Digital insurance and long term risk

Chaire d'Excellence

(D)IGITAL (I)NSURANCE (A)ND (LO)N(G)-TERM RISKS

Chaire de Recherche DIALOG - 2020-2025 - CNP Assurances

MLISTRAL Seminar

Including Customer Lifetime Value in tree-based lapse management strategy

CIRM, 27th September 2022

Mathias VALLA - UCBL1/KU Leuven

http://chaire-dialog.fr

Here are the 3 chaire's research axes :

AI and risk dynamic:
How to include time considerations when modeling risk factors ?

AI and interpretability:
Risk factor analysis and interpretability: how to use it from the insurer's PoV ?

AI and specific data structures:
How to handle outliers, missing or incomplete values, longitudinal data...

Summary

Research question, data and framework

Lapse management

We want to be able:

- to define what a lapse management strategy is
- to target policyholders who should be targetted by a given LMS
- to measure the gain generated by a given LMS
- to optimize a given LMS
- to bring complex tree based models to the actuarial literature

Definition:

A T-years lapse management strategy is modeled by the offer of an incentive δ_i to subject i if he/she is targetted. The incentive is expressed as a percentage of its face amount and is accepted with probability γ_i. 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 its probability of being active at year $t \in [0, T]$ is denoted $r_{i,t}^{\text{acceptant}}$. Conversely, a subject who refuse the incentive and prefers to lapse will be considered as a "lapser" and its probability of being active at year t is denoted $r_{i,t}^{\text{lapser}}$. A lapse management strategy is uniquely defined by the parameters $(\boldsymbol{p}, \boldsymbol{\delta}, \boldsymbol{\gamma}, c, T)$

Disclaimer: For privacy reasons:

- all the data, statistics, product names and perimeters presented in this paper have been either anonymised or modified.
- All analysis, disussions and conclusions remain unchanged.

- Real-world french life insurance portfolio
- From 1998 to today (almost!)
- 249k unique policies
- 235k unique policyholders
- 43 covariates

- Originaly we have one row for every policyholder's movement: (payments, lapses, fees, profit sharing, claims...)
- Here we decided to keep only the most recent information for every policyholder
- 1 row for every (policyholder/policy) pair

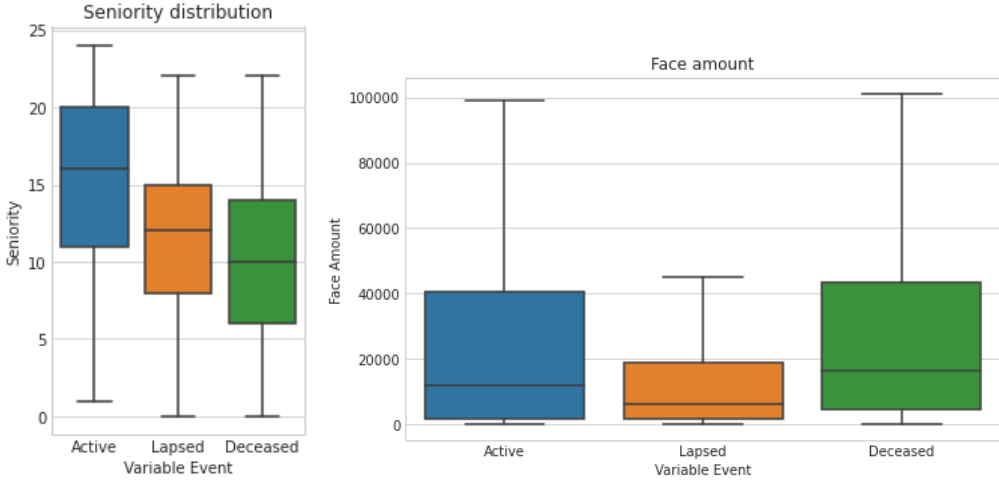


FIGURE: Seniorities and face amounts distributions

$$^{F}CLV_i\left(\boldsymbol{p}_i, \boldsymbol{F}_i, \boldsymbol{r}_i, \boldsymbol{d}\right) = \sum_{t=0}^{T} \frac{p_{i,t} F_{i,t} r_{i,t}}{\left(1+d_t\right)^t}.$$

From which we derive:

- the control portfolio future value ($^{F}CPV(T)$)
  $\Rightarrow$ The hypothetical value of the portfolio, considering that:
  - every subject that did not lapse up to time $t = 0$ has a vector of retention probabilities of $\boldsymbol{r}_i^{\text{acceptant}}$;
  - every subject that lapsed before time $t = 0$ has a vector of retention probabilities of $\boldsymbol{r}_i^{\text{lapser}}$
- the lapse-managed portfolio future value ($^{F}\text{LMPV}(\boldsymbol{\delta}, \gamma, c, T)$)
  $\Rightarrow$ The hypothetical value of the portfolio after a given LMS is applied, considering that:
  - every targeted subject has a γ probability of accepting the incentive and behave with a vector of retention probabilities of $\boldsymbol{r}_i^{\text{acceptant}}$;
  - every targeted subject has a $(1-\gamma)$ probability of refusing the incentive and behave with a vector of retention probabilities of $\boldsymbol{r}_i^{\text{lapser}}$;
  - every non-targetted subject that did not lapse has a vector of retention probabilities of $\boldsymbol{r}_i^{\text{acceptant}}$;
  - every non-targetted subject that lapsed has a vector of retention probabilities of $\boldsymbol{r}_i^{\text{lapser}}$

$$
\begin{aligned}
{}^{F}CPV = & \sum_{i=1}^{n} {}^{F}CLV\left(p_i, F_i, \boldsymbol{r}_i^{\text{acceptant}}, d, T\right) \cdot \mathbf{1}\left(y_i = 0\right) \\
& + \sum_{i=1}^{n} {}^{F}CLV\left(p_i, F_i, \boldsymbol{r}_i^{\text{lapser}}, d, T\right) \cdot \mathbf{1}\left(y_i = 1\right)
\end{aligned}
$$

In the following section, we consider a simplified version of this framework by assuming that:

- $p_{i,t}, F_{i,t}$, and d_t remain constant across time, and denoted p_i, F_i and d hereafter,
- with F_i being the most recent face amount observed for subject i,
- we set γ_i and δ_i to be the same for all subjects and denoted as γ and δ hereafter.

$$
\begin{aligned}
{}^{F}LMPV(\boldsymbol{\delta}, \gamma, c, T) = & \sum_{i=1}^{n} {}^{F}CLV\left(\boldsymbol{p}, F_i, \boldsymbol{r}_i^{\text{acceptant}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 0, \hat{y}_i = 0\right) \\
& + \sum_{i=1}^{n} {}^{F}CLV\left(\boldsymbol{p}, F_i, \boldsymbol{r}_i^{\text{lapser}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 1, \hat{y}_i = 0\right) \\
& + \sum_{i=1}^{n} {}^{F}CLV\left(\boldsymbol{p} - \boldsymbol{\delta}, F_i, \boldsymbol{r}_i^{\text{acceptant}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 0, \hat{y}_i = 1\right) \\
& + \gamma \cdot \sum_{i=1}^{n} {}^{F}CLV\left(\boldsymbol{p} - \boldsymbol{\delta}, F_i, \boldsymbol{r}_i^{\text{acceptant}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 1, \hat{y}_i = 1\right) \\
& + (1 - \gamma) \cdot \sum_{i=1}^{n} {}^{F}CLV\left(\boldsymbol{p}, F_i, \boldsymbol{r}_i^{\text{lapser}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 1, \hat{y}_i = 1\right) \\
& - c(N(-, 1))
\end{aligned}
$$

Remark: It is important to note that this does not reflect the actual future value of the portfolio - as the future CLV of lapsers should be 0 - but rather its hypothetical expected future value given the nature (lapser or not) of every subject but not their real states (actually lapsed or not).

We can now derive:

$$RG(\delta, \gamma, c, T) = {}^{F}\mathrm{LMPV}(\boldsymbol{\delta}, \gamma, c, T) - {}^{F}CPV(T),$$

that can be simplified as follows:

$$\begin{aligned} RG(\delta, \gamma, c, T) = \sum_{i=1}^{n} \Bigg[& \gamma \Big[{}^{F}CLV\left(\boldsymbol{p} - \boldsymbol{\delta}, F_i, \boldsymbol{r}_i^{\text{acceptant}}, \boldsymbol{d}, T\right) \\ & - {}^{F}CLV\left(\boldsymbol{p}, F_i, \boldsymbol{r}_i^{\text{lapser}}, \boldsymbol{d}, T\right) \Big] \cdot \mathbf{1}\left(y_i = 1, \hat{y}_i = 1\right) \\ & - {}^{F}CLV\left(\boldsymbol{\delta}, F_i, \boldsymbol{r}_{\text{stay}}, \boldsymbol{d}, T\right) \cdot \mathbf{1}\left(y_i = 0, \hat{y}_i = 1\right) \Bigg] - c(N(-, 1)) \end{aligned}$$

That allows ud to define:

$$z_i = \begin{cases} -\,{}^{F}CLV_i\left(\boldsymbol{\delta}_j, F_i, \boldsymbol{r}_i^{\text{acceptant}}, d, T\right) - c & \text{if } y_i = 0 \\ \gamma \cdot \Big[\,{}^{F}CLV_i\left(\boldsymbol{p} - \delta, F_i, \boldsymbol{r}_i^{\text{acceptant}}, d, T\right) & \\ \qquad -\,{}^{F}CLV_i\left(p, F_i, \boldsymbol{r}_i^{\text{lapser}}, d, T\right)\Big] - c & \text{if } y_i = 1 \end{cases}$$

$\Rightarrow$ That represents the expected profit or loss that would results from targetting policyholder i with a given lapse management strategy.
And:

$$\tilde{y}_i = \begin{cases} 1 \text{ if } \hat{z} > 0 \\ 0 \text{ if } \hat{z} \leq 0 \end{cases},$$

$\Rightarrow$ indicating wheter or not targetting policyholder i is expected to generate a profit for the insurer.

We define the lapse management strategy modelisation as a 2-steps framework:

- $r^{\text{acceptant}}$ and r^{lapser} modelisation: $\Rightarrow$ aka the survival part
- Computation of $\tilde{y}_i$ and classification, with RG as the evaluation metric: $\Rightarrow$ aka the LMS part

In parallel, we run a classification on y_i with accuracy as the evaluation metric, for comparison sake

What's new ?

- individualized $r^{\text{acceptant}}$ and r^{lapser} modelisation
- $r^{\text{acceptant}}$ and r^{lapser} modelisation that take the risk of death into consideration throught a competing risk survival models
- usage of survival tree-based models for actuarial purpose
- keeping the problem as a classification one

MODELISATION

- Survival part
- LMS part

We are aware that the context of our modelization requires competing risk setting.

Several regression models exist to estimate the global hazard and the hazard of one risk in such settings: cause-specific and subdistribution modelizations.

They account for competing risks differently, obtaining different hazard functions and thus have distinct advantages, drawbacks and interpretations.

After discussions, the simplicity of a cause-specific approach and the fact that it can be adapted to any survival method including tree-based ones, oriented our choice towards it.

In Cause-specific regression, each cause-specific hazard is estimated separately, in our case, the cause-specific hazards of lapse and death, by considering all subjects that experienced the competing event as censored.
The cause-specific hazard rates regarding the j-th risk ($j \in [1, \dots J]$) are defined as:

$$\lambda_{T,j}(t) = \lim_{dt \to 0} \frac{P\left(t \leq T < t + dt, J_T = j \mid T \geq t\right)}{dt}$$

We can recover the global hazard rate as $\lambda_{T,1}(t) + \cdots + \lambda_{T,J}(t) = \lambda_T(t)$, and derive the global survival distribution of T as

$$\begin{aligned} P(T > t) &= 1 - F_T(t) = S_T(t) \\ &= \exp\left(-\int_0^t \left(\lambda_{T,1}(s) + \cdots + \lambda_{T,J}(s)\right) ds\right) \end{aligned}$$

This approach aims at analysing the cause-specific "distribution" function:
$F_{T,j}(t) = P\left(T \leq t, J_T = j\right)$. In practice, it is called the Cumulative Incidence Function (CIF) for cause j and not a distribution function since $F_{T,j}(t) \to P\left(J_T = j\right) \neq 1$ as $t \to +\infty$.

Subdistribution hazard function/Fine and Gray regression, works by considering a new competing risk process τ. Without loss of generality, let's consider death as our cause of interest:

$$\tau = T \times \mathbb{1}_{J_T=2} + \infty \times \mathbb{1}_{J_T \neq 2}.$$

It has the same as T regarding the risk of death, $P(\tau \leq t) = F_{T,2}(t)$ and a mass point at infinity $1 - F_{T,2}(\infty)$, probability to observe other causes $(J_T \neq 2)$ or not to observe any failure.
In other words, if the previous approach considered every subject that experienced competing events as censored, this approach considers a new and artificial at-risk population. This last consideration is made clear when deriving the hazard rate of τ:

$$\lambda_\tau(t) = \lim_{dt \to 0} \frac{P\left(t \leq T < t + dt, J_T = 2 \mid \{T \geq t\} \cup \{T \leq t, J_T \neq 2\}\right)}{dt}.$$

Finally, we obtain the CIF for the risk of death as:

$$F_{T,2}(t) = 1 - \exp\left(-\int_0^t \lambda_\tau(s)ds\right).$$

$\Rightarrow$ resolves the most important drawback of cause-specific regression, as the coefficients resulting from it do have a direct relationship with the cumulative incidence: estimating the CIF for a specific cause does not depend on the other causes, which makes the interpretation of CIF easier.

We considered:

- Cause-specific Cox proportional hazard model,
- Random survival forest
- Gradient boosting survival model

Survival trees work similarly as regular decision trees, except for the splitting criterion.

In regular decision trees: we choose the split that maximize the within node homogeneity.

In survival trees : we choose the split that maximize the between nodes heterogeneity. The heterogenerity between the two child nodes can be measured with a log rank statistic.

- A random survival forest is a random forest of survival trees.
- A gradient boosting survival model is the equivalent of xgboost with survival trees.

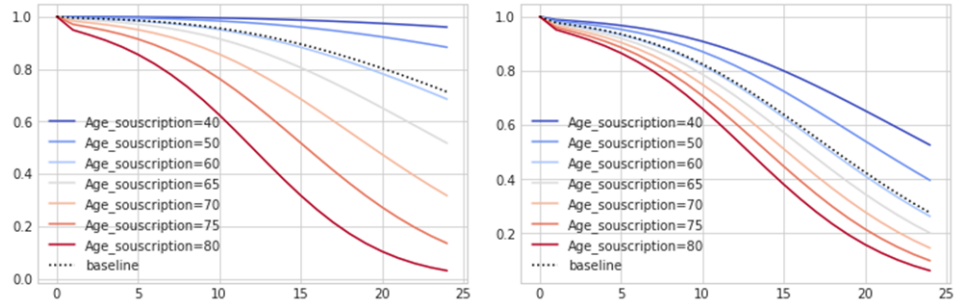


FIGURE: $r^{acceptant}$ and r^{lapser} survival curves at different policyholder ages

Covariate importance:

- Age and gender are common important covariates for the 3 models for $r^{acceptant}$.
- Age and face amount are common important covariates for the 2 tree-based models for r^{lapser}.

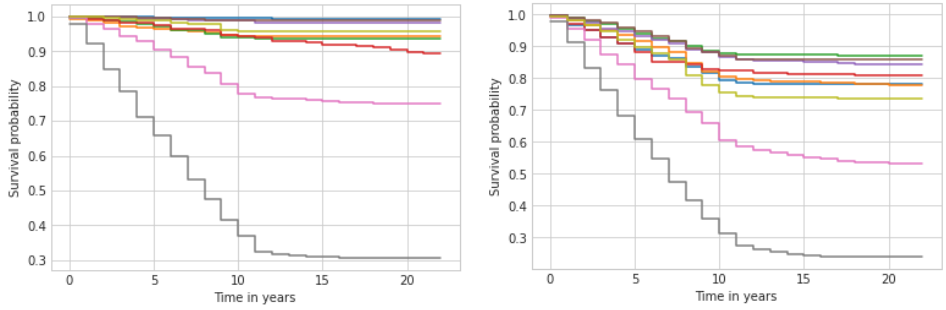


FIGURE: $r^{acceptant}$ and r^{lapser} survival curves for different policyholders

Covariate importance:

- Age and gender are common important covariates for the 3 models for $r^{acceptant}$.
- Age and face amount are common important covariates for the 2 tree-based models for r^{lapser}.

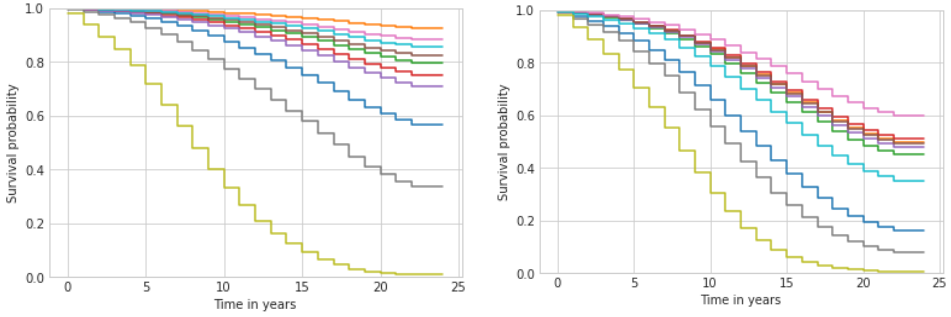


FIGURE: $r^{acceptant}$ and r^{lapser} survival curves for different policyholders

Covariate importance:

- Age and gender are common important covariates for the 3 models for $r^{acceptant}$.
- Age and face amount are common important covariates for the 2 tree-based models for r^{lapser}.

We use the concordance index to compare the results of our considered models:

	Concordance Index	
	r^{lapser}	$r^{acceptant}$
Cox model	69,5%	80,7%
RSF	71,6%	83,7%
GBSM	73,0%	84,1%

- Non-parametric approaches perform better;
- GBSM in particular yields the best results
  $\Rightarrow$ we choose GBSM to model our retention probabilities
- Computation times: tens of seconds for CPH, tens of hours for the tree-based models

We now have all the elements to run LMS models:

- We compute $r^{acceptant}$ and r^{lapser} with GBSM;
- We compute the z_i's and $\tilde{y}_i$'s for every policyholder;
- We can run classification models on y_i and $\tilde{y}_i$
- We can compare the RG's estimated with different models

In order to measure the performance of our framework, from the insurer's PoV, we will compare two frameworks:

Classical framework: Classification on y, with accuracy as evaluation metric

CLV-augmented framework : Classification on $\tilde{y}$, with RG as evaluation metric

Tree based models:

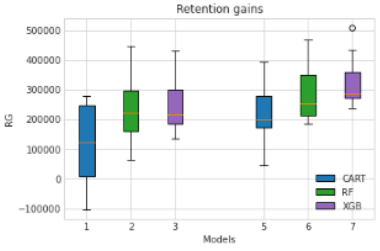
- CART
- RF
- XGBoost

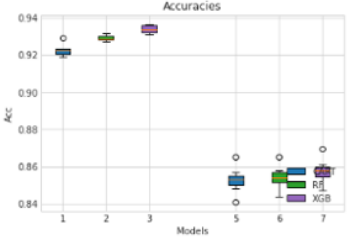
5 scenarios:

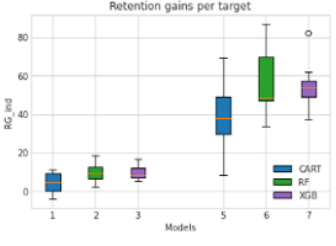
Scenarios	p	δ	γ	c	d	T
1	2,50%	0,04%	25%	10	1,50%	5
4	2,50%	0,04%	5%	10	1,50%	5
13	1,50%	0,20%	20%	10	1,50%	5
21	2,50%	0,08%	10%	10	1,50%	20
30	1,50%	0,20%	20%	100	1,50%	5

Results and discussion

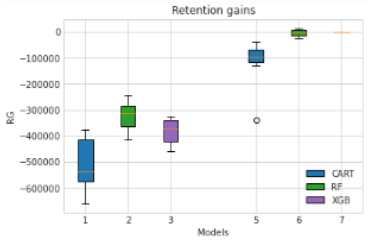
Positive result on y_i and an improved result on $\tilde{y}_i$.

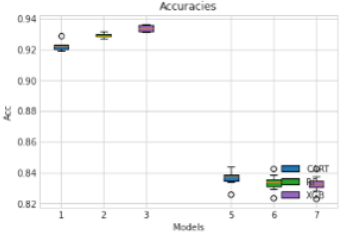


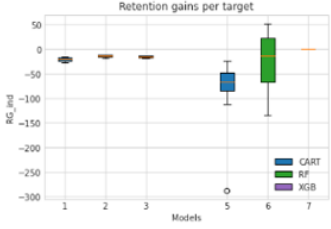




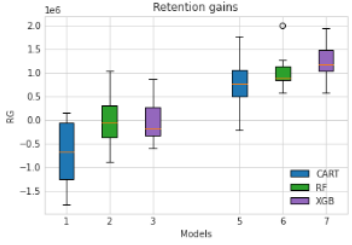
Very negative result on y_i and a loss-limiting result on $\tilde{y}_i$.

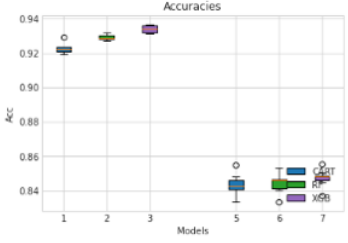


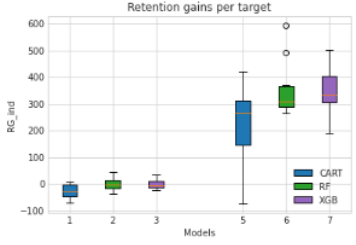




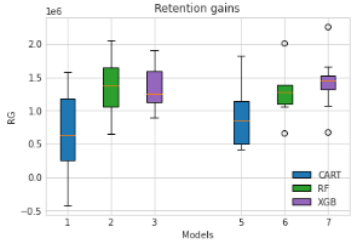
Negative result on y_i and positive one on $\tilde{y}_i$

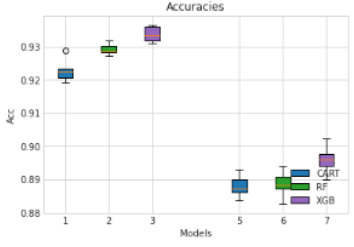


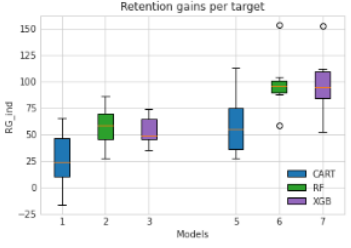




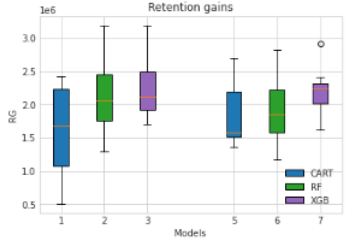
High positive result on y_i slightly improved with $\tilde{y}_i$.

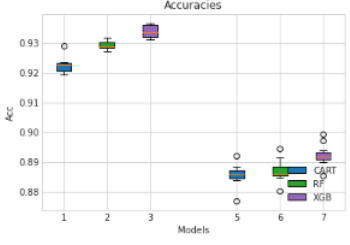


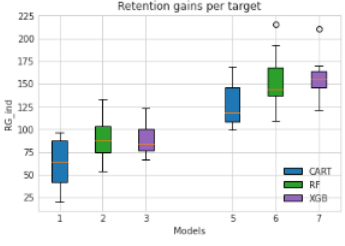




Results on y_i better than results on $\tilde{y}_i$.







We can note that:

- RF and Xgboost perform globally better than CART
- XGB is more consistent and is the best model in most scenarios both with and without the CLV-based measure
- All the realistic ones shows that a classification on $\tilde{y}_i$ produces a targetting that yields better RG than a classification on y_i
- A classification on y_i produces a targetting that yields better accuracies regarding whether a policyholder will churn than a classification on $\tilde{y}_i$
- The model shows to yield very high improvement when classification on y_i gives negative RG.
- The model can turn a negative RG into a positive one

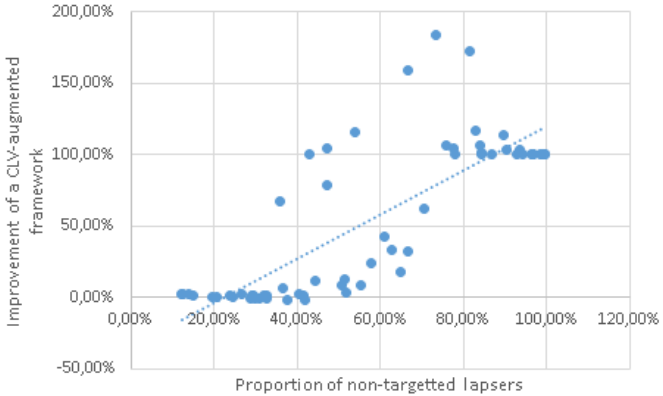
The average observed improvement of a CLV-augmented framework over the classical lapse one is 57,9%[a]. If we weight these results by the expected RGs, the average improvement is still 31,7%.

[a] Using XGBoost

The modelization:

- showed to be loss-limiting (Strategy n°4)
- showed that the improvement of a lapse management strategy including CLV grows with the proportion of lapsers with a negative CLV

- showed to be loss-limiting (Strategy n°4)
- showed that the improvement of a lapse management strategy including CLV grows with the proportion of lapsers with a negative CLV



This framework can be used in several ways by the insurer. It can help:

- understand what differentiate a subject for which $y_i = 1$ and $\tilde{y}_i = 0$ from the others
- interpret the results at an individualized level
- compares future hypothetical lapse management strategies in order to chose the best one
- answer questions like:
  - For which incentive δ the retention strategy becomes profitable ?
  - For which acceptance probability γ the retention strategy becomes profitable ?
  - at which horizon T, the retention strategy become profitable ? In other words, when can the insurer expect a return on investement ?
- Measure the expected gain of a real retention campaign from the past, at various time horizons

This framework can be used in several ways by the insurer.

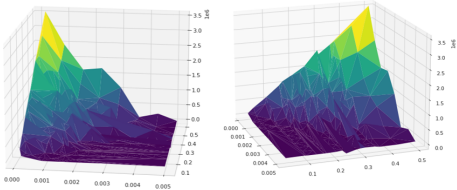


FIGURE: 3d plot (δ, γ, RG)

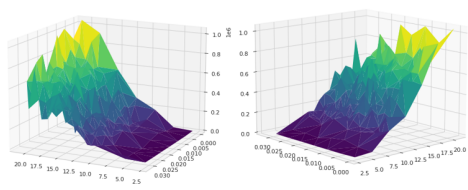


FIGURE: 3d plot (d, T, RG)

PERSPECTIVES AND CONCLUSIONS

This framework has some obvious limitations:

- following one single fixed strategy for every policyholder is not realistic
- policyholder's behavior is dynamical, $r^{acceptant}$ could include some lapse risk
- gamma and delta interdependency could be taken into account
- subdistribution competing risk modeling - using Fine and Gray - was not carried out and could benefit our framework
- could be improved by the use of longitudinal data that would yield time-dynamic results
  $\Rightarrow$ future work

- trying to predict whether targetting a policyholder would benefit the insurer or trying to predict whether he/she will lapse are two very different things !
- including CLV in lapse management strategy can largely benefit an insurer's decision making ability regarding lapse management strategy
- survival tree-based models can outperform parametric approaches in such actuarial contexts
- our CLV-based framework lead to increased predicted gains for any realistic scenario and acted as a loss limiting targetting approach regardless of the retention strategy
- our modelisation can give insights to the life insurer regarding commercial and strategical decision making

Work(s) conducted within the Research Chair DIALog under the aegis of the Risk Foundation, an initiative by CNP Assurances. The authors would like to express their very great appreciation to Marie Hyvernaud and Stéphanie Dosseh for their valuable and constructive suggestions during the development of this research work. Special thanks should be given to Marie Hyvernaud for her contribution regarding code writting.

Thanks for your attention

Any questions ?