Predicting the Lapse Rates of AllSecur

Marcel Geschiere
Erasmus University Rotterdam

2017

Abstract

The lapse rate of the clients of AllSecur is predicted by three different methodologies. The current method used by AllSecur is the Generalized Linear Model and serves as benchmark. The two new methodologies are survival analysis and machine learning. I select the best method by evaluating its predictive performance in an out-of-sample dataset. The best survival analysis method is the Cox Proportional Hazards model with variables selected by a Lasso regression. The best machine learning method is the Stochastic Gradient Boosting algorithm. I find that the Stochastic Gradient Boosting algorithm outperforms the GLM and the Lasso regression, and that the GLM outperforms the Lasso regression.
Keywords: Survival Analysis; Machine Learning; Generalized Linear Model; Prediction

[0]I would like to thank Andreas Alfons and Wendun Wang for their useful insights and comments on this research. I also thank the pricing team of AllSecur for the extra guidance and supplying the data.

Contents

1 Introduction

Insurance companies receive premium paid by their clients and they pay out the claims of their clients. Therefore, an insurance company must accurately assess future risk in order to determine correct premium. The premium is determined by many risk factors which must be modelled and for which forecasts have to be made. One of these factors is the expected lifetime of a policy, which is determined by the rate at which clients leave the client base, called the lapse rate. The premium the clients have to pay consists of two parts, namely the technical price and the commercial price. The technical price is the price of the premium which must cover the claims of the client. The commercial price is the price that is offered to the client. This price also covers all the other expenses and potential profit of the company. The expected lifetime of a policy is indirectly part of the commercial price, indirectly in the sense that a company checks whether a change in the commercial price changes the expected lifetime of the policies.

This research is conducted for the car insurance company AllSecur. The current model used by AllSecur to make forecasts of the lapse rates is the Generalized Linear Model (GLM), but it is unknown whether this method produces the best forecasts. This research will focus on two different methodologies to forecast the lapse rates, while the GLM method will serve as benchmark. The two other methods in this research are: survival analysis and machine learning. For the survival analysis two types of models are considered, namely the commonly used Cox Proportional Hazards (CPH) model and the random effects version of the CPH model, the so-called frailty model. The usefulness of one machine learning concept is investigated, namely decision tree algorithms. The different decision tree algorithms are a CART decision tree, a Random Forest (RF) and Stochastic Gradient Boosting (SGB). The reason to use decision trees is twofold: 1) They are relatively easy to implement 2) Fernández-Delgado et al. (2015) find that random forests perform the best of a total of 179 classifiers. These machine learning methods are often labelled as 'off-the-shelf' methods, since there is not much data pre-processing nor manual tuning of the learning procedure needed. The tuning that is needed is data-driven and therefore does not require a subjective intervention of the researcher.

At the end of this research the following two questions are answered:

- What is the best method to forecast the lapse rates?
- Do 'black-box' methods, such as machine learning algorithms provide added value over more interpretable econometric techniques?

I select the best method to forecast the lapse rates based on quantitative and qualitative measures. The first quantitative measure is the predictive performance of

the method in an out-of-sample data set. The second quantitative measure is the computational time of the fit of the model. If the forecast accuracy is only slightly better for one of the machine learning algorithms, but the computation time is significantly greater, it may desirable to choose the faster method. The qualitative measure is the interpretability of the model and its predictions. Interpretability is needed so we can determine the type of clients who are more likely to lapse.

The contribution of this research for AllSecur is clear. AllSecur will have a better forecasting method for the lapse rates or they know that their current forecasting method is very hard to improve. If one of the two new methods outperforms the old method, AllSecur has a better understanding about the expected lifetime of a policy and a fairer price for the premium can be set. Besides this practical contribution, this research also contributes to the academic world in the sense that different methods are compared in terms of their forecasting power.

Survival analysis is a tool not often used by statisticians (except for biostatisticians), but can provide useful insights in more financial problems. Also, the added value of frailty models in terms of forecasting is further assessed. Machine learning may also be interesting for academics due the findings of e.g. Ishwaran et al. (2008) who find that their machine learning algorithms outperform the Cox regression in real and simulated data sets. Gepp and Kumar (2015) on the other hand find that machine learning algorithms and survival models had roughly the same predictive accuracy as discriminant analysis, but they also find that these methods significantly outperform logistic regression for predicting financial distress. Kattan (2003) finds that the Cox model produces better or comparable predictions compared to several machine learning algorithms on urological data sets. The literature learns us that the performance of each type of modelling technique depends heavily on the problem at hand. This research will provide information on the performance of the different methods for a new problem, namely predicting lapse rates.

AllSecur makes a distinction between four different type of lapses: Afterthought, AllSecur, Mid Term Cancellation (MTC) and Renewal lapse. I select the best method by evaluating their performance on these type of lapses. The best methods of the two new methodologies are: the CPH model with variables selected via the Lasso regression and the SGB algorithm. For all the different type of lapses, the Survival Analysis model is outperformed by the benchmark model, the GLM. The GLM on the other hand is outperformed by the SGB algorithm for the MTC, Renewal and AllSecur lapse. For the Afterthought lapse, the results are mixed. Since there are relatively few Afterthought lapses, I conclude that the SGB algorithm outperforms the GLM. The SGB algorithm also outperforms the Survival Analysis model.

The data is described in detail in section 2. This section also defines the different type of lapses. Section 3 explains the different methodologies: survival analysis, machine learning and the GLM. Section 4 discusses the results and section 5 concludes.

2 Data

This section discusses the data and how I construct different data sets from the original data. The data is provided by AllSecur. The data consists of contracts of clients who had a contract in 2014. In this period there are a total of 184,061 contracts and 47,539 lapses.

Survival data for survival analysis is of the form $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$, where y_i denotes the survival time, x_i is a vector of covariates and δ_i equals 1 if the event of interest occurred for individual i and 0 if the event has not occurred. In case $\delta_i = 0$, the survival time is right-censored, i.e., the client has not lapsed. The survival time is the number of days between the starting date of the contract and the cancellation date if the person has lapsed. In the case the person did not lapse, the censored survival time is 365 days. I refer to the data set constructed from this set up as the *overall* data set.

There are 48 variables included in the analysis. The variables can be grouped in four categories. The first group of variables are the individual characteristics. Examples of this type of variable are age of driver, number of years without a claim and mileage. The second type of variables are car specific variables. This includes: age of car, listed price and motor specifications. The third set of variables are determined based on the home address. For each contract holder, information about the neighbourhood is present. Examples are: mode income, average age and degree of urbanisation. The last type of variables are competition variables. For every customer we know the premium he has to pay if he decides to sign a contract with one of the competitors. Based on this information AllSecur constructs several other variables, such as competitive indices and a ranking. Appendix A shows a list of all different variables. For the regression models in survival analysis, I covert a factor with k levels to $k-1$ dummy variables. This results in a total of 684 variables.

AllSecur makes a distinction between four different lapses.

1. Afterthought lapse: the client cancels the contract within 30 days after the first contract is made;

2. Mid Term Cancellation (MTC) lapse: the client cancels the contract before expiration, but after the first 30 days;

3. Renewal lapse: the client lapses 30 days before or after the contract expires;

4. AllSecur lapse: AllSecur cancels the contract because the client defaults on his payments.

Around the expiration date of the contract AllSecur sends a new proposal contract to the client. This is a natural moment for client to think about changing from insurance company. The new offered premium can be higher (due to a claim) or lower (due to one extra claim free year) than the premium of the previous year. In case of a Renewal lapse, there are three extra variables available, namely the new offered premium, the absolute and relative difference between the old and new premiums. Of these four lapses, the MTC and Renewal lapses are the most interesting to accurately forecast since they account for approximately 64% and 19% of the lapses respectively.

I construct four other data sets corresponding to the four different type of lapses. In each dataset, I only select the contracts which have exposure to that particular type of lapse. The second contract of a client has for example no exposure to the Afterthought lapse, since a lapse in the first 30 days after renewal is defined as a Renewal lapse. I refer to these data sets as the *Afterthought* (75,464), *MTC* (181,692), *Renewal* (143,674) and *AllSecur* (181,851) data set, where the number in the parentheses represents the number of observations in each data set.

3 Methodology

3.1 Survival Analysis

Survival analysis is a method to analyse the expected duration until the event of interest happens and to describe the effects of variables on the survival time. Here, the event of interest is the moment a customer cancels his contract with AllSecur.

I start with introducing some concepts of survival analysis. The *survival function* $S(t)$ is defined as follows: let T denote the time until the event of interest happens and let it be a continuous random variable with cumulative distribution $F(t)$ on interval $[0, \infty)$, then

$$S(t) = P(\{T > t\}) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t), \tag{1}$$

which is a monotonically decreasing function. The survival function captures the probability that the event of interest has not yet happened beyond a specified

point in time. The lapse rate at time t is thus given by $1 - S(t)$. Closely related to the the survival function is the *hazard rate* $h(t)$ and is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t}, \tag{2}$$

which is interpreted as the probability that the subject experiences the event within a small time frame, given that this subject has survived until the beginning of that time frame. A more intuitively interpretation is that the hazard rate represents the negative tangent of $S(t)$. Furthermore, the hazard function and the survival function are related in the following way

$$S(t) = \exp\{-\int_0^t h(u)\mathrm{d}u\} = \exp\{-H(t)\}, \tag{3}$$

where $H(t) = \int_0^t h(u)\mathrm{d}u$ denotes the cumulative hazard function.

This section about survival analysis is build up in the following manner. Section 3.1.1 introduces the Cox Proportional Hazards (CPH) model, the most commonly used model in survival analysis. Section 3.1.2 extends the CPH model by considering the random effects variant of the CPH model, the so-called frailty models.

3.1.1 Cox Proportional Hazards Model

The CPH model models the hazard rate $h(t)$ and consists of two parts. 1) The baseline hazard function $h_0(t)$, which is the risk of experiencing the event of interest for a baseline level of covariates and 2) parameter effects, which describes how the hazard is varied due to the covariates.

The CPH model, introduced by Cox (1972), is given by

$$h(t) = h_0(t) * \exp\{x'\beta\}, \tag{4}$$

where β is a vector of coefficients and x is a vector containing the p covariates. Suppose that the uncensored event times are given by $0 < t_1 < \cdots < t_m$, the coefficients of the CPH model are estimated by maximizing the partial likelihood, which is given by

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp\{x'_{i(j)}\beta\}}{\sum_{i \in R_j} \exp\{x'_i\beta\}}, \tag{5}$$

where R_i denotes the risk set (i.e. the individuals which have not experienced the event of interest at time t_i). The partial likelihood is based only on the order in which the events of interest occur instead of the actual times at which the events occur (Cox (1975)). This parameter estimation method has the advantage

that it does not require a pre-specified functional form for $h_0(t)$, since it cancels in the numerator and denominator in the derivation of the partial likelihood of equation (5).

The cancellation of $h_0(t)$ is a strong advantage of the CPH model. Other survival analysis models often need to assume a distribution for $h_0(t)$, the Weibull distribution for example. This is also the reason to choose the CPH model and not the other very popular survival analysis model, the Accelerated Failure Time (AFT) model. It is out of the scope of this paper to discuss this model in detail, but an important assumption that has to be made in this model is that an assumption has to be made about the distribution of the survival time. In other words, the AFT model is a fully parametric model, while I prefer the semiparametric CPH model.

Note that the partial likelihood of equation (6) is the partial likelihood when there are no ties in the survival times, i.e. no multiple events of interest at a specific time point. If there are tied event times, the true partial (log) likelihood becomes very time-consuming to compute. There are tied event times present in all the data sets considered in this research and therefore I use the Breslow approximation (Breslow (1974)) of the partial likelihood. Suppose that there are d_j tied survival times at the j^{th} survival time and that D_j denotes the event set at the j^{th} distinct survival time. The Breslow approximation is then given by

$$L(\beta) \approx \prod_{j=1}^{m} \frac{\exp\{\sum_{l \in D_j} x_l' \beta\}}{[\sum_{l \in R_j} \exp\{x_l' \beta\}]^{d_j}}. \tag{6}$$

There are two key assumptions that have to be satisfied in order to make a CPH model that makes statistical sense. The first assumption is that the censoring is non-informative. Non-informative censoring occurs when the distribution of censorship times is not influenced by the distribution of survival times. The violation of this assumptions is often a problem in medical studies which have a long time span, as Hakulinen (1982) points out, since ageing people are more likely to die due to other causes than the cause researched in a study. There can be other causes for informative censoring, but there is no clear indication that informative censoring occurs for problem researched in this paper.

The second key assumption is that the covariates meet the PH assumption. The PH assumption means that the survival functions of different individuals are proportional over time. This means that the hazard ratios for individual i and j are independent with respect to time $\forall i, j$ and for every covariate. Assume there is only one covariate, the hazard ratio between individual i and j is then

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\{x_{1i}\beta_1\}}{h_0(t)\exp\{x_{1j}\beta_1\}} = \exp\{(x_{1i} - x_{1j})\beta_1\}, \tag{7}$$

which is independent with respect to time. Now assume that there is a second covariate, which is a time-dependent version of the first covariate, i.e., $x_{2i}(t) = g(t)x_{1i}$, where $g(t)$ is some time-dependent function. The hazard ratio is now

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp\{x_{1i}\beta_1 + x_{2i}\beta_2\}}{h_0(t)\exp\{x_{1j}\beta_1 + x_{2j}\beta_2\}} = \exp\{(x_{1i} - x_{1j})\beta_1 + g(t)(x_{1i} - x_{1j})\beta_2\}, \quad (8)$$

which is not independent with respect to time due to the presence of $g(t)$. Equation (8) can be used to test whether the PH assumption holds for a specific covariate. The test tests whether $\beta_2 = 0$. When $\beta_2 = 0$ the PH assumption holds and when $\beta_2 \neq 0$ the hazards are not proportional.

The inclusion of the time dependent covariate in equation (8) alters the estimation method slightly. The Breslow approximation of the partial likelihood (equation (6)) is still maximized, but the values for the covariates can now change each time the risk set R_i changes. Equation (8) is both the test and the solution for non proportional hazards. If a certain covariate violates the PH assumption, the time-dependency can be added to the model in order to ensure that the PH assumption holds. The only remaining issue is the functional form of $g(t)$. To avoid numerical problems, often $g(t) = ln(t)$ is assumed (e.g. Quantin et al. (1996)). This is however only a technical solution and has no theoretical foundation.

3.1.2 Frailty Model

The CPH model is based on the assumption that the survival data is independent and that the survival times of all the individuals come from the same distribution. This assumption can be restated as the assumption that the subjects in the data set are homogeneous. However, this assumption may be unrealistic if we look in the longitudinal direction. In bad economic times, people are more inclined to search for a cheaper insurance compared to when the economic climate is good. Therefore, one can expect more lapses during a bad state of the economy and less during a good state of the economy. Another example is that after an effective marketing campaign people are more inclined to remain with AllSecur. There are many other possible explanations for heterogeneity over time, a new budget competitor entered the market or there was really good or really bad publicity in a particular period, etc.

Vaupel et al. (1979) suggest to use a random effects model for durations, which they named the (univariate) frailty model, to counter the problem of heterogeneity in a population where the heterogeneity cannot be captured by a covariate. The idea is that different subjects in the data set are more 'frail' tent to encounter the event of interest earlier than those who are less 'frail'. In this model the hazard rate is conditional on the random effect Z, which is called the frailty. Let

Z be an unobservable non-negative random variable which has a multiplicative effect on the baseline hazard function $h_0(t)$, i.e.,

$$h_i(t, Z_i) = Z_i * h_0(t), \tag{9}$$

Z_i is thus a scaling variable for subject i of the baseline hazard function. The CPH model of equation (4) can be extended to a Cox frailty model in the following manner

$$h_i(t|Z_i) = Z_i * h_0(t) * \exp\{x_i'\beta\}, \tag{10}$$

this model is thus a generalization of the CPH. Consequently, the conditional survival function can be found in the same manner as in equation (3)

$$S_i(t|Z_i) = \exp\{-Z_i H_i(t)\}, \tag{11}$$

where $H_i(t)$ denotes the cumulative hazard function at time t for individual i. Note that $H_i(t) = \int_0^t h_i(u)\mathrm{d}u = \int_0^t h_0(u)\exp\{x_i'\beta\}\mathrm{d}u = \exp\{x_i'\beta\}\int_0^t h_0(u)\mathrm{d}u = \exp\{x_i'\beta\}H_0(t)$. To make the notation clearer, denote $\exp\{x_i'\beta\}$ as κ.

The conditional survival function describes the model at the individual level, however models at the individual level are not observable. Therefore, the model at the population level must be considered. To find the unconditional survival function, the frailty term must be integrated out the conditional survival function,

$$S(t) = \int_0^\infty S(t|z)g(z)\mathrm{d}z, \tag{12}$$

where $g(z)$ denotes the frailty distribution. To find the solution of equation (12), the *Laplace transform* can be used. Let $\mathcal{L}(s)$ denote the Laplace transform for variable s, then the Laplace transform is given by

$$\mathcal{L}(s) = \int_0^\infty \exp\{sx\}f(x)\mathrm{d}x. \tag{13}$$

If one thinks of $f(x)$ as the frailty distribution $g(z)$ and s as the $\kappa H_0(t)$, the following expression is obtained

$$S(t) = \int_0^\infty \exp\{\kappa H_0(t)z\}g(z)dz = \mathcal{L}(\kappa H_0(t)) \tag{14}$$

which is the same as plugging equation (11) into equation (12). This useful relationship was first exploited by Hougaard (1984) and Hougaard (1986). Because of this relationship, a distribution for Z is chosen for which an explicit Laplace transform exists.

The distribution that is most used as frailty distribution is the Gamma distribution (Clayton (1978)). The popularity of the Gamma distribution is due

to a couple of, purely mathematical, reasons. First of all, the frailty term Z must be non-negative, otherwise the hazard rate becomes negative (see equation (10)) and the survival function is not a monotonically decreasing function. The Gamma (along the Log-Normal) distribution is one of the most commonly used distributions to model non-negative variables. Second, if we assume a Gamma distribution with shape parameter k for the frailty term, the frailty term for the survivors is also gamma distributed with shape parameter k, implying that the hazard ratio is independent of time. Third, the Gamma distribution has easy derivatives of the Laplace transform.

The Gamma distribution has two parameters, the shape parameter k and the scale parameter θ. For identification issues, it makes sense to restrict the parameters as follows $k = \theta$, so $Z \sim Ga(\frac{1}{k}, \frac{1}{k})$. It then follows that $\mathrm{E}[Z] = 1$, which makes sense since this way there is no bias in the hazard function due to the frailty. For this distribution, the Laplace transform to find the unconditional survival function is given by

$$S(t) = \mathcal{L}(\kappa H_0(t)) = (1 + k\kappa H_0(t))^{-\frac{1}{k}}. \tag{15}$$

There are other distributions that can be used as frailty distribution with easy derivatives of the Laplace transform. I will give an overview of distributions which are used the most in the literature. One of them is the Log-normal distribution, but this distribution is mostly used in modelling multivariate frailty models (McGilchrist and Aisbett (1991)). The compound Poisson distribution can also be used. The reason to not use this distribution is that it yields a subgroup which will never experience the event of interest (Aalen (1992)). There are three other type of distributions that are often used as frailty distribution and can be useful to consider: the inverse Gaussian, a positive stable or a power variance function distributions. However, the Gamma distribution is used because it is the most commonly used frailty distribution in the literature.

3.2 Variable Selection

To apply the survival analysis methods of section 3.1 a number of covariates from the data set must be chosen to include in the survival models. I use two variable selection procedures. In section 3.2.1 I describe the commonly used stepwise regression and in section 3.2.2 I describe how penalized models can be used for variable selection.

3.2.1 Forward Selection

A common used method for variable selection is stepwise regression based on a statistical metric such as an information criteria or on the R^2. To estimate the

survival analysis models I use a forward stepwise regression procedure based on the Bayesian Information Criterion (BIC) of Schwarz (1978).

The BIC is defined as

$$\text{BIC} = -2\ln(\hat{L}) + p\ln(n), \tag{16}$$

where $\hat{L}$ denotes the estimated likelihood, p the number of parameters and n the number of observations in the model. Adding parameters to a model increases the likelihood of the model, but also increases the chances of overfitting resulting in poor predictive performance. To reduce the chance of overfitting, the BIC adds a penalty factor which penalizes the number of parameters included in the model.

The forward stepwise regression procedure is an iterative procedure. To initialize the procedure, one starts with an empty base model ($p = 0$). In each iteration one of the p variables is added to the base model and the BIC is computed. The variable is then deleted from the model and another variable is added. For this model the BIC is again computed. The deletion and addition of variables in the model is repeated until all p variables are included once in the model. The model with the lowest BIC is chosen as base model for the next iteration. In the next iteration, repeat the addition and deletion of the remaining $p - 1$ variables to the base model ($p = 1$). This procedure is repeated until all p variables are included in the model or until an extra variable does not provide any added value. Sometimes the procedure must be stopped earlier, namely if p exceeds n. However, this is not a problem in this research, see section 2.

This approach to select the variables to include in the model is intuitive and relatively easy to implement. However, there is ample evidence that forward selection methods leads to biased parameter estimates and problems with multiple hypothesis testing (see for example Wilkinson (1979), McIntyre et al. (1983) and Huberty (1989)). Another disadvantage of this variable selection method is that this simple method leads to an over simplified model of the real model of the data.

3.2.2 Penalized Models

The forward variable selection procedure of the previous subsection is commonly applied due to its simplicity, but also has some serious drawbacks as already mentioned. Luckily, there are alternatives such as penalized regression models.

Penalized regression models add a penalty term to a cost function. Suppose that $C(\beta|x)$ denotes the cost function (e.g. residual sum of squares) for estimated parameters β. The objective function then becomes

$$\min\{C(\beta|x) + \lambda P(\beta)\}, \tag{17}$$

where $P(\beta)$ is the penalizing function and λ determines the size of the trade-off between the cost function and the penalizing function. If $P(\beta) = 2\|\beta\|_1 = 2\sum_{j=1}^{p}|\beta_j|$ is chosen, we end up with the *Lasso* regression (see Tibshirani (1996) for linear regressions and Tibshirani (1997) for time-to-event data). If $P(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is chosen, we end up with the *Ridge* regression. These two regressions can be combined, i.e. include both penalizing functions in the objective function, to form the *Elastic Net* regression (Zou and Hastie (2005)). The Elastic Net is characterized by the minimization of the following objective function

$$\min\{C(\beta|x) + \lambda[\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1]\}, \tag{18}$$

where α is a mixing parameter determining to what extend the regression is a Lasso or a Ridge regression. For α equal to 1, the regression is the Lasso regression and as α approaches 0, the regression becomes the Ridge regression.

I use the Lasso regression as a method for variable selection. To see how the Lasso regression applies variable selection, consider the follow example. Suppose we have the cost function of the linear model, i.e., $C(\beta|x) = (y-\beta x)^T(y-\beta x)$ and the Lasso penalizing function $P(\beta) = 2\|\beta\|_1$. Equation (17) then becomes: $\min\{y^Ty - 2y^T\hat{\beta}x + x^T\hat{\beta}\hat{\beta}x + 2\lambda|\hat{\beta}|\}$. If we assume $\hat{\beta} > 0$, the solution to this minimization function is $\hat{\beta} = \frac{y^Tx-\lambda}{x^Tx}$. If we apply a relatively small value for λ, i.e. $0 \leq \lambda < y^Tx$, $\hat{\beta}$, is set to a non zero value. However, if we increase λ to y^Tx it is clear that $\hat{\beta}$ goes to zero. Setting λ to a value greater than y^Tx does not make $\hat{\beta}$ negative, because the solution to the minimazation function now changes to $\hat{\beta} = \frac{y^Tx+\lambda}{x^Tx}$. The flip in sign before λ is due to the absolute value in the Lasso penalizing function. So if we set λ to a value greater than y^Tx, there is an increase in both $P(\hat{\beta})$ and $C(\hat{\beta}|x)$, which can not be the optimal solution for the minimization function. Therefore, $\hat{\beta}$ will not become negative, but equal to 0. The same line of reasoning holds when we assume $\hat{\beta} < 0$.

The ridge regression is unable to set the coefficients equal to zero due to the penalizing function $P(\beta) = \|\beta\|_2^2$. If we use this penalizing function together with the cost function of the linear model in equation (17), we end up with: $\min\{y^Ty - 2y^Tx\hat{\beta} + \hat{\beta}^Tx^Tx\hat{\beta} + \|\hat{\beta}\|_2^2$. The solution this minimization function is $\hat{\beta} = \frac{y^Tx}{x^Tx+\lambda}$. From this follows that adjusting λ is not able to set $\hat{\beta}$ equal to zero. However, it is only able to shrink $\hat{\beta}$ to zero as λ increases.

The cost function for the CPH of section 3.1.1 is the negative partial likelihood divided by the number of observations. The purpose of the penalized models in this research is only variable selection. Since the Ridge regression is unable to set the coefficients to zero, this regression is not suitable as a variable selection procedure. On the other hand, Lasso automatically applies parameter shrinkage and variable selection. Therefore, α is set 1 in equation (18).

Although the Lasso regression is a powerful method for variable selection, it also has its drawbacks. If there are more covariates (p) than samples (n), the Lasso regression is only able to set at most n covariates equal to a non-zero value. However, in this paper this is not an issue. Another disadvantage is that the Lasso regression does not utilize any correlation between the covariates and is in general inconsistent for variable selection.

To see why the Lasso regression is in general inconsistent, suppose we have p covariates of which p_0 are relevant. Furthermore, define a positive definite matrix $\mathbf{M}$ as $\frac{1}{n}X^TX$, where n is the number of observations in the design matrix X. We can break $\mathbf{M}$ down to:

$$M = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

where Q_{11} consists of the p_0 relevant predictors. The necessary condition for the Lasso regression to be consistent is: There must be a sign vector $s = (s_1, \ldots, s_{p_0})^T$x, $s_j = 1$ or -1, such that $|Q_{21}Q_{11}^{-1}s| \leq 1$. (See Zou (2006) for more details.)

If the variable selection is inconsistent, it is then advisable to use the adaptive Lasso of Zou (2006). He suggests to adjust the penalizing function to counter this inconsistency problem. The penalizing function $P(\beta) = \sum_{j=1}^{p} |\beta_j|$ is changed to $P(\beta) = \sum_{j=1}^{p} \hat{w}_j|\beta_j|$, where $\hat{w}_j$ is defined as $\frac{1}{|\hat{\beta}_j|^\gamma}$, with $\gamma > 0$ and $\hat{\beta}_j$ is a $\sqrt{n}$ consistent estimator, such as the estimators resulting from a Ridge regression. It is also shown that the adaptive Lasso enjoys the oracle properties, i.e., it is consistent in variable selection and satisfies asymptotic normality. To estimate the adaptive Lasso, I use two steps. First I estimate a Ridge regression, i.e., I set α equal to 0 in equation (18). The resulting estimated parameter $\hat{\beta}_j$ is then used to determine the weight $\hat{w}_j = \frac{1}{|\hat{\beta}_j|}$. I set γ equal to one since this value is often chosen in the literature (e.g. Ivanoff et al. (2016)). In the final step, I estimate a Lasso regression with the altered penalizing function as described above.

The prediction error of the penalized models are estimated using k-fold cross-validation. The prediction error is then used to determine the correct value for the tuning parameter (λ for penalizing regressions). Cross-validation is a method to counter problem such as overfitting and is mostly used when forecasting is the goal of the researcher.

The procedure for k cross-validation for penalized models is as follows:

1. Create k random equal sized subsets of the data.

2. Pick one of the k subsets as test set and the remaining $k - 1$ subsets as training set.

3. Calculate the deviance.

4. Repeat step 2 and 3 k times where each time another subset is the test set.

5. Repeat step 2, 3 and 4 with a different value for the tuning parameter λ.

The deviance that is calculated in step 3 is equal to minus two times the log partial likelihood ratio for the survival analysis models (Cox and Snell (1989)).

By using different values for λ the optimal value for λ and number of nonzero coefficients can be determined. One way to do this is to chose the model with the lowest deviance. However, this can result in a model with many covariates which is prone to overfitting. Therefore, I use the one standard error rule. This means that the chosen model is the most parsimonious model whose deviance is no more than one standard deviation larger than the overall lowest deviance. The intuition behind this rule is that this model is not significantly worse than the model with the lowest deviance. Also, in general people prefer parsimonious models, because these models are easier to interpret. This rule is recommended by Breiman et al. (1984) and they show that it is a successful rule in screening out noise variables.

3.3 Machine Learning

Decision tree learning uses a decision tree to model the data and ultimately predict the value of a target variable based on prediction variables. A decision tree is a tree in which each leaf node (i.e. final node on the tree) contains the value of the target variable. All the internal nodes (i.e. non-leaf nodes) are labelled with one of the predictors. At each node, the path to the node one layer lower in the tree is determined based on the value of the predictor. This procedure is repeated from the top of the tree until the leaf node is reached. The target variable is a Boolean variable indicating whether the client has or has not lapsed. This means that the mean score at each leaf node is the percentage of clients that did lapse.

Section 3.3.1 discusses the CART algorithm. This algorithm constructs a single tree, while the ensemble methods of sections 3.3.2 and 3.3.3 construct multiple decision trees. The ensemble methods are Random Forest (RF) and Stochastic Gradient Boosting (SGB) respectively.

3.3.1 CART

This section discusses the decision tree technique CART (Classification And Regression Tree), introduced by Breiman et al. (1984). CART uses the Gini impurity measure to determine how the splits are made. Gini impurity measures how often

a randomly chosen observation would be incorrectly labelled if it was randomly labelled following the distribution of the labels in the training set. Denote $p(j|t)$ as the probability that an observation in node t belongs to class j, $j \in \{0, m\}$, the Gini impurity is then given by $i(t) = 1 - \sum_{j=0}^{m} p^2(j|t)$. Denote p_L and p_R as the fraction of observations in the left and right node respectively. CART selects the splits that maximizes the decrease in impurity $i(t) - p_L i(t_l) - p_R i(t_R)$ at each node, this is an example of a greedy algorithm, i.e., it makes the locally optimal decision. The procedure is stopped when there are less than the minimum number of observations required in a node. Another strategy is to construct every potential tree possible based on the set of covariates available and then choose the tree that minimizes the cost function. This tree would find the absolute minimum of the cost function, while a greedy algorithm can get stuck in a local minimum. However, if the set of covariates is large, a greedy algorithm is a lot faster. For this reason, I choose to use the greedy CART algorithm.

The resulting tree is possibly quite large and not really interpretable. Furthermore, a large tree is prone to overfitting and leads to inaccurate predictions. To counter these problems, I use pruning to decrease the size of the tree and the probability of overfitting. Pruning only removes the sections of the decision tree that hardly provide additional information. This reduces the chance of overfitting and also makes the tree more interpretable. Denote $C(T)$ as the cost of tree T, the cost for complexity parameter α is then given by

$$C_\alpha(T) = C(T) + \alpha|T|, \tag{19}$$

where $|T|$ denotes the number of leaf nodes. The complexity parameter α is estimated using k-fold cross validation.

3.3.2 Random Forest

To understand the RF (Breiman (2001)) algorithm, two other techniques must be understood. The first one is the bootstrap method. The bootstrap is a method to estimate a quantity (e.g. the mean) of a data sample. The bootstrap method randomly selects different sub-samples with replacement, i.e., each observation in the data set can be present in multiple sub-samples. The quantity of interest is then estimated for each sub-sample. Suppose that we use B sub-samples, we can then estimate the quantity of interest B times. These B estimates represent the empirical distribution of the quantity of interest and estimates for e.g. the standard errors can be made.

The second technique is bagging. Bagging is the application of the bootstrap method applied to high-variance machine learning algorithms. Bagging uses multiple machine learning algorithms in order to improve the accuracy and stability

of the individual algorithms. Furthermore, it reduces variance and the probability of overfitting. The CART algorithm for example has a high variance due to the fact that it depends heavily on the training set that is used. A couple of extra outliers in the training sample can change the whole tree. The bagging procedure applied to the CART algorithm is as follows:

1. Create B sub-samples;
2. Fit a CART tree on each different sub-sample to obtain the ensemble of trees $\{CART_b\}_1^B$;
3. Use each individual tree to make a prediction;
4. The final prediction is obtained by taking the average prediction (for regression trees) or to take the value that is predicted the most (for classification trees).

By using many different trees, the effects of outliers will be averaged out and hence, the predictions are less volatile than the predictions of a single CART tree.

Now we can turn to the RF algorithm. RF is a modification of the bagging technique applied to decision trees. In this research I will use the CART decision tree in the RF algorithm. The difference between RF and bagging lies in the amount of different models that are fitted. Bagging uses the same model throughout for each bootstrap sample, but random forests constructs many different models. In fact, a random forest constructs another model for each bootstrap sample. The difference between the models is which covariates are included in the model. Suppose there are in total p covariates, RF randomly selects m covariates and grows the decision tree. It then repeats this for all B bootstrap samples to obtain the ensemble of trees $\{CART_b\}_1^B$. The value for m is estimated using k-fold cross validation]. The prediction for the target variable, say x, is obtained in the same manner as with bagging. That is, for regression the prediction based in the B trees is $\hat{f}^B(x) = \frac{1}{B}\sum_{b=1}^{B} CARTb(x)$, and for classification $\hat{C}^B(x) = \text{mode}\{\hat{C}_b(x)\}_1^B$, where $\hat{C}_b(x)$ denotes the estimated classification according to tree b.

3.3.3 Stochastic Gradient Boosting

SGB is a method proposed by Friedman (2002) to enhance the gradient boosting technique. Gradient boosting produces a prediction model based on multiple weak prediction models. It then builds the final model in a stage wise manner just as other boosting methods do, such as the popular AdaBoost algorithm of Freund and Schapire (1997).

The goal of (stochastic) gradient boosting is to find a function $y = F^*(x)$ that maps the covariates x to the response variable y in such a way that the expected value of a cost function $C(y, F(x))$ is minimized, i.e.

$$F^*(x) = \arg\min_{F(x)} \mathrm{E}_{y,x}[C(y, F(x))]. \tag{20}$$

Boosting methods approximate $F^*(x)$ by an additive expansion which is given by

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m), \tag{21}$$

where $h(x; a_m)$ are the 'base learner' functions, with parameters a_m. Given an initial guess for $F_0(x)$, the coefficients and parameters can be estimated in the following manner stage wise manner

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{n} C(y_i, F_{m-1}(x_i) + \beta h(x_i; a)) \tag{22}$$

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m). \tag{23}$$

Gradient boosting approximates equation (22) in the following way. First, the base learner $h(x; a)$ is fitted by the CART decision tree of section 3.3.1 to the pseudo-residuals, which are defined as

$$\bar{y}_i m = -\Big[\frac{\partial C(y_i, F(x_i))}{\partial F(x_i)}\Big]_{F(x)=F_{m-1}(x)}. \tag{24}$$

Given these pseudo-residuals, the optimal value of β_m is found by

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{n} C(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)). \tag{25}$$

The following step is to compute the multiplier γ by solving the following equation

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} C(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{26}$$

and then update the model according to

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{27}$$

To obtain the SGB method, one small modification must be made. The base learner in gradient boosting, which was fitted by a CART decision tree

to the pseudo-residuals of equation (24) is estimated using the whole sample. SGB on the other hand, selects a subsample of the data, which are drawn at random without replacement. Friedman (2002) finds that the accuracy of gradient boosting is substantially improved by this small modification. He also finds that the fraction of observations to be used in the subsample is in the range of $[0.5; 0.8]$ to obtain good result. Therefore, the size of the subsample is commonly set to half the size of the whole data set.

SGB requires a couple of parameters to be set or to be estimated by cross validation. The parameters to be estimated by k-fold cross validation are the number of boosting iterations (the number of constructed trees) and the depth of the trees (the number of layers in the tree). The learning rate can also be estimated by cross validation, however I set the learning rate equal to 0.01. I choose this value because low values makes the final model more robust. On the other hand, a low learning rate requires more boosting iterations and this increases the computational time.

3.4 Generalized Linear Model

The classical linear model can be written as

$$Y = \mathrm{E}[Y|X] + \epsilon, \mathrm{E}[Y|X] = X\beta, \tag{28}$$

where Y is the response variable, X consists of p covariates, β consist of the estimated coefficients and the ϵ is a random shock. The linear model has some underlying assumptions. McCullagh and Nelder (1989) define these assumptions as follows:

1. *Random Component Linear Model:* Each component of Y is independent and is normally distributed. The mean of the normal distribution may vary for each component, but the variance must be the same.

2. *Systematic Component Linear Model:* The combination of the p covariates $(x_1, \ldots x_p)$ give the linear predictor η: $\eta = X\beta$.

3. *Link Function Linear Model:* The random and the systematic component are linked by a link function. The link function for the linear model is the identity function so that: $\mathrm{E}[Y|X] \equiv \mu = \eta$

These assumptions are not easily satisfied in many different problems. Normality for the response variable Y can not always be guaranteed or the additivity effects implied by the second and third assumption may not be realistic. The effect could also be multiplicative for example and another link function may be

more realistic. Nonetheless, an assumption about the type of linearity must be made.

Generalized Linear Models (GLMs) have the following set of assumptions:

1. *Random Component Generalized Linear Model:* Each component of Y is independent and follows a distribution of the family of exponential distributions.

2. *Systematic Component Generalized Linear Model:* The combination of the p covariates $(x_1, \ldots x_p)$ give the linear predictor η: $\eta = X\beta$.

3. *Link Function Generalized Linear Model:* The random and the systematic component are linked by a link function g. The link function is a differentiable and monotonic function such that: $\mathrm{E}[Y] \equiv \mu = g^{-1}(\eta)$

From this set of assumptions and the set of assumptions for the linear model, it follows that the linear model is a special case of the GLM. More specifically, if we choose a normal distribution for the random component and the identity function for the link function, GLM boils down to the linear model of equation (28).

At the moment AllSecur uses a GLM with a logit link function and assumes that the random component follows a binomial distribution. The binomial distribution supports binary dependent variables, such as individuals that have not lapsed versus individuals that did lapse. The logit link function is chosen because this link function has the characteristic that the effects of the covariates are multiplicative related to the dependent variable. Furthermore, the logit link function in combination with a binomial distribution predicts a probability and is thus the appropriate choice for predicting the lapse probability. This method is the benchmark for the other methods.

The parameters in the GLM can be found by using maximum likelihood (ML). To find ML estimates for β we need to use an iterative procedure.

1. Make an initial guess for $\hat{\beta}$

2. Use a polynomial approximation of the likelihood

3. Calculate the difference, D, between step 1 and 2

4. Update the initial guess with the approximation of step 2 and repeat until $D < \tau$, where τ is the convergence criterion

4 Results

This section discusses the results. All the models are fitted on an estimation sample which corresponds to 70% of the data. Predictions are made for the hold-out sample, which consists of the remaining 30%. Section 4.1 discusses the results for the survival analysis techniques and section 4.2 does this for the machine learning techniques. Section 4.3 compares the results of these two methodologies and the GLM.

4.1 Survival Analysis

This section presents the results for the survival techniques and discusses them. I start with comparing the three different methods for selecting variables. The variables chosen by the best method are then included in a frailty model and I compare this model with the best CPH model. Lastly, I investigate the PH assumption for the best method.

The goodness of fit of the survival analysis models is determined by two metrics. The first metric is an in-sample metric called concordance index (Harrell et al. (1982)). The concordance is the proportion of pairs of individuals of which the individual with the highest hazard, experienced the event of interest the earliest. This metric measures the ability of the model to predict which individual of a pair dies earlier. The higher the concordant index, the better the in-sample fit.

The concordance index for the survival analysis models is given in table 1 with the standard deviation in parentheses. In the following, I use a significance level of 5%. For the MTC lapse the Lasso and the adaptive Lasso significantly outperform the forward selection method. There is no significant difference between these two methods. The adaptive Lasso performs rather poorly for the Renewal lapse. This method is significantly outperformed by both the Lasso and the forward selection method. The Lasso method preforms the best for the Renewal lapse. For the Afterthought lapse the adaptive Lasso significantly outperforms the forward selection method, but is unable to significantly outperform Lasso method. In contrast with the MTC and Renewal lapse, the Lasso is now unable to significantly outperform the forward selection method. Lastly, the Lasso and adaptive Lasso are not significantly different for the AllSecur lapse, but both significantly outperform the forward selection method. Overall, I conclude that the Lasso method is the best method based on the concordant index.

The second metric assesses the predictive performance of the models and is the Brier score (Brier (1982)). The Brier score is a scoring rule that can be used to asses the performance of any model that makes probabilistic predictions and

Table 1: **Concordant Index for the Cox Proportional Hazards Models**
This table shows the concordant index for the variable selection methods for each different type of lapse. The standard deviation is given in parentheses.

	Lasso	**Adaptive Lasso**	**Forward Selection**
MTC	0.672	0.671	0.639
	(0.002)	(0.002)	(0.002)
Renewal	0.694	0.513	0.630
	(0.004)	(0.001)	(0.004)
Afterthought	0.697	0.733	0.645
	(0.024)	(0.022)	(0.021)
AllSecur	0.824	0.821	0.769
	(0.005)	(0.005)	(0.005)

is defined as

$$BS = \frac{1}{n}\sum_{i=1}^{n}(Y_i - P_i)^2, \tag{29}$$

where Y_i is a 0/1 variable indicating whether individual i has lapsed, P_i is the predicted lapse probability. The lower the Brier score, the better the forecasting power of the model. To calculate the Brier score, the lapse probability must be determined. Since the lapse probability is defined as $1 - S(t)$, where $S(t)$ is the survival curve, I must choose a suitable t. I set t equal to the average survival time of the estimation sample.

Table 2 shows the Brier scores of the different variable selection methods and for each type of lapse. For the MTC lapse, the Lasso method produces the lowest Brier score. The differences with adaptive Lasso and forward selection are small. The differences for the Renewal lapse are even smaller, the Lasso has the lowest Brier score, then the forward selection and lastly the adaptive Lasso. For the Afterthought lapse the differences are only visible at four decimals, the Brier score for only the forward selection differs, but not that much. This is somewhat expected due to the results of table 1, where the three methods were not significantly (at a 5% significance level) different from each other. Lastly, for the AllSecur lapse the Lasso and the adaptive Lasso produce the same Brier score. Both the Lasso and adaptive Lasso outperform the forward selection method.

Table 2: **Brier Score for the Cox Proportional Hazards Models**
This table shows the Brier for the variable selection methods for each different lapse.

	Lasso	**Adaptive Lasso**	**Forward Selection**
MTC	0.1396	0.1400	0.1432
Renewal	0.0544	0.0559	0.0548
Afterthought	0.0459	0.0459	0.0461
AllSecur	0.0242	0.0242	0.0255

It is clear that Lasso method outperforms the forward selection method. The adaptive Lasso performs almost the same as the Lasso, except for the Renewal lapse where the adaptive Lasso method performs poorly. Therefore, I select the Lasso method as best CPH model. An additional reason to prefer the Lasso method over the adaptive Lasso is computational time. The adaptive Lasso requires two estimation steps. One ridge regression to estimate the weights in the second Lasso regression, while the Lasso method only needs the Lasso regression. The computation time for the adaptive Lasso is therefore approximately twice as long compared to the Lasso method.

Since the Lasso method outperforms the other methods, I compare the frailty model only with the Lasso method. The variables included in the frailty model are the same as in the Lasso method. I add the random effect to the renewal number of the contract for the MTC, Renewal and AllSecur lapse, since loyal clients tend to behave the same with respect to lapse. For the Afterthought lapse, the renewal number is always zero and thus the random effect must be added to another variable. The variable with the random effect for the Afterthought lapse is the age of the driver. Table 3 shows the concordant index for the frailty model along with the Lasso method. In this table, we observe that the Lasso method outperforms the frailty model for all different type of lapses, except the renewal lapse.

Table 3: **Concordant Index for the Frailty Model**
This table shows the concordant index for the CPH model with variables selected according to the Lasso method and for the frailty model. The concordance index is given for each different type of lapse. The standard deviation is given in parentheses.

	Lasso	**Frailty**
MTC	0.672	0.668
	(0.002)	(0.002)
Renewal	0.694	0.701
	(0.004)	(0.004)
Afterthought	0.697	0.683
	(0.024)	(0.006)
AllSecur	0.824	0.812
	(0.005)	(0.005)

Table 4 compares the Brier score for the Lasso method and the frailty model. The Lasso method outperforms the frailty model for all the different type of lapses. The Lasso method is especially better at forecasting the afterthought lapse. Based on table 3, I conclude that the in-sample performance is not significantly different for the two methods and based on table 4, I conclude that the Lasso method outperforms the frailty model in terms of predictive power. Overall I conclude that the Lasso method is the best survival analysis method.

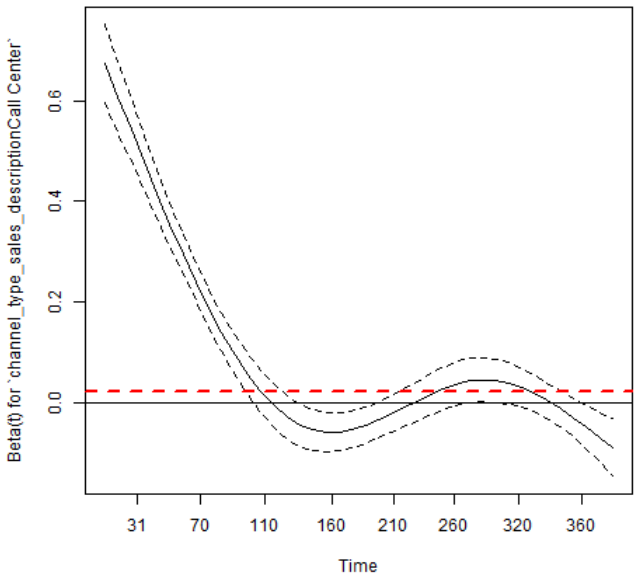
Table 4: **Brier Score for Frailty Model**
This table shows the Brier for the CPH model with variables selected according to the Lasso method and for the frailty model. The Brier score is given for each different type of lapse.

	Lasso	**Frailty**
MTC	0.1396	0.1503
Renewal	0.0544	0.0546
Afterthought	0.0459	0.0657
AllSecur	0.0242	0.0259

So far, no attention is given to possible violations of the PH assumption of the CPH model. To investigate the PH assumption, I fit a CPH model with variables chosen with the Lasso method on the overall data set. The null hypothesis for testing the PH assumption is that the PH assumption is valid. In table 5 we observe that the majority of variables included for the overall data set do not meet the PH assumption, namely 73.33%. In figures 1 and 2 we see two examples of variables that do not meet the PH assumption in the overall data set.

Figure 1: **Time-varying Coefficient of Call Center**

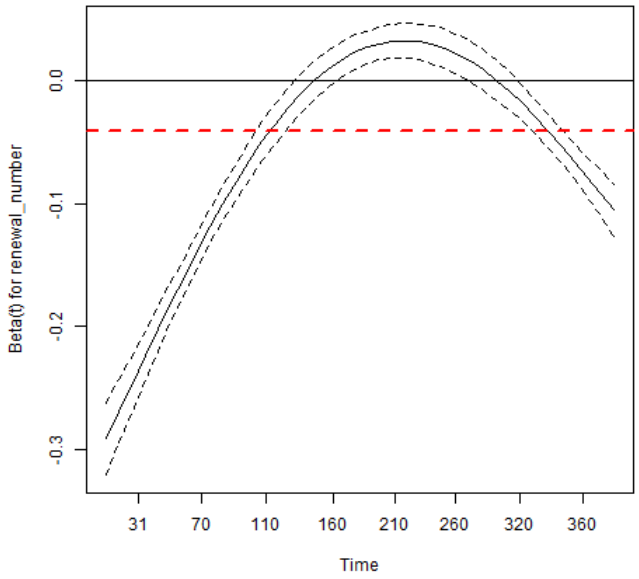
This figure shows the coefficient over time for the variable Call Center. The black line is the coefficient over time and the two black dotted lines show the confidence interval. The red dotted line represents the time-fixed effect as is included in the model.



In figure 1 we observe that there are three periods where the coefficent acts differently. In the first four months, it appears that the coefficient is declining. The following five months the coefficient increases and declines again in the remaining three months. The coefficient of the variable Renewal Number in figure 2 acts differently in two periods. In this first seven months the coefficient is increasing and in the remaining five months the coefficient decreases over time. For the remaining variables that do not meet the PH assumption, their behaviour is also different in two or three periods.

Figure 2: **Time-varying Coefficient of Renewal Number**

This figure shows the coefficient over time for the variable Renewal Number. The black line is the coefficient over time and the two black dotted lines show the confidence interval. The red dotted line represents the time-fixed effect as is included in the model.



This is an indication that a distinction between different type of lapses, such as the type of lapses as defined by AllSecur, makes statistical sense. When I check the PH assumption over different type of lapses, which can occur in distinct time periods, the PH assumption is violated for only a relatively few variables. The Afterthought lapse has relatively the most violations, namely 33.33%, but this is already a large reduction compared to 73.33%. Note that the AllSecur lapse is almost the same as the overall lapse since almost all contracts can be ended by AllSecur due to a defaulting client. Only contracts that have lapsed in the afterthought period, do not appear in the AllSecur data set.

Due to the large reduction in variables that do not meet the PH assumption, I choose to not add a time-varying coefficient as in equation (8). There will still be a bias in the estimated coefficients that do not meet the PH assumption, however introducing time varying coefficients is mostly not enough to completely satisfy the PH assumption due to the complex shapes the coefficients over time have (see figure 1). Furthermore, adding time varying coefficients also leads to a loss of interpretability.

Table 5: **Proportional Hazards Assumption**
This table shows the number of variables that violates the Proportional Hazards assumption for the different type of lapses.

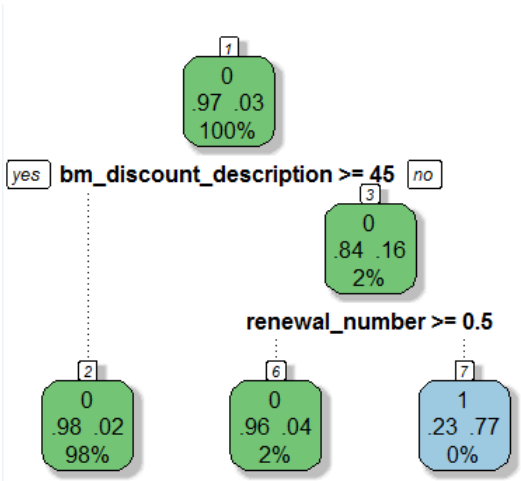
	# variables included	# violated	# non violated	proportion violated
Overall	15	11	4	73.33%
MTC	62	18	44	29.03%
Renewal	586	27	559	4.6%
Afterthought	9	3	6	33.33%
AllSecur	15	11	4	73.33%

4.2 Machine Learning

This section discusses the results for the machine learning techniques. There are three machine learning techniques: a CART decision tree, RF and SGB.

Figure 3 shows the pruned decision tree built via the CART algorithm for the AllSecur lapse. In this tree, we observe that if a contract holder has a Bonus Malus discount larger than 45%, there is a 2% chance that the contract holder will lapse. If the Bonus Malus discount is smaller than 45% and the renewal number is larger or equal to one, the probability of an AllSecur lapse is 2%. On the other hand, if the renewal number is equal to zero, the probability of an AllSecur lapse is 77%.

Figure 3: **CART Decision Tree for the AllSecur Lapse**
This figure shows the tree build by the CART algorithm. The upper number in each node denotes which class is the most dominant class in that node. The two middle number denote the proportion non-lapser and lapser respectively. The lower percentage denotes the fraction of total observations in that node.



This example shows the basic workings of a decision tree. I only show the tree for the AllSecur lapse, since this tree is the smallest and most interpretable compared to the trees which model the other type of lapses. The reason that tree for the AllSecur lapse is this small is due to the definition of the AllSecur lapse, namely default of payments. If we take a look at the node 3, we conclude that the new clients (renewal number of 0) are more likely to default on the payments than people who renewed their contract (renewal number larger than 0). This makes sense, since existing customers have proven that they pay their premium, otherwise they would not be a client anymore. New customers on the other hand have not and are therefore more likely to default on their payments. The split at node 1 is less clear, but an explanation could be that young people have less claim free years and thus a lower Bonus Malus discount. Young people in general have less money than older people and therefore are more likely to default on their payments. Another explanation could be that a smaller Bonus Malus discount makes the premium higher and thus harder to pay, resulting in more AllSecur lapses. The first mentioned possible explanation is probably not the right one, otherwise age would be the variable included in the tree. However, it is also possible that the split is made due to a combination of the two explanations. This shows that the interpretation of the tree is not always straightforward, however the data tells us that these two variables are the most important variables to describe and explain the AllSecur lapse.

Because the RF and SGB construct many different trees, there is not a nice visualization compared to a single decision tree. However, there are other methods to see which variables are used in the models and which of those variables are deemed as most important for determining the target value. One of these methods is the use of relative influence plots ((Friedman, 2001)). In the last part of this section I show that the SGB algorithm outperforms the RF, therefore I only show the relative influence plots of the SGB.

Figure 4 shows the relative influence plots of the MTC and Renewal lapse for the ten most important variables. This figure shows that for the MTC lapse that the vehicle age contributes the most to the SGB model. The first three variables, vehicle age, the precense of an accessory coverage and renewal number contribute about 50% to the SGB model. This plot unfortunately does not tell us why these variables contribute so much to the model. An explanation for vehicle age can be that old cars are mostly insured by the minimum required by law, while new cars probably have additional insurances such as an 'All Risk' insurance. Furthermore, the people who have the minimum car insurance are probably also the most price sensitive segment, while the people with an All Risk insurance are not. Therefore, people with old cars will probably check more often whether they can buy a cheaper insurance, resulting in more MTC lapses for AllSecur. The opposite, i.e. less lapses, holds for the people with the younger

cars.

For the Renewal lapse the first three variables contribute even more than 50%. The explanation for their high contribution is intuitive. When the contract is renewed, a new premium is offered to the client. If this premium is higher than the old premium (due to a claim or a change in tariff), the client will probably search for a new insurance company. On the other hand, if the new premium is lower than the old premium, the client has not much incentive to take action. This is most probably the reason that the first two variables, the difference in absolute and relative premium, contribute so much. It is a natural moment for clients to check whether their insurance can be bought more cheap at a competitor at the moment they receive their new offered premium. This causes the hgih contribution of the third variable, the price rank of AllSecur in the market. The decision of the client to go to another insurance company is most often based on the premium the client has to pay. If the ranking is good for AllSecur, the chances are low that the client will leave.

Figure 4: **Relative Influence Plots MTC and Renewal Lapse**
This figure shows relative importance plots for both the MTC and the Renewal lapse. Only the first ten variables are shown.

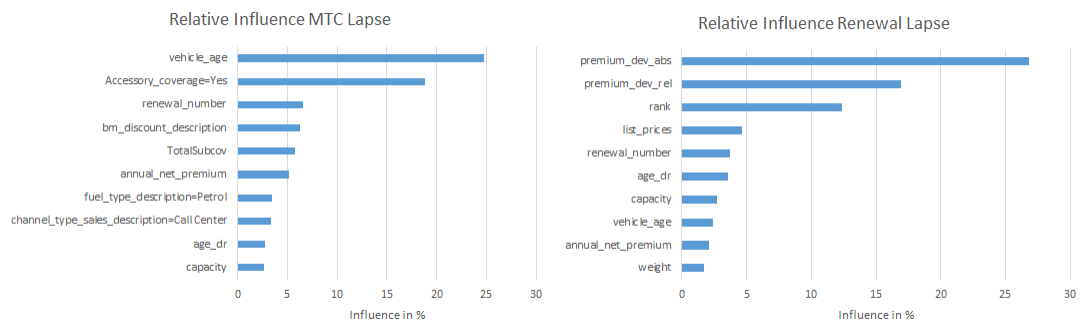
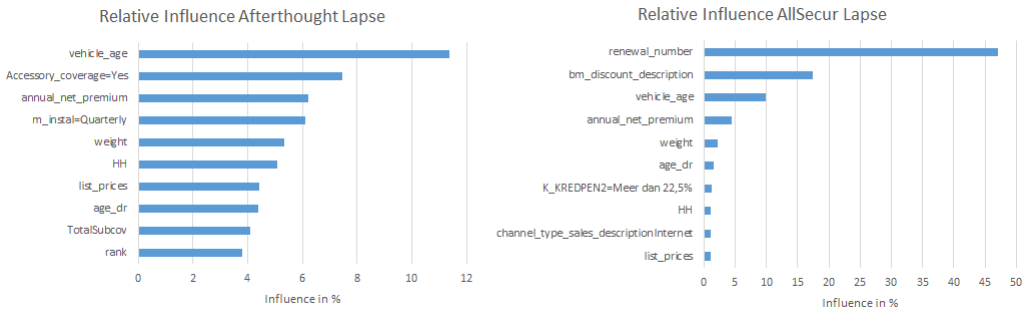


Figure 5 shows the relative influence plots for the Afterthought and AllSecur lapse. For the Afterthought lapse we observe that we need 8 variables to reach a total contribution of 50%. This is an indication that the Afterthought lapse is rather hard to explain, as there is no variable that is clearly explaining why people leave within the first two weeks. For the AllSecur lapse it is exactly the opposite. The variable that contributes the most, renewal number, contributes alone almost 50%. Together with the Bonus Malus discount, the contribution total 65%. It should be of no surprise that these two variables are the most important variables for the AllSecur lapse, as the basic pruned CART tree in figure 3 only included these two variables.

Figure 5: **Relative Influence Plots Afterthought and AllSecur Lapse**
This figure shows relative importance plots for both the Afterthought and the AllSecur lapse. Only the first ten variables are shown.



To assess the predictive performance of the machine learning techniques, I also use the Brier score. Table 6 shows the Brier score for the different techniques and for the different type of lapses. The CART decision tree is outperformed by the RF and SGB algorithms for each type of lapse. It is clear that the two more advanced techniques provide added value over the simple CART tree. The Brier score for the RF and SGB algorithm are quite close to each other, however the SGB algorithm slightly outperforms the RF algorithm for almost all the lapses. Since the SGB ties with the RF for the Afterthought lapse, but outperforms the RF for all other type of lapses. Overall, I conclude that the SGB method is the best performing machine learning technique.

Table 6: **Brier Score for the Decision Tree Trechniques**
This table shows the Brier for the CART, Random Forest (RF) and the Stochastic Gradient Boosting (SGB) technique.

	CART	**RF**	**SGB**
MTC	0.1490	0.1390	0.1383
Renewal	0.0560	0.0535	0.0532
Afterthought	0.0455	0.0470	0.0470
AllSecur	0.0233	0.0230	0.0225

4.3 Comparison of Methodologies

This section compares the results obtained from the best survival analysis and machine learning techniques with the benchmark. The best survival analysis is the CPH model with the parameters estimated by the Lasso method. For the

machine learning, the best method is the SGB algorithm. The benchmark model is the GLM and is provided by AllSecur. The first two subsections discuss the MTC and Renewal lapse respectively. The last subsection gives a short overview of the most important results for the Afterthought and AllSecur lapse. The full analysis for the Afterthought and AllSecur lapse is given in Appendix B, since they only account for 15% of the total lapses.

4.3.1 MTC Lapse

The best survival analysis method found in section 4.1 is the CPH model with variables chosen by the Lasso method, in the remainder of this section I refer to this model as the Survival Analysis model. In this section I compare this model to the GLM model provided by AllSecur and the SGB algorithm. The following three graphs compare the models for the MTC lapse. The horizontal axis shows the relative difference between the two prediction methods. The graph only shows the differences if the exposure is larger than 100. The left vertical axis shows the lapse rate and the right vertical axis the exposure.

Figure 6 shows that the MTC lapse forecasts of the Survival Analysis model and the GLM has highest exposure for the forecasts that do not differ. For the forecasts that do differ, most of the differences are less than 40%. Most predictions differ less than 5 percentage points in absolute terms. For the negative differences, the GLM outperforms the Survival Analysis model as the predictions of the GLM lie closer to the actual than the Survival Analysis predictions. For the positive differences, the actual lies between the GLM and the Survival Analysis model. Based on these two observations, I conclude that the GLM model slightly outperforms the Survival Analysis model.

Figure 6: **Comparison of Survival Analysis and GLM for the MTC lapse**
This figure shows the performance of the Survival Analysis model and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.

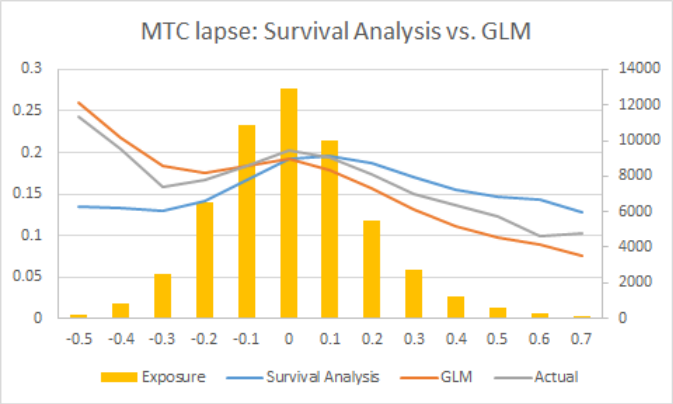
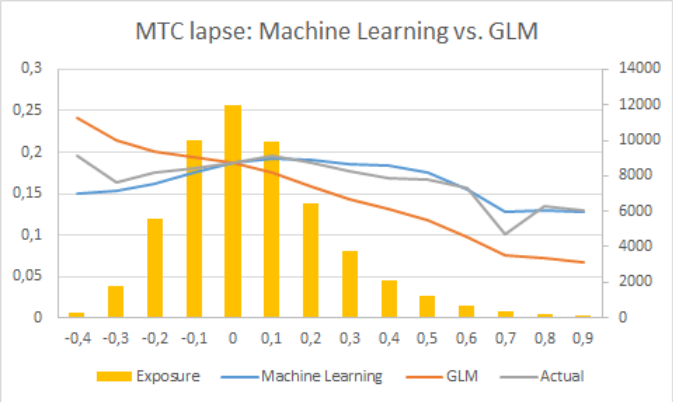


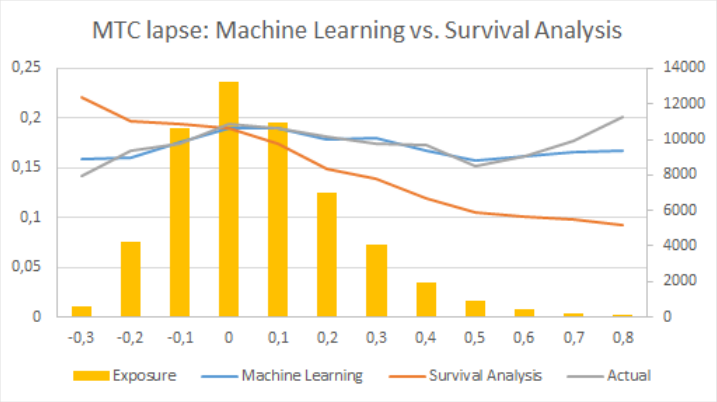
Figure 7 shows the performance of the SGB algorithm and the GLM. In this graph we observe that most forecasts of the two methods do not differ a lot. Most of the forecasts differ less than 30%. For the forecasts that differ, we observe that the SGB method overall performs better than the GLM. Especially for the positive differences, the SGB predictions are more in line with the actual compared to the GLM predictions. This also the case for the negative differences, however the SGB predictions exhibit some degree of underestimation. Based on this, I conclude that the SGB method outperforms the GLM model for the MTC lapse.

Figure 7: **Comparison of Machine Learning and GLM for the MTC lapse**
This figure shows the performance of the SGB algorithm and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



The last figure of this subsection, figure 8, shows the comparison between the MTC lapse predictions of the SGB algorithm and the Survival Analysis model. The SGB algorithm makes almost perfect forecasts and clearly outperforms the Survival Analysis model. The Survival Analysis model only makes accurate predictions when the relative difference is equal to zero. Based on the conclusions drawn from figures 6 and 7, it is expected that the SGB algorithm outperforms the Survival Analysis model, since the SGB outperforms the GLM, while the GLM (slightly) outperforms the Survival Analysis model.

Figure 8: **Comparison of Machine Learning and Survival Analysis for the MTC lapse**
This figure shows the performance of the SGB algorithm and the Survival Analysis model. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



The overall conclusion based on the three graphs above is that the SGB algorithm performs the best, then the GLM and lastly the Survival Analysis model. This conclusion is for the larger part supported by the Brier Score for the three different methods, see table 7. The Brier Score is the lowest for the SGB algorithm, but is unable to differentiate between the GLM and the Survival Analysis model. However, in figure 6 we saw that the GLM only slightly outperformed the Survival Analysis and therefore it is expected that the Brier Scores would be close to each other.

Table 7: **Brier Scores for the MTC lapse**
This table shows the Brier for the three methods for the MTC lapse.

	MTC lapse
Survival Analysis	0.1396
SGB	0.1383
GLM	0.1396

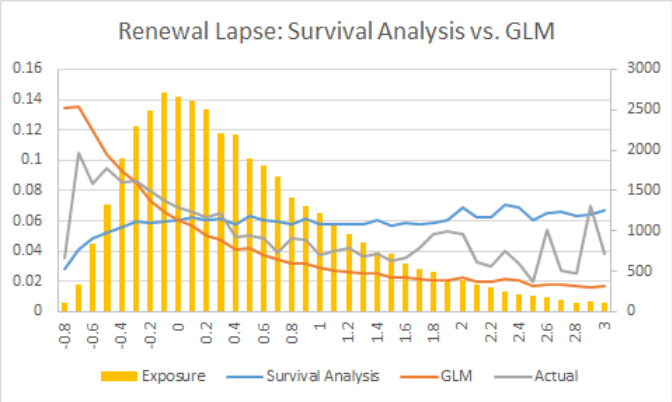
4.3.2 Renewal Lapse

This section repeats the analysis of the previous section, but now for the Renewal lapse.

The relative differences for the Survival Analysis model and GLM are shown in figure 9. The differences are in general larger than for the MTC lapse. There

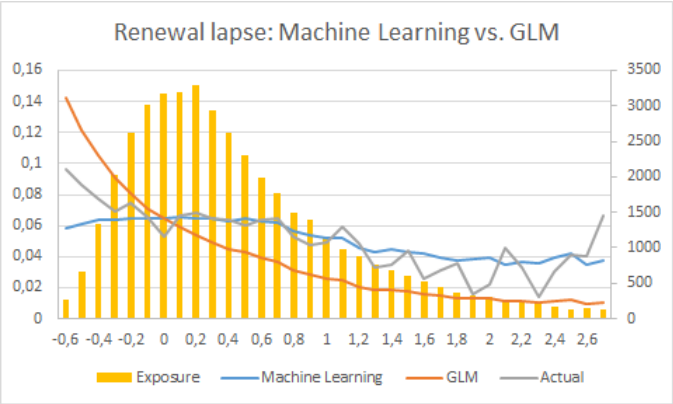
are Survival Analysis model predictions, with a decent exposure, which are three times larger than the GLM predictions. In absolute terms the differences are not that large, since the Renewal lapse rate is in the 4% to 8% range while the MTC lapse is the 15% to 20% range. The Survival Analysis model predicts for most individuals a lapse rate of approximately 6%, while the GLM model predictions show more variability. The actual lapse rates also show more variability and the shape of the actual lapse rate and the GLM predictions are approximately the same. The GLM model seems to underestimate the lapse rate by roughly 1 percentage point for the positive differences and the Survival Analysis model seems to overestimate the lapse rate by roughly 2 percentage points for the same contracts. Due to the smaller underestimation of the GLM and the presence of variability in the predictions, I conclude that the GLM model also outperforms the Survival Analysis model for the Renewal lapse.

Figure 9: **Comparison of Survival Analysis and GLM for the Renewal lapse**
This figure shows the performance of the Survival Analysis model and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



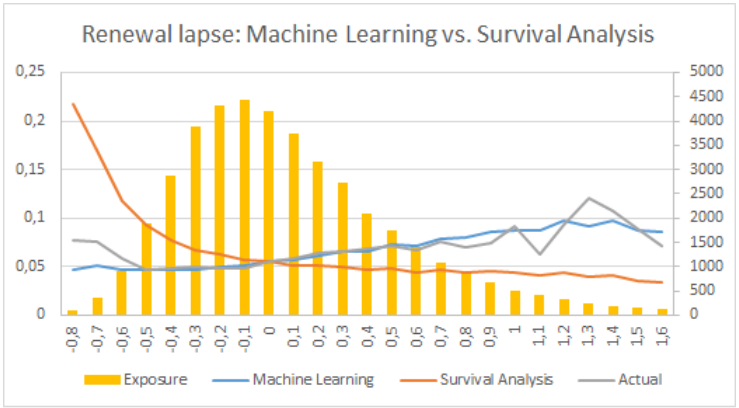
The comparison between the SGB algorithm and the GLM is given in figure 10. In this figure we observe that the SGB algorithm produces better forecasts than the GLM model. When we observe negative differences, the seriousness of the overestimation of the GLM is approximately of the same size as the underestimation of the SGB. For the positive differences, the SGB predictions are clearly better than the GLM predicitons. Only when we observe a relatively low exposure to the relative differences, the GLM sometimes seem to outperform the SGB. However, overall I conclude that the SGB algorithm outperforms the GLM.

Figure 10: **Comparison of Machine Learning and GLM for the Renewal lapse**
This figure shows the performance of the SGB algorithm and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



The results for the Renewal lapse in figure 11 are the same as for the MTC lapse. The SGB algorithm almost perfectly predicts the Renewal lapse, while this is very hard for the Survival Analysis model. This is also expected based on the two previous figures where the GLM outperforms the Survival Analysis model, while the SGB algorithm outperforms the GLM.

Figure 11: **Comparison of Machine Learning and Survival Analysis for the Renewal lapse**
This figure shows the performance of the SGB algorithm and the Survival Analysis model. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



Based on the three graphs above, I conclude that the SGB algorithm outperforms both the Survival Analysis model the GLM, similar to the MTC lapse. Contrary to the MTC lapse, the GLM now clearly outperforms the Survival Analysis model. The above conclusion is fully supported by the Brier scores of table 8. The Brier score is the lowest for the SGB, then the GLM and lastly the Survival Analysis model.

Table 8: **Brier Score for the MTC and Renewal Lapse**
This table shows the Brier for the three methods for the MTC and Renewal lapse.

	MTC	**Renewal**
Survival Analysis	0.1396	0.0544
SGB	0.1383	0.0532
GLM	0.1396	0.0539

4.3.3 Afterthought and AllSecur Lapse

This section gives a summary of the results for the Afterthought and AllSecur lapse. The full analysis is given in Appendix B.

For the Afterthought lapse the results are mixed. Based on the figures, the GLM clearly outperforms the Survival Analysis model and is outperformed by the SGB. However the Survival Analysis model seems to outperform the SGB. If we use the Brier score as metric, table 9, the GLM performs the best, then the Survival analys model and lastly the SGB. In section 4.2 I showed that the SGB found it difficult to explain the Afterthought lapse, which may be the cause of the highest Brier score.

The results for the AllSecur lapse are comparable to the results of the MTC and Renewal lapse. The SGB outperforms both the Survival Analysis model and the GLM. The GLM outperforms the Survival Analysis model. The cause of the bad performance of the Survival Analysis model is due to the many variables which do not meet the PH assumption (see section 4.1).

Table 9: **Brier Score for all the Lapses**
This table shows the Brier for the three methods for the MTC, Renewal, Afterthought and AllSecur lapse.

	MTC	**Renewal**	**Afterthought**	**AllSecur**
Survival Analysis	0.1396	0.0544	0.0459	0.0242
SGB	0.1383	0.0532	0.0470	0.0225
GLM	0.1396	0.0539	0.0455	0.0238

5 Conclusion

Lapse rates are one of the risk factors that determine the premium clients have to pay to their insurance company. AllSecur makes a distinction between four different type of lapses the reason to lapse changes over time. This is especially true for the lapses around the renewal date. The lapse rates are currently modelled by the GLM, one of the most commonly used models used in the insurance business. I focus on two different methodologies in order to make forecasts for the lapse rate. The two methodologies are survival analysis and machine learning.

I compare different variable selection procedures for the survival analysis models. I consider, forward regression and two penalized regression approaches, namely Lasso and adaptive Lasso. I find that the penalized regression approaches outperform the forward regression approach. Furthermore, the Lasso method is preferred over the adaptive Lasso due to a significantly lower computational time and near equal performance. For the machine learning I compare the predictive performance of a CART decision tree, a RF and the SGB algorithm. I find that the SGB algorithm outperforms the other two machine learning methods.

The different best methods of each methodology are then selected and I compare the performance between the methods to find the best prediction method. For all the different type of lapses, the Survival Analysis model is outperformed by the GLM. The GLM is outperformed by the GBM algorithm for the MTC, Renewal and AllSecur lapse. For the Afterthought lapse, the results are mixed. However, the Afterthought data set is the smallest of all the data sets and the Afterthought lapses account for only a small portion of the total lapses. Therefore, I conclude that the SGB algorithm outperforms the GLM. I also compare the SGB algorithm with the Survival Analysis model and find that the SGB algorithm also outperforms the Survival Analysis model.

The main reason of the strong performance of the SGB algorithm is that it in the procedure all the covariates are analyzed. In the two other methodologies, a subset of variables are selected and the selection procedures can contain

drawbacks. For the Survival Analysis model, the variables are chosen by the Lasso regression on a training set. Suppose that in the training set there are two highly correlated variables, x_i and x_j, but the correlation is less strong in the test set. Furthermore, assume that x_i is the covariate with real predictive power. The Lasso regression can choose to use x_j and set the coefficient for x_i equal to zero. The performance of the survival analysis model is then lessened due to the selection of x_j instead of x_i. This also holds for the GLM. The GLM is provided by AllSecur and the variables included in the model may not be best variables to include.

I have a few recommendations for further research. Some of the variables violate the PH assumption. To counter this problem, I repeat the analysis for different lapses over time and find that then only a relatively few variables violate the PH assumption. This number can be decreased even further, if one finds the optimal time periods to define the lapses. Also, time-varying coefficients can be included to achieve this, but then the function over time must be accurately determined.

Appendices

A Variables

Table 10: **Individual Variables Summary**

	Description	Type
channel_type_sales_description	How the contract is established	factor (3 levels)
renewal_number	Number of times client renewed	numeric
annual_net_premium	Annual premium the client has to pay	numeric
bm_discount_description	Discount percentage due to BM	numeric
annual_kilometers_description	Kilometers the client drives yearly	factor (3 levels)
bm_step	Step in the BM system	numeric
no_claim_free_years_description	Number of claim free years	numeric
coverage_description	Type of main cover	factor (3 levels)
age_dr	Age of the driver	numeric
premium_dev_rel	Relative difference old and new premium	numeric
premium_dev_abs	Absolute difference old and new premium	numeric
Roadside_Assistance	Extra subcover	dummy
Foreign_countries	Extra subcover	dummy
Legal_aid	Extra subcover	dummy
Passenger_accident	Extra subcover	dummy
Replacement_as_new	Extra subcover	dummy
Bonus_saver	Extra subcover	dummy
Accessory_coverage	Extra subcover	dummy
Free_choice_repairer	Extra subcover	dummy
Purchase_Arrangement	Extra subcover	dummy
TotalSubcov	Number of extra subcovers	numeric
m_instal	Frequency of payments	factor (3 levels)
m_premium_offered	New offered premium	factor (183 levels)

Table 11: **Car-specific Variables Summary**

	Description	Type
weight	Weight of the car	numeric
capacity	Capacity of the car	numeric
fuel_type_description	Type of fuel of the car	factor (4 levels)
list_prices	Price of the car	numeric
drive_description	Type of drive of the car	factor (5 levels)
gear_description	Type of gear of car	factor (3 levels)
turbo_description	Turbo in car	dummy
vehicle_age	Age of vehicle	numeric
m_Acceleration_gr	Acceleration of the car	factor (28 levels)
m_Bodywork	Bodywork of the car	factor (12 levels)
m_Top_speed_gr	Top speed of the car	factor (31 levels)
m_num_doors	Number of doors of the car	factor (6 levels)
m_automatic_transmission	Autmatic transmission	dummy
m_Make	Brand of car	factor (56 levels)
m_int_kw_gr	KiloWatt of the car	factor (25 levels)

Table 12: **Adress Variables Summary**

	Description	Type
URB	Degree of urbanisation	factor (8 levels)
HH	House Holds	numeric
INKOMEN	Average income	factor (7 levels)
MODUS_LFT	Mode of age	factor (6 levels)
K_KREDPEN2	Creditworthiness	factor (6 levels)
m_PROVINCIE	Province	factor (12 levels)

Table 13: **Competition Variables Summary**

	Description	Type
rank	Price ranking with competitiors	numeric
CI_3_b	Competitive Index based on 3 cheapest profiles	factor (92 levels)
CI_5_b	Competitive Index based on 5 cheapest profiles	factor (89 levels)
CI_10_b	Competitive Index based on 10 cheapest profiles	factor (89 levels)

B Comparison Afterthought and AllSecur lapse

B.1 Afterthought Lapse

This section compares the three methodologies for the Afterthought lapse.

Figure 12 compares the predictions of the Survival and GLM model for the Afterthought lapse. The GLM clearly outperforms the Survival Analysis model. The GLM has near perfect predictions for the lapse rate. The bulk of the Survival Analysis model predictions deviate less than 30%, which is in absolute terms, less than 1 percentage point. Compared to the Renewal and AllSecur lapse, the relative difference is small for the Afterthought lapse.

Figure 12: **Comparison of Survival Analysis and GLM for the Afterthought lapse**
This figure shows the performance of the Survival Analysis model and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.

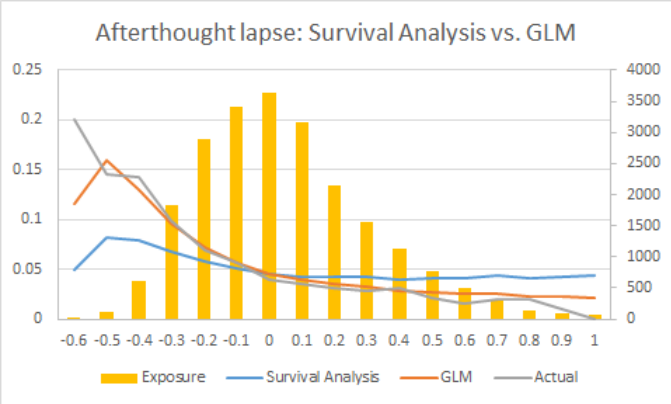
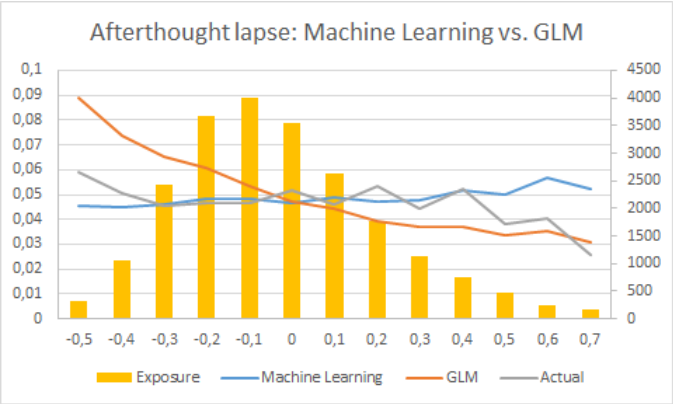


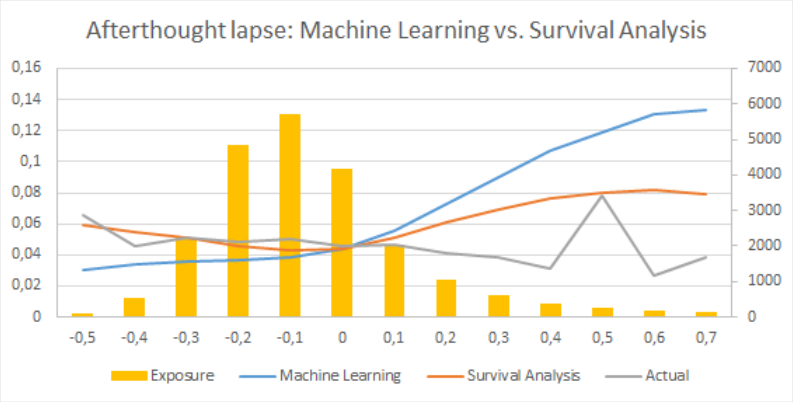
Figure 13 repeats the above analysis, but now for the SGB algorithm and the GLM. In this figure, we observe that the SGB algorithm provides better forecasts than the GLM. It seems that the GLM has the tendency to overpredict the Afterthought lapse, since most the group with the largest exposure is on the left side of the zero difference in the figure. There is less exposure on the right side of the zero difference and when the differences are positive and large, the performance of the GLM is a little bit better. However, overall the SGB algorithm outperforms the GLM.

Figure 13: **Comparison of Machine Learning and GLM for the Afterthought lapse**
This figure shows the performance of the SGB algorithm and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



The last figure of this section, figure 14, compares the predictions of the SGB algorithm and the Survival Analysis model. In this graph we observe that both method overpredict the lapse rate when the differences are postive and that the overprediction is more serious for the SGB algorithm. However, note that the exposure is rather small for these overpredictions and that the absolute differences are also rather small. When the relative differences in the predictions are negative, it seems that the Survival Analysis model is a little bit better at predicting the Afterthought lapse. Overall, I conclude that the Survival Analysis model is better at predicting the Afterthought lapse.

Figure 14: **Comparison of Machine Learning and Survival Analysis for the Afterthought lapse**
This figure shows the performance of the SGB algorithm and the Survival Analysis model. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



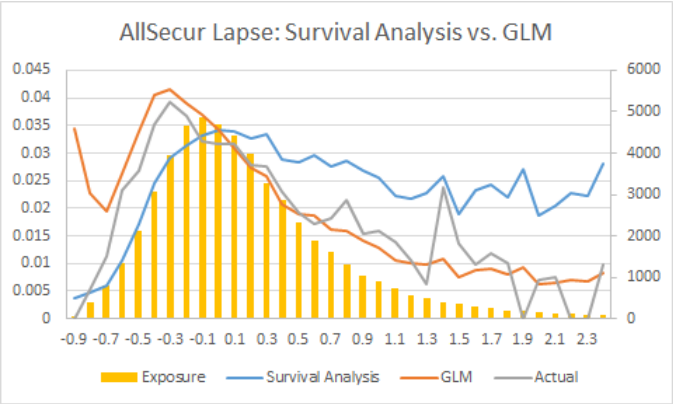
B.2 AllSecur Lapse

This section compares the three methodologies for the AllSecur lapse.

The GLM does a better job than the Survival Analysis model for the AllSecur lapse. In figure 15 we observe that the GLM has an almost perfect fit to the actual lapse rate. The Survival Analysis model underestimates the lapse rate when the differences are negative and seriously overestimates the lapse rate when the differences are positive. It is no surprise that the Survival Analysis model does not perform well for the AllSecur lapse. In table 5 of section 4.1 I show that 73.33% of the included variables does not meet the PH assumption. This introduces a bias in estimating the parameters and results in the poor predictive performance.

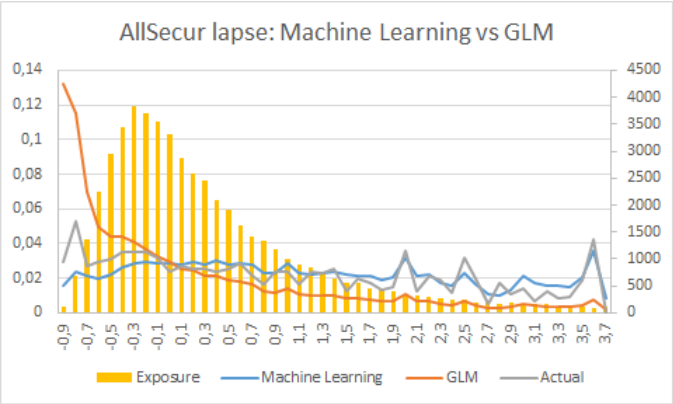
Figure 15: **Comparison of Survival Analysis and GLM for the AllSecur lapse**

This figure shows the performance of the Survival Analysis model and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



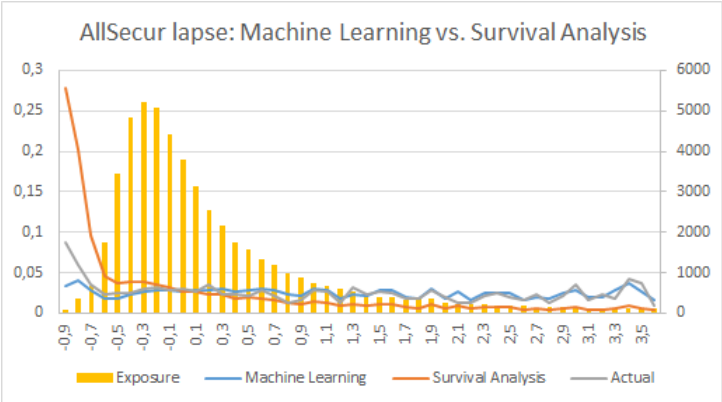
The predictions between the SGB algorithm and the GLM for the AllSecur lapse are more dispersed than the Afterthought lapse, see figure 16. There is still a reasonable exposure in cases where the differences are more than 200%. The largest exposures in differences are in the -0.3 to -0.1 range and in the groups in this range the GLM outperforms the SGB algorithm. However, for all the other predictions, the SGB algorithm outperforms the GLM. Overall, I conclude that the SGB algorithm outperforms the GLM.

Figure 16: **Comparison of Survival Analysis and GLM for the AllSecur lapse**
This figure shows the performance of the SGB algorithm and the GLM. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



In figure 17 we observe that the SGB algorithm almost perfectly predicts the AllSecur lapse. The Survival Analysis model on the other hand, does not. I show in section 4.1 that most of the variables in the Survival Analysis model do not the PH assumption. This is most probably the cause for the rather bad performance of the Survival Analysis model for the AllSecur lapse.

Figure 17: **Comparison of Machine Learning and Survival Analysis GLM for the AllSecur lapse**
This figure shows the performance of the SGB algorithm and the Survival Analysis model. The left axis shows the lapse rate and the right axis shows the exposure. The horizontal axis represents the relative differences between the two forecasting methods.



References

Aalen, O. (1992). *Modelling Heterogeneity in Survival Analysis by the Compound Poisson Distribution.* Annals of Applied Probability, Vol. 4, No.2, pp. 951-972.

Breiman, L. (2001). *Random Forests.* Machine Learning, 45, pp. 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees.* Wadsworth.

Breslow, N. (1974). *Covariance Analysis of Censored Survival Data.* Biometrics, Vol. 30, pp. 89-99.

Brier, G. (1982). *Verification of Forecasts Expressed in Terms of Probability.* Monthly Weather Review, Vol. 78, pp. 1-3.

Clayton, D. (1978). *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence.* Biometrika, Vol. 65, pp. 141-151.

Cox, D. (1972). *Regression Models and Life Tables.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2, pp. 187-220.

Cox, D. (1975). *Partial Likelihood.* Biometrika, Vol. 62, No. 2, pp. 269-276.

Cox, D. and E. Snell (1989). *Analysis of Binary Data, Second Edition.* Chapman & Hall/CRC.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim (2015). *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?* Journal of Machine Learning Research, Vol. 15, pp. 3133-3181.

Freund, Y. and R. Schapire (1997). *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of Computer and System Sciences, Vol. 55, pp. 119-139.

Friedman, J. (2001). *Greedy Function Approximation: A Gradient Boosting Machine.* Annals of Statistics, Vol. 29, pp. 1189-1232.

Friedman, J. (2002). *Stochastic Gradient Boosting.* Computational Statistics & Data Analysis, Vol. 38, No. 4, pp. 367-378.

Gepp, A. and K. Kumar (2015). *Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques.* Procedia Computer Science 54, pp. 396-404.

Hakulinen, T. (1982). *Cancer survival corrected for heterogeneity in patient withdrawal*. Biometrics, Vol. 38, pp. 933-942.

Harrell, F., R. Califf, D. Pryor, K. Lee, and R. Rosatie (1982). *Evaluating the yield of medical tests*. Journal of the American Medical Association, Vol. 247, pp. 2543-2546.

Hougaard, P. (1984). *Life table methods for heterogeneous populations: Distributions describing the heterogeneity*. Biometrika, Vol. 71, No. 1, pp. 75-83.

Hougaard, P. (1986). *Survival models for heterogeneus population derived from stable distributions*. Biometrika, Vol. 73, No. 2, pp. 387-396.

Huberty, C. (1989). *Problems with stepwise methods: Better alternatives*. Advances in Social Science Methodology, Vol. 1, pp. 41-70.

Ishwaran, H., U. Kogalur, E. Blackstone, and M. Lauer (2008). *Random Survival Forest*. The Annals of Applied Statistics, Vol. 2, No. 3, pp. 841-860.

Ivanoff, S., F. Picard, and V. Rivoirard (2016). *Adaptive Lasso and group-Lasso for functional Poisson regression*. Journal of Machine Learning Research, Vol. 17, pp. 1-46.

Kattan, M. (2003). *Comparison of Cox Regression With Other Methods for Determining Prediction Models and Nomograms*. The Journal of Urology, Vol. 16, No. 3, pp. S6-S10.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall.

McGilchrist, C. and C. Aisbett (1991). *Regression with Frailty in Survival Analysis*. Biometrics, Vol. 47, pp. 461-466.

McIntyre, S., D. Montgomery, V. Srinivasan, and B. Weitz (1983). *Evaluating the statistical significance of models developed by stepwise regression*. Journal of Marketing Research, Vol. 20, No. 1, pp. 1-11.

Quantin, C., T. Moreau, B. Asselain, J. Maccario, and J. Lellouch (1996). *A Regression Survival Model for Testing the Proportional Hazards Hypothesis*. Biometrics, Vol. 52, pp. 874-885.

Schwarz, G. (1978). *Estimating the Dimension of a Model*. Annals of Statistics, Vol. 55, No. 2, pp. 461-464.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society Series B, Vol 58, No. 1, pp 267-288.

Tibshirani, R. (1997). *The lasso method for variable selection in the cox model.* Statistics in Medicine, Vol. 16, pp. 385-395.

Vaupel, J., K. Manton, and E. Stellard (1979). *The impact of heterogeneity in individual frailty on the dynamics of mortality.* Demography, Vol. 16, pp. 439-454.

Wilkinson, L. (1979). *Tests of significance in stepwise regression.* Psychological Bulletin, Vol. 86, No. 1, pp. 168-174.

Zou, H. (2006). *The Adaptive Lasso and Its Oracle Properties.* Journal of the American Statistical Association, Vol. 101, No. 476, pp. 1418-1429.

Zou, H. and T. Hastie (2005). *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society Series B, Vol. 67, No. 2, pp. 301-320.