The added value of dynamically updating motor insurance prices with telematics collected driving behavior data

Roel Henckaerts, Katrien Antonio

The added value of dynamically updating motor insurance prices with telematics collected driving behavior data

Roel Henckaerts [*,a,c] and Katrien Antonio[a,b,c]

[a] *Faculty of Economics and Business, KU Leuven, Belgium.*
[b] *Faculty of Economics and Business, University of Amsterdam, The Netherlands.*
[c] *LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.*

Abstract

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.

Key words: Usage-based insurance, Pricing, Telematics, Driving behavior, Profits, Client retention

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 an accurate and fair 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). Customers 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 speeding).

[*]Corresponding author: roel.henckaerts@kuleuven.be, Department of Accounting, Finance and Insurance (AFI), Naamsestraat 69 - box 3525, Leuven 3000, Belgium. Declarations of interest: none.

The motor UBI literature discusses multiple potential benefits for 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). A 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 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, 13.1.2012). Ayuso et al. (2016a) explain women's lower accident risk by a lower driving intensity and less risky behavior compared to men. Ayuso et al. (2016b) 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, 19.5.2015). 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 speed-acceleration heatmaps and groups similar profiles via K-means clustering. Ma et al. (2018) study driving performance measures to assess the effect on claim occurrence and frequency with GLMs. He et al. (2018) use sensor data from a vehicle's on-board diagnostics (OBD) unit to compile driver profiles and to measure accident risk.

In this paper, 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. 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 telematics 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 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 show the added value of telematics via improved risk classification and we put focus on managerial insights by analyzing profits and retention rates under the new telematics paradigm. Our updating mechanism results in a true UBI system where the price of insurance coverage is adjusted to the actual vehicle use. 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. Guillén et al. (2021) use Poisson regression models to update a baseline premium with extra charges for near-miss events, recorded via telematics devices. To the best of our knowledge, this is one of the first papers to explore the full spectrum of pricing (frequency/severity) and driving behavior (habits/style) including a profit and retention analysis. From a different angle, Frees et al. (2021) study the association between pricing and customer loyalty with a copula model for longitudinal and time-to-event data.

The rest of this paper is structured as follows. Section 2 provides a description of the dataset and outlines our methodology. Section 3 details our baseline models for pricing and customer churn prediction. Section 4 describes how we update the baseline pricing model with telematics information, highlighting the improvement in risk classification and resulting price adjustments. Section 5 investigates the managerial impact of telematics pricing by analyzing profits and retention rates under various price elasticity settings. Section 6 concludes this paper.

2 Overview of our data structure and updating methodology

We analyze a novel motor third party liability (MTPL) portfolio followed over the years 2017, 2018 and 2019. Figure 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 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 5974, 9383 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.

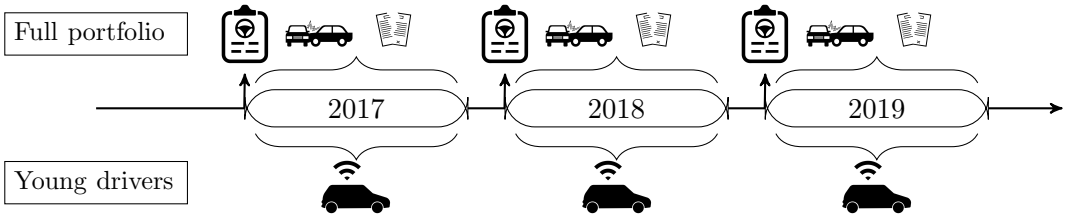

Figure 1: Timeline with policy, claim and telematics information over the years 2017 - 2019.

Sections 2.1 and 2.2 describe the policy and telematics data respectively. Section 2.3 investigates the presence of a selection effect and Section 2.4 outlines our price updating methodology.

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 1 lists the claim and policy information available in the portfolio. The policy information consists of self-reported risk characteristics about the driver(s), payment method, geographical location and insured vehicle. Figure 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 (in a non-obvious way) during the year. The middle panel indicates the number of claims filed by a policyholder. Most policyholders do not file a claim (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 4067 and 1259 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`	Contract condition specifying a 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 1: Description of the claim and policy information in the portfolio data.

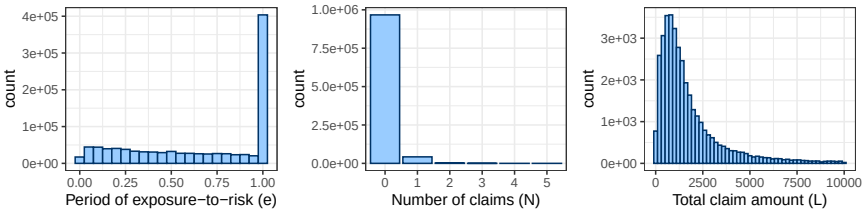

Figure 2: Distribution of the exposure period e (left), claim counts N (middle) and amounts L (right) in the combined training portfolios of 2017 and 2018.

2.2 Telematics data

Driving behavior data is available for a selection of young policyholders in the portfolio. Table 2 lists the information recorded by the telematics 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 telematics box registration period to a full year. Imagine a telematics 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 (estimated in case the telematics 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 2: Description of the available telematics data.

Figure 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 telematics 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 compositional 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).

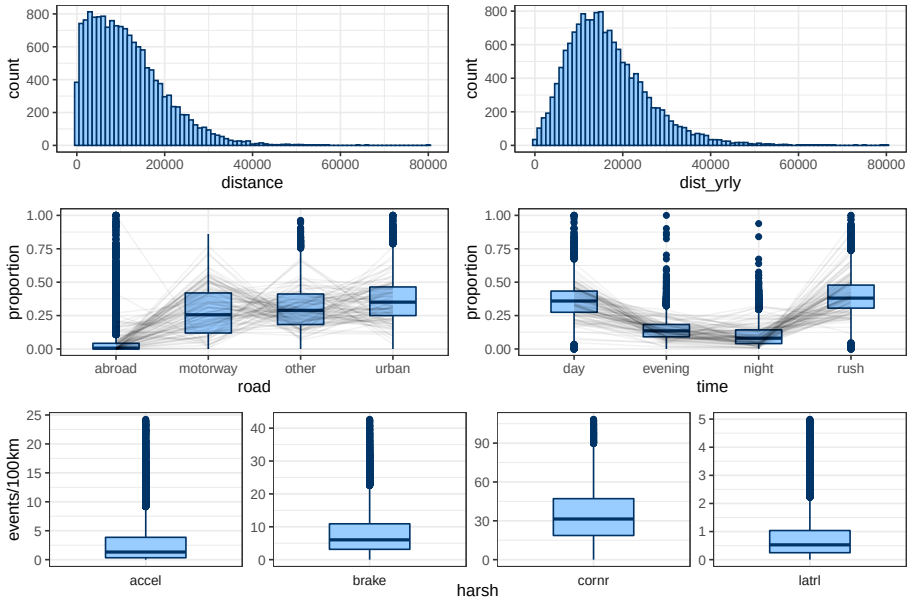

Figure 3: Distribution of the actual distance (top left), yearly distance (top right), road types (middle left), times of day (middle right) and harsh movement events (bottom) in the training data.

2.3 Selection effect

In our portfolio, the installation of a telematics box to record driving behavior is a choice offered to young drivers only. Figure 4 shows the age distribution for policyholders who have a telematics box installed (green) and those who do not (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 telematics 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).

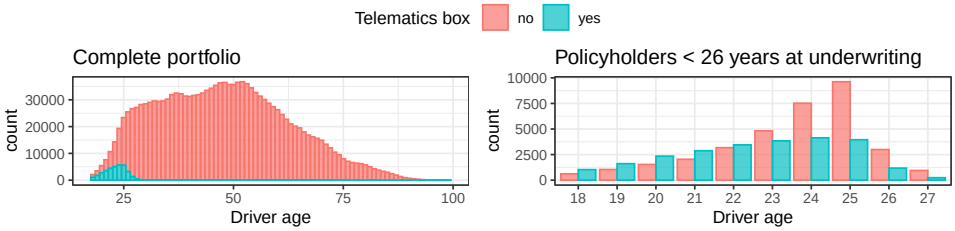

Figure 4: Age distribution for policyholders with/without (green/red) a telematics box installed.

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 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 3 and test the null hypothesis $H_0 : \mu_{\text{yes}} = \mu_{\text{no}}$ or equivalently $H_0 : \mu_{\text{no}}/\mu_{\text{yes}} = 1$. 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 $[0.900, 1.034]$.

telematics box	$\sum_i N_i$	$\sum_i e_i$	$\hat{\mu}$
No	1817	17,984.03	0.1010
Yes	1477	14,104.14	0.1047

Table 3: Claim risk statistics for young policyholders without and with a telematics box.

The empirical observation $\hat{\mu}_{\text{yes}} > \hat{\mu}_{\text{no}}$ might look counter-intuitive. However, the right panel of Figure 4 indicates that policyholders without a telematics box are older on average in our sample, and therefore typically less risky compared to younger ones. We test the presence of a selection effect by fitting the following Poisson GLM, investigating the effect of choosing for telematics via the dummy variable `tbox` while controlling for the driver's age `driv_age`:

$$\ln[\mathbb{E}(N)] = \ln[e] + \beta_0 + \beta_{\text{age}}\texttt{driv\_age} + \beta_{\text{box}}\texttt{tbox} + \beta_{\text{int}}\texttt{driv\_age} : \texttt{tbox}. \tag{1}$$

Table 4 shows the results with (left) and without (right) the interaction term included. The telematics box coefficient β_{box} is negative in both GLMs, indicating 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 telematics box, but also not significantly. Figure 5 shows the fitted GLM effects (lines), 95% confidence intervals (shades) and the empirical claim frequencies (points) by group (color).

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

	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	< 0.01
`driv_age`	-0.078	0.011	-6.85	< 0.01	-0.074	0.008	-9.11	< 0.01
`tbox`	-0.222	0.374	-0.59	0.55	-0.054	0.037	-1.48	0.140
`driv_age:tbox`	0.007	0.016	0.45	0.65	-	-	-	-

Table 4: Selection effect in a GLM for young policyholders with/without (left/right) the interaction.

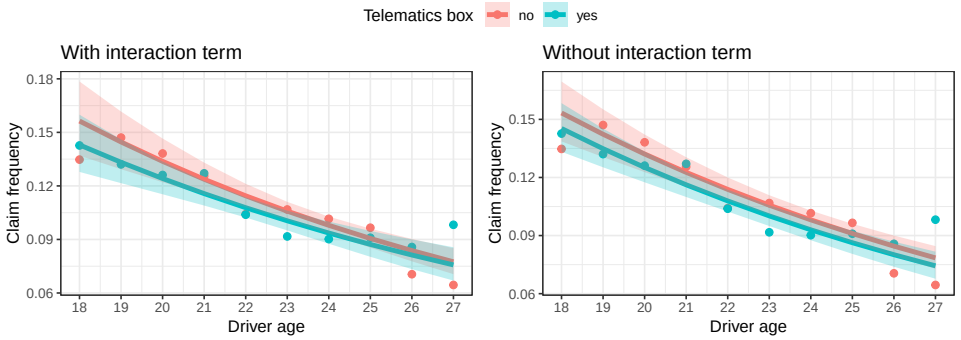

Figure 5: Age effect for young drivers with/without a box (green/red) and the interaction (left/right).

2.4 A methodology to update pricing

Figure 6 outlines our proposed updating mechanism 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: $\pi^*(\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{x}) \times \delta^{\pi}(\boldsymbol{z})$.

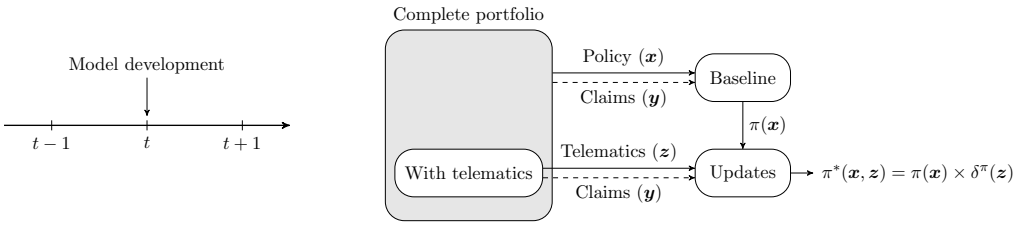

Figure 6: Methodology of our mechanism to update baseline premiums with telematics information.

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 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 $[t, t+1]$.

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 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 1 as features $\boldsymbol{x}$. We opt for stochastic gradient boosting machines or GBMs (Friedman, 2002) to determine the prediction function. This choice is based on the good performance of GBMs in insurance pricing (Henckaerts et al., 2021) and churn (Spedicato et al., 2018) applications.

Section 3.1 details the GBM development process. Section 3.2 proposes a slight adjustment that restores the balance between observed and predicted targets. Section 3.3 provides insights into the optimal GBMs. The frequency and severity GBMs are used in Section 4 as a baseline pricing model, while the churn GBM is used in Section 5 as baseline for retention rates.

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 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 6 are obtained via a random grid search and 5-fold cross-validation on the combined training portfolios of 2017 and 2018.

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

Table 5: Summary of the distributional assumptions for claim frequency, severity and customer churn.

	ntrees	learn_rate	max_depth	sample_rate	col_sample_rate
Claim frequency	4700	0.02	4	1.0	0.6
Claim severity	3900	0.01	1	0.5	0.7
Customer churn	4100	0.02	5	0.7	0.6

Table 6: Optimal settings of the GBM tuning parameters for the different predictive models.

3.2 Balance property

Calibrating the regression parameters in a GLM with canonical link via maximum likelihood estimation (MLE) 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 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 7: Biased ($\sum \hat{y}_i / \sum y_i$) and balanced ($\sum \hat{y}_i^b / \sum y_i$) ratios for the frequency, severity and premium.

3.3 Insights in the optimal GBMs

Table 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 is insightful to predict 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.

	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 8: The most important features in the training data for the frequency, severity and churn GBMs.

Figure 7 shows partial dependence (PD) effects (Friedman, 2001) for the highlighted features in Table 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 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.

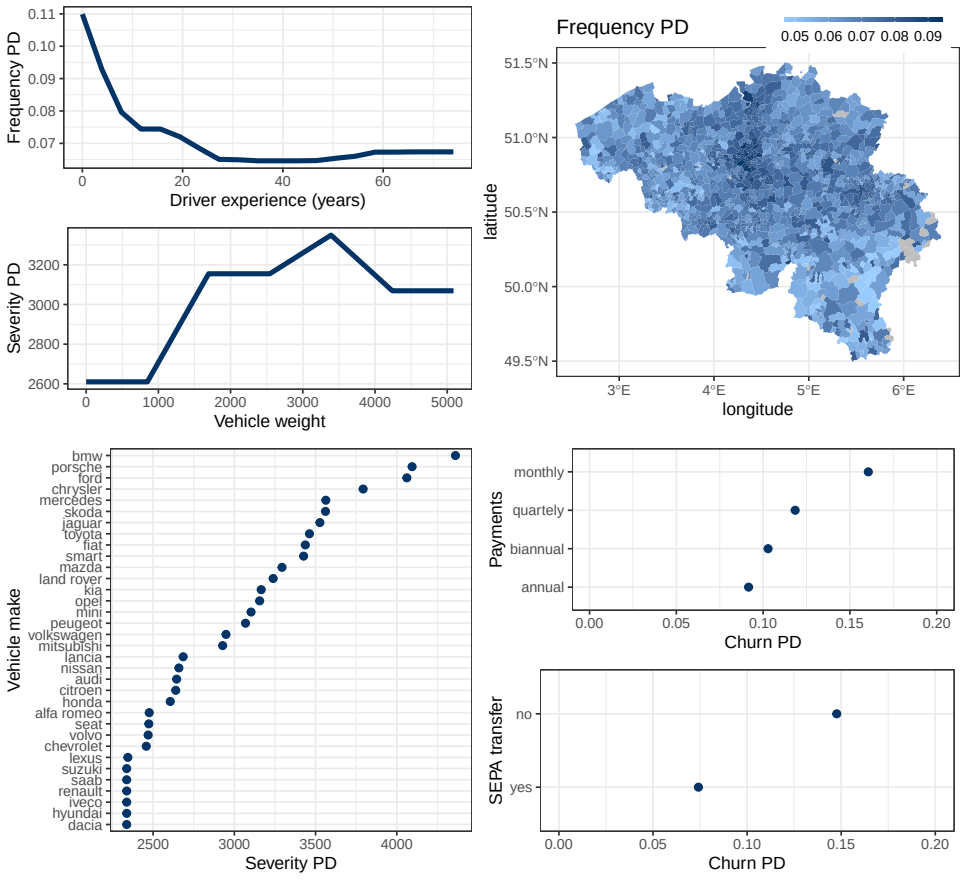

Figure 7: PD effect for driving experience (top left) and postal code (top right) in the claim frequency GBM, the vehicle's make (bottom left) and weight (middle left) in the severity GBM and the payment frequency (middle right) and SEPA indicator (bottom right) in the churn GBM.

4 Towards a usage based pricing mechanism: updating the baseline tariff with driving behavioral information

Our goal is to update the baseline pricing structure, consisting of the combined frequency and severity GBMs developed in Section 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 1 and telematics information $\boldsymbol{z}$ from Table 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 from 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 (i.e., term with a coefficient fixed to one) in the linear predictor is as follows:

$$\begin{aligned} \ln[\mathbb{E}(y \,|\, \boldsymbol{x}, \boldsymbol{z})] &= \ln[f(\boldsymbol{x})] + \beta_0 + \sum_{j=1}^{p} \beta_j z_j \\ \mathbb{E}(y \,|\, \boldsymbol{x}, \boldsymbol{z}) &= f(\boldsymbol{x}) \times \exp(\beta_0) \times \prod_{j=1}^{p} \exp(\beta_j z_j) \end{aligned} \tag{2}$$

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 that the target y represents N and L/N, while $f(\boldsymbol{x})$ equals $\mathbb{E}(N \,|\, \boldsymbol{x}, e)$ and $\mathbb{E}(L/N \,|\, \boldsymbol{x})$ for the frequency and severity GBM respectively. Figure 8 visualizes our updating methodology, applied to the claim frequency (left) and severity (right) components.

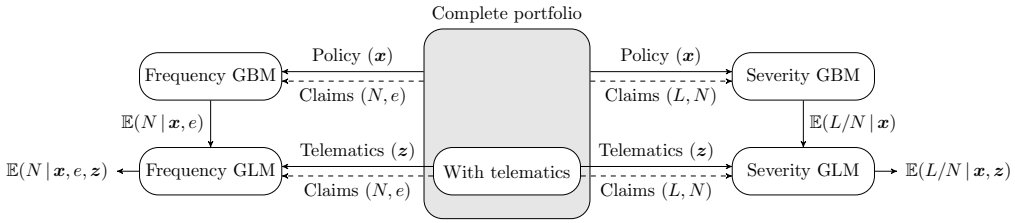

Figure 8: Methodology of our mechanism to update baseline premium components with telematics.

Our proposed updating mechanism in Equation (2) allows for intuitive explanations about the impact of telematics data on the price, since the final prediction is multiplicative in the following 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.

The updated GLM predictions satisfy the balance property, as described in Section 3.2, while we deliberately enforce this property in the baseline GBMs for young drivers. This implies that the multiplicative adjustments result in a pure redistribution of risk in the updated models.

We perform a feature selection procedure to unravel the effect of driving behavior on claim risk in Section 4.1. Focusing on the most informative features, we develop our explainable updating mechanism in Section 4.2. Finally, we highlight the added value of telematics for risk classification in Section 4.3.

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 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 (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 (2) and the distributional assumptions outlined in Table 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. We can therefore assume that those features are most informative and reliable to update the baseline predictions. Figure 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 incorporate telematics information in the pure premium solely via the claim frequency component. The LASSO procedure on the full training data without sampling, and with the "one standard error rule" (Hastie et al., 2009), leads to the same feature selection results for frequency and severity.

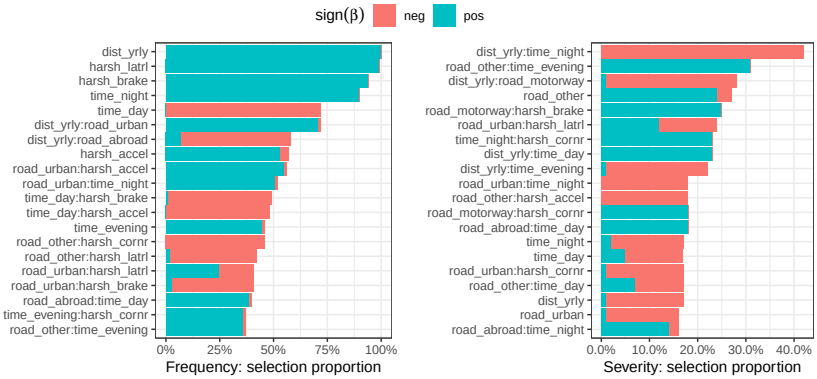

Figure 9: Feature selection proportions in the 100 LASSO GLM experiments for claim frequency (left) and severity (right), where red/green indicates a negative/positive β coefficient if selected.

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:

$$\begin{aligned} \ln[\mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z}^*)] &= \ln[\mathbb{E}(N \mid \boldsymbol{x}, e)] + \beta_0 + \sum_{j=1}^{4} \beta_j \log(z_j^* + 1) \\ \mathbb{E}(N \mid \boldsymbol{x}, e, \boldsymbol{z}^*) &= \mathbb{E}(N \mid \boldsymbol{x}, e) \times \exp(\beta_0) \times \prod_{j=1}^{4} (z_j^* + 1)^{\beta_j}. \end{aligned} \tag{3}$$

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 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 (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 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.

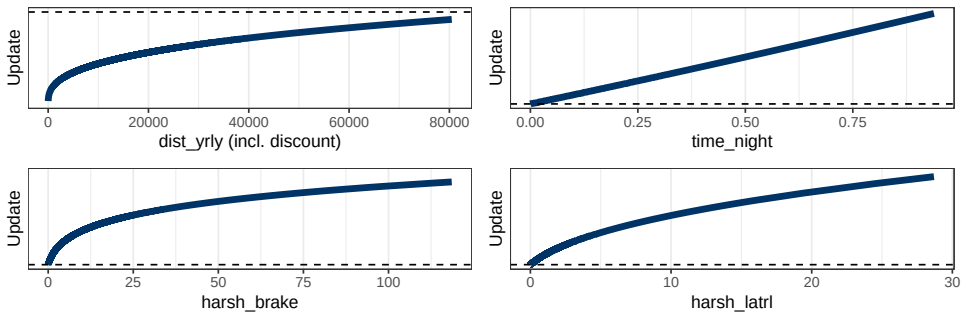

Figure 10: Multiplicative update effects for the mileage including the fixed discount (top left), night-time driving (top right), harsh braking (bottom left) and lateral movements (bottom right).

Figure 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 (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 we discuss how to transform this technical analysis into a commercial UBI product with update limits on discounts and penalties.

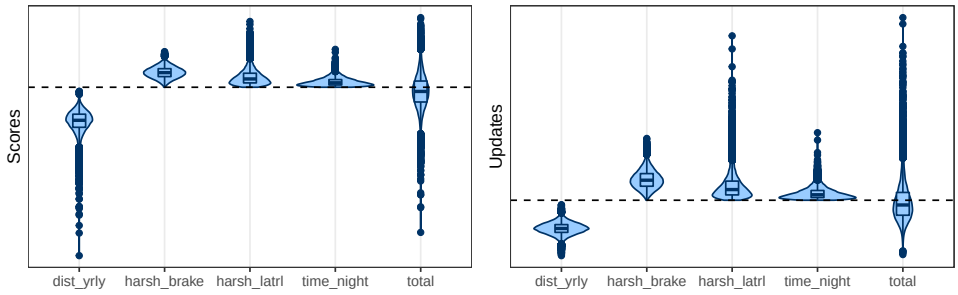

Figure 11: Distribution of scores $\beta_j \log(z_j^* + 1)$ (left) and updates $(z_j^* + 1)^{\beta_j}$ (right) in the train data.

Figure 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.

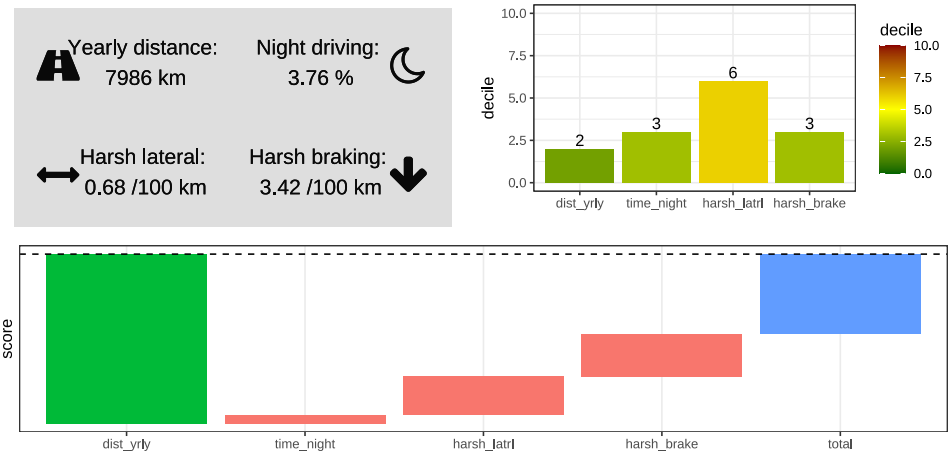

Figure 12: Dashboard with recorded driving information (top left), ranking within the portfolio (top right) and influence of each component on the final price (bottom).

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 analyzes the effects on an insurer's profits and retention rates.

Table 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 (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 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):

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

The Lorenz curve accumulates observed claims from low to high risks as perceived by model m (i.e., $r_i^m : 0 \to 1$). Better risk classification means that claims accumulate at a slower/faster rate for low/high values of r_i^m. Figure 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:

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

Figure 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 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.

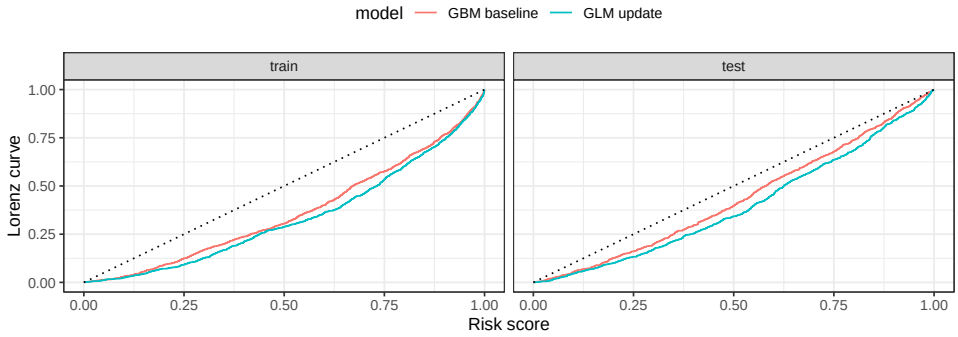

Figure 13: Lorenz curves for the GBM (red) and GLM (green) on the train (left) and test (right) data.

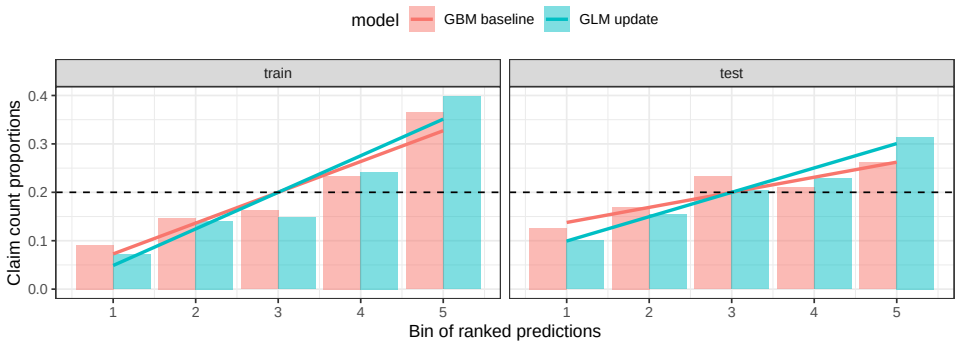

Figure 14: Claim bins for the GBM (red) and GLM (green) on the train (left) and test (right) data.

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 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 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 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 5.1 details our assumptions regarding changes in the churn probability following price updates via the price elasticity of demand. Section 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 5.3 we optimize the product design for maximal profits under retention constraints and for maximum retention under profitability constraints.

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 to 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 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.

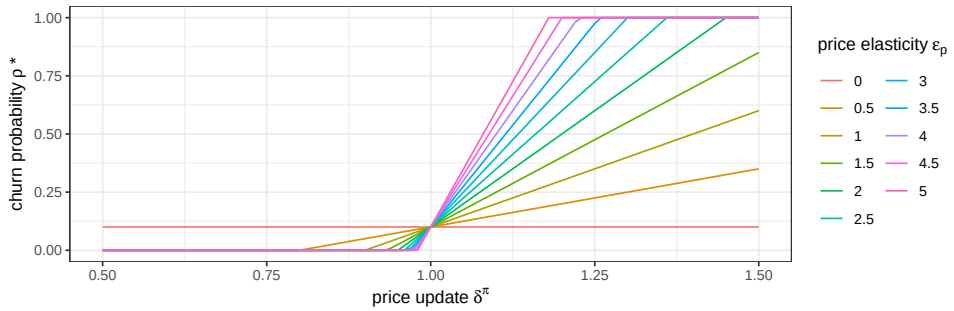

Figure 15: Effect of price updates δ^π on the churn probability ρ^* for a price elasticity $\epsilon_p \in [0, 5]$.

5.2 Profits and retention rates with fairness constraints

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

$$P = \frac{1}{n}\sum_{i=1}^{n}(1-(\rho_i+\delta_i^{\rho}))\cdot(\delta_i^{\pi}\pi_i - L_i) \qquad \text{and} \qquad R = \frac{1}{n}\sum_{i=1}^{n} 1-(\rho_i+\delta_i^{\rho}). \tag{4}$$

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 (4). This baseline results in profits of 12.45 Euro per policyholder and a retention rate of 90.85%. To allow for a fair and realistic comparison of telematics versus the baseline, we propose a solidarity/commercial constraint via update limits and a redistribution constraint via a scale factor α:

$$\delta_{lo}^{\pi} \leq \delta^{\pi} \leq \delta_{hi}^{\pi} \qquad \text{and} \qquad \sum_{i=1}^{n}(1-\rho_i)\cdot\pi_i = \sum_{i=1}^{n}(1-\rho_i)\cdot\alpha\cdot\delta_i^{\pi}\cdot\pi_i. \tag{5}$$

Figure 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) 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) 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 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 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 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 on the total premium inflow.

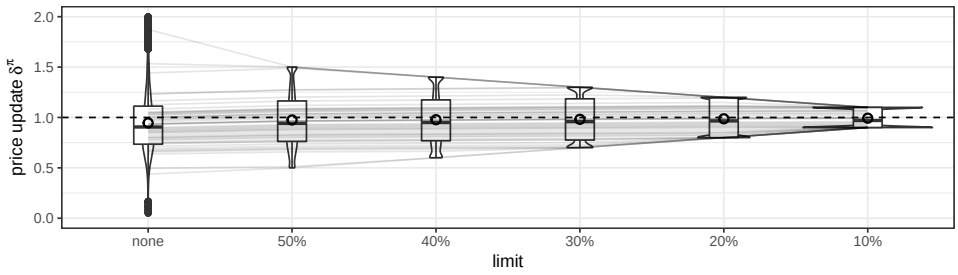

Figure 16: Distribution of the price updates δ^π for different symmetrical lower and upper limits.

		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 10: Statistics on updates δ^π and prices $\delta^\pi\pi$ for different symmetrical lower and upper limits.

Figure 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 an 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 by the median values in Table 10. 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.

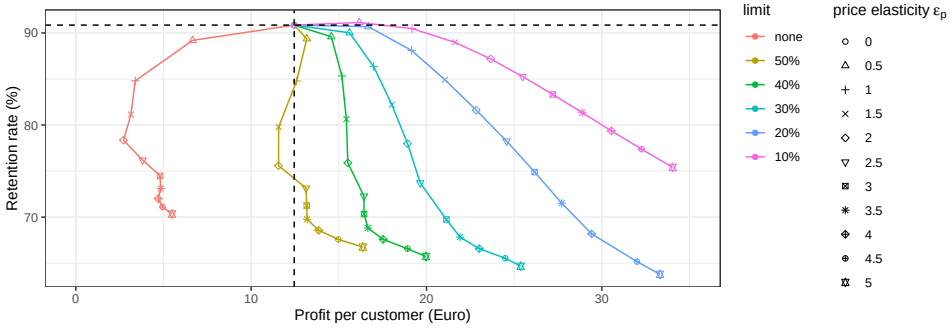

Figure 17: Profit and retention rate by values of the update limit (color) and price elasticity ϵ_p (shape).

5.3 Constrained optimization for 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:

$$\max_{\alpha} P(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(1-(\rho_i+\delta_i^{\rho}))\cdot(\alpha\delta_i^{\pi}\pi_i - L_i) \quad \text{subject to} \quad R(\alpha) = \frac{1}{n}\sum_{i=1}^{n} 1-(\rho_i+\delta_i^{\rho}) \geq R^*. \quad (6)$$

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 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 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.

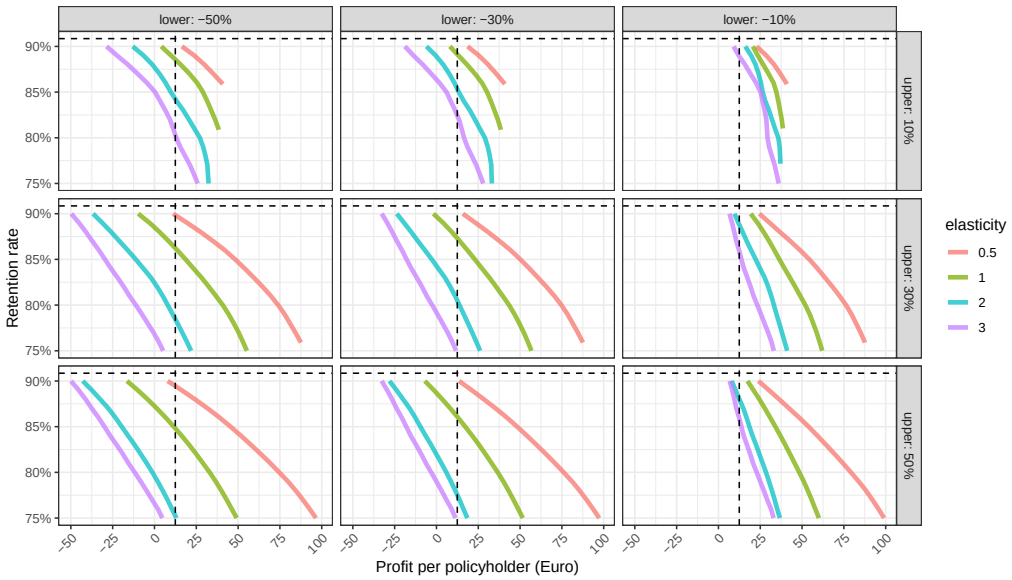

Figure 18: Profits and retention rates by values of the update limit (grid) and price elasticity (color).

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:

$$\max_{\alpha} R(\alpha) = \frac{1}{n}\sum_{i=1}^{n} 1-(\rho_i+\delta_i^{\rho}) \quad \text{subject to} \quad P(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(1-(\rho_i+\delta_i^{\rho}))\cdot(\alpha\delta_i^{\pi}\pi_i - L_i) \geq P^*. \quad (7)$$

Again, the churn update δ^{ρ} implicitly depends on α. Figure 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 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 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.

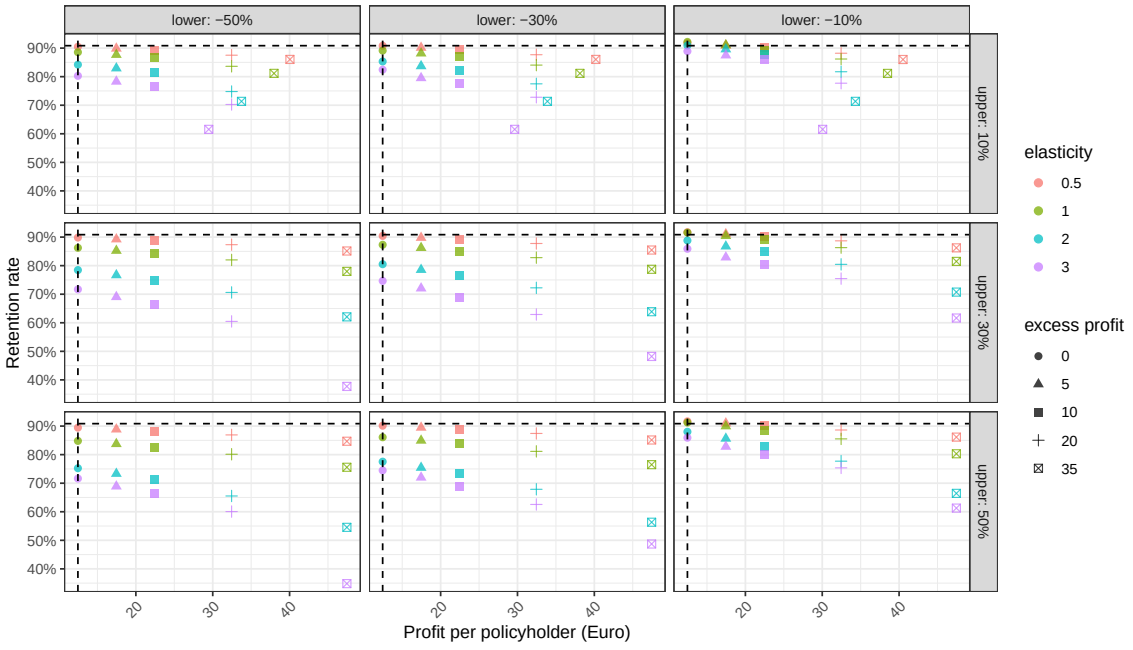

Figure 19: Profits and retention rates by values of the limit (grid), elasticity (color) and P^* (shape).

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.

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. In this paper we combine both worlds. We first develop a baseline 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. 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. In this paper, 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 may consider the development of 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.

Another direction for future work is in the connection between the churn and pricing models. 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. Yet 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, 2012). These systems are common in the European insurance market and result in a fixed discount/penalty for the next period. 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.

Funding

This research is supported by the Research Foundation Flanders [SB grant 1S06018N].

References

M. Ayuso, M. Guillén, and A. M. Pérez-Marín. 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, 2016a.

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

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

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

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

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

J. P. Boucher, A. M. Pérez-Marín, and M. Santolino. 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, 2013.

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

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

C. Click, M. Malohlava, V. Parmar, A. Candel, and H. Roark. *Gradient boosting machine with H2O*, 2021. URL https://www.h2o.ai/resources/booklet/gradient-boosting-machine-with-h2o/.

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

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

P. Desyllas and M. Sako. Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. *Research Policy*, 42(1):101–116, 2013.

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

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

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

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

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

E. W. Frees, C. Bolancé, M. Guillén, and E. A. Valdez. Dependence modeling of multivariate longitudinal hybrid insurance data with dropout. *Expert Systems with Applications*, 185(115552):1–11, 2021.

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

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

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

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

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

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

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

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

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

M. Guillén, J. P. Nielsen, and A. M. Pérez-Marín. Near-miss telematics in motor insurance. *Journal of Risk and*

Insurance, 88(3):569–589, 2021.

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

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

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

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

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

E. LeDell, N. Gill, S. Aiello, A. Fu, A. Candel, C. Click, T. Kraljevic, T. Nykodym, P. Aboyoun, M. Kurka, and M. Malohlava. *h2o: R Interface for the H2O Scalable Machine Learning Platform*, 2020. URL https://CRAN.R-project.org/package=h2o. R package version 3.32.0.1.

J. Lemaire. *Bonus-malus systems in automobile insurance*, volume 19. Springer science & business media, 2012.

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

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

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

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

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

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

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

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

OJ/C11. 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, 13.1.2012.

OJ/L123. 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, 19.5.2015.

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

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

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

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

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

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

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

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

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

D. I. Tselentis, G. Yannis, and E. I. Vlahogianni. Innovative insurance schemes: pay as/how you drive. *Trans-*

portation Research Procedia, 14:362–371, 2016.

R. Verbelen, K. Antonio, and G. Claeskens. 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, 2018.

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

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

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

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