A longitudinal framework for lapse management in life insurance

Mathias Valla

A longitudinal Machine Learning framework for lapse management in life insurance

Mathias Valla*[1,2]

[1] *Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, Lyon, France.*
[2] *Faculty of Economics and Business, KU Leuven, Belgium.*

Abstract

Developing an informed lapse management strategy (LMS) is critical for life insurers to improve their profitability, and gain insight into the risk of their global portfolio. When designing a retention campaign, prior research in actuarial science (see Loisel et al. (2021); Valla et al. (2023)) has shown that targeting policyholders by maximizing their individual Customer Lifetime Value is more advantageous and informative for the insurer than targeting all those who are likely to lapse. However, most existing lapse analyses are not taking advantage of the fact that features and targets may vary over time. We propose to define a longitudinal LMS framework, that provides time-informed insights and leads to increased precision in targeting. The strengths and flaws of this new methodology are discussed in various settings. This paper contributes to the field of lapse analysis for life insurers and highlights the importance of using the complete past trajectory of policyholders, which is often available in insurers' information systems but has yet to be exploited.

Key words: Lapse management strategy, longitudinal, Machine learning, life insurance, Customer lifetime value

1 Introduction

In this article, we present a novel methodology developed to address the retention challenges faced by life insurers in a French insurance portfolio consisting of equity-linked whole-life insurance policies (see Hardy (2003) for an extensive review on such insurance products). Whole-life insurance provides coverage for the entire lifetime of the insured individual, rather than a specified term and when contracting such an insurance plan, policyholders can choose how the outstanding face amount of their policy is invested between "euro funds" and unit-linked funds. Understanding the fundamental differences between these investment vehicles is essential to comprehending the dynamics of the whole-life insurance market. For savings invested in euro funds, the coverage amount is determined by deducting the policy costs from the total premiums paid, the financial risk associated with these funds is borne by the insurance company itself. The underlying assets of euro funds primarily consist of government and corporate bonds, limiting the potential returns, thus the performance of these funds is directly influenced by factors such as the composition of the euro fund, fluctuations in government bond yields, and the insurance company's profit distribution policy. Additionally, early termination of the policy by the policyholder incurs exit penalties, as determined by the insurance company. In contrast, unit-linked insurance plans operate under a different framework. The coverage amount is determined by the number of units of accounts held by the policyholder, and the financial risk is assumed by the policyholders themselves. Unit-linked funds offer a wide range of underlying assets, among all types of financial instruments, enabling potentially unlimited performance

*Email: mathias.valla@univ-lyon1.fr , URL: https://mathias-valla.com

based on the market performance of these assets. The investment strategy is tailored to the specific investment objectives of the policyholder and while certain limitations exist in terms of asset selection, policyholders generally face no exit penalties for their underlying investments.

Lapse is a critical risk for whole-life insurance products (see Bacinello (2005) or MacKay et al. (2017)), thus, policyholders represent a critical asset for life insurers. Therefore, the ability to retain profitable ones is a significant determinant of the insurer's portfolio value (and more generally a firm's value, see Gupta et al. (2004)). If some historical explanations for lapse are liquidity needs (Outreville (1990)) and rise of interest rates, it also appears that individual characteristics are also insightful (see Eling and Kochanski (2013) for a complete review). Consequently, policyholder retention is a strategic imperative, and lapse prediction models are a crucial tool for data-driven policyholder lapse management strategy in any company operating in a contractual setting such as a life insurer. We define an LMS as in Valla et al. (2023):

Definition 1 (Lapse management strategy (LMS))
A lapse management strategy for a life insurer is modeled by offering an incentive $\boldsymbol{\eta} = (\eta^{(1)}, ..., \eta^{(N)})$ *to policyholders* $(1, ...N)$. *Their policies yield a profitability ratio of* $\boldsymbol{p} = (p^{(1)}, ..., p^{(N)})$. *The incentive is accepted with probability* $\boldsymbol{\gamma} = (\gamma^{(1)}, ..., \gamma^{(N)})$. *and contacting the targeted policyholder has a fixed cost* c. *A targeted subject who accepts the incentive will be considered as an "acceptant" who will never lapse, and her probability of being active at year* $t \in [0, T]$ *is denoted* $r_{acceptant}(t)$. *Conversely, a subject who refuses the incentive and prefers to lapse will be considered as a "lapser", and her probability of being active at year* t *is denoted* $r_{lapser}(t)$. *The parameters* $(\boldsymbol{p}, \boldsymbol{\eta}, \boldsymbol{\gamma}, c, T)$ *uniquely define a lapse management strategy, while* $r_{acceptant}(t)$ *and* $r_{lapser}(t)$ *need to be estimated from the portfolio.*

Our goal is not only to model the lapse behavior but also to select which policyholder to target with a given retention strategy to generate an optimized profit for the insurer. Such a lapse management strategy requires estimating what can be considered as the future profit generated by a given policyholder: the individual customer lifetime value or CLV (see Donkers et al. (2007)). The individual CLV over horizon T, for the i-th subject aims at capturing the expected profit or loss that will be generated in the next T years and is expressed as follows, in the general time-continuous case:

$$CLV^{(i)} = \int_{\tau=0}^{T} \frac{p^{(i)}(\tau) \cdot F^{(i)}(\tau) \cdot r^{(i)}(\tau)}{e^{d(\tau)\cdot\tau}} \, d\tau, \tag{1}$$

with the profitability ratio $p_t^{(i)}$ being represented as a proportion of the face amount, $F_t^{(i)}$, observed at time t. The conditional individual retention probability, $r_t^{(i)}$, is the i-th observation's probability of still being active at time t. In practice, the individual CLV is often discretized and computed as a sum of annual flows, thus with t, the time in years,

$$CLV^{(i)}\left(\boldsymbol{p}^{(i)}, \boldsymbol{F}^{(i)}, \boldsymbol{r}^{(i)}, \boldsymbol{d}, T\right) = \sum_{\tau=0}^{T} \frac{p_\tau^{(i)} \cdot F_\tau^{(i)} \cdot r_\tau^{(i)}}{(1+d_\tau)^\tau}. \tag{2}$$

Equation 2 is primarily used in the marketing and actuarial literature (see Berger and Nasr (1998) or Loisel et al. (2021)). If we only consider the future T years of CLV, after time t, the sum becomes

$${}^{F}CLV^{(i)}\left(t, \boldsymbol{p}^{(i)}, \boldsymbol{F}^{(i)}, \boldsymbol{r}^{(i)}, \boldsymbol{d}, T\right) = \sum_{\tau=t+1}^{T+t} \frac{p_\tau^{(i)} \cdot F_\tau^{(i)} \cdot r_\tau^{(i)}}{(1+d_\tau)^\tau}. \tag{3}$$

Given an LMS, a policyholder can either be likely to accept the offer of an incentive and behave with an "*acceptant*" risk profile or she can be likely to reject the offer and thus behave with a "*lapser*" risk profile. In this context, *acceptants* and *lapsers* will not generate the same CLV as their respective retention probabilities differ. The CLV of an *acceptant* or a *lapser* are estimated using respectively $\boldsymbol{r}^{(i)}_{acceptant}$ and $\boldsymbol{r}^{(i)}_{lapser}$ as retention probabilities. All the expected future financial flows are discounted, with d_t representing the annual discount rate at year t. In definitive, ${}^{F}CLV^{(i)}(t, ...)$ represents the future T years of profit following observation at time t.

The analysis of a lapse management strategy, as described in Loisel et al. (2021), then in Valla et al. (2023), is a two-step framework. The first step consists of using the insurer's data to train survival models and predict yearly retention probabilities for any subject in the portfolio: we will refer to it as the *survival step*. The retention probabilities are used to compute an individual CLV-based estimation of the profit generated from targetting any policyholder. This estimation is eventually used as a response variable to fit a model predicting which kind of subject is likely to generate profit for the insurer: we will refer to it as the *regression step*. As in Ascarza et al. (2018) or Guelman et al. (2012), the goal of such a CLV-based methodology is not only to model the lapse behavior but rather to select which policyholder is worth targetting with a given retention strategy in order to generate an optimized profit for the insurer. This existing framework relies on the analysis of the time-to-death and time-to-lapse that can be updated regularly with new information from the policies. It can be summarized as follows:

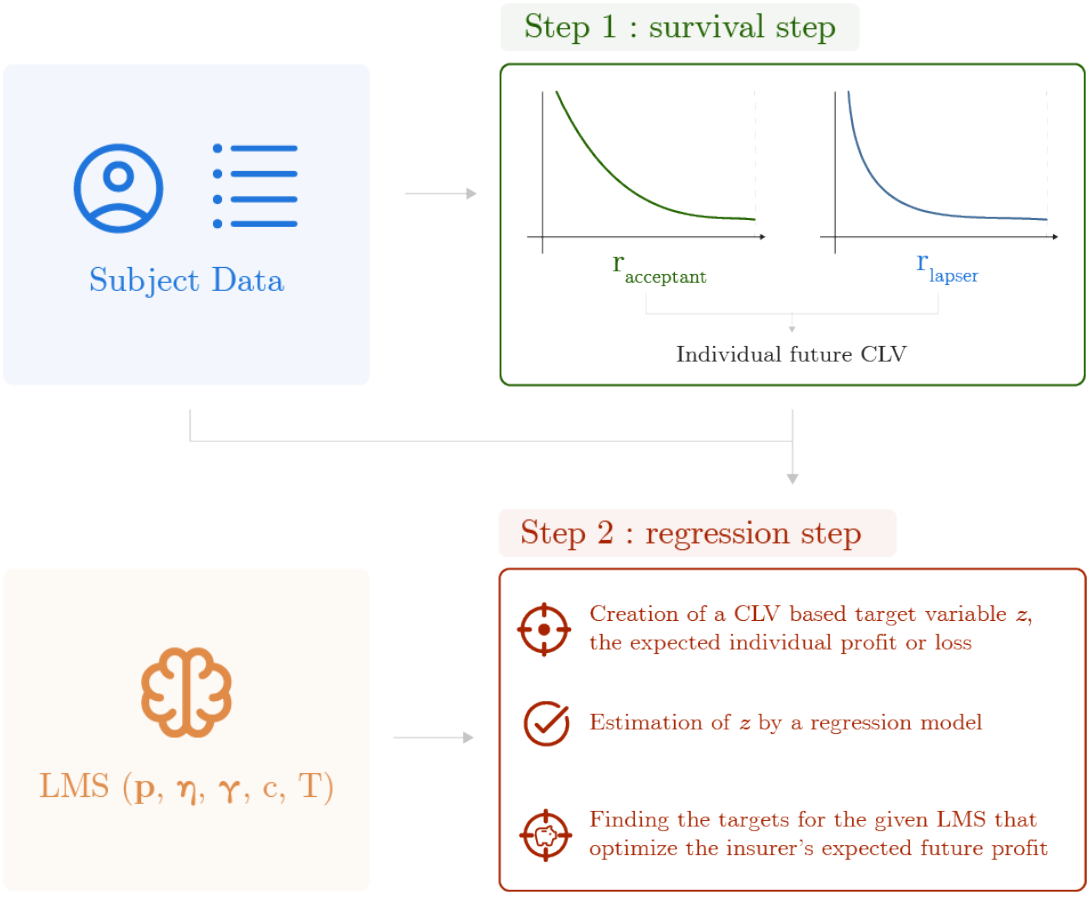


Figure 1: General framework for lapse management strategy

At least three limitations of that framework can be addressed. Firstly, it does not consider that an *acceptant* can lapse in the future, which is at best a very optimistic assumption, and at worst a great oversimplification. Secondly, it does not give any information on whether the timing of the retention campaign is optimal or not. Thirdly, it does not allow tightening the criteria on which the targeting of each policyholder is decided, depending on the risk the insurer is willing to take on the uncertainty of the predictions. This work addresses these limitations.

Throughout the lifetime of such insurance policies, a series of significant time-dependent events shape the interactions between policyholders and insurers. Firstly, premium payments play a pivotal role in sustaining the policy: these payments are highly flexible, allowing policyholders to choose their amount and frequency, thus they can be adjusted according to the policyholder's financial circumstances and preferences. Additionally, policyholders may decide to reduce their coverage by withdrawing a portion of their policy. We refer to these events as partial lapses: they involve a voluntary decrease in the face amount of the policy, enabling policyholders to adjust their coverage to better align with their changing needs. Such flexibility caters to policyholders' evolving financial situations and offers them greater control over their insurance plans. Over the policy's lifetime, other financial operations can occur, such as the payment of interest or profit sharing to the policyholder, and the payment of fees to the insurer. Insurance companies' information systems are usually designed to keep track of those operations at the policy-level, thus actuaries and life insurers often have access to the complete history of their policyholders as the information system is updated in real-time.

In certain instances, a policyholder may choose to lapse their insurance policy entirely. Complete policy lapse typically occurs when the policyholder decides to terminate her policy and receives a surrender value, which represents the accumulated value of the premiums paid, adjusted for fees, expenses, and potential surrender charges. Moreover, the occurrence of a policyholder's death also terminates the policy and triggers the payment of the policy's value, often referred to as the death benefit or claim, to the designated beneficiaries.

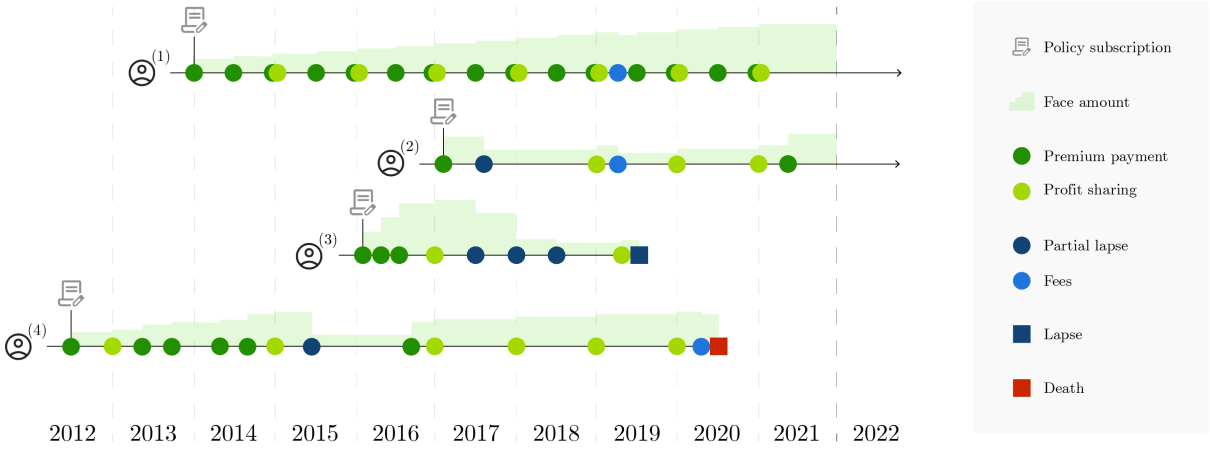


Figure 2: Example of policyholders timelines

In the context of our research, a policy can only terminate with a complete lapse or the death of the policyholder, which will be considered as competing risks in the following developments. If none of these events has happened to a policy, it is still active. The cumulated sum of all the financial flows occurring during one's policy timeline, including premiums, claims, fees,

interests, profit-sharing, and lapses, is commonly known as the face amount of the policy. This face amount represents the total value of the policy over its duration and serves as a measure of the policy's coverage and financial benefits. By comprehensively understanding and analyzing these events and their impact on the face amount of a life insurance policy, insurers can effectively develop lapse management strategies that align with policyholders' preferences and financial goals. Through our research, we aim to shed light on these dynamics and provide insights to optimize the design of such strategies, ultimately enhancing customer retention and overall portfolio performance in the life insurance industry.

In practice, actuaries often have access to the complete trajectories of every policy and it seems that not using them in models is ignoring a significant part of the available information. A data structure where time-varying covariates are measured at different time points is called longitudinal and individual policyholders' timelines can be illustrated as in Figure 2. The dynamical aspects of covariates have an impact on the performance of lapse prediction models and Risselada et al. (2010) concludes in favor of the development of dynamic churn models. They showed how the predictive performance of different types of churn prediction models in the insurance market decays quickly over time: this conclusion arguably applies to life insurers and in the case of lapse management strategy, we argue that using the complete longitudinal trajectories of every individual is also justified. Firstly, a change in financial behaviour - recent and frequent withdrawals for instance - can be informative lapse predictors. As an illustration of this point, we can imagine making predictions for two individuals with the exact same characteristics at the time of study but completely different past longitudinal trajectories: one is consistently paying premiums for instance, whereas the other stopped all payments for months and has been withdrawing part of her face amount lately. A prediction model ignoring longitudinal information would produce the exact same lapse prediction for both individuals. Conversely, an appropriate model, trained on longitudinal data is likely to seize the differences between the individuals over time and provide different predictions for the future. Secondly, a longitudinal lapse management framework allows for dynamic predictions with new information. It proves to be insightful in terms of decision-making for the insurer, as it shows how a change in the policy induces a change in the lapse behavior. Eventually, existing lapse management strategy approaches can only provide the insurer with information on whether targetting a given individual now is expected to yield profit, not on whether the timing of targeting is optimal. A longitudinal framework can help answer that last question.

In this paper, we want to account for the time-varying aspect of this problem in both steps of that framework. Firstly, we want to take advantage of the information contained in the historical data from the portfolio and obtain more accurate predictions for $\boldsymbol{r}^{(i)}$ and thus ${}^{F}CLV^{(i)}$: that is a gain of precision on the survival step. Secondly, we want to evaluate the expected individual retention gains over time to derive the optimal timing to offer the incentive: that is a gain of flexibility and expected profit on the regression step. For that purpose, we introduce tree-based models which are, to the best of our knowledge, yet to be explored in the actuarial literature. Those models, such as left-truncated and right-censored (LTRC) survival trees and LTRC forests by Fu and Simonoff (2016b) and Yao et al. (2020), or mixed-effect tree-based regression models (Sela and Simonoff (2012), Hajjem et al. (2014), Fu and Simonoff (2015), Capitaine et al. (2021)) are considered state-of-the-art and have yet to be exploited in the actuarial literature. We propose an application of that framework with data-driven tree-based models but other types of models exist and could fit in this framework (see Appendix A.1)

This extension is not trivial, as time-dependent features and time-dependent response

variables are difficult to implement in parametric or tree-based models. Indeed, conventional statistical or machine learning models do not readily accommodate time-varying features. This is the case for most tree-based models as they assume that records are independently distributed. Of course, this is unrealistic as observations of any given individual are highly correlated. Moreover, time-varying features can generate bias if not dealt with carefully (see Fisher and Lin (1999) for instance). The use of longitudinal data is already a well-studied topic (see Molenberghs and Verbeke (2006)), with rare examples within the actuarial literature (see Frees et al. (2021) for instance) and, to the best of our knowledge, only a few actuarial uses of time-varying survival trees or mixed-effect tree-based models have been tried or suggested (Dal Pont (2020), Campo and Antonio (2022) or Moradian et al. (2022)) and no longitudinal lapse analysis framework based on CLV has been described.

In summary, this work presents a longitudinal lapse analysis framework with time-varying covariates and target variables. This framework accommodates for competing risks and relies on tree-based machine learning models. This work focuses on a lapse management strategy and retention targeting for life insurers and extends the existing lapse management framework proposed in Loisel et al. (2021) and Valla et al. (2023). It defers from the latter by taking advantage of time-varying features, introducing different tree-based models to the lapse management literature, including the possibility for an *acceptant* to lapse in the future, yielding insights regarding individual targeting times, and adding the possibility to adjust the level of risk which the insurer is willing to take in a retention campaign. The rest of this paper is structured as follows. We describe the specifics of longitudinal analysis and a new longitudinal and time-dynamic lapse management framework which is the main contribution of this work in Section 2. This section also includes a brief description of models that can fit in this framework. In Section 3, we show a concrete application of our framework on a real-world life insurance portfolio with a discussion of our methodology and results. Eventually, Section 4 concludes this paper.

2 Longitudinal framework

2.1 Preliminaries on time-varying covariates and longitudinal notations

We aim to enrich the existing lapse management frameworks (see Definition 1) with time-varying covariates. To do so, we decide to adapt LMS methods to longitudinal analysis. In order to be perfectly clear on what we mean by *time-varying covariates* or *longitudinal data*, let us introduce some notations. This section borrows notations from the existing literature including Rizopoulos (2012) or Yao et al. (2022) for instance. Let us assume a very general setting where we want to build a dataset $\mathcal{D}$, encompassing the information of N individuals from which features are repeatedly measured over time. These covariates may come in many forms, some of them are time-varying, and others are time-invariant. We denote p_{tv}, p_{ti} the number of covariates in those respective categories, with $p = p_{tv} + p_{ti}$, the total number of covariates. At time t, the covariates matrix is $\boldsymbol{X}(t) = (x_1, x_2, \ldots, x_{p_{ti}}, x_{p_{ti}+1}(t), \ldots, x_p(t))$. In order to simplify the notations, we write $\boldsymbol{X}(t) = (x_1(t), x_2(t), \ldots, x_p(t))$ with $x_k(t) = x_k$, $\forall t$ and $\forall k \in [1, \ldots, p_{ti}]$.

These covariates are available for the N individuals, or subjects, which are observed at discrete time points. Subject i has been observed $n^{(i)}$ times, at $t_j^{(i)}, j = 0, 1, \ldots, n^{(i)} - 1$. In our life insurance context $t_0^{(i)}$ represents the first measurement of the covariates, i.e the subscription and times $t_j^{(i)}, j = 1, 2, \ldots, n^{(i)} - 1$ are the movement dates, i.e times at which a change in the policy has been recorded. If $t_0^{(i)} \neq 0$, this means that the baseline information at subscription is missing and the observation is left-truncated. A given subject i, at time $t_j^{(i)}$ has a vector of covariates denoted $\boldsymbol{x}_j^{(i)} = \left(x_{j,1}^{(i)}, \ldots, x_{j,p}^{(i)}\right)$ and generally, has a matrix of covariates denoted

$$\boldsymbol{X^{(i)}} = \left(\boldsymbol{x}_0^{(i)}, \ldots, \boldsymbol{x}_{n^{(i)}-1}^{(i)}\right)^\top. \tag{4}$$

As stated in Definition 1, the probability of still having an active policy at time t depends on the policyholder's risk profile. *Acceptants* are only at risk for death whereas *lapsers* are at risk for both lapse and death and we consider as the event of interest respectively death and the end of the policy (whatever the cause). Regardless of our outcome of interest, we study the time to an event ending the policy, thus we use the classical survival notations: subject i will eventually experience the event at time $T_*^{(i)}$ and she is no longer observed after censoring time $C^{(i)}$. We let $T^{(i)}$ denote the observed event time for subject i, defined as $t_{n^{(i)}}^{(i)} = T^{(i)} = \min\left(T_*^{(i)}, C^{(i)}\right)$.

The notations regarding the time dynamics of our data are now clear, we decide to structure this information in a longitudinal dataset. In order to do so, we assume that the time-varying covariates are constant between the observed time points, that is,

$$\boldsymbol{x}^{(i)}(t) = \boldsymbol{x}_j^{(i)}, \quad t \in \left[t_j^{(i)}, t_{j+1}^{(i)}\right), \quad j = 0, 1, \ldots, n^{(i)} - 1.$$

This assumption is perfectly consistent in an actuarial context where time-varying covariates such as financial flows are immediately updated. Any covariate update leads to a new observation and all variables are in fact constant between two consecutive observations. The only limit of this assumption is that updating the insurer's database usually takes some time and it proves to be unrealistic if a policy change has been reported but not yet processed in the information system.

An insurance policy at any time point is either active or ended. Moreover, it can only end in two ways: the policyholder either lapses her policy or dies. Thus we define three

event indicators. $\Delta^{(i)}$ is the event indicator, defined at the subject level, it denotes whether individual (i) has experienced an event (and which one) before censoring time,

$$\Delta^{(i)} = \begin{cases} 0 & \text{if } T_*^{(i)} \leq C^{(i)} \\ 1 & \text{if } T_*^{(i)} > C^{(i)} \text{ and EVENT = lapse} \\ 2 & \text{if } T_*^{(i)} > C^{(i)} \text{ and EVENT = death.} \end{cases} \tag{5}$$

We also introduce $\delta^{(i)}(t)$, the event indicator defined at the observation level, it denotes whether individual (i) has experienced an event (and which one) by time t:

$$\delta^{(i)}(t) = \Delta^{(i)} \cdot \mathbb{I}\left\{t \geq T^{(i)}\right\}. \tag{6}$$

At time $t = T_*^{(i)}$, the true event has occurred and we define the ultimate event indicator as

$$\Delta_*^{(i)} = \begin{cases} 1 & \text{if EVENT = lapse at time } T_*^{(i)} \\ 2 & \text{if EVENT = death at time } T_*^{(i)}. \end{cases} \tag{7}$$

It is constant over the observations for a given subject and represents the final value of $\Delta^{(i)}$ when the subject's policy eventually ends. It can be either equal to 1 or 2. For a subject with an active policy at the censoring time, the value of $\Delta_*^{(i)}$ is unknown.

Eventually, let $\mathcal{X}^{(i)}(t)$ denote the covariate individual information up to time t, and we define $\pi_*^{(i)}$ as the probability that the policy will eventually end with lapse, given all available information at observation time t. Mathematically speaking, we have

$$\pi_*^{(i)} = P(\Delta_*^{(i)} = 1 | \mathcal{X}^{(i)}(T^{(i)})). \tag{8}$$

We can now build $\mathcal{D}$, a longitudinal dataset encompassing the complete past information of all N subjects. For a given subject i, covariates are stored in rows, one row per observation window $[t_j^{(i)}, t_{j+1}^{(i)})$. Each row contains the unique $\left(t_j^{(i)}, t_{j+1}^{(i)}, \delta^{(i)}(t_j^{(i)}), \boldsymbol{x}_j^{(i)}\right)$ element and is completed by the subject unique identifier i and her event indicator $\Delta^{(i)}$: each row is called an *observation*. It is critical to include all those elements in the longitudinal dataset as all columns are inputs of longitudinal models used for the *survival step*.

Any observation only corresponds to one subject and conversely, any subject can be linked to a set of $n^{(i)}$ observations. We build $\mathcal{D}$ as the collection of all observations structured longitudinally :

$$\mathcal{D} = \left\{ \left(i, \left\{ t_j^{(i)}, t_{j+1}^{(i)}, \boldsymbol{x}_j^{(i)}, \delta^{(i)}(t_j^{(i)}) \right\}_{j=0}^{n^{(i)}-1}, \Delta^{(i)} \right) \right\}_{i=1}^{N},$$

or, if displayed in a table:

Table 1 A longitudinal dataset, in all generality

ID	Time window Start	Time window End	Covariate 1	...	Covariate p	Observation event indicator	Subject event indicator
1	$t_0^{(1)}$	$t_1^{(1)}$	$x_{0,1}^{(1)}$	...	$x_{0,p}^{(1)}$	$\delta^{(1)}(t_0^{(1)})$	Δ^1
1	$t_1^{(1)}$	$t_2^{(1)}$	$x_{1,1}^{(1)}$	...	$x_{1,p}^{(1)}$	$\delta^{(1)}(t_1^{(1)})$	Δ^1
1	$t_2^{(1)}$	$t_3^{(1)}$	$x_{2,1}^{(1)}$	...	$x_{2,p}^{(1)}$	$\delta^{(1)}(t_2^{(1)})$	Δ^1
1	$t_3^{(1)}$	$C^{(1)}$	$x_{3,1}^{(1)}$	...	$x_{3,p}^{(1)}$	$\delta^{(1)}(t_3^{(1)})$	Δ^1
2	$t_0^{(2)}$	$t_1^{(2)}$	$x_{0,1}^{(2)}$	...	$x_{0,p}^{(2)}$	$\delta^{(2)}(t_0^{(2)})$	Δ^2
3	$t_0^{(3)}$	$t_1^{(3)}$	$x_{0,1}^{(3)}$	...	$x_{0,p}^{(3)}$	$\delta^{(3)}(t_0^{(3)})$	Δ^3
3	$t_1^{(3)}$	$t_2^{(3)}$	$x_{1,1}^{(3)}$	...	$x_{1,p}^{(3)}$	$\delta^{(3)}(t_1^{(3)})$	Δ^3
3	$t_2^{(3)}$	$t_3^{(3)}$	$x_{2,1}^{(3)}$	...	$x_{2,p}^{(3)}$	$\delta^{(3)}(t_2^{(3)})$	Δ^3
...	...	...	...	...	...	...	...

Table 1 precisely illustrates what we call a longitudinal dataset, and a real-world example of such a dataset can be found in Section 3, Table 3. Adapting a lapse management strategy framework to a longitudinal setting means we take such a dataset as input and produce enriched predictions of the individual retention probabilities in the *survival step*, but also of individual profit or loss estimated in the *regression step*.

2.2 LMS longitudinal framework

We adopt Valla et al. (2023)'s framework and suggest some modifications and improvements to accommodate for longitudinally structured data. Instead of a top-down approach that consists of estimating the individual contributions to the insurer's profit from a global measure of the portfolio value, we suggest a bottom-up approach and directly evaluate the former and then derive the latter. Thus, we define the control future value of the policy, ${}^{F}CV^{(i)}(t,\dots)$, which represents the expected T-year individual profit or loss generated by subject i, after time t:

$$\begin{aligned} {}^{F}CV^{(i)}(t,\boldsymbol{p},\boldsymbol{\eta},\boldsymbol{\gamma},c,T) = \quad & {}^{F}CLV^{(i)}\left(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\right)\cdot(1-\pi_*^{(i)}) \\ & + {}^{F}CLV^{(i)}\left(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{lapser}},\boldsymbol{d},T\right)\cdot\pi_*^{(i)}. \end{aligned} \tag{9}$$

In other words, it simply represents an individual expected future CLV, if no lapse management is carried out. It highly depends on the probability for the policyholder to be a lapser.

Let us consider an LMS, let $\circledcirc^{(i)}(t)$ be the individual target vector indicator, designating if subject i is to be targetted at any time t. Our framework aims to find the optimal list of policyholders to target, $\mathcal{T}(t)=\{i \mid \circledcirc^{(i)}(t)=1\}$ that maximizes the expected profit for the insurer. In order to evaluate the profit or loss generated by an LMS, we must compare the expected profit obtained if no LMS was applied, with the expected profit generated by the lapse-managed portfolio. The former is given by Equation 9 and to obtain the latter, we define the lapse managed observation future value as

$$
\begin{aligned}
&{}^{F}LMV^{(i)}(t,\boldsymbol{p},\boldsymbol{\eta},\boldsymbol{\gamma},c,T) = \\
&\quad\Big[\,{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big)\cdot(1-\pi_*^{(i)})+{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{lapser}},\boldsymbol{d},T\Big)\cdot\pi_*^{(i)}\,\Big]\cdot(1-\circledcirc^{(i)}(t)) \\
&+\Big[\,{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}}-\boldsymbol{\eta^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big)\cdot(1-\pi_*^{(i)})+\gamma^{(i)}\cdot{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}}-\boldsymbol{\eta^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big)\cdot\pi_*^{(i)} \\
&\quad+(1-\gamma^{(i)})\cdot{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{lapser}},\boldsymbol{d},T\Big)\cdot\pi_*^{(i)}-c\,\Big]\cdot\circledcirc^{(i)}(t).
\end{aligned}
\tag{10}
$$

In simple terms, it is equal to the control future value of the policy (given by Equation 9) when subject i is not targetted, otherwise, it depends on whether she intended to lapse in the first place and if so, if she accepts the incentive η. If a policyholder that would not have lapsed (with probability $(1-\pi_*^{(i)})$) is targetted, she will rationally accept the incentive and generate the future CLV of an acceptant with profitability $p-\eta$. Conversely, for a policyholder that would have ultimately lapsed, she either accepts the incentive (with probability $\gamma^{(i)}$) and generates the future CLV of an acceptant with profitability $p-\eta$, or she refuses (with probability $(1-\gamma^{(i)})$) and generates profitability p with the risk profile of a lapser.

It follows that the individual expected retention gain obtained by applying an LMS is the difference between the expected individual CLVs with and without lapse management:

$$
RG^{(i)}(t,\boldsymbol{p},\boldsymbol{\eta},\boldsymbol{\gamma},c,T) = {}^{F}\mathrm{LMV}^{(i)}(t,\boldsymbol{p^{(i)}},\boldsymbol{\eta^{(i)}},\boldsymbol{\gamma^{(i)}},c,T) - {}^{F}CV^{(i)}(t,\boldsymbol{p^{(i)}},\boldsymbol{\eta^{(i)}},\boldsymbol{\gamma^{(i)}},c,T). \tag{11}
$$

that can be simplified as

$$
\begin{aligned}
RG^{(i)}(t,\boldsymbol{p},\boldsymbol{\eta},\boldsymbol{\gamma},c,T) = \circledcirc^{(i)}(t)\cdot\Big[&\pi_*^{(i)}\gamma^{(i)}\Big[{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}}-\boldsymbol{\eta^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big) \\
&\qquad-{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{lapser}},\boldsymbol{d},T\Big)\Big] \\
&-(1-\pi_*^{(i)})\cdot{}^{F}CLV^{(i)}\Big(t,\boldsymbol{\eta^{(i)}},\boldsymbol{F_i},\boldsymbol{r}_{\text{acceptant}},\boldsymbol{d},T\Big)\Big]-c\cdot\circledcirc^{(i)}(t).
\end{aligned}
\tag{12}
$$

An evaluation metric is finally derived to obtain the retention gain, at any observation time, if the policyholder i is targetted. We define $z^{(i)}(t)$ as

$$
\begin{aligned}
z^{(i)}(t) &= RG^{(i)}(t,\boldsymbol{p},\boldsymbol{\eta},\boldsymbol{\gamma},c,T|\circledcirc^{(i)}(t)=1) \\
&=\Big[\pi_*^{(i)}\gamma^{(i)}\Big[{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}}-\boldsymbol{\eta^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big) \\
&\qquad-{}^{F}CLV^{(i)}\Big(t,\boldsymbol{p^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{lapser}},\boldsymbol{d},T\Big)\Big] \\
&\quad-(1-\pi_*^{(i)})\cdot{}^{F}CLV^{(i)}\Big(t,\boldsymbol{\eta^{(i)}},\boldsymbol{F^{(i)}},\boldsymbol{r}^{(i)}_{\text{acceptant}},\boldsymbol{d},T\Big)\Big]-c.
\end{aligned}
\tag{13}
$$

In terms of intuition, it shows that if a policyholder that would have lapsed

(with probability $\pi_*^{(i)}$) is targetted and accepts the incentive (with probability $\gamma^{(i)}$), she generates the future CLV of an acceptant with profitability $p - \eta$ instead of her initial future CLV with profitability p and the risk profile of a lapser. The gain generated by targeting this policyholder is then the difference between the two. On the other hand, if the policyholder is wrongfully targetted and would not have lapsed (with probability $(1 - \pi_*^{(i)})$), she rationally accepts the incentive which is then lost for the insurer. In any case, the contact cost of c is spent.

From a practical point of view, we can see that the value of $z^{(i)}(t)$ depends on parameters that are observed in the porfolio ($\boldsymbol{F^{(i)}}$), or assumed by the insurer ($\boldsymbol{p^{(i)}}, \boldsymbol{\eta^{(i)}}, \boldsymbol{d}, T$), and that only $\boldsymbol{r}^{(i)}_{\text{acceptant}}$ and $\boldsymbol{r}^{(i)}_{\text{lapser}}$ need to be estimated. This estimation is the *survival step* mentioned in Section 1. We will show in Section 3.2.1 how to concretely estimate these retention probabilities using time-varying covariates.

Assuming that $\boldsymbol{z^{(i)}}$ has been estimated for every observation in the *survival step*, we can move forward to the *regression step* and use $\boldsymbol{z^{(i)}}$ as a target variable in a regression model handling time-varying covariates to predict whether targeting any policyholder will generate profit, given her previous observations if any. We will show in Section 3.3.1 how to concretely obtain $\hat{\boldsymbol{z}}^{(\boldsymbol{i})}$ with mixed-effect tree-based models.

With that in mind, we can update Definition 1 and define our LLMS as follows:

Definition 2 (Longitudinal lapse management strategy (LLMS))
A T-years lapse management strategy is modeled by offering an incentive $\eta^{(i)}$ to subject i if she is targeted. The incentive offered is expressed as a percentage of her face amount at the observation time and is accepted with probability $\gamma^{(i)}$. Contacting the targeted policyholder has a fixed cost of c. A targeted subject who accepts the incentive will be considered an "acceptant" who will be ***less likely to lapse****, and her probability of being active at year $t \in [0, T]$, given the information available until then, is denoted $r^{(i)}_{acceptant}(t \mid \mathcal{X}^{(i)}(t))$. Conversely, a subject who refuses the incentive and prefers to lapse will be considered a "lapser", and her probability of being active at year t, given the information available until then, is denoted $r^{(i)}_{lapser}(t \mid \mathcal{X}^{(i)}(t))$.*

Those probabilities are used to derive a dynamical profit-driven measure $z^{(i)}(t)$ based on CLV (see Equation 13). A regression model, allowing for longitudinal data is then used with $z^{(i)}(t)$ as a target variable, which allows us to estimate $\hat{z}^{(i)}(t)$ for any new observations (new observations of known subjects or observations of new subjects). Denoting the standard error of such a model σ_z and any confidence level α, we define the optimal longitudinal LMS at time t as

$$\circledcirc_*^{(i)}(t) = \mathbb{I}\left\{\hat{z}^{(i)}(t) > \alpha \cdot \sigma_z\right\}. \tag{14}$$

This is an indicator variable representing whether it is worth targeting policyholder i at time t, thus, the corresponding list of targetted policyholders is defined as

$$\mathcal{T}(t) = \left\{i \mid \circledcirc_*^{(i)}(t) = 1\right\}. \tag{15}$$

For any targeted policyholder and any confidence level α desired by the insurer, there is a unique future time $t_*^{(i)} \geq T^{(i)}$ when offering the incentive is optimal, which yields a maximal profit of $\hat{z}_*^{(i)}$. If all policyholders in $\mathcal{T}(t)$ are targetted at time t, the LLMS generates a profit of

$$RG(t, \boldsymbol{p}, \boldsymbol{\eta}, \boldsymbol{\gamma}, c, T, \alpha) = \sum_{i \in \mathcal{T}(t)} \hat{z}^{(i)}(t). \tag{16}$$

If all policyholders are targetted at the optimal time $t_*^{(i)} \geq t$*, the LLMS induces a gain for the life insurer of*

$$\overset{*}{RG}(t, \boldsymbol{p}, \boldsymbol{\eta}, \boldsymbol{\gamma}, c, T, \alpha) = \sum_{i \in \mathcal{T}(t)} \frac{\hat{z}_*^{(i)}}{\left(1 + d_{t_*^{(i)}}\right)^{\Delta t}} \text{ , with } \Delta t = t_*^{(i)} - t. \tag{17}$$

The addition of a confidence level α contrasts with previous approaches (see Loisel et al. (2021); Valla et al. (2023)). Setting $\alpha = 0$ means that the prediction $\hat{z}^{(i)}(t)$ is trusted with 100% confidence by the insurer, whereas letting α take higher values ensures that $\hat{z}^{(i)}(t)$ is positive with a given confidence interval. Another novelty here is the time dynamic of those results. Not only can we predict whether it is worth targeting a given policyholder, but we can also predict whether there will be some point in the future when targeting her will be more profitable. Predicting the trajectory of $z^{(i)}(t)$ at future time points requires projecting the time-varying covariates at those future time points. It can be done by either modeling such covariates individually or setting assumptions. It is trivial for covariates such as age or year but more complex for stochastic covariates such as the face amount. This framework does not aim to answer this question, and we assume in our application that stochastic covariates remain constant and equal to their last observed value. Regardless of the assumptions, the framework allows adding a time dimension to the LMS optimization and marketing decision-making. It is also worth noting that our developed framework is consistent in the time-invariant case. By design, it is also fully applicable with uncensored observations, or left-truncated ones. That shows our two-step framework's broad effectiveness and applicability regardless of right-censorship, left-truncation, risk factor, time-varying covariates, or time-varying effects. In that sense, it is a generalized framework for lapse management strategy in life insurance.

3 Application

3.1 Data

Our framework is inspired by a real-world life insurance dataset used in Valla et al. (2023). It initially contains the most recent information from 248 737 unique policies contracted between 1997 and 2018 and 235 076 unique policyholders. A single row originally represented a unique pair policy/policyholder, identified by a unique ID and denoted as a *subject*. Due to great computation times, we restain our application on a 10,000-subjects subset of this original dataset, but the astute reader will find more information about the complete one in the original article. The 10,000 rows dataset containing the last available information for the 10,000 selected subjects will be denoted $\mathcal{D}^{last}$, here is a subset for illustrative purposes:

Table 2 $\mathcal{D}^{last}$ random subset

ID	EVENT	PRODUCT	SEX	SENIORITY	F_i	CLAIM	CNTRCTS	AGE	YEAR
25737	1	1	1	17	0,73	0	2	76	2015
117322	1	1	2	10	4,32	0	1	63	2012
1322	0	1	2	20	9,82	0	1	75	2019
37433	2	1	2	14	0,99	-50,49	1	88	2011
23902	0	1	1	20	32,66	-13,12	2	71	2019
219281	0	2	2	8	7,08	0	2	71	2019
160112	0	1	2	15	0,04	0	1	51	2019
53108	2	1	2	12	13,11	0	1	92	2010
166078	1	2	2	5	9,02	0	1	64	2013
139644	0	1	1	16	5,65	-107,59	1	66	2019

Here, we were able to retrieve the longitudinal history of every subject present in $\mathcal{D}^{last}$: this means that for every policy and policyholder, we observe every payment, lapse, fee, profit sharing or discount rate from the policy subscription to the most updated information to date along with baseline covariates such as gender or age at subscription. For operational reasons, the longitudinal data are measured and reported yearly and organized as follows[1]:

Table 3 $\mathcal{D}^{long}$ random subset

ID	EVENT	START	END	PRODUCT	SEX	SENIORITY	F_i	CLAIM	CNTRCTS	AGE	YEAR
46784	0	0	1	3	2	0	8,38	0	1	66	2013
46784	0	1	2	3	2	1	8,40	0	1	67	2014
46784	0	2	3	3	2	2	8,57	0	1	68	2015
46784	0	3	4	3	2	3	11,90	0	1	69	2016
46784	0	4	5	3	2	4	12,10	0	1	70	2017
46784	0	5	6	3	2	5	12,28	0	1	71	2018
46784	1	6	7	3	2	7	15,06	-15,06	1	72	2019
7825	0	0	1	2	2	0	3,02	0	1	81	2016
7825	0	1	2	2	2	1	3,05	0	1	82	2017
7825	0	2	3	2	2	2	3,10	0	1	83	2018
7825	0	3	5	2	2	5	3,15	0	1	84	2019
264309	0	0	1	3	2	0	2,61	0	1	66	2016
264309	0	1	2	3	2	1	2,64	0	1	67	2017
264309	0	2	3	3	2	2	2,67	0	1	68	2018
264309	0	3	5	3	2	5	3,48	0	1	69	2019

Moreover, all the covariates describing financial flows are observed as cumulated over the years.

[1]But it is worth mentioning that covariates in actuarial datasets are usually updated continuously. In that case, we could build a continuous longitudinal dataset with one observation per policy change, and not one per year. The framework detailed here still applies in the continuous case.

As an example, let us assume that a subject subscribed in the year 2000: her payment variable for the year 2000 observation contains the sum of all payments that occurred in that year, her payment variable for the year 2001 contains the sum of all payments that occurred up to the year 2001 included (hence 2000 and 2001), and so on for the years after. This longitudinal dataset will be denoted $\mathcal{D}^{long}$. It contains 126,865 observations, in other words, almost 13 for each subject.

For privacy reasons, all the data, statistics, product names, and perimeters presented in this paper have been either anonymized or modified. All analyses, discussions, and conclusions remain unchanged.

3.2 Application: survival step

3.2.1 Survival analysis with time-varying covariates

The survival step, described in Section 2 requires survival tree-based models that can handle longitudinal time-varying covariates. Most survival tree-based models are analogous to regular tree-based models: survival trees work similarly to regular decision trees, creating partitions of the covariate space. What differentiates them is the splitting criterion that splits by maximizing the difference between two considered child nodes. Typically, at each node and for each split considered, a log-rank test is used to test the null hypothesis that there is no difference between the child nodes in the probability of an event at any time. The split that minimizes the p-value is then selected. By extension, a random survival forest is a random forest of survival trees.

As regression and classification trees, most survival trees are unable to deal with time-varying and longitudinal covariates. Indeed, let $x_1(t)$ be a numerical time-varying covariate. For a single tree, the splitting rule should be able to split subjects into two child nodes at each node. It would then be a rule of the form "$x_1 \leq s$". A subject for which this rule is true $\forall t$ will go in one child node without any ambiguity. On the other hand, the general case where the rule is true for some periods but false for anywhere else is unclear and needs to be addressed. Note that the same reasoning can be applied to categorical time-varying covariates as well. A simple idea is that the subject's observations in periods where the splitting rule is true would go to the left node, and the other would go to the right node, thus dividing one subject into several pseudo-subjects. With a longitudinal dataset, that method just implies considering all rows as independent which creates correlated right-censored and left-truncated (LTRC) observations that need special treatment. In such models, any individual can be spread in many different tree leaves - even if, at any fixed time, any individual will have a single observation that will fall into one unique leaf. Fu and Simonoff (2016a) proposed a model based on those ideas: they allowed subjects to be divided into pseudo-subjects and adjusted the log-rank test in the splitting procedure to accommodate for left truncation and ensure that the independence implicit assumption does not lead to biased results[2].

LTRC trees and forests yield an estimate of the survival function:

$$\widehat{S}\left(t \mid \mathcal{X}^{(i)}(t)\right) = P(T^{(i)} > t \mid \mathcal{X}^{(i)}(t)),$$

that can directly be used to evaluate the conditional incidence functions for competing risks. Bagging models of such trees then emerged (Yao et al. (2020)), with the usual prediction

[2]See Fu and Simonoff (2016a) for details on that point.

advantages and interpretability drawbacks of such bagging techniques[3]. In order to evaluate the survival models' performance, we chose to use the time-dependent Brier score (td-BS), integrated Brier score (td-IBS), Brier skill score (td-BSS) and integrated Brier skill score (td-IBSS) for longitudinal data (as in Yao et al. (2020)). More details about these metrics can be found in Appendix A.3.

3.2.2 Comparison settings

We propose here a comparison framework to measure the benefits of including the historical data in $\mathcal{D}^{long}$, compared to using $\mathcal{D}^{last}$. The matrices r_{lapser} and $r_{acceptant}$ are estimated with the algorithms LTRCRRF and LTRCCIF[4] from the R package LTRCforests, described in Section 3.2.1. In order to assess the advantages of that longitudinal model, we compare its results with those obtained with gradient boosting survival Model (GBSM) as it proved to be a high-performing non-longitudinal model on that dataset (See Valla et al. (2023)). With $T^{(i)}$, the "any event" time for subject i (that is the censoring time for active policies and the termination time, whatever the cause, for all others), r_{lapser} and $r_{acceptant}$ are estimated from the respective survival functions

$$\widehat{S}_{\text{lapser}}\left(t \mid \mathcal{X}^{(i)}(t)\right) = P(T^{(i)} > t \mid \mathcal{X}^{(i)}(t)),$$

$$\widehat{S}_{\text{acceptant}}\left(t \mid \mathcal{X}^{(i)}(t)\right) = P(T^{(i)} > t, \text{EVENT} = \text{death} \mid \mathcal{X}^{(i)}(t)),$$

with observations that ended with lapse considered as censored in the estimation of $\widehat{S}_{\text{acceptant}}$.

We want to compare the performance of all models trained with and without longitudinal data but also compare them on different tasks. Typically, predictions on $\mathcal{D}^{last}$ and $\mathcal{D}^{long}$ do not answer the same questions. The former aims at predicting the last observation of the target variable, and the latter aims at predicting its value at any given point in time. Graphically, depending on whether or not the model has been trained on longitudinal data or only on the most recent observation and with the different prediction goals described, this naturally designs four settings that answer four prediction problems:

(a) Models are trained on $\mathcal{D}^{last}_{train}$ and evaluated on predictions from $\mathcal{D}^{last}_{test}$

(b) Models are trained on $\mathcal{D}^{long}_{train}$ and evaluated on predictions from $\mathcal{D}^{last}_{test}$

(c) Models are trained on $\mathcal{D}^{last}_{train}$ and evaluated on predictions from $\mathcal{D}^{long}_{test}$

(d) Models are trained on $\mathcal{D}^{long}_{train}$ and evaluated on predictions from $\mathcal{D}^{long}_{test}$

Setting (a) is the classical setting, where any subject has only one measurement, and the prediction task is also to predict a variable at one given time point. Conversely, setting (d) represents the longitudinal setting, where models are trained with longitudinal time-varying

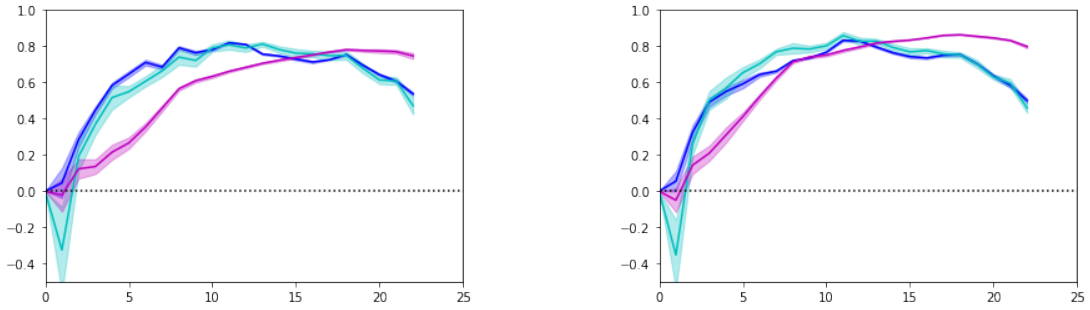
[3]Both methods have been implemented in the R packages `LTRCtrees` and `LTRCforests`, and are considered state-of-the-art methods for tree-based survival analysis with time-varying covariates.

[4]In the following sections, we consider LTRCRRF and LTRCCIF: LTRC forests respectively based on regular CART and conditional inference survival tree algorithms. More insights about those models can be found in the references detailed in Section 3.2.1

covariates and where the prediction task aims at retrieving the value of a target variable at any given time point during a subject's lifetime. Setting (c) is uninsightful as a model trained on aggregated data cannot retrieve longitudinal information and is expected to perform poorly by design. Intermediate setting (b) is also insightful as it can be used to highlight the added value of the information contained in longitudinal data when training a model. The comparison is made on a time-varying survival evaluation metric: the time-dependent Brier Skill Score (td-BSS) for longitudinal data (see Appendix A.3).

3.2.3 Results

First of all, in order to assess the superiority of longitudinal models in a longitudinal context, we need to compare all our considered models in the classical aggregated setting: with training and testing phases on subsets of $\mathcal{D}^{last}$. We can see that in this non-longitudinal setting, GBSM and LTRC models (LTRCRRF and LTRCCIF) are close in terms of BSS. Figure 3 displays the td-BSS on the y-axis, for which a value of 0 means that the score for the predictions is merely as good as that of a naïve prediction[5] and a value of 1 is the best score possible. BSSs are computed for every time point, meaning that we can observe and compare the performance of models on estimating retention probabilities for low-seniority policies or high-seniority ones independently.



td-BSS for $r_{acceptant}$ td-BSS for r_{lapser}

Figure 3: td-BSS of models trained on $\mathcal{D}^{last}_{train}$ and tested on $\mathcal{D}^{last}_{test}$

The IBSS, the mean of BSSs over all time points (see Appendix A.3), indicates that LTRCRRF performs slightly better than LTRCCIF, hence we will drop LTRCCIF for the rest of this application. In practice, the underlying true survival distribution may possess a complex structure with time-varying covariates and time-varying effects. The cross-validated Brier scores and Brier Score Skills graphs can potentially lead decision makers to choose different survival estimations at different time points and not a unique choice of method for all time points.

By contrast, the difference between those models is evident and significant whenever they are trained on longitudinal data. The graphs below show the difference in terms of BSS over time in prediction settings (b) and (d):

The conclusion regarding prediction richness contained in longitudinal data and accuracy benefits from using dedicated longitudinal methods is clear. Longitudinal models perform significantly better, and GBSM brings minor improvement over naïve models.
In the end, we select LTRCRRF for estimating the retention probabilities in the *survival step*

[5] in our application, the empirical estimate of the survival function has been chosen as the naïve prediction.

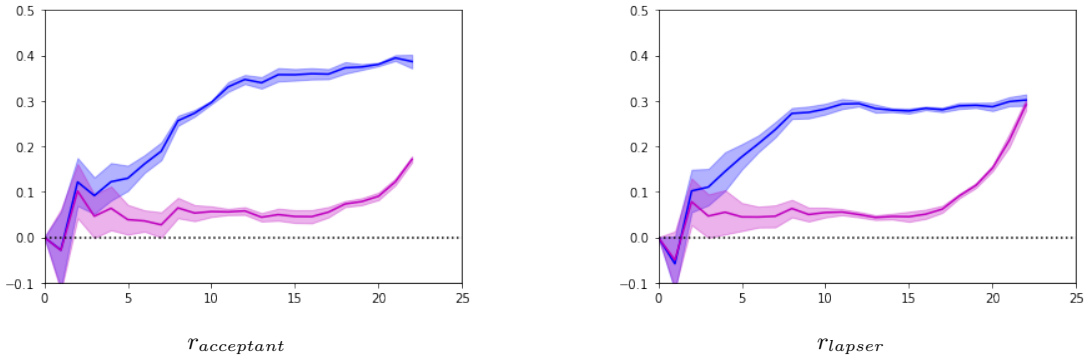


Figure 4: td-BSS of models trained on $\mathcal{D}_{train}^{long}$ and tested on $\mathcal{D}_{test}^{last}$ - Setting (b)

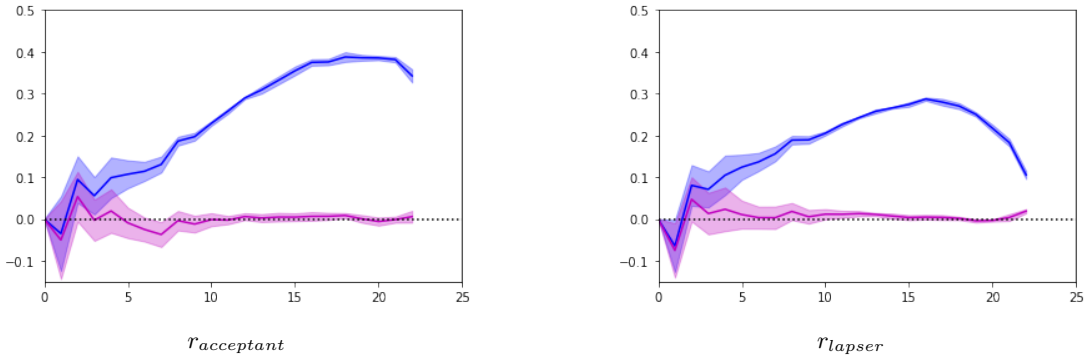


Figure 5: td-BSS of models trained on $\mathcal{D}_{train}^{long}$ and tested on $\mathcal{D}_{test}^{long}$ - Setting (d)

as it shows to be the best model when trained on longitudinal data.

It is to be noted that the results of that modeling approach in terms of global retention gain (Equation 16) are not necessarily better than the results obtained without the use of longitudinal data in the estimation of r_{lapser} and $r_{acceptant}$. In other words, a better performance of the model used in the *survival step* does not lead to an increase in the insurer's expected profit, for a given LMS but to a more realistic estimation of it as they model the CLV more accurately.

With that, we determine r_{lapser} and $r_{acceptant}$, the conditional retention probabilities for every observation to derive the trajectory of the observed individual CLV, RG, and eventually $z^{(i)}(t)$ (see Equation 13). The latter can then be used as a longitudinal target variable in a regression model: this constitutes the *regression step*, introduced in Section 1 and detailed within this application in the next Section.

Another advantage of using longitudinal data for survival analysis is that it helps study how a given subject's retention probabilities are updated with time. We take the example of a randomly selected subject and plot her retention probability at every observation time:

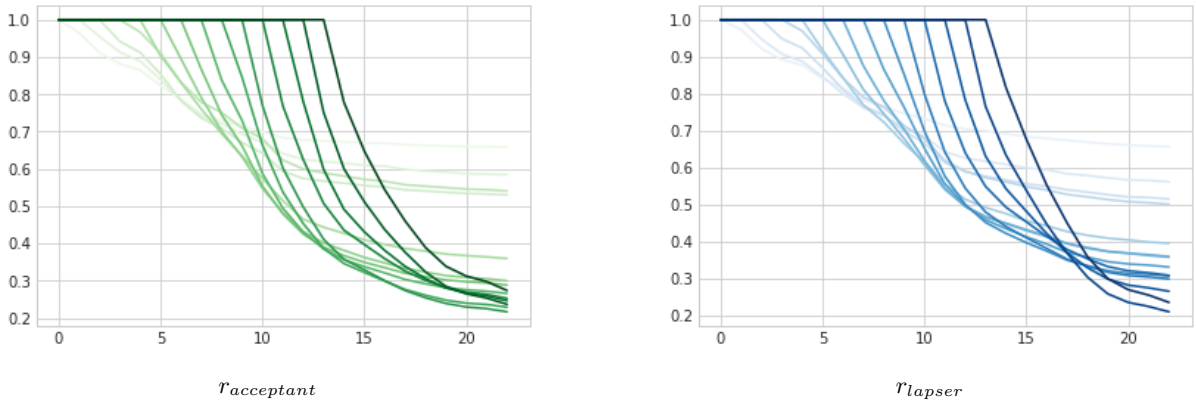


Figure 6: Longitudinally updated retention trajectories for a random subject

The further in time the observation is, the more pellucid the survival curve is. The individual retention curves are updated as new measurements are available.

3.3 Application: regression step

3.3.1 Regression analysis with time-varying covariates

The *regression step* of the framework introduced in Section 2.2 requires using a regression model allowing for longitudinal data to produce an estimate of $z^{(i)}(t)$. We chose to use mixed-effect tree-based models (METBM). First of all, a mixed-effect model is designed to work on clustered data in general, including longitudinal data (see Verbeke et al. (1997)). Sela and Simonoff (2012), Capitaine et al. (2021), Fu and Simonoff (2015) and Hajjem et al. (2014) describe a procedure to fit a mixed effect model using tree-based models through an iterative two-step process[6]. Mixed effect tree-based algorithms are designed to take clustered data as input. By considering subjects as clusters, they can grasp the dependence structure within the different observations of a single subject and can be used for longitudinal analysis (see Verbeke et al. (1997)). The underlying idea behind mixed effect tree-based algorithms is to assume a mixed model for the longitudinal outcome and estimate the random effect parameters with a tree-based model. Such approaches estimate the random effects of a mixed model in the first step, then construct a regression tree with the fixed-effect covariates on the original outcome, excluding the estimated random effect. The idea is to repeat these two steps: the model parameters and the random effects are estimated iteratively until convergence, similar to the two-step well-known EM optimization procedure. Suppose that we have p_f covariates with a fixed effect and p_s covariates with a random effect. Initially, a parametric linear mixed-effect model is given by

$$\boldsymbol{z^{(i)}} = F^{(i)\top}\boldsymbol{\beta} + S^{(i)\top}\boldsymbol{b^{(i)}} + \boldsymbol{\epsilon^{(i)}}. \tag{18}$$

where $\boldsymbol{z^{(i)}}$ is the $n^{(i)} \times 1$ longitudinal vector outcome of subject i, $\boldsymbol{\beta}$ is the $p_f \times 1$ vector of the fixed effect coefficients and $F^{(i)}$ is the $n^{(i)} \times p_f$ design matrix of the covariates with a fixed effect. The quantity $\boldsymbol{b^{(i)}}$ is the $p_s \times 1$ vector of random effects and $S^{(i)}$ is the $n^{(i)} \times p_s$ design matrix of the covariates with a subject-specific effect. By construction, $F^{(i)}$ and $S^{(i)}$ are subdivisions of the covariate space. The error term $\boldsymbol{\epsilon^{(i)}}$ is the $n^{(i)} \times 1$ vector of residuals, assumed to come from a normal distribution with mean 0 and variance σ^2, and we assume $\boldsymbol{b^{(i)}} \sim \mathcal{N}(0, D)$, $\boldsymbol{\epsilon^{(i)}} \sim \mathcal{N}(0, \sigma^2 \cdot \mathbb{I}_{n^{(i)}})$. Eventually, D is the $p_s \times p_s$ variance-covariance matrix

[6]The algorithms corresponding to their respective work are available in the R packages `REEMtree` and `LongituRF`, the R function "REEMctree" and the Python library `MERF`.

for the random effects.

In order to model a longitudinal outcome with non-linear fixed effects, a tree-based model is included in Equation 18, as follows:

$$\boldsymbol{z^{(i)}} = f(F^{(i)}) + S^{(i)\top}\boldsymbol{b^{(i)}} + \boldsymbol{\epsilon^{(i)}}. \tag{19}$$

Here the linear structure of the fixed effect part of the model is generalized: the fixed effects are described by a function of the fixed-effect covariates f, which is the part that a tree-based model will estimate. In MERT (Hajjem et al. (2011)), the tree-based model is a single regression tree, in MERF (Hajjem et al. (2014)), it is a random forest, whereas in RE-EM (Sela and Simonoff (2012); Fu and Simonoff (2015)) it can be both. A general algorithm for such mixed effect tree-based models can be described as follows:

Algorithm 1 Mixed effect tree-based model pseudo-code

Input: $\mathcal{D}$, a longitudinal dataset with an outcome $\boldsymbol{z^{(i)}}$, $\forall i \in [1 \dots N]$
Output: $\hat{\boldsymbol{z}}^{(\boldsymbol{i})}$, $\widehat{f}$, $\hat{\boldsymbol{b}}^{(\boldsymbol{i})}$, $\hat{\boldsymbol{\epsilon}}^{(\boldsymbol{i})}$, $\hat{\sigma}^{(i)^2}$, $\hat{D}^{(i)}$, $\forall i \in [1 \dots N]$

Initialize: $\widehat{b} \leftarrow 0$, $\widehat{\sigma^2} \leftarrow 1$, $\widehat{D} \leftarrow \mathbb{I}_{p_s}$
while GLL $<$ some convergence threshold **do**
    1. $\boldsymbol{z^{(i)}} \leftarrow \boldsymbol{z^{(i)}} - S^{(i)\top}\boldsymbol{b_i}$
    2. Fit a tree-based model on $\boldsymbol{z^{(i)}}$ and obtain $\widehat{f}$
    3. Infer the updated random effects parameters $\hat{\boldsymbol{b}}^{(\boldsymbol{i})}$
    4. Compute $\hat{\boldsymbol{\epsilon}}^{(\boldsymbol{i})} = \boldsymbol{z^{(i)}} - \hat{f}(F^{(i)}) - S^{(i)\top}\hat{\boldsymbol{b}}^{(\boldsymbol{i})}$
    5. Update $\hat{\sigma}^{(i)^2}$ and $\hat{D}^{(i)}$
    6. Update GLL, the generalized log-likelihood criterion used to control for convergence
end while

For further details about all these elements - and notably, the update formulas for $\hat{\sigma}^{(i)^2}$, $\hat{D}^{(i)}$ and GLL - we refer the astute reader to the work of Hajjem et al. (2014) (see Section 2 for details on how the between-subject standard error can be estimated from a METBM). Once fit, the mixed-effect tree-based model can be used to predict the vector $\hat{\boldsymbol{z}}^{(\boldsymbol{i})}$, the longitudinal predicted trajectory of an LMS-induced profit for any subject. For subjects with past observations included in the training dataset, the prediction includes the random effect correction:

$$\hat{\boldsymbol{z}}^{(\boldsymbol{i})} = \hat{f}(F^{(i)}) + S^{(i)\top}\hat{\boldsymbol{b}}^{(\boldsymbol{i})}.$$

For a new subject, with a first observation in the testing set, the mixed-effect prediction only includes the fixed effect:

$$\hat{\boldsymbol{z}}^{(\boldsymbol{i})} = \hat{f}(F^{(i)}).$$

Moreover, making predictions with such models at given times imposes that we know the value of the longitudinal covariates at those times. This implies that to compute future values of $z^{(i)}(t)$, future unknown values of the longitudinal covariates are needed. In other words, no predictions for any subject are made beyond that subject's last observation time value unless we assume future values of the longitudinal covariates.

3.3.2 Results

This section contains the results of the *regression step* of our framework. In order to model whether a policyholder is worth targeting or not, we fit a mixed-effect tree-based regression

model to our longitudinal dataset with $\boldsymbol{z^{(i)}}$, the vector of $n^{(i)}$ observations as a longitudinal target variable for every subject i. As $\boldsymbol{z^{(i)}}$ can take any real value, the mean squared error (MSE) in the tree-based part of the mixed-effect model is to be preferred. For a given LLMS, the survival step allows us to compute $\boldsymbol{z^{(i)}}$, the longitudinal variable representing the expected trajectory of the profits or losses generated by subject i. Then, by estimating $\boldsymbol{z^{(i)}}$ on various LLMS with a mixed effect tree-based model, we can hope to find an optimal retention strategy in the sense that it will maximize the expected gain for the life insurer. For this application, we assume parameters p, η, γ, and d to be constant over all policyholders and over time and we fit a mixed effect random forest (MERF). We suggest testing five LLMS:

- one that is obviously and extremely bad and would lead to a loss for the insurer, if applied to a large number of subjects (LLMS n°1)
- one that is unrealistically good, with a small incentive largely accepted and would lead to a sure profit for the insurer (LLMS n°2)
- three realistic strategies, with various degrees of aggressivity (LLMS n°3, 4 and 5)

We train our targeting mixed-effect random forest model on all observations and their respective retention probabilities up to 2020 and test it on all subjects with an observation in 2021. We can note that in 2021, there are predictions on subjects with past observations prior to 2021 but also predictions on new subjects not included in the training set. Overall, the testing set contains "only" 4,472 unique policyholders, hence the order of magnitude of the retention gains presented below. We also chose a very conservative risk parameter, that greatly reduces the number of subjects targetted.

Here are the five strategies, and the corresponding expected profit or loss[7] they induce:

Table 4 Various LMS results with our framework

LMS n°	p	η	γ	c	d	T	RG	# targets	campain investement
1	1%	1%	90%	200	2.00%	10	0	0	0
2	5%	0.01%	80%	5	2.00%	20	134,347.54	141	705
3	3%	0.009%	40%	15	1.50%	20	3,112.03	98	1470
4	2.5%	0.005%	15%	10	1.50%	20	2,940.51	94	940
5	3%	0.001%	5%	5	1.50%	20	2,962.68	122	610

Evidently, the main feature proposed by this framework is that it allows the decision maker to choose the best LLMS among realistic ones. In our application, we immediately see that in terms of profit for the insurer, strategy n°3 is optimal, compared to LLMS n°4 and 5. On the other hand, other factors, such as the number of policyholders to target or the cost of the campaign, are also displayed. they can prove to be critical elements of decisions in a real-world context, as some life insurers could have a limited commercial workforce or investment budget. For instance, an insurer that can only contact up to 95 policyholders this year would choose LLMS n°4, and another that would be limited by a 1,000€ budget for retention would choose LLMS n°5. Moreover, the bad LMS n°1 demonstrates that this framework allows us to detect

[7] As defined in Definition 2

whenever a strategy should not be carried out. In that case, the conclusion of the targeting step is not to target any policyholder, thus limiting the insurer's loss to 0, which is arguably a desirable feature. Finally, the unrealistically good LLMS n°2 shows that this framework cannot detect a "too good to be true" strategy with an unrealistic pair of parameters (η, γ). This emphasizes the fact that taking this interdependency into account directly in the framework should prevent such unrealistic scenarios and avoid the life insurer the task of selecting in advance a consistent set of LLMS parameters. Another novelty in this framework is the longitudinal structure of the results. Indeed, we can easily retrieve the expected individual loss or profit at any future time. For example, here is a plot of the expected profits generated by targeting randomly selected policyholders:

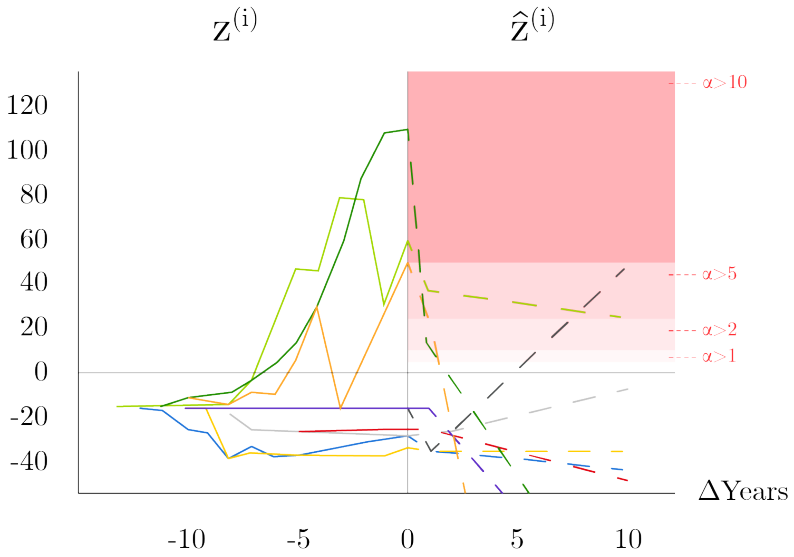


Figure 7: Projections of targetted profits over time

Most policyholders have a $\hat{z}_i$ with a decreasing future trajectory. It makes sense as time is positively correlated with one's policy probability to end: the more the insurer waits to offer an incentive to a subject, the less profitable it becomes. Usually, if a policyholder does not generate profit by being targeted now, it is even less relevant to target her later in time. For specific profiles, the lapse risk grows faster than the death risk. It can then become more profitable to offer an incentive as the lapse risk increases if the death risk is insignificant.

In any case, we show graphically that depending on the level of risk α that the insurer consent to take, the time at which it is optimal to apply an LLMS to a given policyholder changes. The longitudinal trajectory being estimated with a linear model, the framework as it stands should not be used to evaluate the time when offering an incentive is optimal. It rather yields information about individual tendencies and answers strategical questions: is it profitable to target a given policyholder now? If not now, is it likely to become profitable in the future? And if it is, should the insurer decide quickly or can it wait? The individual intercepts and slopes of the future estimations of $\hat{z}^{(i)}$ answer those questions.

This example of a time-dynamic application shows that including longitudinal data in a lapse management strategy can benefit a life insurer in terms of prediction accuracy and decision-making.

4 Conclusion, limitations and future work

In conclusion, this paper presents a novel longitudinal lapse management framework that is tailored specifically for life insurers. The framework enhances the targeting stage of retention campaigns by selectively applying it to policyholders who are likely to generate long-term profits for the life insurer. Our key contribution is the adaptation of existing methodologies to a longitudinal setting through the use of tree-based models. The results of our application demonstrate the advantages of approaching lapse management in a longitudinal context. The use of longitudinally structured data significantly improves the precision of the models in predicting lapse behavior, estimating customer lifetime value, and evaluating individual retention gains. The implementation of mixed-effect random forests enables the production of time-varying predictions that are highly relevant for decision-making. The framework is designed to prevent the application of loss-inducing strategies and allows the life insurer to select the most profitable LMS, under constraints.

However, our work has several limitations that must be acknowledged:
Firstly regarding the framework: the longitudinal lapse management strategy is defined with fixed incentive, probability of acceptance and cost of contact, regardless of the time in the future. Moreover, the γ parameter is constant for a given policyholder, but it could be seen as the realization of a random variable following a chosen distribution. Those points may restrict the framework's practical effectiveness. Moreover, we did not account for the interdependence between different LLMS parameters, which could lead to the implementation of unrealistic strategies. Additionally, the introduction of the confidence level α could be discussed further as it could be linked with actuarial risk measures such as the Value-at-Risk. Eventually, the article describes a discrete-time longitudinal methodology, but in general, the insurer has access to the precise dates of any policy's financial flows. Thus, a continuous-time framework could also be implemented.
Secondly regarding the application: a lot of assumptions have been formulated in the application we propose such as constant parameters where the framework allows them to vary across time and policyholders, or the use of MERF where more complex and completely non-linear models could be tried.
Finally regarding longitudinal tree-based models: the use of LTRC and MERF requires the management of time-varying covariates with the pseudo-subject approach, which has practical limitations and prevents the longitudinal data from being predicted alongside the target variable. Future works could address the latter remark using joint models (see Appendix A.1 for references).

The limitations of the general framework should be discussed and tackled in forthcoming research. Other use-case and applications, with sensitivity analysis over various sets of parameters, models and datasets could constitute an engaging following work. Pseudo-subjects limitations are inherent in the current design of longitudinal tree-based models. Future work will involve developing innovative algorithms to address these issues. Overall, this article contributes to the field of lapse analysis, and our framework has the potential to improve retention campaigns and increase long-term profitability for a life insurer.

Acknowledgments and competing interests

Work(s) conducted within the Research Chair DIALog under the aegis of the Risk Foundation, an initiative by CNP Assurances.

References

Eva Ascarza, Scott A. Neslin, Oded Netzer, Zachery Anderson, Peter S. Fader, Sunil Gupta, Bruce Hardie, Aurelie Lemmens, Barak Libai, David T. Neal, Foster Provost, and Rom Schrift. In pursuit of enhanced customer retention management: Review, key issues, and future directions, 2018. Special Issue on 2016 Choice Symposium. Customer Needs and Solutions 5,.

Peter Austin, Aurélien Latouche, and Jason Fine. A review of the use of time-varying covariates in the Fine-Gray subdistribution hazard competing risk regression model. *Statistics in Medicine*, October 2019. doi: 10.1002/sim.8399. URL https://hal-cnam.archives-ouvertes.fr/hal-02349043.

Anna Rita Bacinello. Endogenous model of surrender conditions in equity-linked life insurance. *Insurance: Mathematics and Economics*, 37(2):270–296, 2005. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2005.02.002. URL https://www.sciencedirect.com/science/article/pii/S0167668705000156. Papers presented at the 8th IME Conference, Rome, 14-16 June 2004.

Paul Berger and Nada Nasr. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12:17 – 30, 12 1998. doi: 10.1002/(SICI)1520-6653(199824)12:1⟨17::AID-DIR3⟩3.0.CO;2-K.

Bavo D. C. Campo and Katrien Antonio. Insurance pricing with hierarchically structured data: An illustration with a workers' compensation insurance portfolio, 2022. URL https://arxiv.org/abs/2206.15244.

Louis Capitaine, Robin Genuer, and Rodolphe Thiébaut. Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, 30(1):166–184, 2021. doi: 10.1177/0962280220946080. PMID: 32772626.

Manon Dal Pont. *Construction d'une table de mortalité d'expérience en assurance emprunteur*. PhD thesis, ISFA, Université Lyon 1, 2020. URL https://www.institutdesactuaires.com/docs/mem/1e992efa93786a498553e9a184326e4a.pdf.

Bas Donkers, Peter Verhoef, and Martijn Jong. Modeling clv: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*, 5:163–190, 02 2007. doi: 10.1007/s11129-006-9016-y.

Martin Eling and Michael Kochanski. Research on lapse in life insurance: what has been done and what needs to be done? *Journal of Risk Finance*, 14(4):392–413, 2013. URL https://EconPapers.repec.org/RePEc:eme:jrfpps:v:14:y:2013:i:4:p:392-413.

Lloyd D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1):145–157, 1999. doi: 10.1146/annurev.publhealth.20.1.145. PMID: 10352854.

Edward W. Frees, Catalina Bolancé, Montserrat Guillen, and Emiliano A. Valdez. Dependence modeling of multivariate longitudinal hybrid insurance data with dropout. *Expert Systems with Applications*, 185:115552, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115552. URL https://www.sciencedirect.com/science/article/pii/S0957417421009581.

Wei Fu and Jeffrey Simonoff. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics (Oxford, England)*, 18, 06 2016a. doi: 10.1093/biostatistics/kxw047.

Wei Fu and Jeffrey S. Simonoff. Unbiased regression trees for longitudinal and clustered data. *Computational Statistics & Data Analysis*, 88:53–74, 2015. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2015.02.004. URL https://www.sciencedirect.com/science/article/pii/S0167947315000432.

Wei Fu and Jeffrey S. Simonoff. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2):352–369, 12 2016b. ISSN 1465-4644. doi: 10.1093/biostatistics/kxw047.

Leo Guelman, Montserrat Guillén, and Ana M. Pérez-Marín. Random forests for uplift modeling: An insurance customer retention case. In Kurt J. Engemann, Anna M. Gil-Lafuente, and José M. Merigó, editors, *Modeling and Simulation in Engineering, Economics and Management*, pages 123–133, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-30433-0.

Sunil Gupta, Donald R Lehmann, and Jennifer Ames Stuart. Valuing customers. *Journal of marketing research*, 41(1):7–18, 2004.

Ahlem Hajjem, François Bellavance, and Denis Larocque. Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, 81(4):451–459, 2011. ISSN 0167-7152. doi: https://

doi.org/10.1016/j.spl.2010.12.003. URL https://www.sciencedirect.com/science/article/pii/S0167715210003433.

Ahlem Hajjem, François Bellavance, and Denis Larocque. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328, 2014. doi: 10.1080/00949655.2012.741599.

Mary Hardy. *Investment guarantees: modeling and risk management for equity-linked life insurance*, volume 168. John Wiley & Sons, 2003.

Dennis M. Heisey and Brent R. Patterson. A review of methods to estimate cause-specific mortality in presence of competing risks. *The Journal of Wildlife Management*, 70 (6):1544–1555, 2006. doi: https://doi.org/10.2193/0022-541X(2006)70[1544:AROMTE]2.0.CO;2. URL https://wildlife.onlinelibrary.wiley.com/doi/abs/10.2193/0022-541X%282006%2970%5B1544%3AAROMTE%5D2.0.CO%3B2.

Stéphane Loisel, Pierrick Piette, and Cheng-Hsien Jason Tsai. Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin*, 51(3):839–871, 2021. doi: 10.1017/asb.2021.10.

Anne MacKay, Maciej Augustyniak, Carole Bernard, and Mary R. Hardy. Risk management of policyholder behavior in equity-linked life insurance. *Journal of Risk and Insurance*, 84(2):661–690, 2017. doi: https://doi.org/10.1111/jori.12094. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jori.12094.

G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer Series in Statistics. Springer New York, 2006. ISBN 9780387289809. URL https://books.google.fr/books?id=LjfyKpw36S8C.

Hoora Moradian, Weichi Yao, Denis Larocque, Jeffrey S. Simonoff, and Halina Frydman. Dynamic estimation with random forests for discrete-time survival data. *Canadian Journal of Statistics*, 50(2):533–548, 2022. doi: https://doi.org/10.1002/cjs.11639. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11639.

J. Francois Outreville. Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance: Mathematics and Economics*, 9(4):249–255, 1990. URL https://EconPapers.repec.org/RePEc:eee:insuma:v:9:y:1990:i:4:p:249-255.

Hans Risselada, Peter C. Verhoef, and Tammo H.A. Bijmolt. Staying power of churn prediction models. *Journal of Interactive Marketing*, 24(3):198–208, 2010. doi: 10.1016/j.intmar.2010.04.002.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Chapman & Hall/CRC, Boca Raton, 2012.

Thomas H. Scheike and Torben Martinussen. *Dynamic Regression models for survival data*. Springer, NY, 2006.

Thomas H. Scheike and Mei-Jie Zhang. Analyzing competing risk data using the R timereg package. *Journal of Statistical Software*, 38(2):1–15, 2011. URL https://www.jstatsoft.org/v38/i02/.

Rebecca Sela and Jeffrey Simonoff. Re-em trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86:169–207, 02 2012. doi: 10.1007/s10994-011-5258-3.

Mathias Valla, Xavier Milhaud, and Anani Ayodélé Olympio. Including individual Customer Lifetime Value and competing risks in tree-based lapse management strategies. *European Actuarial Journal*, September 2023. doi: 10.1007/s13385-023-00358-0. URL https://hal.science/hal-03903047.

Geert Verbeke, Geert Molenberghs, and Geert Verbeke. *Linear mixed models for longitudinal data*. Springer, 1997.

Weichi Yao, Halina Frydman, Denis Larocque, and Jeffrey S. Simonoff. Ensemble methods for survival function estimation with time-varying covariates, 2020. URL https://arxiv.org/abs/2006.00567.

Weichi Yao, Halina Frydman, Denis Larocque, and Jeffrey S. Simonoff. Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, 31(11):2217–2236, 2022. doi: 10.1177/09622802221111549. PMID: 35895510.

A Appendix

A.1 Note on parametric models

This work focuses on the ability of non-parametric tree-based approaches to perform in both steps of our framework. For comparison's sake, a semi-parametric survival model had been fitted in Valla et al. (2023); it is important to explain why we did not investigate such models here. Time-varying Cox-like models also exist and can even take competing risks into account. They can be compared and yield survival curves for any individual but only up to their last observed time. Predicting survival probabilities at future time points is not possible. For the astute reader, a complete implementation of those techniques can be found in the R package `timereg` by Scheike and Martinussen (2006); Scheike and Zhang (2011).
Moreover, other prediction biases can appear in the presence of endogenous longitudinal covariates, with Cox-like models Austin et al. (2019), which is typically our situation. This is why we decided to leave such modeling approaches out of this paper.

It is to be noted that a statistical learning approach addressing research questions involving the association structure between longitudinal data and an event time exists: joint models. This type of modeling technique is primarily used in time-to-event contexts, with censored data and can handle multiple exogenous and endogenous longitudinal covariates with possibly multiple competing risks. Joint models outweigh time-dependent Cox models in terms of prediction; by predicting both the longitudinal trajectories and the survival probabilities simultaneously, it is possible to compute the conditional probability of surviving later than the last observed time for which a longitudinal measurement was available. They have been extensively studied and extended and have proved to yield competitive predictive results for relatively small datasets. A complete overview of such models can be found in Rizopoulos (2012), and their implementation is available in R packages `JM`, `JMBayes` and `JMBayes2`. Joint models are performant but computationally expensive for large datasets and multiple longitudinal covariates or outcomes. We did not implement this approach in this paper for those reasons and instead implement tree-based models handling time-varying covariates that we will compare to tree-based models with time-fixed covariates.

A.2 Model selection methodology

Regardless of their size, $\mathcal{D}^{last}$ and $\mathcal{D}^{long}$ both relate to 10,000 subjects. In order to tune the models detailed in the next Sections, we adopt a 5-fold Monte-Carlo cross-validation methodology. We randomly select 80% of subjects' observations in $\mathcal{D}^{last}$ and $\mathcal{D}^{long}$ as training sets, and the remaining 20% of subjects' observations go in testing sets. Models are trained on the training sets and tested on both training and testing sets to control for overfitting. We repeat this step 5 times such that we obtain 20 different datasets: ${}^{k}\mathcal{D}^{last}_{train}$, ${}^{k}\mathcal{D}^{last}_{test}$, ${}^{k}\mathcal{D}^{long}_{train}$ and ${}^{k}\mathcal{D}^{long}_{test}$ for $k \in [1, ..., 5]$. We can illustrate this as follows:

In the following Sections, this will be our methodology for studying the mean and variance of all considered models' performances. All presented conclusions are the results of a 5-fold Monte-Carlo cross-validation.

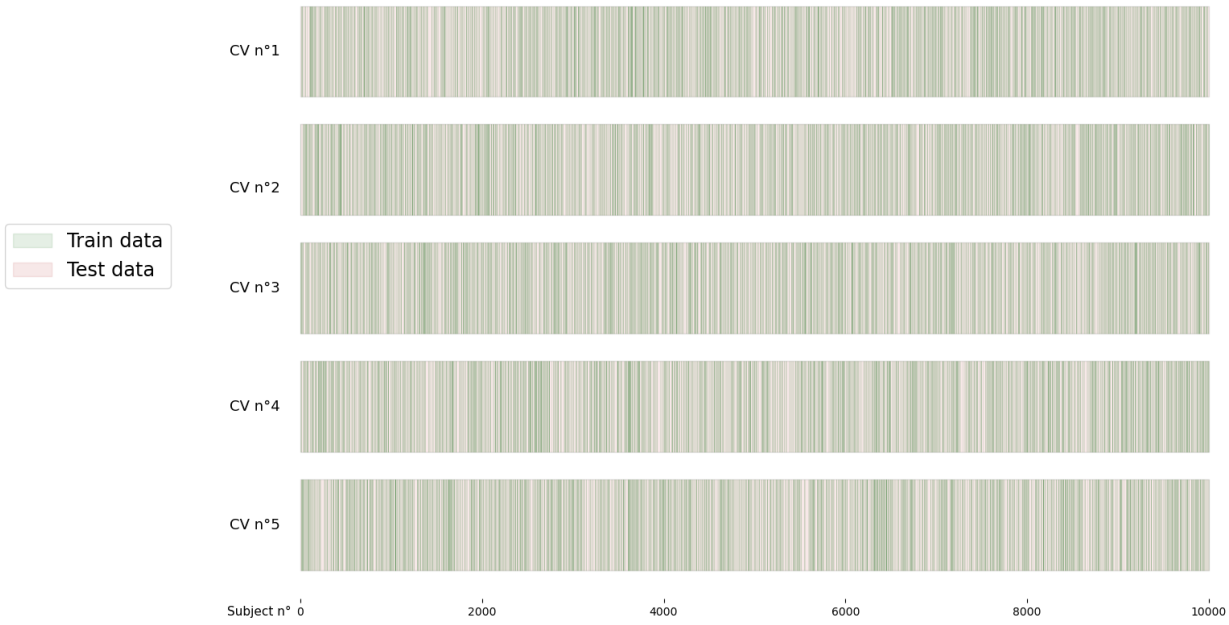


Figure 8: Monte-Carlo cross-validation

A.3 Brier scores and Brier skill scores

A.3.1 BS

Let $\mathcal{D}$ be a longitudinal life insurance dataset and let us assume that only one event can occur. Survival models yield $\widehat{S}(t|\mathcal{X}^{(i)}(t))$ the predicted probability of staying active up to time t given all past observations $\boldsymbol{x}^{(i)}$, $\widehat{G}$ the Kaplan-Meier estimate of the censoring distribution and $\widehat{W}^{(i)}(t)$ the corresponding inverse probability of censoring weights (IPCW), the time-dependent Brier Score is given by:

$$\widehat{\mathrm{BS}}(t, \widehat{S}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \widehat{W}^{(i)}(t) \left[\delta^{(i)}(t) - \widehat{S}\left(t \mid \mathcal{X}^{(i)}(t)\right)\right]^2.$$

With the notations introduced in Section 2.1, the IPCW are being computed as follows:

$$\widehat{W}^{(i)}(t) = \frac{\left(1 - \delta^{(i)}(t)\right) \Delta^{(i)}}{\widehat{G}\left(T^{(i)}\right)} + \frac{\delta^{(i)}(t)}{\widehat{G}(t)}.$$

The td-BS yields a vector of scores, each evaluated at different time points. In order to get a unidimensional evaluation metric, we also compute the time-dependent integrated Brier Score (td-IBS), defined as :

$$\widehat{\mathrm{IBS}}(\widehat{S}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{T^{(i)}} \int_0^{T^{(i)}} \widehat{W}^{(i)}(t) \left[\delta^{(i)}(t) - \widehat{S}\left(t \mid \mathcal{X}^{(i)}(t)\right)\right]^2 dt.$$

A.3.2 BSS

The td-BS and td-IBS are consistent with our framework but have some drawbacks as they do not give insights regarding the models' performance compared to naïve approaches. In order

to do so, we use the Brier Skill Score (BSS) and integrated Brier Skill Score (IBSS), defined as follows:

$$\widehat{\text{BSS}}(t, \widehat{S}; \mathcal{D}) = 1 - \frac{\widehat{\text{BS}}(t, \widehat{S}; \mathcal{D})}{\widehat{\text{BS}}(t, \widehat{S_{ref}}; \mathcal{D})}$$

$$\widehat{\text{IBSS}}(\widehat{S}; \mathcal{D}) = 1 - \frac{\widehat{\text{IBS}}(\widehat{S}; \mathcal{D})}{\widehat{\text{IBS}}(\widehat{S_{ref}}; \mathcal{D})}.$$

BSS measures the BS improvement over some reference metric. We see that it takes positive (or negative) values whenever the $\widehat{\text{BS}}(t, \widehat{S}; \mathcal{D})$ -respectively $\widehat{\text{IBS}}(\widehat{S}; \mathcal{D})$ - is inferior (or superior) to $\widehat{\text{BS}}(t, \widehat{S_{ref}}; \mathcal{D})$ - respectively $\widehat{\text{IBS}}(\widehat{S_{ref}}; \mathcal{D})$. Those reference metrics are the BS - respectively IBS - obtained with $\widehat{S_{ref}}$, a predicted probability of staying active up to time t constant and equal to the proportion of active policies in $\mathcal{D}$. The BSS and IBSS represent the improvement in terms of Brier Score over a reference model: the higher, the better.

A.4 Estimation of π_*

Very intuitively, for policyholders linked to a non-active policy, the last observation ended with either lapse or death and $\Delta^{(i)} \neq 0$. For any observation related to a policyholder that eventually lapsed $\pi_*^{(i)} = 1$. For any observation related to a policy that eventually ended with the policyholder's death, we have $\pi_*^{(i)} = 0$. Deriving $\pi_*^{(i)}$ is more complex for policyholders with an active policy where we have

$$\pi_*^{(i)} = P(\Delta_*^{(i)} = 1 | \Delta^{(i)} = 0, \mathcal{X}^{(i)}(T^{(i)})) = \frac{P\left(\Delta_*^{(i)} = 1, \Delta^{(i)} = 0 \mid \mathcal{X}^{(i)}(T^{(i)})\right)}{P\left(\Delta^{(i)} = 0 \mid \mathcal{X}^{(i)}(T^{(i)})\right)}. \tag{20}$$

By treating the competing risks within the cause-specific framework, we have that the probability of having an active policy, in other words having survived every cause of events, is the product of the cause-specific probabilities (See Heisey and Patterson (2006)). Given the risk profiles that we introduced in Section 1, we define $r_{lapser}^{(i)}(t)$ the all-causes survival probability of subject i at time t and $r_{acceptant}^{(i)}(t)$ the death survival probability of subject i at time t. Moreover, in practice, we only have access to a limited history $T_{max} = max(T^{(i)})$, corresponding to the longest time a policy was ever observed to last. In order to estimate $\pi_*^{(i)}$, we will consider that the ultimate event time $T_*^{(i)}$ is bounded by T. Thus we have

$$\pi_*^{(i)} = \frac{1 - {r_{lapser}^{(i)}(T_{max})}/{r_{acceptant}^{(i)}(T_{max})}}{{r_{lapser}^{(i)}(T^{(i)})}/{r_{acceptant}^{(i)}(T^{(i)})}} = \frac{r_{acceptant}^{(i)}(T^{(i)})}{r_{lapser}^{(i)}(T^{(i)})} \cdot \left(1 - \frac{r_{lapser}^{(i)}(T_{max})}{r_{acceptant}^{(i)}(T_{max})}\right). \tag{21}$$