LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER OF SCIENCE IN
ACTUARIAL SCIENCE

MASTERS FINAL WORK
INTERNSHIP REPORT

MODELLING RENEWAL PRICE ELASTICITY:
AN APPLICATION TO THE MOTOR PORTFOLIO OF OCIDENTAL

GUILHERME FILIPE PALMA MOUSINHO

OCTOBER - 2016

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Master of Science in
Actuarial Science

Masters Final Work

Internship Report

Modelling Renewal Price Elasticity:
An Application to the Motor Portfolio of Ocidental

Guilherme Filipe Palma Mousinho

Supervisors:

Prof. Doutor José Manuel de Matos Passos
Dr. João Filipe Azevedo dos Santos

October - 2016

Acknowledgements

I would like to thank my supervisor at Ocidental, Filipe Santos, for his indispensable guidance and availability during this project, along with introducing me to the captivating world of machine learning.

I express my gratitude to my supervisor at ISEG, Professor José Passos, for his availability and care for my work.

I am thankful to Professor Lourdes Centeno, for being available to help me secure this internship.

I am indebted to Sjoerd Smeets and André Rufino, for providing me with the opportunity of doing a curricular internship at Ocidental, as well as to the whole Non-Life Pricing & Business Analytics team, for making me feel like part of it since day one.

I am grateful to Paula Santos, Diana Duarte and Tiago Cavaleiro, for being available to answer all my questions on the Motor business and much more.

I would also like to thank Rita Costa, for helping me with the inner workings of the company's database during the early stages of this project.

Last but not least, I thank my family and friends, for all their support.

Abstract

The increase in competition in the Portuguese Motor insurance market has lead insurers to consider a more demand-based approach to ratemaking, as a complement to the usual risk-based approach. Insurance companies now want to have a better understanding of who their clients are, how they behave, and what actions can insurers take, during the policy renewal period, in order to prevent their clients from leaving while maintaining profitability.

This report is the result of a curricular internship that took place at Ocidental Seguros, with the main goals of modelling the company's Motor insurance lapse rate during the renewal period and studying how different covariates influence renewals. We considered logistic regression, a special case of Generalized Linear Models, to model the binary response variable renewal/lapse.

By modelling the response as a function of premium change and other covariates, the lapse probability for each client per amount of premium variation can then be estimated. As premium change is the only covariate the company has direct control over, obtaining such knowledge on each client's price elasticity will allow the insurer to make better decisions, so that a finer balance between customer satisfaction and profitability can be achieved.

The model's capacity to predict which clients will cancel their policy was also analysed. In order to transform the output probabilities into binary classifications, several threshold optimisation criteria were compared, to find the threshold generating the best overall discriminatory performance.

Keywords: Motor insurance; Lapse rate; Renewal price elasticity; Logistic regression; Binary classification

Resumo

O aumento da competitividade no mercado segurador automóvel em Portugal tem levado as seguradoras a considerar uma abordagem de tarifação mais assente na procura, como um complemento à tradicional abordagem baseada no risco. As companhias de seguros querem actualmente saber mais sobre quem são os seus clientes, como estes se comportam e que medidas podem as seguradoras tomar, durante o período de renovação de apólice, de modo a evitar a saída dos seus clientes sem prejudicar a rentabilidade.

Este relatório é o resultado de um estágio curricular que teve lugar junto da Ocidental Seguros, tendo como principais objectivos modelar a taxa de anulação na renovação do seguro automóvel da companhia e analisar como diversas variáveis influenciam as renovações. Considerámos a regressão logística, um caso particular dos Modelos Lineares Generalizados, para modelar a variável de resposta binária renovação/anulação.

Modelando a variável de resposta como uma função da variação do prémio e de outras variáveis explicativas, é possível estimar a probabilidade de anulação por valor da alteração do prémio para cada cliente. Como a variação do prémio é a única variável que a companhia pode controlar directamente, obter tal informação sobre a elasticidade preço de cada cliente permitirá à seguradora tomar melhores decisões, com o objectivo de aperfeiçoar o equilíbrio entre o grau de satisfação dos clientes e a rentabilidade.

A capacidade do modelo em prever que clientes irão anular as suas apólices foi também examinada. Para converter as probabilidades obtidas pelo modelo em classificações binárias, foram comparados vários critérios de optimização de ponto de corte, de modo a encontrar o valor que resulta na melhor capacidade discriminatória global.

Palavras-chave: Seguro automóvel; Taxa de anulação; Elasticidade preço nas renovações; Regressão logística; Classificação binária

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This Masters Final Work is the result of a curricular internship, which took place between February and June 2016 at Ocidental Seguros.

When non-life insurance companies develop their pricing strategy, two crucial moments in their relationship with their clients are taken into account. The first one is the moment of risk acquisition and signing the contract (conversion). The second one, which is the focus of this work, is the moment to renew that contract (renewal).

As in most countries, Motor Third-Party Liability insurance is mandatory in Portugal, which has lead to a very competitive Motor insurance market where the average premium per vehicle systematically decreased over 10 years up to 2014 (Associação Portuguesa de Seguradores, 2015). Consequently, very different sets of prices are available to customers, meaning that, besides the usual risk-based approach to ratemaking, insurers also need to be very careful with the premium variations presented to policyholders, when the time comes to renew their contract. In other words, the increasing level of competition has compelled insurance companies to start looking into a more demand-based approach to pricing, by trying to understand who their clients are, how they behave, and what actions can insurers take, during the policy renewal period, in order to prevent their clients from leaving while maintaining profitability.

It was under this setting that this work emerged, having as its two main goals the modelling of the company's Motor insurance lapse rate and studying how different covariates influence renewals. By modelling the binary response variable renewal/lapse as a function of premium change and other covariates, the lapse probability for each client per amount of premium variation can then be estimated.

As premium change is the only covariate the company has direct control over, such a model can help the company understand how price elasticity varies from client to client and what amount of premium change is most adequate for its customers. Obtaining this knowledge would allow the insurer to make better decisions in order to achieve a finer balance between customer satisfaction and profitability.

Furthermore, the model can also be adapted to provide predictions of which clients will cancel their policy. Using these predictions the company could, for instance, promote proactive retention measures for the customers whose policies were predicted to be cancelled.

Different approaches to the subject have been studied. Yeo *et al.* (2001) used clustering and neural networks to model price sensitivities, while more recently Guelman and Guillén (2014) have proposed using a causal inference framework, where the response isn't modelled directly. In this work however we followed the more conventional methodology of using Generalized Linear Models, particularly the special case of logistic regression, to model the response (Bland *et al.*, 1997; Murphy *et al.*, 2000; Guven and McPhail, 2013).

Although some of the previous works also dealt with the application of the models in pricing optimisation, our work focused only on the model building and testing process and on the insights gained from it.

After creating the model, a secondary goal was to evaluate the model's predictive capacity and transform its output probabilities into binary predictions. Consequently, we followed the work of Freeman and Moisen (2008a) by comparing several threshold optimisation criteria, in order to assess which threshold (above which all policies are predicted to be cancelled) results in the best overall predictive performance.

The logistic regression approach has also been recently applied by Garraio (2015), but our work mainly differs in the tests done on the model, the analysis of the model's predictive capacity and optimal thresholds, the data (coming from a company operating in different distribution channels) and the software used.

Regarding the software employed in our work, SAS Enterprise Guide was used for dataset building and making simple bivariate analyses prior to modelling; Emblem was used for building the logistic regression model and R was used for testing the model, evaluating the model's predictive capacity and optimising thresholds. Use was made of the R packages 'PresenceAbsence' (Freeman and Moisen, 2008b), 'ResourceSelection' and ‘ggplot2’.

This report is organised as follows. In chapter 2 we discuss the variables used in our work, describing the response variable and presenting the various covariates analysed. Chapter 3 begins with an overview of Generalized Linear Models, followed by an exposition of logistic regression and a description of several measures of binary classification performance. Our modelling strategy and goodness of fit assessment of the model are presented in chapter 4, followed by a discussion of the more important covariates in the model. We evaluate the model's predictive capacity and analyse different threshold optimisation criteria in chapter 5. Our results and suggestions for future work are discussed in chapter 6.

Chapter 2

The data

For the purpose of this work, a training set containing 86 344 policies that went through the renewal process was built and prepared for modelling, using SAS Enterprise Guide (Slaughter and Delwiche, 2006). It contains 12 months of data and includes policies that were up for renewal between March 2015 and February 2016. In addition, a test set containing 9 714 policies that went through the renewal process in March 2016 was also created, considering the same assumptions and procedures that went into the creation of the training set. Only individual clients owning passenger cars, commercial cars or vans were under the scope of this work. For details on the data preparation stage (the most time consuming part of this work) see section A.1.

In this chapter we start by discussing the response variable in section 2.1. We present the various covariates in section 2.2, explaining our motivations and expectations behind them.

2.1 Response variable

For a policy to go through the renewal process, it needs to be in force *45 days before* the expiry date of the current term, when the company sends the client a letter indicating the new premium for the next policy term, in case the client decides to renew. The policy's status is then checked *50 days after* the expiry date; if it's still in force it's considered to be a *renewal*, otherwise it's a *cancellation*.[1]

[1] This implies that, for a policy with monthly payments, if the client pays the 1st instalment and then cancels it before the 2nd instalment, it's considered a cancellation.

In this work we code the response as 0 if a policy is renewed and as 1 if a policy is cancelled, meaning that we shall model the *lapse* or *cancellation rate.*

Note that, under special circumstances, a policy may be cancelled outside of the time frame mentioned above. However, as these cancellations aren't caused by premium variations they were obviously not considered in our analysis. Keeping the same idea in mind, policies that were cancelled *during* that time frame but were motivated by something certainly not related to premium variations were also discarded from both datasets.[2] This includes motives such as total vehicle losses and errors in the company's database.

On the other hand, if a policy was cancelled so that the client could purchase a new one in the company (policy cannibalisation), it was still kept under analysis as a cancellation, even though the client hasn't left the company. This is because this sort of policy lapse is almost always due to the newer policy having a lower premium for that customer, so we consider it to be an effect of overpricing on renewals.

2.2 Covariates

In this section we present the different covariates that were included in our datasets, along with some of our expectations regarding them. Amounting to almost 50 covariates, most were suggested in meetings with several key areas in the company dealing with Motor renewals, such as Actuarial, Underwriting and Marketing. By reviewing the literature and considering our own intuition, additional ones were added to the list of variables to analyse. The covariates in our final datasets are marked in **bold**. For further details see section A.2.

2.2.1 Premium-related covariates

The most important covariate is the premium change, since it's the only one the company has direct control over. Thus, we collected for each policy the **absolute premium change** and the **percentage premium change**. Since these two covariates are obviously correlated, it was decided at start that, when creating the model, testing would be done to understand which has more explanatory power.

[2] These disregarded policies amounted to about 10% of all cancellations.

It was discussed whether the full amount of premium being paid should be included in the list of covariates to analyse since, for example, an increase of €20 will have a different impact depending on whether the current premium is €100 or €1000. However, since it's a function of several rating factors that were also under the scope of our analysis, the model's interpretability could be compromised. It was then decided not to include the current premium in our datasets and, in order to still capture some of its effect, analyse additional rating factors which weren't initially considered.

Due to the increasing market competitiveness, it makes sense to account not only for the company's own prices but also for the competitors' prices or, in more general terms, to how competitive the company's prices are for different clients. Obviously, it's not possible to obtain such information, so it was important to look for a different way to measure the competitiveness of Ocidental. We followed one of the propositions given by Murphy *et al.* (2000) which is to do a conversion analysis, where we estimated the **conversion rate** per time of year and customer profile. The motivation behind this idea is that if for a type of client the conversion rates are currently low, then the company's price isn't competitive for those customers and they can find better prices elsewhere.

It should be noted that using the expected conversion rate as a competitive index isn't perfect, as conversion rates reflect how competitive the company is at acquiring new business, and prices for new customers can be very different from prices for renewing clients of the same type (recall that policy cannibalisations weren't removed from our analysis). For further details on our conversion analysis see subsection A.2.1.

2.2.2 Customer satisfaction and service levels

In most cases, the customer's only contact with the company occurs when a policy term ends and a new one begins. The main exception for this is of course when the client reports a **claim**, which is the scenario where the company's true value presents itself to the client. It was therefore important to analyse whether the client reported a claim in the previous term and how satisfied they were with the service, as claimants may focus more on service quality than on premium (Bond and Stone, 2004). Associated with claims is the level of ***bonus-malus*** discount, which depends on the client's claim history.

To analyse how the company's service level has an impact on its customers, the **average time to accept a claim** and the **average time to close a claim** were collected, considering only data on Motor claims. To understand how the *perception* of service quality by the customers impacts their decision, two Net Promoter Scores (NPS) (Reichheld, 2003) were used, the **claims handling NPS** and the **call centre NPS** (as most clients aren't claimants but may still contact the company for other reasons).

Still on the topic of customer satisfaction, we observed whether the client had a **rejected claim** in the previous term and whether the client made a **complaint** in the previous term. Complaints were analysed on a policyholder level and not on a policy level, as the client's dissatisfaction with, for example, their Household insurance may have an impact on their decision to renew the Motor policy.

2.2.3 Client/policy characteristics

Other measures of the relationship between the clients and the company were investigated. It's expected that clients that have been with the company for a longer time are less inclined to cancel their policies, so we collected the **policy age** and the policyholder's **tenure** with the company. Another way to assess customer loyalty is by analysing the **number of other policies in force** or the **number of other lines of business** in the company where the clients have policies in force. On the negative side, we collected the **number of other policies cancelled** in the previous term, as we expect that if a client has recently cancelled some other policy then the probability of cancelling the Motor policy is higher.

Several policy characteristics were also taken into account. An obvious one is the **tariff** associated with the policy, as different pricing approaches and retention measures should have a large impact on renewal rates. When analysing Motor policies, another important factor is whether the policyholder has any **own damage** cover, since they lead to higher premiums and clients that request them tend to be more interested in what their insurance provides and more sensitive to claims handling. Besides the previous covariate, we considered the **number of covers**, the **sum insured** and whether the client has **collision** coverage (the main own damage cover). The

number of objects[3] in the policy and the amount of **third-party liability capital** also made their way into our final datasets.

Another interesting covariate is whether the client made any **mid-term changes** to the policy. One example of this may be a policyholder that added own damage coverage to the policy half way through the term. This policy's premium has then increased during the mid-term and the client may only become fully aware of the amount of this increase during the renewal period, when informed of the new premium for the full year. The client's willingness to renew could then decrease.

As for the **distribution channel**, two levels were considered in this study: "Bancassurance" (the company's main channel) and "Other". We also looked at the number of **days between the policy issue and start dates**, as customers that buy their policy some time before it comes into force may pay more attention to what they're buying.

When it comes to the clients themselves, covariates such as their **gender** and **marital status** were analysed. With respect to the former, regulatory constraints prohibit gender-based pricing discrimination, so extra care was taken when dealing with this covariate.

Regarding ages, besides the **client age** we also kept the **driver age** in our datasets. Intuitively, older clients have lower premiums and are financially better off, thus should be less likely to cancel their policy. The **driving licence age**, which is usually a rating factor, was analysed as a way to measure the client's experience in dealing with insurers.

Covariates related to payments were deemed very important. With respect to the **payment frequency**, we anticipate that clients making more payments are less sensitive to premium changes, as they don't feel the variation all at once. Also, clients paying annually usually only cancel their policy at the end of a policy term (since they've already paid for all of it), so lapse rates for these customers tend to be higher during the renewal period than for others. As for the **payment method**, we expect clients paying by direct debit to have a lower cancellation probability, as less effort goes into making the payments.

Besides the two previous covariates, we looked at whether the client had already

[3] Additional objects on a Motor policy include, for instance, trailers.

missed payments during the previous term. We considered payments previous to the sending of the letter with the new premium, since failing to make payments is one of the motives for policy cancellation. Only payments that were missed because the client couldn't or didn't want to pay were taken into account in our work. Still on this topic, the company has developed a model for classifying customers according to a **financial risk score**.

The geographical area where the policyholders live may also offer some explanation for their behaviour during the renewal period. Besides the **district**, we obtained data from the company's database on the characteristics of the different postal codes[4], including covariates such as **income deciles**, **education level deciles**, **unemployment level** and **urban-rural** classification. A **demographic score**, using information from the previous factors and others not considered here, was available per postal code. The Portuguese **unemployment rate** (Instituto Nacional de Estatística, 2016) was also collected and added to the datasets.

Vehicle characteristics were also considered, such as the **type of vehicle**, the **power-to-weight ratio**, the **weight**, the **engine displacement** and the **fuel**. These rating factors were used mostly as a replacement for the full amount of premium, which as explained previously wasn't considered in our work. We also analysed the **vehicle age**, as we expect clients with older vehicles to no longer feel the need for own damage coverage and therefore being more concerned with the price, compelling them to shop around for different premiums.

[4] Portuguese 7-digit postal codes.

Chapter 3

Methodology

In this chapter we start by presenting the Generalized Linear Models (GLM) in section 3.1. The special case of logistic regression, appropriate for modelling binary data, is described in section 3.2, including some considerations on suitable goodness of fit tests. Section 3.3 contains an exposition of the most commonly used metrics of binary discriminatory performance.

3.1 Generalized Linear Models

Generalized Linear Models (Nelder and Wedderburn, 1972) are, as the name suggests, a generalisation of the classical linear model and have been used extensively in actuarial work. Just like the classical model, GLM are used to analyse the effect that the different *covariates* (or *factors* for categorical covariates) have on the *response variable* of interest, with the additional benefit that non-normal data can now be considered. For more detailed expositions on GLM see McCullagh and Nelder (1989) or De Jong and Heller (2008).

A Generalized Linear Model is composed of three components:

- The distribution of the response variable Y, belonging to the exponential family of distributions.
- The linear predictor $\eta = \sum_{j=1}^{p} x_j \beta_j$, a linear combination of the p covariates, where x_j are the covariates and β_j are the parameters.

- The link function $g(\mu) = \eta$, where $g(.)$ is a monotonic differentiable function and $\mu = E[Y]$.

The covariates are incorporated into the model through the linear predictor, while the link function connects the linear predictor with the mean of the response. The link function must then be chosen so that the fitted values fall inside the interval of possible values of μ.

3.1.1 Exponential family

Definition 3.1. A random variable Y belongs to the exponential family if its probability function may be written as

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.1}$$

for some specific functions $a(.)$, $b(.)$ and $c(.)$.

θ is called the *natural parameter* and ϕ is the *dispersion parameter*. The expected value of a distribution belonging to this family is given by

$$E[Y] = \mu = b'(\theta), \tag{3.2}$$

and its variance by

$$Var(Y) = b''(\theta)a(\phi). \tag{3.3}$$

For a proof of these results see, for example, Nelder and Wedderburn (1972).

The exponential family includes several well known distributions, including the Normal, the Poisson and the Binomial, which can therefore be considered when designing a GLM. For some distribution belonging to this family, the link function where $g(\mu) = \theta$ is called the *canonical link*.

3.1.2 Parameter estimation

Maximum likelihood estimation is used to estimate the parameters of a GLM. Considering a distribution from the exponential family, the log-likelihood of a random sample $(y_1, \ldots, y_n)$ is given by

$$\ell(\theta,\phi;y_1,\ldots,y_n) = \sum_{i=1}^{n} \ln f(y_i;\theta_i,\phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right].$$

The maximum likelihood estimates of the coefficients in the linear predictor can then be estimated, by solving the system of equations

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\partial \ell(\beta_j, y_i)}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \left[\frac{y_i - b'(\theta_i)}{a(\phi)} \right] \frac{\partial \theta_i}{\partial \beta_j} = 0,\ j = 1, \ldots, p,$$

where $\boldsymbol{\beta}$ is the parameter vector.

Since there is usually no closed-form solution, numerical methods such as the Newton-Raphson or the Fisher scoring methods are used, which are identical when the canonical link is selected (McCullagh and Nelder, 1989).

3.1.3 Deviance

Consider the so-called *saturated model*, where the number of parameters equals the number of observations and therefore the fitted values equal the observed ones. Such a model will perform poorly when applied to new data, but it can be compared with a currently fitted model to assess how far this second model is from a perfect fit.

The current model's *deviance* is computed as

$$D = 2\phi(\check{\ell} - \hat{\ell}), \tag{3.4}$$

where $\check{\ell}$ is the log-likelihood of the saturated model and $\hat{\ell}$ is the log-likelihood of the current model.

For large samples, the distribution of the deviance approximates a $\chi^2(n-p)$ distribution (under certain conditions), with this result being usually used to assess the goodness of fit of the model.

3.1.4 Hypothesis testing

To test restrictions on the parameters, including significance testing, the likelihood ratio test can be used, where the null hypothesis (H_0) is that the restrictions hold. The corresponding test statistic is

$$LR = -2(\ell_0 - \ell_1), \tag{3.5}$$

where ℓ_0 and ℓ_1 are the log-likelihoods of the models with and without the restrictions, respectively.

This test statistic asymptotically follows a $\chi^2(q)$ distribution, where q is the number of restricted parameters. We remark that it's identical to the difference in the deviance of both models when $\phi = 1$.

The Wald test is an alternative that doesn't require an additional model to be fitted. The Wald test statistic follows a standard normal distribution for large samples and, in the simplest case of testing the significance of a parameter, is given by

$$Z = \frac{\hat{\beta}}{se(\hat{\beta})},$$

where $se(\hat{\beta})$ is the standard error estimate of $\hat{\beta}$.

3.1.5 Non-nested model selection

While the previous tests present a useful way to chose between two nested models, when comparing non-nested models other means must be contemplated. One commonly used criterion for model selection is the *Akaike information criterion* (AIC) (Akaike, 1974), defined as

$$\text{AIC} = -2\ell + 2p.$$

It balances the model's goodness of fit (measured by the log-likelihood) with a penalty for the number of parameters. When comparing a set of different models, all based on the same observations, the one with the lowest AIC is selected.

3.2 Logistic regression

Consider a binary random variable Y, with the two outcomes denoted by 0 and 1, and define $\mu = P[Y = 1]$ as the probability of success. Then, $Y \sim Bernoulli(\mu)$ and

$$f_Y(y; \theta, \phi) = \mu^y (1 - \mu)^{1-y} \tag{3.6}$$

is its probability function.

It can be easily shown that the Bernoulli distribution belongs to the exponential family, by writing (3.6) as in (3.1):

$$\begin{aligned} f_Y(y; \theta, \phi) &= \mu^y (1 - \mu)^{1-y} \\ &= (1 - \mu) \left[\frac{\mu}{1 - \mu} \right]^y \\ &= \exp \left\{ y \ln \left(\frac{\mu}{1 - \mu} \right) + \ln(1 - \mu) \right\}. \end{aligned}$$

The natural parameter is then $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$, resulting in $\mu = \frac{e^\theta}{1+e^\theta}$. We also have $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = -\ln(1-\mu) = \ln(1+e^\theta)$ and $c(y, \phi) = 0$.

Consequently, by (3.2) and (3.3), we have $E[Y] = \frac{e^\theta}{1+e^\theta} = \mu$ and $Var(Y) = \frac{e^\theta}{(1+e^\theta)^2} = \mu(1-\mu)$, respectively.

A binary response can then by modelled through a GLM by considering the Bernoulli distribution. Regarding the choice of the link function, any appropriate link must bound the probability μ between 0 and 1. Such is the case of the canonical link $g(\mu) = \theta = \ln\left(\frac{\mu}{1-\mu}\right)$, known as the *logit.*

A GLM with the Bernoulli distribution and the logit link defines *logistic regression*, which we'll use in this work. Still, other possibilities for the link include the probit and the complementary log-log functions. For more on logistic regression, refer to Hosmer and Lemeshow (2000) or Kleinbaum and Klein (2010).

3.2.1 Goodness of fit testing

Since it falls under the scope of GLM, the results on parameter estimation and model selection presented in section 3.1 also apply to logistic regression. However, care must be taken when deciding how to test the fit of the model, as the deviance statistic presented in (3.4) may not be an appropriate measure to assess the fit of logistic regression.

We denote by *covariate pattern* the combination of the covariates in the model for a particular observation. For instance, if the model only included two binary covariates, we would have four possible covariate patterns in the data.[1]

When the number of distinct covariate patterns in the data (J) is close to the number of observations (n), the deviance can no longer be assumed to asymptotically follow a χ^2 distribution (McCullagh and Nelder, 1989). We remark that comparing the deviances for hypothesis testing as in (3.5) is however still applicable in this scenario (De Jong and Heller, 2008).

A possible alternative in this case is the Hosmer-Lemeshow (HL) test (Hosmer and Lemeshow, 2000). Rather than considering each distinct covariate pattern, the HL test groups the observations based on the quantiles of the estimated probabilities. Usually the *deciles of risk* are considered, meaning that the observations are divided into 10

[1] For each observation, we could only observe the patterns (0,0), (1,0), (0,1) or (1,1).

groups of equal size, with the first group containing the 10% of observations with the smallest estimated probabilities and so on.

The number of observed responses in group k is given by

$$o_k = \sum_{j=1}^{n'_k} y_j,$$

where we consider just the n'_k observations in group k.

The average estimated probability in group k is computed as

$$\bar{\mu}_k = \sum_{j=1}^{n'_k} \frac{\hat{\mu}_j}{n'_k},$$

where $\hat{\mu}_j$ is the fitted probability for observation j.

The HL statistic is then defined as

$$HL = \sum_{k=1}^{g} \frac{(o_k - n'_k\bar{\mu}_k)^2}{n'_k\bar{\mu}_k(1-\bar{\mu}_k)},$$

where g is the number of groups considered.

Under the null hypothesis of good fit, the distribution of the statistic is well approximated by a $\chi^2(g-2)$ distribution, being more appropriate when $J \approx n$ (Kleinbaum and Klein, 2010).

Although residual checking is a crucial point when assessing the appropriateness of a GLM, the usually considered residual plots are uninformative when using logistic regression and $J \approx n$, as for almost all distinct covariate patterns the number of observed responses is either 0 or 1 (Agresti, 2002).

3.3 Binary classification measures

In this section we discuss binary classification and how to evaluate a model's discriminatory performance (how good are its class predictions for the different observations). For an introduction on the topic, refer to Metz (1978).

A *classification model* will, in the case of binary outcomes, classify each observation as either positive or negative (1 or 0), leading to some very simple definitions. A *true positive* (TP) indicates an observation that was correctly predicted as being positive, while a *true negative* (TN) indicates a correctly predicted negative instance. Conversely, a *false positive* (FP) denotes an observation that was wrongly predicted as positive, while a *false negative* (FN) denotes an observation that was incorrectly predicted as negative. We shall denote the total number of observed positive instances by P and the total number of observed negative instances by N.

The observed and predicted outcomes can be summarised in a *confusion matrix*, as shown in Table 3.1.

	Observed Positive	**Observed Negative**	**Total**
Predicted Positive	*True Positive*	*False Positive*	TP+FP
Predicted Negative	*False Negative*	*True Negative*	TN+FN
Total	P	N	P+N

Table 3.1: Confusion matrix

We observe that summing over the main diagonal gives us the number of cases where the model was right. The fraction of correctly predicted cases is known as *accuracy*:

$$Accuracy = \frac{TP + TN}{P + N}.$$

While it might seem that this is the main index of classification performance, it isn't appropriate when in the presence of *class imbalance* (Chawla, 2005), where one of the classes is much more prevalent in the data. For instance, if 99% of the cases were positive, a model could predict every observation to be positive and yield an accuracy of 99%. Obviously this model would have no practical use, so other measures besides accuracy should be taken into account.

The *sensitivity* is given by the proportion of true positives among all positive cases:

$$Sensitivity = \frac{TP}{P}.$$

The *specificity* on the other hand is the proportion of true negatives among all negatives cases:

$$Specificity = \frac{TN}{N}.$$

Increasing one of these last two measures usually results in decreasing the value of the other, with this trade-off being one of the main concerns when building classification models. In other words, to obtain more true positives the model will generate fewer true negatives, and vice-versa.

The proportion of true positives among all predicted positive cases is denoted by *precision*:

$$Precision = \frac{TP}{TP + FP}.$$

Assuming that the positive cases are the class of interest, the main goal would be to have both a high sensitivity (capture most of the positives) and a high precision (avoid many false positives). However, these two statistics also tend to have opposite behaviours, as trying to capture more positives usually leads to an increase in the number of false positives.

One way to represent this second trade-off is using the *F-measure*, which is simply the harmonic mean of sensitivity and precision:

$$F\text{-}measure = \frac{2}{\frac{1}{Sensitivity} + \frac{1}{Precision}}.$$

Another way to measure the overall quality of the predictions is using the *kappa* statistic, which compares the model's accuracy with the *expected accuracy* in case predictions were done by chance. The expected accuracy is computed based on the marginal totals of the confusion matrix as

$$Expected\ Accuracy = \frac{P \times (TP + FP) + N \times (TN + FN)}{(P + N)^2}.$$

Kappa is then computed as

$$Kappa = \frac{Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}.$$

A higher value of kappa indicates a better model, with 1 indicating perfect agreement between predictions and observations and 0 indicating a model performing no better than chance. It's possible for a model with very high accuracy to have a very low kappa, if the expected accuracy is also high. This means that the model has poor predictive capacity despite the accuracy pointing to the contrary, demonstrating the importance of kappa as a classification measure. For more details on kappa see, for example, Agresti (2002).

3.3.1 ROC analysis

While the statistics discussed thus far are computed from actual class predictions, *probabilistic models* such as logistic regression yield a probability rather than a class prediction. In these cases it's then necessary to define a threshold or cut-off point c, so that an instance is classified as positive if its associated probability is higher than c and classified as negative otherwise. Nevertheless, it's still possible to evaluate a probabilistic model's discriminatory ability without choosing a threshold, with the most common approach being ROC analysis (Fawcett, 2006).

Receiver Operating Characteristics (ROC) graphs display the sensitivity on the vertical axis and 1-specificity on the horizontal axis, representing the trade-off between true and false positives. Consequently, the point (0,1) represents perfect discrimination while points (0,0) and (1,1) represent the performance associated with thresholds of 1 and 0, respectively.

For probabilistic models, an ROC curve can then be constructed by evaluating the sensitivity and 1-specificity for the whole range of thresholds and plotting these points on an ROC graph. An example of an ROC curve is presented in Figure 3.1.

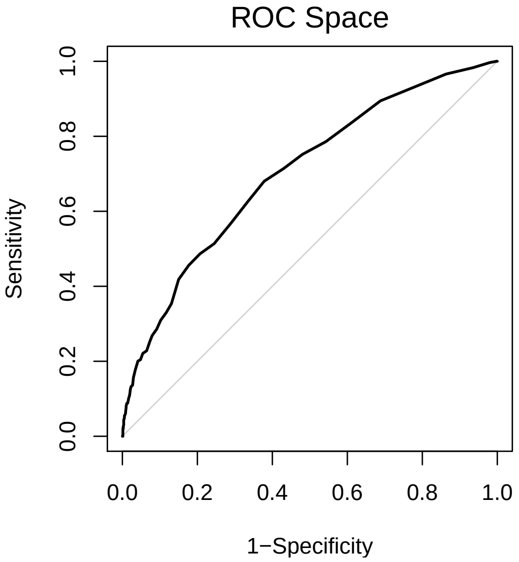

Figure 3.1: Example of an ROC curve

ROC curves can be used to compare different models for the same data. If one model's curve is always above the other, then the first model has higher sensitivity and specificity for all possible thresholds, indicating superior performance.

Alternatively, the area under the ROC curve (AUC) can be computed, with a higher value indicating better performance. The AUC takes values between 0 and 1 and can be used to assess the discriminatory capacity of an individual model, with most realistic models having an AUC greater than 0.5. It's equivalent to the probability that a randomly selected positive instance has a higher associated probability (given by the model) than a randomly selected negative instance (Hanley and McNeil, 1982).

Chapter 4

Modelling the lapse rate

In this chapter we apply the GLM/logistic regression methodology described in chapter 3 to build and test a model for the lapse rate. The model building process and the decisions that were made during it are presented in section 4.1 and the adequacy of the model in terms of fit is discussed in section 4.2. The main covariates included in the model are then examined in section 4.3.

A simple bivariate analysis was performed prior to modelling. See section B.1 for more details.

4.1 Building the model

We shall now describe the main guidelines followed in our modelling process, where we made use of the variables and the training set described in chapter 2. Emblem, a software designed specifically for GLM modelling, was used to build the logistic regression model.

The *backward elimination* procedure was chosen, where we start with an initial model including every covariate and no interactions. However, in order to prevent collinearity and making sure that this initial model ran without any problem, pairs of highly correlated covariates had to be initially identified and only one covariate out of each pair was selected.

We used Cramér's V to identify these pairs, considering a threshold of 0.5. Most of the high correlations were already anticipated, such as the correlation between client age and driver age ($V = 0.913$). In this particular case the client age was chosen, as it makes more sense (the client is the one paying the premiums) and had better data quality.

Nevertheless, there were cases of more than two covariates being correlated between them, which could lead to multicollinearity in the model. The demographic covariates were all correlated with each other, as areas with higher income tend to have lower unemployment, for instance. We decided in this case to keep only the district in the initial model.

The decision was made to revisit these removed covariates further down the road, for instance if the initial chosen covariate was deemed not significant, had less explanatory power or, for factors, ended up with so few levels that the initial correlations no longer had an impact on the model.

Having our initial list of covariates to input in the model, we ran the logistic regression in Emblem, using the binary variable lapse(1)/renew(0) as our response variable, with fixed $\phi = 1$. Due to the large number of covariates, in order to test joint significance of a covariate's parameters we used the Wald test, which unlike the likelihood ratio test doesn't require additional models to be fitted. The different covariates were ordered by their associated p-values, the one with the largest p-value was removed and a new model was fitted. This process continued until finding the model where all covariates were significant at a 5% significance level.

Covariates that didn't exhibit sensible trends were also removed, as was the case of the NPS and the unemployment rate. We believe that even if they presented sensible trends, having only one year of data prevents us from reaching any meaningful conclusions for these covariates. We thus recommend a future analysis of these covariates when more years of data are made available.

We emphasise that, for a factor, testing $H_0 : \beta_j = 0$ only tells us whether level j is statistically different from the level in the intercept. Therefore we needed to test $H_0 : \beta_j = \beta_k$ for all pairs (j, k), $j \neq k$ of levels of a factor. Emblem provides a matrix with the result of the corresponding Wald test for each pair of levels. The two levels that were most statistically similar were then aggregated and a new model was fitted. This procedure went on until all levels of a factor were statistically different from the rest, at a 5% significance level.

Regarding the quantitative covariates, we introduced polynomials of degree 4 to capture non-linear effects in the data.[1] The polynomial terms would then be removed according to the significance tests as well as by visually inspecting the curve fitting.

Due to the characteristics of the software, some quantitative covariates had to be categorised and converted into a score, with this replacement score taking the place of the original covariate in the model. Additionally, we used the default setting in Emblem and used orthogonal polynomials to prevent collinearity. Section B.2 has more details on both of these topics.

Our next step was testing for interactions, with several pairs of covariates being considered. Due to the nature of our work, there was a larger focus on interactions between the premium change and other covariates. To prevent the coding of a factor from impacting tests for interactions terms, the main effects were kept in the model, even if these weren't significant after including the interaction (Kleinbaum and Klein, 2010).

As a final step, and as mentioned previously, we revisited the covariates that were initially discarded due to large correlations. Our strategy here was to introduce each covariate one by one and test its significance, using the likelihood ratio test. When the correlations were still an issue, the AIC was used to asses which of the correlated covariates had more explanatory power, by comparing models including just one of them (along with the non-correlated ones).

Regarding the absolute and percentage premium change, we observed that the percentage change was highly correlated with factors such as tariff or *bonus-malus*. This lead to difficulties in obtaining sensible fits when percentage rather than absolute change was in the model, since we were obtaining decreases in lapse probability for higher positive premium variations. As the model with the absolute change no longer presented such nonsensical results, this was obviously the one we selected.

As mentioned in subsection 2.2.3, gender-based pricing discrimination isn't allowed. So, even though this factor was significant, it couldn't stay in the final model and was removed.

As for the financial risk score, for scores greater than 10 there was a "jump" in lapse rates. Instead of using a second degree polynomial, a better fit was obtained by considering just a linear effect with a jump at the score of 11. Emblem makes this by

[1] Notice that the linear predictor is linear in the *parameters*, not in the *covariates*.

simply introducing a binary variable indicating whether the score is greater than 10, which we kept in our final model.

For the final model summary, see section B.3.

4.2 Goodness of fit assessment

After completing the previous steps, the final stage in the model building process was assessing the goodness of fit of the model. Since the training set contains 86 306 distinct covariate patterns out of 86 344 observations, the Hosmer-Lemeshow test is preferable, as stated in subsection 3.2.1. We stress that these covariate patterns take into account only the covariates in the final model and respective final level groupings.

The R package 'ResourceSelection' was used for this purpose, for the R output see section C.1. The value of the resulting HL statistic is 5.2389, corresponding to a p-value of 0.7318 for a $\chi^2(8)$ distribution.

While the null hypothesis of good fit isn't rejected, we also want the model to perform well in an out-of-time sample, so we applied the HL test on the test set (with 9 710 distinct covariate patterns out of 9 714). We remark that when applying the test on validation data the number of degrees of freedom of the χ^2 distribution increases to 10 (Hosmer and Lemeshow, 2000). The value of the resulting test statistic is 10.271, corresponding to a p-value of 0.4170 for a $\chi^2(10)$ distribution, indicating once again no evidence of poor fit.

Figure 4.1 shows two calibration plots, one for the training set and another for the test set, with the observed and average estimated lapse rates plotted against each other. While the HL test inspects the *deciles of risk*, each dot in the plots represents a *percentile of risk*.

As expected, the fit is much better on the training set, as it contains the observations used in creating the model. Nevertheless, the graphical analysis doesn't point to the model consistently over or underestimating the lapse rates and, in conjunction with the results of the HL test, we conclude that the model provides an adequate fit on both datasets.

Furthermore, to check if the model is well calibrated for specific subpopulations of the portfolio, we estimated the lapse rate for 10 clusters of clients defined by the company, using the policies in the test set. The results in Table 4.1 show that the

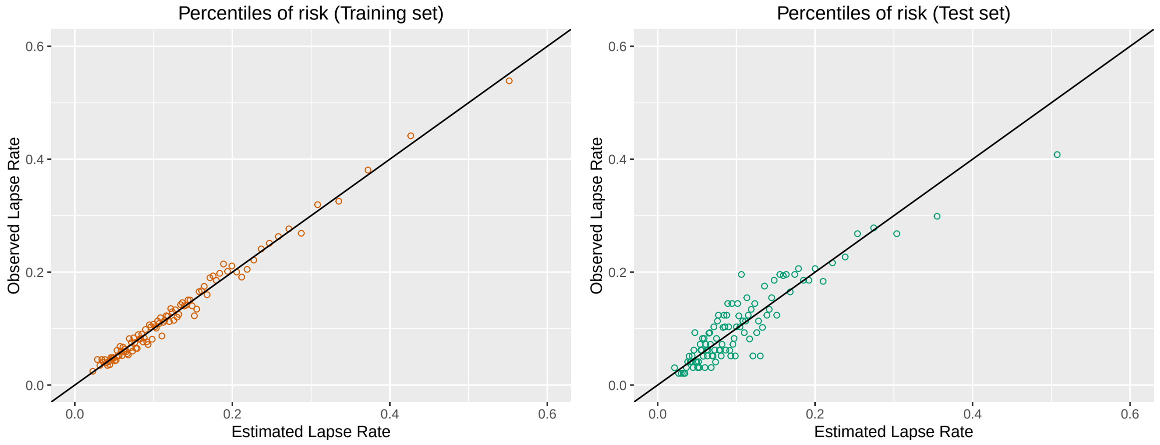

Figure 4.1: Calibration plots (both datasets)

estimated rate was close to the observed one on most clusters, with our main concern being cluster 7, where the rate was almost 3pp off and we have a large number of observations. After further inspection, we discovered that a modification on the tariff associated with cluster 7 had occurred, during the month where the test data originates from, prompting the previously remarked difference. This constitutes an obvious example that, as business conditions change, so must this sort of models be recalibrated over time.

	Clusters									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Difference in rates	1.40	0.58	2.05	2.48	1.70	0.65	2.79	0.99	1.02	0.90
No. of policies	794	689	200	268	446	1740	2086	382	313	2796

Table 4.1: Difference between observed and estimated lapse rates per cluster (*in pp*)

4.3 Empirical results

We shall now comment the results obtained regarding the most remarkable covariates in the model. The probabilities shown here were computed with all other covariates set to their base level. To preserve the confidentiality of the data, probabilities/rates are presented as a percentage of the highest probability/rate in each axis.

Figure 4.2 shows the lapse probability per interval of absolute premium change.

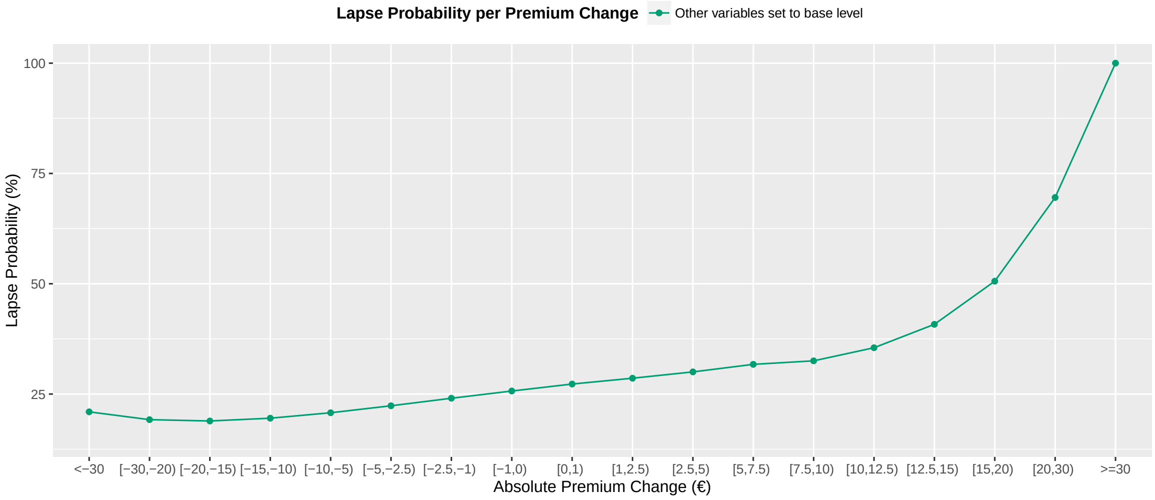

Figure 4.2: Effect of premium change on cancellations

Higher increases in premium lead to higher cancellation probabilities, as expected. However, the lapse probability flattens off where premium decreases are around €15 and then increases slightly. Similar effects were observed by Bland *et al.* (1997) and by Garraio (2015), with Murphy *et al.* (2000) stating that this effect is typical of elasticity curves. Since large decreases are often already anticipated by the customers that receive them (e.g. by moving to another *bonus-malus* level), one possible justification for this effect is that those clients feel that the decrease isn't large enough (Garraio, 2015). Another possibility is that clients may not understand why they paid so much in the previous term, compared to what the new premium is, triggering them to look for alternative prices elsewhere (Murphy *et al.*, 2000; Guven and McPhail, 2013).

Since our final model has interaction terms between the absolute premium change and three other factors, each client's estimated price elasticity curve won't necessarily have the same shape as the one shown in Figure 4.2. This way, more insights are gained regarding how the sensitivity to premium variations varies between different subpopulations of the portfolio.

Figure 4.3 shows the same as the previous figure, only this time distinguishing between clients with and without own damage coverage.

The first interaction effect with the premium change is clear, as policyholders with additional covers pay higher premiums and are therefore less sensitive to large absolute variations. Additionally, clients with this coverage tend to care more about the product characteristics than clients with only the basic third-party liability coverage, which may

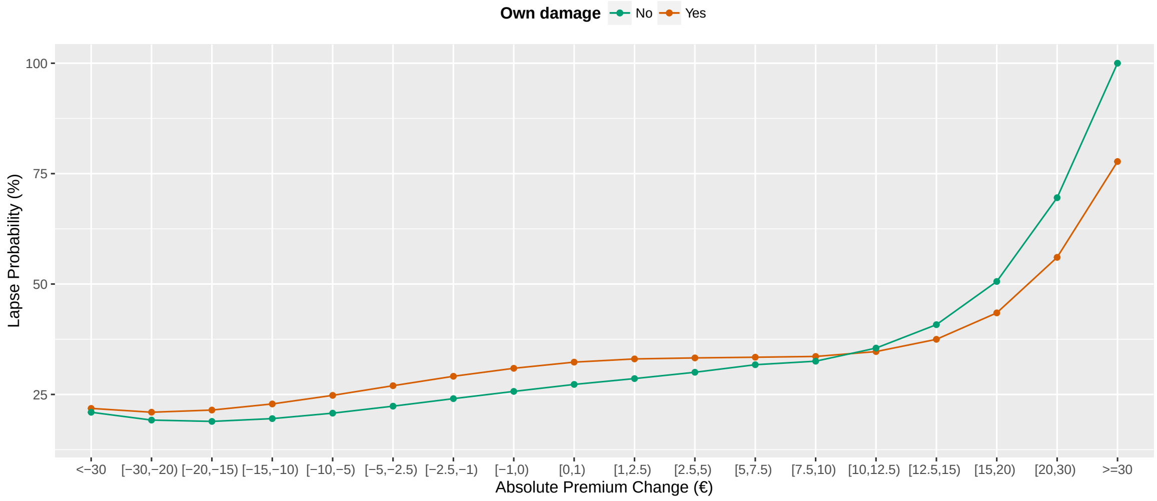

Figure 4.3: Effect of own damage on cancellations

also help explain this effect.

Similarly, claimants are less sensitive to large increases in premium, as can be seen in Figure 4.4. This effect is in agreement with the results of Guelman and Guillén (2014) and occurs since those customers already expect some premium increase. We remark that the number of clients with claims and who had premium decreases larger than €15 is very small (approximately 2% of the dataset), making it harder to retrieve any conclusion from the estimated increase in price sensitivity for these clients.

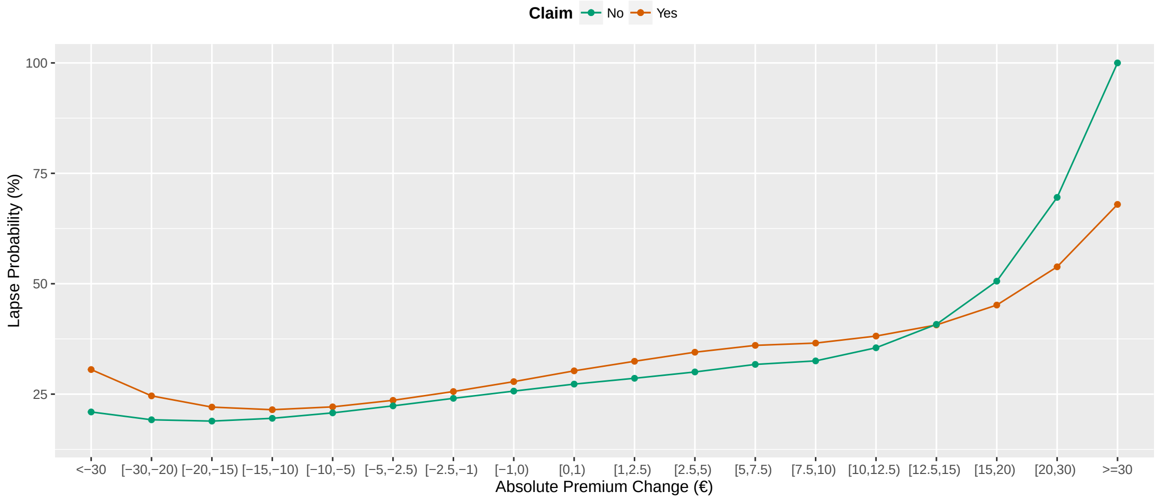

Figure 4.4: Effect of claims on cancellations

Regarding the payment frequency, Figure 4.5 confirms our expectations that the price sensitivity of clients making annual payments rises as the premium increases,

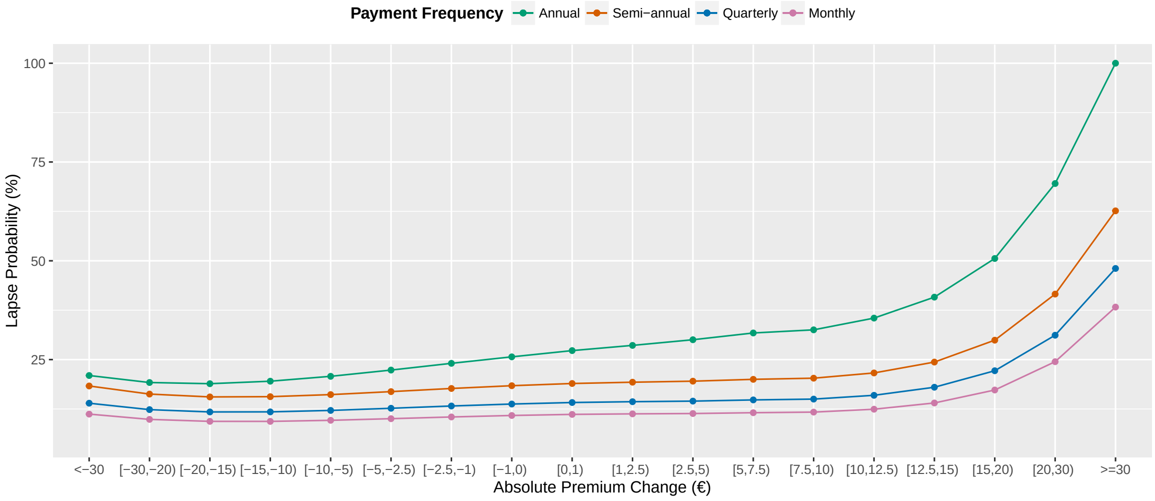

Figure 4.5: Effect of payment frequency on cancellations

since they receive the full price increase at once.

We emphasise that the *bonus-malus* discount and the number of other cancellations, besides exhibiting their expected effects, have a noticeable impact on the lapse probability. Also, customers with several policies in the company have lower cancellation probabilities, showing that clients care about keeping their different insurance products in one company (Barone and Bella, 2004).

The effects of the different tariffs on the response reinforce the notion that, when designing new tariffs or products, the long term impact on retention should be taken into account, and not just the short term impact on new business.

We recall that we estimated the conversion rate per time of year and customer profile, with Figure 4.6 showing the lapse probability per conversion rate. Since higher conversion rates are associated with competitive premiums, we expected the lapse probability to decrease as the estimated conversion rate increased. While this effect can be observed at first, we see the opposite effect occurring for very high conversion rates.

As mentioned in subsection 2.2.1, conversion rates are a measure of competitiveness on new business, not of renewal competitiveness. We thus believe that policy cannibalisations (clients cancelling their old policy to make a new one) and clients that tend to shift companies very often are what's causing lapse rates to increase for high conversion profiles.

Also of note is the effect of mid-term changes that lead to premium increases,

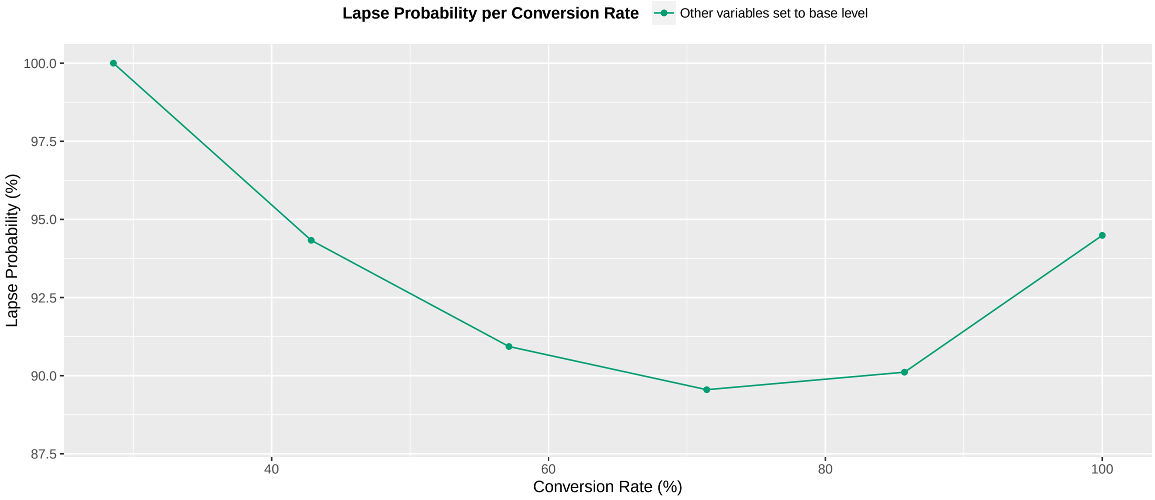

Figure 4.6: Lapse probability per conversion rate

as clients who change their policy and consequently see their premium rise are more prone to cancelling it during the renewal period. As mentioned in subsection 2.2.3, one possible reason for this may be that clients only become fully aware of the impact on the premium when the new policy term begins.

As for the district where the customer lives, the final aggregation resulted in groups with districts which are, in several cases, very far apart from each other. We believe that this result may be linked to how the level of competition in Motor insurance differs from one to district to another.

Overall, the effects of the remaining covariates matched our expectations and the results of the previous works mentioned in this section.

Chapter 5

Obtaining binary predictions

Our two main objectives for this work were creating a model for the probability of policy cancellation and, at the same time, gaining insights on how different factors influence renewals. As a consequence, if a covariate wasn't significant and/or didn't make sense it was promptly removed from the model, even if keeping it would have increased the model's predictive ability (as measured by AUC).

Still, it's clear that obtaining predictions on which clients will cancel their policies would be advantageous for the insurer, as it could, for instance, help the company select which clients should be the target of retention measures during the renewal period.

In this chapter we analyse the model's predictive ability and study different threshold optimisation criteria, to obtain the best threshold based on overall performance, using the binary classification measures discussed in section 3.3.

5.1 Evaluating predictive ability

The ROC curves resulting from the application of the model on the training and test sets are shown in Figure 5.1. The value of the AUC on the training set is 0.702, indicating fair discriminatory capacity (Kleinbaum and Klein, 2010). As expected, when applying the model on the test set the AUC decreases, to 0.678. It's a small decrease nevertheless, indicating no strong sign of overfitting and leading us to conclude that model's predictive ability translates well into unseen data.

More insights on the topic are given by Figure 5.2. For a model with excellent

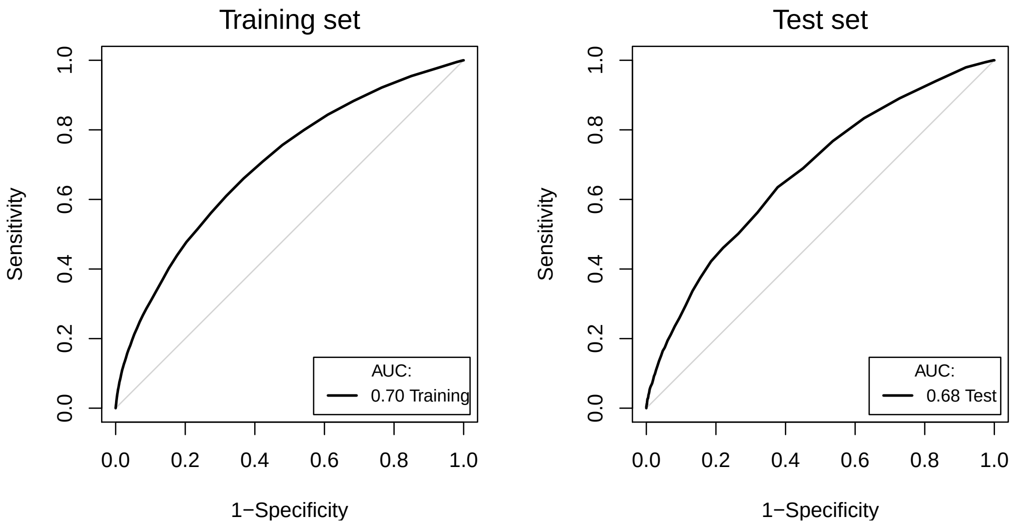

Figure 5.1: ROC curves (both datasets)

discrimination, this histogram would have almost all renewals on the left side, with low probability, and almost all lapses on the right side with high probability, showing clear separation between classes. In that case it would be possible to find a threshold that could easily discriminate between both classes. However, what we observe is that while renewals are gathered on the left side of the graph, lapses are spread out over a large range of probabilities. This is a consequence of the severe class imbalance in our dataset (the lapse rate is lower than 15%). As most clients renew, the model can easily achieve high specificity, but predicting cancellations (our event of interest) is much harder.

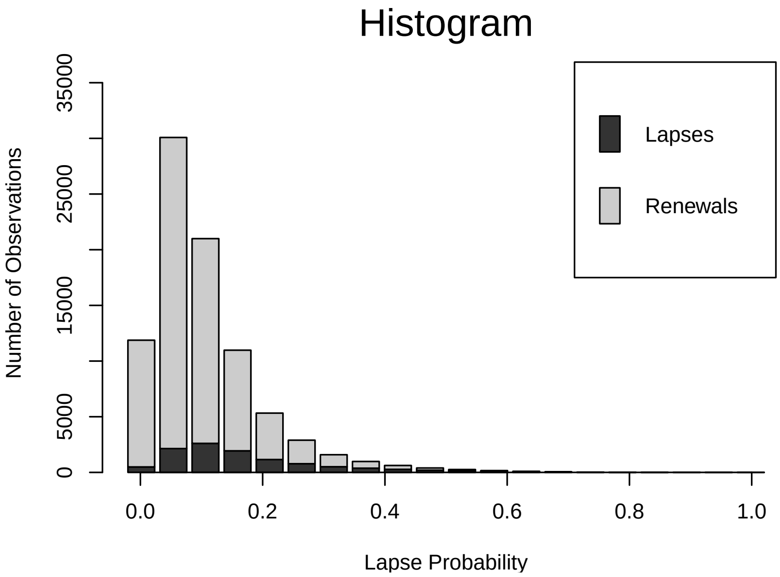

Figure 5.2: Histogram plot (training set)

Figure 5.3 shows sensitivity, specificity and kappa on the training set as functions of threshold. As expected, specificity has a sharp increase, due to the renewals having been assigned mostly small probabilities. On the other hand, sensitivity decreases quickly - to capture most of the lapses the model yields a large number of false positives. Kappa, which evaluates the model's discriminatory ability overall, is consequently low for most of the threshold range, barely going above 0.2 at its maximum.

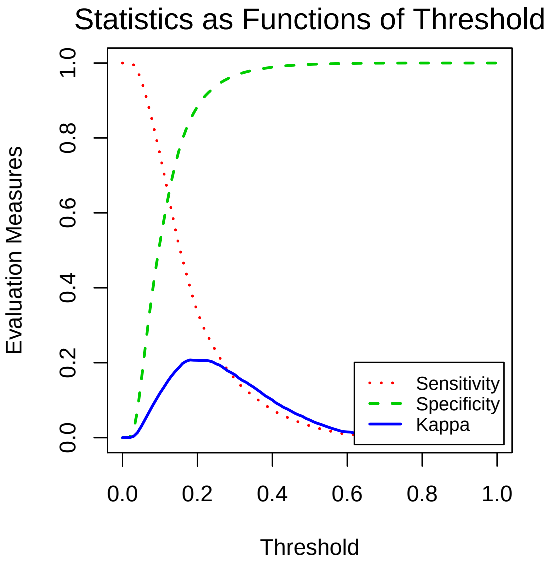

Figure 5.3: Sensitivity, specificity and kappa (training set)

5.2 Threshold optimisation

Since the model displays a reasonable predictive capacity, we now need to define a threshold to obtain actual lapse predictions from the model. Several threshold optimisation criteria exist, and in our work we consider 7 of such criteria which have been previously analysed by Freeman and Moisen (2008a).[1] Besides the evaluation measures discussed in section 3.3, we also look at the difference between the observed and predicted prevalence (lapse rate) given by each threshold. The different criteria are presented below in **bold**.

Usually, the thresholds should be determined using a set other than the training or test sets, especially when dealing with small samples. This is to avoid an overly

[1] Freeman and Moisen (2008a) studied additional criteria, such as criteria where the sensitivity or specificity are user defined, which weren't considered in our work.

optimistic assessment of the model's quality and to keep the test set completely independent from the model building process (Kuhn and Johnson, 2013). In this work however, the training set was chosen to examine the criteria (since we have a large enough sample), making use of the R package 'PresenceAbsence' (Freeman and Moisen, 2008b). The results are shown in Table 5.1 (see section C.2 for the R output).

Criterion	Cut-off point	Difference in prevalence	Accuracy	Sensitivity	Specificity	Precision	Kappa	F-measure
PredPrev=Obs	0.212	0.0452	0.8265	0.3050	0.9011	0.3061	0.2064	0.3055
MaxKappa	0.184	4.9534	0.7972	0.3877	0.8558	0.2778	0.2082	0.3237
MaxPCC	0.479	11.6210	0.8758	0.0394	0.9954	0.5511	0.0578	0.0736
Default	0.500	11.8040	0.8757	0.0317	0.9964	0.5577	0.0472	0.0601
MaxSens+Spec	0.132	21.9911	0.6802	0.6011	0.6916	0.2180	0.1670	0.3200
Sens=Spec	0.123	26.3203	0.6476	0.6438	0.6482	0.2075	0.1536	0.3138
MinROCdist	0.120	27.9220	0.6358	0.6603	0.6322	0.2044	0.1496	0.3121

Table 5.1: Threshold optimisation results (training set)

The **Default** criterion, simply consisting of a cut-off point of 0.5, and the **MaxPCC** criterion, maximising accuracy[2], give similar results. While accuracy is very high in both cases, due to the class imbalance this is done by "focusing" on renewals, as they're much more prevalent, resulting in exceptionally high specificity and extremely low sensitivity. Precision is higher than for other criteria, due to only the policies with very high probabilities actually being predicted as lapses. This "unbalance" in the statistics is reflected by very low values of kappa and F-measure, indicating poor discriminatory performance.

Sens=Spec finds the threshold where sensitivity equals specificity, while **MaxSens+Spec** maximises the sum of the two. As mentioned in subsection 3.3.1, the point (0,1) in ROC space implies perfect discrimination. The **MinROCdist** criterion therefore finds the threshold minimising the Euclidian distance to that point. While precision is now lower for these three criteria, sensitivity is much higher than before, which is reflected by the increase in F-measure. Still, the predicted and observed prevalence are too far apart and while kappa is higher than before it's still possible to improve it.

[2] Percent Correctly Classified (PCC) is the denomination given by Freeman and Moisen (2008a) to accuracy.

PredPrev=Obs finds the threshold where the predicted prevalence equals the observed one and **MaxKappa** maximises kappa. They result in the two highest values of kappa and the two lowest differences between observed and predicted prevalence, with Freeman and Moisen (2008a) observing similar results in their work. **MaxKappa** however resulted in a better balance between sensitivity and precision, as given by F-measure.

If the company wishes to use this model for prediction, we therefore recommend applying a threshold of 0.184 (MaxKappa), based on overall performance.

The test set has an even lower prevalence than the training set (by 1.88 pp). Besides the expected decrease in performance when moving to new data, this higher proportion of renewals means that all the previously computed thresholds lead to lower sensitivity, precision, kappa and F-measure (when compared with the training set). Once again, see section C.2 for the R output.

Regarding our chosen threshold of 0.184, the results from the test set are presented in Table 5.2. Despite the predicted and observed prevalence now being much closer than when we used it on the training set, there is an obvious decrease in overall performance, as seen by the low values of kappa and F-measure.

Criterion	Cut-off point	Difference in prevalence	Accuracy	Sensitivity	Specificity	Precision	Kappa	F-measure
MaxKappa	0.184	0.0515	0.8393	0.2469	0.9098	0.2457	0.1563	0.2463

Table 5.2: Application of the chosen threshold (test set)

We stress that prediction wasn't our main goal and that class imbalance in the data notably leads to worse model performance. We believe however that using simply remedies for class imbalance, such as basic resample methods to construct more balanced datasets, could lead to significant improvements in model performance, if prediction is the ultimate goal. For more on dealing with the class imbalance problem in lapse prediction, including different resampling methods and other models more flexible than logistic regression, see Burez and Van den Poel (2009).

Chapter 6

Conclusions

In this work we used logistic regression to model the Motor lapse probability (during the renewal period) as a function of premium change and other covariates. In addition, the model's ability to correctly predict which policies will be cancelled was studied and several threshold optimisation criteria were compared, to obtain the best overall binary predictions.

Our model allows the lapse probability to be estimated for each client per interval of premium change. This way, an understanding of which customers are more elastic was obtained and, in the context of the whole portfolio, better pricing decisions can be made so that the company may improve its renewal rate and/or renewal profitability.

Preparing the data was notably the most time-consuming stage, with almost 50 covariates being initially considered to help explain the behaviour of the response variable. Owning to the nature of our work, a greater focus was given to premium-related covariates, with the conversion rate being estimated to measure the company's competitiveness.

Our model was built in Emblem, using the GLM/logistic regression methodology described in chapter 3. Logistic regression was chosen as it allows a binary response to be modelled and provides interpretable results, regarding the effects of the covariates on the response. Because of the large number of correlated covariates, extra care was taken to prevent multicollinearity.

Due to the large number of distinct covariate patterns in the data, the Hosmer-Lemeshow test was used to assess the model's goodness of fit and, considering the purpose of the model, a test set was used to evaluate its performance on new data.

The HL test, along with additional graphical analysis, lead us to conclude that the model provides an adequate fit to the data.

The covariates included in the final model provided some insights on how different clients react during the renewal period, with the results matching most of our expectations. Main covariates besides the premium change include the payment frequency, the *bonus-malus* level or the number of other cancellations and, in addition, some interactions with the premium change were left in the final model, allowing the shape of price elasticity curves to differ from client to client.

We also concluded that the model presents a fair overall predictive ability (as measured by AUC) and, if the company wants to use this model for prediction, we recommend using a threshold of 0.184, resulting from the maximization of the kappa statistic. The class imbalance in our data however caused difficulties on predicting cancellations and further developments on this matter can be undertaken, if prediction is the ultimate goal.

In the end, we have achieved our main goal of creating an adequate model to estimate renewal price elasticity. Still, as business conditions change, old clients leave and new ones arrive, we expect the quality of the model to worsen over time, with one such example already being reported in section 4.2. We therefore advise that a periodic model recalibration must be done, as more data becomes available.

One limitation of our work was that our training set only included one year of data, preventing us from gathering meaningful conclusions from covariates such as the NPS or the unemployment rate. Another limitation was that no rigorous method was used when categorising quantitative covariates. We thus suggest that proper methods for minimising information loss, when banding quantitative covariates, be considered in future analyses.

While we estimated renewal price elasticity, one recommendation for future work is to consider the other aspect of demand modelling, by analysing the moment of risk acquisition and estimating *conversion* price elasticity. By having at its disposal a model for each of the two sides of demand modelling, a company can use them along with the usual cost models to improve future tariff design.

Also, while most of the previous works (including ours) focused on individual Motor clients, the same approach we used can be applied to corporate clients and to other competitive lines of business, such as Household insurance.

Appendix A

Additional notes on the data

A.1 Data preparation

Although the initial strategy was to use 3 years of data (from late 2012 to late 2015), the lack of historical information on premiums led to a change of plans. More precisely, before March 2015, the proposed premiums for the new policy term were not recorded in the system for policies that were cancelled before the renewal date. This meant that, in order to obtain a good representative sample of the portfolio and be able to compute the past premium variations for all policies under analysis, our time horizon became shorter and we used more recent observations, with data from March 2015 to February 2016 on the training set and from March 2016 on the test set.

Creating the datasets was the most time consuming part of this work (taking about half of the project's duration), as the amount of data sources, the idiosyncrasies of the company's database and the occasional errors or nonsensical values demanded a lot of care and attention to detail. A thorough univariate analysis was done on each covariate, in order to detect and consequently handle these flaws, which included making histograms and box plots for quantitative covariates and bar plots for factors.

The more relevant issues found in the data and the decisions that were taken follow:

- Several policies had missing premiums, usually caused by changes in those policies that were wrongly processed by the system and as such no premiums were loaded into the database. These policies were removed from the analysis.

- Other policies had premiums of €0, which was once again the result of errors and as such we also didn't consider these policies in our datasets.

- Although under some older tariffs it's possible for a policy to maintain its premium from one term to the next, more than 80% of the policies with no premium change in our dataset were classified as such due to the database not saving the new premium and instead keeping the old one. After careful consideration, it was decided to remove policies without any premium change from the analysis.

- Some policies didn't have an associated driver, and in this case the driver's date of birth and postal code are given default values. However, these default values change from tariff to tariff and, consequently, to identify policies in this situation it was necessary to obtain for each tariff the corresponding pair (date of birth, postal code) of default values.

- For policies with more than one object, the information associated with the main object was the one selected (in most cases the main object was a car and the second object was something like a trailer).

- In order to cross data related to the client and not just the one policy under analysis (such as complaints made or number of other policies cancelled), a code identifying each customer was used. A small number of policies were missing it and were thus ruled out from the analysis.

- Requests done by the clients to the company are sometimes classified as complaints, in order to accelerate their process. Since these "complaints" don't represent any dissatisfaction that the client may have with the company, these false complaints were identified and discarded from our analysis.

- For the remaining covariates, when the number of errors/missing values was small and could be corrected manually then that was our course of action; if they could not be corrected then those problematic observations were removed from the dataset. On the other hand, when a covariate had a considerable amount of missing values or errors we decided to create a new level for that covariate, "Unknown" or "NA", and aggregate the faulty data under it.

- Examples of problems in the data include negative vehicle ages, policies associated with tariffs that didn't even exist at the time or policyholders younger than 18 years old.

- Regarding missing values, while the "Unknown" level was created for several covariates such as the driver age, the marital status or the vehicle fuel, due to missing postal codes all demographic factors had the same policies with missing values. This issue would then be dealt with during the modelling stage of the project.

We emphasise that, even after dealing with the previous issues and consequently removing various policies, the lapse rates in our final datasets were similar to the lapse rates that were recorded by the company (in each respective time frame).

A.2 Further details on the covariates

This section presents some more details on how we defined each covariate and which values they can take. Whenever the missing value percentage for a covariate is presented, it is the percentage of missing values on the training set.

Regarding the **absolute** and **percentage premium change**, they were both computed using the last premium paid before the end of the term and then either the first premium paid in the next term (for renewed policies), or the proposed premium which was rejected by clients who cancelled their policy.

As for the ***bonus-malus***, the possible levels were "Malus", "0% ", "1% - 10%", "11% - 30%", "31% - 49%" and "50%".

The **average times** (to accept or close a claim) and the **Net Promoter Scores** were computed on a rolling 12 months basis for each month of information in the dataset. The NPS is a measure that tries to assess how loyal clients are to a company, being computed as the difference between the percentage of promoters (clients considered likely to recommend the company) and the percentage of detractors (clients considered unlikely to recommend the company to others). Clients are classified according to their answer (on a scale of 0 to 10) to the question "How likely is it that you would recommend [company X] to a friend or colleague?". See Reichheld (2003) for more details.

Rejected claims were defined as closed claims (during the previous policy term) with zero cost *after* removing administrative costs.

The client's **tenure** with the company was defined as the age of the client's oldest policy and each policy in our dataset was associated with one out of 6 different **tariffs**. In this work we have denoted them with the letters A through F, where A is the oldest tariff and F is the most recent one.

The factor **mid-term changes** was computed by comparing the premiums on the second month after the start of the term and on the second to last month of the policy term. It contains three levels: "Same Premium", "Reduced Premium" and "Increased Premium". In the example given in subsection 2.2.3 the client added some covers to the policy during the mid-term. Its premium would then increase and this policy would fall under "Increased Premium". A threshold of €5 was used to disregard possible premium variations due to small changes or system errors.

As for the **marital status**, we consider the categories "Divorced", "Married", "Single", "Widowed" and "Unknown"(1.34% of observations). Due to some policies not having an associated driver, the covariates **driver age** and **driving licence age** had some missing values in our final dataset (1.54% of observations). These policies were kept since the missing values weren't the result of system errors.

Regarding the **payment frequency**, four options are available to clients: "Annual", "Semi-annual", "Quarterly" or "Monthly". The **payment method** has two classes, "Direct Debit" and "Other", and the company's **financial risk score** model takes values from 1 to 15, with a higher value indicating a higher risk.

District includes the 18 districts and 2 autonomous regions of Portugal as well as the level "Unknown", with this last one existing due to policies missing the corresponding postal code (1.34% of observations). The covariate **unemployment rate** (quarterly data, per NUTS II region) also had the same policies under the "NA" level. The remaining demographic factors had additional policies under "NA" (5.89% of observations), as their data source was also missing some new postal codes. **Income** and **education level deciles** take values from 1 to 10 and "NA", with higher values indicating areas with lower average income and more inhabitants with higher education levels, respectively. The **unemployment level** has the levels "Very High", "High", "Above Average", "Average", "Below Average", "Low" and "NA", while the urban-rural classification is either "Urban", "Mixed", "Rural" or "NA". The **demographic**

score takes values from 1 to 9 and “NA”, with a lower value indicating areas with higher income, further education, lower unemployment, etc.

We only considered clients owning commercial cars, passenger cars or vans, with these being the levels of the factor **type of vehicle**. The factor **fuel** includes the levels “Gasoline”, “Diesel”, “Electricity”, “Mixed”, “Gas” and “Unknown” (1.86% of observations).

A.2.1 Conversion analysis

Estimating the **conversion rate**, per customer profile and time of year, was our way of trying to assess the effect of the company’s competitiveness on the lapse rate. The conversion rate is the proportion of quotations that were successfully converted into policies. Naturally, a higher rate for some client profile indicates that the company’s premium for that profile is very competitive, and therefore clients are responding well to it.

Quotation data from the same time frame as our training set was collected and a Poisson model with a log link, a particular case of the GLM described in section 3.1, was used to model the conversion rate, using the number of quotations made as an offset. Since we were mostly interested in identifying conversion profiles and not so much in conversion price elasticity, the premium was not included as a covariate (Murphy *et al.*, 2000).

Following the same modelling techniques described in section 4.1, a final model with 8 covariates was obtained. The covariates were the quarter of the year, the driver age, the district, the type of discount on entry, the type of vehicle, the fuel, the engine displacement and the power of the vehicle. The effects of these covariates on the conversion rate, as given by this final model, matched our expectations and the business experience.

Using this very simple model the conversion rates were then estimated per customer profile and time of year, for all policies in our dataset. For more on conversion modelling, refer to Murphy *et al.* (2000).

Appendix B

Bivariate analysis and modelling results

B.1 Bivariate analysis

A bivariate analysis involving our variable of interest, the lapse rate, and each individual covariate was done prior to modelling. This simple analysis was meant as a way to create some expectations for the modelling process, to identify possible flaws in variable design and mostly to ensure that we were on the right track.

A $2 \times k$ contingency table was created for each pair (dependent variable Y, covariate X), where 2 is the number of possible outcomes (lapse/renew) and k is the number of levels of covariate X. For the purposes of this analysis quantitative covariates were categorised at their quintiles.

The Pearson chi-squared test was then used to test the null hypothesis that the variables were independent. For an exposition on the Pearson chi-squared test and further analysis of contingency tables see Agresti (2002).

The results of the tests were encouraging, with the hypothesis of independence being rejected in almost all cases. Notable exceptions were the unemployment level and the Net Promoter Scores.

Still, the chi-squared test only tests for independence and doesn't give any details on the strength or form of the association between the two variables. A simple graphical analysis was then done with that purpose in mind. A graph was produced for each pair showing the lapse rate for each level of the covariate.

Most of our expectations were strengthened by this graphical analysis. The main exceptions were the covariates “number of other policies in force” and “number of other lines of business”, where it appeared that the lapse rate increased with the number of policies/lines of business, contrary to what we expected. We later discovered that this was due to a system error that was rectified prior to modelling. Also, after a careful look at how those covariates were designed, we decided to count the number of policies in force 45 days before the expiry date rather than on the expiry date, as was initially done. This prevents policy cannibalisations from being accounted for in the number of policies, which explained in part what we were observing. After the dataset update and the change in design the results finally began making sense.

The analysis described in this section was done using SAS Enterprise Guide. For more on statistical and graphical analysis using this software see Slaughter and Delwiche (2006).

B.2 Handling quantitative covariates in Emblem

Since Emblem has an upper limit of 255 values that a covariate can take, covariates such as the premium change or the power-to-weight ratio had to be categorised before the model building process. The intervals considered for each covariate were designed in order to balance exposure and interval size. Previous categorisations done by the company were also taken into account while preparing these covariates.

Emblem takes this now categorical covariate and transforms it into a numerical score, allocating a different value for each interval (with their natural ordering being taken into consideration). This score is what is then used to model the effect of the original covariate.

It is very simple to obtain the corresponding effect on the response per interval of the initial categorical covariate, as Emblem exports the additive effect on the linear predictor per interval, rather than per numerical score.

Additionally, the default setting in Emblem is to use orthogonal polynomials for quantitative covariates. As higher order terms of a covariate may be highly correlated, considering orthogonal polynomials prevents collinearity and consequently numerical instability (De Jong and Heller, 2008).

We remark that, as a first step to compute orthogonal polynomials, Emblem normalises the values of the covariate x as follows:

$$x^* = \frac{2x - (x_{max} + x_{min})}{x_{max} - x_{min}},$$

where x_{max} is the maximum value and x_{min} is the minimum value in the data. The new covariate x^* takes values between -1 and 1.

B.3 Model summary

Coefficients	Estimate	Std. Error	z value	p-value
(Intercept)	———	———	−66.300887	<2E-16
Absolute premium change	0.385359	0.016896	22.807436	<2E-16
Absolute premium changeˆ2	0.192023	0.012275	15.643333	<2E-16
Absolute premium changeˆ3	0.067463	0.008780	7.683895	1.55E-14
Absolute premium changeˆ4	0.064217	0.008371	7.670969	1.70E-14
Claim (Yes)	0.056454	0.038314	1.473475	0.140623
Claim (Y)*Abs. prem. change	−0.128166	0.026635	−4.811921	1.49E-06
Claim (Y)*Abs. prem. changeˆ2	−0.070124	0.023988	−2.923303	0.003463
Claim (Y)*Abs. prem. changeˆ3	−0.088253	0.022243	−3.967576	7.26E-05
Own damage (Yes)	0.021476	0.031462	0.682601	0.494859
Own damage (Y)*Abs. prem. change	−0.097772	0.019000	−5.145844	2.66E-07
Own damage (Y)*Abs. prem. changeˆ2	−0.088248	0.018262	−4.832316	1.35E-06
Payment frequency (Semi-annual)	−0.428927	0.032655	−13.135159	<2E-16
Payment frequency (Quarterly)	−0.745058	0.057108	−13.046356	<2E-16
Payment frequency (Monthly)	−0.998959	0.034948	−28.584267	<2E-16
Payment freq. (S-a)*Abs. prem. change	−0.135869	0.023410	−5.803770	6.48E-09
Payment freq. (Q)*Abs. prem. change	−0.148201	0.039055	−3.794675	0.000148
Payment freq. (M)*Abs. prem. change	−0.158984	0.022673	−7.012122	2.35E-12
Conversion rateˆ2	0.023140	0.009680	2.390548	0.016823
Bonus-Malus (Malus)	1.386526	0.111136	12.475914	<2E-16
Bonus-Malus (0% - 10%)	0.694658	0.075105	9.249213	<2E-16
Bonus-Malus (11% - 30%)	0.513874	0.052447	9.797966	<2E-16
Bonus-Malus (31% - 49%)	0.135926	0.036741	3.699543	0.000216
No. other policies in force	−0.076374	0.007882	−9.689171	<2E-16
No. other policies in forceˆ2	0.016159	0.006523	2.477362	0.013236

Coefficients	Estimate	Std. Error	z value	p-value
No. other cancellations (1)	0.591512	0.029096	20.329617	<2E-16
No. other cancellations (2)	0.891541	0.049779	17.909941	<2E-16
No. other cancellations ($\geq$3)	1.076756	0.066526	16.185408	<2E-16
Tariff (D)	0.139827	0.047141	2.966127	0.003016
Tariff (F)	−0.325368	0.039211	−8.297792	<2E-16
Payment method (Other)	0.337066	0.068427	4.925934	8.40E-07
Missed payments (Yes)	0.348300	0.059657	5.838379	5.27E-09
Financial risk score	0.072401	0.013002	5.568538	2.57E-08
Financial risk score (>10)	0.448815	0.047941	9.361860	<2E-16
Distribution channel (Other)	−0.173712	0.041664	−4.169324	3.06E-05
Client age	−0.193834	0.014241	−13.611405	<2E-16
Client age^2	0.033875	0.014999	2.258522	0.023913
Client age^3	0.054079	0.014704	3.677817	0.000235
Client age^4	0.040958	0.014635	2.798564	0.005133
Complaint (Yes)	0.814899	0.112363	7.252406	4.10E-13
Rejected claim (Yes)	0.311017	0.101016	3.078878	0.002078
Mid-term changes (Increased prem.)	0.501087	0.079264	6.321747	2.59E-10
District (Group 1)	0.181511	0.033888	5.356153	8.50E-08
District (Group 2)	0.312692	0.046483	6.727041	1.73E-11
District (Group 3)	0.564115	0.060679	9.296703	<2E-16
District (Group 4)	0.082149	0.029195	2.813790	0.004896
Income decile (Bottom 70%, NA)	0.096043	0.028981	3.314034	0.000920
No. of objects ($\geq$2)	−0.518593	0.100408	−5.164841	2.41E-07
Type of vehicle (Commercial car)	0.122082	0.044557	2.739883	0.006146
Vehicle age	0.147008	0.015592	9.428656	<2E-16
Fuel (Diesel, Gas, Mixed)	0.087820	0.024725	3.551836	0.000383
Engine displacement^2	−0.032770	0.008526	−3.843359	0.000121
Engine displacement^3	−0.018911	0.006779	−2.789652	0.005277
Vehicle weight^2	−0.022241	0.008147	−2.729840	0.006337

Table B.1: Model summary

- Dispersion parameter ϕ fixed at 1
- Deviance: 59 836.34524

- Degrees of freedom: 86 290
- AIC: 59 944.34524
- AUC: 0.7017718
- Orthogonal polynomials were used for quantitative covariates
- Coefficients in *italic* were kept in the model as explained in section 4.1
- To preserve the confidentiality of the data, the estimate and standard error of the intercept are omitted

Levels included in the intercept: Claim (No); Own damage (No); Payment frequency (Annual); Bonus-Malus (50%); No. other cancellations (0); Tariff (A, B, C, E); Payment method (Direct debit); Missed payments (No); Financial risk score (≤10); Distribution channel (Bancassurance); Complaint (No); Rejected claim (No); Mid-term changes (Same prem., Reduced prem.); District (Group 5); Income decile (Top 30%); Type of vehicle (Passenger car, Van); No. of objects (1); Fuel (Gasoline, Electricity, Unknown).

District groups (group 5 is included in the intercept):

1. Aveiro, Braga, Castelo Branco, Évora, Leiria, Santarém, Viana do Castelo
2. Açores, Beja, Bragança, Portalegre, Viseu
3. Coimbra
4. Guarda, Porto, Setúbal, Vila Real, Unknown
5. Faro, Lisboa, Madeira

Appendix C

R output

C.1 Hosmer-Lemeshow test

```
> library(ResourceSelection)
>
> # Applying the HL test on the training set
> hl.train = hoslem.test(Training_set$Lapsed, Training_set$Prob_Lapse)
> hl.train

    Hosmer and Lemeshow goodness of fit (GOF) test

data:  Training_set$Lapsed, Training_set$Prob_Lapse
X-squared = 5.2389, df = 8, p-value = 0.7318
>
> # Applying the HL test on the test set
> hl.test = hoslem.test(Test_set$Lapsed, Test_set$Prob_Lapse)
> hl.test

    Hosmer and Lemeshow goodness of fit (GOF) test

data:  Test_set$Lapsed, Test_set$Prob_Lapse
X-squared = 10.271, df = 8, p-value = 0.2465

> # Since we are using the test set, the statistic follows instead a
> # chi-square distribution with 10 d.f. The corresponding p-value is then:
> p.value = pchisq(10.271, 10, lower.tail = FALSE)
> p.value
[1] 0.417048
```

C.2 Threshold optimisation

```
> library(PresenceAbsence)
>
> # Computing the optimal thresholds
> opt = optimal.thresholds(Training_set, which.model = 1,
+                           opt.methods = c(1:6,9), threshold = 1001)
> opt
       Method Prob_Lapse
1     Default      0.500
2   Sens=Spec      0.123
3 MaxSens+Spec     0.132
4    MaxKappa      0.184
5      MaxPCC      0.479
6 PredPrev=Obs     0.212
7  MinROCdist      0.120
>
> # Computing the difference between the observed and predicted prevalence
> pred.train = predicted.prevalence(Training_set, threshold = opt$Prob_Lapse,
+                             which.model = 1)
> diff.train = abs(pred.train$Obs.Prevalence - pred.train$Prob_Lapse)
> data.frame(opt$Method, pred.train$threshold, diff.train)
    opt.Method pred.train.threshold   diff.train
1     Default                0.500 0.1180394700
2   Sens=Spec                0.123 0.2632030019
3 MaxSens+Spec               0.132 0.2199110535
4    MaxKappa                0.184 0.0495344205
5      MaxPCC                0.479 0.1162095803
6 PredPrev=Obs               0.212 0.0004516816
7  MinROCdist                0.120 0.2792203280
```

```
>
> # PCC (Percent Correctly Classified) is what we've denoted by Accuracy
> paa.train = presence.absence.accuracy(Training_set,
+               threshold = opt$Prob_Lapse, which.model = 1, st.dev = FALSE)
> paa.train
       model threshold       PCC sensitivity specificity      Kappa       AUC
1 Prob_Lapse     0.500 0.8756602  0.03173869   0.9963991 0.04721762 0.7017718
2 Prob_Lapse     0.123 0.6476188  0.64374942   0.6481724 0.15356075 0.7017718
3 Prob_Lapse     0.132 0.6802326  0.60109188   0.6915551 0.16695802 0.7017718
4 Prob_Lapse     0.184 0.7971949  0.38771167   0.8557793 0.20818837 0.7017718
5 Prob_Lapse     0.479 0.8757528  0.03941890   0.9954062 0.05783159 0.7017718
6 Prob_Lapse     0.212 0.8264732  0.30498751   0.9010816 0.20638822 0.7017718
7 Prob_Lapse     0.120 0.6357477  0.66031276   0.6322332 0.14957482 0.7017718
>
> # Computing the precision (pre) and the F-measure (fme)
> pre.train = pred.train$Obs.Prevalence*paa.train$sensitivity/
+                    (pred.train$Obs.Prevalence*paa.train$sensitivity+
+                    (1-pred.train$Obs.Prevalence)*(1-paa.train$specificity))
> fme.train = 2/(1/paa.train$sensitivity+1/pre.train)
> data.frame(opt$Method, pre.train, fme.train)
    opt.Method pre.train  fme.train
1      Default 0.5577236 0.06005953
2    Sens=Spec 0.2074673 0.31380244
3 MaxSens+Spec 0.2180232 0.31998424
4     MaxKappa 0.2777778 0.32366459
5       MaxPCC 0.5510996 0.07357513
6 PredPrev=Obs 0.3060921 0.30553882
7   MinROCdist 0.2043762 0.31214050
>
> # Applying the thresholds on the test set
> pred.test = predicted.prevalence(Test_set, threshold = opt$Prob_Lapse,
+                                  which.model = 1)
> diff.test = abs(pred.test$Obs.Prevalence - pred.test$Prob_Lapse)
> data.frame(opt$Method, pred.test$threshold, diff.test)
    opt.Method pred.test.threshold   diff.test
```

```
1      Default               0.500 0.102120651
2    Sens=Spec               0.123 0.169857937
3 MaxSens+Spec               0.132 0.133312744
4     MaxKappa               0.184 0.000514721
5       MaxPCC               0.479 0.101091209
6 PredPrev=Obs               0.212 0.033045090
7   MinROCdist               0.120 0.183137739
>
> # PCC (Percent Correctly Classified) is what we’ve denoted by Accuracy
> paa.test = presence.absence.accuracy(Test_set, threshold = opt$Prob_Lapse,
+                            which.model = 1, st.dev = FALSE)
> paa.test
       model threshold       PCC sensitivity specificity      Kappa       AUC
1 Prob_Lapse     0.500 0.8935557  0.01936108   0.9975809 0.02936318 0.6775735
2 Prob_Lapse     0.123 0.7218448  0.49080348   0.7493376 0.14096018 0.6775735
3 Prob_Lapse     0.132 0.7513897  0.45788964   0.7863149 0.15732531 0.6775735
4 Prob_Lapse     0.184 0.8393041  0.24685382   0.9098030 0.15632354 0.6775735
5 Prob_Lapse     0.479 0.8935557  0.02420136   0.9970050 0.03648424 0.6775735
6 Prob_Lapse     0.212 0.8611283  0.19167473   0.9407902 0.15347297 0.6775735
7 Prob_Lapse     0.120 0.7108297  0.50145208   0.7357447 0.13487674 0.6775735
>
> # Computing the precision (pre) and the F-measure (fme)
> pre.test = pred.test$Obs.Prevalence*paa.test$sensitivity/
+                    (pred.test$Obs.Prevalence*paa.test$sensitivity+
+                    (1-pred.test$Obs.Prevalence)*(1-paa.test$specificity))
> fme.test = 2/(1/paa.test$sensitivity+1/pre.test)
> data.frame(opt$Method, pre.test, fme.test)
    opt.Method  pre.test   fme.test
1      Default 0.4878049 0.03724395
2    Sens=Spec 0.1889676 0.27287406
3 MaxSens+Spec 0.2031787 0.28146385
4     MaxKappa 0.2456647 0.24625785
5       MaxPCC 0.4901961 0.04612546
6 PredPrev=Obs 0.2780899 0.22693410
7   MinROCdist 0.1842105 0.26944083
```

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6), 716–723.

[3] Associação Portuguesa de Seguradores (2015). *Insurance Market Overview 14/15.* Available at: `https://www.apseguradores.pt/Portal/Content_Show.aspx?ContentId=2248&PageId=8&MicrositeId=1&CategoryId=70` [Accessed: 21 May 2016].

[4] 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.

[5] Bland, R., Carter, T., Coughlan, D., Kelsey, R., Anderson, D., Cooper, S. and Jones, S. (1997). Customer selection and retention. *Institute of Actuaries General Insurance Convention 1997*, 493–514.

[6] Bond, A. and Stone, M. (2004). How the automotive insurance claims experience affects customer retention. *Journal of Financial Services Marketing* 9 (2), 160–171.

[7] Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications* 36 (3), 4626–4636.

[8] Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O. and Rokach, L., (Eds.) *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer, 853–867.

[9] De Jong, P. and Heller, G. Z. (2009). *Generalized Linear Models for Insurance Data*, 1st ed. Cambridge: Cambridge University Press.

[10] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), 861–874.

[11] Freeman, E. A. and Moisen, G. G. (2008a). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling* 217 (1-2), 48–58.

[12] Freeman, E. A. and Moisen, G. G. (2008b). PresenceAbsence: An R Package for Presence Absence Analysis. *Journal of Statistical Software* 23 (11), 1–31.

[13] Garraio, J. (2015). *Modelação da Taxa de Anulação no Seguro Automóvel.* Master Thesis, Universidade de Lisboa - Faculdade de Ciências.

[14] 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.

[15] Guven, S. and McPhail, M. (2013). Beyond the Cost Model: Understanding Price Elasticity and Its Applications. *Casualty Actuarial Society Forum 2013:Spring*, 1–29.

[16] Hanley, J. A. and McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143 (1), 29–36.

[17] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.

[18] Instituto Nacional de Estatística (2016). *Taxa de desemprego (Série 2011 - %) por Local de residência (NUTS - 2013) e Sexo; Trimestral* [Database], 11 May 2016. Available at: `https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0005598&contexto=bd&selTab=tab2`.

[19] Kleinbaum, D. G. and Klein, M. (2010). *Logistic Regression: A Self-Learning Text*, 3rd ed. New York: Springer.

[20] Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, 1st ed. New York: Springer.

[21] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

[22] Metz, C. E. (1978). Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine* 8 (4), 283–298.

[23] Murphy, K. P., Brockman, M. J. and Lee, P. K. W. (2000). Using Generalized Linear Models to Build Dynamic Pricing Systems. *Casualty Actuarial Society Forum 2000:Winter*, 107–140.

[24] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* 135 (3), 370–384.

[25] Reichheld, F. F. (2003). The One Number You Need to Grow. *Harvard Business Review* 81 (12), 46–54.

[26] Slaughter, S. J. and Delwiche, L. D. (2006). *The Little SAS® Book for Enterprise Guide® 4.1*, 1st ed. Cary, NC: SAS Institute Inc.

[27] Yeo, A. C., Smith, K. A., Willis, R. J. and Brooks, M. (2001). Modelling the effect of premium changes on motor insurance customer retention rates using neural networks. *Lecture Notes in Computer Science* 2074, 390–399.