FACULTY OF ECONOMICS
AND BUSINESS

KU LEUVEN

Insurance pricing in the era of machine learning and telematics technology

Dissertation presented to
obtain the degree of
Doctor in Business Economics
by

Roel HENCKAERTS

Serial Number 760 2021

Committee

Advisor:	
Prof. Dr. Katrien Antonio	*KU Leuven and University of Amsterdam*
Chair:	
Prof. Dr. Dieter Smeulders	*KU Leuven*
Members:	
Prof. Dr. Bart Baesens	*KU Leuven and University of Southampton*
Prof. Dr. Jan De Spiegeleer	*KU Leuven*
Prof. Dr. Jan Dhaene	*KU Leuven*
Prof. Dr. Marie-Pier Côté	*Université Laval*
Prof. Dr. Xavier Milhaud	*Aix-Marseille Université*

Daar de proefschriften in de reeks van de Faculteit Economie en Bedrijfswetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

Acknowledgements

Five years ago, I embarked on a journey to obtain a PhD. This thesis marks the end of said journey. The road to get to this point was usually bright and full of achievements, sometimes dark and full of obstacles, but always very exciting and definitely worth the travel. And as usually is the case, none of these achievements were earned alone and none of these obstacles were overcome alone. I therefore want to take some time to thank everyone who joined me on this journey and thereby supported me in getting to the finish line.

Before thinking about reaching the finish, you need to get to the starting point. And as most PhDs do, mine also started with a supervisor and a topic. Katrien, I can not express how grateful I am that you were my supervisor. You provided me with interesting topics to work on, but you did so much more. Over the past years, I have had the pleasure to learn a lot on many fronts, meet amazing people and discover new locations all around the world, explore my own interests and get to know myself better. None of this would have been possible without you, your enthusiasm and your tremendous support. You were always there to accommodate me in the best way possible on every aspect of the job. Whenever I needed guidance on a research project, you were always ready to listen, think with me and provide directions towards a solution. You also empowered me to step out of the purely academic research setting by working together with companies on projects and deliver in-company courses. I will never forget how I was able to go to conferences at exotic locations like New York, Australia and Brazil thanks to you. So long story short, I could not have wished for a better supervisor, thank you so very much!

I would like to thank the members of my internal supervisory committee at KU Leuven, namely Bart Baesens, Jan De Spiegeleer and Jan Dhaene. Thank you for your interesting questions and for sharing your expertise during the doctoral seminars and preliminary defense. Your suggestions made me think out-of-the-box about certain aspects and helped to improve the final quality of this thesis. I also would like to thank the external members of my examination committee, namely Marie-Pier Côté and Xavier Milhaud. Thank you for reviewing this whole thesis and bringing new insights, ideas and challenging questions to the preliminary defense. Your constructive feedback allowed me to add another layer of improvement to the thesis. And to Marie-Pier, an extra big thanks for the nice and fruitful collaboration over the past years. You did not only bring a lot of expertise and interesting ideas to the table, but I also really enjoyed working together with you. Finally, I would like to thank the chair Dieter Smeulders for accepting this role and helping to run my public defense in a smooth way. (Hopefully.)

I am very grateful to the Research Foundation - Flanders (FWO) for awarding me with a PhD Fellowship strategic basic research (SB). This did not only financially support this doctorate, but also allowed me to tackle an extra project at the end of my PhD. Furthermore, I would like to thank all private insurers and associated practitioners that I interacted with during the past years in the context of Research Chairs or projects. Thank you for all the interesting problems, educational meetings, business feedback and for helping me bridge the gap between purely academic research and practical business implementations.

People often think that you are still studying when doing a PhD, as you are technically a PhD student. Even though you are learning every day, I would still call it working instead of studying. Now there is one key ingredient that every fun work environment needs, a group of nice colleagues. I feel extremely lucky to have met so many amazing people at the Faculty over the years.

Thank you to all my colleagues from the insurance research group. Roel, you were a great Master thesis supervisor and co-author on later projects. I learned a lot from you, from working methodologically to R coding best practices. Thank you Liivika and Sander for being such great office mates at the start of my PhD. Liivika, I will never forget your quirky sense of humor which always made me laugh and how fun it was to attend your defense in Estonia after which we had an amazing trip there. Also thanks to Roel for this, and another thanks to the both of you for hosting me in Sydney. Sander, thank you for all your guidance, for always being there to listen to research problems and brainstorm solutions and for our interesting chats on life in general. Jonas, thank you for being a great office mate and fill the void left behind by Liivika. You were always up for a nice chat and ready with solutions for every problem that I encountered. Your delicious desserts and baked goods that you brought to the office were a very welcome bonus to all of that. Thank you Hamza and Karim, it was always fun to just enter your office for a fascinating chat on research or anything else. Thank you María and Tom, you were part of our lab for a (too) short postdoc period, but we still had an abundance of fun and interesting times together. Thank you Eva, Bavo and Jens for the nice office and hallway talks, for our fun lunches in the park and other activities to relax outside of work. The future of insurance relies on you, but I am sure that you are all going to excel!

Thank you to all my colleagues from AFI. The soup breaks, lunch gatherings, common room meetings, after-work drinks and after-after-work parties were always very fun. The yearly AFI weekends were amazing experiences where we could unwind as friends instead of just being colleagues. Thank you Ines, Anne-Florence, Charlotte, Elizaveta, Nela, Liliana, Mathilde, Hien, Phoung, Tineke, Viola, Claudia, Sander, Mathijs, Alvaro, Nejat, Jeffrey, Kristof, Nick, Ditmir, Benjamin, Victor and Ivan for sharing all these amazing moments and memories with me. Thank you Mieke and Joachim for offering me refuge at your office while they were repainting ours, it was nice to step out of the "dark" and experience the "light" side of the first floor for a while. Mieke, thank you for all the amazing and endless chats over tea. Joachim, thanks for all the random facts and additions to these chats. Thank you Carola, Karen and Maxim for being the heart of our little PhD-techno-group and for all the amazing parties, raves and festivals that we did (and will do) together. Not to forget the mind-blowing experience of celebrating carnival in Rio de Janeiro. Thank you Lars, Mathias and Dieter for being such cool guys, for all the great times and the awesome weekends away. I am sure that more will follow in the future! Lars, you make a mean cocktail and super tasty food, but you can't tell directions at all. Mathias, I enjoy cracking jokes with you as well as our intelligent discussion on mindless tv shows like temptation island. Dieter, I love all the experiences that we shared and definitely will share in the future. Even though Mol feels like the end of the world, it is always very much worth the trip. And I truly am sorry for Cologne, but it was Lars' fault (see supra).

Thank you to the colleagues from MSI. I immensely enjoyed the evenings/nights at Stapleton's and the challenges we took on, such as the hot-wings-challenge. Not to forget the poker games and house parties with beer pong and punch bowls full of that amazing drink that I keep forgetting the name of (Karibalu?). The Bro weekends topped this all off with more exciting challenges, legendary nights, cool activities and memories to cherish. Thank you Naza, Kieran, Philippe, Jeroen, Mathias, Sven, Michael, Dennis, Maarten, Schaper, Steven and Adrían. The PhD journey would not have been the same without you.

Flashback to my very first day at the office. Apparently I picked an unfortunate day to start, because the whole insurance group was away at a conference in Lyon. Around noon, a finance colleague named José knocked on my door, asking if I was joining the other AFI colleagues for lunch. We immediately hit it off, went for after-work drinks, and became close friends over time. José, thanks for being an amazing friend and your contribution in shaping my PhD experience. I have wonderful memories of bbq's and parties at your terrace and of our trips to Estonia, Croatia and Madrid. Thanks for always being there and having my back. You are a true friend and a true brother. I'm curious to see what the future holds, but I am sure that it will be bright. Thanks for everything dude!

Finally, I want to thank my family. I am extremely lucky for growing up in a warm family that enjoys each other's company. Our family feasts are always very fun, with a lot of laughing involved, and the traditional shouting during the closing session of cards at the end of the evening. Thank you to all my aunts, uncles and cousins for providing such a comfortable environment to grow up. I also want to thank my grandparents, but will continue to do so in Dutch. Dank u oma en opa! Dank u om zo een belangrijk deel uit te maken van mijn leven sinds ik klein was. Om mij als kind ‘s ochtends naar school te brengen, ‘s middags te voorzien van lekker eten en ‘s avonds te komen halen en al wat huiswerk mee te overlopen. Dank u voor alles wat jullie voor mij hebben gedaan en om er nog steeds voor mij te zijn. Verder wil ik ook Tata en Nono bedanken. Als kind keek ik enorm uit naar de vrijdagavonden dat ik naar jullie mocht komen, bedankt voor alles! Tom, my brother, thank you for being such a great guy. As kids we maybe used to fight from time to time, but I really am happy that you are my brother and enjoy our moments together like our trips in Barcelona. Know that I am extremely proud of you and your achievements! Last but certainly not least, there are two persons who deserve the biggest praise of all. I talked earlier about the starting point of this journey, but the true starting point lies of course much further in the past than five years ago. The true starting point lies with my parents almost 30 years in the past. Dear mom and dad, words can't describe how grateful I am for everything that you have done for me during my whole lifetime. You were always there for me, empowering me when I was doing good and picking me up when it was the other way around. You gave me all the opportunities in life which led up to this point. Without your support, I could not even dream of being here. Thank you for being the best parents I can imagine. Thank you, thank you a million times!

To anyone who I forgot to mention, whether you influenced this PhD thesis or not, thank you very much for being part of my life over the years!

Roel Henckaerts Leuven, October 2021.

Contents

Chapter 1

Introduction

Insurance is a risk management strategy to protect an entity against uncertain future financial losses. Property and casualty (P&C) insurance covers one's belongings and liability, for example a person's house or responsibility when crashing into someone else's car. An individual might not be able to carry the financial burden of such unfortunate events, which is why our society relies heavily on insurance. Insurers pool many individuals, all exposed to similar risks, such that *the contributions of the many cover the misfortunes of the few.*

Figure 1.1 shows the risk transfer process from policyholders who buy insurance to insurers who sell insurance. A policy contract stipulates the specific conditions which trigger a compensation from insurer to policyholder. In return, policyholders are required to pay a predetermined premium at the start of the policy period. From the policyholder's point of view, this transforms uncertain financial losses into a certain upfront cost. This allows individuals to live their life without constantly worrying about incurring devastating financial consequences, while of course still avoiding unnecessary risks and moral hazard.

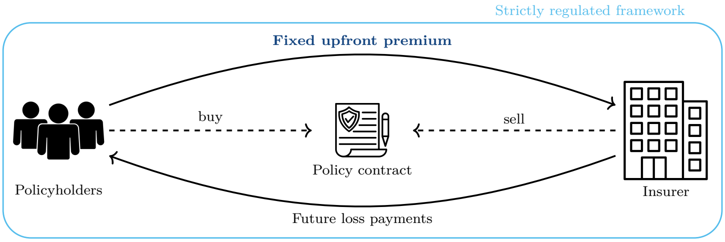


Figure 1.1: The insurance risk transfer process from policyholder to insurer.

The insurer receives a fixed premium and promises to pay future losses to the policyholders. These future payments are uncertain such that the insurer sells a product of which the cost is unknown at the moment of selling. This is known as the *inverse production cycle* and makes it of vital importance for the insurer to properly assess the policyholders' claim risk based on the available information.

Predictive modeling is the discipline of developing a mathematical model on historical data in order to predict the future. Insurers heavily rely on such tools to predict the future number of claims (*frequency*) and claim sizes (*severity*). Insurance pricing combines both components and leads to the *pure premium*, i.e., the premium needed to purely cover the policyholder's underlying claim risk. Heavy competition and anti-selection effects require insurers to constantly improve their risk classification process with more accurate predictive models.

Two components are needed to develop a predictive model: a modeling framework and historical data. In a classical setting, insurers focus on specific statistical models named generalized linear models (GLMs). These types of models are the industry standard due to many advantages such as easy interpretation and simple implementation. Classical data typically comprises a collection of self-reported risk characteristics such as the policyholder's age and residence area, supplemented with vehicle characteristics in motor insurance.

Technological advancements allow for innovations on both the model and data components. Machine learning (ML) algorithms are gaining popularity in many predictive modeling applications due to improved performance results. This however leads to more opaque decision models, which makes interpretation and implementation more difficult in practice. The insurance industry therefore remains reluctant to use ML for pricing due to strict regulations on model explainability. Regarding new data sources, telematics technology allows to measure policyholder behavior on a granular scale. This results in new types of information to use in the risk classification process, for example driving behavior in motor insurance or data from smart sensors in home insurance.

This thesis comprises four chapters. In the first three chapters we focus on new modeling paradigms within the insurance industry. We investigate ML approaches to insurance pricing, thereby also focusing on interpretability issues by *opening the black box*. Furthermore, we work on general data-driven procedures to develop a GLM when starting from more flexible and complicated models. The knowledge extracted from these complex models can help actuaries to design a simple yet accurate GLM. The fourth and last chapter puts focus on a new data paradigm with telematics technology. A baseline premium is suggested by using self-reported risk characteristics and driving behavior data is used to update those prices. This leads to a usage-based insurance (UBI) system where policyholders can earn rebates by driving less and more safely.

1.1 Machine learning: innovating insurance models

The first three chapters of this thesis focus on new modeling paradigms for insurance pricing with the use of ML techniques. Figure 1.2 shows a visual summary of these three chapters and their relation to each other.

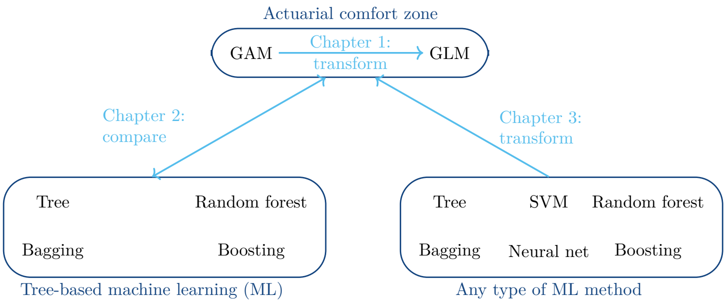


Figure 1.2: Overview of the first three chapters in this thesis.

Chapter 2 stays within the actuarial comfort zone of well-known statistical models, namely generalized additive models (GAMs) and GLMs. GAMs extend the framework of GLMs by allowing to include smooth effects of risk factors. Our procedure starts from a flexible GAM with smooth effects for continuous and spatial risk factors. We then use these smooth effects to construct insurance tariff classes in a data-driven way via decision trees and clustering techniques. In the end we obtain a GLM with all risk factors in a categorical format, easy to explain and easy to be implemented in a practical business environment.

Chapter 3 breaks out of the actuarial comfort zone and into the world of tree-based ML techniques. We compare the GAM/GLM from Chapter 2 with decision trees, random forests and gradient boosting machines (GBMs). We stress the importance of using proper loss functions in line with distributional assumptions for insurance frequency and severity data. Our comparison goes from pure statistical out-of-sample performance to a more managerial evaluation with lift measures. We furthermore put focus on model interpretability by looking under the hood of our ML models and show how to discover interesting interaction effects in the data. In our case study, GBMs outperform the GAM/GLM approach on both statistical and managerial measures.

Chapter 4 builds on insights from Chapter 3 with the goal of returning to the actuarial comfort zone. GBMs typically have superior predictive performance over GAMs/GLMs, but at the cost of model transparency. Insurance is a highly regulated industry and deals with high-stakes decisions on insurance coverage. We therefore start from a complex black box model and extract knowledge via model interpretation techniques. These insights are used to perform smart feature engineering and allow to fit a transparent global surrogate model which approximates the black box behavior. It is important to note that this procedure is model-agnostic, i.e., it can be applied to any model class. This approach can be seen as a more general and improved version of the procedure in Chapter 2.

1.2 Telematics: innovating insurance data sources

The fifth and last chapter of this thesis shifts focus to new data paradigms for insurance pricing with the use of telematics technology. Figure 1.2 shows an overview of possible data features for motor insurance. In the classic setting, insurers use static demographic information on the policyholder, vehicle and area of residence. Telematics allows to track driving behavior of policyholders, typically split into two broad categories. Firstly we measure driving habits or *pay as you drive* (PAYD) features. This contains information on for example the total mileage and distances driven on different road types and during distinct times of day. Secondly we measure driving style or *pay how you drive* (PHYD) features. This includes information on for example speeding, acceleration and braking events, possibly combined with external weather data.

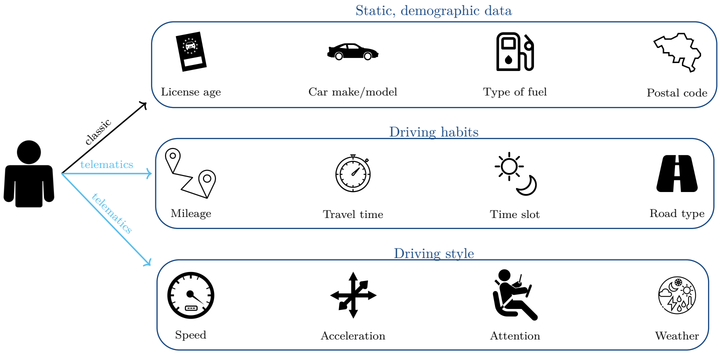


Figure 1.3: Overview of different possible data features for motor insurance.

Chapter 5 investigates the added value of telematics information for insurance pricing. We start by developing baseline pricing models with only classical self-reported risk characteristics. In a next step we update those prices with the driving behavior registered during the policy period. This results in a UBI product framework in which policyholders can directly influence their premium by adjusting their driving behavior. We show how telematics improves the risk classification process, reduces the policyholder's premiums on average and allows the insurer to realize higher profits.

1.3 Research contributions

The thesis chapters are based on the following publications and working papers:

(1) Henckaerts, R., Antonio, K., Clijsters, M., and Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, 2018(8):681–705

(2) Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021b). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285

(3) Henckaerts, R., Antonio, K., and Côté, M.-P. (2021a). When stakes are high: balancing accuracy and transparency with Model-Agnostic Interpretable Data-driven suRRogates. *arXiv preprint arXiv:2007.06894*

(4) Henckaerts, R. and Antonio, K. (2021). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Working paper*

The author also contributed to the following working paper:

(i) Holvoet, F., Henckaerts, R., Antonio, K., and Gielis, S. (2021). Neural networks for non-life insurance pricing. *Working paper*

The author contributed to the R universe with the following packages:

Henckaerts, R. (2020). *distRforest: Distribution-based Random Forest*. R package version 1.0.0

Henckaerts, R. (2021). *maidrr: Model-Agnostic Interpretable Data-driven suRRogate*. R package version 1.0.0

Chapter 2

A data-driven binning strategy to construct tariff classes

We present a fully data-driven strategy to incorporate continuous risk factors and geographical information in an insurance tariff. A framework is developed that aligns flexibility with the practical requirements of an insurance company, the policyholder and the regulator. Our strategy is illustrated with an example from property and casualty (P&C) insurance, namely a motor insurance case study. We start by fitting generalized additive models (GAMs) to the number of reported claims and their corresponding severity. These models allow for flexible statistical modeling in the presence of different types of risk factors: categorical, continuous and spatial risk factors. The goal is to bin the continuous and spatial risk factors such that categorical risk factors result which capture the effect of the covariate on the response in an accurate way, while being easy to use in a generalized linear model (GLM). This is in line with the requirement of an insurance company to construct a practical and interpretable tariff that can be explained easily to stakeholders. We propose to bin the spatial risk factor using Fisher's natural breaks algorithm and the continuous risk factors using evolutionary trees. GLMs are fitted to the claims data with the resulting categorical risk factors. We find that the resulting GLMs approximate the original GAMs closely, and lead to a very similar premium structure.

This chapter is based on joint work with Katrien Antonio, Maxime Clijsters and Roel Verbelen, as published in the Scandinavian Actuarial Journal (Henckaerts et al., 2018). We thank Michel Denuit for providing the data and for his valuable suggestions. Furthermore, we acknowledge financial support from the Ageas Research Chair at KU Leuven and from KU Leuven's research council (COMPACT C24/15/001).

2.1 Introduction

An insurance portfolio offers protection against a specified type of risk to a collection of policyholders with various risk profiles. Insurance companies differentiate premiums to reflect the heterogeneity of risks in their portfolio. A flat premium across the entire portfolio would encourage good risks to leave the company and accept a better offer elsewhere such that the insurer is left with bad risks which pay a too low premium. To avoid such lapses, insurance companies use risk factors (or: rating factors) to group policyholders with similar risk profiles in tariff classes. Premiums are equal for policyholders within the same tariff class and should reflect the inherent riskiness of each class. The process of constructing these tariff classes is also known as risk classification, see Denuit et al. (2007); Antonio and Valdez (2012); Paefgen et al. (2013). Pricing (or: ratemaking, tarification) through detailed risk classification is the mechanism for insurance companies to compete and to reduce the cost of insurance contracts. In a highly competitive market many rating factors are used to classify risks and to differentiate the price of an insurance product.

Property and casualty (P&C, or: non-life, general) insurance pricing typically makes use of categorical, continuous and spatial risk factors. Categorical risk factors have a discrete number of possible outcomes or levels. Examples of categorical risk factors for motor insurance are the type of coverage and type of fuel of the car. Continuous risk factors can attain all values within a specified range. Examples of continuous risk factors for motor insurance are the age of the policyholder and the horsepower of the car. A spatial risk factor contains information about the policyholder's residency. To capture the spatial heterogeneity one can for example use the postal code of the municipality where the policyholder resides as a rating factor. In motor insurance, this serves as a proxy for the region where a policyholder drives his car.

Constructing tariff classes is rather straightforward when all risk factors are categorical; each tariff class then represents a certain combination of levels of the categorical risk factors. The continuous and spatial risk factors can be interpreted as categorical factors with many levels, also called multi-level factors by Ohlsson and Johansson (2010). It is however inefficient to take all these levels into account separately since this will result in too many tariff classes with very few policyholders. A better approach is to transform the continuous and spatial risk factors with many levels in categorical risk factors with fewer levels, also called binning by Kuhn and Johnson (2013). In this chapter we present a data-driven strategy to bin continuous and spatial risk factors in order to obtain categorical risk factors with a limited number of levels. After this binning procedure it is again straightforward to construct the corresponding tariff classes.

Actuaries examine historical claims data to estimate the cost of offering the insurance cover, i.e. the premium, to policyholders in a specific tariff class. Insurance companies maintain large databases with policy(holder) characteristics and claim histories which enable the actuary to build risk-based pricing models. Actuarial models for P&C insurance pricing put focus on two components: a predictive model for the frequency of claims and a predictive model for the severity of claims (see Denuit et al., 2007; Frees et al., 2014; Parodi, 2014). Claim frequency refers to the number of claims per unit of exposure. Exposure, as described in McClenahan (2001), can be seen as a rating unit and measures to which degree the policyholder is exposed to the insured risk. An example of exposure in an insurance product is the fraction of the year for which premium has been paid and therefore coverage is provided. Severity is the average claim cost, expressed as the ratio of the total loss to the corresponding number of claims causing this total loss, over a specific period of insurance.

Frequency and severity are typically assumed to be independent and the resulting pure premiun (or: risk premium) is the product of the expected value of the frequency and the expected value of the severity (see Klugman et al., 2012). Alternatives for this independence assumption are investigated in the literature, allowing dependence between frequencies and severities (see Gschlößl and Czado, 2007; Czado et al., 2012). A risk margin taking model risk and pure randomness into account, as well as other premium elements (e.g. profit, commissions, taxes), is added on top of the pure premium to end up with a commercial tariff (see Wüthrich, 2016).

Generalized linear models (GLMs), developed by Nelder and Wedderburn (1972), have become the industry standard to develop predictive models for frequency and severity (see Haberman and Renshaw, 1997; Denuit et al., 2007; De Jong and Heller, 2008; Frees, 2015). GLMs allow the response variable to follow any distribution in the exponential family. The Poisson distribution is particularly interesting for claim frequency models whereas the gamma and lognormal distributions are often used for claim severity modeling. Covariates enter a GLM through a linear predictor, leading to interpretable effects of the risk factors on the response. Such a linear predictor is however less suited for continuous risk factors that relate to the response in a non-linear way, since transformations of the covariate are needed to capture a non-linear effect. Generalized additive models (GAMs), developed by Hastie and Tibshirani (1990), extend the framework of GLMs and allow for smooth continuous effects in the predictor structure. This results in a statistically more flexible model compared to the GLM. In practice however, actuaries tend to prefer the simplicity of GLMs with categorical risk factors over GAMs with smooth effects, because pricing models should be interpretable, intuitive, explainable to clients and regulators, easy to program and adjustable to marketing needs and benchmark studies with competitors. Therefore our contribution designs a strategy to construct tariff classes in GLMs in a data-driven way.

This chapter should be framed in between two existing approaches to handle different types of risk factors in the literature on insurance pricing. One strand of literature uses predefined bins for the continuous and spatial risk factors (see Frees and Valdez, 2008; Antonio et al., 2010). These bins, which are constructed without much motivation, are then used in GLMs. Dougherty et al. (1995) gives an overview of methods to bin variables to be used in a (generalized) linear model, but a disadvantage of those methods is that the response variable is not taken into account in the binning process. Another strand of literature develops GAMs for pricing with flexible effects of continuous and spatial risk factors (see Denuit and Lang, 2004; Klein et al., 2014). What is lacking is a general framework that aligns the statistical advantages of flexible modeling with GAMs to the requirements of a production environment in an insurance company. This chapter tries to fill this gap by starting from GAMs with smooth effects and transforming these models into GLMs with categorical effects that satisfy the practical needs of an insurance company. Our strategy bins the continuous and spatial risk factors based on their GAM effects, resulting in categorical risk factors which are easily deployed in a GLM.

This chapter is structured as follows. In Section 2.2 we present the claims dataset and in Section 2.3 we fit flexible GAMs for frequency and severity to this dataset. In Section 2.4 we bin the spatial and continuous effects using Fisher's natural breaks and evolutionary trees. In Section 2.5 we fit GLMs with the binned risk factors and illustrate that the GLMs approximate the GAMs closely.

2.2 Claims dataset

We illustrate our methodology with a motor third party liability (MTPL) insurance portfolio from a Belgian insurer in 1997. A sample from this dataset is analyzed in Denuit and Lang (2004) and Klein et al. (2014). Each record in the dataset represents a unique policyholder who is observed during a certain policy period, ranging from one day to one year. The risk factors are registered at the start of the policy period and remain constant during this period. The dataset contains 163,231 policyholders and the available variables are listed in Table 2.1. In the Test-Achats Ruling, the Court of Justice of the EU prohibited the use of gender in insurance tariffs to avoid discrimination between males and females regarding pricing as from 21 December 2012. Notice of the European Commission: `http://ec.europa.eu/justice/newsroom/gender-equality/news/121220_en.htm`. Gender is therefore only investigated for use within an internal, technical tariff, but can not be used in a commercial tariff.

Variable	Description
`nclaims`	The number of claims filed by the policyholder.
`exp`	The fraction of the year 1997 during which the policyholder was exposed to the risk.
`amount`	The total amount claimed by the policyholder in Euros.

`coverage`	Type of coverage provided by the insurance policy:
	TPL = only third party liability,
	PO = partial omnium = TPL + limited material damage,
	FO = full omnium = TPL + comprehensive material damage.

`fuel`	Type of fuel of the vehicle: gasoline or diesel.
`sex`	Gender of the policyholder: male or female.
`use`	Main use of the vehicle: private or work.
`fleet`	The vehicle is part of a fleet: yes or no.
`ageph`	Age of the policyholder in years.
`power`	Horsepower of the vehicle in kilowatt.
`agec`	Age of the vehicle in years.
`bm`	Level occupied in the former compulsory Belgian bonus-malus scale. From 0 to 22, a higher level indicates a worse claim history (see Lemaire, 1995).
`long`	Longitude coordinate of the center of the municipality where the policyholder resides.
`lat`	Latitude coordinate of the center of the municipality where the policyholder resides.

Table 2.1: Overview of the available variables in the `MTPL` dataset.

Figure 2.1 illustrates how `nclaims`, `exp` and `amount` from Table 2.1 are distributed in the `MTPL` dataset. Most policyholders (88.79%) are claim-free during their insured period. A substantial number of policyholders (10.14%) files one claim and the remaining ones (1.07%) file two, three, four or five claims. Most policyholders (77.33%) have an exposure equal to one and are therefore covered by the insurance and exposed to the risk during the entire year. The exposure of the other policyholders (22.67%) is equally spread out between zero and one. Policyholders with an exposure lower than one have surrendered the policy during the year or started the policy in the course of the year. The overall claim frequency of the portfolio, calculated as the ratio of the total number of claims and the total exposure in years, is equal to 13.93%. Claims mainly involve small amounts. The total claim amount exceeds 10,000 Euro for only 2% of the claiming policyholders. The overall claim severity of the portfolio, calculated as the ratio of the total claim amounts and the total number of claims, is equal to 1,620.06 Euro.

Figure 2.2 illustrates how the risk factors from Table 2.1 are distributed in the

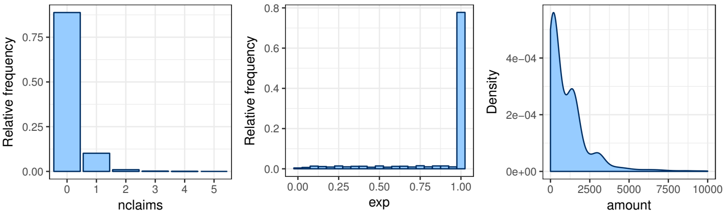


Figure 2.1: Relative frequency of `nclaims` and `exp` and density estimate of `amount`.

`MTPL` dataset. The `MTPL` dataset contains five categorical risk factors: `coverage`, `fuel`, `sex`, `use` and `fleet`. Most policyholders (58.28%) have only TPL coverage, which means that only their liability with respect to a third party (that is: another person) is covered. As in many developed countries, this coverage is compulsory in Belgium. The other policyholders have chosen for a policy which covers material damage on top of the TPL; either limited material damage in the form of a partial omnium (28.17%) or comprehensive coverage in the form of a full omnium (13.54%). Two types of fuel are used in the cars of the policyholders: gasoline (69.12%) and diesel (30.88%). Most policyholders are males (73.55%), they use their car mainly for private reasons (95.17%) and most cars are not part of a fleet (96.83%).

The `MTPL` dataset contains four continuous risk factors:, `ageph`, `power`, `agec` and `bm`. Almost all policyholders (93.53%) are aged between 25 and 75, which means that there are few young and old drivers in the insurance portfolio. Most of the cars in the insurance portfolio have less than 100 kilowatt of horsepower (97.35%) and are younger than 20 years old (99.53%). The rather low range of horsepower is nowadays outdated, but fits the less powerful cars from 1997. The left panel of Figure 2.3 shows a two dimensional density estimate for `ageph` and `power`. This gives additional intuition about the distribution of the policyholders over these continuous risk factors and the interplay between `ageph` and `power`.

More than half of the policyholders reside in the two lowest bonus-malus levels (level 0: 37.77% and level 1: 16.52%). Most of the other policyholders (42.90%) have a bonus-malus level between 2 and 11 and almost no policyholders (2.81%) occupy a bonus-malus level higher than 11. It should be noted that the bonus-malus level is usually not incorporated as a risk factor in an a priori tariff. However, we keep this variable in our analysis to investigate the information contained in this risk factor, much in line with the work of Denuit and Lang (2004); Klein et al. (2014). In reality, we distinguish between *a priori* and *a posteriori* pricing. The *a priori* premium takes into account policyholder information known at this point in time, while the *a posteriori* premium adjusts the *a priori* price based on historical claims information as it becomes available over time. The bonus-malus level is therefore a typical example of *a posteriori* information and is normally incorporated in an insurance tariff via credibility models or via bonus-malus systems in a more commercial setting (Antonio and Valdez, 2012). In traditional actuarial practice, *a priori* pricing deals with cross-sectional data via GLMs or GAMs, while *a posteriori* pricing deals with longitudinal panel data via Generalized Linear Mixed Models (GLMMs). Such GLMMs extend GLMs with random effects in the linear predictor to take into account unobserved heterogeneity and to determine the correlation structure.

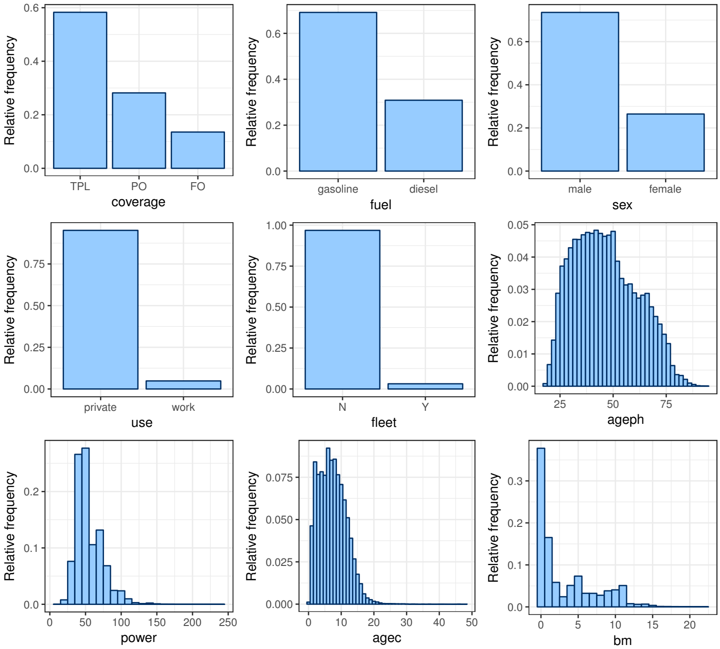


Figure 2.2: Relative frequency of the risk factors `coverage`, `fuel`, `sex`, `use`, `fleet`, `ageph`, `power`, `agec`, `bm`.

The `MTPL` dataset contains geographical information in the form of longitude and latitude coordinates, `long` and `lat`, of the municipality (or: postal code area) where the policyholder resides. The map of Belgium in Figure 2.3 visualizes the exposure in each municipality relative to the area of the municipality. White municipalities are those where the insurer has no policyholders and is therefore not exposed to the risk of filing a claim. Municipalities in light (dark) blue represent the 20% of municipalities containing the lowest (highest) relative exposure. Few policyholders are living in the southeastern part of Belgium, the Ardennes, while a lot of policyholders are living near some big cities of the French Community in Belgium; Brussels, Liège, Charleroi and Mons.

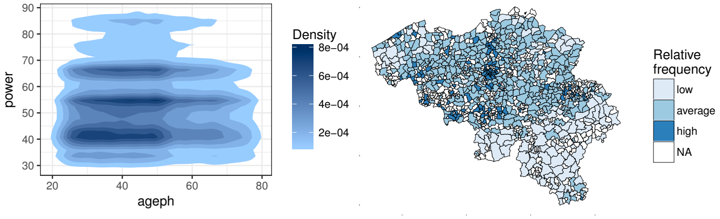


Figure 2.3: Density of `ageph` - `power` (left) and exposure map of Belgium (right).

2.3 Flexible models for P&C pricing using GAMs

Following McClenahan (2001); Antonio and Valdez (2012) we denote with F_i and S_i respectively the frequency and severity of policyholder i. Frequency is expressed as the number of claims N_i per unit of exposure e_i, while severity is expressed as the average claim amount over the number of claims N_i. The severity S_i is therefore only defined if policyholder i files a claim, i.e. if $N_i > 0$. We define the pure premium π_i as follows: $\pi_i = \mathbb{E}[F_i] \times \mathbb{E}[S_i]$, by assuming independence between F_i and S_i. In this setting we construct a predictive model for F_i using the claim history of all policyholders in the portfolio, including those who did not file a claim, and one for S_i using the history of policyholders who filed at least one claim.

GAMs are a suitable tool for actuarial regression modeling due to their flexibility in handling different types of risk factors. These models allow for the incorporation of smooth effects of continuous and spatial risk factors. The predictor η of the GAMs is expressed as follows:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}^d + \sum_{j=1}^{q} f_j(x_{ij}^c) + \sum_{j=1}^{r} f_j(x_{ij}^s, y_{ij}^s). \tag{2.1}$$

where μ_i is the mean of a response variable with a distribution from the exponential family and $g(.)$ is the link function. The 0/1-valued dummy variables x^d represent the typical way to code categorical risk factors in the GLM or GAM framework: a categorical risk factor with z levels requires the choice of a reference level and $z-1$ dummy variables to model the differences between the other levels and the reference level. The regression coefficient β_j captures the effect of dummy variable x_j^d on the predictor η. GAMs extend GLMs by including smooth functions of continuous risk factors. Main effects are captured by the univariate smooth functions $f(x^c)$, while interaction and spatial effects are expressed by bivariate smooth functions $f(x^s, y^s)$.

GAMs form the starting point of our pricing strategy and we search for the optimal model by using the Akaike information criterium (AIC, see Akaike,

1974) and the Bayesian information criterion (BIC, see Schwarz, 1978). Both take goodness of fit and model complexity into account and are defined as follows:

$$\begin{aligned} \text{AIC} &= -2 \cdot \log \mathcal{L} + 2 \cdot \text{EDF} \\ \text{BIC} &= -2 \cdot \log \mathcal{L} + \log(n) \cdot \text{EDF} \end{aligned} \tag{2.2}$$

where $\log \mathcal{L}$ is the log-likelihood of the model, n is the number of observations in the dataset and EDF represents the effective degrees of freedom which corresponds to the number of parameters in a GLM. Both AIC and BIC measure the goodness of fit by minus two times the log-likelihood supplemented with a complexity penalty. The BIC penalty is more severe and BIC will therefore favor less complex models. We continue our search for the optimal GAM with BIC as model selection criterion since we want to favor well performing models that are as simple as possible.[1] Note that lower AIC/BIC values indicate better models.

We fit the GAMs to the claims data on frequency and severity separately and follow - for both predictive models - a two-step strategy to select the appropriate set of risk factors to be included in (2.1). The first step performs an exhaustive search for the optimal GAM without taking into account interactions between the risk factors. In the second step, we perform an additional exhaustive search to search for meaningful interactions which improve the model fit. We only try to add interactions between continuous risk factors which have been selected in the first step.

We use `R` and the `mgcv` package developed by Wood (2006) to fit the GAMs. The smooth functions f from (2.1) are represented by penalized thin plate regression splines, wich are low rank approximations of the thin plate splines of Duchon (1977). For details on thin plate (regression) splines we refer to Section 2 of Wood (2003) and Section 4.1.5 of Wood (2006). We construct the interaction effects as tensor product interactions which exclude the main effects of the continuous risk factors, see Section 4.1.8 of Wood (2006) for details on tensor product smooths. We define the interactions in such a way that we can interpret them as corrections on top of the main effects which are included separately in the model. The model parameters are estimated by maximizing the penalized log-likelihood via penalized iteratively reweighted least squares (P-IRLS), see Section 4.3 of Wood (2006) for details on this procedure. The smoothness of the splines is controlled by a smoothing parameter, which will make a trade-off between penalizing a bad fit to the data and penalizing the 'wiggliness' of the spline. Smoothing parameters are estimated via Generalized Cross Validation (GCV, see Craven and Wahba, 1978) or via an Un-Biased

[1] For illustrative purposes we use BIC to select the optimal GAM, which serves as the starting point of our strategy. In the subsequent binning steps we use AIC as selection criterion.

Risk Estimator (UBRE, see Wahba, 1990) when the scale parameter in the distribution of the response is unknown or when it is known, see Section 4.5.4 of Wood (2006).

2.3.1 Frequency

In this section we focus on developing a flexible regression model for claim frequencies. We assume a Poisson distribution for `nclaims` and demonstrate our approach within this distributional setting. This is a common assumption in the insurance pricing industry and is in line with earlier work on this dataset (see Denuit et al., 2007). An actuary can however easily apply our approach to other distributional settings (e.g. negative binomial).

The goal is to explain the number of claims `nclaims` reported by a policyholder, for given exposure `exp`, using different types of risk factors. Our starting point is a Poisson GAM which includes all categorical risk factors `coverage`, `fuel`, `sex`, `use` and `fleet` together with main effects of all continuous risk factors `ageph`, `power`, `agec` and `bm` and a spatial effect based on `long` and `lat`. This GAM, which is not yet using any interaction terms, is formulated as follows:

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1\texttt{coverage}_{PO} + \beta_2\texttt{coverage}_{FO} + \\ &\beta_3\texttt{fuel}_{diesel} + \beta_4\texttt{sex}_{female} + \beta_5\texttt{use}_{work} + \\ &\beta_6\texttt{fleet}_{Y} + f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{agec}) + \\ &f_4(\texttt{bm}) + f_5(\texttt{long}, \texttt{lat}).\end{aligned} \tag{2.3}$$

The logarithm of exposure is included in the model as an offset, such that the expected number of claims is proportional to the exposure. The five categorical risk factors are coded with dummy variables by taking the level with the largest amount of exposure as reference level: $\texttt{coverage}_{TPL}$, $\texttt{fuel}_{gasoline}$, $\texttt{sex}_{male}$, $\texttt{use}_{private}$ and $\texttt{fleet}_{N}$. The functions f_1, f_2, f_3 and f_4 are univariate smooth effects of continuous risk factors. The spatial effect, f_5, is a bivariate smooth function of the latitude and longitude coordinates.

Our modeling choice for the spatial effect is in line with Denuit and Lang (2004) and Klein et al. (2014) in the field of P&C insurance pricing. This approach for predictive models involving spatial information is also used in other domains of statistics (see, among others, Vieira et al., 2005; Bristow et al., 2014; Chen et al., 2015). The first law of geography, introduced by Tobler (1970), states that *"everything is related to everything else, but near things are more related than distant things"*. The GAM framework adheres this law and allows us to smooth the spatial effect over neighboring municipalities and interpolate to unobserved districts, while controlling for other confounding risk factors such as the age of the policyholder and power of the car. When socio-economic characteristics are available per postal code, such as the average salary in a region, these can be controlled for in the GAM as well. An alternative approach (see Ohlsson, 2008) would be to model the spatial effect as a multi-level factor in a credibility framework.

We perform an exhaustive search over all possible combinations of explanatory variables in order to find the best GAM fit. The full model in (2.3) contains 10 risk factors: 5 categorical, 4 continuous and 1 spatial. All 1024 different models that can be formed by including or excluding these 10 risk factors are evaluated.[2] The model with the lowest BIC value of all 1024 investigated models is given by:

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \\ &\beta_3 \texttt{fuel}_{diesel} + f_1(\texttt{ageph}) + f_2(\texttt{power}) + \\ &f_3(\texttt{bm}) + f_4(\texttt{long}, \texttt{lat}).\end{aligned} \tag{2.4}$$

Two categorical risk factors, `coverage` and `fuel`, three continuous risk factors, `ageph`, `power` and `bm`, and the spatial risk factor are included in the optimal specification for the predictor.

We now investigate whether the model in (2.4) can further be improved by adding interaction effects between the continuous risk factors. Such interaction effects are not considered in the studies of Denuit and Lang (2004); Klein et al. (2014). For demonstration purposes we only include interaction effects among continuous risk factors and not among categorical risk factors or between a continuous and a categorical risk factor. An interaction effect between a continuous and categorical risk factor will give rise to a smooth effect of the continuous risk factor for every level of the categorical risk factor. Adding these types of interactions will therefore only result in a more complex model without contributing added value to the demonstration of our strategy for the construction of tariff classes.

We examine interactions between the continuous risk factors already included in (2.4). The only possible interactions between `ageph`, `power` and `bm` are: `ageph-power`, `ageph-bm` and `power-bm`. Incorporating the interaction `ageph-power` results in a decrease of BIC whereas adding the other two interaction effects always results in an increase of BIC. The interaction `ageph-power` is therefore added to the model in (2.4) and our resulting GAM

[2] This operation takes approximately 20 hours on one core of a 2.7 GHz Intel Core i5 processor.

for claim frequency is given by:

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = \log(\texttt{exp}) + \beta_0 + \beta_1 \texttt{coverage}_{PO} + \beta_2 \texttt{coverage}_{FO} + \\ \beta_3 \texttt{fuel}_{diesel} + f_1(\texttt{ageph}) + f_2(\texttt{power}) + f_3(\texttt{bm}) + \\ f_4(\texttt{ageph}, \texttt{power}) + f_5(\texttt{long}, \texttt{lat}).\end{aligned} \tag{2.5}$$

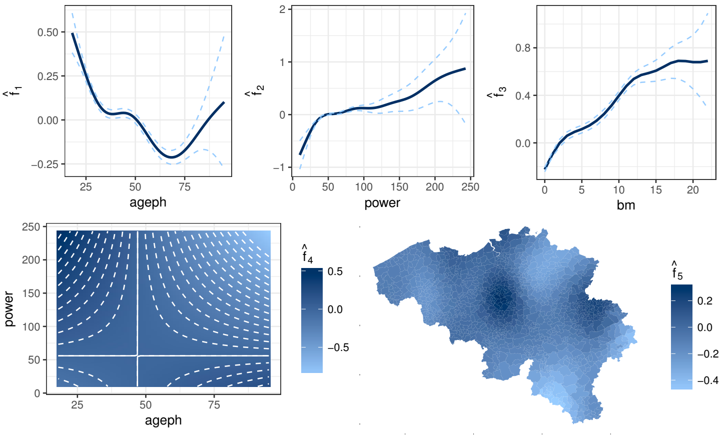


Figure 2.4: Fitted smooth frequency GAM effects from (2.5). Top row: main effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$ and $\hat{f}_3(\texttt{bm})$. Bottom row: interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$ and spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$.

Figure 2.4 displays the five fitted smooth functions: $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$ and $\hat{f}_5(\texttt{long}, \texttt{lat})$ from (2.5). The top row shows the fitted smooth effects of the risk factors `ageph`, `power` and `bm` in solid lines. The dashed lines represent the 95% pointwise confidence intervals, which are wider in regions with scarce data. Young policyholders appear to be risky drivers, which might be explained by their driving style or lack of experience behind the wheel. This riskiness decreases over increasing ages and stabilizes around the age of 35. It increases slightly between ages 45 and 50, possibly due to the fact that children of policyholders in their late 40s - early 50s start to drive with their parents' car. After age 50 the riskiness decreases again until the age of 70, after which it starts increasing again. This implies that seniors report more car accidents when growing older. Note however the widening confidence interval for these high ages due to the rarity of old policyholders in our portfolio. The smooth effect of `power` shows a steep increase over the interval from 0 to 50 kilowatt and a more gradual increase from 50 kilowatt onwards. This implies that policyholders driving a more powerful vehicle are more likely to report a claim. The smooth effect of bm shows a steady increase over increasing bonus-malus levels. This effect is in line with our intuition, since policyholders occupying high bonus-malus levels have worse claim histories compared to policyholders with low bonus-malus levels.

The fitted interaction effect between ageph and power is displayed in the bottom left panel of Figure 2.4. A negative (positive) correction, coloured in light blue (dark blue), indicates that the combined main effects of ageph and power overestimate (underestimate) the annual expected claim frequency. The combinations low ageph - low power and high ageph - high power are therefore less risky than the two main effects predict. The combinations high ageph - low power and low ageph - high power are therefore more risky than the two main effects predict. Among others, the results of our preferred GAM show that young policyholders driving a more powerful car imply a high risk for the insurer, at least in terms of the claim frequency.

The fitted spatial effect is displayed in the bottom right panel of Figure 2.4. Note that this map does not indicate how likely claims are to occur in each municipality, but it reflects in which municipalities the more risky policyholders reside. Although a policyholder can have an accident in any municipality, we can assume that he will drive quite often in his own municipality. Moreover, the municipality serves as a proxy for socio-economic characteristics that characterize the neighborhood where the policyholder resides. The municipalities are colour coded where light blue (dark blue) indicates a municipality where policyholders reside which have, on average, few (many) car accidents. The region around Brussels, in the center of Belgium, is associated with the highest accident risk. Traffic is very dense in this area, which is reflected in a higher expected annual claim frequency for policyholders who live here. The southeastern, northeastern and western parts of Belgium are less densely populated, which is reflected in a lower expected annual claim frequency for policyholders who live here.

2.3.2 Severity

We now focus on developing a flexible regression model for claim severities. Our dataset does not contain individual claim amounts, but we have the total claim amount and the number of claims at our disposal. We therefore work with the average cost of a claim where avg is defined as the ratio of amount and nclaims. We use nclaims as a weight in our regression model and assume a lognormal distribution for avg. This is a common assumption in the insurance pricing industry and is in line with earlier work on this dataset (see Denuit and Lang, 2004). An actuary can however easily apply our strategy with other severity distributions (e.g. gamma).

We follow the fitting procedure outlined in Section 2.3.1 and start with finding an optimal lognormal GAM without interation effects. In a next step we look for interactions between continuous risk factors that improve the model fit. In the severity fitting procedure we can only use observations of policyholders who actually filed a claim, i.e. $\texttt{nclaims} > 0$, which accounts for 18,295 records in our MTPL dataset. The very large claims are excluded from our analysis since these are not the focus when developing a tariff structure. Using techniques from Extreme Value Theory (EVT) Denuit and Lang (2004); Klein et al. (2014) obtain a threshold of 81,000 Euro which separates small, attritional losses from large losses. For 19 records the average claim cost exceeds this threshold. We therefore obtain 18,276 records below the threshold to fit our severity model.

Our preferred model for claim severity is the lognormal GAM given by:

$$\begin{aligned}\mathbb{E}(\log(\texttt{avg})) =& \gamma_0 + \gamma_1 \texttt{coverage}_{PO} + \gamma_2 \texttt{coverage}_{FO} + \\ & g_1(\texttt{ageph}) + g_2(\texttt{bm}).\end{aligned} \tag{2.6}$$

A Gaussian distribution is assumed for the response $\log(\texttt{avg})$, such that the average amount of a claim follows a lognormal distribution. Only one categorical risk factor, coverage, and two continuous risk factors, ageph and bm, are selected. We find no relevant interaction or spatial effect for severity. As documented in actuarial pricing literature (see Charpentier, 2014), severity models tend to have fewer relevant risk factors compared to frequency models. Claim severity is more difficult to explain by risk factors than claim frequency for at least two reasons. First of all, one has less data available to fit a severity model. Secondly, the driver has almost no control over the cost of an accident.

Figure 2.5 displays the two fitted smooth functions: $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$ from (2.6). Going from ages 18 to 35, we can observe a decrease of the average claim cost. This indicates that very young drivers are involved in more severe car accidents. The average claim cost starts to increase again for policyholders older than 35, stabilizes in the age interval 45 to 60 after which it starts to increase again. A possible explanation might be the fact that older policyholders drive more expensive cars and repairing costs increase. The age of the policyholder might be operating as a proxy for the price of the car. Unfortunately we do not have that information in our dataset to confirm this. The average claim cost increases with increasing bonus malus levels. There is however a stabilizing region around level 5 and the average claim cost decreases from bonus malus level 13 onwards. Because of the scarceness of data for the high bonus malus levels one can not conclude much about this region, as the widening confidence bounds illustrate.

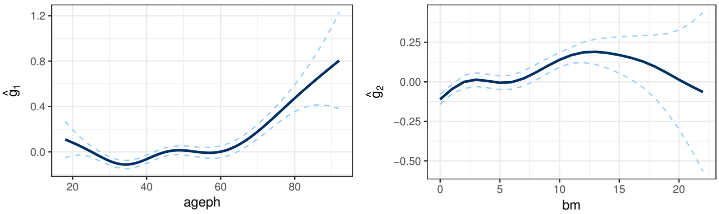


Figure 2.5: Fitted smooth severity GAM effects $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$ from (2.6).

2.4 Data-driven binning methods for the smooth GAM effects

The GAM model formulas (2.5) and (2.6) are optimal, according to BIC, for claim frequency and severity in the MTPL dataset. These models offer a high degree of flexibility for the spatial and continuous risk factors, which is very appealing from a statistical modeling point of view. For practical purposes, as discussed in Section 2.1, insurers prefer a pricing model where each risk factor is categorical. This makes the price list easy to implement, explain and adjust. In this section we present a data-driven approach to bin the spatial and continuous risk factors of the predictors (2.5) and (2.6). In Section 2.4.1 we bin the spatial risk factor, which is only present in the frequency model. In Section 2.4.2 we bin the continuous risk factors, which are present in both the frequency and severity models. Once all risk factors are categorical, it is straightforward to estimate a GLM with the risk factors coded by dummy variables (see Section 2.5).

2.4.1 Spatial effect

We first put focus on binning the fitted continuous spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ from the frequency model (2.5). For each of the 1,146 Belgian municipalities we have a single number which represents the spatial riskiness of that municipality: $s_i = \hat{f}_5(\texttt{long}_i, \texttt{lat}_i)$ for $i \in \{1, ..., 1146\}$. The goal therefore is to group the municipalities with similar spatial riskiness together. We use the classInt package in R, developed by Bivand (2015), to compare four different binning methods:

- **Equal intervals.** The range of the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ is divided in k bins of equal length: $\frac{\max(s_i)-\min(s_i)}{k}$, where the maximum and minimum are taken over all i. This approach can give good results for uniformly distributed data, but tends to perform poorly for skewed data.

- **Quantile binning.** Each bin will contain approximately $\frac{1146}{k}$ municipalities where k equals the number of bins. This method is often the default in statistical software packages, though it can give very misleading results. Similar observations can be assigned to different bins in order to make sure that each bin contains the same number of observations.

- **Complete linkage.** This method performs agglomerative hierarchical clustering (see Kaufman and Rousseeuw, 1990). Initially each municipality forms its own bin and in every iteration the two bins closest to each other are merged. The distance between bins i and j is equal to the distance between their most distant points: $d(i,j) = \max |s_u^{(i)} - s_v^{(j)}|$ $\forall u, v : s_u^{(i)} \in i$, $s_v^{(j)} \in j$, where u and v run over all possible combinations of points from bin i and bin j. Bins with remote observations will only be merged in a late stage of the iteration process.

- **Fisher's natural breaks.** This iterative algorithm, developed by Fisher (1958) and discussed in Slocum et al. (2005), maximizes the homogeneity within bins. Bins are created such that every observation $s_u^{(i)}$ in bin i is as close as possible to the average of its bin $\bar{s}^{(i)}$. This is done by minimizing the sum of squared distances between observations $s_u^{(i)}$ and the respective bin means $\bar{s}^{(i)}$: $\sum_{i=1}^{k} \sum_{u=1}^{n_i} (s_u^{(i)} - \bar{s}^{(i)})^2$. Here, i runs over the different bins, u runs over the municipalities within each bin, k is the number of bins and n_i the number of municipalities within bin i.

We compare the results of the different binning methods using two measures: the goodness of variance fit (GVF) and the tabular accuracy index (TAI). The GVF and TAI are defined as follows by Armstrong et al. (2003):

$$\text{GVF} = 1 - \frac{\sum_{i=1}^{k} \sum_{u=1}^{n_i} (s_u^{(i)} - \bar{s}^{(i)})^2}{\sum_{u=1}^{1146} (s_u - \bar{s})^2} \tag{2.7}$$

$$\text{TAI} = 1 - \frac{\sum_{i=1}^{k} \sum_{u=1}^{n_i} |s_u^{(i)} - \bar{s}^{(i)}|}{\sum_{u=1}^{1146} |s_u - \bar{s}|} \tag{2.8}$$

where k and n_i indicate the number of bins and the number of municipalities in bin i. The denominator in (2.7) and (2.8) measures the deviation of each municipality from the global average. The numerator in (2.7) and (2.8) measures the deviation of each municipality from its bin average. For both measures a value closer to 1 indicates small variance within the bins compared to the global variance and hence a more homogeneous binning of the municipalities.

Table 2.2 compares the performance of the four binning methods for different number of bins k. We denote the highest values for the GVF and TAI in bold in every column where we use the fourth digit behind the decimal point in case of a tie in the presented values. Fisher's natural breaks algorithm outperforms the other methods in eleven out of twelve cases. This method clearly results in the most homogeneous binning and is therefore the preferred method to bin the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ from the frequency model (2.5). The equal intervals method is performing rather well despite its simplicity; it attains the second best GVF in five out of six cases. Quantile binning also performs well, attaining the second best TAI in five out of six cases. The complete linkage method attains the lowest GVF/TAI in nine out of twelve cases.

	k=2		k=3		k=4		k=5		k=6		k=7	
	GVF	TAI	GVF	TAI	GVF	TAI	GVF	TAI	GVF	TAI	GVF	TAI
Equal	0.610	**0.357**	0.778	0.492	0.882	0.628	0.913	0.675	0.940	0.730	0.955	0.767
Quantile	0.615	0.356	0.773	0.526	0.854	0.642	0.894	0.694	0.921	0.751	0.937	0.778
Complete	0.558	0.314	0.680	0.395	0.857	0.613	0.892	0.657	0.936	0.726	0.952	0.761
Fisher	**0.615**	0.356	**0.822**	**0.562**	**0.892**	**0.654**	**0.927**	**0.724**	**0.951**	**0.769**	**0.963**	**0.795**

Table 2.2: The GVF and TAI for the four methods and different values for k to bin the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$.

In order to get a better understanding of the resulting bins for different methods we show two visual comparisons with $k = 5$. Figure 2.6 shows the empirical cumulative distribution function of the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ in combination with the bins produced by the different methods. Figure 2.7 visualizes the different spatial binning results on the map of Belgium. The four methods result in very different bins for the spatial effect and therefore give rise to very different groupings of the municipalities. The method of equal intervals clearly divides the range of the spatial effect in five equally sized bins. A lot of municipalities are therefore grouped in the middle bin (499) whereas few municipalities are grouped in the first and last bin (39 and 111). Quantile binning produces very wide bins in the extreme ends of the support where data are scarce. Every bin contains approximately 230 municipalities. Complete linkage groups a lot of municipalities in the second bin (465) and few municipalities in the first and last bin (39 and 55). Fisher's natural breaks algorithm results in the most homogeneous binning and seems to act as a middle ground between equal intervals and quantile binning. We obtain 300 municipalities in the middle bin and respectively 86 and 176 in the first and last bin.

From inspecting Table 2.2 it follows that increasing the number of bins k results in a monotonic increase of both the GVF and TAI. These measures can therefore not be used to choose the number of bins k since more bins will always result in a more homogeneous binning. As motivated in Section 2.1, pricing actuaries ultimately prefer a GLM where all types of risk factors are coded as categorical

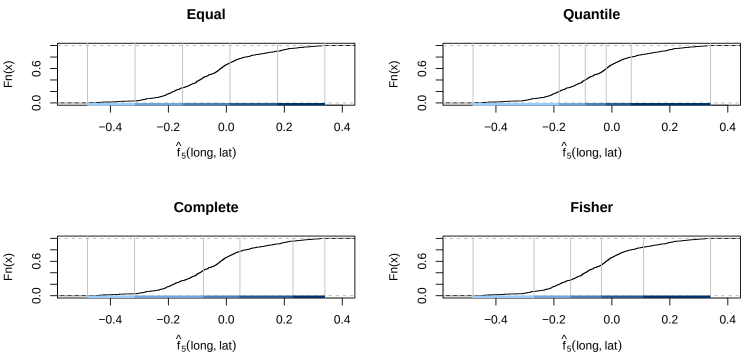


Figure 2.6: The empirical cumulative density function of the fitted spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ in combination with the five bins produced by the four different binning methods.

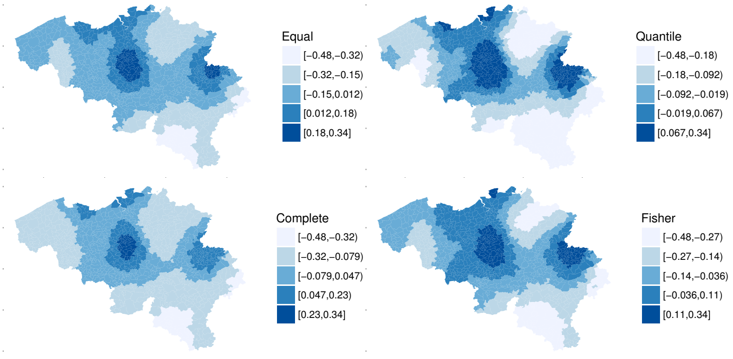


Figure 2.7: Maps of Belgium with the municipalities grouped into five distinct bins based on the intervals produced by the four different binning methods for the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$.

variables. We therefore propose to tune the optimal number of bins for the spatial effect based on the binned spatial effect which will be used in a GLM. We tune the number of bins by considering a set of possible values for the number of spatial bins, e.g. $k \in \{2, 3, 4, 5, 6, 7\}$. For each value in this set the procedure listed in Table 2.3 is applied.

After applying this procedure we choose the number of bins for the spatial effect that results in the lowest AIC for the GAM with a binned spatial effect. Our approach requires the estimation of a GAM with binned spatial effect for each value in the considered set, but this extra effort allows us to find the best

Procedure:	Find the optimal number of bins for the spatial effect
Step 1	Apply Fisher's natural breaks algorithm to calculate the bin intervals for the spatial effect, $\hat{f}_5(\texttt{long}, \texttt{lat})$ from (2.5), where the number of bins is chosen equal to the current value of the predefined set of values. These bin intervals are used to transform the continuous spatial effect into a categorical spatial effect.
Step 2	Estimate a new GAM where we use a predictor structure similar to (2.9). This GAM contains the spatial effect in a categorical format, but still uses flexible effects to model the continuous risk factors.
Step 3	Calculate AIC of the GAM with a binned spatial effect.

Table 2.3: Procedure to find the optimal number of bins for the spatial effect.

fitting GLM in the end. The results in Table 2.4 illustrate that choosing five bins result in the lowest AIC for the GAM with a binned spatial effect. We also report the BIC values of the GAMs with a binned spatial effect and conclude that choosing BIC as evaluation measure also results in five bins for binning the spatial effect.

# bins	AIC	BIC
2	124778.9	125047.6
3	124753.1	125023.9
4	124652.3	124928.4
5	**124621.3**	**124907.2**
6	124627.7	124921.6
7	124639.1	124942.9

Table 2.4: AIC and BIC for the fitted GAM with binned spatial effect, as obtained via Fisher's natural breaks, evaluated over a predefined set for the number of bins.

From this section we conclude that it is optimal to bin the spatial effect $\hat{f}_5(\texttt{long}, \texttt{lat})$ from the frequency model (2.5) with Fisher's natural breaks algorithm in five bins. This results in the most homogeneous binning for the spatial effect and the lowest AIC value for the GAM with a binned spatial effect. For the frequency model we continue our study with a GAM which specifies the spatial effect as a categorical risk factor (geo). The bin $[-0.036, 0.11)$ is chosen as the reference class since it contains the highest amount of exposure, namely 354 municipalities. This gives the following GAM specification for the frequency model:

$$\begin{aligned}\log(\mathbb{E}(\texttt{nclaims})) = &\log(\texttt{exp}) + \beta_0 + \beta_1\texttt{coverage}_{PO} + \beta_2\texttt{coverage}_{FO} + \\ &\beta_3\texttt{fuel}_{diesel} + \beta_4\texttt{geo}_{[-0.48,-0.27)} + \beta_5\texttt{geo}_{[-0.27,-0.14)} + \\ &\beta_6\texttt{geo}_{[-0.14,-0.036)} + \beta_7\texttt{geo}_{[0.11,0.34]} + f_1(\texttt{ageph}) + \\ &f_2(\texttt{power}) + f_3(\texttt{bm}) + f_4(\texttt{ageph},\texttt{power}).\end{aligned} \tag{2.9}$$

2.4.2 Continuous risk factors

We now put focus on binning the main and interaction effects of the continuous risk factors: $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph},\texttt{power})$ from the frequency model (2.9) and $\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$ from the severity model (2.6).

We want to create bins where consecutive values of a continuous risk factor are grouped together. The approach followed for binning the spatial effect is therefore no longer appropriate, since it might create bins where - for example - policyholders younger than 30 are grouped with policyholders older than 80. We propose the use of regression trees as a technique to perform the binning since these models produce intuitive splits in line with our requirement of grouping consecutive values of the continuous variables, e.g. ages. Classic regression tree methods, such as Classification And Regression Trees (CART) from Breiman et al. (1984), are recursive partitioning methods that fit a model in a forward stepwise search. Splits are chosen to maximize the homogeneity of the partitions at every step and these consecutive splits are kept fixed in all the following steps. This forward stepwise search is an efficient heuristic, but the resulting binning is only locally optimal. We refer to Breiman et al. (1984); Hastie and Tibshirani (1990); for details and terminology regarding tree-based models.

In our procedure, we choose to use evolutionary trees from the R package evtree developed by Grubinger et al. (2014). These evolutionary trees combine the framework of regression trees with genetic algorithms. An evolutionary process mutates the tree structure at any possible node until convergence is reached towards an optimal solution which can not be improved anymore. Evolutionary trees therefore allow us to adapt earlier splits in a later stage of the fitting procedure. Thanks to this extra flexibility, evolutionary trees are capable of finding a global optimum (see Grubinger et al., 2014), which results in a more robust binning of our continuous risk factors as we showcase later on.

The data that serves as input to the evolutionary trees are the main and interaction effects from the GAMs in (2.9) and (2.6): ($\texttt{ageph}$,$\hat{f}_1$), ($\texttt{power}$,$\hat{f}_2$), (bm,$\hat{f}_3$), (ageph, power,$\hat{f}_4$), (ageph,$\hat{g}_1$), (bm,$\hat{g}_2$). Hence we grow a regression tree for each of the fitted flexible effects, with the flexible effect (e.g. $\hat{f}_1$(ageph)) as response and the corresponding risk factor (e.g. ageph) as covariate. This results in a binned version of the flexible effect which takes into account the ordering of the levels of the continuous risk factors. It is important to take the composition of the insurance portfolio into account when deciding where to split or bin the risk factors. For policyholders older than 75, for example, the fitted smooth effect $\hat{f}_1$(ageph) in Figure 2.4 is strongly increasing, but Figure 2.2 indicates that the portfolio does not contain many policyholders aged over 75. It is not desirable to obtain many splits in this region, since it will only affect a small portion of the portfolio. Therefore the distribution of the policyholders with respect to a specific risk factor is taken into account by using the number of policyholders as weights. Table 2.5 shows a snippet of input data for the $\hat{f}_1$(ageph) tree. For example, the MTPL portfolio contains 393 policyholders aged 20. The smooth effect for ageph $= 20$, $\hat{f}_1(20)$, therefore obtains an integer weight of 393. The evolutionary tree will interpret this weight as if the observation pair (ageph $= 20, \hat{f}_1(20)$) occurs 393 times in its input data. A constraint is imposed to make sure that bins are not too sparsely populated: each bin should at least contain 5% of the policyholders in the entire portfolio. Modifying this constraint gives insurers flexibility over the granularity of the bins.

Covariate: ageph	Response: $\hat{f}_1$(ageph)	Weight: w
18	0.495	16
19	0.459	116
20	0.424	393

Table 2.5: Snippet of the input data for the evolutionary tree that bins $\hat{f}_1$(ageph).

The evaluation function to measure the performance of a tree has the following form:

$$n \cdot \log(\text{MSE}) \; + \; 4 \cdot \alpha \cdot (m+1) \cdot \log(n) \tag{2.10}$$

where n is the number of observations in the input data, m is the number of leaf nodes in a tree and α is a tuning parameter (see Grubinger et al., 2014). Note that n is equal to the sum of all the weights since a tree interprets a weight as being the number of times that the respective (response, covariate) pair occurs in the input data. The first term in evaluation function (2.10) measures the accuracy of the tree by means of the mean squared error (MSE). For $\hat{f}_1$(ageph) - for example - this MSE looks as follows:

$$\text{MSE} = \frac{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i} \left(\hat{f}_1(\texttt{ageph}_i) - \hat{f}_1^b(\texttt{ageph}_i)\right)^2}{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i}} \tag{2.11}$$

with $w_{\texttt{ageph}_i}$ the number of policyholders with $\texttt{ageph} = i$, $\hat{f}_1(\texttt{ageph}_i)$ the fitted GAM effect for a policyholder with $\texttt{ageph} = i$ and $\hat{f}_1^b(\texttt{ageph}_i)$ the fitted value obtained from the regression tree. This last value is obtained as the mean of $\hat{f}_1(\texttt{ageph}_i)$ over all policyholders in the age bin where $\texttt{ageph} = i$ belongs to. The second term of evaluation function (2.10) represents a complexity penalty in terms of the size of the tree, measured by the number of leaf nodes m, where each leaf node corresponds to a bin. The tuning parameter α takes care of the trade-off between accuracy and complexity.

The evaluation function in (2.10) scales in a comparable manner over the four trees that bin the frequency effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$, $\hat{f}_4(\texttt{ageph}, \texttt{power})$ because of two reasons. Firstly, $n = 163{,}231$ for all four trees since this is the total number of policyholders in the MTPL dataset. Secondly, the MSEs are on the same level since they deal with the variance of the effects at the level of the predictor of the frequency model. This motivates the use of a single tuning parameter α_{freq} to construct the trees. Likewise we define α_{sev} as the single tuning parameter for both trees that bin the severity effects $\hat{g}_1(\texttt{ageph})$, $\hat{g}_2(\texttt{bm})$. For both these trees $n = 18{,}276$ and the MSEs are again at the same level, that is the level of the predictor of the severity model. This implies that we have two tuning parameters which can be optimized independently: α_{freq} and α_{sev}. Tuning these α's follows the same approach as tuning the number of bins for the spatial effect in Section 2.4.1. We consider a set of possible values for both α_{freq} and α_{sev} and search for the optimal ones. After some initial exploration, we choose to use the unequally spaced set $\{1,1.5,2,\ldots,9.5,10,15,20,\ldots,95,100,150,200,\ldots,950\}$ for both α_{freq} and α_{sev}. These values allow us to find the right balance between too complex trees (low α's) and too simplistic trees (high α's). For each value in this set the procedure listed in Table 2.6 is applied.

Procedure:	Find the optimal tuning parameters α_{freq} and α_{sev} for the evolutionary trees
Step 1	Fit evolutionary trees to the main and interaction effects of the continuous risk factors, ($\texttt{ageph}$, $\hat{f}_1(\texttt{ageph})$) et cetera, where α is chosen equal to the current value of the predefined set of values. The splits produced by these trees are used to transform the continuous effects into categorical effects.
Step 2	Estimate a frequency and severity GLM with the resulting categorical risk factors.
Step 3	Calculate AIC of the frequency GLM and the severity GLM.

Table 2.6: Procedure to find the optimal tuning parameters α_{freq} and α_{sev} to bin the main and interaction effects, ($\texttt{ageph}$, $\hat{f}_1(\texttt{ageph})$) et cetera, via the resulting GLM.

After applying this procedure we choose the values of α_{freq} and α_{sev} that result in the lowest AIC for respectively the frequency and severity GLM. Figure 2.8 shows the splits produced by the preferred evolutionary trees with $\alpha_{freq} = 550$ and $\alpha_{sev} = 70$.

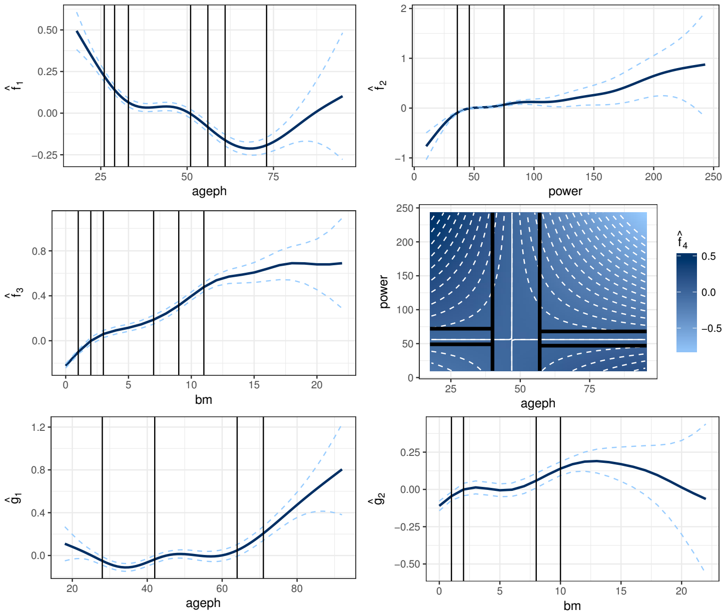


Figure 2.8: Binning intervals for the continuous effects. Top and middle row: main and interaction effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_2(\texttt{power})$, $\hat{f}_3(\texttt{bm})$ and $\hat{f}_4(\texttt{ageph}, \texttt{power})$ from the frequency model, binned by evolutionary trees with $\alpha_{freq} = 550$. Bottom row: main effects $\hat{g}_1(\texttt{ageph})$ and $\hat{g}_2(\texttt{bm})$ from the severity model, binned by evolutionary trees with $\alpha_{sev} = 70$.

By means of example we discuss the splits obtained for the main effects $\hat{f}_1(\texttt{ageph})$, $\hat{f}_3(\texttt{bm})$ and the interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$. Similar observations hold for the other effects. The main effect $\hat{f}_1(\texttt{ageph})$ is split into eight bins. All policyholders with an age in the interval $[33, 51)$ are grouped together since $\hat{f}_1(\texttt{ageph})$ is rather flat over this interval. Younger policyholders have a higher risk profile and three bins are formed for policyholders younger than 33: $[18, 26)$, $[26, 29)$ and $[29, 33)$. The first bin is wider because there are very few young policyholders present in the portfolio. The smooth effect $\hat{f}_1(\texttt{ageph})$ decreases for policyholders aged over 51. These policyholders typically have more driving experience and therefore a lower risk profile. Two bins are created in this region: $[51, 56)$ and $[56, 61)$. The smooth effect $\hat{f}_1(\texttt{ageph})$ stabilizes after age 61 before it starts increasing again for senior policyholders.

This results in two bins; $[61, 73)$ for the stabilizing region and $[73, 95]$ for the senior policyholders with a higher risk profile.

The main effect $\hat{f}_3(\texttt{bm})$ is split into seven bins. The first three bonus-malus levels end up in separate bins: $[0, 1)$, $[1, 2)$ and $[2, 3)$. The next bin, $[3, 7)$, is wider because $\hat{f}_3(\texttt{bm})$ increases less steeply over this range. The next two bins, $[7, 9)$ and $[9, 11)$, are less wide because $\hat{f}_3(\texttt{bm})$ starts to increase more steeply again over this region. The higher bonus-malus levels are grouped into one bin: $[11, 22]$. This bin is so wide for two reasons: the slope of $\hat{f}_3(\texttt{bm})$ decreases for higher bonus-malus levels and only a few policyholders have such high levels.

The interaction effect $\hat{f}_4(\texttt{ageph}, \texttt{power})$ is split into seven bins. The solid white contour lines indicate where $\hat{f}_4(\texttt{ageph}, \texttt{power}) = 0$. These combinations of ageph and power therefore indicate a neutral region where the main effects of ageph and power are sufficient to fully grasp the riskiness. Our method detects this neutral region by forming three bins around the solid white contour lines. The vertical bin for $40 \leq \texttt{ageph} < 57$ represents the neutral zone around the vertical contour line. Two horizontal bins, one for $49 \leq \texttt{power} < 72$ on the left and one for $47 \leq \texttt{power} < 68$ on the right, take the neutral zone around the horizontal contour line into account. The bottom left graph of Figure 2.3 shows the two-dimensional distribution of the number of policyholders over ageph and power. This figure explains why the tree chooses the neutral zone in the vertical direction as largest; most policyholders can be captured as such. The other four bins represent policyholders with lower/higher risk profiles compared to the neutral region. Two bins represent policyholders with a lower risk profile: $\texttt{ageph} < 40, \texttt{power} < 49$ and $\texttt{ageph} \geq 57, \texttt{power} \geq 68$. Two bins represent policyholders with a higher risk profile: $\texttt{ageph} < 40, \texttt{power} \geq 72$ and $\texttt{ageph} \geq 57, \texttt{power} < 47$.

Figure 2.9 motivates our preference for the more flexible evolutionary trees over the classic recursive partitioning trees when binning the continuous risk factors. The R package rpart, developed by Therneau et al. (2019), is used to fit the recursive trees and weights are applied in the same manner as for the evolutionary trees. The left panel of Figure 2.9 shows the first split for binning the interaction effect. Although this split at $\texttt{ageph} = 74$ is an optimal way to bin the interaction effect in two regions, it does not imply that this split is still optimal when we bin the interaction effect in three or more regions. The right panel of Figure 2.9 shows the binning result with recursive partitioning for seven bins. As such we obtain the same number of bins as with the evolutionary trees in Figure 2.8. Evolutionary trees lead to a much more stable binning result which is globally optimal instead of only locally optimal. Furthermore, compared to the pattern found in Figure 2.9 with a recursive tree, the neutral region with areas of increased and decreased risk from Figure 2.8 in an evolutionary tree is much easier to explain to customers and managers in practice.

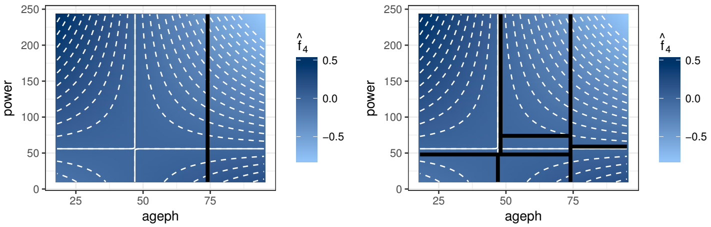


Figure 2.9: The optimal first split for the interaction effect (left) and the resulting seven bins for the interaction effect (right) obtained with recursive partitioning.

2.5 Analysis of the premium structure

2.5.1 From GAM to GLM

We now use the categorical risk factors, as constructed in Section 2.4, to fit a Poisson GLM for the frequency and a lognormal GLM for the severity of the claims data. For each risk factor we choose the bin with the largest exposure as reference class in our GLMs. A full specification of the frequency and severity regression models is in Appendix A.1.

The top and bottom row of Figure 2.10 compare the original main GAM effects with the resulting estimated GLM coefficients for the frequency and severity model respectively. Note that the GLM coefficients of Figure 2.10, indicated by dots, are a rescaled version of the actual coefficients. This rescaling is performed to make the GAM effects and the GLM coefficients comparable. Indeed, a GAM effect such as $\hat{f}_1(\texttt{ageph})$ is estimated in such a way that the weighted mean of the smooth effect, with the number of policyholders as weights, is equal to zero:

$$\frac{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i} \hat{f}_1(\texttt{ageph}_i)}{\sum_{i=\min(\texttt{ageph})}^{\max(\texttt{ageph})} w_{\texttt{ageph}_i}} = 0 \tag{2.12}$$

with $w_{\texttt{ageph}_i}$ the number of policyholders with $\texttt{ageph} = i$ and $\hat{f}_1(\texttt{ageph}_i)$ the corresponding fitted GAM effect. For a categorical risk factor, the GLM coefficient for the reference class is equal to zero and the GLM coefficients of the other classes are expressed relative to this reference class. The weighted mean of the GLM coefficients, with the number of policyholders as weights, therefore depends on the choice of the reference class. We adjust the GLM coefficients such that the weighted mean of the rescaled GLM coefficients is equal to zero:

$$\tilde{\beta}_{\texttt{ageph}_j} = \hat{\beta}_{\texttt{ageph}_j} - \frac{\sum_{j=1}^{k_{\texttt{ageph}}} m_{\texttt{ageph}_j} \hat{\beta}_{\texttt{ageph}_j}}{\sum_{j=1}^{k_{\texttt{ageph}}} m_{\texttt{ageph}_j}} \tag{2.13}$$

with $k_{\texttt{ageph}}$ the number of bins for the risk factor `ageph`, $m_{\texttt{ageph}_j}$ the number of policyholders in bin j and $\hat{\beta}_{\texttt{ageph}_j}$ the fitted GLM coefficient for the policyholders in bin j. This rescaling is of course only performed to enable a visual comparison between the GAM effects and the GLM coefficients; the actual GLM is not adjusted in any way.

The piecewise constant functions formed by the GLM coefficients in Figure 2.10 approximate the smooth GAM effects very closely, especially for bins with high exposure. For example, the GLM coefficients for age bin $[33, 51)$ approximate $\hat{f}_1(\texttt{ageph})$ very well. This indicates that our resulting GLMs are a good approximation of the original GAMs. We trade flexibility for simplicity and some discrepancies are therefore impossible to avoid. For example: both $\hat{f}_1(\texttt{ageph})$ and $\hat{g}_1(\texttt{ageph})$ are underestimated by the GLM coefficients for the youngest and oldest policyholders. Such discrepancies only occur in the extreme ends of the range of the risk factors, where the exposure is very low, and therefore few policyholders are affected by this.

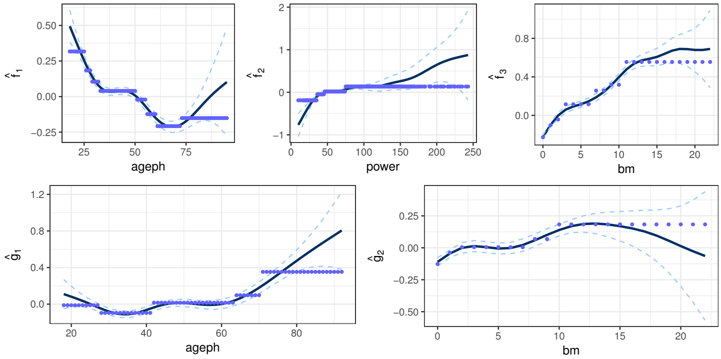


Figure 2.10: Comparison between the GAM effects and GLM coefficients. Top row: risk factors `ageph`, `bm` and `power` in the frequency model. Bottom row: risk factors `ageph` and `bm` in the severity model.

Figure 2.11 shows the GAM interaction effect between `ageph` and `power` in the frequency model together with the approximation obtained with the GLM coefficients. The plus-shaped (+) region can be interpreted as a neutral zone while the top left and bottom right (bottom left and top right) indicate regions of increased (decreased) risk. Figure 2.12 shows both the GAM estimate for the spatial effect in the frequency model and its approximation in the GLM. As expected, the GLM coefficients reflect the riskiness as captured by the spatial effect from the GAM.

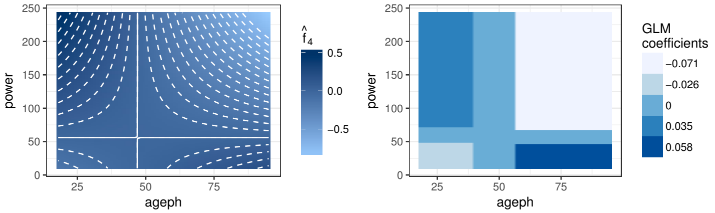


Figure 2.11: Comparison between the GAM effect and GLM coefficients for the interaction between `ageph` and `power` in the frequency model.

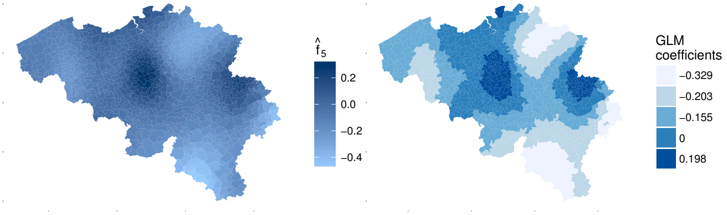


Figure 2.12: Comparison between the GAM effect and GLM coefficients for the spatial effect in the frequency model.

2.5.2 A comparative analysis

Table 2.7 compares the statistical performance of the original GAMs and the resulting GLMs via both the AIC and BIC measures. The GAMs attain a lower AIC value for both the frequency and severity models, whereas the GLMs attain a lower BIC value for both the frequency and severity models. This comparison clearly shows the trade-off between flexibility and simplicity in the modeling process. GAMs are the preferred tool for flexibility based on the low complexity penalty of AIC, while GLMs are the preferred tool for simplicity based on the high complexity penalty of BIC.

AIC	Frequency	Severity
GAM	**124 630**	**65 593**
GLM	124 646	65 603

BIC	Frequency	Severity
GAM	125 121	65 706
GLM	**124 926**	**65 696**

Table 2.7: Comparison of statistical performance between the original GAMs and resulting GLMs via both the AIC and BIC measures.

We calculate the pure premium π_i for every policyholder i as the product of the expected values of the frequency F_i and the severity S_i: $\pi_i = \mathbb{E}[F_i] \times \mathbb{E}[S_i]$, taking the actual exposure and risk factors of insured i into account. We use the GAM (resp. GLM) frequency and severity models to obtain the GAM (resp. GLM) premium structure. By summing these premiums over all policyholders we obtain the total pure premium inflow: $\Pi = \sum_{i=1}^{n} \pi_i$. At portfolio level, the GLM and the GAM result in a pure premium cash inflow of 29,867,987 and 29,865,859 Euro respectively. The GLM premium inflow is 1,706 Euro higher compared to the GAM inflow, but this difference represents only 0.007% of the total pure premium income. An insurance company therefore obtains nearly the same cash inflow by using any of the two models. The premium inflow is in both cases sufficient to cover the actual losses in the portfolio, which amount to 26,464,970 Euro. Note that policyholders with very large losses are excluded from this comparison in order to stay in line with the modeling process as outlined in Section 2.3.2.

Figure 2.13 shows a comparison between the annual GLM and GAM pure premiums by setting `exp` $= 1$ for all policyholders in the insurance portfolio. The policyholders are ordered, from left to right, according to increasing GAM premium in the top panel. We observe that the GLM premiums (light blue) are scattered around the GAM premiums (dark blue), but both show the same increasing trend. Differences between both premiums for a specific policyholder are attributable to differences between both the underlying frequency as well as severity models. The bottom left panel in Figure 2.13 shows violin plots of both the GLM and GAM premiums. We can clearly observe that the premiums resulting from both models are very similar. There is no visual difference in the bodies of both distributions. This indicates that, on average, there is no difference in premiums for the bulk of the policyholders. Our findings are confirmed by examining the relative pure premium differences, defined as $(\pi_i^{GLM} - \pi_i^{GAM})/\pi_i^{GAM}$, in the bottom right panel of Figure 2.13. A negative (positive) difference therefore indicates that the policyholder pays less (more) under the GLM compared to the GAM. The differences are centered around zero, again indicating that policyholders, on average, pay the same premium in the GLM and GAM case. Note that the difference in both premiums lies between -4.7% and 5.3% for half the insurance portfolio and that the median difference is equal to 0.17%.

Figure 2.14 shows the distribution of the relative premium difference over the continuous risk factors `ageph`, `power` and `bm`. The dots indicate the average of the relative differences for policyholders with that specific risk characteristic while the error bars indicate plus and minus two times the standard deviation. The top panel shows that for most ages the premium difference is close to zero. Both very young and very old policyholders pay less under the GLM compared to the GAM, which is in line with our findings in Figure 2.10. The middle panel indicates that policyholders driving a car with low (high) horsepower pay more (less) under the GLM compared to GAM, while for the other policyholders the premium difference is close to zero, again in line with our findings in Figure 2.10.

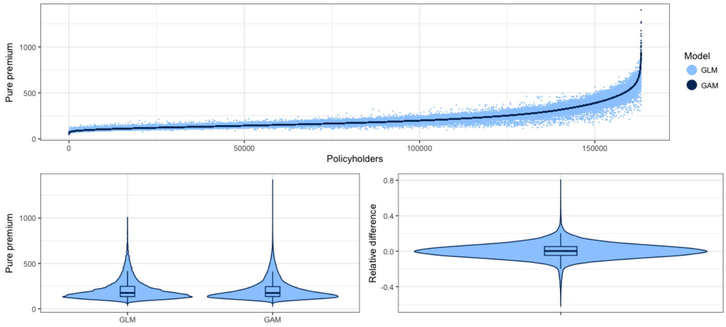


Figure 2.13: Comparison between the GAM and GLM pure premiums.

The bottom panel shows that for nearly every `bm` level the premium difference is close to zero. The GLM coefficients approximate the GAM effects very closely for the low `bm` levels and the approximation errors of the frequency and severity models offset each other for high `bm` levels. For all continuous risk factors we can conclude that the premium differences are close to zero in areas containing a large number of policyholders (cfr. Figure 2.2).

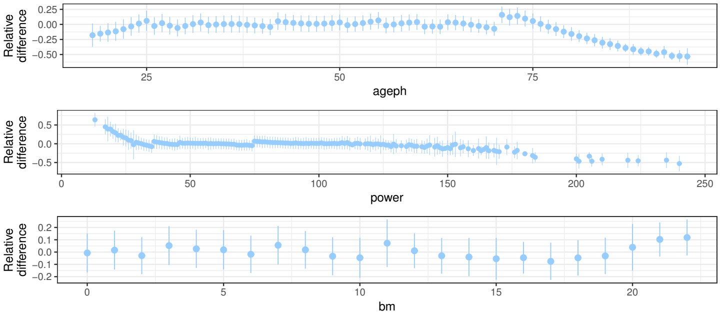


Figure 2.14: Distribution of the premium differences over the continuous risk factors `ageph`, `power` and `bm`.

For solvency purposes it is very important to hold enough capital such that the insurance company can fulfill its obligations towards its policyholders. We simulate 5,000,000 GLM and GAM scenarios by sampling observations from the estimated Poisson and lognormal distributions for frequency and severity respectively. The left panel of Figure 2.15 shows the distribution of the portfolio losses in these scenarios for both the GLM and GAM. These distributions look very similar, indicating that both models predict similar portfolio losses. The right panel of Figure 2.15 shows the high empirical quantiles of these losses, from the 90% quantile onwards. These high quantiles can be interpreted as a Value at Risk (VaR), a very popular measure to calculate capital requirements. We observe that the GLM and the GAM approach result in very similar values for the VaR. Table 2.8 shows the numerical values for the VaR measure at four often used quantiles. The GLM results in slightly higher capital requirements than the GAM in three out of four cases, but differences between both are extremely small ($< 0.01\%$).

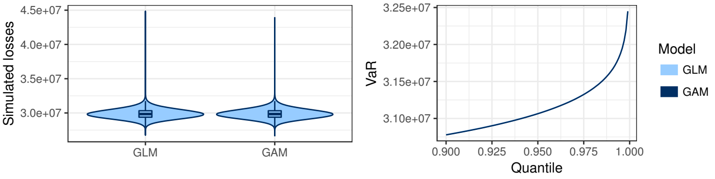


Figure 2.15: Comparison of the simulated losses and quantiles for GAM/GLM.

VaR	95%	99%	99.5%	99.9%
GAM	31,064,090	31,641,996	31,875,657	32,449,219
GLM	31,066,851	31,644,391	31,878,921	32,448,961

Table 2.8: Comparison of the 95%, 99%, 99.5% and 99.9% VaR for GAM/GLM.

2.6 Conclusions

This chapter presents a fully data-driven strategy to bin spatial and continuous risk factors with the goal of deploying these in a practical insurance tariff. We develop a predictive model for claim frequency as well as severity. Combining both models allows the actuary to calculate a pure risk premium for an insurance contract. We start from the framework of generalized additive models and build close approximations to these models. As such an easy to understand predictive model results which is formulated within the well-known framework of generalized linear models. Starting from a GAM with flexible effects for continuous risk factors, their interactions and a spatial effect, we propose to bin the spatial effect using Fisher's natural breaks algorithm and the effects of the continuous risk factors using evolutionary trees.

With Fisher's natural breaks we minimize the within-bin variance and hence produce a homogeneous binning of the spatial effect. This method is very similar to the K-means clustering algorithm (see MacQueen, 1967), though uses dynamic programming and will always return the same (optimal) binning result.

K-means starts optimizing after a random initialization and the end result therefore depends on this initialization. Fisher's method is only applicable to one-dimensional binning problems whereas K-means generalizes towards higher dimensional settings. This is not a problem in our case since binning the spatial effect is a one-dimensional problem.

The evolutionary trees are very good at binning the main effects of continuous risk factors. They focus on splitting a smooth effect in ranges where it has a large slope, but only if sufficient policyholders are present in these ranges. As such, the composition of the current portfolio is taken into account and our approach will automatically adapt to changing portfolio compositions. The evolutionary trees also perform well at discriminating between areas of increased and decreased risk in a two-dimensional setting for interaction effects of continuous risk factors. In our case study, this leads to a neutral zone grouping policyhoders with very similar risk profiles and regions containing policyholders with increased or decreased risk profiles. A downside of using a single regression tree to split a bivariate interaction effect is the fact that the resulting bins will always have a rectangular shape. It is therefore not possible to produce a split along a non-linear contour line such as those encountered in Figure 2.4. More complex models - for example boosted trees or support vector machines - are needed to obtain non-linear borders. The problem with such models however is their lack of interpretability; a non-linear border is harder to explain than the straight borders produced by a regression tree.

Our approach leads to GLMs which involve a simpler tariff structure compared to the original GAMs. We observed that premiums, simulated losses and capital requirements calculated from the resulting GLM approximate those from the original GAM quite closely. We therefore end up with a simpler model that can be deployed in practice as a close substitute for a more complex, flexible model.

Our study did not consider car brands and models as risk factors in the tariff. These are examples of multi-level factors, i.e. factor variables with a huge number of levels. Ohlsson (2008) demonstrates how (hierarchical) multi-level factors, such as car brands and models, can be estimated using credibility theory in an iterative algorithm. The spatial effect can be processed in the same way if one chooses not to model it as a bivariate smooth function in the GAM framework. A possible strategy is then to bin all the continuous risk factors in the GAM in order to obtain a GLM with only categorical risk factors. Afterwards all multi-level factors, such as the car brand, the car model and the municipality, can be incorporated in the tariff structure by using the iterative procedure outlined in Ohlsson (2008).

Our approach illustrates the use of data analytics within insurance pricing. This field is rapidly gaining importance in the era of big data. We focus on the interplay between the traditional toolkit of the pricing actuary (e.g. GLMs) and tools from the machine learning community (e.g. regression trees and genetic algorithms). We illustrate the complementarity of these techniques in pricing practice and how they can assist an actuary in finding the right balance between flexibility and simplicity. The application of our approach is not limited to the case of car insurance or P&C insurance in general. It can be used in every business environment where it is useful to approximate flexible smooth effects with more interpretable and simpler predictive models that are easier to explain to stakeholders. An obvious example is credit scoring, since a credit scoring model needs to be transparent and easy to explain to customers.

Chapter 3

Tree-based machine learning with a focus on interpretability

Pricing actuaries typically operate within the framework of generalized linear models (GLMs). With the upswing of data analytics, our study puts focus on machine learning methods to develop full tariff plans built from both the frequency and severity of claims. We adapt the loss functions used in the algorithms such that the specific characteristics of insurance data are carefully incorporated: highly unbalanced count data with excess zeros and varying exposure on the frequency side combined with scarce, but potentially long-tailed data on the severity side. A key requirement is the need for transparent and interpretable pricing models which are easily explainable to all stakeholders. We therefore focus on machine learning with decision trees: starting from simple regression trees, we work towards more advanced ensembles such as random forests and boosted trees. We show how to choose the optimal tuning parameters for these models in an elaborate cross-validation scheme, we present visualization tools to obtain insights from the resulting models and the economic value of these new modeling approaches is evaluated. Boosted trees outperform the classical GLMs, allowing the insurer to form profitable portfolios and to guard against potential adverse risk selection.

This chapter is based on joint work with Marie-Pier Côté, Katrien Antonio and Roel Verbelen, as published in the North-American Actuarial Journal (Henckaerts et al., 2021b). This research was supported by the Research Foundation - Flanders (FWO) [SB grant 1S06018N] and by the Society of Actuaries James C. Hickman Scholar Program. Katrien Antonio acknowledges financial support from the Ageas Research Chair at KU Leuven and from KU Leuven's research council (COMPACT C24/15/001). We are grateful for the access to the computing resources of the Department of Mathematics and Statistics at McGill University and the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI. Thanks are also due to Christian Genest for supporting this project in various ways and to Michel Denuit for sharing the data analyzed in this chapter.

3.1 Introduction

Insurance companies bring security to society by offering protection against financial losses. They allow individuals to trade uncertainty for certainty, by transferring the risk to the insurer in exchange for a fixed premium. An insurer sets the price for an insurance policy before its actual cost is revealed. Due to this phenomenon, known as the reverse production cycle of the insurance business, it is of vital importance that an insurer properly assesses the risks in its portfolio. To this end, tools from predictive modeling come in handy.

The insurance business is highly data driven and partly relies on algorithms for decision making. In order to price a contract, property and casualty (P&C, or: general, non-life) insurers predict the loss cost y for each policyholder based on his/her observable characteristics $\boldsymbol{x}$. The insurer therefore develops a predictive model f, mapping the risk factors $\boldsymbol{x}$ to the predicted loss cost $\hat{y}$ by setting $\hat{y} = f(\boldsymbol{x})$. For simplicity, this predictive model is usually built in two stages by considering separately the frequency and severity of the claims. Generalized linear models (GLMs), introduced by Nelder and Wedderburn (1972), are the industry standard to develop state-of-the-art analytic insurance pricing models (Haberman and Renshaw, 1997; De Jong and Heller, 2008; Frees, 2015). Pricing actuaries often code all risk factors in a categorical format, either based on expert opinions (Frees and Valdez, 2008; Antonio et al., 2010) or in a data-driven way (see Chapter 2). GLMs involving only categorical risk factors result in predictions available in tabular format, that can easily be translated into interpretable tariff plans.

Technological advancements have boosted the popularity of machine learning and big data analytics, thereby changing the landscape of predictive modeling in many business applications. However, few papers in the insurance literature go beyond the actuarial comfort zone of GLMs. Dal Pozzolo (2010) contrasts the performance of various machine learning techniques to predict claim frequency in the Allstate Kaggle competition. Guelman (2012) compares GLMs and gradient boosted trees for predicting the accident loss cost of auto at-fault claims. Liu et al. (2014) approach the claim frequency prediction problem using multi-class AdaBoost trees. Wüthrich and Buser (2019) and Zöchbauer et al. (2017) show how tree-based machine learning techniques can be adapted to model claim frequencies. Yang et al. (2018) predict insurance premiums by applying a gradient boosted tree algorithm to Tweedie models. Pesantez-Narvaez et al. (2019) employ XGBoost to predict the occurrence of claims using telematics data. Ferrario et al. (2018) and Schelldorfer and Wüthrich (2019) propose neural networks to model claim frequency, either directly or via a nested GLM. Machine learning techniques have also been used in other insurance applications, such as policy conversion or customer retention (Spedicato et al., 2018), renewal pricing (Krasheninnikova et al., 2019) and claim fraud detection (Wang and Xu, 2018).

Insurance pricing models are heavily regulated and they must meet specific requirements before being deployed in practice, posing some challenges for machine learning algorithms. Firstly, the European Union’s General Data Protection Regulation (GDPR, 2016), effective May 25, 2018, establishes a regime of “algorithmic accountability" of decision-making machine algorithms. By law, individuals have the right to an explanation of the logic behind the decision (Kaminski, 2018), which means that pricing models must be transparent and easy to communicate to all stakeholders. Qualified transparency (Pasquale, 2015) implies that customers, managers and the regulator should receive information in different degrees of scope and depth. Secondly, every policyholder should be charged a fair premium, related to his/her risk profile, to minimize the potential for adverse selection (Dionne et al., 1999). If the heterogeneity in the portfolio is not carefully reflected in the pricing, the good risks will be prompt to lapse and accept a lower premium elsewhere, leaving the insurer with an inadequately priced portfolio. Thirdly, the insurer has the social role of creating solidarity among the policyholders. The use of machine learning for pricing should in no way lead to an extreme “personalization of risk” or discrimination, e.g., in the form of extremely high premiums (O’Neil, 2016) for some risk profiles that actually entail no risk transfer. By finding a trade-off between customer segmentation and risk pooling, the insurer avoids adverse selection while offering an effective insurance product involving a risk transfer for all policyholders. In a regime of algorithmic accountability, insurers should be held responsible for their pricing models in terms of transparency, fairness and solidarity. It is therefore very important to be able to “look under the hood” of machine learning algorithms and the resulting pricing models. That is exactly one of the goals of this chapter.

In this chapter, we study how tree-based machine learning methods can be applied to insurance pricing. The building blocks of these techniques are decision trees, covered in Friedman et al. (2001). These are simple predictive models that mimic human decision-making in the form of yes-no questions. In insurance pricing, a decision tree partitions a portfolio of policyholders into groups of homogeneous risk profiles based on some characteristics. The partition of the portfolio is directly observable, resulting in high transparency. For each subgroup, a constant prediction is put forward, automatically inducing solidarity among the policyholders in a subgroup (as long as the size of this subgroup is large enough). These aspects of decision trees make them good candidates for insurance pricing. However, the predictive performance of such simple trees tends to be rather low. We therefore consider more complex algorithms that combine multiple decision trees in an ensemble, i.e., tree-based machine learning. These ensemble techniques usually provide better predictive performance, but at the cost of less transparency. We employ model interpretation tools to understand these “black boxes”, allowing us to interpret the resulting models and underlying decision process. Tree-based machine learning techniques are often praised for their ability to discover interaction effects in the data, a very useful insight for insurers that will be explored in this chapter.

Insurance claim data typically entails highly imbalanced count data with excess zeros and varying exposure-to-risk on the frequency side, combined with long- or even heavy-tailed continuous data on the severity side. Standard machine learning algorithms typically deal with data that is more normal-like or balanced. Guelman (2012) models the accident loss cost by simplifying the frequency count regression problem into a binary classification task. This however cannot factor in varying exposure-to-risk and leads to a loss of information regarding policyholders who file more than one claim in a period. Wüthrich and Buser (2019) and Zöchbauer et al. (2017) show how the specific data features on the frequency side can be taken into account for regression.

We extend the existing literature by also putting focus on the severity side of claims and obtaining full tariff plans on real-world claims data from an insurance portfolio. We develop an elaborate cross-validation scheme instead of relying on built-in routines from software packages and we take into account multiple types of risk factors: categorical, continuous and spatial information. The goal of this chapter is to investigate how tree-based pricing models perform compared to the classical actuarial approach with GLMs. This comparison puts focus on statistical performance, interpretation and business implications. We go beyond a purely statistical comparison, but acknowledge the fact that the resulting pricing model has to be deployed, after marketing adjustments, in a business environment with specific requirements.

The rest of this chapter is structured as follows. Section 3.2 introduces the basic principles and guidelines for building a benchmark pricing GLM. Section 3.3 consolidates the important technical details on tree-based machine learning. Section 3.4 presents interpretations from the optimal frequency and severity models fitted on a Belgian insurance data set, together with an out-of-sample performance comparison. Section 3.5 reviews the added value from a business angle and Section 3.6 concludes this chapter. In an accompanying online supplement, available at `https://github.com/henckr/sevtree`, we provide more details on the construction and interpretation of tree-based machine learning methods for the severity.

3.2 State-of-the-art insurance pricing models

To assess the possible merits of tree-based machine learning for insurance pricing, we first have to establish a fair benchmark pricing model that meets industry standards. GLMs are by far the most popular pricing models in today's industry. This section outlines the basic principles and steps for creating a benchmark pricing GLM with the strategy from Chapter 2.

A P&C insurance company is interested in the total loss amount L per unit of exposure-to-risk e, where L is the total loss for the N claims reported by a policyholder during the exposure period e. P&C insurers usually opt for a so-called frequency-severity strategy to price a contract (Denuit et al., 2007; Frees et al., 2014; Parodi, 2014). Claim frequency F is the number of claims N filed per unit of exposure-to-risk e. Claim severity S refers to the cost per claim and is defined by the average amount per claim filed, that is the total loss amount L divided by the number of claims N. The pure (or: risk) premium π then follows as:

$$\pi = \mathbb{E}\left(\frac{L}{e}\right) \overset{\text{indep.}}{=} \mathbb{E}\left(\frac{N}{e}\right) \times \mathbb{E}\left(\frac{L}{N} \mid N > 0\right) = \mathbb{E}(F) \times \mathbb{E}(S),$$

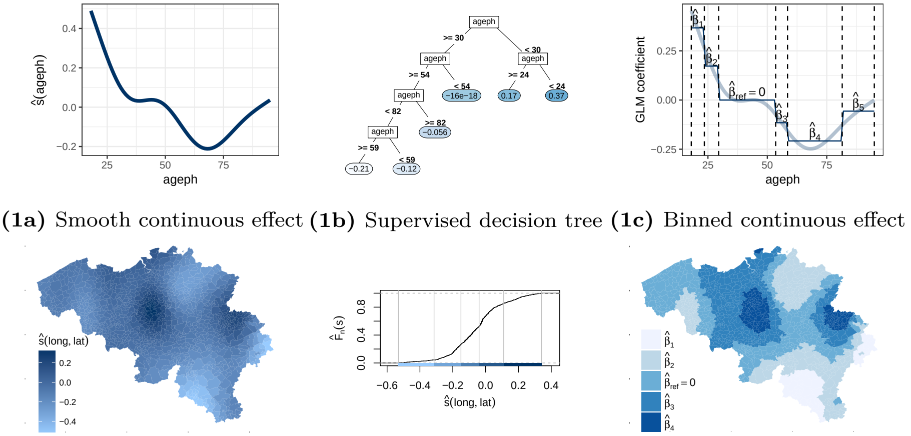
assuming independence between the frequency and the severity component of the premium (Klugman et al., 2012). Alternatives, allowing dependence between F and S, are investigated in the literature (Gschlößl and Czado, 2007; Czado et al., 2012; Garrido et al., 2016).

Predictive models for both F and S are typically developed within the framework of GLMs. Let Y, the response variable of interest, follow a distribution from the exponential family. The structure of a GLM with all explanatory variables in a categorical format is:

$$\eta = g(\mu) = \boldsymbol{z}^\top \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{q} \beta_j z_j \,, \tag{3.1}$$

with η the linear predictor, $g(\cdot)$ the link function and μ the expectation of Y. The $q+1$ dimensional $0/1$-valued vector $\boldsymbol{z}$ contains a 1 for the intercept together with the q dummy variables expressing the policyholder's risk profile. In a claim frequency model, the response variable N typically follows a count distribution such as the Poisson. Assuming $g(\cdot) = \ln(\cdot)$, the model may account for exposure-to-risk through an offset $\ln(e)$ such that the risk premium is proportional to the exposure. In a claim severity model, the response variable L/N typically follows a right skewed distribution with a long right tail such as the gamma or log-normal. Only policyholders filing at least one claim, i.e., $N > 0$, contribute to the severity model calibration and the number of claims N is used as a case weight in the regression (Denuit and Lang, 2004).

Chapter 2 details a data-driven strategy to build a GLM with all risk factors in a categorical format. This strategy aligns the practical requirements of a business environment with the statistical flexibility of generalized additive models (GAMs, documented by Wood, 2006). GAMs extend the linear predictor in Eq. (3.1) with (multidimensional) smooth functions. After an exhaustive search, the starting point for both frequency and severity is a flexible GAM with smooth effects for continuous risk factors, including two-way interactions, and a smooth spatial effect. These smooth effects are used to bin the continuous and spatial risk factors, thereby transforming them to categorical variables. Figure 3.1 schematizes how decision trees and unsupervised clustering are applied to achieve this binning. The output of this framework is an interpretable pricing GLM, which serves as benchmark pricing model in this study.



(1a) Smooth continuous effect **(1b)** Supervised decision tree **(1c)** Binned continuous effect

(2a) Smooth spatial effect **(2b)** Unsupervised clustering **(2c)** Binned spatial effect

Figure 3.1: Schematic overview of the binning strategy of Chapter 2 for a continuous risk factor (1a - 1c) and a spatial risk factor (2a - 2c). The smooth GAM effect of a continuous risk factor (1a) is fed as a response to a decision tree (1b), which splits the continuous risk factor into bins (1c). The smooth spatial effect (2a) is clustered in an unsupervised way (2b), resulting in groups of postcode areas (2c). These categorical risk factors are used in a GLM.

3.3 Tree-based machine learning methods

Section 3.3.1 introduces the essential algorithmic details needed for understanding the tree-based modeling techniques used in this chapter. We consider regression trees (Breiman et al., 1984), random forests (Breiman, 2001) and gradient boosting machines (Friedman, 2001) as alternative modeling techniques for insurance pricing. These models rely on the choice of a loss function, which has to be tailored to the characteristics of insurance data as we motivate in Section 3.3.2. In Section 3.3.3 and 3.3.4, we explain our tuning strategy and present interpretation tools.

3.3.1 Algorithmic essentials

Regression tree Decision trees partition data based on yes-no questions, predicting the same value for each member of the constructed subsets. A popular approach to construct decision trees is the Classification And Regression Tree (CART) algorithm, introduced by Breiman et al. (1984). The predictor space R is the set of possible values for the p variables $x_1, \ldots, x_p$, e.g., $R = \mathbb{R}^p$ for p unbounded, continuous variables and $R = [\min(x_1), \max(x_1)] \times [\min(x_2), \max(x_2)]$ for two bounded, continuous variables. A tree divides the predictor space R into J distinct, non-overlapping regions $R_1, \ldots, R_J$. In the jth region, the fitted response $\hat{y}_{R_j}$ is computed as a (weighted) average of the training observations falling in that region. The regression tree predicts a (new) observation with characteristics $\boldsymbol{x}$ as follows:

$$f_{\text{tree}}(\boldsymbol{x}) = \sum_{j=1}^{J} \hat{y}_{R_j} \, \mathbb{1}(\boldsymbol{x} \in R_j). \tag{3.2}$$

The indicator $\mathbb{1}(A)$ equals one if event A occurs and zero otherwise. As the J regions are non-overlapping, the indicator function differs from zero for exactly one region for each $\boldsymbol{x}$. A tree therefore makes the same constant prediction $\hat{y}_{R_j}$ for the entire region R_j.

It is computationally impractical to consider every possible partition of the predictor space R in J regions, therefore CART uses a top-down greedy approach known as recursive binary splitting. From the full predictor space R, the algorithm selects a splitting variable x_v with $v \in \{1, \ldots, p\}$ and a cut-off c such that $R = R_1(v, c) \cup R_2(v, c)$ with $R_1(v, c) = \{R \,|\, x_v \leqslant c\}$ and $R_2(v, c) = \{R \,|\, x_v > c\}$. This forms two nodes in the tree, one containing the observations satisfying $x_v \leqslant c$ and the other containing the observations satisfying $x_v > c$. For a categorical splitting variable, the corresponding factor levels are replaced by their empirical response averages, see Section 8.8 in Breiman et al. (1984). These averages are sorted from low to high and a cut-off c is chosen such that the factor levels are split into two groups. The splitting variable x_v and cut-off c are chosen such that their combination results in the largest improvement in a carefully picked loss function $\mathscr{L}(\cdot\,, \cdot)$. For $i = \{1, \ldots, n\}$, where n is the number of observations in the training set, the CART algorithm searches for x_v and c minimizing the following summations:

$$\sum_{i\,:\,\boldsymbol{x}_i \in R_1(v,c)} \mathscr{L}(y_i, \hat{y}_{R_1}) \;+\; \sum_{i\,:\,\boldsymbol{x}_i \in R_2(v,c)} \mathscr{L}(y_i, \hat{y}_{R_2}).$$

A standard loss function is the squared error loss, but we present more suitable loss functions for claim frequency or severity data in Section 3.3.2. In a next iteration, the algorithm splits R_1 and/or R_2 in two regions and this process is repeated recursively until a stopping criterion is satisfied. This stopping criterion typically puts a predefined limit on the size of a tree, e.g., a minimum improvement in the loss function (Breiman et al., 1984), a maximum tree depth or a minimum number of observations in a tree node (Friedman et al., 2001).

A large tree is likely to overfit the data and does not generalize well to new data, while a small tree is likely to underfit the data and fails to capture the general trends. This is related to the bias-variance tradeoff (Friedman et al., 2001) meaning that a large tree has low bias and high variance while a small tree has high bias but low variance. To prevent overfitting, the performance of a tree is penalized by the number of regions J as follows:

$$\sum_{j=1}^{J} \sum_{i\,:\,\boldsymbol{x}_i \in R_j} \mathscr{L}(y_i, \hat{y}_{R_j}) \;+\; J \cdot cp \cdot \sum_{i\,:\,\boldsymbol{x}_i \in R} \mathscr{L}(y_i, \hat{y}_R)\,, \tag{3.3}$$

where the first part assesses the goodness of fit and the second part is a penalty measuring the tree complexity. The strength of this penalty is driven by the complexity parameter *cp*, a tuning parameter (see Section 3.3.3 for details on the tuning strategy). A large (small) value for *cp* puts a high (low) penalty on extra splits and will result in a small (large) tree. The complexity parameter *cp* is usually scaled with the loss function evaluated for the root tree, which is exactly the last summation in Eq. (3.3); see Remark 3.8 in Zöchbauer et al. (2017). This ensures that $cp = 1$ delivers a root tree without splits capturing an overall y estimate (denoted $\hat{y}_R$ in Eq. (3.3)) and $cp = 0$ results in the largest possible tree allowed by the stopping criterion.

Figure 3.2 depicts an example of a regression tree for claim frequency data. The rectangles are internal nodes which partition observations going from top to bottom along the tree. The top node, splitting on the bonus-malus level `bm`, is called the root node. The ellipses are leaf nodes, containing prediction values for the observations ending in that specific leaf. Going from left to right, the leaf nodes are ordered from low (light blue) to high (dark blue) prediction values. Decision trees have many advantages because their predictions are highly explainable and interpretable, both very important criteria for regulators. The downside of trees however is that the level of predictive accuracy tends to be lower compared to other modeling techniques. This is mainly driven by the high variance of a tree, e.g., slight changes in the data can result in very different trees and therefore rather different predictions for certain observations. The predictive performance can be substantially improved by aggregating multiple decision trees in ensembles of trees, thereby reducing the variance. That is the idea behind popular ensemble methods, such as random forests and gradient boosting machines, which are discussed next.

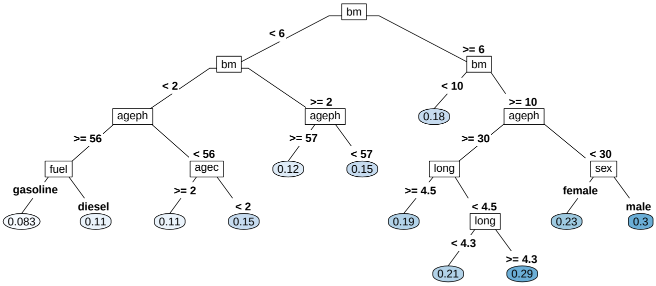


Figure 3.2: Visual representation of a regression tree for claim frequency with nodes (rectangles) containing the splitting variable x_v, edges (lines) representing the splits with cut-off c and leaf nodes (ellipses) containing the prediction values $\hat{y}_{R_j}$. The variable names are defined in Table B.1.

Random forest Bagging, which is short for **b**ootstrap **agg**regat**ing** (Breiman, 1996), and random forests (Breiman, 2001) are similar ensemble techniques combining multiple decision trees. Bagging reduces the variance of a single tree by averaging the forecasts of multiple trees on bootstrapped samples of the original data. This stabilizes the prediction and improves the predictive performance compared to a single decision tree. Starting from the data set $\mathcal{D}$, the idea of bagging is to take bootstrap samples $\{\mathcal{D}_t\}_{t=1,\dots,T}$ and to build T decision trees, one for each $\mathcal{D}_t$ independently. The results are then aggregated in the following way:

$$f_{\text{bagg}}(\boldsymbol{x}) = \frac{1}{T} \sum_{t=1}^{T} f_{\text{tree}}(\boldsymbol{x} \,|\, \mathcal{D}_t)\,, \tag{3.4}$$

where the condition $(\,|\,\mathcal{D}_t)$ indicates the tree being developed on sample $\mathcal{D}_t$.

The performance improvement through variance reduction gets bigger when there is less correlation between the individual trees, see Lemma 3.25 in Zöchbauer et al. (2017). For that reason, the trees are typically grown deep (i.e., $cp = 0$ in Eq. (3.3)), until a stopping criterion is satisfied. Taking bootstrap samples of smaller sizes $\delta \cdot n$, with n the number of observations in $\mathcal{D}$ and $0 < \delta < 1$, decorrelates the trees further and reduces the model training time. However, a lot of variability remains within a bagged ensemble because the trees built on the bootstrapped data samples are still quite similar. This is especially the case when some explanatory variables in the data are much more predictive than the others. The important variables will dominate the first splits, causing all trees to be similar to one another. To prevent this, a random forest further decorrelates the individual trees by sampling variables during the growing process. At each split, m out of p variables are randomly chosen as candidates for the optimal splitting variable. Besides this adaptation, a random forest follows the same strategy as bagging and predicts a new observation according to Eq. (3.4). The random forest procedure is detailed in Algorithm 1 where T and m are treated as tuning parameters (see Section 3.3.3 for details on the tuning strategy).

```
for $t = 1, \dots, T$ do
    generate bootstrapped data $\mathcal{D}_t$ of size $\delta \cdot n$ by sampling with replacement from data $\mathcal{D}$;
    while stopping criterion not satisfied do
        randomly select $m$ of the $p$ variables;
        find the optimal splitting variable $x_v$ from the $m$ options together with cut-off $c$;
$f_{\text{rf}}(\boldsymbol{x}) = \frac{1}{T} \sum_{t=1}^{T} f_{\text{tree}}(\boldsymbol{x} \mid \mathcal{D}_t)$;
```

Algorithm 1: Procedure to build a random forest model.

A random forest improves the predictive accuracy obtained with a single decision tree by using more, and hopefully slightly different, trees to solve the problem at hand. However, the trees in a random forest are built independently from each other (i.e., the `for` loop in Algorithm 1 can be run in parallel) and do not share information during the training process.

Gradient boosting machine In contrast to random forests, boosting is an iterative statistical method that combines many weak learners into one powerful predictor. Friedman (2001) introduced decision trees as weak learners; each tree improves the current model fit, thereby using information from previously grown trees. At each iteration, the pseudo-residuals are used to assess the regions of the predictor space for which the model performs poorly in order to improve the fit in a direction of better overall performance. The pseudo-residual $\rho_{i,t}$ for observation i in iteration t is calculated as the negative gradient of the loss function $-\partial \mathscr{L}\{y_i, f(\boldsymbol{x}_i)\}/\partial f(\boldsymbol{x}_i)$, evaluated at the current model fit. This typical approach called stepwise gradient descent ensures that a lower loss is obtained at the next iteration, until convergence. The boosting method learns slowly by fitting a small tree of depth d (with a squared error loss function) to these pseudo-residuals, improving the model fit in areas where it does not perform well. For each region R_j of that tree, the update $\hat{b}_j$ is calculated as the constant that has to be added to the previous model fit to minimize the loss function, namely b that minimizes $\mathscr{L}\{y_i, f(\boldsymbol{x}_i) + b\}$ over this region. A shrinkage parameter λ controls the learning speed by shrinking updates for $\boldsymbol{x} \in R_j$ as follows: $f_{new}(\boldsymbol{x}) = f_{old}(\boldsymbol{x}) + \lambda \cdot \hat{b}_j$. A lower λ usually results in better performance but also increases computation time because more trees are needed to converge to a good solution. Typically, λ is fixed at the lowest possible value within the computational constraints (Friedman, 2001). The collection of T trees at the final iteration is used to make predictions.

Stochastic gradient boosting, introduced by Friedman (2002), injects randomness in the training process. In each iteration, the model update is computed from a randomly selected subsample of size $\delta \cdot n$. This improves both the predictive accuracy and model training time when $\delta < 1$. The (stochastic) gradient boosting machine algorithm is given in Algorithm 2 where T and d are treated as tuning parameters (see Section 3.3.3 for details on the tuning strategy).

```
initialize fit to the optimal constant model: $f_0(\boldsymbol{x}) = \arg\min_b \sum_{i=1}^{n} \mathscr{L}(y_i, b)$;
for $t = 1, \dots, T$ do
    randomly subsample data of size $\delta \cdot n$ without replacement from data $\mathcal{D}$;
    for $i = 1, \dots, \delta \cdot n$ do
        $\rho_{i,t} = -\left[\frac{\partial \mathscr{L}\{y_i, f(\boldsymbol{x}_i)\}}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{t-1}}$
    fit a tree of depth $d$ to the pseudo-residuals $\rho_{i,t}$ resulting in regions $R_{j,t}$ for
     $j = 1, \dots, J_t$;
    for $j = 1, \dots, J_t$ do
        $\hat{b}_{j,t} = \arg\min_b \sum_{i : \boldsymbol{x}_i \in R_{j,t}} \mathscr{L}\{y_i, f_{t-1}(\boldsymbol{x}_i) + b\}$
    update $f_t(\boldsymbol{x}) = f_{t-1}(\boldsymbol{x}) + \lambda \sum_{j=1}^{J_t} \hat{b}_{j,t} \mathbb{1}(\boldsymbol{x} \in R_{j,t})$;
$f_{\text{gbm}}(\boldsymbol{x}) = f_T(\boldsymbol{x})$;
```

Algorithm 2: Procedure to build a (stochastic) gradient boosting machine.

3.3.2 Loss functions for insurance data

The machine learning algorithms discussed in Section 3.3.1 require the specification of a loss (or: cost) function that is to be minimized during the training phase of the model. We first present a general discussion on the loss function choice, followed by details on the R implementation.

Loss functions The standard loss function for regression problems is the squared error loss:

$$\mathscr{L}\{y_i, f(\boldsymbol{x_i})\} \propto \{y_i - f(\boldsymbol{x}_i)\}^2 \,,$$

where y_i is the observed response and $f(\boldsymbol{x}_i)$ is the prediction of the model for variables $\boldsymbol{x}_i$. However, the squared error is not necessarily a good choice when modeling integer-valued frequency data or right-skewed severity data. We use the concept of deviance to make this idea clear. The deviance is defined as $D\{y, f(\boldsymbol{x})\} = -2 \cdot \ln[\mathcal{L}\{f(\boldsymbol{x})\}/\mathcal{L}(y)]$, a likelihood ratio where $\mathcal{L}\{f(\boldsymbol{x})\}$ is the model likelihood and $\mathcal{L}(y)$ the likelihood of the saturated model (i.e., the model in which the number of parameters equals the number of observations). The condition $\mathcal{L}\{f(\boldsymbol{x})\} \leqslant \mathcal{L}(y)$ always holds, so the ratio of likelihoods is bounded from above by one. For competing model fits, the best one obtains the lowest deviance value on holdout data. We therefore use a loss function $\mathscr{L}(\cdot\,,\cdot)$ such that $D\{y, f(\boldsymbol{x})\} = \sum_{i=1}^{n} \mathscr{L}\{y_i, f(\boldsymbol{x}_i)\}$. This idea was put forward by Venables and Ripley (2002) for general classification and regression problems.

Assuming constant variance, the normal (or: Gaussian) deviance is expressed as follows:

$$\begin{aligned} D\{y, f(\boldsymbol{x})\} &= 2 \ln \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - y_i)^2\right\} - 2 \ln \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}\{y_i - f(\boldsymbol{x}_i)\}^2\right] \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \{y_i - f(\boldsymbol{x}_i)\}^2 \,, \end{aligned}$$

which boils down to a scaled version of the sum of squared errors. This implies that a loss function based on the squared error is appropriate when the normal assumption is reasonable. More generally, the squared error is suitable for any continuous distribution symmetrical around its mean with constant variance, i.e., any elliptical distribution. However, claim frequency and severity data do not follow any elliptical distribution, as we show in Section 3.4.1. Therefore, in an actuarial context, Wüthrich and Buser (2019) and Zöchbauer et al. (2017) propose more suitable loss functions inspired by the GLM pricing framework from Section 3.2.

Claim frequency modeling involves count data, typically assumed to be Poisson distributed in GLMs. Therefore, an appropriate loss function is the Poisson deviance, defined as follows:

$$\begin{aligned} D(y, f(\boldsymbol{x})) &= 2 \ln \prod_{i=1}^{n} \exp(-y_i) \frac{y_i^{y_i}}{y_i!} - 2 \ln \prod_{i=1}^{n} \exp\{-f(\boldsymbol{x}_i)\} \frac{f(\boldsymbol{x}_i)^{y_i}}{y_i!} \\ &= 2 \sum_{i=1}^{n} \left[y_i \ln \frac{y_i}{f(\boldsymbol{x}_i)} - \{y_i - f(\boldsymbol{x}_i)\} \right] . \end{aligned} \tag{3.5}$$

When using an exposure-to-risk measure e_i, $f(\boldsymbol{x}_i)$ is replaced by $e_i \cdot f(\boldsymbol{x}_i)$ such that the exposure is taken into account in the expected number of claims. Thus, the Poisson deviance loss function can account for different policy durations.

Predictions from a Poisson regression tree in Eq. (3.2) are equal to the sum of the number of claims divided by the sum of exposure for all training observations in each leaf node: $\hat{y}_{R_j} = \sum_{i \in I_j} N_i / \sum_{i \in I_j} e_i$ for $I_j = \{i : \boldsymbol{x}_i \in R_j\}$. This optimal estimate is obtained by setting the derivative of Eq. (4.5) with respect to f equal to zero. As a tree node without claims leads to a division by zero in the deviance calculation, an adjustment can be made to the implementation with a hyper-parameter that will be introduced in Section 3.3.3.

Right-skewed and long-tailed severity data is typically assumed to be gamma or log-normally distributed in GLMs. In Section 3.4, we present the results obtained with the gamma deviance as our preferred model choice, but a discussion on the use of the log-normal deviance is available in the supplementary material. The gamma deviance is defined as follows:

$$\begin{aligned} D\{y, f(\boldsymbol{x})\} &= 2 \ln \prod_{i=1}^{n} \frac{1}{y_i \Gamma(\alpha)} \left(\frac{\alpha y_i}{y_i} \right)^{\alpha} \exp\left(-\frac{\alpha y_i}{y_i} \right) \\ &\qquad - 2 \ln \prod_{i=1}^{n} \frac{1}{y_i \Gamma(\alpha)} \left\{ \frac{\alpha y_i}{f(\boldsymbol{x}_i)} \right\}^{\alpha} \exp\left\{ -\frac{\alpha y_i}{f(\boldsymbol{x}_i)} \right\} \qquad (3.6) \\ &= 2 \sum_{i=1}^{n} \alpha \left\{ \frac{y_i - f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} - \ln \frac{y_i}{f(\boldsymbol{x}_i)} \right\}. \end{aligned}$$

The shape parameter α acts as a scaling factor and can therefore be ignored. When dealing with case weights, α can be replaced by the weights w_i. In severity modeling, the response is the average claim amount over N_i observed claims and N_i should be used as case weight. Predictions from a gamma regression tree in Eq. (3.2) are equal to the sum of the total loss amount divided by the sum of the number of claims for all training observations in each leaf node: $\hat{y}_{R_j} = \sum_{i \in I_j} L_i / \sum_{i \in I_j} N_i$ for $I_j = \{i : \boldsymbol{x}_i \in R_j\}$. This optimal estimate is obtained by setting the derivative of Eq. (3.6) with respect to f equal to zero.

Implementation in R Our results are obtained with two special purpose packages for tree-based machine learning in the statistical software R. For the regression trees and random forests, we developed our own package called `distRforest` (Henckaerts, 2020). For stochastic gradient boosting, we chose the implementation from Southworth (2015) of the `gbm` package, originally developed by Ridgeway (2014). Our `distRforest` package extends the `rpart` package by Therneau et al. (2019) such that it is capable of developing regression trees and random forests with our specific desired loss functions for both claim frequency and severity. We had to go beyond the standard implementations especially because of the loss functions appropriate for actuarial applications.

Although the `rpart` package supports the Poisson deviance for regression trees, it did not facilitate the use of a suitable loss function for severity data.

3.3.3 Tuning strategy

Tuning and hyper-parameters Table 3.1 gives an overview of the parameters used by the algorithms described in Section 3.3.1. Some of these are chosen with care (tuning parameters), while others are less influential and are set to a sensible predetermined value (hyper-parameters). Instead of relying on the built-in tuning strategies of the R packages mentioned in Section 3.3.2, we perform an extensive grid search to find the optimal values among a predefined tuning grid displayed in Table B.2 in Appendix B.2. We prefer a grid search above other tuning strategies, such as Bayesian optimization (Xia et al., 2017), for its ease of implementation while being a sound approach. The hyper-parameter κ enforces a stopping criterion for trees used across the three algorithms, ensuring that a split is not allowed if a resulting node would contain less than 1% of the observations. The hyper-parameter δ in Algorithms 1 and 2 specifies to develop the trees in the ensemble techniques on 75% of the available training data. The shrinkage parameter λ in Algorithm 2 is set at a low value for which computation time is still reasonable, namely 0.01. The parameter γ helps to avoid division by zero when optimizing the Poisson deviance in Eq. (4.5). This parameter is therefore only used when growing a regression tree and random forest for claim frequency. We refer the reader to Section 8.2 in Therneau and Atkinson (2019) for details on the `rpart` implementation. In short, a gamma prior is assumed on the Poisson rate parameter to keep it from becoming zero when there is no claim in a node. With $I_j = \{i : \boldsymbol{x}_i \in R_j\}$, the prediction in a node is adapted as follows:

$$\hat{y}^{\gamma}_{R_j} = \frac{\gamma^{-2} + \sum_{i \in I_j} N_i}{\gamma^{-2}/\hat{y}_R + \sum_{i \in I_j} e_i} \,.$$

Note that $\hat{y}^{\gamma}_{R_j} = \hat{y}_R$ for $\gamma = 0$ and $\hat{y}^{\gamma}_{R_j} = \hat{y}_{R_j} = \sum_{i \in I_j} N_i / \sum_{i \in I_j} e_i$ for $\gamma = \infty$.

Cross-validation Machine learning typically relies on training data to build a model, validation data to tune the parameters and test data to evaluate the out-of-sample performance of the model. In this chapter, we develop an extensive cross-validation scheme, inspired by K-fold cross-validation (Friedman et al., 2001), that serves two purposes. First, we tune the parameters in the algorithms under study with a 5-fold cross-validation approach. Second, we evaluate the predictive performance of the algorithms investigated on multiple data sets, instead of on a single test set. Algorithm 3 outlines the basic principles of our approach and Figure 3.3 gives a schematic representation. The full data

	Tuning parameters	Hyper-parameters
Regression tree	complexity parameter cp coefficient of variation gamma prior γ	$\kappa = 0.01$
Random forest	number of trees T number of split candidates m	$cp = 0$ $\gamma = 0.25$ $\kappa = 0.01$ $\delta = 0.75$
Gradient boosting machine	number of trees T tree depth d	$\lambda = 0.01$ $\kappa = 0.01$ $\delta = 0.75$

Table 3.1: Overview of the parameters for the different machine learning techniques.

$\mathcal{D}$ is split in six disjoint and stratified (Neyman, 1934) subsets $\mathcal{D}_1, \ldots, \mathcal{D}_6$ by ordering first on claim frequency, then on severity. The ordered observations are assigned to each of the subsets in turn. Stratification ensures that the distribution of response variables is similar in the six subsets, as we illustrate in Table 3.2 for the data introduced in Section 3.4.1. The `foreach` and inner `for` loop in Algorithm 3 represent the typical approach to perform 5-fold cross-validation on data from which we already separated a hold-out test set $\mathcal{D}_k$. The `foreach` loop iterates over the tuning grid and the `for` loop allows the validation set $\mathcal{D}_\ell$ to vary. The optimal tuning parameters are those that minimize the cross-validation error, which is obtained by averaging the error on the validation sets. The outer `for` loop in Algorithm 3 allows the hold-out test set to vary and model performance is evaluated on this test set $\mathcal{D}_k$. Advantages of evaluating a trained model on multiple test sets are threefold. First, we obtain multiple performance measures per model class which results in a more accurate performance assessment. Second, it allows to perform sensitivity checks to assess the stability of different algorithms. Third, it exempts us from the choice of a specific test set which could bias results.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$
$\sum N_i / \sum e_i$	0.1391687	0.1391433	0.1392443	0.1392213	0.1391517	0.1393045
$\sum L_i / \sum N_i$	1,296.165	1,302.894	1,324.667	1,312.619	1,330.884	1,287.832

Table 3.2: Summary statistics of response variables in the different data subsets $\mathcal{D}_1$ to $\mathcal{D}_6$.

3.3.4 Interpretability matters: opening the black box

The GDPR's regime of "algorithmic accountability" and the resulting "right to explanation" highlight the vital importance of interpretable and transparent

```
Input: model class (mclass) and corresponding tuning grid (tgrid)
split data $\mathcal{D}$ into 6 disjoint stratified subsets $\mathcal{D}_1, \ldots, \mathcal{D}_6$;
for $k = 1, \ldots, 6$ do
    leave out $\mathcal{D}_k$ as test set;
    foreach parameter combination in tgrid do
        for $\ell \in \{1, \ldots, 6\} \setminus k$ do
            train a model $f_{k\ell}$ of mclass on $\mathcal{D} \setminus \{\mathcal{D}_k, \mathcal{D}_\ell\}$;
            evaluate the model performance on $\mathcal{D}_\ell$ using loss function $\mathscr{L}(\cdot, \cdot)$;
            $\text{valid\_error}_{k\ell} \leftarrow \frac{1}{|\mathcal{D}_\ell|} \sum_{i \in \mathcal{D}_\ell} \mathscr{L}\{y_i, f_{k\ell}(\boldsymbol{x}_i)\}$;
        $\text{valid\_error}_k \leftarrow \frac{1}{5} \sum_{\ell \in \{1,\ldots,6\} \setminus k} \text{valid\_error}_{k\ell}$;
    optimal parameters from tgrid are those that minimize $\text{valid\_error}_k$;
    train a model $f_k$ of mclass on $\mathcal{D} \setminus \mathcal{D}_k$ using the optimal parameters;
    evaluate the model performance on $\mathcal{D}_k$ using loss function $\mathscr{L}(\cdot, \cdot)$;
    $\text{test\_error}_k \leftarrow \frac{1}{|\mathcal{D}_k|} \sum_{i \in \mathcal{D}_k} \mathscr{L}\{y_i, f_k(\boldsymbol{x}_i)\}$;
Output: optimal tuning parameters + performance measure for each of the six folds.
```

Algorithm 3: Cross-validation scheme for model tuning and performance evaluation.

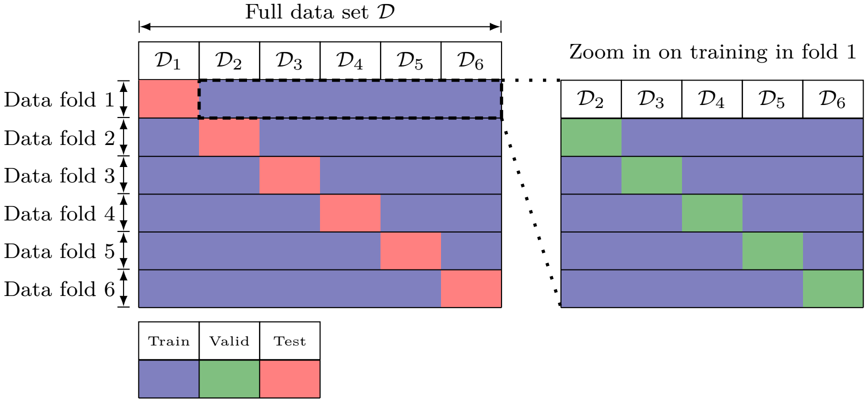


Figure 3.3: Graphical representation of the cross-validation scheme. The holdout test set for data fold k is $\mathcal{D}_k$, indicated in red. Within data fold k, we tune the parameters by 5-fold cross-validation on $\mathcal{D} \setminus \mathcal{D}_k$ with the validation sets $\mathcal{D}_\ell$ in green and the training data $\mathcal{D} \setminus \{\mathcal{D}_k, \mathcal{D}_\ell\}$ in blue. After tuning, we train the model on $\mathcal{D} \setminus \mathcal{D}_k$ using the optimal parameters for data fold k.

pricing models. However, machine learning techniques are often considered black boxes compared to statistical models such as GLMs. In a GLM, parameter estimates and their standard errors give information about the effect, uncertainty and statistical relevance of all variables. Such quick and direct interpretations are not possible with machine learning techniques, but this section introduces tools to gain insights from a model. A good source on interpretable machine learning is Molnar (2020). These tools are evaluated on the data used to train the optimal models, i.e., $\mathcal{D} \setminus \mathcal{D}_k$ for data fold k. A subset of the training data can be used to save computation time if needed.

Variable importance Variable selection and model building is often a time consuming and tedious process with GLMs (see Chapter 2). An advantage of tree-based techniques is their built-in variable selection strategy, making a priori design decisions less critical. Unraveling the variables that actually matter in the prediction is thus crucial. For $\ell \in \{1, \dots, p\}$, Breiman et al. (1984) measure the importance of a specific feature x_ℓ in a decision tree t by summing the improvements in the loss function over all the splits on x_ℓ:

$$\mathcal{I}_\ell(t) = \sum_{j=1}^{J-1} \mathbb{1}\{v(j) = \ell\} \ (\Delta \mathscr{L})_j \,.$$

The sum is taken over all $J - 1$ internal nodes of the tree, but only the nodes where the splitting variable x_v is x_ℓ contribute to this sum. These contributions $(\Delta \mathscr{L})_j$ represent the difference between the evaluated loss function before and after split j in the tree. The idea is that important variables appear often and high in the decision tree such that the sum grows largest for those variables. We normalize these variable importance values such that they sum to 100%, giving a clear idea about the relative contribution of each variable in the prediction.

We can easily generalize this approach to the ensemble techniques by averaging the importance of variable x_ℓ over the different trees that compose the ensemble:

$$\mathcal{I}_\ell = \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}_\ell(t) \,,$$

where the sum is taken over all trees in the random forest or gradient boosting machine.

Partial dependence plots Besides knowing which variables are important, it is meaningful to understand their effect on the prediction target. Partial dependence plots, introduced in Friedman (2001), show the marginal effect of a variable on the predictions obtained from a model. Hereto, we evaluate the prediction function in specific values of the variable of interest x_ℓ for $\ell \in \{1, \dots, p\}$, while averaging over a range of values of the other variables $\boldsymbol{x}^*$:

$$\bar{f}_\ell(x_\ell) = \frac{1}{n} \sum_{i=1}^{n} f_{\text{model}}(x_\ell, \boldsymbol{x}_i^*) \,. \tag{3.7}$$

The vector $\boldsymbol{x}_i^*$ holds the realized values of the other variables for observation i and n is the number of observations in the training data. Interaction effects between x_ℓ and another variable in $\boldsymbol{x}^*$ can distort the effect (Goldstein et al., 2015). Suppose that half of the observations show a positive association between x_ℓ and the prediction outcome (higher x_ℓ leads to higher predictions), while the other half of the observations show a negative association between x_ℓ and the prediction outcome. Taking the average over all observations will cause the partial dependence plot to look like a horizontal line, wrongly indicating that x_ℓ has no effect on the prediction outcome. Individual conditional expectations can rectify such wrong conclusions.

Individual conditional expectation Individual conditional expectations, introduced by Goldstein et al. (2015), also show the effect of a variable on the predictions obtained from a model, but on an individual level. We evaluate the prediction function in specific values of the variable of interest x_ℓ for $\ell \in \{1, \ldots, p\}$, keeping the values of the other variables $\boldsymbol{x}^*$ fixed:

$$\tilde{f}_{\ell,i}(x_\ell) = f_{\text{model}}(x_\ell, \boldsymbol{x}_i^*) \,, \tag{3.8}$$

where $\boldsymbol{x}_i^*$ are the realized values of the other variables for observation i. We obtain an effect for each observation i, allowing us to detect interaction effects when some (groups of) observations show different behavior compared to others. For example, two distinct groupings will emerge when half of the observations have a positive association and the other half a negative association between x_ℓ and the prediction outcome. Individual conditional expectations can also be used to investigate the uncertainty of the effect of variable x_ℓ on the prediction outcome. The partial dependence plot can be interpreted as the average of this collection of individual conditional expectations, i.e., $\bar{f}_\ell(x_\ell) = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{\ell,i}(x_\ell)$.

3.4 Case study on claim frequency and severity

An insurer's pricing team uses proprietary data to deliver a fine-grained tariff plan for a portfolio. As a typical example of such data, we study a motor third party liability (MTPL) portfolio from a Belgian insurer in 1997. This section puts focus on the claim frequency and severity models that are developed with the different modeling techniques. We briefly introduce the data and we report the optimal tuning parameters for the frequency and severity models. Afterwards, we use the tools from Section 3.3.4 to gain some insights in these optimal models. We conclude this section with an out-of-sample deviance comparison to assess the statistical performance of the different modeling techniques.

3.4.1 Quick scan of the MTPL data

The data used here is also analyzed in Denuit and Lang (2004), Klein et al. (2014) and Chapter 2. We follow the same data pre-processing steps as the aforementioned papers, e.g., regarding the exclusion of very large claims. Table B.1 in Appendix B.1 lists a description of the available variables. The portfolio contains 163,212 unique policyholders, each observed during a period of exposure-to-risk expressed as the fraction of the year during which the policyholder was exposed to the risk of filing a claim. Claim information is known in the form of the number of claims filed and the total amount claimed (in euro) by a policyholder during her period of exposure. The data set lists five categorical, four continuous and two spatial risk factors, each of them informing about specific characteristics of the policy or the policyholder. A detailed discussion on the distribution of all variables is available in Chapter 2. Regarding spatial information, we have access to the 4-digit postal code of the municipality of residence and the accompanying latitude/longitude coordinates of the center of this area. The GAM/GLM benchmarks employ spatial smoothing over the latitude/longitude coordinates. In line with this approach, we use the coordinates as continuous variables in the tree-based models.

Figure 3.4 displays the distribution of the claims information (`nclaims` and `amount`) and the exposure-to-risk measure (`expo`). Most policyholders in the portfolio are claim-free during their insured period, some file one claim and few policyholders file two, three, four or five claims. The majority of all these claims involve small amounts, but very large claims occur as well. Most policyholders are exposed to the risk during the full year, but there are policyholders who started the policy in the course of the year or surrendered the policy before the end of the year. Figure 3.4 motivates the use of loss functions which are not based on the squared error loss. We work with the Poisson and gamma distribution/deviance for frequency and severity respectively as our preferred distributional assumption (for GAM/GLM) and loss function choice (for tree-based techniques). Note that earlier work on this data, such as Denuit et al. (2007) and Chapter 2, assumed a log-normal distribution for severity. We illustrate the difference and motivate our choice in the supplementary material.

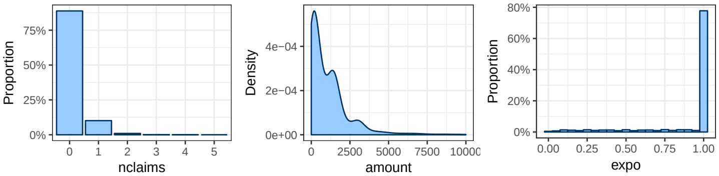


Figure 3.4: Distribution of the claim counts, amounts and the exposure-to-risk measure in the `MTPL` data.

3.4.2 Optimal tuning parameters

Table 3.3 lists the optimal tuning parameters for the different machine learning techniques in each of the six data folds. Comparing the number of splits in the trees and the number of trees in the ensembles, we conclude that the frequency models are more extensive compared to the severity variants. This is driven by the lower sample size for severity modeling and the fact that the severity of a claim is typically harder to predict than the frequency (Charpentier, 2014).

The complexity parameter cp does not give much information about the size of a regression tree and therefore Table 3.3 also lists the number of splits in the tree. All frequency trees contain between 20 and 38 splits while the severity trees comprise of only one or two splits. The coefficient of variation for the gamma prior γ remains stable over the different data folds.

The number of trees T in the random forest is very unstable over the different folds for both frequency and severity. This shows that the size of the eventual model highly depends on the training data when simply averaging independently grown trees. In four out of six cases, the number of split candidates m is equal to 5 for frequency models and 2 for the severity models. The low value of m for severity random forests indicates that variance reduction is the main performance driver, as opposed to finding the best split out of multiple candidates.

The number of trees T in the gradient boosting machine is more stable over the folds compared to the random forest. This shows that the sequential approach of growing a boosted model is less affected by the specific data fold. The tree depth d ranges from 3 to 5 in the frequency models. Table 3.3 reveals that the largest values of T correspond to the smallest values of d and vice versa, highlighting the interplay between these tuning parameters. In five out of six cases, the severity models use stumps (i.e., trees with only one split) as weak learners. A tree depth of $d = 1$ makes these models completely additive, without interaction by construct.

		Regression tree			Random forest		Boosting machine	
	data fold	cp	γ	splits	T	m	T	d
Frequency	1	7.3×10^{-5}	0.125	38	4,900	5	2,600	3
	2	1.4×10^{-4}	0.125	24	900	5	2,000	4
	3	1.1×10^{-4}	0.125	31	400	8	1,400	5
	4	1.2×10^{-4}	0.250	27	5,000	5	1,500	5
	5	1.8×10^{-4}	0.250	20	600	10	1,900	4
	6	1.7×10^{-4}	0.250	23	100	5	2,700	3
Severity	1	3.3×10^{-3}	-	2	4,300	2	600	1
	2	5.8×10^{-3}	-	1	200	2	300	1
	3	3.7×10^{-3}	-	2	600	1	500	1
	4	7.3×10^{-3}	-	1	100	2	400	2
	5	5.4×10^{-3}	-	1	3,600	2	600	1
	6	5.4×10^{-3}	-	1	100	1	600	1

Table 3.3: Overview of the optimal tuning parameters for the tree-based machine learning techniques.

We also tune the benchmark GLMs for each of the six data folds separately, i.e., we perform the binning strategy from Chapter 2 in each fold k such that the optimal bins are chosen for the training data at hand $\mathcal{D} \setminus \mathcal{D}_k$. A grid is used for the two tuning parameters involved, one for the continuous variables and one for the spatial effect, thereby avoiding the two-step binning procedure initially proposed in Chapter 2. Examples of the resulting benchmark GLMs for frequency and severity are presented in Appendix B.4.2.

3.4.3 Model interpretation

We will use the variable importance measure to find the most relevant variables in the frequency and severity models. Afterwards, we will make use of partial dependence plots and individual conditional expectations to gain understanding on a selection of interesting effects for the claim frequency. Similar results on claim severity can be found in the supplementary material.

Variable importance To learn which variables matter for predicting claim frequency and severity, we compare in Figure 3.5 the variable importance plots for the different machine learning techniques over the six data folds. The variables are ranked from top to bottom, starting with the most important one as measured by the average variable importance over the folds (multiple variables with zero importance are ordered alphabetically from Z to A). By contrasting the graphs in the left column of Figure 3.5, we see that the important variables (mostly bonus-malus scale and age) are similar across all methods for the frequency model. Other relevant risk factors are the power of the vehicle and the spatial risk factor (combining the longitude and latitude information). The frequency GLM, presented in Table B.3 in Appendix B.4.2, contains the top seven variables together with `coverage`, which is ranked at the ninth place for all methods.

The right column of Figure 3.5 shows the variable importance for severity models. The ranking is very dissimilar across the different modeling techniques. The regression tree models for severity contain only one split using the type of coverage in four out of the six folds, while the other two trees have an additional split on the age of the car. The random forest and gradient boosting machine include more variables, but they both lead to rather different rankings of the importance of those variables. The severity GLM, presented in Table B.4 in Appendix B.4.2, contains three variables: `coverage`, `ageph` and `agec`. An interesting observation is that the most important risk factors in the gradient boosting machines are those selected in the GLMs.

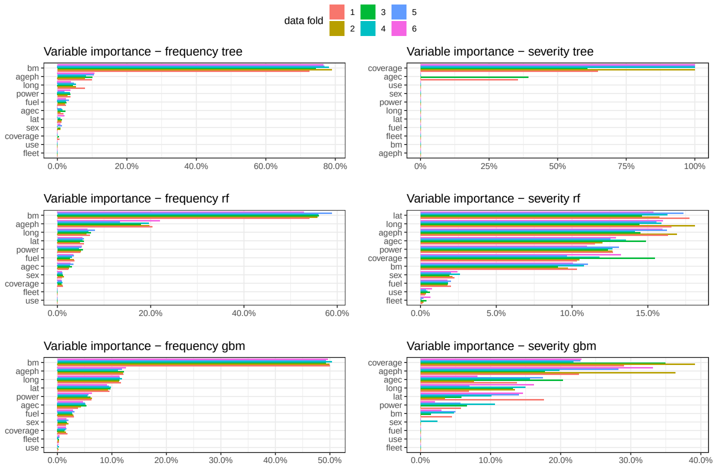


Figure 3.5: Variable importance in the optimal regression tree (top), random forest (middle) and gradient boosting machine (bottom) per data fold (color) for frequency (left) and severity (right).

Partial dependence plot of the age effect Figure 3.6 compares the partial dependence effect of the age of the policyholder in frequency models. The two top panels of Figure 3.6 show the GLM and GAM effects on the left and right respectively. As explained in Section 3.2, due to our proposed data-driven approach, the GLM effects are a step-wise approximation of the GAM effects. The risk of filing a claim is high for young policyholders and gradually decreases with increasing ages to stabilize around the age of 35. The risk starts decreasing again around the age of 50 and increases for senior policyholders around the age of 70. The bottom left panel of Figure 3.6 shows the age effect captured by the regression trees. The effect is less stable across the folds compared to the other methods, this is a confirmation and illustration of the variability of a single regression tree. There is also no increase in risk for senior policyholders in the regression trees. The bottom right panel of Figure 3.6 shows the age effect according to the gradient boosting machines. This looks very similar to the smooth GAM effect with one important distinction, namely the flat regions at the boundaries. This makes the tree-based techniques more robust with respect to extrapolation and results in less danger of creating very high premiums for risk profiles at edges. Note that the gradient boosting machine predicts a wider range of frequencies than the regression tree, namely 0.12 to 0.20 versus 0.12 to 0.165 respectively. The shape of the age effect in the random forest, available in Appendix B.3, is rather similar to the gradient boosting machine effect but on a slightly more compact range.

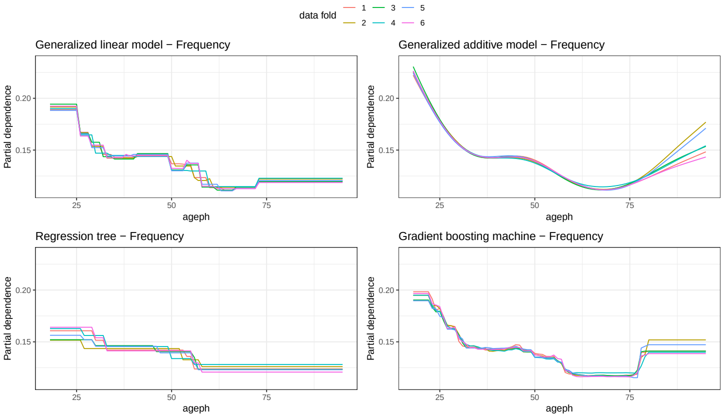


Figure 3.6: Partial dependence plot to visualize the effect of the age of the policyholder on frequency for the optimal model obtained per data fold (color) in a GLM (top left), GAM (top right), regression tree (bottom left) and gradient boosting machine (bottom right).

Partial dependence plot of the spatial effect Figure 3.7 compares the spatial effect in frequency models, more specifically the models trained on the data where fold $\mathcal{D}_3$ was kept as the hold-out test set. We choose a specific data fold because we otherwise need to show six maps of Belgium per method as opposed to overlaying six effects as in Figure 3.6. The two top panels show the GLM and GAM effects on the left and right respectively. Brussels, the capital of Belgium located in the center of the country, is clearly the most accident-prone area to live and drive a car because of heavy traffic. The southern and northeastern parts of Belgium are less risky because of sparser population and more rural landscapes. The bottom left panel of Figure 3.7 shows the spatial effect as it is captured with a regression tree. Splitting on longitude and latitude coordinates results in a rectangular split pattern on the map of Belgium. The bottom right panel of Figure 3.7 shows the spatial effect for the gradient boosting machine. The underlying rectangular splits are still visible but in a smoother way compared to the regression tree. Brussels still pops out as the most risky area and the pattern looks similar to the GLM and GAM effects. The shape of the random forest spatial effect (see Appendix B.3) is again similar to that of the gradient boosting machine on a more compact range.

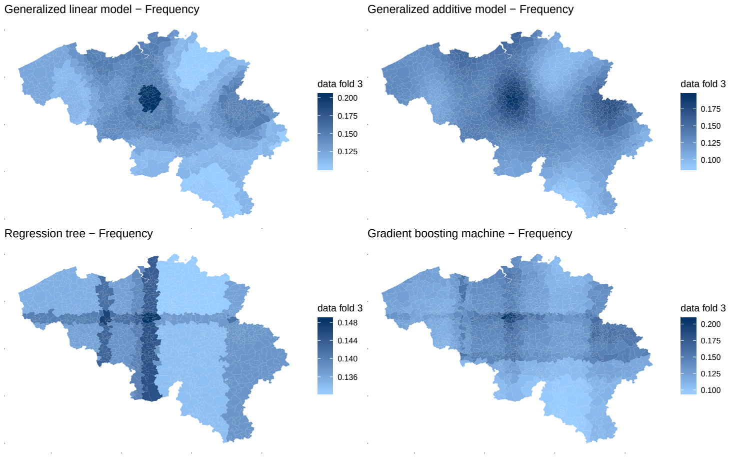

Generalized linear model − Frequency
data fold 3
0.200
0.175
0.150
0.125
Generalized additive model − Frequency
data fold 3
0.175
0.150
0.125
0.100
Regression tree − Frequency
data fold 3
0.148
0.144
0.140
0.136
Gradient boosting machine − Frequency
data fold 3
0.200
0.175
0.150
0.125
0.100

Figure 3.7: Partial dependence plot to visualize the effect of the municipality of residence on frequency in a GLM (top left), GAM (top right), regression tree (bottom left) and gradient boosting machine (bottom right).

Figures 3.6 and 3.7 teach us that a single tree is not able to capture certain aspects in the data, resulting in a coarse approximation of the underlying risk effect. The ensemble techniques are able to capture refined risk differences in a much smoother way. The differences in the range of the predicted effects imply that the gradient boosting machine performs more segmentation while the random forest puts more focus on risk pooling.

Individual conditional expectation of the bonus-malus effect Figure 3.8 compares the bonus-malus effect captured with a regression tree (left) and a gradient boosting machine (right) for frequency data. As in Figure 3.7, we show the effect for the models trained on the data where fold $\mathcal{D}_3$ was kept as the hold-out test set. The gray lines are individual conditional expectations for 1,000 random policyholders and the blue line shows the partial dependence curve. The values for $\boldsymbol{x}^*$ in Eq. (3.8) are those registered for the selected policyholders. On average, we observe an increase in frequency risk as the blue line surges over the bonus-malus levels, which is to be expected because higher bonus-malus levels indicate a worse claim history. We can get an idea about the sensitivity of the bonus-malus effect across the different policyholders in the portfolio by comparing the steepness of the gray lines. Keeping all other risk factors fixed, a steeper effect indicates that a policyholder's risk is more sensitive to changes in the bonus-malus scale. This effect is driven by the combination of all risk factors registered for this policyholder.

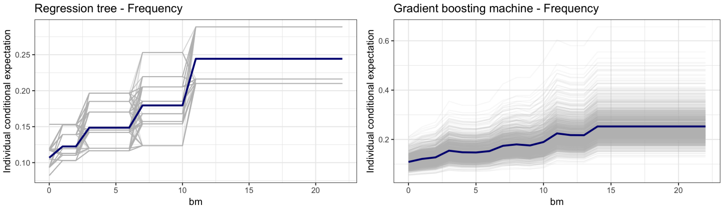


Figure 3.8: Effect of the bonus-malus scale on frequency in a regression tree (left) and gradient boosting machine (right) as partial dependence (blue) and individual conditional expectations (gray).

The previous figures show some counterintuitive results regarding the monotonicity of a fitted effect. For example, the bonus-malus individual conditional expectations for the regression tree in Figure 3.8 reveal decreases in risk over increasing bonus-malus levels for certain policyholders. This poses problems for the practical implementation of such a tariff because it assigns a lower premium to policyholders with a worse claim history. In practice, an actuary would specify monotonicity constraints on such a risk factor, either by an a posteriori smoothing of the resulting effect or by using an implementation that allows to specify such constraints a priori, e.g., the `gbm` package has this functionality. Our analysis does not enforce such constraints.

3.4.4 Hunting for interaction effects

Tree-based models are often praised for their ability to model interaction effects between variables (Buchner et al., 2017; Schiltz et al., 2018). The predictions of a model can not be expressed as the sum of the main effects when interactions are present, because the effect of one variable depends on the value of another variable. Friedman's H-statistic, introduced by Friedman and Popescu (2008), estimates the interaction strength by measuring how much of the prediction variance originates from the interaction effect. We will put focus on two-way interactions between variables x_k and x_ℓ, but in theory this measure can be applied to arbitrary interactions between any number of variables. Let $\bar{f}_k(x_k)$ and $\bar{f}_l(x_\ell)$ represent the one-dimensional partial dependence of the variables as defined in Section 3.3.4 and $\bar{f}_{kl}(x_k, x_\ell)$ the two-way partial dependence, defined analogously to Eq. (3.7). The H-statistic is expressed as:

$$H^2_{k\ell} = \frac{\sum_{i=1}^{n} \{\bar{f}_{kl}(x_k^{(i)}, x_\ell^{(i)}) - \bar{f}_k(x_k^{(i)}) - \bar{f}_l(x_\ell^{(i)})\}^2}{\sum_{i=1}^{n} \bar{f}^2_{kl}(x_k^{(i)}, x_\ell^{(i)})} \, ,$$

where $x_k^{(i)}$ indicates that the partial dependence function is evaluated at the observed value of x_k for policyholder i. Assuming the partial dependence is centered at zero, the numerator measures the variance of the interaction while the denominator measures the total variance. The ratio of both therefore measures the interaction strength as the amount of variance explained by the interaction. The H-statistic ranges from zero to one, where zero indicates no interaction and one implies that the effect of x_k and x_ℓ on the prediction is purely driven by the interaction.

Table 3.4 shows the fifteen highest two-way H-statistics among the variables available in the data set (as listed in Table B.1 in Appendix B.1) for the frequency gradient boosting machine trained on the data where fold $\mathcal{D}_3$ was kept as the hold-out test set. The strongest interaction is found between the longitude and latitude coordinates, which is not a surprise seeing how these two variables together encode the region where the policyholder resides.

The H-statistic informs us on the strength of the interaction between two variables, but gives us no idea on how the effect behaves. Figure 3.9 shows grouped partial dependence plots to investigate the interactions highlighted in gray in Table 3.4. The partial dependence plots of a specific variable are grouped into five equally sized groups based on the value of another variable. Interaction effects between both variables can be discovered by comparing the evolution of the curves over the five different groups. An interaction is at play when this evolution is different for policyholders in different groups. In order to focus purely on the evolution of the effect, we let all the curves in Figure 3.9 start at zero by applying a vertical shift.

Variables	H-statistic	Variables	H-statistic	Variables	H-statistic
(lat, long)	0.2687	(agec, coverage)	0.1185	(bm, power)	0.0800
(fuel, power)	0.1666	(ageph, power)	0.1062	(ageph, lat)	0.0799
(agec, power)	0.1319	(ageph, bm)	0.0961	(agec, ageph)	0.0785
(ageph, sex)	0.1293	(power, sex)	0.0829	(long, sex)	0.0732
(coverage, long)	0.1203	(fuel, long)	0.0828	(agec, bm)	0.0678

Table 3.4: H-statistic of the 15 strongest two-way interactions between all the variables in the gradient boosting machine for frequencies, trained on data with $\mathcal{D}_3$ as hold-out test set.

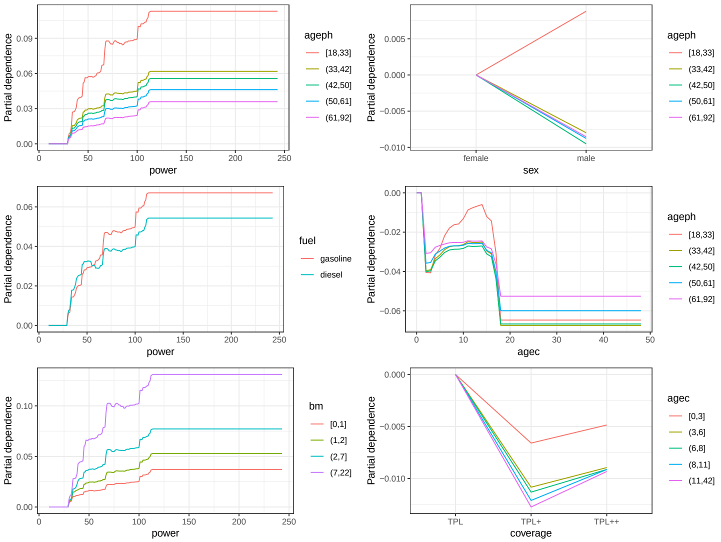


Figure 3.9: Grouped partial dependence plots for the gradient boosting machine on frequency, trained on data with $\mathcal{D}_3$ as hold-out test set. The effect is binned in five equally sized groups. The left column shows the effects for the power of the car grouped by the age of the policyholder (top), the type of fuel (middle) and the bonus-malus scale (bottom). The right column shows the effects for the sex of the policyholder (top), age of the car (middle) and type of coverage (bottom), grouped by the age of the policyholder or car.

An important and well-known effect in insurance pricing is the interaction between the age of the policyholder and the power of the vehicle. Our benchmark GLM and GAM use this interaction effect in the predictor and Figure 11 in Chapter 2 shows that young policyholders with high power vehicles form an increased risk for the insurer regarding claim frequency. The top left panel of Figure 3.9 shows the partial dependence of the power of the vehicle, grouped by the age of the policyholder. The power effect is steepest for young policyholders, indicated by the red line. The steepness of the effect decreases for increasing ages. The difference in the steepness of the effect between young and old policyholders is a visual confirmation of the interaction at play between the variables `ageph` and `power`. The top right panel of Figure 3.9 shows the partial dependence for the sex of the policyholders, grouped by their age. For young policyholders, aged 18 to 33, we observe that males are on average more risky drivers compared to females, while for the other age groups the female drivers are perceived more risky than males. European insurers are not allowed to use gender in their tariff structure nowadays, implying that young female drivers might be partly subsidizing their male peers.

The middle left panel of Figure 3.9 shows the partial dependence of the power of the vehicle, grouped by the type of fuel. We observe that the steepness of the power effect is slightly higher for gasoline cars. Drivers typically choose a diesel car when their annual mileage is above average, which would justify their choice of buying a bigger and more comfortable car with higher horsepower. However, drivers who own a high powered gasoline car might choose such a car to accommodate for a more sportive driving style, making them more prone to the risk of a claim. The middle right panel of Figure 3.9 shows the partial dependence of the age of the vehicle, grouped by the policyholder's age. We observe a big bump for young policyholders in the vehicle age range from 5 to 15. This could indicate an increased claim frequency risk for starting drivers who buy their first car on the second-hand market. The sharp drop around 19 could relate to vintage cars that are used less often and are thus less exposed to the claim risk.

The bottom left panel of Figure 3.9 shows the partial dependence of the power of the vehicle, grouped by the bonus-malus scale. We observe that the power effect grows steeper for increasing levels occupied in the bonus-malus scale. This indicates that driving a high powered car becomes more risky for policyholders in higher bonus-malus scales. The bottom right panel of Figure 3.9 shows the partial dependence of the type of coverage, grouped by the age of the vehicle. For vehicles in the age range zero to three, we observe that adding material damage covers decreases the claim frequency risk less compared to other age ranges. This might indicate that policyholders who buy a new car add a material damage cover because they worry about damaging their newly purchased vehicle, while policyholders with older cars who still add damage covers are more risk-averse and also less risky drivers.

3.4.5 Statistical out-of-sample performance

Figure 3.10 compares the out-of-sample performance of the different models investigated over the six data folds. We evaluate the Poisson deviance for frequency models and the gamma deviance for severity models, see Eq. (4.5) and (3.6) respectively, on the holdout test data. In the left panel of Figure 3.10, we observe a clear ranking of the out-of-sample Poisson deviance among the different methods. The gradient boosting machine is most predictive, consistently leading to the lowest deviance values. The performance of GLMs and GAMs is very similar, which is expected because the GLM is a data driven approximation of the GAM, as explained in Section 3.2. Next in line is the random forest, and the regression tree is the least predictive for frequency. These results are very stable over the six data folds. The right panel of Figure 3.10 shows the out-of-sample gamma deviance for severity. The methods perform rather similarly and there is no clear winner or loser over the different folds. The peak at the fourth fold reveals a weakness of the GAM: the extrapolation of smooth effects, see Figure 3.6, combined with out-of-sample testing can lead to huge deviance values. In the severity GAM trained on $\mathcal{D} \setminus \mathcal{D}_4$ and evaluated on $\mathcal{D}_4$, the problem occurred with the age of the car. Specifically, the maximal value for `agec` in $\mathcal{D} \setminus \mathcal{D}_4$ for severity is 32 while the maximal value in $\mathcal{D}_4$ is 37, thus requiring an extrapolation of the calibrated smooth effect. This motivates to cap continuous variables at a certain cut-off in a pre-processing stage for a GAM. A tree-based method automatically deals with this problem thanks to the flat region at the outer ends of the effect, see Figure 3.6.

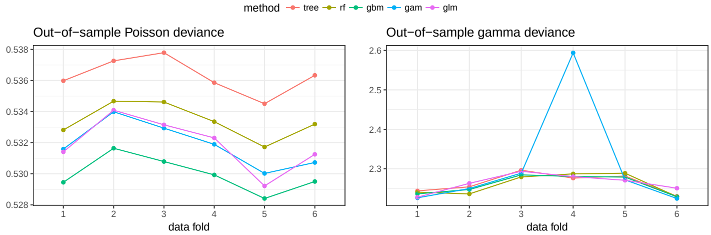


Figure 3.10: Out-of-sample Poisson deviance for frequency (left) and gamma deviance for severity (right), each color representing a different modeling technique.

This comparison only puts focus on the statistical performance of the frequency and severity models. In the next section, we combine both in a pricing model and compare the different tariff structures using practical business metrics relevant for an insurance company.

3.5 Model lift: from analytic to managerial insights

Choosing between two tariff structures is an important business decision. This creates the need to translate our model findings to evaluation criteria that capture a manager's interest. We evaluate the economic value of a tariff using tools proposed in the literature to measure model lift (Frees et al., 2013; Goldburd et al., 2016, Section 7.2). In this context, model lift refers to the ability of a model to prevent adverse selection. An insurer might become victim hereof when a competitor refines its tariff structure via innovation such that good risks switch to the competitor and the insurer is left with the bad risks which are more prone to high losses.

We combine the claim frequency and severity models from Section 3.4 to obtain the pure premium for each policy under consideration, allowing us to compare different tariff structures. As Figure 3.3 illustrates, each observation is out-of-sample in exactly one of the data folds, more specifically the observations in $\mathcal{D}_k$ are out-of-sample for data fold k. We then use the optimal model trained on $\mathcal{D} \setminus \mathcal{D}_k$ to predict the policyholders in holdout test set $\mathcal{D}_k$. Following this strategy, we obtain one premium per modeling technique for each policyholder in the full data $\mathcal{D}$. Table 3.5 shows a comparison between the predicted premium totals and the observed losses, both on the portfolio level and split by data fold. On average, every method is perfectly capable of replicating the total losses. Section 3.5.2 compares the model lift measures from Section 3.5.1 on the portfolio level. We also analyzed each of the data folds separately (not shown), leading to the same ranking of models as in Section 3.5.2, thereby validating the consistency of our results.

Data fold	GLM	CART	RF	GBM	Losses	GLM	CART	RF	GBM
1	4,396,698	4,341,397	4,407,389	4,376,619	4,365,483	1.01	0.99	1.01	1.00
2	4,420,933	4,339,615	4,419,903	4,384,513	4,388,147	1.01	0.99	1.01	1.00
3	4,369,876	4,313,768	4,380,972	4,337,848	4,461,478	0.98	0.97	0.98	0.97
4	4,370,502	4,374,666	4,383,748	4,324,014	4,422,213	0.99	0.99	0.99	0.98
5	4,405,369	4,374,368	4,399,226	4,357,937	4,485,079	0.98	0.98	0.98	0.97
6	4,397,372	4,375,151	4,412,852	4,363,588	4,342,569	1.01	1.01	1.02	1.01
portfolio	26,360,750	26,118,966	26,404,091	26,144,519	26,464,970	1.00	0.99	1.00	0.99

Table 3.5: Comparison of the predicted premiums and observed losses on portfolio level and by data fold. We show the premium and loss totals (left) and the ratio of premiums to losses (right).

3.5.1 Tools to measure model lift

Suppose that an insurance company has a tariff structure P^{bench} in place and a competitor introduces a tariff structure P^{comp} based on a new modeling technique or a different set of rating variables. We define the relativity r_i as the ratio of the competing premium to the benchmark premium for policyholder i:

$$r_i = \frac{P_i^{\text{comp}}}{P_i^{\text{bench}}} \,. \tag{3.9}$$

A small relativity indicates a profitable policy which can potentially be lost to a competitor offering a lower premium. A high relativity reveals an underpriced policy which could benefit from better loss control measures such as renewal restrictions. These statements make the assumption that P^{comp} is a more accurate reflection of the true risk compared to P^{bench}.

Loss ratio lift The loss ratio (LR) is the ratio of total incurred claim losses and total earned premiums. Following Goldburd et al. (2016), we assess the loss ratio lift in the following way:

1. sort the policies from smallest to largest relativity r_i;
2. bin the policies into groups containing the same amount of total exposure e;
3. within each bin, calculate the overall loss ratio using the benchmark premium P^{bench}.

The bins should have loss ratios around 100% if the benchmark tariff is a technically accurate reflection of the risk. However, an upward trend in the loss ratios would indicate that policies with a lower (higher) premium under the competing tariff are those with a lower (higher) loss ratio in the benchmark tariff, pointing out that the competing tariff better aligns the risk.

Double lift A double lift chart facilitates a direct comparison between two potential tariff structures. Following Goldburd et al. (2016), this chart is created in the following way:

1. sort the policies from smallest to largest relativity r_i;
2. bin the policies into groups containing the same amount of total exposure e;
3. within each bin, calculate the average actual loss amount (L) and the average predicted pure premium for both the models (P^{bench} and P^{comp});
4. within each bin, calculate the percentage error for both models as $P/L-1$.

The best tariff structure is the one with the percentage errors closest to zero, indicating that those premiums match the actual losses more closely.

Gini index Frees et al. (2013) introduced the ordered Lorenz curve to compare the tariffs P^{bench} and P^{comp} by analyzing the distribution of losses versus premiums, where both are ordered by the relativities r from Eq. (3.9). The ordered Lorenz curve is defined as follows:

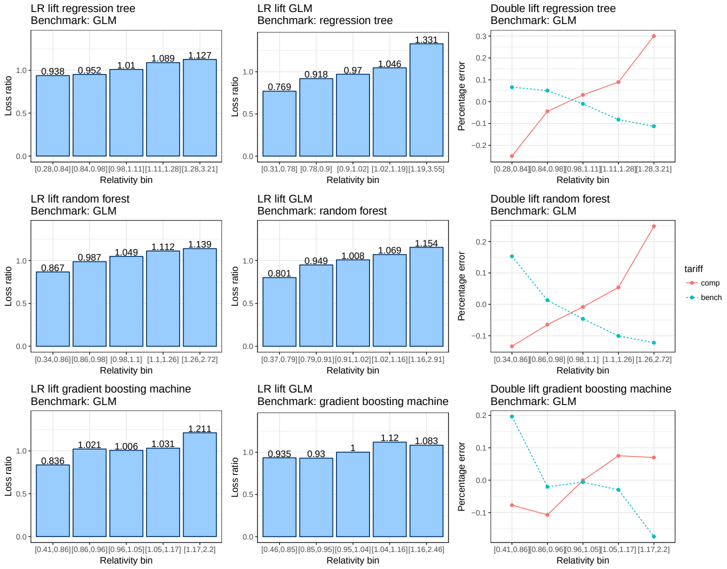
$$\left(\frac{\sum_{i=1}^{n} L_i \, \mathbb{1}\{F_n(r_i) \leq s\}}{\sum_{i=1}^{n} L_i}, \frac{\sum_{i=1}^{n} P_i^{\text{bench}} \, \mathbb{1}\{F_n(r_i) \leq s\}}{\sum_{i=1}^{n} P_i^{\text{bench}}} \right),$$

for $s \in [0,1]$ where $F_n(r_i)$ is the empirical cumulative distribution function of the relativities r. This curve will coincide with the 45 degree line of equality when the technical pricing is done right by the benchmark premium. However, the curve will be concave up when P^{comp} is able to spot tariff deficiencies in P^{bench}. The cumulative distributions are namely taken from the most overpriced policies towards the most underpriced policies in P^{bench}. The Gini index, introduced by Gini (1912) and computed as twice the area between the ordered Lorenz curve and the line of equality, has a direct economic interpretation. A tariff structure P^{comp} that yields a larger Gini index is likely to result in a more profitable portfolio because of better differentiation between good and bad risks. The insurer can decide to only retain the policies with a relativity value below a certain threshold. Averaging this decision over all possible thresholds, Frees et al. (2013) show that the average percentage profit for an insurer equals one half of the Gini index.

3.5.2 Adverse selection and profits

The panels in the left column of Figure 3.11 show the loss ratio lift charts for the regression tree, random forest and gradient boosting machine respectively with the GLM as benchmark tariff (i.e., the GLM premium is the denominator in Eq. (3.9)). All tree-based methods show an increasing trend in the loss ratios. This implies that policies which would receive a lower premium under the competing tariff, those in the first bins, are policies with favorable loss ratios. At the same time, policies having a higher premium under the competing tariff, those in the last bins, exhibit detrimental loss ratios. The tree-based techniques are therefore able to spot deficiencies in the GLM benchmark tariff. One should not draw conclusions from these graphs too fast however. The middle panels of Figure 3.11 show the loss ratio lifts for the GLM with the three respective tree-based techniques as a benchmark tariff. Comparing these lift charts side by side, we can observe that the upwards trend is now steeper in the cases of the regression tree and random forest. Thus, the GLM is better in spotting deficiencies in those tree-based tariffs compared to the other way around. The gradient boosting machine and GLM result in rather complementary tariffs. The GLM is very good in the three middle relativity bins, but the gradient boosting machine is clearly outperforming in the first and last bin.

These findings are confirmed by the double lift charts in the right panels of Figure 3.11. These show the double lift charts obtained with the GLM as the benchmark tariff in the relativities. The red and turquoise line respectively show the percentage error for the tree-based model and the GLM. For both the regression tree and the random forest, the percentage error for the GLM benchmark tariff is more closely centered around zero compared to the competitor percentage error. We again notice the complementarity of the gradient boosting machine and GLM tariffs. The percentage error for the gradient boosting machine is closer to zero for the first and last relativity bin, but the GLM is closer to zero for the other three relativity bins.



LR lift regression tree
Benchmark: GLM

LR lift GLM
Benchmark: regression tree

Double lift regression tree
Benchmark: GLM

LR lift random forest
Benchmark: GLM

LR lift GLM
Benchmark: random forest

Double lift random forest
Benchmark: GLM

LR lift gradient boosting machine
Benchmark: GLM

LR lift GLM
Benchmark: gradient boosting machine

Double lift gradient boosting machine
Benchmark: GLM

Figure 3.11: Assessment of model lift for the regression tree (top), random forest (middle) and gradient boosting machine (bottom). The left column shows the loss ratio lift for the tree-based techniques with the GLM as benchmark. The middle column shows the loss ratio lift for the GLM with the tree-based techniques as benchmark. The right column shows the double lift chart for the tree-based techniques with the GLM as benchmark.

The gradient boosting machine tariff clearly holds economic value over the GLM benchmark. However, in the bottom left panel of Figure 3.11, we observe that the gradient boosting machine is slightly over-correcting the GLM premium in the extreme ends of the tariff. The loss ratio in the first bin is 0.84 while the average relativity in that bin is equal to 0.78. Likewise, the average loss ratio in the last bin is 1.21 while the average relativity in that bin is equal to 1.30.

Table 3.6 shows a two-way comparison of Gini indices for the machine learning methods and the GLM. The row names indicate the model generating the benchmark tariff structure P^{bench} while the column names indicate the model generating the competing tariff structure P^{comp}. The row-wise maximum values are indicated in bold. We observe that the gradient boosting machine achieves the highest Gini index when the benchmark is either the GLM, the regression tree or the random forest. When the gradient boosting machine serves as benchmark, the GLM attains the highest Gini index. We use the mini-max strategy of Frees et al. (2013) where we search for the benchmark model with the minimal maximum Gini index. In other words, we look for the benchmark model with the lowest value in bold in Table 3.6. The gradient boosting machine achieves this minimal maximum Gini index, indicating that this approach leads to a tariff structure that is the least probable to suffer from adverse selection. Note that the GLM tariff achieves the second place, before the random forest and regression tree.

Frees et al. (2013) explain that the average profit for an insurer is equal to half the Gini index. Let us assume that the insurance company uses state-of-the-art GLMs to develop their current tariff structure on this specific data. This implies that developing a competing tariff structure with gradient boosting machines would result in an average profit of around 3.3% for the insurer. The average is taken over all possible decision-making strategies that the insurance company can take to retain policies based on the relativities. Therefore, by following an optimal strategy, the profit can even be higher for a specific choice of portfolio. We suspect that the improvement in profits could be even greater if there were more explanatory variables in the data.

	competitors:	GLM	CART	RF	GBM
benchmark	GLM		4.07	5.99	**6.57**
	CART	10.86		10.10	**12.02**
	RF	7.07	0.53		**7.59**
	GBM	**3.93**	0.67	2.30	

Table 3.6: Two-way comparison of Gini indices for the different tree-based techniques and GLM.

3.5.3 Solidarity and risk differentiation

From a social point of view, it is crucial for everybody to be able to buy insurance cover at a reasonable price. A tariff structure that follows from a complex machine learning algorithm should not lead to the "personalization of risk" with excessively high premiums for some policyholders. Figure 3.12 shows violin plots of the annual (i.e., exposure equals one) premium distribution in both the gradient boosting machine tariff P^{gbm} and the GLM tariff P^{glm}. We only consider the gradient boosting machine because Sections 3.4.5 and 3.5.2 teach us that only this method holds added value over the GLM. The left panel shows the annual premium amounts and we observe that both distributions look very similar. The minimum, median and maximum premium is 43, 155 and 1138 Euro in P^{gbm} and 41, 156 and 1230 Euro in P^{glm} respectively. The right panel shows the relative difference between both premiums, namely $(P^{\text{gbm}} - P^{\text{glm}})/P^{\text{gbm}}$. The difference is centered around zero and for half the policyholders the difference lies in the range $[-12\%, +12\%]$. This implies that, overall, P^{gbm} and P^{glm} trade off segmentation and risk pooling in a similar way, thereby finding a balance between differentiation and solidarity. For a small selection of policyholders, the gradient boosting machine leads to considerable discounts compared to the GLM.

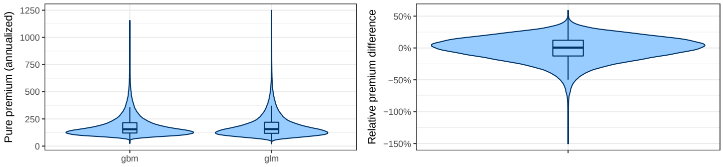


Figure 3.12: Comparison of the annual premium distribution in the gradient boosting machine tariff P^{gbm} and the GLM tariff P^{glm}: absolute premiums (left) and relative differences (right).

The gradient boosting machine and GLM result in similar premiums on a portfolio level (see Table 3.5), but on a coarser scale, they could lead to different approaches for targeting specific customer segments. Figure 3.13 are the relative premium differences between P^{gbm} and P^{glm} over the age of the policyholder in the left panel and the power of the car in the right panel. The blue dots show the average premium difference for policyholders with that specific characteristic, e.g., all policyholders aged 25. We observe that younger policyholders obtain a slightly lower premium in the gradient boosting machine tariff, while senior policyholders obtain a slightly higher premium compared to the GLM tariff. For middle aged policyholders there are some fluctuations which can be explained by analyzing the age effects in Figure 3.6. For the group of dots around the age of 75, P^{gbm} gives an average 30% premium discount over P^{glm}. Figure 3.6

shows that the age effect starts increasing before the age of 75 in the GLM (top left), but only after the age of 75 for the gradient boosting machine (bottom right). Therefore, policyholders around the age of 75 obtain a better deal in the gradient boosting machine tariff. In the right panel of Figure 3.13, the premium differences increase roughly monotonically with the power of the car. Low powered cars obtain a lower premium in P^{gbm} while high powered cars get a lower premium in P^{glm}. In between the differences are close to zero, indicating that both tariffs treat those cars in a similar fashion.

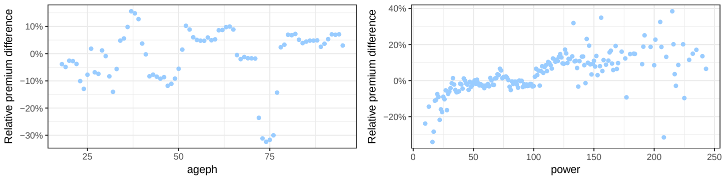


Figure 3.13: Comparison of premium differences between P^{gbm} and P^{glm} over the age of the policyholder (left) and the power of the car (right).

We conclude that gradient boosting machines can be a valuable tool for the insurer, while the other tree-based techniques under investigation show little added value on our specific portfolio. The gradient boosting machine is able to deliver a tariff that assesses the underlying risk in a more accurate way, thereby guarding the insurer against adverse selection risks which eventually can result in a profit. The gradient boosting machine also honored the principle of solidarity in the portfolio, offering affordable insurance cover for all policyholders with premiums in the same range as the benchmark GLM.

3.6 Conclusions and outlook

In this study, we have adapted tree-based machine learning to the problem of insurance pricing, thereby leaving the comfort zone of both traditional ratemaking and machine learning. State-of-the-art GLMs are compared to regression trees, random forests and gradient boosting machines. These tree-based techniques can be used on insurance data, but care has to be taken with the underlying statistical assumptions in the form of the loss function choice. This chapter brings multiple contributions to the existing literature. First, we develop complete tariff plans with tree-based machine learning techniques for a real-life insurance portfolio. In this process, we use the Poisson and gamma deviance because the classical squared error loss is not appropriate for a frequency-severity problem. Second, our elaborate cross-validation scheme gives a well thought and careful tuning procedure, allowing us to assess not only the performance of different methods, but also the stability of our results across multiple data folds. Third, we go beyond a purely statistical comparison and also focus on business metrics used in insurance companies to evaluate different tariff strategies. Fourth, we spend a lot of attention on the interpretability of the resulting models. This is a very important consideration for insurance companies within the GDPR's regime of algorithmic accountability. Fifth, our complete analysis is available in well-documented R functions, readily applicable to other data sets. This includes functions for training, predicting and evaluating models, running the elaborate cross-validation scheme, interpreting the resulting models and assessing the economic lift of these models. Sixth, we extended the `rpart` package such that it is now possible to build regression trees with a gamma deviance as loss function and random forest with both the Poisson and gamma deviance as loss functions. This package is available at https://github.com/henckr/distRforest.

The gradient boosting machine is consistently selected as best modeling approach, both by out-of-sample performance measures and model lift assessment criteria. This implies that an insurer can prevent adverse selection and generate profits by considering this new modeling framework. However, this might be impossible because of regulatory issues, e.g., filing requirements (see Appendix B.4). In that case, an insurance company can still learn valuable information on how to form profitable portfolios from an internal, tech nical model and translate this to a commercial product which is in line with all those requirements. A possible approach would be to approximate a gradient boosting machine with a GLM, much in line with the strategy to develop the benchmark pricing GLM in this study. The gradient boosting machine can be used to discover the important variables and interactions between those variables, which can then be included in a GLM for deployment. Although we present the tools to detect potentially interesting variables and interactions, we leave for future work the building of a competitive GLM inspired by the gradient boosting machine.

Chapter 4

Model-Agnostic Interpretable Data-driven suRRogates

Technological advancements allow to develop high-performance black box predictive models. However, strictly regulated industries (like banking and insurance) ask for transparent decision-making algorithms. We therefore present a procedure to develop a Model-Agnostic Interpretable Data-driven suRRogate (maidrr) suited for structured tabular data. Knowledge is extracted from a black box via partial dependence effects. These are used to perform smart feature engineering by grouping variable values. This results in a segmentation of the feature space with automatic variable selection. A transparent generalized linear model (GLM) is fit to the features in categorical format and their relevant interactions. This GLM serves as a global surrogate to the original black box and replaces it in production. We demonstrate our R package `maidrr` with a case study on general insurance claim frequency modeling for six publicly available datasets. Our maidrr GLM closely approximates a gradient boosting machine (GBM) black box and outperforms both a linear and tree surrogate as benchmarks.

This chapter is based on joint work with Katrien Antonio and Marie-Pier Côté, which is publicly available as an arXiv preprint Henckaerts et al. (2021a). This research is supported by the Research Foundation Flanders [SB grant 1S06018N] and by the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-04190]. Furthermore, Katrien Antonio acknowledges financial support from the Ageas Research Chair at KU Leuven and from KU Leuven's research council [COMPACT C24/15/001].

4.1 Introduction

The big data revolution opened the door to highly complex artificial intelligence (AI) technology and black box models in search for top performance (Caruana and Niculescu-Mizil, 2006). However, at the same time, there is growing public awareness for the issues of interpretability, explainability and fairness of AI systems (O'Neil, 2016). The General Data Protection Regulation (GDPR, 2016) introduces "the right to an explanation" of decision-making algorithms, thereby pushing for transparent communication on the underlying decision process. An explainable AI (XAI) algorithm enables human users to understand, trust and manage its decisions (Gunning, 2017). Explainability is gaining attention in many industries, such as automotive (Meteier et al., 2019), banking (Bracke et al., 2019), healthcare (Ahmad et al., 2018), insurance (OECD, 2020), manufacturing (Hrnjica and Softic, 2020) and critical systems (Gade et al., 2019). Full transparency is essential for high-stakes decisions with a big impact on a person's life (Rudin, 2019). High-stakes examples include medical diagnosis, insurance pricing, education admission, loan applications, criminal justice, job recruitment and autonomous transportation.

A lack of algorithmic transparency can hinder AI implementations in business practice due to regulatory compliance requirements (Arrieta et al., 2020). XAI is therefore especially important in highly regulated industries with an extensive review of algorithms by supervisory authorities. Examples from the financial sector include the key information documents (KIDs) for packaged retail and insurance-based investment products (PRIIPs, 2014), detailed motivations for credit actions under the Equal Credit Opportunity Act (ECOA, 1974) and filing requirements for general insurance rates to the National Association of Insurance Commissioners (NAIC, 2012). Our case study in Section 4.3 puts focus on general insurance pricing as one of the high-stakes XAI application areas where transparent decision-making is essential due to strict regulations.

4.1.1 Related works

Surrogate models aim to copy the behavior of a complex system by capturing its essence in a simpler format. Approaches like model compression (Buciă et al., 2006), mimic learning (Ba and Caruana, 2014) and distillation (Hinton et al., 2015) transfer knowledge from a complex/slow model into a simple/fast approximation. The resulting surrogate is still an opaque model, but is easier to deploy in environments with stringent space and time requirements. These methods purely focus on simpler implementations with faster execution times and lower memory requirements. Within XAI applications it is however very important to explain a system's underlying decision process. Explainability is often a subjective notion without a formal definition according to Lipton (2018). In effort to make this notion more specific, several classification schemes have emerged in recent literature (Adadi and Berrada, 2018; Carvalho et al., 2019; Arrieta et al., 2020; Burkart and Huber, 2021). We use the classification of Guidotti et al. (2018, Figure 4) to discuss the different roads towards explainability of black box models.

Model inspection The goal of model inspection is to understand the inner workings of a black box and how its behavior changes for different input conditions. Examples of popular tools are feature importance (Breiman, 2001), partial dependence plots (Friedman, 2001), individual conditional expectations (Goldstein et al., 2015) and accumulated local effects (Apley and Zhu, 2019). The provided level of information is typically rather low as these tools assess the effect of one or two features on the prediction target, thereby making it hard to obtain a global view and understanding of the black box behavior. Furthermore, Wachter et al. (2018) argue that these kinds of tools are hard to interpret by non-expert users. Methods for model inspection are however very valuable for experts (e.g., data scientists) to debug or validate a black box.

Model explanation With model explanation, we aim to understand the whole decision logic inside a back box. The internal decision process is explained via a *global surrogate*, i.e., a simple white box model that is constructed by using the black box predictions as targets with the goal to mimic its behavior. Example tools are TREPAN (Craven and Shavlik, 1995) and born again trees (**?**), both extracting a tree structure from a black box model. This type of explanation is intuitive for a non-expert user as its logic can be described via a decision tree or a list of rules. Fidelity then measures to which extent the surrogate is able to mimic the black box behavior. The surrogate needs to capture the correct feature interactions with a high degree of fidelity to avoid misleading explanations, which can be a challenging task.

Outcome explanation Outcome explanation has the goal to explain the black box prediction for a specific data instance. Individual predictions are explained via a *local surrogate*, i.e., a simple model constructed in the vicinity of the observation of interest. Examples of popular tools are LIME (Ribeiro et al., 2016), K-LIME (Hall et al., 2017), SHAP (Lundberg and Lee, 2017), Anchors (Ribeiro et al., 2018) and SLIM (Hu et al., 2020). These methods each have their own approach to derive the local model, possibly resulting in different explanations for the same data instance. The user therefore needs to understand the inner workings, validity domain and limitations of these outcome explanations to avoid wrong conclusions. These observations make it hard for non-experts to trust and correctly interpret the provided explanations. Furthermore, Laugel et al. (2018) show that there is a trade-off between the stability of explanations across similar instance predictions and fidelity to the original black box model.

In summary, the discussed explanation methods suffer from two main drawbacks. First, the need for expert knowledge to draw correct conclusions limits the usability for layman users. Second, explanations are based on a surrogate model while the complex black box remains in production. High fidelity is therefore crucial to obtain valid and non-misleading explanations. Guidotti et al. (2018) mention a fourth approach towards model explainability, namely a *transparent design.*

Transparent design An interpretable model is used from the start in a transparent design. As such, the model can be understood by non-experts while avoiding the issue of misleading explanations. Decision trees, rules and linear models are transparent by design, meaning they are easily comprehensible for human users. In linear models, the contribution (sign and strength) of feature x_j to the prediction target y is directly observable from the model coefficient β_j (Doran et al., 2017). Rudin (2019) advocates to only use interpretable models for high-stakes decisions because correct and clear explanations are crucial in such situations.

4.1.2 Research goals

The transparent design approach has two potential drawbacks. First, a simple interpretable model is expected to have reduced accuracy compared to a complex black box. However, Rudin (2019) argues that this is often not the case on structured data with meaningful features after careful pre-processing. Second, domain expertise might be needed to develop an interpretable model that makes sense and performs well. For a linear model, one has to define the model structure upfront by selecting the features, interactions and their representation. In this paper, we aim to streamline and automate the transparent design process. We extract knowledge from a complex black box via model inspection techniques and perform smart feature engineering. This allows to develop a transparent model which approximates the complex system. The resulting surrogate is simple enough to allow for easy/fast deployment and is highly transparent which eliminates the need for extra model explanation tools.

This paper presents a novel procedure to develop a *global surrogate* for a complex system, with the goal of implementing the surrogate in production. The surrogate inherits the strengths of a sophisticated black box algorithm, delivered in a simpler format that is easier to understand, manage and implement. For certain applications, it might be sufficient to have an explanation surrogate model while the black box remains in production. However, high-stakes decisions or strict regulations ask for full transparency of the model in production. We deliver a surrogate that aims to closely approximate the black box model such that it can be used as a substitute with (model and outcome) explanations readily available. The resulting high degree of model transparency can boost AI business applications, especially in highly regulated sectors such as banking and insurance.

Our procedure to construct a global surrogate first extracts knowledge from the complex system via model inspection techniques. Next, using these insights, it performs smart feature engineering on the training data. In the end, a transparent design model is fit to the engineered training data. We put forward the following three desirable properties for our procedure. Firstly, a *model-agnostic* procedure is preferred due to the ever increasing variety of black box algorithms. We rely on partial dependence (PD) effects to extract knowledge from the black box, thereby covering a vast amount of different model types (Friedman, 2001). Secondly, a *data-driven* procedure avoids the need for ad hoc model choices and assists model developers. This fully automates the transformation from black box to transparent surrogate. Thirdly, the resulting surrogate should be *interpretable* such that it is easy to comprehend by humans. Here, we use generalized linear models (GLMs), formulated by Nelder and Wedderburn (1972), as the transparent design model that is fit in the final step of our procedure. This versatile model class covers a broad range of classification and regression models and allows to represent its output as a decision table. Huysmans et al. (2011) perform a user study on the comprehensibility of several representation formats and show that decision tables outperform trees and rules with respect to accuracy, response time, answer confidence and ease of use. GLMs are widely used in the insurance industry thanks to the high degree of transparency (Goldburd et al., 2016).

We introduce maidrr: a procedure to develop a Model-Agnostic Interpretable Data-driven suRRogate for a black box model on structured tabular data. The complete procedure is implemented in the open source R package `maidrr` (Henckaerts, 2021). The rest of this paper is structured as follows. Section 4.2 details the maidrr methodology. Section 4.3 shows an application to insurance claim frequency modeling, where transparency is essential due to strict regulations imposed to insurance pricing models put in production. We demonstrate that our maidrr surrogate GLM is able to closely approximate a gradient boosting machine (GBM) black box, while outperforming a linear and tree benchmark surrogate. Section 4.4 concludes this paper.

4.2 Methodology

We first give an overview of the process behind maidrr, as schematized in Figure 4.1 and described in Algorithm 1. All the steps are detailed later in this section. The starting point of maidrr is a black box that we want to transform into a simpler and more comprehensible global surrogate. We extract knowledge from the black box in the form of partial dependence (PD) effects for all features x_j that pass an (optional) importance screening, with $j \in \{1, \dots, p\}$. These PD effects $\bar{f}_j$, describing the relation between a feature x_j and the target, are used to group values/levels within a feature via dynamic programming (DP). The grouping approach slightly differs depending on the type of feature. For continuous or ordinal features, only adjacent values may be binned together, whereas any two levels within a nominal feature can be clustered. The binning/clustering via DP leads to an optimal and reproducible grouping of levels. This results in a full segmentation of the feature space as all features are transformed to a categorical format x_j^c with k_j^* groups. Automatic feature selection is performed as only categorical features x_j^c with $k_j^* > 1$ groups are withheld. The same feature engineering process is followed to include interactions in the surrogate, with slight modifications detailed later on. Finally, a generalized linear model (GLM) is fit to the segmented data with selected features in a categorical format and their relevant interactions. The end product is an interpretable global surrogate which approximates the black box model and replaces it in production.

Algorithm 1 maidrr

```
Input: data, $f_{\text{pred}}$, $\lambda_{\text{marg}}$, $\lambda_{\text{intr}}$, $k$, $v$ and $h$
marginal
upfront feature selection: $F = \{j \mid j \in \{1, \dots, p\} \ , \ V(x_j) \geq v\}$
for all $j$ in $F$ do
    calculate the PD effect $\bar{f}_j$ via Eq. (4.1)
    apply the DP algorithm to feature $x_j$ with $k_j^* = \underset{k_j \in \{1,\dots,k\}}{\arg\min}$ Eq. (4.2)
    $x_j^c$ represents the grouped version of $x_j$ in categorical format with $k_j^*$ groups
end for
final feature selection: $F^* = F \setminus \{j \mid k_j^* = 1\}$
interaction
upfront interaction selection: $I = \{(l, m) \mid l \in F^* \ , \ m \in F^* \ , \ l \neq m \ , \ H(x_l, x_m) \geq h\}$
for all $(a, b)$ in $I$ do
    calculate the PD effect $\bar{f}_{a:b}$ via Eq. (4.3)
    apply the DP algorithm to interaction $x_{a:b}$ with $k_{ab}^* = \underset{k_{ab} \in \{1,\dots,k\}}{\arg\min}$ Eq. (4.4)
    $x_{a:b}^c$ represents the grouped version of $x_{a:b}$ in categorical format with $k_{ab}^*$ groups
end for
final interaction selection: $I^* = I \setminus \{(l, m) \mid k_{lm}^* = 1\}$
fit a GLM to the target with features $x_j^c$ for $j \in F^*$ and interactions $x_{a:b}^c$ for $(a, b) \in I^*$
Output: surrogate GLM
```

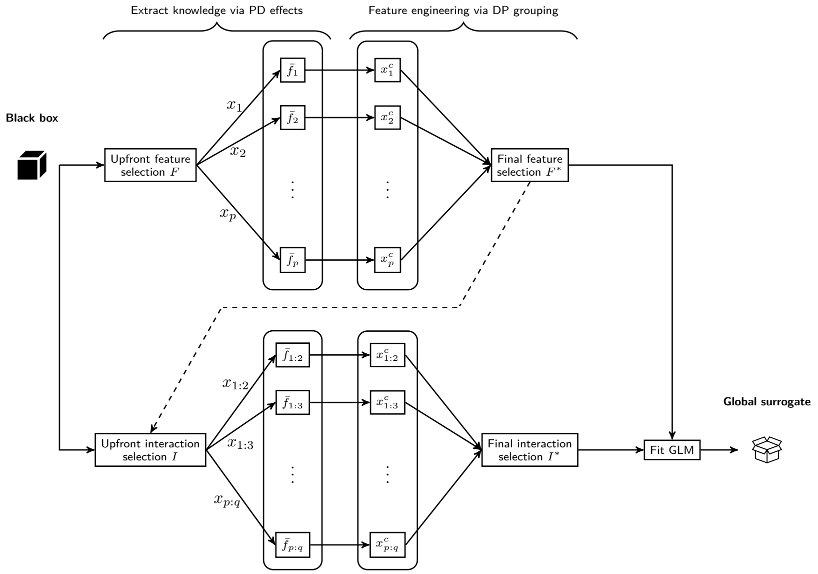


Figure 4.1: The maidrr process for transforming a black box algorithm into a transparent GLM.

4.2.1 Marginal effects: knowledge extraction

Any black box model giving a prediction function $f_{\text{pred}}(\boldsymbol{x})$ for features $\boldsymbol{x} \in \mathbb{R}^p$ can be used as a starting point, because maidrr is a model-agnostic procedure. With $j \in \{1, \ldots, p\}$, upfront feature selection is possible by only considering the features $x_j \in F$ for which an importance measure $V(x_j)$ exceeds a pre-specified threshold value v in Algorithm 1. We use the permutation approach of Breiman (2001) to calculate the measure $V(x_j)$. This approach quantifies the decrease in model performance after randomly changing a feature's values. Features that cause a high decrease in model performance after permutation are considered to be important.

We extract knowledge from the black box via partial dependence (PD) effects. A univariate PD captures the marginal relation between a feature x_j and the model predictions (Friedman, 2001). The PD effect $\bar{f}_j(x_j)$ evaluates the prediction function f_{pred} for a given value of feature x_j, while averaging over n observed values of the other features $\boldsymbol{x}^i_{-j}$ for observation $i \in \{1, \ldots, n\}$:

$$\bar{f}_j(x_j) = \frac{1}{n} \sum_{i=1}^{n} f_{\text{pred}}(x_j, \boldsymbol{x}^i_{-j}). \tag{4.1}$$

4.2.2 Marginal effects: feature segmentation

The PD effect $\bar{f}_j$ is used to group values/levels within feature x_j, as a similar PD indicates a similar relation to the prediction target. This grouping reduces the complexity of the feature with a limited loss of information. In theory, PDs can be misleading for correlated features and accumulated local effects (ALE) serve as an alternative (Apley and Zhu, 2019). However, Appendix C.1 compares the resulting PDs and ALEs for highly correlated features in our case study, justifying the use of PDs for grouping purposes.

For feature x_j, let m_j denote the unique number of observed values and let $x_{j,q}$ denote its qth value for $q \in \{1, \dots, m_j\}$. We then define $z_{j,q} = \bar{f}_j(x_{j,q})$ as the PD effect of feature x_j evaluated in $x_{j,q}$. The goal is now to arrange the values $x_{j,q}$ in k_j groups based on $z_{j,q}$. This represents a one-dimensional clustering problem of $z_{j,q}$ for $q \in \{1, \dots, m_j\}$. Wang and Song (2011) developed a dynamic programming (DP) algorithm for optimal and reproducible one-dimensional clustering problems. Elements of an m_j-dimensional vector $z_{j,q}$ are assigned to k_j clusters by minimizing the within-cluster sum of squares, that is, the sum of squared distances from each element to its corresponding cluster mean. This follows the same spirit as the K-means algorithm (MacQueen, 1967), but the DP algorithm guarantees reproducible and optimal groupings by progressively solving the sub-problem of clustering u elements in v clusters with $1 \leq u \leq m_j$ and $1 \leq v \leq k_j$. This algorithm is implemented in the R package `Ckmeans.1d.dp` (Song, 2019) and allows for the inclusion of adjacency constraints in the clustering problem. We impose such constraints for continuous and ordinal features to group adjacent values. Nominal features with no specific ordering are clustered without constraints such that any two levels can be grouped. The DP algorithm requires the specification of the number of groups k_j for feature x_j to group $z_{j,q}$.

In theory, we can perform a p-dimensional grid search to find the optimal k_j for each feature x_j with $j \in \{1, \dots, p\}$. However, this would cause the computation time to grow exponentially with p, harming maidrr's scalability. We thus propose a penalized loss function to find the optimal number of groups k_j in $\{1, \dots, k\}$, where k is the hyperparameter in Algorithm 1 that allows to specify the maximum number of groups per feature. After grouping feature x_j in k_j groups, let $\tilde{z}_{j,q}$ represent the average PD effect for the group to which $x_{j,q}$ belongs. The penalized loss function is then defined as follows:

$$\sum_{q=1}^{m_j} w_{j,q} \left(z_{j,q} - \tilde{z}_{j,q}\right)^2 + \lambda_{\text{marg}} \log(k_j). \tag{4.2}$$

The first part of this loss function measures how well the PD effect is approximated by the grouped variant as a weighted mean squared error (wMSE) over all unique values of feature x_j. The weight $w_{j,q}$ represents the proportion of observations that equal value $x_{j,q}$ for feature x_j. This forces the procedure to focus on closely approximating frequently occurring feature values as opposed to rare cases. The second part of Eq. (4.2) measures the complexity by means of the common logarithm of the number of groups k_j. The penalty parameter λ_{marg} in Eq. (4.2) acts as a bias-variance trade-off. A low (high) value of λ_{marg} allows for many (few) groups, resulting in an accurate (coarse) approximation of the PD. Note that λ_{marg} does not depend on j because the PD effects reside on the same scale, namely the scale of the predictions, see Eq. (4.1). The original p-dimensional tuning problem in this way reduces to be one-dimensional over λ_{marg}. We minimize Eq. (4.2) for each feature x_j, resulting in a full segmentation F^* of the feature space. Automatic feature selection is enabled as x_j is excluded from the surrogate when $k_j = 1$.

4.2.3 Interaction effects: knowledge and segmentation

So far we focused on grouping features via their marginal PDs, but feature interactions can play a major role in explaining the data. Interaction strength in the black box model is measured via the H-statistic (Friedman and Popescu, 2008). We find a set of relevant interactions I by selecting all pairwise combinations of the features present in F^* whose realized values of the H-statistic exceed a pre-specified threshold value h in Algorithm 1.

The pure interaction effect between features x_a and x_b is captured by subtracting both one-dimensional PDs from the two-dimensional PD:

$$\bar{f}_{a:b}(x_a, x_b) = \frac{1}{n}\sum_{i=1}^{n} f_{\text{pred}}(x_a, x_b, \boldsymbol{x}^i_{-a,-b}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in\{a,b\}} f_{\text{pred}}(x_\ell, \boldsymbol{x}^i_{-\ell}). \quad (4.3)$$

We define feature $x_{a:b}$ as the interaction containing all combinations of features x_a and x_b in the original data. The DP algorithm clusters levels in $x_{a:b}$ that have similar $\bar{f}_{a:b}(x_a, x_b)$ values, without any adjacency constraints. We allow for such maximum flexibility because interactions represent a correction on top of the marginal effects. Defining $z_{a,c:b,d} = \bar{f}_{a:b}(x_{a,c}, x_{b,d})$ and $\tilde{z}_{a,c:b,d}$ as the average PD effect for the group to which $(x_{a,c}, x_{b,d})$ belongs, we determine the number of groups k_{ab} by minimizing the two-dimensional equivalent of Eq. (4.2) as follows:

$$\sum_{c=1}^{m_a}\sum_{d=1}^{m_b} w_{a,c:b,d}\left(z_{a,c:b,d} - \tilde{z}_{a,c:b,d}\right)^2 + \lambda_{\text{intr}}\log(k_{ab}). \quad (4.4)$$

We minimize Eq. (4.4) for each interaction $x_{a:b}$, resulting in a segmentation of select interactions I^*. A distinct value of λ is advised for marginal (λ_{marg}) and interaction (λ_{intr}) effects, as the PDs in Eq. (4.1) and (4.3) reside on different scales. We tune both λ's from Eq. (4.2) and (4.4) via a grid search and K-fold cross-validation. The optimal λ values minimize loss functions that balance the original target observations and the surrogate GLM predictions, resulting in a data-driven procedure. Section 4.3.2 further details our sequential tuning approach for λ_{marg} and λ_{intr}.

4.2.4 Global surrogate

In a final step, we fit a transparent model to the original target with selected features F^* and interactions I^* in a categorical format. Generalized linear models (GLMs) allow for the specification of a diverse set of target distributions (Nelder and Wedderburn, 1972). This facilitates the application of maidrr to classification tasks and many types of regression problems, for example linear, Poisson and gamma regression. We refer to Appendix C.2 for details on the GLM formulation. GLMs with only categorical features lead to fixed-size decision tables, see Appendix C.3. Even with many features they remain transparent, fileable in a tabular format and easy to use by business intermediaries, so the complexity of the GLM is not a concern.

4.2.5 Computational complexity

We detail the computational cost of the steps in Algorithm 1. For training data with n records, the calculation of the permutation feature importance measure $V(x_j)$ for $j \in \{1, \ldots, p\}$ requires $n \times (1+p)$ evaluations of the prediction function f_{pred}, namely once for the original data and once for each of the p permuted datasets. Calculating a PD effect $\bar{f}_j$ for feature x_j with m_j unique values via Eq. (4.1) requires $n \times m_j$ evaluations of f_{pred}. This can become computationally expensive for high values of n and/or m_j, but several strategies exist to ease the computational burden (Greenwell, 2017). Firstly, the PD calculation can be parallelized over different values of x_j. Secondly, it is generally not necessary to evaluate the prediction in each of the m_j values as reasonable results are obtained with a reduced number of points. Thirdly, a subset of the full training data of size smaller than n can be used to speed up computations. The main PD effects are calculated for all features in the upfront selection of size $|F| \leq p$. The interaction PD effect for $x_{a:b}$ requires at most $n \times m_a \times m_b$ evaluations of f_{pred} and is calculated for all pairwise feature combinations obtained from the final (univariate) feature selection. These main and interaction PD effects only need to be calculated once and can then be reused in the calculation of the H-statistic. Overall, the maximum total number of f_{pred} evaluations equals $n \times (1 + p + \sum_{j=1}^{p} m_j + \sum_{j=1}^{p} \sum_{i=1, i \neq j}^{p} m_j \times m_i)$ with a complexity of at most $O(np^2M^2)$ where $M = \max(m_1, \ldots, m_p)$. The actual number of evaluations will typically be much lower due to feature selection and the strategies mentioned earlier. The DP clustering algorithm of Wang and Song (2011) has a complexity of $O(k_j m_j^2)$ to cluster m_j values in k_j groups. Standard GLM algorithms have a complexity of $O(p^3 + np^2)$ for n records and p features (Nykodym et al., 2016).

4.3 Case study for the insurance industry

In most jurisdictions, insurers are required by law to document their pricing or rating model to the regulator (in detail) and to customers (high-level). Determining a fair insurance quote is high-stakes with a big impact on a person's life. This creates a clear need for transparency in the underlying decision-making process. GLMs are currently a widely used pricing tool within the strictly regulated insurance industry. Their high degree of transparency, thanks to observable coefficients, allows intuitive model post-processing by industry experts. Actuaries first construct a technical tariff with GLMs and then, in dialogue with product managers and marketing, they manually tweak the coefficients to develop the commercial tariff that is eventually put into production. A crucial part in this pricing process is the accurate modeling of the number of claims reported by a policyholder. We therefore apply maidrr to a general insurance claim frequency prediction problem. Section 4.3.1 introduces the model setting and datasets. Section 4.3.2 details the model construction for the black box and our GLM surrogate. Section 4.3.3 evaluates the performance of the GLM with respect to the black box against two benchmark surrogates.

4.3.1 Claim frequency modeling with insurance data

We analyze six motor third party liability (MTPL) insurance portfolios, which are available in the `R` packages `CASdatasets` (Dutang and Charpentier, 2019) or `maidrr` (Henckaerts, 2021). All datasets contain an MTPL portfolio followed over a period of one year, with the number of policyholders n and the number of features p detailed in Table 4.1. Each dataset holds a collection of different types of risk features, for example the age of the policyholder (numeric), the region of residence (nominal) and the type of insurance coverage (ordinal).

We model the number of claims filed during a given period of exposure-to-risk, defined as the fraction of the year for which the policyholder was covered by the insurance policy. Exposure is vital information, as filing one claim during a single month of coverage represents a higher risk than filing one claim during a full year. Table 4.1 shows the distribution of the number of claims in the portfolios. Most policyholders do not file a claim, some file one claim and a small portion files two or more claims. Such count data is often modeled via Poisson regression, a specific form of GLM with a Poisson assumption for the target y and a logarithmic link function. In this setting, the industry standard is to incorporate the logarithm of exposure t via an offset term: $\ln(\mathbb{E}[y]) = \ln(t) + \beta_0 + \sum_j \beta_j x_j$. This leads to $\mathbb{E}[y] = t \times \exp(\beta_0 + \sum_j \beta_j x_j)$, that is, predictions are proportional to exposure and have a multiplicative structure: $\mathbb{E}[y] = t \times \exp(\beta_0) \times \prod_j \exp(\beta_j x_j)$.

			number of claims						
dataset	n	p	0	1	2	3	4	5	6
`ausprivauto`	67,856	5	63,232	4,333	271	18	2	0	0
`bemtpl`	163,210	10	144,936	16,539	1,554	162	17	2	0
`freMPL`	137,254	9	106,577	26,068	4,097	448	62	2	0
`freMTPL`	677,925	8	643,874	32,175	1,784	82	7	2	1
`norauto`	183,999	4	175,555	8,131	298	15	0	0	0
`pricingame`	99,859	19	87,213	11,232	1,262	134	16	1	1

Table 4.1: Overview of the number of policyholders n, number of features p and distribution of the number of claims in the portfolios. The dataset names correspond to those in the `CASdatasets` or `maidrr` R packages.

4.3.2 Finding a transparent model by opening the black box

Section 4.3.2 describes the construction of a gradient boosting machine or GBM as black box. Section 4.3.2 details the maidrr procedure to obtain a GLM surrogate and illustrates the automatic feature selection and segmentation for several datasets.

GBM as black box

We opt for a gradient boosting machine or GBM (Friedman, 2001) as the black box to start from. More specifically, we use stochastic gradient boosting (Friedman, 2002) as implemented in the R package `gbm` (Greenwell et al., 2019). This choice is justified by the good performance of GBMs in Chapter 3 to predict claim frequency and severity data. The model-agnostic nature of maidrr allows any model to be used as input, including deep neural networks.

We tune the number of trees T in the GBM via 5-fold cross-validation, see Table 4.2. Other hyperparameters are fixed to a sensible value. Following Hastie et al. (2009, Section 10.11), we use decision trees of depth two, which are able to model up to third-order interactions. Each tree is built on randomly sampled data of size $0.75n$ and the learning rate is set to 0.01. To take into account the distributional characteristics of the count data, we use the Poisson deviance as loss function in the GBM tuning process. The Poisson deviance is defined as follows:

$$D^{\text{Poi}} \left\{ y, f_{\text{pred}}(\boldsymbol{x}) \right\} = \frac{2}{n} \sum_{i=1}^{n} \left[y_i \times \ln \left\{ \frac{y_i}{f_{\text{pred}}(\boldsymbol{x}_i)} \right\} - \left\{ y_i - f_{\text{pred}}(\boldsymbol{x}_i) \right\} \right] . \tag{4.5}$$

	ausprivauto	bemtpl	freMPL	freMTPL	norauto	pricingame
T	474	3,214	1,377	3,216	793	1,198

Table 4.2: Overview of the optimal number of trees (T) in the GBM for the different datasets.

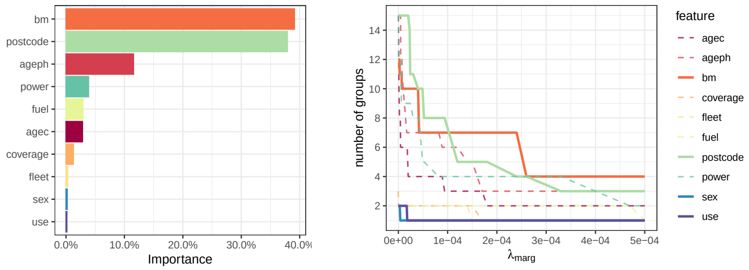
GLM surrogate via maidrr

We build a surrogate GLM to approximate the optimal GBM for each dataset. The function `maidrr::autotune` (Henckaerts, 2021) implements a tuning procedure for Algorithm 1 which requires five input parameters: λ_{marg}, λ_{intr}, k, v and h. The λ values determine the granularity of the resulting segmentation and GLM. We define a search grid for both λ's, ranging from 10^{-10} to 1. This range is sufficiently wide for our application, as indicated by the optimal values in Table 4.3. These λ values are determined by performing 5-fold cross-validation on the resulting GLM with the Poisson deviance in Eq. (4.5) as loss function. Tuning of the λ values is done in two stages. First, a grid search over λ_{marg} is performed by running the “marginal” part of Algorithm 1 and fitting a GLM. This results in the optimal GLM with only marginal effects. Then, a grid search over λ_{intr} determines which interactions to include in that marginal GLM by running the “interaction” part of Algorithm 1. This requires two one-dimensional grid searches of length $grid\_size$ instead of one two-dimensional search of length $grid\_size^2$, thereby saving computation time. We perform no upfront feature selection by setting $v = 0$. The value of h determines the set of interactions that are considered for inclusion in the GLM by excluding meaningless interactions with a low H-statistic. This value is calculated automatically to consider the minimal set of interactions for which the empirical distribution function of the H-statistic exceeds 50%. We set the maximum number of groups $k = 15$.

	ausprivauto	bemtpl	freMPL	freMTPL	norauto	pricingame
λ_{marg}	4.2×10^{-5}	4.2×10^{-5}	1.6×10^{-4}	1.3×10^{-7}	1.1×10^{-5}	2.0×10^{-6}
λ_{intr}	8.5×10^{-6}	4.1×10^{-6}	4.6×10^{-5}	3.1×10^{-6}	3.1×10^{-5}	2.8×10^{-6}

Table 4.3: Overview of the optimal λ_{marg} and λ_{intr} values for the different datasets.

Figure 4.2 illustrates the automatic feature selection of maidrr for the bemtpl portfolio. Figure 4.2a shows feature importance scores according to the GBM and Figure 4.2b shows the number of groups for each feature in function of λ_{marg}. Important features, such as bm and postcode, retain a higher number of groups for increasing values of λ_{marg}. Levels of uninformative features, like use and sex, are quickly placed in one group, effectively excluding these variables from the GLM. This is how maidrr performs automatic feature selection via the data-driven tuning of λ_{marg}.

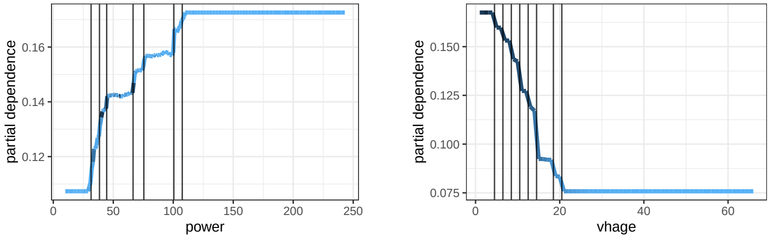


(a) Feature importance in the GBM **(b)** Number of groups in function of λ_{marg}

Figure 4.2: Illustration of the automatic feature selection process in maidrr for bemtpl.

Figure 4.3 displays the resulting segmentation for two continuous features: vehicle power for bemtpl in Figure 4.3a and vehicle age for pricingame in Figure 4.3b. Both show the GBM PD effect, where darker blue indicates a higher observation count in the portfolio. The features are grouped into 8 and 9 bins respectively, indicated by the vertical lines. The bins are wide wherever the PD effect is quite stable and narrow where the effect is steeper. We observe that claim risk increases for increasing vehicle power, while it decreases for increasing vehicle age.

Figure 4.4 displays the resulting segmentation for three categorical features. Groups are indicated by different plotting characters, with size proportional to the observation count in the portfolio. Figure 4.4a shows that claim risk decreases with increasing age of the policyholder in the ausprivauto portfolio. Due to similar PD effects, both levels containing the oldest policyholders are grouped together as well as both levels containing the people of working age. This results in four age segments: youngest, young, working and older people. Figure 4.4b shows that claim risk decreases for a decreasing driving distance limit in the `norauto` portfolio. The PD effects are dissimilar enough not to be grouped together, so each level remains in a separate segment. Figure 4.4c shows the PD effects and resulting grouping for vehicle makes in the `norauto` portfolio. The 41 different makes are divided in 11 segments with {Mazda, Jeep} and {Lada, Unic, Other} as the most and least risky segments respectively. Categorical features with many levels are often hard to deal with in practice. Appendix C.4 demonstrates how maidrr greatly reduces the complexity for geographical information in the `bemtpl` and `pricingame` portfolios.

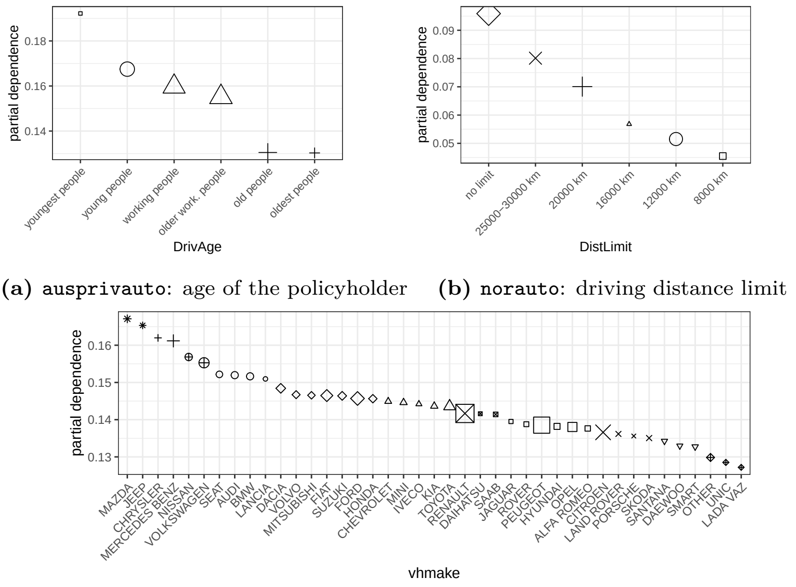


(a) `bemtpl`: power of the vehicle in hp **(b)** `pricingame`: age of the vehicle in years

Figure 4.3: PD effect and the resulting segmentation for two continuous features. Groups are separated by vertical lines and darker blue indicates a higher observation count in the portfolio.

4.3.3 Evaluation of the GLM surrogate

This section evaluates the performance of our maidrr GLM based on three desiderata for a surrogate model: accuracy, fidelity and interpretability (Guidotti et al., 2018, Section 3.2). Since accurate predictions matter highly for a model to remain competitive and relevant in production, Section 4.3.3 puts focus on accuracy. Section 4.3.3 evaluates fidelity as the extent to which the surrogate is able to mimic the behavior of the original black box. Section 4.3.3 evaluates interpretability because the surrogate should be comprehensible and easy to use in practice. We benchmark our GLM against two transparent global surrogates: a linear model (LM) and a decision tree (DT) of four levels depth. We fit both with the original data as features and the GBM predictions as target to capture the GBM's underlying decision process (Molnar, 2020, Section 5.6.1).



(a) `ausprivauto`: age of the policyholder **(b)** `norauto`: driving distance limit

(c) `pricingame`: make of the vehicle

Figure 4.4: PD effect and the resulting segmentation for three categorical features. Groups are indicated by plotting characters, with size proportional to the observation count in the portfolio.

Accuracy

The goal of our maidrr GLM surrogate is to approximate a complex black box and replace it in the production pipeline. It is therefore vital that the GLM results in accurate predictions with minimal accuracy loss compared to the black box. We measure prediction accuracy via the Poisson deviance from Eq. (4.5), where smaller deviance values indicate higher accuracy. With f_{surro} and f_{gbm} the surrogate and GBM prediction function, we assess the accuracy loss via percentage differences as follows: $\Delta D^{\text{Poi}} = 100 \times \big(D^{\text{Poi}}\{y, f_{\text{surro}}(\boldsymbol{x})\}/D^{\text{Poi}}\{y, f_{\text{gbm}}(\boldsymbol{x})\} - 1\big)$.

Table 4.4 shows the Poisson percentage differences ΔD^{Poi} for the GLM, LM and DT surrogates with respect to the GBM black box. Results are shown for each dataset separately and the last column contains the average over all datasets. The maidrr GLM attains the lowest accuracy loss and outperforms the benchmark surrogates on each dataset. The GLM's accuracy loss stays below 0.5% on four out of six datasets, with an average of 0.64% over all datasets. On average, the GLM is 3 and 7.5 times as accurate as the DT and LM surrogates.

	ausprivauto	bemtpl	frempl	fremtpl	norauto	pricingame	avg.
GLM	**0.10**	**0.49**	**1.80**	**0.92**	**0.03**	**0.48**	**0.64**
LM	0.22	1.15	18.39	6.35	0.07	2.53	4.79
DT	0.25	1.68	4.82	2.66	0.28	2.13	1.97

Table 4.4: Poisson percentage differences ΔD^{Poi} for the different surrogate models.

Fidelity

This section investigates how closely the maidrr GLM mimics the behavior of the GBM black box by assessing how well the surrogates replicate the GBM's predictions. Firstly, we compute Pearson's linear and Spearman's rank correlation coefficients ρ (Weaver et al., 2017, Chapter 10) between the GBM and surrogate predictions. We average these coefficients to consolidate both types of correlation in one number, but the results below also hold for each coefficient separately. Secondly, the R^2 measure represents the percentage of variance that the surrogate model is able to capture from the black box (Molnar, 2020, Section 5.6.1). With μ_{gbm} the mean GBM prediction, the $R^2 \in [0,1]$ is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \{f_{\mathrm{surro}}(\boldsymbol{x}_i) - f_{\mathrm{gbm}}(\boldsymbol{x}_i)\}^2}{\sum_{i=1}^{n} \{f_{\mathrm{gbm}}(\boldsymbol{x}_i) - \mu_{\mathrm{gbm}}\}^2}.$$

Table 4.5 shows the averaged ρ for the GLM, LM and DT surrogates on each dataset separately and averaged over all datasets in the last column. The GLM ranks first in all datasets, thereby outperforming both benchmark surrogates. The correlation between the GBM and GLM is at least 95% on four out of six datasets, with an average of 95% over all datasets. On average, the GLM's correlation to the GBM is 12% and 9% higher compared to the DT and LM surrogates.

	ausprivauto	bemtpl	frempl	fremtpl	norauto	pricingame	avg.
GLM	**0.95**	**0.97**	**0.91**	**0.92**	**0.99**	**0.97**	**0.95**
LM	0.95	0.93	0.74	0.60	0.98	0.95	0.86
DT	0.86	0.83	0.75	0.78	0.91	0.87	0.83

Table 4.5: Average correlation coefficient ρ for the different surrogate models.

Table 4.6 shows the R^2 for the GLM, LM and DT surrogates on each dataset separately and averaged over all datasets in the last column. The GLM ranks first in five datasets and second in `ausprivauto`. The GLM captures more than

90% of variance on four out of six datasets, with an average of 90% over all datasets. On average, the GLM captures an extra 12% and 15% of variance compared to the DT and LM surrogates.

	ausprivauto	bemtpl	frempl	fremtpl	norauto	pricingame	avg.
GLM	0.86	**0.94**	**0.91**	**0.78**	**0.99**	**0.93**	**0.90**
LM	**0.89**	0.83	0.62	0.30	0.95	0.88	0.75
DT	0.75	0.74	0.88	0.75	0.84	0.76	0.78

Table 4.6: R^2 measure for the different surrogate models.

For both the ρ and R^2 measures, DT outperforms LM on the `frempl` and `fremtpl` datasets while LM outperforms DT on the remaining four datasets. This is driven by the fact that the DT puts focus on interactions while the LM puts focus on marginal effects. Our maidrr GLM combines both marginal and interaction effects, resulting in better performance overall.

We conclude that our GLM constructed with maidrr outperforms the benchmark DT and LM surrogates when it comes to both prediction accuracy and mimicking the GBM's underlying behavior. Remember that the DT and LM are trained with the GBM's predictions as target. The maidrr procedure extracts knowledge from the GBM to perform smart feature engineering, but afterwards the GLM is fit to the original target. The observation that the GLM is better at mimicking the GBM compared to the benchmark surrogates is therefore especially interesting.

Interpretability

Global interpretations A GLM is globally interpretable as the model coefficients, relating the features to the predictions, are easily observable. Appendix C.3 details global interpretations for the GLM to model the number of claims in the `norauto` dataset. Our maidrr procedure outputs a GLM with all features in a categorical format. This allows to summarize the full working regime of the GLM in a decision table, see Table C.1 in Appendix C.3. In practice, such a tabular model is easy to represent and maintain in a spreadsheet with responsive filters. Decision tables are very comprehensible for human users and outperform both trees and rules in accuracy, response time, answer confidence and ease of use (Huysmans et al., 2011).

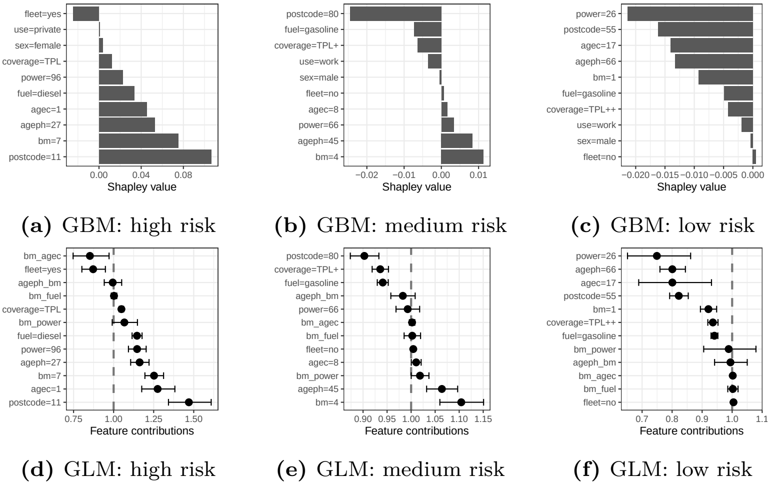
Local interpretations We now turn to explaining individual predictions and illustrate this with the three artificial instances in the `bemtpl` dataset listed in Table 4.7. Based on the GBM and GLM predictions, these instances represent a high/medium/low risk profile. We want to assess how the features influence the riskiness of each individual. Feature contributions in a GLM can be extracted via the fitted coefficients, as implemented in `maidrr::explain` (Henckaerts, 2021). For comparison purposes we use Shapley (1953) values to explain the GBM predictions, with the efficient implementation of Štrumbelj and Kononenko (2010, 2014) available in the R package `iml` (Molnar et al., 2018).

	high risk	medium risk	low risk
`bm`	7	4	1
`postcode`	11	91	55
`ageph`	27	45	66
`power`	96	66	26
`fuel`	diesel	gasoline	gasoline
`agec`	1	8	15
`coverage`	TPL	TPL+	TPL++
`fleet`	yes	no	no
`sex`	female	male	male
`use`	private	work	work
GBM	0.2847	0.1398	0.0502
GLM	0.3861	0.1231	0.0413

Table 4.7: Artificial instances in the `bemtpl` portfolio for which we explain the individual predictions.

Figures 4.5a, 4.5b and 4.5c show the Shapley values for the GBM prediction of each instance. The sum of these values equals the difference between the instance prediction, shown in Table 4.7, and the average GBM prediction of 0.1417. The presence of mainly positive/negative Shapley values in Figure 4.5a/4.5c thus represents a high/low risk profile respectively. Figures 4.5d, 4.5e and 4.5f show the GLM's feature contributions on the response scale after taking the inverse link function, namely $\exp(\beta_j)$ for feature x_j in our Poisson GLMs with log link. The contributions are multiplicative with respect to the baseline prediction of $\exp(-2.04) = 0.13$ from the intercept. The gray dashed line indicates the point of "no contribution" at $\exp(0)$. Furthermore, the GLM allows to split the contributions over marginal effects and interactions with other features, while 95% confidence intervals indicate the uncertainty around each contribution.

The GBM and GLM explanations are very similar. For example, Figures 4.5a and 4.5d attribute this profile's high risk to a residence in Brussels, young age, high bonus-malus level and driving a new high-powered diesel vehicle. The interaction between the bonus-malus level and age of the vehicle puts a negative correction on both positive marginal effects in the GLM, while the other interactions have limited impact on the prediction. The GLMs show no



(a) GBM: high risk **(b)** GBM: medium risk **(c)** GBM: low risk

(d) GLM: high risk **(e)** GLM: medium risk **(f)** GLM: low risk

Figure 4.5: Explanations for the high (left), medium (middle) and low (right) risk instance predictions from Table 4.7 in the GBM via Shapley values (top) and the GLM via β coefficients (bottom).

contribution from gender as this feature is not selected by maidrr, while it has negligibly small Shapley values in all cases. An insurance rate is determined by the product of claim frequency and severity, such that the contributions can be directly interpreted as a percentage premium/discount on the price. Living in Brussels increases the baseline frequency, and thus the price, by almost 50% in the technical analysis for this dataset. One can assess the fairness of this penalty, possibly followed by a manual adjustment to intervene in the decision-making process via expert judgment.

Other interpretations We focused on the GBM's Shapley values and GLM's β_j coefficients to answer the question *"How does each feature contribute to an instance prediction?"*. As Section 4.1 indicated, there exists a wide variety of other interpretation tools to answer different questions. We briefly discuss two options and connect these to GLM explanations. LIME fits a local surrogate to observations in the neighborhood of an instance of interest. This local model is then used to explain the black box model's behavior and to answer the question *"How does the prediction change in the vicinity of this instance?"* (Ribeiro et al., 2016). The observable β_j coefficients in a GLM allow to directly quantify how a prediction changes if a continuous feature increases/decreases or when a categorical feature switches levels. Furthermore, these prediction changes relate directly to the GLM in production instead of a surrogate which merely approximates the black box's behavior. Counterfactual explanations answer the question *"How do we need to change an instance's feature values to change the prediction to a predefined value?"* (Wachter et al., 2018). This indicates how a data instance needs to change in order to obtain a desired prediction output from a model. Given an instance's current prediction and the desired target value, the observable β_j coefficients allow to reverse engineer which feature changes are necessary. For practical usability, one can limit the options to those features that one actually has control over to change.

4.4 Conclusions

Decision-making algorithms in business practice can become highly complex in order to gain a competitive advantage. However, transparency is critical for any high-stakes decision or for companies active in strictly regulated industries. To balance accuracy and explainability, we present maidrr: a procedure to develop a Model-Agnostic Interpretable Data-driven suRRogate for a complex system. The paper is accompanied by an R package in which the procedure is implemented (Henckaerts, 2021). We apply maidrr to six real-life general insurance portfolios for claim frequency prediction, with insurance pricing as an example of a high-stakes decision in a strictly regulated industry. We thereby put focus on a highly relevant count regression problem, which is not often dealt with in classical machine learning literature. Our maidrr procedure results in a surrogate GLM which closely approximates the performance of a black box GBM in terms of accuracy and fidelity, while outperforming two benchmark surrogates. The resulting GLM can then be deployed in the production pipeline with minimal performance loss. In the process, maidrr automatically performs feature selection and segmentation, providing a possibly useful by-product for customer or market segmentation applications.

Both global and local interpretations are easily extracted from our maidrr GLM. Explanations only depend on the fitted coefficients, which are easily observable and presentable on the response scale. This representation boosts the ability to understand the feature contributions on the scale of interest and allows for manual intervention when deploying the model in practice. This gives some important advantages to maidrr with respect to the following XAI goals (see Arrieta et al., 2020, Table 1). 1) *Trustworthiness:* a GLM with only categorical features always acts as intended since all the possible working regimes can be listed in a decision table of fixed size. 2) *Accessibility/Interactivity:* manual post-processing of the model becomes very easy and intuitive by tweaking the GLM coefficients. This allows users to intervene and be more involved in the development and improvement of the model. 3) *Fairness:* the clear influence of each feature allows for an ethical analysis of the model, which becomes especially important for high-stakes decisions which influence people's lives. In our insurance setting, it is important that every policyholder receives a fair insurance quote. The direct interpretation of the feature contributions as a penalty/discount to the baseline tariff further serves this cause. 4) *Confidence:* the uncertainty of the contributions is quantifiable via confidence intervals such that the model's robustness, stability and reliability can be assessed. 5) *Informativeness:* contributions are split across marginal effects and interactions of features, thereby increasing the amount of information available to the user on the underlying decision of the model.

Our maidrr procedure combines the inherent interpretability of a GLM with the data knowledge extracted from a sophisticated black box. We therefore believe that maidrr can serve as a useful tool in any situation where a competitive, yet transparent model is needed.

Chapter 5

Dynamically updating motor insurance prices with driving behavior data

We analyze a novel dataset collecting the driving behavior of young policyholders in a motor third party liability (MTPL) portfolio, followed over a period of three years. Driving habits are measured by the total mileage and the distance driven on different road types and during distinct time slots. Driving style is characterized by the number of harsh acceleration, braking, cornering and lateral movement events. First, we develop a baseline pricing model for the complete portfolio with claim history and self-reported risk characteristics of approximately 400,000 policyholders each year. Next, we propose a methodology to update the baseline price via the telematics information of young drivers. Our approach results in a truly usage-based insurance (UBI) product, making the premium dependent on a policyholder's driving habits and style. We highlight the added value of telematics via improvements in risk classification and we put focus on managerial insights by analyzing expected profits and retention rates under our new UBI pricing structure.

This chapter is based on joint work with Katrien Antonio. This research is supported by the Research Foundation Flanders [SB grant 1S06018N]. We are very grateful to the contact persons at the insurance company for providing the data analyzed in this work, for the insightful meetings with interesting feedback and for the smooth co-operation.

5.1 Introduction

Property and casualty (P&C) insurance is a highly data-driven business, where proper risk assessment is fundamental in several applications. Insurance pricing is the process of determining a fair accurate premium through risk classification. Traditional pricing relies on a policyholder's self-reported risk characteristics, for example driver age, vehicle power or residence location in motor insurance. These characteristics allow an actuary to form groups of policyholders with similar perceived risk. However, these features merely act as proxy measurements for the actual risk. Vickrey (1968) was the first to express critique towards the static pricing structure in motor insurance, advocating to link premiums to vehicle use. With the advent of digitization and big data, telematics technology allows to access new sources of information via the integrated use of *tele*communications and infor*matics* (Husnjak et al., 2015).

Telematics can serve as a monitoring tool for risk prevention, for example via smart wearables which stimulate a healthy lifestyle in health insurance or smart sensors which detect fires, leaks or intrusion in home insurance (Eling and Kraft, 2020). Personalized feedback on risky behavior and financial incentives motivate positive behavioral changes (Ellison et al., 2015). Customer are generally willing to share personal information for new pricing paradigms or additional services within motor and home insurance, whereas sharing health-data is less accepted (Maas et al., 2008). Telematics has great application potential within motor insurance and other innovative mobility services (Longhi and Nanni, 2020). Usage-based motor insurance (UBI) makes the price of a policy dependent on the vehicle use and corresponding driving behavior via *pay-as-you-drive* (PAYD) and *pay-how-you-drive* (PHYD) schemes (Tselentis et al., 2016). PAYD puts focus on driving habits (e.g., distance driven, time of day or road type) while PHYD takes into account driving style (e.g., aggressive acceleration, sudden lane shifts or excessive speeding).

Motor UBI products provide significant benefits to insurers, customers and society in general. Monitoring driving behavior allows to reduce asymmetric information between the insurer and its policyholders, thereby mitigating the problems of moral hazard and adverse selection (Filipova-Neumann and Welzel, 2010). UBI gives insurance companies the chance to innovate and profit from new business models by increasing revenues (e.g., by tapping into underexploited market segments) and/or decreasing costs (e.g., by a reduction of crash rates, claim costs and fraud) (Desyllas and Sako, 2013). The benefits of reduced crashes and other operational gains outweigh the system's costs, making telematics economically viable (Pitera et al., 2013). More accurate assessment of the underlying claim risk leads to higher actuarial accuracy, fairness and economic efficiency, which in turn reduces cross-subsidies between groups and premium leakage (Litman, 2011). UBI has the opportunity to stimulate responsible driving by providing interactive feedback that motivates and engages users, making the customer experience more exciting (Toledo et al., 2008). Progressive pricing towards low income drivers increases insurance affordability through consumer savings, resulting in less uninsured driving (Litman, 2004). Reduced vehicle travel leads to many societal benefits such as increased road safety with less crashes and a reduction in traffic congestion, fuel consumption, oil dependence, CO_2 emissions, air pollution and road costs (Parry, 2005; Bordhoff and Noel, 2008; Greenberg, 2009).

Recent regulatory developments in Europe are, indirectly, endorsing the use of telematics in insurance. Following the Test-Achats Ruling, the European Commission adopted Guidelines to prohibit price discrimination at the individual level between men and women (OJ/C11, 2012). Ayuso et al. (2016b) explain women's lower accident risk by a lower driving intensity and less risky behavior compared to men. Ayuso et al. (2016a) show that, when taking driving intensity into account, gender no longer has a significant effect in explaining the time to the first accident at fault. Verbelen et al. (2018) find that driving behavior renders gender redundant as a rating factor. This suggests that gender differences regarding claim risk are, to a certain extent, attributable to differences in driving behavior between men and women. Telematics can leverage this new information and reduce the need to rely on, possibly discriminatory, proxy characteristics. Next to this, all new motor vehicles in the EU are required to be equipped with eCall technology as of April 2018 (OJ/L123, 2015). This system automatically sends location data to emergency services in case of an accident and facilitates to offer UBI services.

State-of-the-art P&C insurance pricing follows a *frequency-severity* approach: modeling claim counts and sizes independently with generalized linear or additive models (GLM/GAM) (Denuit et al., 2019b). Various actuarial studies compare predictive model performance when using 1) only traditional features, 2) only telematics information and 3) the combination of both in a hybrid set-up. The occurrence of a claim in these studies is predicted with logistic regression (LR), random forests (RF) and neural networks (NNs) by both Baecke and Bocca (2017) and Huang and Meng (2019), where the latter also include support vector machines (SVMs) and extreme gradient boosting (XGBoost) in their comparison. Gao et al. (2019) predict claim frequency with Poisson GAMs and telematics features extracted from speed-acceleration heatmaps (Wüthrich, 2017) with dimension reduction techniques (Gao and Wüthrich, 2018). Verbelen et al. (2018) use Poisson and negative binomial GAMs with compositional predictors to model claim frequency. Ayuso et al. (2019) and Guillén et al. (2019) model claim frequency using standard and zero-inflated Poisson GLMs respectively. So et al. (2020) develop a cost-sensitive multi-class adaptive boosting (AdaBoost) algorithm to predict claim frequency. All aforementioned studies find that the hybrid approach results in the best predictive performance and that predictive models using only telematics information outperform those with only traditional features. This clearly indicates the added value of driving behavior to improve current risk classification practices. Paefgen et al. (2013) find that mileage is most valuable to predict accident risk, even more than all other driving features in their study combined.

Several studies find an increasing non-proportional relationship between distance driven and accident risk, stabilizing for high mileage. Boucher et al. (2013) and Boucher et al. (2017) use Poisson GLMs and GAMs respectively to assess the impact of distance on claim frequency. Paefgen et al. (2014) perform a case-control study with logistic regression to distinguish drivers with and without an accident. Guillén et al. (2019) find a positive relation between the driving distance and the excess zeros in observed claim counts with a zero-inflated Poisson GLM. The stabilization of accident risk for high-mileage drivers might be due to a learning effect after gaining more experience, different driving habits (e.g., less risky roads or time slots) or other safety factors (e.g., newer vehicles). In addition to similar results for claim frequency, Lemaire et al. (2016) find a slight positive linear effect of mileage on claim severity and Ferreira and Minikel (2012) find that the *per mile* pure premium decreases with annual mileage.

Another set of studies puts focus on deriving driving profiles from high-frequency GPS data. Wüthrich (2017) designs so called speed-acceleration heatmaps from GPS data and performs K-means clustering to group similar driver profiles. Ma et al. (2018) use GPS data to calculate driving performance measures, including contextual elements such as traffic speed and volume, to assess the influence on claim occurrence and frequency with logistic and Poisson GLMs respectively. He et al. (2018) use GPS and other sensor data from a vehicle's on-board diagnostics (OBD) unit to compile driver profiles and to measure accident risk.

In this chapter, we analyze a novel dataset on telematics motor insurance which consists of two components. The first data component is a large insurance portfolio followed over the years 2017, 2018 and 2019 with claim history and self-reported risk characteristics of approximately 400,000 policyholders each year. This component is used to develop a baseline a priori tariff which represents the status quo of pricing a (new) policyholder using only a priori directly observable characteristics. The second data component contains information on the driving behavior of young drivers in the portfolio. Policyholders younger than 26 can opt to install a black box in their vehicle in return for a one-time price discount. The recorded driving behavior has no influence on future premiums charged under this contract. Driving habits are registered by measuring the total mileage and the distance driven on different road types and during distinct time slots. Driving style is characterized by recording the number of sudden movement events such as harsh acceleration, braking, cornering and lateral movements.

Our goal is to start from a pricing model with only self-reported characteristics and to develop an updating mechanism that adjusts the proposed baseline price by means of the available telematics information. This approach allows incumbent insurers to incorporate insights on driving behavior into their current in-house pricing expertise. We showcase the added value of telematics via the resulting improvement in risk classification. Furthermore, we put focus on managerial insights by analyzing profits and retention rates under the new telematics paradigm. Denuit et al. (2019a) propose an update mechanism that accounts for driving habits in claim frequency via a multivariate mixed Poisson model, a typical actuarial approach to incorporate a posteriori information in a credibility framework. To the best of our knowledge, this is the first work to explore the full spectrum of pricing (frequency/severity) and driving behavior (habits/style) including a profit and retention analysis. Our updating mechanism results in a true UBI system where the price of insurance coverage is adjusted to the actual vehicle use.

The rest of this chapter is structured as follows. Section 5.2 provides a description of the dataset and outlines our methodology. Section 5.3 details our baseline models for pricing and customer churn prediction. Section 5.4 describes how we update the baseline pricing model with telematics information, highlighting the resulting improvement in risk classification. Section 5.5 investigates the managerial impact of telematics pricing by analyzing profits and retention rates under various price elasticity settings. Section 5.6 concludes this chapter.

5.2 Overview of our data and methodology

We analyze a novel motor third party liability (MTPL) portfolio followed over the years 2017, 2018 and 2019. Figure 5.1 shows a timeline indicating the collection of policy, claim and telematics information. Self-reported risk characteristics are typically known at the start of the policy period, with changes (e.g., replacing the insured vehicle) reported during the policy period. During the course of the year, the insured can surrender the policy and claims can occur. Both policy and claim information are available for the complete portfolio of approximately 400,000 policyholders each year, with 68,196 reported claims in total. Young policyholders have the option to sign up for a telematics black box, registering driving behavior information on mileage, driving habits (by road type and time of day) and driving style (via harsh movements). We aggregate the driving behavior measurements on the yearly policy level, resulting in telematics information for 5,974, 9,383 and 10,481 policyholders in the portfolios observed in 2017, 2018 and 2019 respectively. In total, more than 308 million kilometers are driven by these policyholders. We split the dataset in train (2017 and 2018) and test (2019) data for assessment purposes.

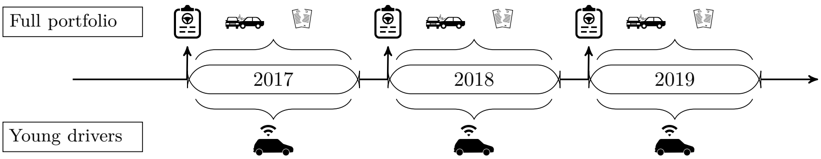


Figure 5.1: Timeline with policy, claim and telematics information over the years 2017 - 2019.

Sections 5.2.1 and 5.2.2 describe the policy and telematics data respectively. Section 5.2.3 investigates the presence of a selection effect and Section 5.2.4 outlines our price updating methodology.

5.2.1 Classic insurance pricing with portfolio data

The pure premium π is the price required to purely cover a policyholder's claim risk. The calculation of this premium is typically split into two components, namely the expected claim frequency F and severity S. Suppose that a policyholder files N claims during a period of exposure-to-risk e for a total amount of L, then $\mathbb{E}(F) = \mathbb{E}(N/e)$ and $\mathbb{E}(S) = \mathbb{E}(L/N \,|\, N > 0)$. Both components are then combined to result in the pure premium as follows: $\pi = \mathbb{E}(F) \times \mathbb{E}(S)$.

Table 5.1 lists the claim and policy information available in the portfolio data. The policy contains self-reported risk characteristics about the driver(s), payment method, geographical location and insured vehicle. Figure 5.2 shows the distribution of claim information in the training portfolios of 2017 and 2018. The left panel shows the exposure-to-risk as the fraction of the year that a policyholder was covered by the policy. A large portion of the policyholders is exposed to the risk of filing a claim during the full year (38.9%), while the others have an exposure between zero and one. An exposure below one occurs when a policyholder starts a policy after the start of the year, surrenders the contract before the end of the year or when one of the self-reported characteristics changes during the year. The middle panel indicates the number of claims filed by a policyholder. Most policyholders do not file a claim with the insurance company (95.6%), some file one claim (4.2%) and the remaining policyholders file two, three, four or five claims. The right panel shows the distribution of the claim amounts up to 10,000 Euro. Claims are typically of a moderate size, with the mean and median amount respectively equal to 4,067 and 1,259 Euro, but extremely large claims occur with the maximum equal to 3,422,728 Euro.

Claims	
`claim_expo`	Fraction of the year that a policyholder is covered by the policy.
`claim_count`	Number of claims reported by a policyholder during the exposure period.
`claim_amount`	Total amount in Euros for all reported claims during the exposure period.
Driver(s)	
`driv_age`	Age of the main driver in years.
`driv_experience`	Years of driving experience.
`driv_seniority`	Years of seniority as a client.
`driv_number`	Number of registered drivers.
`driv_add_younger`	Registered driver younger than the main driver: yes or no.
`driv_add_younger26`	Registered driver younger than the age of 26: yes or no.
Payment method	
`paym_split`	Frequency of payments: annual, biannual, quarterly, monthly or other.
`paym_sepa`	Payment via SEPA (Single Euro Payments Area) bank transfer: yes or no.
Geographical location	
`geo_postcode`	Postal code of the policyholder's residence.
`geo_mosaic`	Customer segment based on demographic and socioeconomic characteristics.
Vehicle	
`veh_age`	Age of the vehicle in years.
`veh_power`	Power of the vehicle in kilowatts.
`veh_weight`	Weight of the vehicle in kilos.
`veh_value`	Value of the vehicle in Euros.
`veh_seats`	Number of seats in the vehicle.
`veh_fuel`	Type of fuel: diesel, petrol, hybrid, gas, electricity or other.
`veh_use`	Type of use: personal (with or without commute), professional or transport.
`veh_type`	Type of vehicle: car, van, mobile home or minibus
`veh_segment`	Vehicle segment, with small urban, medium family, sports and 21 others.
`veh_make`	Vehicle make, with 34 different levels.
`veh_mileage_limit`	Limit on the driving mileage: yes or no.
`veh_garage`	Garage to park the vehicle: yes or no.
`veh_adas`	Vehicle equipped with advanced driver-assistance systems: yes or no.
`veh_trailer`	Trailer insured together with the vehicle: yes or no.

Table 5.1: Description of the claim and policy information in the portfolio data.

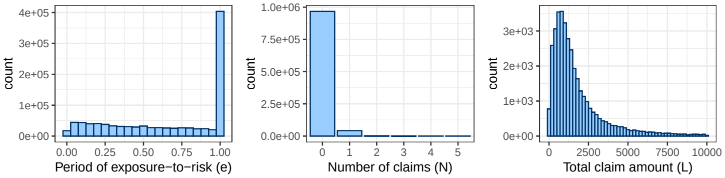


Figure 5.2: Distribution of the exposure period e (left), claim counts N (middle) and amounts L (right).

5.2.2 Telematics data

Driving behavior data is available for a selection of young policyholders in the portfolio. Table 5.2 lists the information recorded by the telematics black box. Driving habits are measured by the total mileage, the proportional distance driven on different road types (abroad, motorway, urban and other) and the proportional distance driven during different time slots (day, rush hour, evening and night). These proportions sum to one and indicate where and when a policyholder usually drives. Verbelen et al. (2018) discuss how to deal with such compositional data from a statistical perspective. Driving style is measured by recording different types of harsh movement events (acceleration, braking, cornering and lateral), which we transform to the number of occurrences per 100 kilometers (km). We also define a measure for the mileage on a yearly basis by scaling the black box registration period to a full year. Imagine a black box that was active for 4 out of 12 months, then the yearly mileage equals three times the recorded distance.

Mileage	*Driving distance in kilometers for each calendar year.*
`distance`	The actual recorded mileage during the year under consideration.
`dist_yrly`	Yearly mileage (in case the black box did not register the full year).
Road type	*Proportion of the total distance driven on different road types.*
`road_abroad`	Roads outside of Belgium.
`road_motorway`	Belgian motorways.
`road_urban`	Belgian urban areas.
`road_other`	Other road types in Belgium.
Time of day	*Proportion of the total distance driven during different time slots.*
`time_day`	Day: 9.30AM - 4PM.
`time_evening`	Evening: 7PM - 10PM.
`time_night`	Night: 10PM - 6AM.
`time_rush`	Rush hours: 6AM - 9.30AM and 4PM - 7PM.
Harsh events	*Number of sudden movement events recorded per 100 kilometers.*
`harsh_accel`	Acceleration: high positive g-force in the direction of travel.
`harsh_brake`	Deceleration: high negative g-force in the direction of travel.
`harsh_latrl`	Lateral: high g-force orthogonal to the direction of travel, e.g., lane shifts.
`harsh_cornr`	Cornering: high g-force in multiple directions.

Table 5.2: Description of the available telematics data.

Figure 5.3 details the distribution of the telematics features in the training data. The top panels show the recorded (left) and yearly (right) distance driven. The rightwards shift indicates how most low mileage recordings are due to inactive black boxes and we observe an average yearly mileage of 16,502 kilometers. The middle left panel indicates that a large proportion of kilometers is driven in urban areas, followed by other roads and motorways. Abroad driving accounts for a small part of the distance driven. The middle right panel shows that daytime and rush hour driving are frequent, with less kilometers driven during the evening and at night. Gray lines emphasize the composition nature of the data for 100 random drivers. The bottom panels indicate the number of harsh movement events recorded per 100 kilometer driven. Harsh cornering occurs most often (35.5 events/100km on average), followed by braking (8.7), acceleration (3.3) and lateral movements (0.9).

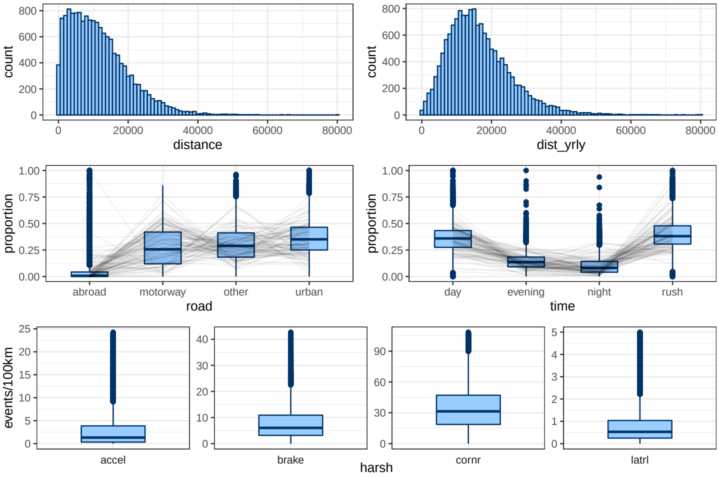


Figure 5.3: Distribution of the actual distance (top left), yearly distance (top right), road types (middle left), times of day (middle right) and harsh movement events (bottom) in the training data.

5.2.3 Selection effect

In our portfolio, the installation of a black box to record driving behavior is a choice offered to young drivers only. Figure 5.4 shows the age distribution for policyholders who have a black box installed (green) and those who do not have a black box (red). The left panel displays the full portfolio and indicates that only young policyholders have the option to sign up for the telematics device. The right panel zooms in on policyholders aged younger than 26 at underwriting time. For the ages 18 up to 22 there is a higher number of drivers with a black box, while the situation is reversed for the ages 23 up to 27. In total, around 42% of the young policyholders opted for the telematics device. We therefore focus our analysis of a possible selection effect on young policyholders with the telematics option (< 26 years at underwriting).

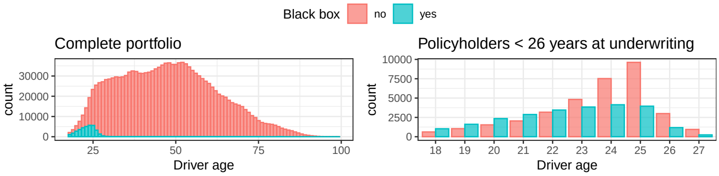


Figure 5.4: Age distribution for policyholders with/without (green/red) a black box.

We use the two-sample Poisson test of Fay (2010) to compare the observed claim risk for a control group of young policyholders without a black box (μ_{no}) and a test group with a box (μ_{yes}). For each group we calculate $\hat{\mu} = \sum_i N_i / \sum_i e_i$ in Table 5.3 and test the null hypothesis $H_0 : \hat{\mu}_{\text{yes}} = \hat{\mu}_{\text{no}}$ or equivalently $H_0 : \hat{\mu}_{\text{no}} / \hat{\mu}_{\text{yes}} = 1$. The p-value equals 0.315 such that we do not reject the null H_0. The observed value of $\hat{\mu}_{\text{no}} / \hat{\mu}_{\text{yes}} = 0.965$ with a 95% confidence interval of $[0.900, 1.034]$.

Black box	$\sum_i N_i$	$\sum_i e_i$	$\hat{\mu}$
No	1,817	17,984.03	0.1010
Yes	1,477	14,104.14	0.1047

Table 5.3: Claim risk statistics for young policyholders without/with a black box.

The empirical observation $\hat{\mu}_{\text{yes}} > \hat{\mu}_{\text{no}}$ might sound surprising. However, the right panel of Figure 5.4 indicates that policyholders without a black box are older on average in our sample. Older drivers are typically less risky compared to younger ones. We therefore further investigate the presence of a selection effect by fitting the following Poisson GLM, investigating the effect of having a black box via the dummy variable `bbox` while controlling for the driver's age `driv_age`:

$$\ln[\mathbb{E}(N)] = \ln[e] + \beta_0 + \beta_{\text{age}}\texttt{driv\_age} + \beta_{\text{box}}\texttt{bbox} + \beta_{\text{int}}\texttt{driv\_age} : \texttt{bbox}. \quad (5.1)$$

Table 5.4 shows the results with (left) and without (right) the interaction term included. The black box coefficient β_{box} is negative in both GLMs, indicating a lower claim risk for policyholders with the box. Since $\exp(-0.054) = 0.95$, having a box installed decreases claim risk with 5%. However, the effect is not statistically significant according to the p-values in both GLMs. The fitted interaction term reveals that the age effect decreases less steep for policyholders with a black box, but this is also not significant. Figure 5.5 shows the fitted GLM effects (lines), 95% confidence intervals (shades) and the empirical claim frequencies (points) by group (color).

	With interaction term				Without interaction term			
	Coefficient β	Std. error	z-value	p-value	Coefficient β	Std. error	z-value	p-value
`intercept`	-0.452	0.268	-1.68	0.09	-0.536	0.193	-2.78	0.006
`driv_age`	-0.078	0.011	-6.85	$7e^{-12}$	-0.074	0.008	-9.11	$< 2e^{-16}$
`bbox`	-0.222	0.374	-0.59	0.55	-0.054	0.037	-1.48	0.140
`driv_age:bbox`	0.007	0.016	0.45	0.65	-	-	-	-

Table 5.4: Summary of the selection effect in a GLM with (left) and without (right) the interaction term.

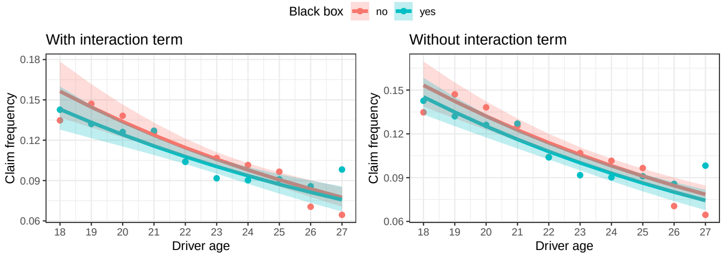


Figure 5.5: Age effect for young drivers with/without a box (green/red) and the interaction (left/right).

These findings point to the absence of a significant selection effect. This could be due to the fact that signing up for the telematics device is not coupled to future premium changes. Furthermore, young policyholders might be persuaded by their parents to install the black box.

5.2.4 A methodology to update pricing

Figure 5.6 outlines our updating mechanism proposal to include telematics information into a pricing structure that already uses self-reported policy characteristics. We take position at time t and consider yearly policy periods, as is customary in motor insurance, but this scheme is applicable to any policy duration (e.g., quarters or months). For now, we simply denote claim, policy and telematics features by $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively. First, a baseline pricing model $\pi(\boldsymbol{x})$ is developed for the complete portfolio using policy and claim information recorded in period $[t-1, t]$. Next, this premium is updated for policyholders with a back box using telematics and claim information in period $[t-1, t]$. These updates are modeled as a multiplicative adjustment $\delta^{\pi}(\boldsymbol{z})$ to the baseline such that the updated price follows as: $\pi^*(\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{x}) \times \delta^{\pi}(\boldsymbol{z})$.

We propose to implement a commercial UBI product where the premium for coverage in $[t, t+1]$ is paid at two different moments in time. The baseline

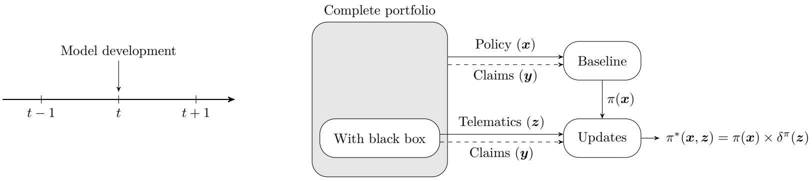

Model development
$t-1$
t
$t+1$
Complete portfolio
With black box
Policy ($\boldsymbol{x}$)
Claims ($\boldsymbol{y}$)
Baseline
$\pi(\boldsymbol{x})$
Telematics ($\boldsymbol{z}$)
Claims ($\boldsymbol{y}$)
Updates
$\pi^*(\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{x}) \times \delta^\pi(\boldsymbol{z})$

Figure 5.6: Methodology of our mechanism to update baseline premiums with telematics information.

premium $\pi(\boldsymbol{x})$ is paid at time t based on the actual policy characteristics registered at that time. The ex post update $\delta^\pi(\boldsymbol{z})$ is calculated at time $t+1$ based on the driving behavior in period $[t, t+1]$. Clients have the opportunity to directly influence their insurance premium and earn a rebate with good driving if $\delta^\pi(\boldsymbol{z}) < 1$. Risky behavior is discouraged as bad driving results in a price penalty via $\delta^\pi(\boldsymbol{z}) > 1$. The insurer still receives the base premium a time t to cover claims and other costs during period $[t, t+1]$.

5.3 Baseline pricing and churn models

We first put focus on developing a baseline insurance pricing model for the complete portfolio using the self-reported policy data from Table 5.1. This represents the status quo for incumbent insurance companies who are thinking about incorporating telematics into their pricing strategies. We also develop a baseline model to predict the churn (or: lapse) behavior of customers, defining the churn rate ρ as the probability that a policyholder surrenders the policy. Suppose that a binary indicator $C \in \{0, 1\}$ equals one for policyholders who lapse their contract during the year, then $\rho = \mathbb{E}(C)$. We therefore develop a predictive model for the claim frequency F, severity S and churn probability ρ with the risk characteristics listed in Table 5.1 as features $\boldsymbol{x}$. We opt for stochastic gradient boosting machines or GBMs (Friedman, 2002) to determine the prediction function, based on the good GBM performance in insurance pricing (see Chapter 3) and churn applications (Spedicato et al., 2018).

Section 5.3.1 details the GBM development process. Section 5.3.2 proposes a slight adjustment that restores the balance between observed and predicted targets. Section 5.3.3 provides insights into the optimal GBMs. The frequency and severity GBMs are used in Section 5.4 as a baseline pricing model, while the churn GBM is used in Section 5.5 as baseline for retention rates.

5.3.1 GBM training process

Given features $\boldsymbol{x}$ and a target y, our goal is to train a GBM to accurately predict $\hat{y} = f(\boldsymbol{x})$. We model integer-valued count data for claim frequency, skewed long-tailed data for claim severity and binary 0/1-valued data for customer churn. Table 5.5 summarizes our distributional assumptions and the accompanying deviance loss functions used in the GBM training process. The exposure-to-risk e is taken into account via an offset term in the frequency model to obtain expected claim frequencies proportional to the duration of the policy contract. Furthermore, the number of claims N is used as a weight in the claim severity model. We train our GBMs via the R interface to H2O: an open source machine learning (ML) platform (LeDell et al., 2020). Many parameters are available to tune the performance of GBMs, see Click et al. (2021) for a complete list. The selected parameters listed in Table 5.6 are obtained via a random grid search and 5-fold cross-validation on the combined training portfolios of 2017 and 2018. It is interesting to note that the optimal tree depths for claim frequency and severity are consistent with the results from Table 3.3 in Section 3.4.2 of Chapter 3, especially because completely different datasets are analyzed in both case studies.

	Distribution	Prediction $f(\boldsymbol{x})$	Loss function $D(y, f(\boldsymbol{x}))$
Claim frequency	$N \sim$ Poisson	$\mathbb{E}(N \mid \boldsymbol{x}, e)$	$\frac{2}{n}\sum_{i=1}^{n}\left[y_i \ln\left\{\frac{y_i}{f(\boldsymbol{x}_i)}\right\} - \{y_i - f(\boldsymbol{x}_i)\}\right]$
Claim severity	$L/N \sim$ gamma	$\mathbb{E}(L/N \mid \boldsymbol{x})$	$\frac{2}{\sum_i N_i}\sum_{i=1}^{n} N_i\left[\frac{y_i - f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} - \ln\left\{\frac{y_i}{f(\boldsymbol{x}_i)}\right\}\right]$
Customer churn	$C \sim$ Bernoulli	$\mathbb{E}(C \mid \boldsymbol{x})$	$-\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln\{f(\boldsymbol{x}_i)\} + (1 - y_i)\ln\{f(\boldsymbol{x}_i)\}\right]$

Table 5.5: Distributional assumptions for claim frequency, severity and client churn.

	ntrees	learn_rate	max_depth	sample_rate	col_sample_rate
Claim frequency	4,700	0.02	4	1.0	0.6
Claim severity	3,900	0.01	1	0.5	0.7
Customer churn	4,100	0.02	5	0.7	0.6

Table 5.6: Optimal settings of the GBM tuning parameters for the different models.

5.3.2 Balance property

The maximum likelihood estimator (MLE) in a GLM framework (with canonical link) leads to $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$ (Wüthrich, 2020, Corollary 2.4). This is known as the balance property and implies that the sum of predicted targets $\hat{y}_i$ equals the sum of the observed targets y_i for $i \in 1, \ldots, n$ in the training data.

This unbiasedness is very important for insurance pricing as we need to cover total losses at the portfolio level. GBMs, as most predictive models, focus purely on accurate individual predictions. We therefore enforce the balance property in our portfolio of young drivers by scaling the frequency and severity GBM predictions from $\hat{y}_i$ to $\hat{y}_i^b$. Table 5.7 shows the (possibly) biased ratio $\sum \hat{y}_i / \sum y_i$ and the balanced ratio $\sum \hat{y}_i^b / \sum y_i$ for claim frequency F, severity S and the resulting premium $\pi = \mathbb{E}(F) \times \mathbb{E}(S)$. On the train data we observe an underestimation of total claim frequency (0.3%) and severity (5.4%), leading to an underestimation of the premium inflow to cover losses. Scaling the predictions with aforementioned percentages leads to perfect balance for frequency and severity, while total losses are now covered by the premium inflow. On the test data we observe an over/underestimation for frequency/severity respectively. Perfect balance for these components is not achieved as the scaling is based on the train data. However, both components offset each other, resulting in a premium inflow that covers total losses on the test data as well.

	Claim frequency F		Claim severity S		Premium π	
	biased	balanced	biased	balanced	biased	balanced
Train	0.997	1.000	0.946	1.000	0.948	1.004
Test	1.045	1.048	0.907	0.958	0.961	1.019

Table 5.7: Biased and balanced ratios for the frequency, severity and premium.

5.3.3 Insights in the optimal GBMs

Table 5.8 lists the ten most important features in each GBM. Postal code and driving experience are most important to predict claim frequency, while vehicle characteristics (e.g., the weight, make and segment) are most informative to predict severity. The various ways of paying premiums learns us a lot about the churn behavior of customers. The top ten features carry around 90% (or even more) of the total information contained in the collection of 24 features.

Figure 5.7 shows partial dependence (PD) effects (Friedman, 2001) for the highlighted features in Table 5.8. Claim frequency decreases as the driver gains more experience behind the wheel (top left panel). This decrease is rather steep in the first 10 years, emphasizing the high claim risk of young, inexperienced drivers. The effect becomes stable after 30 years, with a slight increase for senior policyholders. The top right panel shows the frequency PD for each postal code area in Belgium. Claim risk is highest in densely populated cities (e.g., the capital Brussels in the center) and lowest in spacious rural areas (e.g., the Ardennes in the south-east). Claim severity increases with the vehicle's weight (middle left panel). This is likely due to the fact that heavier cars cause more

	Claim frequency		Claim severity		Customer churn	
Rank	Feature	%	Feature	%	Feature	%
1	geo_postcode	34.72	veh_weight	23.21	paym_split	43.48
2	driv_experience	14.08	veh_make	21.37	geo_postcode	11.67
3	driv_seniority	8.52	geo_postcode	10.54	veh_age	9.85
4	veh_make	6.25	veh_segment	10.48	paym_sepa	9.44
5	geo_mosaic	5.85	geo_mosaic	6.59	driv_seniority	6.90
6	veh_fuel	5.09	driv_seniority	5.83	veh_make	3.43
7	veh_segment	4.66	veh_value	3.50	driv_experience	2.85
8	paym_split	3.91	veh_age	3.44	geo_mosaic	2.45
9	driv_add_younger26	3.29	driv_experience	2.98	driv_age	2.43
10	driv_age	2.75	driv_add_younger26	2.91	veh_use	1.99
$\sum$		89.12		90.86		94.48

Table 5.8: Top ten most important features in the different GBMs.

damage to other cars in an accident. Some of the more expensive brands (e.g., BMW, Porsche, Mercedes and Jaguar) lead to higher severities, maybe due to a more sturdy build compared to cheaper cars. The churn probability increases with the payment frequency (middle right panel) and is higher for policyholders not paying via a SEPA transfer (bottom right panel). Policyholders who pay an annual premium might be quite loyal and convinced to stay with the company, while monthly payments may indicate that someone is browsing for better offers elsewhere in the meantime. SEPA transfers are often automatically credited from an account. Policyholders who prefer to actively pay the invoice might not be ready to enter a long-term commitment with the company and prefer to be able to switch insurance swiftly.

5.4 Towards a usage-based pricing mechanism

Our goal is to update the baseline pricing structure, consisting of the combined frequency and severity GBMs developed in Section 5.3, by using the driving behavior of policyholders with a telematics box. For this selection of drivers we have access to claim targets y, a baseline prediction $f(\boldsymbol{x})$ based on self-reported policy characteristics $\boldsymbol{x}$ from Table 5.1 and telematics information $\boldsymbol{z}$ from Table 5.2. Explainability of the updating mechanism is a key requirement, as the resulting price adjustments should be comprehensible and easy to communicate to all stakeholders (e.g., regulators, managers and clients). We therefore opt to use generalized linear models or GLMs (Nelder and Wedderburn, 1972). Such a GLM leads to an interpretable model structure and is applicable to targets following any distribution of the exponential family (e.g., Bernoulli, Poisson and gamma). The general formulation of a log-link GLM with $\ln[f(\boldsymbol{x})]$ as an offset

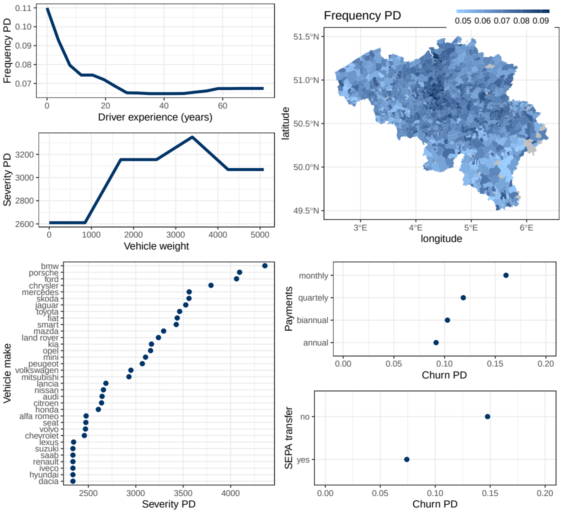


Figure 5.7: PD effect for driving experience (top left) and postal code (top right) in the claim frequency GBM, the vehicle's make (bottom left) and weight (middle left) in the severity GBM and the payment frequency (middle right) and SEPA indicator (bottom right) in the churn GBM.

(i.e., term with a coefficient fixed to one) in the linear predictor is as follows:

$$
\begin{aligned}
\ln[\mathbb{E}(y \mid \boldsymbol{x}, \boldsymbol{z})] &= \ln[f(\boldsymbol{x})] + \beta_0 + \sum_{j=1}^{p} \beta_j z_j \\
\mathbb{E}(y \mid \boldsymbol{x}, \boldsymbol{z}) &= f(\boldsymbol{x}) \times \exp(\beta_0) \times \prod_{j=1}^{p} \exp(\beta_j z_j)
\end{aligned}
\tag{5.2}
$$

with β_0 the intercept and β_j the coefficient for telematics feature z_j with $j \in \{1, \ldots, p\}$. Recall from Table 5.5 that the target y represents N and L/N, while $f(\boldsymbol{x})$ equals $\mathbb{E}(N \mid \boldsymbol{x}, e)$ and $\mathbb{E}(L/N \mid \boldsymbol{x})$ for the frequency and severity GBM respectively. Figure 5.8 visualizes our update methodology, applied to the claim frequency (left) and severity (right) components.

Our proposed updating mechanism in Equation (5.2) allows for intuitive price effects since the final prediction is multiplicative in three contributions:

- the baseline GBM prediction $f(\boldsymbol{x})$ for a policyholder with risk characteristics $\boldsymbol{x}$,
- an overall update factor $\exp(\beta_0)$ via the intercept and
- an update $\exp(\beta_j z_j)$ from each individual telematics feature z_j.

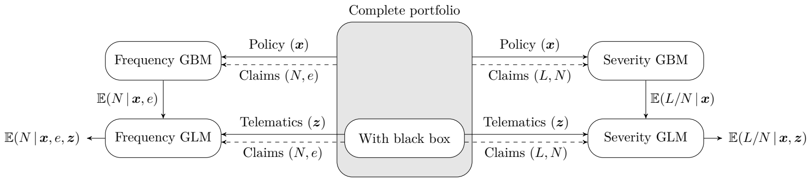

Complete portfolio
Frequency GBM
Policy ($\boldsymbol{x}$)
Claims (N, e)
Policy ($\boldsymbol{x}$)
Severity GBM
Claims (L, N)
$\mathbb{E}(N \mid \boldsymbol{x}, e)$
$\mathbb{E}(L/N \mid \boldsymbol{x})$
$\mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z})$
Frequency GLM
Telematics ($\boldsymbol{z}$)
Claims (N, e)
With black box
Telematics ($\boldsymbol{z}$)
Claims (L, N)
Severity GLM
$\mathbb{E}(L/N \mid \boldsymbol{x}, \boldsymbol{z})$

Figure 5.8: Methodology of our mechanism to update baseline premium components.

The updated GLM predictions satisfy the balance property, as described in Section 5.3.2, while we deliberately enforce this property in the baseline GBMs. This implies that the multiplicative adjustments result in a pure redistribution of risk in the claim frequency and severity models.

We perform feature selection to unravel the effect of driving behavior on claim risk in Section 5.4.1. Focusing on informative features, we develop our explainable updating mechanism in Section 5.4.2. Finally, we highlight the added value of telematics for risk classification in Section 5.4.3.

5.4.1 Finding the most important telematics features

We search for a small collection of highly informative telematics features $\boldsymbol{z}$ to render the update mechanism simple, yet powerful. The complete set of possible features includes those listed in Table 5.2, supplemented with all possible two-way interactions. We apply the Least Absolute Shrinkage and Selection Operator or LASSO (Tibshirani, 1996) to perform feature selection. LASSO shrinks model coefficients β_j to zero by applying a regularization penalty $\lambda \, ||\boldsymbol{\beta}||_1$ in the maximum likelihood estimation (MLE) of the GLM in Equation (5.2). Only highly informative features z_j with non-zero coefficients β_j remain in the GLM, leading to a sparse structure. The degree of sparseness depends on the value of λ, with higher values leading to more sparsity. All telematics features are continuous but with various scales, so we standardize each z_j before applying LASSO. We fit a frequency and severity GLM with the structure of Equation (5.2) and the distributional assumptions outlined in Table 5.5. The following steps are performed 100 times:

1. sample 50% of the train data and divide the sample in five equally sized sets,
2. standardize the features $\boldsymbol{z}$ by subtracting the mean and dividing by the standard deviation,

3. fit 5 GLMs, each time omitting one data set, for each value of λ in a predetermined grid,
4. find the value of λ that minimizes the 5-fold cross-validation error $D(y, f(\boldsymbol{x}, \boldsymbol{z}))$,
5. register the features z_j with non-zero coefficients β_j in the GLM fit with optimal λ value.

Repeating the LASSO procedure for multiple data samples allows to discover features which are selected consistently, no matter which part of the data is used. We can therefore assume that those features are most informative and reliable to update our baseline predictions. Figure 5.9 shows the selection proportions based on 100 LASSO experiments for the 20 most informative features. A red/green color indicates a negative/positive β coefficient if selected. The left panel shows four dominant telematics features to update claim frequency, namely `dist_yrly` (100), `harsh_latrl` (99), `harsh_brake` (94) and `time_night` (90), all with unanimous positive coefficients across all simulations. We decide to keep these four features as the next feature is selected only 72/100 times. The right panel indicates that none of the telematics features carries much information to update claim severity. The most popular feature is selected in only 42% of the simulations. Telematics features do not seem to be important for predicting claim severity and we therefore decide to incorporate telematics information in the pure premium solely via the claim frequency component. It is interesting to note that the LASSO procedure on the full training data, in combination with the “one standard error rule” (Hastie et al., 2009), leads to the same feature selection results for both frequency and severity.

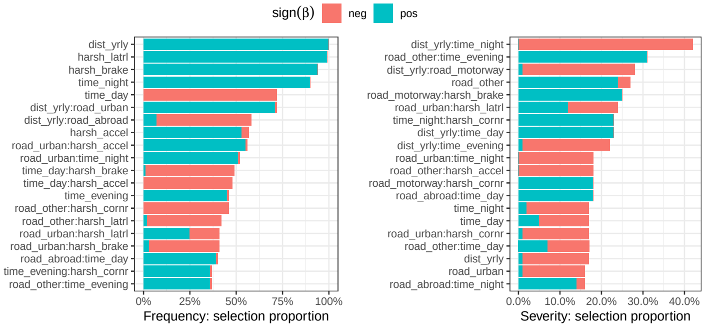


Figure 5.9: Feature selection proportions in the LASSO GLM for claim frequency (left) and severity (right), where red/green indicates a negative/positive β coefficient if selected.

5.4.2 An explainable updating mechanism

Let $\boldsymbol{z}^* \in \mathbb{R}^4$ represent the features `dist_yrly`, `harsh_latrl`, `harsh_brake` and `time_night`. We propose an updating mechanism based on the following Poisson GLM for claim frequency:

$$\begin{aligned} \ln[\mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z}^*)] &= \ln[\mathbb{E}(N \mid \boldsymbol{x}, e)] + \beta_0 + \sum_{j=1}^{4} \beta_j \log(z_j^* + 1) \\ \mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z}^*) &= \mathbb{E}(N \mid \boldsymbol{x}, e) \times \exp(\beta_0) \times \prod_{j=1}^{4} (z_j^* + 1)^{\beta_j}. \end{aligned} \tag{5.3}$$

The updated prediction $\mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z}^*)$ takes self-reported policy characteristics into account via the baseline prediction $\mathbb{E}(N \mid \boldsymbol{x}, e)$. This baseline is multiplied by one fixed term $\exp(\beta_0)$ and four terms that depend on the recorded driving behavior, one for each telematics feature z_j^*. We model the telematics features as $\beta_j \log(z_j^* + 1)$, which is basically the Yeo–Johnson transformation of power zero for non-negative values (Yeo and Johnson, 2000). This choice is based on two reasons: 1) to stabilize the data distributions shown in Figure 5.3 and 2) to obtain an intuitive updating formula where each telematics feature has an effect of the form $(z_j^* + 1)^{\beta_j}$. These terms all equal one when the telematics features equal zero, implying that the update to the baseline is completely determined by $\exp(\beta_0)$ for a policyholder who did not drive at all.

We obtain $\exp(\beta_0) \approx 0.02$ after fitting the GLM from Equation (5.3) to the drivers with telematics. This indicates that policyholders who did not drive during the entire year receive a 98% rebate of their baseline premium. The small fee of 2% can be seen as a fixed subscription payment and is justified by the administrative costs needed to maintain the policy during the full year. Furthermore, the policyholder was covered for the entire policy period and had the freedom to drive on public roads without worrying about insurance. Figure 5.10 shows the multiplicative update effect for each telematics feature, namely $(z_j^* + 1)^{\beta_j}$. We anonymized the y-axis for confidentiality reasons, but every panel contains a horizontal dashed line at the value one. The top left panel shows the non-proportional increase for mileage with the fixed discount already included, namely $\exp(\beta_0) \times (\texttt{dist\_yrly} + 1)^{\beta_{\texttt{dist\_yrly}}}$. Low-mileage drivers receive large discounts and the combined update even remains below one for high-mileage drivers. The top right panel shows an almost linear increase for night-time driving and the bottom left/right panels show non-proportional increases for harsh braking/lateral events. These three components focus on driving safety and the associated updates are always above one. This increases the total update once night driving, harsh braking or lateral events are registered.

Safe driving during the day is therefore the key to earn discounts, with less driving resulting in bigger discounts.

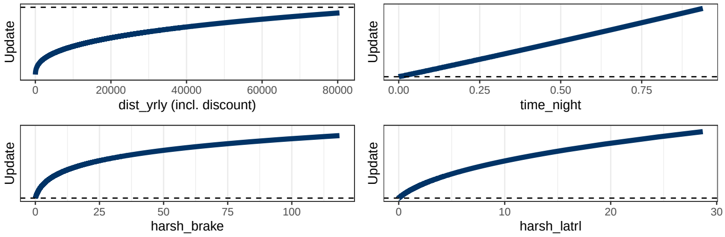


Figure 5.10: Multiplicative update effects for the mileage including the fixed discount (top left), night-time driving (top right), harsh braking (bottom left) and lateral movements (bottom right).

Figure 5.11 shows the distribution of scores $\beta_j \log(z_j^* + 1)$ and updates $(z_j^* + 1)^{\beta_j}$ for policyholders in the train data. The y-axis is again anonymized and a horizontal dashed line represents the value zero/one in the left/right panel. Total scores/updates are additive/multiplicative in the different components, as shown in Equation (5.3). The update for mileage, with fixed discount included, remains below one for every policyholder. An average mileage driver without unsafe events receives a discount of around 50%. The three other telematics components result in updates above one due to their risky nature, thereby increasing the total update. Average night-time driving, harsh braking and lateral movements results in penalties of approximately 10%, 35% and 20%. Total updates range from around 95% discounts to more than 300% penalties, with a 5% discount on average. Around 60% of the drivers are receiving a discount on the baseline premium with our updating mechanism. In Section 5.5 we discuss how to transform this technical analysis into a commercial UBI product with update limits on discounts and penalties.

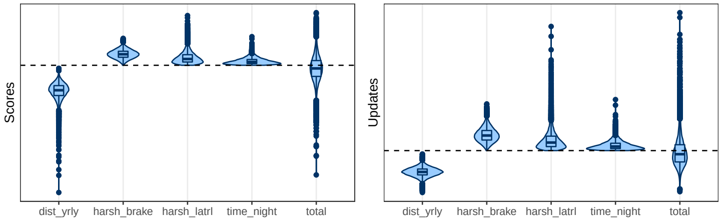


Figure 5.11: Distribution of scores $\beta_j \log(z_j^* + 1)$ (left) and updates $(z_j^* + 1)^{\beta_j}$ (right) in the train data.

Figure 5.12 shows an intuitive dashboard to inform policyholders on their driving behavior and related price effects. The top left panel shows the driving information recorded in 2017 for a random policyholder. The top right panel compares this behavior relative to the full portfolio: a low/high decile indicates better/worse driving behavior. This profile shows an above average number of lateral movement events, but scores well regarding braking, night-time driving and especially mileage. The bottom panel shows the additive score for each component. Low mileage driving (green) results in a big discount, while the other three components (red) decrease the discount. This driver obtains a total discount (blue) of around 35% on the baseline premium.

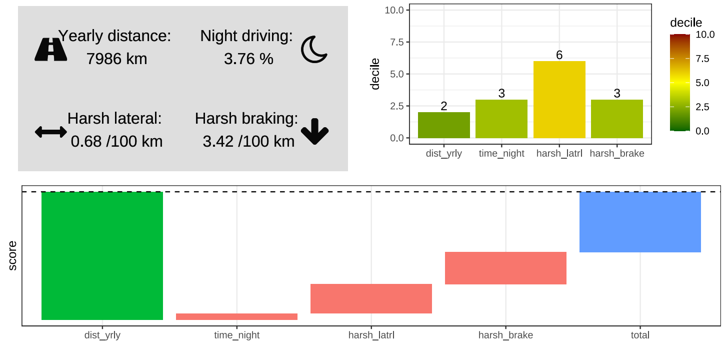


Figure 5.12: Dashboard with recorded driving information (top left), ranking within the portfolio (top right) and influence of each component on the final price (bottom).

5.4.3 The added value of telematics for risk classification

We aim to quantify the value of our telematics updating mechanism. Here the focus lies on predictive performance gains and risk classification improvements by updating the claim frequency component. Section 5.5 analyzes the effects on an insurer's profits and retention rates.

Table 5.9 shows the Poisson deviance values for the GBM baseline and GLM update predictions. The updates result in a relative deviance improvement of 2.58% and 1.50% on the train and test data respectively. This shows that our simple updating mechanism with telematics information is able to improve the predictive performance of an elaborate GBM. We also show the relative improvements when only one telematics feature z_j^* is used to fit the update GLM in Equation (5.3). The mileage and harsh movements show the highest deviance improvements. It is interesting to note how similar the gains in train and test data are for the mileage-only GLM. Mileage might therefore be considered as the most general and consistent indicator of claim risk in our data.

	Poisson deviance (absolute values)		Relative improvement from GBM baseline to GLM update (%)				
	GBM baseline	GLM update	Total	`dist_yrly`	`time_night`	`harhs_brake`	`harsh_latrl`
Train	0.4044	0.3939	2.581	0.905	0.659	0.848	1.154
Test	0.3927	0.3868	1.495	0.881	0.218	0.285	0.305

Table 5.9: Poisson deviance for the baseline GBM and update GLM on the train and test data.

We define a risk score for policyholder i in model m as $r_i^m = F_n\{f^m(\boldsymbol{x}_i, \boldsymbol{z}_i^*)\}$, namely the empirical cumulative distribution function of the predicted claim frequency for policyholder i in model m. Note that $r_i^m \in [0, 1]$ with low/high values for policyholders with a low/high prediction in model m. We visualize improvements in claim risk classification with a Lorenz curve, a tool developed to represent wealth distribution inequalities in welfare economics (Lorenz, 1905):

$$LC^m(s) = \frac{\sum_{i=1}^n N_i \, \mathbb{1}\{r_i^m \leq s\}}{\sum_{i=1}^n N_i} \;\text{ for }\; s \in [0, 1].$$

The Lorenz curve accumulates observed claims from low to high risks as perceived by model m (i.e., $r_i^m : 0 \rightarrow 1$). Better risk classification means that claims accumulate at a slower/faster rate for low/high values of r_i^m. Figure 5.13 shows the Lorenz curves for the GBM baseline (red) and GLM update (green) on both the train (left) and test (right) data. We observe that, in both the train and test data, the green line is shifted further to the bottom right than the red line, indicating the improved risk classification with telematics updates. To quantify this improvement we use the Gini index, defined as two times the area between a Lorenz curve and the 45 degree line of equality (Gini, 1912). We obtain a Gini improvement of 19.6% (going from 0.275 to 0.329) and 52.5% (going from 0.136 to 0.207) for the train and test data respectively.

We now group policyholders in five equally sized bins based on the risk scores r_i^m and calculate the observed claim proportions in each bin as follows:

$$PC^m(s) = \frac{\sum_{i=1}^n N_i \, \mathbb{1}\{\frac{s-1}{5} < r_i^m \leq \frac{s}{5}\}}{\sum_{i=1}^n N_i} \;\text{ for }\; s \in \{1, \ldots, 5\}$$

Figure 5.14 shows the proportional claims for the GBM baseline (red) and GLM update (green) on both the train (left) and test (right) data. Both models show an increasing trend in claim proportions thanks to risk classification. However, the green bars are lower/higher compared to the red ones for low/high risk bins, indicating a better risk classification of the update GLM. To quantify

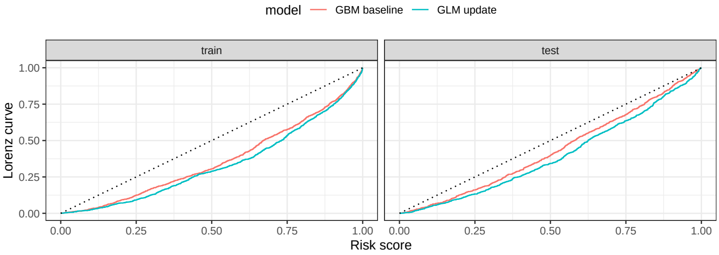


Figure 5.13: Lorenz curves for the GBM (red) and GLM (green) on the train (left) and test (right) data.

the improvement we calculate the slopes of a linear fit to the proportions. We obtain a slope increase of 18.9% (0.064 to 0.076) and 61.7% (0.031 to 0.050) for the train and test data.

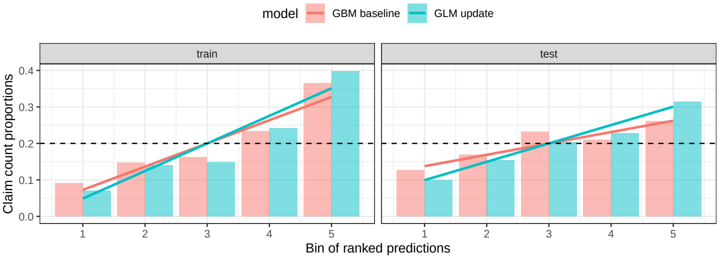


Figure 5.14: Claim bins for the GBM (red) and GLM (green) on the train (left) and test (right) data.

It does not come as a surprise that extra features carry useful information to improve predictive performance and risk classification. The gains are however of a considerable size, even higher on the test compared to train data. This hints that driving behavior is a better measure to extrapolate past claim behavior to the future compared to the self-reported risk characteristics.

5.5 Managerial insights on telematic updates

We now turn to a managerial view on the value of telematics for insurance pricing by analyzing the resulting monetary profits and client retention rates. The GBMs from Section 5.3 result in a baseline price $\pi(\boldsymbol{x})$ and churn probability $\rho(\boldsymbol{x})$ for a policyholder with self-reported risk characteristics $\boldsymbol{x}$ at time t. The GLM from Section 5.4.2 proposes multiplicative premium updates $\delta^{\pi}(\boldsymbol{z})$ based on telematics information $\boldsymbol{z}$ gathered over the period $[t, t+1]$. This results in an updated price $\pi^*(\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{x}) \times \delta^{\pi}(\boldsymbol{z})$, taking the form of a rebate or penalty at time $t+1$. The churn behavior of clients is likely to depend on these price changes, implying a transformation of the baseline churn probability $\rho(\boldsymbol{x})$ to $\rho^*(\boldsymbol{x}, \delta^{\pi})$ over the period $[t, t+1]$. We hereby assume that policyholders can track their driving behavior and the price implications in a dashboard application, directly influencing their churn behavior. Section 5.5.1 details our assumptions regarding changes in the churn probability following price updates via the price elasticity of demand. Section 5.5.2 shows the effect on profits and retention rates in a stylized example with a fair redistribution constraint. This constraint intends to allow for a fair comparison between the baseline and telematics situation, while combating extremely high (and low) premium changes. In Section 5.5.3 we optimize the product design for maximal profits under retention constraints and for maximum retention under profitability constraints.

5.5.1 Price elasticity of demand

We aim to analyze an insurer's profits and retention rates under the new telematics pricing structure. The price elasticity of demand ϵ_p measures how sensitive the demand of a quantity q is to changes in its price π as follows: $\epsilon_p = \frac{\Delta q / q}{\Delta \pi / \pi}$, with $\Delta q / q$ and $\Delta \pi / \pi$ the percentage change in quantity and price respectively. For the vast majority of goods and services, the "law of demand" dictates that the quantity decreases for increasing prices, leading to a negative price elasticity (Gillespie, 2014). We assume insurance follows this law, especially in a highly competitive segment such as motor insurance. Within economics it is customary to drop the minus sign and report on absolute values of ϵ_p, with demand being referred to as elastic when $\epsilon_p > 1$ and inelastic when $\epsilon_p < 1$ (Browning and Zupan, 2020).

Our dataset does not allow to estimate the portfolio's observed price elasticity, as we do not have information on price quotes and the insured's acceptance/decline decision. We therefore develop assumptions based upon relevant empirical research on demand elasticity within motor insurance. Sherden (1984) analyzes elasticity over a range of prices for different types of coverage. He shows that bodily injury covers are rather inelastic over the full price range, i.e. $\epsilon_p < 1$, while collision becomes elastic for prices equal to 1.6 times the average with ϵ_p approaching three for high prices. Barone and Bella (2004) compute the price elasticity for 989 customer segments and find most values ranging from 0.4 (inelastic) to 2.2 (elastic). Guelman and Guillén (2014) find an approximate linear relation between lapse rates and price changes. However, the resulting price elasticity ϵ_p (i.e., the slope) differs per customer segment and they obtain a slightly higher elasticity for price increases compared to price decreases.

Let δ^ρ represent an additive change in a customer's churn probability as follows: $\rho^* = \rho + \delta^\rho$. We assume a linear relationship between the change in churn probability δ^ρ, the price update δ^π and the elasticity ϵ_p as follows: $\delta^\rho = \epsilon_p \cdot (\delta^\pi - 1)$. This leads to the following churn probability, forced to be bounded in the interval $[0, 1]$: $\rho^*(\boldsymbol{x}, \delta^\pi) = \rho(\boldsymbol{x}) + \epsilon_p \cdot (\delta^\pi - 1)$. Figure 5.15 illustrates this relation for a policyholder with a baseline churn probability $\rho(\boldsymbol{x}) = 10\%$ and a price elasticity $\epsilon_p \in [0, 5]$. Notice how $\delta^\rho = 0$ when there is no price change, i.e., when $\delta^\pi = \pi^*/\pi = 1$. The churn probability increases or decreases linearly when $\delta^\pi > 1$ or $\delta^\pi < 1$ respectively, with a slope equal to the price elasticity ϵ_p. Following the aforementioned empirical research, we opt for $\epsilon_p \in [0, 5]$ to cover all examples of realistic motor insurance markets. Our assumption proposes a fixed elasticity for the complete portfolio without taking customer segmentation into account. We believe that this simplification is justifiable as our telematics portfolio of only young drivers is already more homogeneous compared to the complete portfolio with all policyholders. Furthermore, this allows us to focus on the effect of telematics pricing updates on the retention rates and profits.

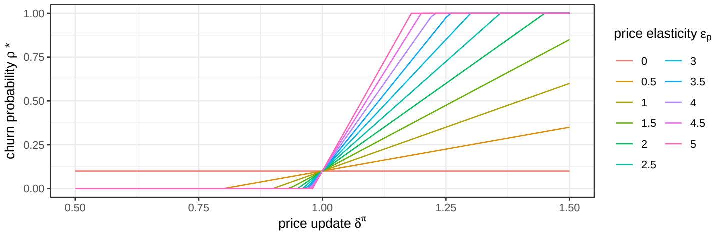


Figure 5.15: Effect of price updates δ^π on the churn probability ρ^* for a price elasticity $\epsilon_p \in [0, 5]$.

5.5.2 Profits and retention rates with fairness constraints

Let us define the expected average profit (P) and retention rate (R) as follows:

$$P = \frac{1}{n}\sum_{i=1}^{n}(1-(\rho_i+\delta_i^\rho))\cdot(\delta_i^\pi\pi_i - L_i) \qquad \text{and} \qquad R = \frac{1}{n}\sum_{i=1}^{n}1-(\rho_i+\delta_i^\rho). \quad (5.4)$$

The expected retention rate R is defined by averaging over n policyholders the probability of retaining policyholder i, namely the term $1 - (\rho_i + \delta_i^\rho)$, with ρ and δ^ρ the baseline churn probability and additive change due to price updates. The profit P is defined by averaging the product of two terms. The second term $(\delta_i^\pi \pi_i - L_i)$ represents the profit (or loss) for contract i with $\delta_i^\pi \pi_i$ the updated premium inflow and L_i the observed claim amount outflow. The first term in P represents the retention probability that this profit/loss is realized for policyholder i. Averaging over all policyholders results in the expected average profit per client in the portfolio. We use all n = 25,838 policyholders with telematics during the period 2017-2019 to evaluate P and R. Both the baseline price π and churn probability ρ are calculated at the beginning of each year, based on the self-reported risk characteristics $\boldsymbol{x}$ available at that time. The price updates δ^π and (indirectly related) churn updates δ^ρ depend on the registered driving behavior $\boldsymbol{z}$ during the year. We assume that this information becomes available to policyholders as the year progresses. Finally, the loss payments L depend on the claim experience during each year.

Our goal is to compare profits and retention rates under the telematics paradigm to the baseline situation without telematics, i.e., when $\delta^\rho = 0$ and $\delta^\pi = 1$ in Equation (5.4). This baseline results in profits of 12.45 Euro per policyholder and a retention rate of 90.85%. We propose two constraints that allow for a fair and realistic comparison of telematics versus the baseline, namely a solidarity/commercial constraint via update limits and a redistribution constraint via a scale factor α:

$$\delta_{lo}^\pi \leq \delta^\pi \leq \delta_{hi}^\pi \qquad \text{and} \qquad \sum_{i=1}^{n} (1-\rho_i) \cdot \pi_i = \sum_{i=1}^{n} (1-\rho_i) \cdot \alpha \cdot \delta_i^\pi \cdot \pi_i. \tag{5.5}$$

Figure 5.11 showed that price updates δ^π result in huge discounts and penalties. We want to refrain from such excessive price increases as this goes against the nature of insurance and the principle of solidarity. From a commercial point of view, it is reasonable to assume that an insurer desires to put a maximum limit on the discount for financial protection. The first constraint in Equation (5.5) therefore restricts price updates by imposing lower and upper limits δ_{lo}^π and δ_{hi}^π. Further, we want to use the updates to redistribute the premium volume among policyholders. This is achieved by scaling the updates δ^π with a fixed factor α to ensure that the equality in the second constraint in Equation (5.5) holds. This redistribution constraint allows for a fair comparison of profits as the telematics and baseline tariff result in the same expected total premium inflow under the assumption of zero price elasticity, i.e., $\epsilon_p = 0$ and $\delta^\rho = 0$.

Figure 5.16 shows the distribution of the updates δ^π for five symmetrical lower and upper bounds, namely $\delta_{hi}^\pi = 1 + \delta_{lo}^\pi$ with $\delta_{lo}^\pi \in \{0.5, 0.4, 0.3, 0.2, 0.1\}$. This results in price increases and decreases of maximum 50%, 40% up to 10% respectively. The gray lines connect updates δ_i^π for random policyholders i under the different limits and indicate how the updates end up in the lower/upper bound for stricter limits. Table 5.10 reports the scale factor α and median/average value of the updates δ^π (respectively indicated by a horizontal bar and open circle in Figure 5.16). Both the median and average updates stay below one, indicating that more than half of the policyholders are receiving a discount thanks to the telematics updates. Furthermore, the median of the resulting price $\delta^{\pi}\pi$ remains below the median baseline price π of 304.7 Euro. The average price is approximately equal to the average baseline price of 342.3 Euro in all the scenarios, a direct consequence of our redistribution constraint.

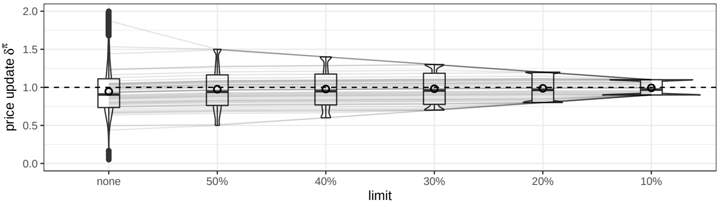


Figure 5.16: Distribution of the price updates δ^{π} for different limits.

		Symmetrical lower and upper limits					
		none	50%	40%	30%	20%	10%
Scale factor α		1.010	1.043	1.053	1.062	1.071	1.075
Price update δ^{π}	Median	0.911	0.941	0.950	0.959	0.966	0.970
	Average	0.968	0.975	0.978	0.982	0.987	0.993
Premium $\delta^{\pi}\pi$	Median	275.2	284.1	286.9	290.7	295.9	301.0
	Average	342.9	342.7	342.7	342.6	342.5	342.4

Table 5.10: Statistics on updates δ^{π} and prices $\delta^{\pi}\pi$ for different limits.

Figure 5.17 shows the expected profits per client on the x-axis and retention rates on the y-axis for different values of the symmetrical update limits (color) and price elasticity ϵ_p (plot shape). The vertical and horizontal dashed lines indicate the baseline profit (12.45 Euro) and retention rate (90.85%) without using telematics ($\delta^{\rho} = 0$ and $\delta^{\pi} = 1$). Notice that all situations lead to the baseline profit and retention for $\epsilon_p = 0$, a direct consequence of our redistribution constraint. Profits and retention rates diverge for different limits when $\epsilon_p > 0$. The limits of 10% up to 40% always result in higher profits compared to the baseline situation, at the cost of lower retention rates. For a moderate price elasticity $\epsilon_p \in [1, 2]$, the 10% and 20% limit result in profits between 7 and 11 Euro per customer on top of the baseline, with retention rates remaining above 87% and 82% respectively. An extra profit of 10 Euro per customer results in a total excess profit of almost 260,000 Euro. A higher price elasticity typically results in more profits but a decrease in client retention. The 50% limit has lower profits compared to the baseline for a price elasticity $\epsilon_p \in [1, 2]$

and no limit results in lower profits over the full range of ϵ_p. This is driven by the relatively low premiums in these cases, as indicated the median values in Table 5.10.

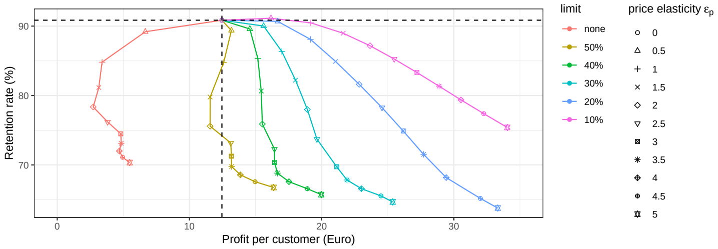


Figure 5.17: Profit and retention rate by limit (color) and price elasticity ϵ_p (shape).

This stylized example indicates that both policyholders and the insurer are able to gain from telematics via lower premiums on average and higher expected profits respectively. We now turn to constrained optimization techniques for profit or retention maximization.

5.5.3 Constrained profit or retention maximization

We maximize the expected profit P, given that we want to retain a minimum proportion of the portfolio R^*. This corresponds to the following constrained optimization problem:

$$\begin{aligned} \max_{\alpha} P(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}(1-(\rho_i+\delta_i^{\rho}))\cdot(\alpha\delta_i^{\pi}\pi_i - L_i), \\ \text{subject to } R(\alpha) &= \frac{1}{n}\sum_{i=1}^{n} 1-(\rho_i+\delta_i^{\rho}) \geq R^*. \end{aligned} \tag{5.6}$$

We explicitly take the dependence on the scale factor α into account via the premium updates δ^{π}, but the churn updates implicitly also depend on α via $\delta^{\rho} = \epsilon_p \cdot (\alpha\delta^{\pi} - 1)$. We find an efficient frontier by varying R^* over a range of values and maximizing $P(R^*)$ via α. Figure 5.18 shows the efficient frontiers when $R^* \in [0.75, 0.9]$ for various combinations of the update limits δ_{lo}^{π} and δ_{hi}^{π} (grid) and price elasticity ϵ_p (color). We no longer focus on symmetrical bounds but allow all combinations in the set $\pm\{10\%, 30\%, 50\%\}$. The profit and retention rate under the baseline without using telematics are again indicated by the dashed lines for comparison purposes.

For an inelastic portfolio ($\epsilon_p = 0.5$), the expected profit is always higher than the baseline. The range of excess profits per policyholder increases with the upper limit going from 28 Euro for 10% to 86 Euro for 50%. The large profits with high upper limits come at the cost of lower retention and losing around 15% of the policyholders. For a unit elastic portfolio ($\epsilon_p = 1$), the maximal profits drop to around 47 Euro per customer. Telematics results in lower profits compared to the baseline (or even losses) for retention rates above 85% when the limits widen (i.e., going to the left bottom of Figure 5.18). The efficient frontier shifts further to the left for elastic portfolios ($\epsilon_p > 1$). For the symmetrical 10% limit the profits remain larger than the baseline, while for the symmetrical 50% limit they never exceed the baseline.

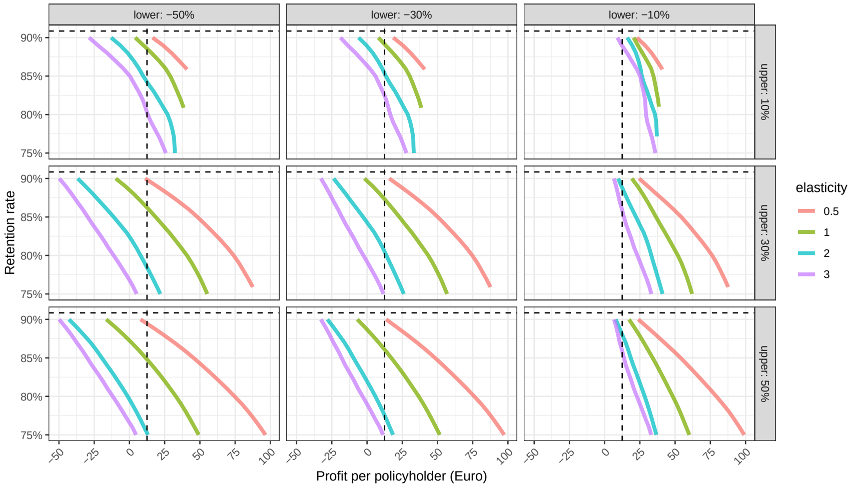


Figure 5.18: Profits and retention rates by values of the update limit (grid) and price elasticity (color).

A company with a clear idea on the price elasticity of its customers can use this analysis to pinpoint the retention rate and update limits in a profit-maximizing strategy. Without an accurate estimate of price elasticity, these results can still be used for a risk-return analysis. The symmetrical 10% limits are almost certain to result in (small) profits, while the symmetrical 50% limits can result in huge profits or detrimental losses depending on the actual price elasticity. A lower limit of 10% and upper limit of 50% give the best of both worlds, high return and low risk, but such a structure with low discounts and high penalties will be hard to sell to customers.

We now maximize the expected retention rate R, given that we expect to make a minimum amount of profit P^*. This corresponds to the following constrained optimization problem:

$$\begin{aligned} \max_{\alpha} R(\alpha) &= \frac{1}{n}\sum_{i=1}^{n} 1 - (\rho_i + \delta_i^{\rho}), \\ \text{subject to}\, P(\alpha) &= \frac{1}{n}\sum_{i=1}^{n} (1 - (\rho_i + \delta_i^{\rho})) \cdot (\alpha\delta_i^{\pi}\pi_i - L_i) \geq P^*. \end{aligned} \tag{5.7}$$

Again, the churn update δ^{ρ} implicitly depends on α. Figure 5.19 shows the retention rates and profits for various combinations of the update limits δ_{lo}^{π} and δ_{hi}^{π} (grid), price elasticity ϵ_p (color) and required excess profits above the baseline (shape). The profit and retention rate without telematics is indicated by the dashed lines. For example, an excess profit of 10 Euro above the baseline profit of 12.45 Euro implies that the minimum profit P^* equals 22.45 Euro. Notice that the combination of a 10% upper limit and excess profit of 35 Euro per client is impossible, as the plotting characters do not attain $P^* = 47.45$ in the top panels of Figure 5.19.

In general, retention rates are decreasing for an increasing price elasticity and excess profit, while retention increases when going from wide to narrow limits (bottom left to top right in Figure 5.19). In some settings it is possible to achieve higher retention than the baseline. This is for example the case with the symmetrical 10% limit in an inelastic market for low excess profits and in an elastic market without excess profit. Retention rates remain relatively high in both inelastic and unit elastic portfolios, but they decrease drastically when using wider limits in elastic portfolios. A solid risk-return analysis is therefore very important in an elastic market.

Our analysis shows that telematics has big economical value for insurers, but care has to be taken in implementing the updating scheme to align risk and return. We believe this helps companies to make decisions on the discount/penalty structure that aligns best with the strategic goals regarding target profits or retention rates. This can be combined together with marketing and consumer studies on which types of structures would be accepted by policyholders.

5.6 Conclusions

On the one hand, insurance companies have an abundance of historical data and in-house expertise on technical risk assessment with self-reported characteristics. On the other hand, new technologies such as telematics offer exciting opportunities to innovate and further improve the pricing practice. This chapter aims at combining the best of both worlds. We first develop a baseline

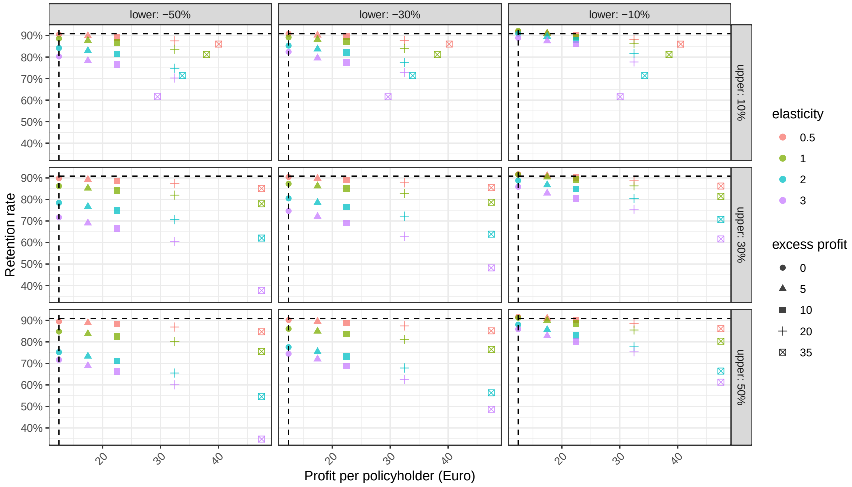


Figure 5.19: Profits and retention rates by values of the limit (grid), elasticity (color) and P^* (shape).

pricing model on a large portfolio with only self-reported features. Next, we propose an explainable updating mechanism to incorporate driving behavior information into the baseline tariff. The yearly mileage, amount of night driving and rate of harsh braking and lateral movement events are used to update the baseline price in an intuitive way. We analyze the added value of telematics for insurance pricing from both a statistical and managerial perspective. The statistical performance shows that telematics improves the risk classification process, resulting in a better assessment of claim risk for both the in-sample train and out-of-sample test data. The managerial evaluation shows the added economic value of telematics with respect to profits and retention rates under different assumptions of the price elasticity of the clients. We show how the updating system's design has an impact on the risk-return profile. We believe this analysis can help managers, actuaries and marketeers to bring a successful commercial telematics product into the market, aligned with the strategic goals and risk-appetite of the company.

The application of telematics technology within the (motor) insurance industry poses many opportunities, but is still in its infancy. In this chapter we take a first step in utilizing the added value of telematics and highlight the improvements in risk classification and pricing of an MTPL product. The resulting gains for both policyholders (e.g., lower premiums) and the insurer (e.g., higher profits) might spark the interest of insurance companies who were on the fence about launching a telematics product until now. In this chapter we take the angle of an incumbent firm with in-house expertise who is interested in updating the current pricing structure with telematics information. In a next project we aim to develop a purely telematics tariff structure based on driving behavior and claims data, without relying on any self-reported risk characteristics. A more dynamic structure of premium payment, for example like a monthly usage-based subscription service, could represent how insurtech startups try to make a disruptive entry in the market.

In our work, the churn and pricing models are not interconnected. It can be interesting to connect insights on churn behavior with price updates from telematics to improve marketing offers. For example by offering a bigger discount to safe drivers with a high probability to surrender the policy, thereby persuading these good risks to stay with the insurance company.

Another path for future research is to analyze post-accident changes in driving behavior and related price implications. Bonus-malus systems reward policyholders with a discount for claim-free years and penalize with a surcharge following an accident at fault (Lemaire, 1995). These systems are common in the European insurance market and result in a fixed discount/penalty for the next period. However, psychological studies show that involvement in a traffic accident can lead to lack of confidence behind the wheel, anxiety and emotional distress, resulting in slower and more cautious driving (Mayou et al., 1991, 1993). The analysis of post-accident driving behavior can lead to more dynamic bonus-malus updates, for example by rewarding improved behavior with less severe penalties or a faster convergence to the initial bonus.

Chapter 6

Conclusions and outlook

This thesis puts focus on new modeling techniques and new data sources to calculate the pure premium. This is the estimated amount that the insurance company needs to pay future claims, without any profit or loss. This pure premium is intended to cover the underlying risk of each profile, and its calculation relies heavily on statistical methods and the available information. Our work contributes to the more accurate calculation of the pure premium, but this is far from the eventual price that is charged to the policyholder. The technical premium is obtained by adding a safety margin and other loading factors to the pure premium. The safety margin takes into account the inherent uncertainty of the future and allows to adjust for differences between the predictions and observed reality. The loading factors cover all sorts of costs such as administration, claims management, commissions, taxes, sales and marketing. These loadings are usually defined as a percentage of the pure premium and can be defined via expert judgment or calculated via statistical approaches (Yang et al., 2020). The commercial premium, the one which is actually charged to the policyholders, is then derived from the technical premium by taking into account the company's goals and extra constraints. This includes the current market positioning, the desired level of profit, competition and premium stability, as well as legal and IT constraints on the information that one can use. This shows that the journey from pure to commercial premium is still a challenging one, but we hope that this thesis already provides a good starting point to calculate accurate pure premiums, from which can be built further.

Recent technological advancements have boosted the performance of many predictive models, leading to innovating business applications across industries. Innovation in the insurance industry is however not an easy task for two main reasons. Firstly, insurance plays a crucial role in protecting our modern society and the people in it against life-ruining financial losses. This makes insurance pricing a high-stakes decision with a big impact on a person's life. Secondly, the insurance industry is heavily regulated to make sure that everyone is treated fairly without any discrimination. These two reasons contribute to the fact that explainability is a key requirement for any practical insurance tariff. Transparent communication of the underlying decisions is a right for policyholders and should be taken into account during the tariff design process.

The inherent need for transparency makes it more difficult to apply black box ML models to insurance. In Chapter 2 we therefore remained within the actuarial comfort zone of well-known statistical white box models. We make use of ML techniques to transform a flexible GAM into a very transparent GLM. Chapter 3 however shows that our GAM/GLM approach is outperformed by a GBM black box, both from a statistical and an economic perspective. The price to pay for this extra performance is reduced interpretability of the pricing model. We show some techniques to open the black box and understand the decision process, but this might not be sufficient for the regulator. In Chapter 4 we bring pricing back to the world of white box models. By extracting knowledge from a complex black box, we perform smart feature engineering which allows to fit a transparent global surrogate. Our approach results in an explainable tariff model which approximates the performance of the original black box closely. We believe this allows actuaries to combine both performance and interpretability, thereby having a competitive solid tariff while complying with regulations.

In our research on new modeling approaches we mainly put focus on tree-based ML techniques. A straightforward extension is to take a journey into the realm of neural networks (NNs) and deep learning. Holvoet et al. (2021) is unpublished work to which we contributed, where we investigate the application of NNs to insurance pricing. In that case study we find that standard NNs perform worse than a GBM for example. We suspect that NNs are great alternatives for unstructured text or image data, but GBMs tend to perform better on structured tabular insurance data. However, we did find that NNs can be used as an adjustment model to improve upon a baseline model such as a GBM or GLM. This is very promising as it means that one might be able to use a transparent GLM in combination with a NN that makes slight adjustments to that tariff. This can increase predictive performance without losing too much of the inherent interpretability of the GLM baseline.

In practice, insurance contracts can stipulate deductibles and/or policy limits, which makes the observed loss amount left-truncated and/or right-censored. Such aspects need be taken into account in the statistical estimation process via actuarial models or tools from survival analysis (Klugman et al., 2012). Lopez et al. (2016) show how to adapt CART to survival data with Kaplan-Meier weights. When there is no policy limit in place, then extreme loss amounts can occur, which need to be addressed appropriately. The central part of the loss distribution can be modeled via approaches outlined in this thesis, while the tail of the distribution should be modeled via tools from Extreme Value Theory (EVT). Recent work brings techniques from EVT to tree-based models via Generalized Pareto Regression Trees (Farkas et al., 2020) and gradient boosting for quantile regression (Velthoen et al., 2021) to model extreme claims. There is still a big opportunity for research that brings these practical policy considerations to the world of machine learning via solid statistical processes.

In Chapter 5 we switch from a focus on new modeling approaches to new data sources for insurance pricing. Telematics technology allows to tap into a whole new category of information on policyholder behavior. This behavioral data can be valuable to understand a policyholder's risk profile in addition to insights obtained from the typical self-reported proxy characteristics. We show the added value of driving behavior data to update prices from a classical baseline tariff. We again put forward the explainability of our updating mechanism as a key requirement. Policyholders need to be able to monitor and understand the price effects of their driving behavior, which might even result in safer or less driving in the end. In our study, telematics information results in improved risk classification, lower premiums on average and possible profits for the insurer. We believe that there are many unexplored roads for insurers to make proper use of this new technology, of course within the bounds of fairness and regulations.

In our research on telematics we take the point of an incumbent firm with strong developed in-house expertise on motor pricing. Such firms might be most interested in using driving behavior to update their current pricing structure. Insurtech players might be more interested in finding a disruptive business model which purely puts focus on driving behavior for pricing. A possible extension on our work would therefore be to go from a hybrid pricing structure to one that uses only telematics information. Policyholders can be clustered together based on their driving habits, making groups of similar driving profiles. Within these profiles one could then propose a fixed price-per-km, possibly adding extra costs once dangerous events are registered. In the end, it would be possibly to switch from the typical static payment structure to a more dynamic subscription-based UBI approach. Besides pricing, telematics can also assist in faster claims handling thanks to detailed crash reports or other services like for example predictive maintenance alerts. This technology can possibly lead to big changes in the insurance value chain and current business models.

The application of machine learning and telematics approaches in the insurance industry is still in its infancy, at least in the Belgian market. Main reasons of adoption reluctance are the fact that these new techniques are harder to explain to stakeholders, implement in an IT infrastructure and manually adjust to business needs. However, as we try to show in this thesis, it might be possible to take advantage of these new technologies while still adhering to the specific requirements of the insurance industry. We hope that this thesis contributes to assist practitioners in translation new insights from research to practice.

Appendix Chapter 2

A.1 Full specification of the resulting GLMs

	Coefficient	SE (σ)	t-stat	p-value
(Intercept)	-2.17531	0.02138	-101.734	< 2e-16 ***
COVERAGEPO	-0.11964	0.01678	7.129	1.01e-12 ***
COVERAGEFO	-0.11399	0.02187	-5.212	1.87e-07 ***
FUELdiesel	0.17802	0.01574	11.313	< 2e-16 ***
AGEPH[18,26)	0.27717	0.02967	9.341	< 2e-16 ***
AGEPH[26,29)	0.14379	0.02925	4.917	8.80e-07 ***
AGEPH[29,33)	0.06492	0.02612	2.485	0.01295 *
AGEPH[51,56)	-0.06093	0.02648	-2.301	0.02142 *
AGEPH[56,61)	-0.16279	0.03330	-4.888	1.02e-06 ***
AGEPH[61,73)	-0.24733	0.02963	-8.348	< 2e-16 ***
AGEPH[73,95]	-0.19082	0.04086	-4.670	3.01e-06 ***
POWER[10,36)	-0.20689	0.03305	-6.259	3.88e-10 ***
POWER[36,46)	-0.06471	0.02268	-2.853	0.00434 **
POWER[75,243]	0.11630	0.02807	4.143	3.43e-05 ***
BM[1,2)	0.12522	0.02296	5.454	4.93e-08***
BM[2,3)	0.18461	0.03314	5.570	2.55e-08 ***
BM[3,7)	0.34292	0.02125	16.140	< 2e-16 ***
BM[7,9)	0.48463	0.02978	16.271	< 2e-16 ***
BM[9,11)	0.54521	0.02702	20.176	<2e-16 ***
BM[11,22]	0.78219	0.02592	30.175	< 2e-16 ***
GEO[-0.48,-0.27)	-0.32882	0.05320	-6.181	6.37e-10***
GEO[-0.27,-0.14)	-0.20339	0.02326	-8.744	< 2e-16 ***
GEO[-0.14,-0.036)	-0.15519	0.01944	-7.985	1.40e-15 ***
GEO[0.11,0.34]	0.19847	0.01822	10.894	< 2e-16 ***
AGEPHPOWER-0.052	-0.07053	0.04641	-1.520	0.12859
AGEPHPOWER-0.029	-0.02585	0.02609	-0.991	0.32176
AGEPHPOWER0.039	0.05843	0.03967	1.473	0.14078
AGEPHPOWER0.047	0.03483	0.03730	0.934	0.35051
Significance codes:	$^{\cdot}$p<0.1;*p<0.05; **p<0.01; ***p<0.001			

Table A.1: Full specification of the frequency GLM.

	Coefficient	SE (σ)	t-stat	p-value
(Intercept)	6.06322	0.02528	239.798	$<$ 2e-16 ***
COVERAGEPO	-0.16280	0.02501	-6.510	7.70e-11 ***
COVERAGEFO	0.10968	0.03236	3.389	0.000703 ***
AGEPH[18,28)	-0.02757	0.03665	-0.752	0.451893
AGEPH[28,42)	-0.11173	0.02535	-4.407	1.06e-05 ***
AGEPH[64,71)	0.08156	0.04409	1.850	0.064313 .
AGEPH[71,92]	0.33702	0.04749	7.096	1.33e-12 ***
BM[1,2)	0.08857	0.03459	2.561	0.010460*
BM[2,8)	0.13165	0.02878	4.574	4.82e-06 ***
BM[8,10)	0.19560	0.04313	4.535	5.80e-06 ***
BM[10,22]	0.31009	0.03457	8.970	$<$ 2e-16 ***
Significance codes:	$^{\cdot}$p<0.1;*p<0.05; **p<0.01; ***p<0.001			

Table A.2: Full specification of the severity GLM.

Appendix Chapter 3

B.1 List of variables in the MTPL data

Claim information and exposure-to-risk measure	
`nclaims`	The number of claims filed by the policyholder.
`amount`	The total amount claimed by the policyholder in euro.
`expo`	The fraction of the year during which the insurer was exposed to the risk.
Categorical risk factors	
`coverage`	Type of coverage provided by the insurance policy: TPL, TPL+ or TPL++. TPL = only third party liability, TPL+ = TPL + limited material damage, TPL++ = TPL + comprehensive material damage.
`fuel`	Type of fuel of the vehicle: gasoline or diesel.
`sex`	Gender of the policyholder: male or female. (As from 21 December 2012, the European Court of Justice prohibited the use of gender in insurance tariffs to avoid discrimination between males and females, known as the Test-Achats Ruling. Gender is therefore only investigated for use within an internal technical tariff, but can not be used in a commercial product.)
`use`	Main use of the vehicle: private or work.
`fleet`	The vehicle is part of a fleet: yes or no.
Continuous risk factors	
`ageph`	Age of the policyholder in years.
`power`	Horsepower of the vehicle in kilowatt.
`agec`	Age of the vehicle in years.
`bm`	Level occupied in the former compulsory Belgian bonus-malus scale. From 0 to 22, a higher level indicates a worse claim history, see Lemaire (1995). (This variable is typically not used as an a priori rating factor, but rather as an a posteriori correction in a credibility framework or bonus-malus scheme. We however keep `bm` in the data to assess the amount of information contained in this variable and to investigate the resulting effect.)
Spatial risk factor	
`long`	Longitude coordinate of the center of the municipality where the policyholder resides.
`lat`	Latitude coordinate of the center of the municipality where the policyholder resides.

Table B.1: Description of the available variables in the `MTPL` data.

B.2 Search grid for the tuning parameters

Regression tree	$cp \in \{1.0 \times 10^{-5}, 1.1 \times 10^{-5}, \ldots, 1.0 \times 10^{-2}\}$ $\gamma \in \{2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}^{*}$
Random forest	$T \in \{100, 200, \ldots, 5{,}000\}$ $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$
Gradient boosting machine	$T \in \{100, 200, \ldots, 5{,}000\}$ $d \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Table B.2: Search grid for the tuning parameters in the different tee-based machine learning techniques.
* Note that the γ tuning parameter is only used in frequency models for the Poisson deviance.

B.3 Random forests for claim frequency data

The left and right panel of Figure B.1 show the partial dependence plot of the age and spatial effect in the claim frequency random forests, respectively. These effects are discussed in Section 3.4.3. The six random forests for frequency contain rather different number of trees, ranging from 100 to 5,000 (see Table 3.3 earlier). However, these random forests exhibit very similar effects for the age of the policyholder in Figure B.1. This indicates that the underlying model structure does not change drastically after including a sufficient number of trees.

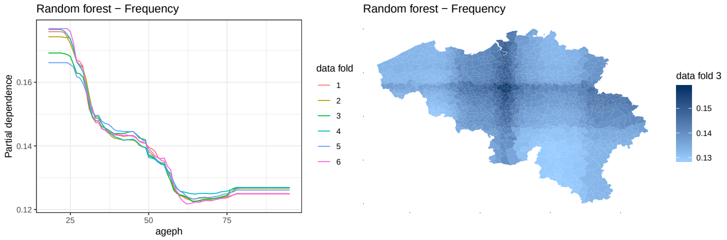


Figure B.1: Effect of the age of the policyholder (left) and the municipality of residence (right) on frequency in a random forest.

B.4 Filing requirements for pricing models

Insurance companies can be required by regulation to file their rating model on paper. This section presents such frequency and severity models, trained on the data where $\mathcal{D}_3$ was kept as hold-out test set. Section B.4.2 and B.4.1 show the GLMs and regression trees respectively. The other five GLMs and regression trees are not shown due to lack of space. This filing requirement is more difficult to satisfy for the ensemble methods, but it is possible by printing the individual trees. However, this would result in a large amount of pages which is not very practical or insightful for the regulator.

B.4.1 Regression tree

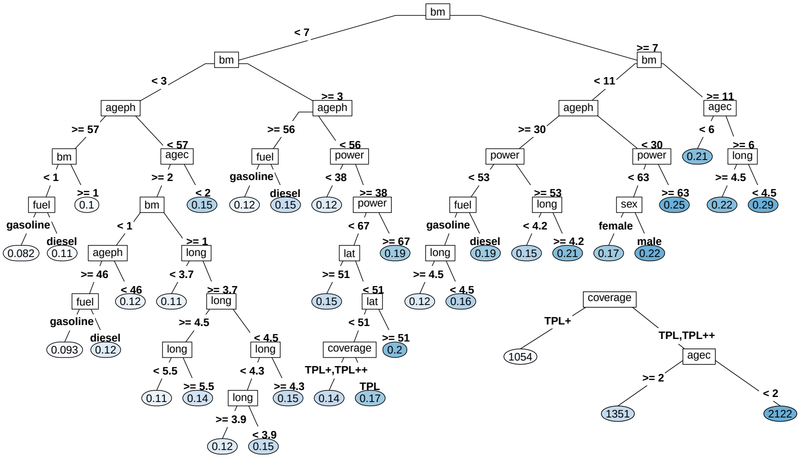


Figure B.2: Regression trees for claim frequency (big, top left) and severity (small, bottom right), both trained on the data where $\mathcal{D}_3$ was kept as hold-out test set.

B.4.2 GLM

	Coefficient	SE (σ)	t-stat	p-value
Intercept	-2.19^{***}	0.03	-64.09	0.00
coverageTPL+	-0.10^{***}	0.02	-5.46	0.00
coverageTPL++	-0.10^{***}	0.03	-3.97	0.00
fueldiesel	0.18^{***}	0.02	10.00	0.00
fleetyes	-0.13^{***}	0.05	-2.72	0.01
ageph[18,26)	0.32^{***}	0.04	8.94	0.00
ageph[26,29)	0.16^{***}	0.04	4.61	0.00
ageph[29,32)	0.11^{***}	0.04	3.11	0.00
ageph[32,35)	0.02	0.04	0.46	0.65
ageph[50,54)	-0.11^{***}	0.03	-3.53	0.00
ageph[54,58)	-0.05	0.04	-1.28	0.20
ageph[58,62)	-0.23^{***}	0.05	-4.93	0.00
ageph[62,73)	-0.25^{***}	0.04	-7.06	0.00
ageph[73,95]	-0.21^{***}	0.05	-4.42	0.00
bm[1,2)	0.14^{***}	0.03	5.46	0.00
bm[2,3)	0.16^{***}	0.04	4.48	0.00
bm[3,5)	0.37^{***}	0.03	11.86	0.00
bm[5,7)	0.32^{***}	0.03	11.39	0.00
bm[7,9)	0.48^{***}	0.03	14.77	0.00
bm[9,11)	0.52^{***}	0.03	17.30	0.00
bm[11,22]	0.75^{***}	0.03	26.07	0.00
agec[9,14)	0.04^{*}	0.02	1.91	0.06
agec[14,48]	-0.02	0.03	-0.75	0.46
power[10,35)	-0.16^{***}	0.04	-3.95	0.00
power[35,42)	-0.05	0.03	-1.54	0.12
power[42,49)	-0.02	0.03	-0.55	0.58
power[59,73)	0.02	0.02	0.89	0.37
power[73,92)	0.07^{*}	0.04	1.76	0.08
power[92,243]	0.13^{***}	0.05	2.88	0.00
latlong[-0.46,-0.36)	-0.31	0.19	-1.62	0.11
latlong[-0.36,-0.26)	-0.33^{***}	0.06	-5.97	0.00
latlong[-0.26,-0.18)	-0.20^{***}	0.03	-5.86	0.00
latlong[-0.18,-0.12)	-0.23^{***}	0.04	-6.38	0.00
latlong[-0.12,-0.061)	-0.13^{***}	0.03	-4.41	0.00
latlong[-0.061,-0.017)	-0.05	0.03	-1.62	0.11
latlong[0.025,0.077)	0.06^{**}	0.03	2.42	0.02
latlong[0.077,0.14)	0.03	0.03	0.86	0.39
latlong[0.14,0.23)	0.04	0.03	1.29	0.20
latlong[0.23,0.33]	0.36^{***}	0.03	12.37	0.00
agephpower-0.05	-0.04	0.06	-0.61	0.54
agephpower-0.02	-0.11^{***}	0.04	-3.15	0.00
agephpower-0.01	-0.03	0.03	-0.79	0.43
agephpower0.01	-0.00	0.03	-0.02	0.98
agephpower0.02	-0.04	0.05	-0.67	0.50
agephpower0.04	0.04	0.04	1.07	0.29
Significance codes:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$			

Table B.3: Frequency GLM specification, trained on the data where $\mathcal{D}_3$ was kept as hold-out test set.

	Coefficient	SE (σ)	t-stat	p-value
Intercept	7.31 ***	0.08	88.50	0.00
coverageTPL+	−0.23 ***	0.05	−4.62	0.00
coverageTPL++	0.18 **	0.07	2.54	0.01
ageph[18,25)	0.13	0.11	1.23	0.22
ageph[25,28)	−0.02	0.10	−0.15	0.88
ageph[28,30)	0.08	0.11	0.70	0.49
ageph[30,33)	0.07	0.10	0.65	0.52
ageph[33,36)	−0.19 *	0.10	−1.87	0.06
ageph[36,39)	−0.26 **	0.10	−2.57	0.01
ageph[39,42)	−0.19 *	0.10	−1.84	0.07
ageph[42,45)	−0.01	0.10	−0.12	0.91
ageph[49,52)	−0.02	0.10	−0.21	0.83
ageph[52,55)	−0.13	0.11	−1.18	0.24
ageph[55,61)	−0.17 *	0.10	−1.69	0.09
ageph[61,66)	−0.20 *	0.11	−1.73	0.08
ageph[66,72)	−0.04	0.11	−0.33	0.75
ageph[72,95]	0.11	0.11	0.93	0.35
agec[0,2)	0.22 **	0.11	2.09	0.04
agec[2,3)	−0.11	0.09	−1.11	0.27
agec[3,4)	−0.08	0.10	−0.82	0.41
agec[4,6)	−0.08	0.07	−1.09	0.28
agec[6,7)	−0.10	0.08	−1.22	0.22
agec[7,8)	−0.10	0.09	−1.21	0.23
agec[10,11)	−0.12	0.09	−1.33	0.18
agec[11,12)	−0.05	0.09	−0.52	0.60
agec[12,14)	−0.14 *	0.08	−1.69	0.09
agec[14,48]	−0.01	0.09	−0.06	0.95
Significance codes:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$			

Table B.4: Severity GLM specification, trained on the data where $\mathcal{D}_3$ was kept as hold-out test set.

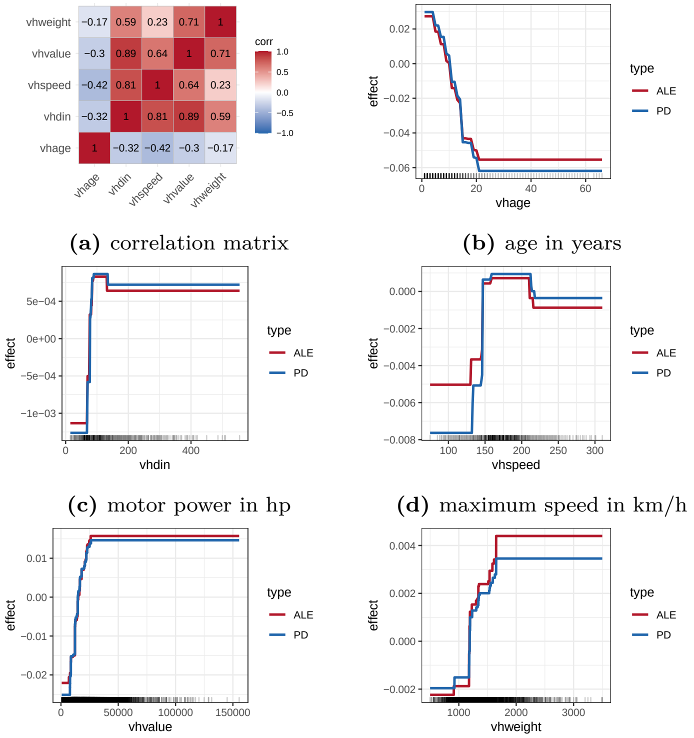
B.5 Supplementary material

Supplementary material related to this chapter can be found online at `https://github.com/henckr/sevtree`.

Appendix Chapter 4

C.1 PD and ALE for correlated features

Figure C.1 compares the PD and ALE for several vehicle characteristics in the `pricingame` dataset, namely the weight, value, maximum speed, horsepower and age. Figure C.1a shows that the vehicle age is negatively correlated with the other characteristics while there is a strong positive correlation between the weight, value, maximum speed and horsepower. Figures C.1b, C.1c, C.1d, C.1e and C.1f show the centered PD (in blue) and ALE (in red) for all the vehicle features. Both effects are very similar for each of the features, especially in the ranges with high observation counts as indicated by the black rugs on the x-axis. We observe some vertical shifts between the PD and ALE in feature ranges with low observation counts. However, these vertical shifts are not a problem for our maidrr procedure as we only use these effects to perform the feature grouping. Furthermore, observation counts are taken into account as weights in the penalized loss function of Eq. (4.2), further reducing the impact of these shifts on the obtained segmentation. This justifies the use of PD effects for grouping, even when dealing with correlated features.



(a) correlation matrix

(b) age in years

(c) motor power in hp

(d) maximum speed in km/h

(e) value in euros

(f) weight in kg

Figure C.1: Comparison of PD and ALE for correlated vehicle characteristics in the `pricingame` dataset.

C.2 General GLM formulation

A GLM allows any distribution from the exponential family for the target of interest y. This includes, among others, the normal, Bernoulli, Poisson and gamma distributions, making GLMs a versatile modeling tool. Denoting by $g(\cdot)$ the link function, the structure of a GLM with all categorical features $\boldsymbol{x}$ is:

$$g(\mathbb{E}[y]) = \boldsymbol{x}^\top \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{d} \beta_j x_j .$$

The $d+1$ dimensional vector $\boldsymbol{x}$ contains a 1 for the intercept β_0 together with d dummy variables $x_j \in \{0,1\}$. A categorical feature x with m levels contains a reference level which is captured by the intercept. The other $m-1$ levels are coded via dummy variables to model the differences between those levels and the reference level, captured by the coefficients β_j.

C.3 Global interpretations for a GLM

The Poisson GLM with logarithmic link function to model the number of claims for the `norauto` dataset has the following structure and fitted coefficients:

$$\begin{aligned}
\ln\left(\mathbb{E}[\texttt{nclaims}]\right) = &-2.40 + 0.54\,\texttt{Male}_{no} + 0.09\,\texttt{Young}_{yes} \\
&- 0.76\,\texttt{DistLimit}_{8000km} - 0.62\,\texttt{DistLimit}_{12000km} \\
&- 0.51\,\texttt{DistLimit}_{16000km} - 0.33\,\texttt{DistLimit}_{20000km} - 0.20\,\texttt{DistLimit}_{30000km} \\
&- 0.17\,\texttt{GeoRegion}_{Low-\ \&\ Low+} - 0.05\,\texttt{GeoRegion}_{Med-} + 0.23\,\texttt{GeoRegion}_{High+} \\
&- 0.08\,\texttt{DistLimit:GeoRegion}_{8000/12000/16000km:High+\ \&\ nolimit:Low-/Low+/Med-}
\end{aligned}$$

where $Male_{yes}$, $Young_{no}$, $DistLimit_{nolimit}$ and $GeoRegion_{Med+\ \&\ High-}$ are the reference levels captured by the intercept. These references are the levels which contain the highest number of policyholders such that the intercept models the claim frequency of an "average" policyholder. Taking the inverse link function, namely the exponential, on both sides results in a multiplicative GLM prediction function with the following global interpretations:

- The predicted claim frequency for an older male policyholder without a driving distance limit and living in the Med+ or High- geographical region equals 0.09 or $\exp(-2.40)$.
- Predictions are 72% higher for female policyholders compared to males as $\exp(0.54) = 1.72$. *Note: In 2012, the EU put forward rules on gender-neutral pricing in the insurance industry such that gender is no longer allowed as a rating factor in a commercial tariff. However, gender is typically still included in the internal technical analysis.*
- As $\exp(0.09) = 1.09$, predictions are 9% higher for young compared to old policyholders.
- For policyholders with a driving distance limit of 8, 12, 16, 20 and 30 thousand kilometers, predictions respectively amount to 47%, 54%, 60%, 72% and 82% of those for someone without a limit. There is a clear increasing trend of claim risk with the distance limit.
- Predictions for policyholders living in the Low or Med- geographical regions amount to respectively 84% and 95% of those for the Med+/High- regions, whereas predictions increase with 26% for those in the High+ region.

- The interaction between the distance limit and geographical region results in a negative correction for policyholders with the most risky level of one of the features and a low-risk level of the other. As exp(0.08) = 0.92, predictions are reduced by 8% for policyholders living in the High+ region with a maximal distance limit of 16,000 kilometers and for those with no distance limit living in the Low-, Low+ or Med- region.

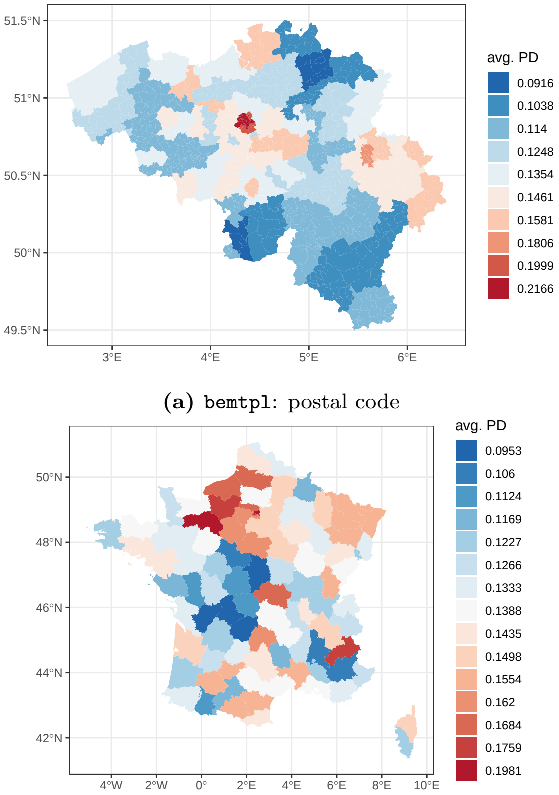
The observations listed above are globally valid for the GLM predictions on every policyholder. It is possible to detail the full working regime of our maidrr GLM as all features are represented in a categorical format. A decision table can be constructed by combining all possible levels of the different features. Table C.1 shows part of the decision table for the `norauto` dataset, with the four lowest and highest predictions indicated in italics and bold respectively. The three other parts for *Male = Yes & Young = No* and *Male = No & Young = Yes/No* are not shown for space reasons.

	Male	Young	DistLimit	GeoRegion	GLM prediction (%)
1	Yes	Yes	8000 km	Low- & Low+	*3.88*
2	Yes	Yes	8000 km	Medium-	*4.41*
3	Yes	Yes	8000 km	Medium+ & High-	*4.62*
4	Yes	Yes	8000 km	High+	5.36
5	Yes	Yes	12000 km	Low- & Low+	*4.47*
6	Yes	Yes	12000 km	Medium-	5.08
7	Yes	Yes	12000 km	Medium+ & High-	5.32
8	Yes	Yes	12000 km	High+	6.17
9	Yes	Yes	16000 km	Low- & Low+	4.99
10	Yes	Yes	16000 km	Medium-	5.67
11	Yes	Yes	16000 km	Medium+ & High-	5.94
12	Yes	Yes	16000 km	High+	6.89
13	Yes	Yes	20000 km	Low- & Low+	5.94
14	Yes	Yes	20000 km	Medium-	6.75
15	Yes	Yes	20000 km	Medium+ & High-	7.07
16	Yes	Yes	20000 km	High+	**8.92**
17	Yes	Yes	30000 km	Low- & Low+	6.78
18	Yes	Yes	30000 km	Medium-	7.70
19	Yes	Yes	30000 km	Medium+ & High-	8.07
20	Yes	Yes	30000 km	High+	**10.18**
21	Yes	Yes	no limit	Low- & Low+	7.63
22	Yes	Yes	no limit	Medium-	8.67
23	Yes	Yes	no limit	Medium+ & High-	**9.87**
24	Yes	Yes	no limit	High+	**12.45**

Table C.1: Part of the GLM predictions in a decision table for the `norauto` dataset.

C.4 Geographical segmentation

Figure C.2 shows the average PD effect for geographical regions where groups are indicated by colors. Figure C.2a shows the postal code areas on the map of Belgium with the initial 80 regions from the `bemtpl` portfolio segmented in 10 clusters. The capital Brussels in the center of Belgium (red colored), together with other big cities (orange colored), are risky due to heavy traffic in those densely populated areas. The rural regions in the northeast and south of Belgium are less risky. Figure C.2b shows the INSEE department code areas on the map of France with the initial 96 regions from the `pricingame` portfolio segmented in 15 clusters. The capital Paris and surrounding departments in the north of France (red/orange colored) are high-risk areas.



(a) `bemtpl`: postal code

(b) `pricingame`: INSEE department code

Figure C.2: Average PD effect for geographical regions where groups are indicated by colors.

List of Figures

List of Tables

Bibliography

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Antonio, K., Frees, E. W., and Valdez, E. A. (2010). A multilevel analysis of intercompany claim counts. *Astin Bulletin*, 40(01):151–177.

Antonio, K. and Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *Advances in Statistical Analysis*, 96(2):187–224.

Apley, D. W. and Zhu, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Armstrong, M. P., Xiao, N., and Bennett, D. A. (2003). Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers*, 93(3):595–623.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Ayuso, M., Guillén, M., and Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46(3):735–752.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016a). Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016b). Using gps data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation research part C: emerging technologies*, 68:160–167.

Ba, J. and Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27*, pages 2654–2662.

Baecke, P. and Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98:69–79.

Barone, G. and Bella, M. (2004). Price-elasticity based customer segmentation in the Italian auto insurance market. *Journal of targeting, measurement and analysis for marketing*, 13(1):21–31.

Bivand, R. (2015). *classInt: Choose Univariate Class Intervals.* R package version 0.1-23.

Bordhoff, J. E. and Noel, P. J. (2008). Pay-as-you-drive auto insurance: A simple way to reduce driving-relayed harms and increase equity. Technical report, The Hamilton Project.

Boucher, J. P., Côté, S., and Guillén, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4):54.

Boucher, J. P., Pérez-Marín, A. M., and Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. In *Anales del Instituto de Actuarios Españoles*, volume 19, pages 135–154. Instituto de Actuarios Españoles.

Bracke, P., Datta, A., Jung, C., and Sen, S. (2019). *Machine learning explainability in finance: an application to default risk analysis.* Bank of England Working Paper No. 816.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, New York.

Bristow, R. E., Chang, J., Ziogas, A., Anton-Culver, H., and Vieira, V. M. (2014). Spatial analysis of adherence to treatment guidelines for advanced-stage ovarian cancer and the impact of race and socioeconomic status. *Gynecologic oncology*, 134(1):60–67.

Browning, E. K. and Zupan, M. A. (2020). *Microeconomics: Theory and applications.* John Wiley & Sons.

Buchner, F., Wasem, J., and Schillo, S. (2017). Regression trees identify relevant interactions: can this improve the predictive performance of risk adjustment? *Health economics*, 26(1):74–85.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge* *discovery and data mining*, pages 535–541.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Charpentier, A. (2014). General insurance pricing. In *Computational Actuarial Science with R*, pages 507–542. Chapman and Hall/CRC, New York.

Chen, K., Huang, L., Zhou, L., Ma, Z., Bi, J., and Li, T. (2015). Spatial analysis of the effect of the 2010 heat wave on stroke mortality in nanjing, china. *Scientific reports*, 5:10816.

Click, C., Malohlava, M., Parmar, V., Candel, A., and Roark, H. (2021). *Gradient boosting machine with H2O*.

Craven, M. and Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305.

Dal Pozzolo, A. (2010). Comparison of data mining techniques for insurance claim prediction (MSc thesis). *Università degli Studi di Bologna.*

De Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data.* Cambridge University Press, Cambridge.

Denuit, M., Guillén, M., and Trufin, J. (2019a). Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, 13(2):378–399.

Denuit, M., Hainaut, D., and Trufin, J. (2019b). *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions.* Springer.

Denuit, M. and Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627–647.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems.* John Wiley & Sons Ltd, West Sussex.

Desyllas, P. and Sako, M. (2013). Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. *Research Policy*, 42(1):101–

116.

Dionne, G., Gouriéroux, C., and Vanasse, C. (1999). Evidence of adverse selection in automobile insurance markets. In *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, pages 13–46. Springer, New York.

Doran, D., Schulz, S., and Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Dougherty, J., Kohavi, R., Sahami, M., et al. (1995). Supervised and unsupervised discretization of continuous features. *Machine learning: proceedings of the twelfth international conference*, 12:194–202.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive Theory of Functions of Several Variables*, pages 85–100.

Dutang, C. and Charpentier, A. (2019). *CASdatasets: Insurance datasets*. R package version 1.0.10.

ECOA (1974). U.S. Code Title 15. Commerce and Trade. *Chapter 41. Consumer Credit Protection*, Subchapter IV. Equal Credit Opportunity (Section 1691).

Eling, M. and Kraft, M. (2020). The impact of telematics on the insurability of risks. *The Journal of Risk Finance*, 21:77–109.

Ellison, A. B., Bliemer, M. C. J., and Greaves, S. P. (2015). Evaluating changes in driver behaviour: a risk profiling approach. *Accident Analysis & Prevention*, 75:298–309.

Farkas, S., Lopez, O., and Thomas, M. (2020). Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *The R journal*, 2(1):53–58.

Ferrario, A., Noll, A., and Wüthrich, M. V. (2018). Insights from inside neural networks.

Ferreira, J. and Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation research record*, 2297(1):97–103.

Filipova-Neumann, L. and Welzel, P. (2010). Reducing asymmetric information in insurance markets: Cars with black boxes. *Telematics and Informatics*, 27(4):394–403.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798.

Frees, E. W. (2015). Analytics of insurance markets. *Annual Review of Financial Economics*, 7:253–277.

Frees, E. W., Derrig, R. A., and Meyers, G. (2014). Predictive modeling in actuarial science. *Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques*.

Frees, E. W., Meyers, G., and Cummings, A. D. (2013). Insurance ratemaking and a Gini index. *Journal of Risk and Insurance*, 81(2):335–366.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, 103(484):1457–1469.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer, New York.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954.

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., and Taly, A. (2019). Explainable AI in industry. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3203–3204.

Gao, G., Meng, S., and Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2019(2):143–162.

Gao, G. and Wüthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. *European Actuarial Journal*, 8(2):383–406.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205–215.

GDPR (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. *O.J. (L 119)*, 1:1–88.

Gillespie, A. (2014). *Foundations of economics*. Oxford University Press, USA.

Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Cuppini, Bologna.*

Goldburd, M., Khare, A., and Tevet, D. (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, 5.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Greenberg, A. (2009). Designing pay-per-mile auto insurance regulatory incentives. *Transportation research part D: transport and environment*, 14(6):437–445.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2019). *gbm: Generalized Boosted Regression Models.* R package version 2.1.6.

Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, 9(1):421–436.

Grubinger, T., Zeileis, A., and Pfeiffer, K.-P. (2014). evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software*, 61(1):1–29.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 2007(3):202–225.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667.

Guelman, L. and Guillén, M. (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, 41(2):387–396.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42.

Guillén, M., Nielsen, J. P., Ayuso, M., and Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk analysis*, 39(3):662–672.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, Tech. rep.

Haberman, S. and Renshaw, A. (1997). Generalized linear models and actuarial science. *Insurance Mathematics and Economics*, 2(20):142.

Hall, P., Gill, N., Kurka, M., and Phan, W. (2017). *Machine learning interpretability with H2O Driverless AI.* H2O.ai.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models.* CRC Press.

He, B., Zhang, D., Liu, S., Liu, H., Han, D., and Ni, L. M. (2018). Profiling driver behavior for personalized insurance pricing and maximal profit. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1387–1396. IEEE.

Henckaerts, R. (2020). *distRforest: Distribution-based Random Forest.* R package version 1.0.0.

Henckaerts, R. (2021). *maidrr: Model-Agnostic Interpretable Data-driven suRRogate.* R package version 1.0.0.

Henckaerts, R. and Antonio, K. (2021). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Working paper.*

Henckaerts, R., Antonio, K., Clijsters, M., and Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, 2018(8):681–705.

Henckaerts, R., Antonio, K., and Côté, M.-P. (2021a). When stakes are high: balancing accuracy and transparency with Model-Agnostic Interpretable Data-driven suRRogates. *arXiv preprint arXiv:2007.06894.*

Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021b). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531.*

Holvoet, F., Henckaerts, R., Antonio, K., and Gielis, S. (2021). Neural networks for non-life insurance pricing. *Working paper.*

Hrnjica, B. and Softic, S. (2020). Explainable AI in manufacturing: A predictive maintenance case study. In *Advances in Production Management Systems. Towards Smart and Digital Manufacturing*, pages 66–73. Springer.

Hu, L., Chen, J., Nair, V. N., and Sudjianto, A. (2020). Surrogate locally-interpretable models with supervised machine learning algorithms. *arXiv preprint arXiv:2007.14528.*

Huang, Y. and Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, 127:113156.

Husnjak, S., Peraković, D., Forenbacher, I., and Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, 100:816–825.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154.

Kaminski, M. E. (2018). The right to explanation, explained. *Berkeley Technology Law Journal*, 34(1).

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data. an introduction to cluster analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Klein, N., Denuit, M., Lang, S., and Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225–249.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Hoboken.

Krasheninnikova, E., García, J., Maestre, R., and Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering Applications of Artificial Intelligence*, 80:8–19.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling.* Springer, New York.

Laugel, T., Renard, X., Lesot, M. J., Marsala, C., and Detyniecki, M. (2018). Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498.*

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., and Malohlava, M. (2020). *h2o: R Interface for the H2O Scalable Machine Learning Platform.* R package version 3.32.0.1.

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance.* Huebner international series on risk, insurance and economic security. Springer science & business media, New York.

Lemaire, J., Park, S. C., and Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin: The Journal of the IAA*, 46(1):39–69.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Litman, T. (2004). Pay-as-you-drive pricing for insurance affordability. Technical report, Victoria Transport Policy Institute.

Litman, T. (2011). Distance-based vehicle insurance feasibility, costs and benefits: Comprehensive technical report. Technical report, Victoria Transport Policy Institute.

Liu, Y., Wang, B., and Lv, S. (2014). Using multi-class AdaBoost tree for prediction frequency of auto insurance. *Journal of Applied Finance and Banking*, 4(5):45.

Longhi, L. and Nanni, M. (2020). Car telematics big data analytics for insurance and innovative mobility services. *Journal of Ambient Intelligence and Humanized Computing*, 11:3989—3999.

Lopez, O., Milhaud, X., and Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10(2):2685–2716.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774.

Ma, Y. L., Zhu, X., Hu, X., and Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, 113:243–258.

Maas, P., Graf, A., and Bieck, C. (2008). Trust, transparency and technology. european customers' perspectives on insurance and innovation. Technical report, IBM Institute for Business Value and I.VW-HSG.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297.

Mayou, R., Bryant, B., and Duthie, R. (1993). Psychiatric consequences of road traffic accidents. *British Medical Journal*, 307(6905):647–651.

Mayou, R., Simkin, S., and Threlfall, J. (1991). The effects of road traffic accidents on driving behaviour. *Injury*, 22(5):365–368.

McClenahan, C. L. (2001). *Ratemaking.* Casualty Actuarial Society, Fourth edition.

Meteier, Q., Capallera, M., Angelini, L., Mugellini, E., Khaled, O. A., Carrino, S., De Salis, E., Galland, S., and Boll, S. (2019). Workshop on explainable ai in automated driving: a user-centered interaction approach. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 32–37.

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable.* Leanpub.

Molnar, C., Bischl, B., and Casalicchio, G. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, 3(26):786.

NAIC (2012). *Model 777 - Guideline 1775 - Guideline 1780 - Product filing review handbook.*

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Nykodym, T., Kraljevic, T., Wang, A., and Wong, W. (2016). *Generalized linear modeling with H2O.* Published by H2O.ai.

OECD (2020). *The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector.*

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, 2008(4):301–314.

Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models.* Springer, Berlin.

OJ/C11 (13.1.2012). Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). *OJ*, C11:1–11.

OJ/L123 (19.5.2015). Regulation (EU) 2015/758 of the European Parliament and of the Council of 29 April 2015 concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC. *OJ*, L123:77–89.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group, New York.

Paefgen, J., Staake, T., and Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61:27–40.

Paefgen, J., Staake, T., and Thiesse, F. (2013). Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems*, 56:192–201.

Parodi, P. (2014). *Pricing in General Insurance.* Chapman and Hall/CRC, New York.

Parry, I. W. H. (2005). Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *American Economic Review*, 95(2):288–293.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press, Cambridge.

Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, 7(2):70.

Pitera, K., Boyle, L. N., and Goodchild, A. V. (2013). Economic analysis of onboard monitoring systems in commercial vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, 2379(1):64–71.

PRIIPs (2014). Regulation (EU) 1286/2014 of the European Parliament and of the Council of 26 November 2014 on key information documents for packaged retail and insurance-based investment products. *O.J. (L 352)*, 1:1–23.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, volume 18, pages 1527–1535. AAAI.

Ridgeway, G. (2014). *gbm: Generalized Boosted Regression Models.* R package version 2.1-06.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Schelldorfer, J. and Wüthrich, M. V. (2019). Nesting classical actuarial models into neural networks.

Schiltz, F., Masci, C., Agasisti, T., and Horn, D. (2018). Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, 50(58):6341–6354.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Sherden, W. A. (1984). An analysis of the determinants of the demand for automobile insurance. *Journal of Risk and Insurance*, pages 49–62.

Slocum, T., McMaster, R., Kessler, F., and Howard, H. (2005). *Thematic cartography and geographic visualization.* Upper Saddle River, New Jersey: Pearson Prentice Hall.

So, B., Boucher, J. P., and Valdez, E. A. (2020). Cost-sensitive multi-class AdaBoost for understanding driving behavior with telematics. *arXiv preprint arXiv:2007.03100.*

Song, J. (2019). *Ckmeans.1d.dp: Optimal, fast, and reproducible univariate clustering.* R package version 4.3.0.

Southworth, H. (2015). *gbm: Generalized Boosted Regression Models.* R package version 2.1-06.

Spedicato, G. A., Dutang, C., and Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. *Variance*, 12(1):69–89.

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Therneau, T., Atkinson, B., and Ripley, B. (2019). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-15.

Therneau, T. M. and Atkinson, E. J. (2019). An introduction to recursive partitioning using the rpart routines (vignette).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(1):234–240.

Toledo, T., Musicant, O., and Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies*, 16(3):320–331.

Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2016). Innovative insurance schemes: pay as/how you drive. *Transportation Research Procedia*, 14:362–371.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2021). Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808.*

Venables, W. N. and Ripley, B. D. (2002). Tree-based methods. In *Modern Applied Statistics with S*, pages 251–269. Springer, New York.

Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304.

Vickrey, W. (1968). Automobile accidents, tort law, externalities, and insurance: An economist's critique. *Law and Contemporary Problems*, 33(3):464–487.

Vieira, V., Webster, T., Weinberg, J., Aschengrau, A., and Ozonoff, D. (2005). Spatial analysis of lung, colorectal, and breast cancer on cape cod: an application of generalized additive models to case-control data. *Environmental Health*, 4(1):11.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887.

Wahba, G. (1990). *Spline models for observational data.* Siam, Philadelphia.

Wang, H. and Song, M. (2011). Ckmeans.1d.dp: optimal K-means clustering in one dimension by dynamic programming. *The R journal*, 3(2):29.

Wang, Y. and Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95.

Weaver, K. F., Morales, V. C., Dunn, S. L., Godde, K., and Weaver, P. F. (2017). *An introduction to statistical analysis in research: with applications in the biological and life sciences.* John Wiley & Sons.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Wüthrich, M. V. (2016). Non-life insurance: mathematics & statistics.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1):89–108.

Wüthrich, M. V. (2020). From generalized linear models to neural networks, and back. *Available at SSRN 3491790.*

Wüthrich, M. V. and Buser, C. (2019). Data analytics for non-life insurance pricing (lecture notes).

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert*

Systems with Applications, 78:225–241.

Yang, L., Li, Z., and Meng, S. (2020). Risk loadings in classification ratemaking. *arXiv preprint arXiv:2002.01798*.

Yang, Y., Qian, W., and Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470.

Yeo, I. K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Zöchbauer, P., Wüthrich, M. V., and Buser, C. (2017). Data science in non-life insurance pricing (MSc thesis). *ETH Zürich*.