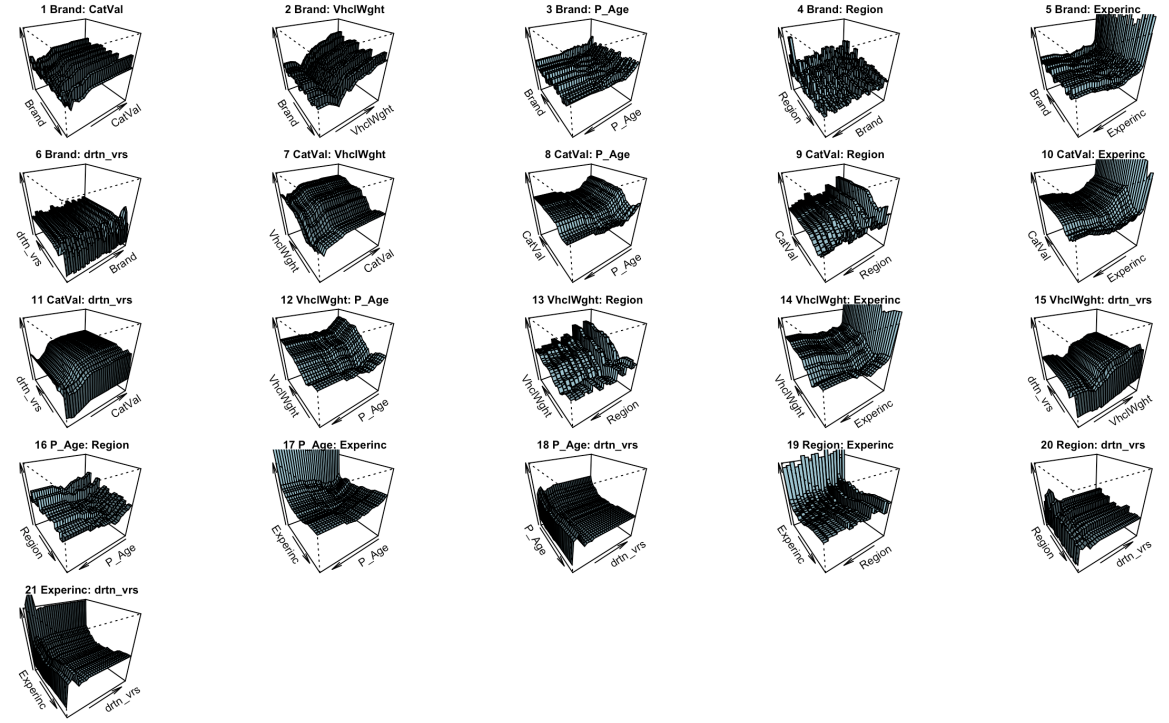
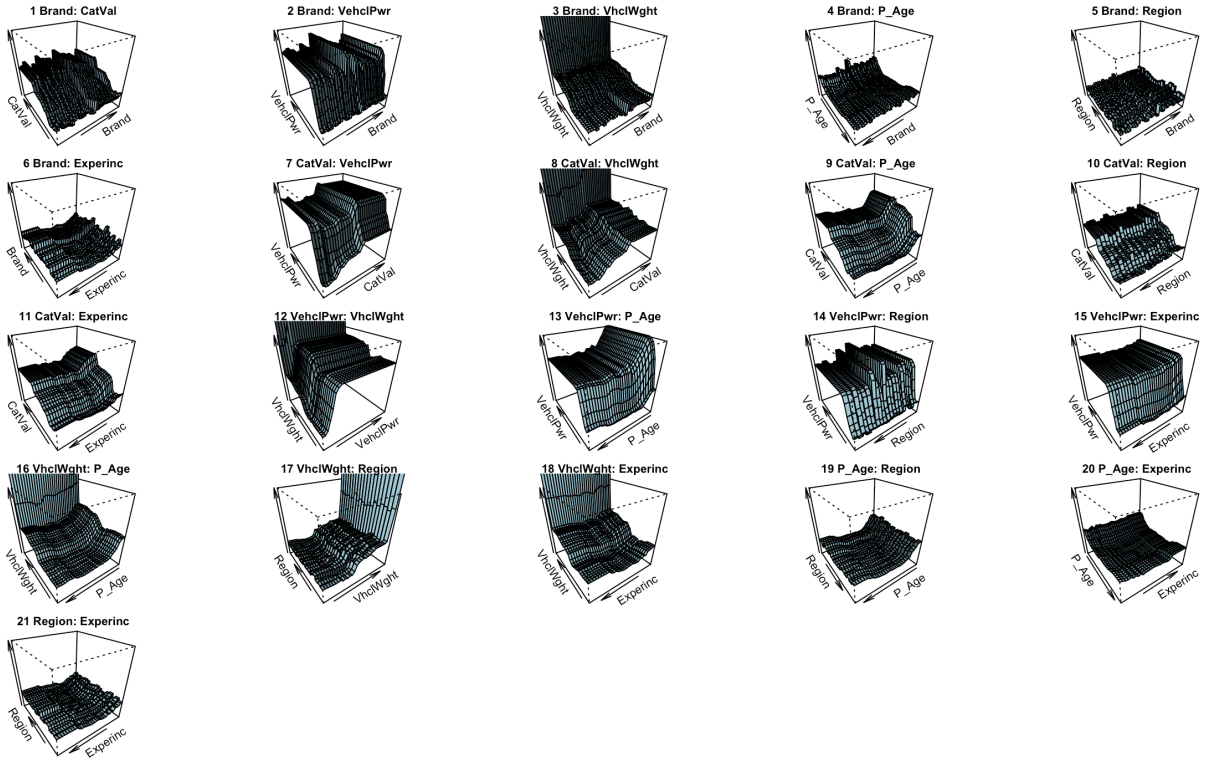


Figure A.5: Marginal effects of the different covariates on the claim severity. The remaining variables of the Random Forest model are illustrated.



(a) 3D partial dependence plots for the number of claims.



(b) 3D partial dependence plots for the number of claims.

Figure A.6: 3D partial dependence plots of the RF model for both the number of claims and claim severity.

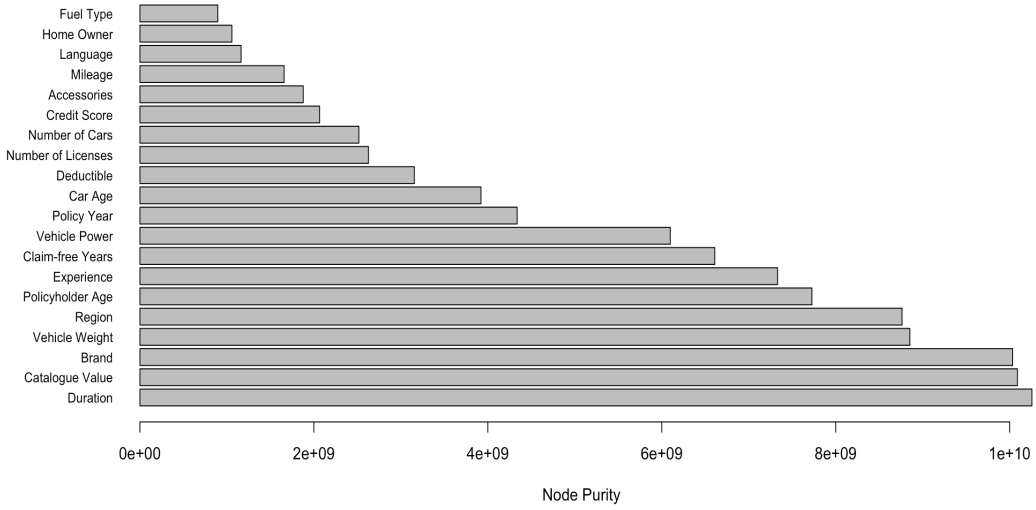


Figure A.7: Variable importance according to the Random Forest Premium model on a subset of the data.

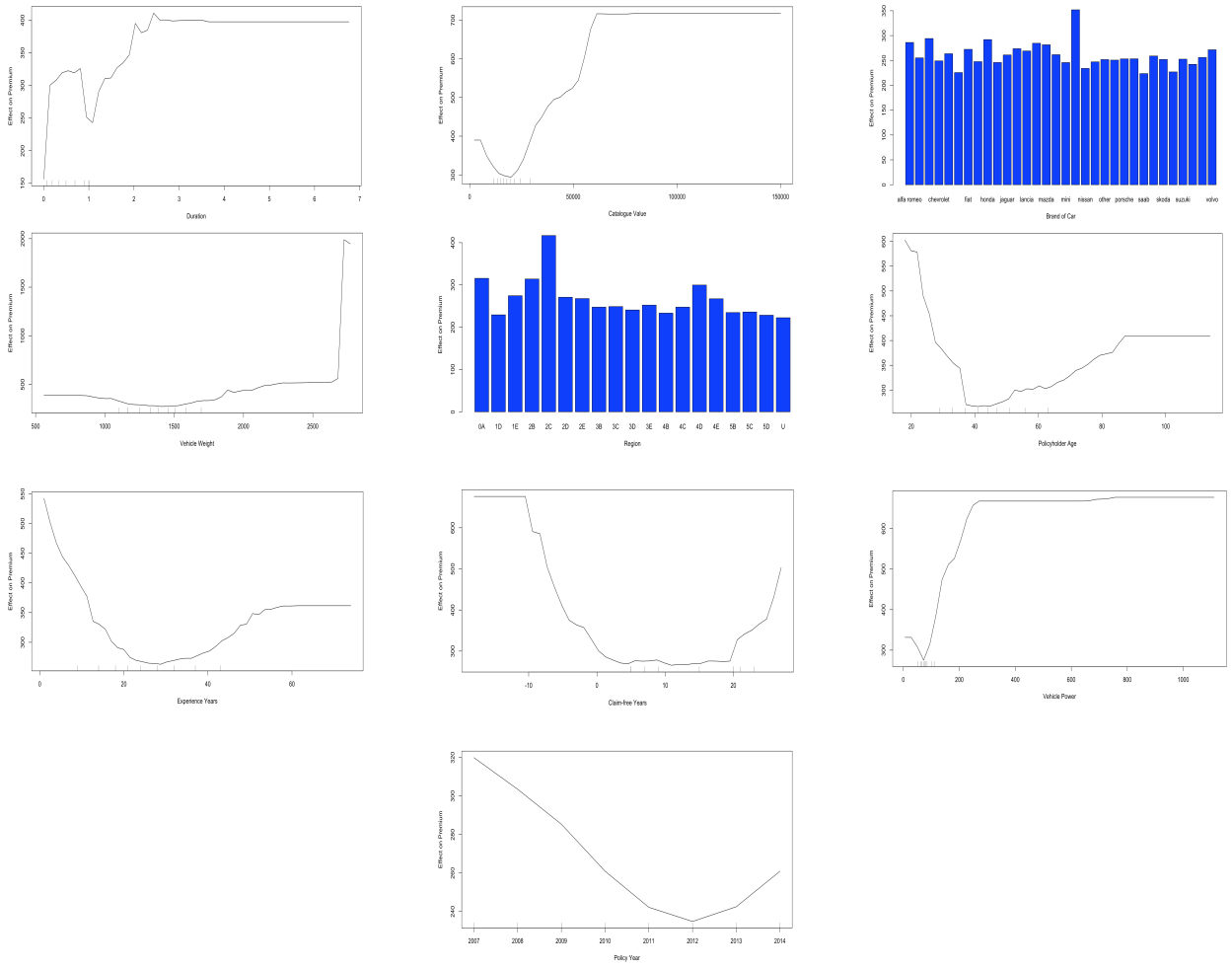


Figure A.8: Marginal effects of the different covariates on the claim size or premium. The ten most important variables of the Random Forest model are illustrated.

A.7 Setup RStudio server on GCC

We followed the following steps to setup a RStudio server on Google Cloud Computing (GCC) with a Macbook.

1. Create a free account and project with GCC.
2. Download the proper tar.gz file from https://cloud.google.com/sdk/docs/quickstart-mac-os-x.
3. Open up the Terminal on your MAC. The following type instructions are all in the terminal. We will now Setup the VM:
   - Type *./google-cloud-sdk/install.sh* .
   - Type *gcloud init* .
   - If command gcloud not found, type *source '[INSERT PROPER PATH]/google-cloud-sdk/completion.bash.inc'* , followed by *source '[INSERT PROPER PATH]/google-cloud-sdk/path.bash.inc'* and then retype the previous item. It will navigate to the proper project.
   - Create a 8-core 52GB RAM Virtual Machine by typing *sudo gcloud compute instances create rstudio --image-family ubuntu-1604-lts --image-project ubuntu-os-cloud --machine-type n1-highmem-8 --zone europe-west1-b* .
   - Allow the Rstudio Server to run on the 8787 port of the external IP by typing *sudo gcloud compute firewall-rules create allow-rstudio --allow=tcp:8787* .
   - Log in via SSH by typing *sudo gcloud compute ssh rstudio --zone europe-west1-b* .
4. The terminal will now navigate to the root of the VM. Next we install R on the VM:
   - Type *sh -c 'echo "deb https://cloud.r-project.org/bin/linux/ubuntu xenial/" >> /etc/apt/sources.list'* .
   - Followed by *apt-key adv --keyserver keyserver.ubuntu.com --recv-keys E084DAB9* .
   - And *apt update* .
   - Then *apt install r-base r-base-dev* .
   - Now we will install some important settings, so type: *apt install libcurl4-openssl-dev libssl-dev libxml2-dev* .
   - Followed by *add-apt-repository -y ppa:ubuntugis/ubuntugis-unstable* .
   - Then *apt update && apt upgrade* .
   - And finally *apt install libgeos-dev libproj-dev libgdal-dev* .
5. R is now installed on the VM. Next we add a user and we are immediately asked to enter a password so we can login to the RStudio server by typing *adduser olivier* .
6. To find out the external IP we have to type into the web browser, we open a new Terminal window (do not close the root window) and type *sudo gcloud compute instances describe rstudio –zone europe-west1-b* .
7. Before we use the RStudio server, we need to install Java. Navigate back to the root Terminal window and type:
   - *sudo apt-get update* .

- *java -version*. You will notice no Java version is installed properly.
- *sudo apt-get install default-jre*.
- *sudo apt-get install default-jdk*.
- *sudo apt-get install oracle-java8-installer*.
- *R CMD javareconf*.
- Finally we can check the proper installation of Java by typing *java -version*.

8. Navigate to http://EXTERNAL-IP-ADDRESS:8787, login with the added user and enter your password.
9. In a new Terminal Window you can stop and restart the VM instance by typing either *sudo gcloud compute instances stop rstudio* or *sudo gcloud compute instances start rstudio*.

A.8 BART

k	nu	q	num_trees	oos_error	% diff with lowest
5	3	0.99	200	1393.977	0.00000000
5	3	0.90	200	1395.039	0.07618018
5	10	0.75	200	1396.473	0.17903345
5	3	0.90	100	1397.013	0.21774965
5	10	0.75	100	1399.630	0.40553325
5	3	0.99	100	1400.400	0.46073669
3	3	0.99	200	1427.527	2.40676458
3	10	0.75	200	1428.741	2.49388199
3	3	0.90	200	1433.842	2.85975577
3	3	0.90	100	1442.292	3.46594695
3	10	0.75	100	1446.977	3.80208438
3	3	0.99	100	1452.308	4.18445616
2	3	0.90	200	1461.347	4.83293605
2	10	0.75	100	1465.534	5.13331509
2	3	0.99	200	1467.220	5.25422929
2	10	0.75	200	1470.484	5.48836476
1	10	0.75	100	1480.569	6.21184411
2	3	0.99	100	1488.280	6.76498747
2	3	0.90	100	1489.568	6.85743449
1	10	0.75	200	1494.738	7.22829292
1	3	0.90	200	1518.208	8.91196796
1	3	0.90	100	1564.639	12.24276835
1	3	0.99	200	1564.880	12.26010614
1	3	0.99	100	1611.893	15.63266182

Figure A.9: Results of the 5-fold cross-validation of the BART model. Best model is chosen by the lowest rmse on the hold-out folds (oos).

A.9 Hierarchical Model

Parameter	Covariate
α	Intercept
$\beta_{i,1}$	Policyholder Age
$\beta_{i,2}$	Catalogue Value
$\beta_{i,3}$	Vehicle Weight
$\beta_{i,4}$	Experience Years
$\beta_{i,5}$	Claim-free Years
$\beta_{i,6}$	Vehicle Power
$\beta_{i,7}$	Car Age
$\beta_{i,8}$	Mileage
$\beta_{i,9}$	Reference year

Table A.7: Covariates in HM.

Parameter	mean	sd	2.50%	25%	50%	75%	97.50%
α^S	7.34	4.55E-02	7.26	7.31	7.34	7.37	7.43
β_1^S	2.76E-02	3.14E-02	-3.16E-02	4.45E-03	2.74E-02	4.91E-02	8.83E-02
β_2^S	1.84E-02	3.02E-02	-3.96E-02	-2.27E-03	1.77E-02	3.76E-02	7.44E-02
β_3^S	-3.24E-02	2.05E-02	-7.58E-02	-4.63E-02	-3.25E-02	-1.82E-02	6.46E-03
β_4^S	-4.81E-02	3.22E-02	-1.11E-01	-6.92E-02	-4.85E-02	-2.58E-02	8.92E-03
β_5^S	-9.62E-02	1.34E-02	-1.21E-01	-1.06E-01	-9.64E-02	-8.66E-02	-7.06E-02
β_6^S	1.62E-01	3.34E-02	9.77E-02	1.39E-01	1.62E-01	1.84E-01	2.31E-01
β_7^S	-7.28E-02	1.28E-02	-9.69E-02	-8.18E-02	-7.28E-02	-6.49E-02	-4.67E-02
β_8^S	4.40E-03	2.41E-02	-4.28E-02	-1.06E-02	3.83E-03	2.16E-02	4.95E-02
β_9^S	5.22E-02	6.19E-03	4.01E-02	4.82E-02	5.21E-02	5.64E-02	6.39E-02
ν	7.79E-01	9.56E-03	7.61E-01	7.72E-01	7.80E-01	7.85E-01	7.98E-01
α^F	-1.94	3.83E-02	-2.01E+00	-1.96	-1.94	-1.92	-1.86
β_1^F	4.24E-02	2.42E-02	-8.19E-03	2.67E-02	4.15E-02	5.70E-02	9.38E-02
β_2^F	3.69E-02	2.19E-02	-5.47E-03	2.18E-02	3.57E-02	5.07E-02	8.11E-02
β_3^F	2.79E-02	1.49E-02	-3.42E-03	1.90E-02	2.86E-02	3.77E-02	5.48E-02
β_4^F	-1.25E-01	2.56E-02	-1.75E-01	-1.40E-01	-1.24E-01	-1.08E-01	-7.40E-02
β_5^F	-4.59E-02	1.14E-02	-6.71E-02	-5.31E-02	-4.59E-02	-3.84E-02	-2.28E-02
β_6^F	8.70E-03	2.43E-02	-4.49E-02	-6.81E-03	9.59E-03	2.51E-02	5.44E-02
β_7^F	-1.07E-01	1.03E-02	-1.26E-01	-1.14E-01	-1.07E-01	-9.89E-02	-8.66E-02
β_8^F	3.02E-01	1.99E-02	2.62E-01	2.90E-01	3.03E-01	3.16E-01	3.39E-01
β_9^F	-4.31E-02	5.08E-03	-5.22E-02	-4.68E-02	-4.34E-02	-3.96E-02	-3.32E-02
deviance	2.54E+05	6.66	2.54E+05	2.54E+05	2.54E+05	2.54E+05	2.54E+05

Table A.8: Summary of HM fit.

Figure A.11: Density (left top), autocorrelation - (right top), moving mean - (right middle) and trace plots (bottom) for the α^F (intercept_f) and β^F parameters of the HM for the number of claims (F).

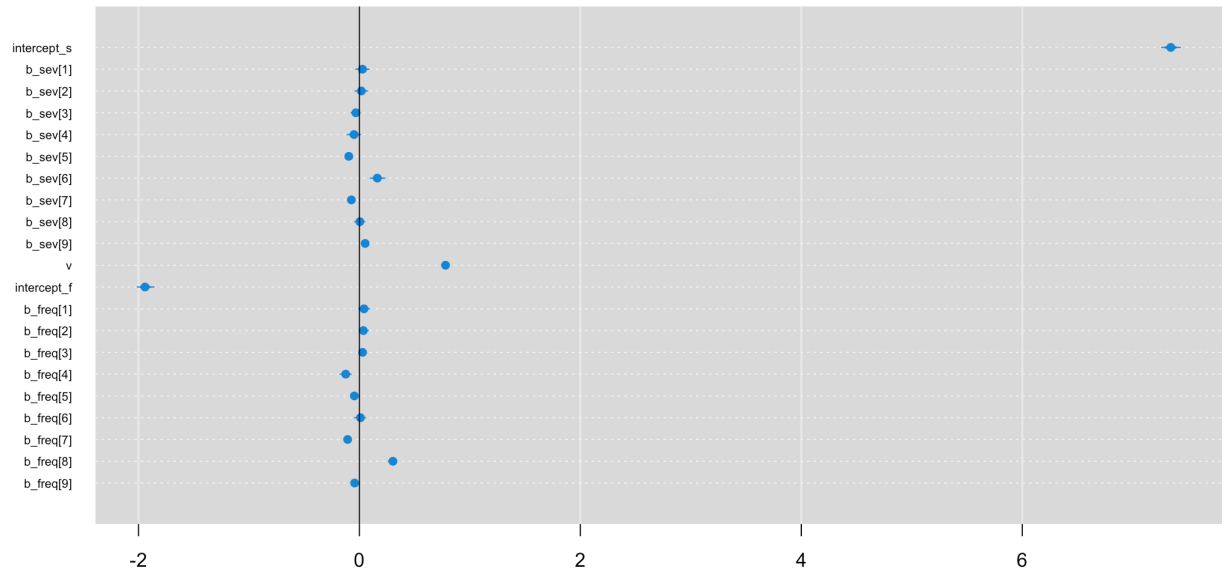


Figure A.10: Visual summary of the HM fit (parameter estimates).

Figure A.12: Density (left top), autocorrelation - (right top), moving mean - (right middle) and trace plots (bottom) for the α^S (intercept_s), ν and β^S parameters of the HM for the claim severity (S).

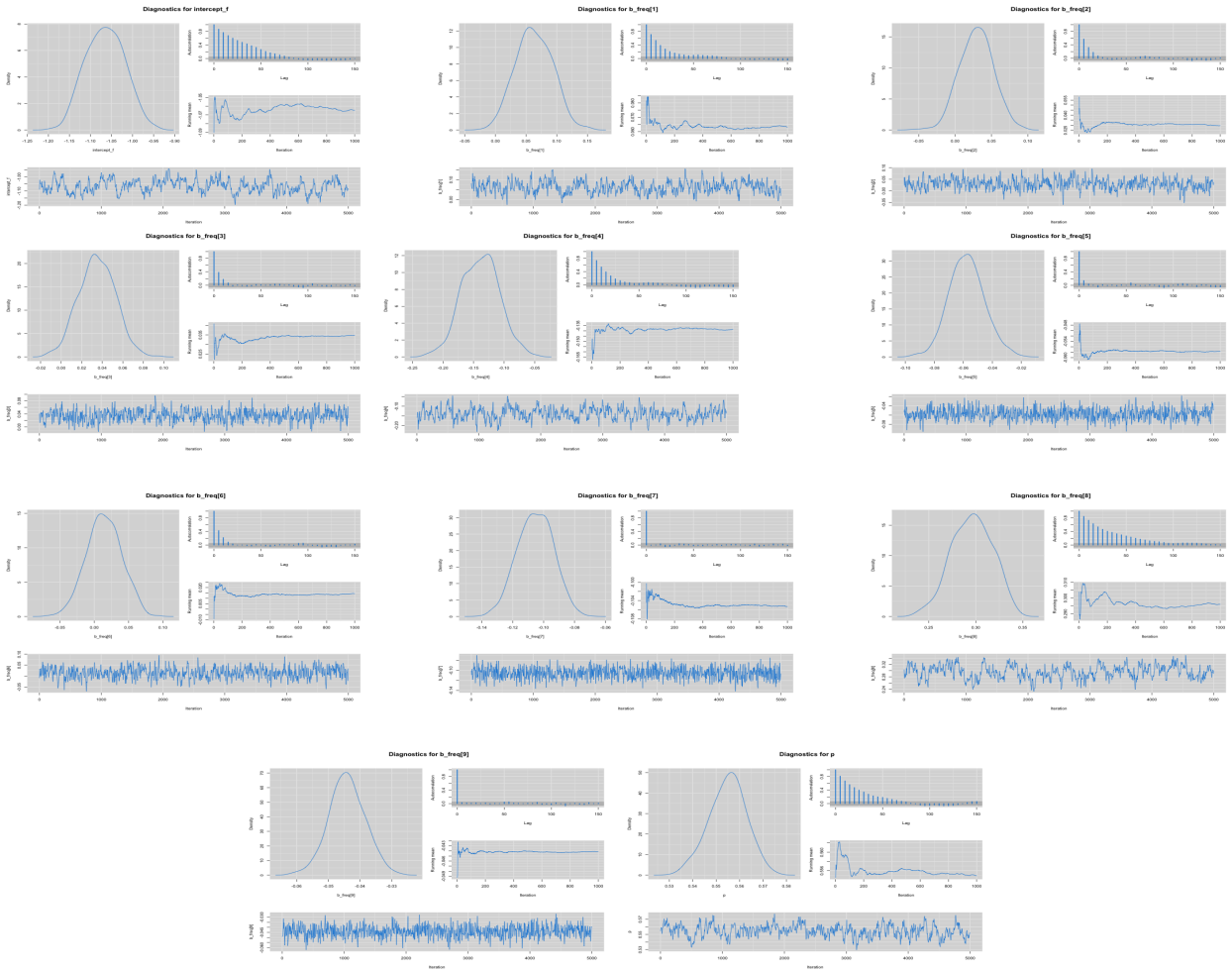


Figure A.13: Density (left top), autocorrelation - (right top), moving mean - (right middle) and trace plots (bottom) for the α^F (intercept_f), β^F and p parameters of the alternative HM for the number of claims (F).

Figure A.14: Density (left top), autocorrelation - (right top), moving mean - (right middle) and trace plots (bottom) for the α^S (intercept_s), ν and β^S parameters of the alternative HM for the claim severity (S).

Parameter	mean	sd	2.50%	25%	50%	75%	97.50%
α^S	7.35	4.48E-02	7.26	7.32	7.34	7.38	7.44
β_1^S	2.18E-02	3.21E-02	-3.92E-02	4.33E-04	2.10E-02	4.19E-02	8.67E-02
β_2^S	1.96E-02	2.92E-02	-3.79E-02	-3.67E-04	1.84E-02	4.07E-02	7.58E-02
β_3^S	-3.26E-02	2.10E-02	-7.36E-02	-4.61E-02	-3.16E-02	-1.83E-02	7.35E-03
β_4^S	-4.16E-02	3.37E-02	-1.09E-01	-6.41E-02	-4.11E-02	-1.84E-02	2.04E-02
β_5^S	-9.65E-02	1.36E-02	-1.23E-01	-1.06E-01	-9.67E-02	-8.68E-02	-7.07E-02
β_6^S	1.61E-01	3.13E-02	1.03E-01	1.39E-01	1.60E-01	1.82E-01	2.21E-01
β_7^S	-7.24E-02	1.32E-02	-9.88E-02	-8.16E-02	-7.22E-02	-6.34E-02	-4.72E-02
β_8^S	2.67E-03	2.31E-02	-4.82E-02	-1.21E-02	4.53E-03	1.80E-02	4.57E-02
β_9^S	5.18E-02	5.94E-03	3.95E-02	4.82E-02	5.19E-02	5.58E-02	6.28E-02
ν	7.80E-01	9.16E-03	7.62E-01	7.74E-01	7.81E-01	7.87E-01	7.98E-01
α^F	-1.06	4.55E-02	-1.15	-1.10	-1.06	-1.03	-9.80E-01
β_1^F	6.31E-02	2.99E-02	7.52E-03	4.25E-02	6.19E-02	8.45E-02	1.20E-01
β_2^F	2.94E-02	2.31E-02	-1.52E-02	1.38E-02	2.99E-02	4.46E-02	7.30E-02
β_3^F	3.48E-02	1.76E-02	8.52E-04	2.31E-02	3.47E-02	4.69E-02	6.95E-02
β_4^F	-1.39E-01	3.11E-02	-2.02E-01	-1.62E-01	-1.38E-01	-1.18E-01	-7.64E-02
β_5^F	-5.76E-02	1.22E-02	-8.06E-02	-6.59E-02	-5.76E-02	-4.96E-02	-3.19E-02
β_6^F	1.43E-02	2.54E-02	-3.63E-02	-2.36E-03	1.40E-02	3.19E-02	6.20E-02
β_7^F	-1.05E-01	1.14E-02	-1.28E-01	-1.13E-01	-1.05E-01	-9.72E-02	-8.38E-02
β_8^F	2.96E-01	2.16E-02	2.51E-01	2.80E-01	2.96E-01	3.12E-01	3.35E-01
β_9^F	-4.41E-02	5.70E-03	-5.55E-02	-4.79E-02	-4.42E-02	-4.04E-02	-3.31E-02
p	5.55E-01	7.96E-03	5.38E-01	5.50E-01	5.55E-01	5.60E-01	5.70E-01
deviance	2.32E+05	4.65E+02	2.31E+05	2.32E+05	2.32E+05	2.32E+05	2.33E+05

Table A.9: Summary of alternative HM fit.

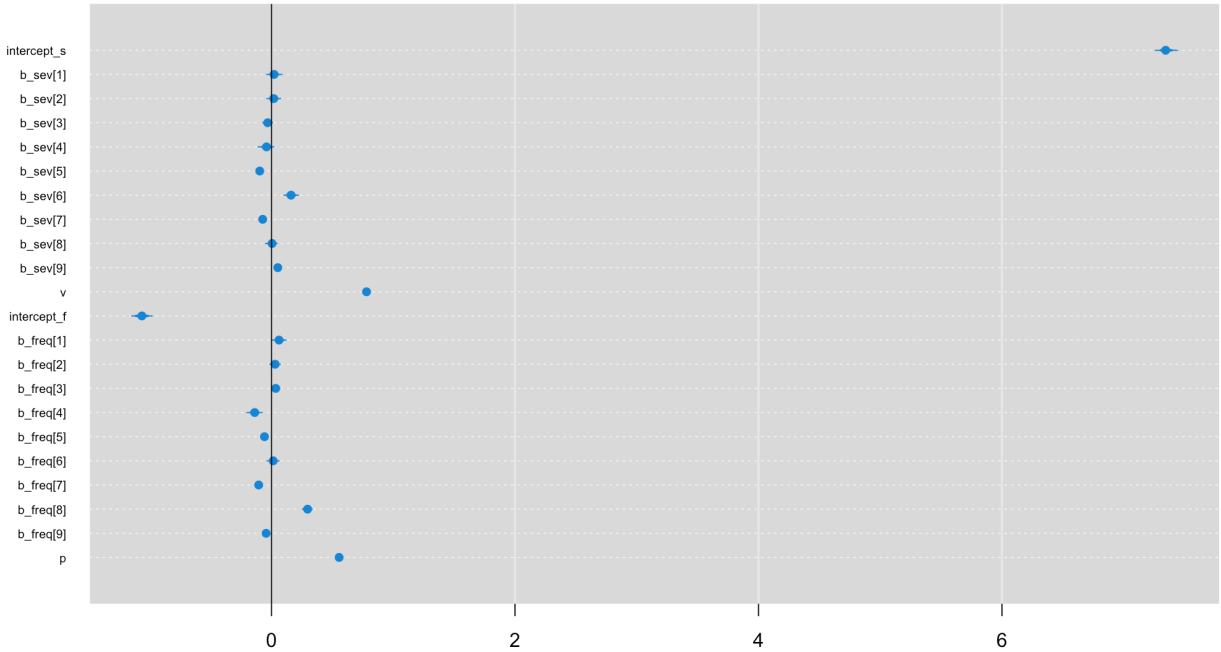


Figure A.15: Visual summary of the alternative HM fit (parameter estimates).

	$E[c]$	$E[s\|c>0]$	$E[\rho]$
Policy 1	0.414	1 498	620
Policy 39351	0.04	5 286	211
Policy 31244	1.83	1 162	2 123
Policy 33614	1.44	19 103	13 229

Table A.10: Estimated number of claims, claim severity and premiums of some extracted policies of the test set.

	GLM	GBM	RF sep	RF dir	HM mean
Earned - Paid	$-3.0 * 10^6$	$-7.3 * 10^5$	$1.9 * 10^6$	$2.2 * 10^6$	$-2.8 * 10^6$

Table A.11: Final values for the Earned premiums minus the paid claims during the holdout year, part one.

	HM 75% ub	HM 80% ub	HM msd
Earned - Paid	$-4.1 * 10^6$	$-8.8 * 10^5$	$2.4 * 10^5$

Table A.12: Final values for the Earned premiums minus the paid claims during the holdout year, part two.

Bibliography

[1] Valdez E.A. Antonio K. "Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance". In: *Advances in Statistical Analysis* 96.2 (2012), pp. 187–224.

[2] Kannan R. Polson N.G. Applegate D. *Random Polynomial Time Algorithms for Sampling from Joint Distributions.* Tech. rep. Pittsburgh, Pennsylvania: Carneggie-Mellon University, 1990.

[3] Simon L.R.J. Bailey R.A. "An actuarial note on the credibility of experience of a single private passenger car". In: *PCAS* XLVI.1 (1959), pp. 41–48.

[4] Simon L.R.J. Bailey R.A. "Two studies in Automobile Insurance Ratemaking". In: *ASTIN Bull.* 1.4 (1960), pp. 192–217.

[5] Coutts S.M. Ross G.A.F. Baxter L.A. "Applications of Linear Models in Motor Insurance". In: *Proceedings of the 21st International Congress of Actuaries*. Vol. 2. 1980, pp. 11–29.

[6] Pentikäinen T. Pesonen E. Beard R.E. *Risk Theory*. London: Chapman & Hall, 1984.

[7] Jackman R.W. Bollen K.A. "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases". In: *Modern Methods of Data Analysis* (1990), pp. 257–291.

[8] L. Breiman. *Random Forests*. Tech. rep. Berkeley: University of California, 2001.

[9] Wrightt T.S. Brockman M.J. "Statistical motor rating: making efficient use of your data". In: *Journal of Institute of Actuaries* 119 (1992), pp. 457–543.

[10] H. Bühlmann. "Experience Rating and Credibility". In: *ASTIN Bull.* 4.3 (1967), pp. 199–207.

[11] George E.I. McCulloch R.E. Chipman H.A. "BART: Bayesian Additive Regression Trees". In: *The Annals of Applied Statistics* 4.1 (2010), pp. 266–298.

[12] Weisberg S. Cook R. D. *Residuals and Influence in Regression*. New York: Chapman & Hall, 1982.

[13] H. Cramér. "On the Mathematical Theory of Risk." In: Skandia Jubilee (1930).

[14] M. David. "A review of theoretical concepts and empirical literature of non-life insurance pricing". In: *7th International Congress on Globalization and Higher Education in Economics and Business Administration*. 2013, p. 314.

[15] J.H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine*. IMS Reitz Lecture. 1999.

[16] J.H. Friedman. "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1 (1991), pp. 1–141.

[17] Carlin J. Stern H. Rubin D. Gelman A. *Bayesian Data Analysis*. 2nd ed. Boca Raton, FL: Chapman & Hall, 2004.

[18] Rubin D.B. Gelman A. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–511.

[19] A. Gut. *An Intermediate Course in Probability*. Berlin: Springer, 1995.

[20] Renshaw A.E. Haberman S. "Generalized Linear Models and Actuarial Science". In: *Journal of the Royal Statistical Society* D (1996).

[21] Tibshirani R. Friedman J. Hastie T. *The Elements of Statistical Learning*. Stanford, California: Springer, 2008.

[22] B. Jörgensen. “Exponential dispersion models”. In: *J.R. Statist. Soc. A. Ser. B* 49 (1987), pp. 127–162.

[23] J. Jung. “On automobile insurance rating”. In: *ASTIN Bull.* 5.1 (1968), pp. 41–48.

[24] Rowe D. Richardson J. Elith J. Hastie T. Leathwick J.R. “Using multivariate adaptive regression splines to predict the distributions of New Zealand’s freshwater diadromous fish”. In: *Freshwater Biology* 50 (2005), pp. 2034–2052.

[25] E.F. Lundberg. *Approximations of the probability function. Reinsurance of collective risks.* Tech. rep. University of Uppsala, 1903.

[26] T. Mikosch. *Non-Life Insurance Mathematics*. Berlin: Springer, 2009.

[27] McCullagh P. Nelder J.A. *Generalized Linear Models*. London and New York: Chapman and Hall, 1983.

[28] Wedderburn R.W.M. Nelder J.A. “Generalized linear models”. In: *Journal of the Royal Statistical Society* A.135 (1972), pp. 370–384.

[29] Lewis S.M. Raftery A.E. “Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo”. In: *Statistical Science* 7.4 (1992), pp. 493–497.

[30] A.E. Renshaw. *Modelling the claims process in the presence of covariates*. Tech. rep. London: The City University, 1994.

[31] M.J. Goovaerts R.J.A. Laeven. *Premium Calculation and Insurance Pricing*. Tech. rep. Tilburg: Tilburg University, 2007.

[32] Freund Y. Schapire R.E. “Experiments with a New Boosting Algorithm”. In: *Machine Learning: Proceedings of the Thirteenth International Conference*. 1996, pp. 148–156.

[33] Best N.G. Carlin B.P. Van der Linde A. Spiegelhalter D.J. “Bayesian measures of model complexity and fits (with discussion)”. In: *Royal Statistical Society* B.64 (2002), pp. 583–639.

[34] H.E. Wittick. “The Canadian merit rating plan for individual automobile risks”. In: *PCAS* XLV.49 (1959), pp. 214–220.