Modeling Dynamic Policyholder Behavior through Machine Learning Techniques

Marco Aleandri
Ph.D. Student in Actuarial Science
Dept. of Statistical Sciences
Univ. "La Sapienza", Rome
Email: marco.aleandri@uniroma1.it

Abstract

In this paper, we focus our attention on the most relevant non-market risk in insurance, that is, lapse risk. It is basically linked to the behavior of policyholders facing various situations such as aging, actual economic condition, contract features, and so on. At the same time, policyholder's retention directly impacts the profitability of the product itself, thus the profitability of the company as a whole.

Through the first part of our analysis, we will recognize some relevant lapse risk factors from a specific dataset including a number of explanatory variables. More importantly, the predictive results from the traditional logistic regression will be compared to those of a bagging classification tree, in order to select the most powerful model.

Furthermore, the goal of the second part of the analysis is the valuation of the impact on the profitability of a specific insurance product based on the predicted lapse rates. We will observe how significant the policyholder behavior can be as soon as it is introduced within the profit valuation in a dynamic fashion.

Keywords: policyholder behavior, lapse risk, machine learning, logistic regression, bagging classification tree, TVOG

1. Introduction

Following Solvency II implementation, the traditional actuarial methods are being replaced, little by little, by more complex and structured models to estimate the economic value (in his several forms) of insurance companies. So far, most effort focused on market risks, which represent the largest component of the capital requirement for life insurers under Solvency II. However, as specified by the Directive itself, non-market risks should be evaluated and monitored as well, especially the so-called policyholder behavior. Typically, policyholders are given a number of embedded options within their insurance contracts, and predicting the exercise's likelihood is crucial to forecast the portfolio profitability.

Indeed, policyholder behavior refers to the policyholder's tendency to exercise any of the options embedded in its insurance contract. They include surrender option, guaranteed annuity option (GAO), dynamic premium increase option, product switching option, fund switching option, paid-up option, and so on. For several reasons, the surrender option leads the greatest impact in the portfolio. Effectively, it is embedded in almost all the insurance products. To some extent, it guarantees the fairness of the contract by allowing the policyholder to receive back its reserve (net of some penalty). As opposed to other aforementioned options, the surrender option can be generally exercised at any time prior to the maturity, just like an American option, so that the insurer is exposed on a continuous basis.

As a result, Solvency II recognizes that policyholder behavior is a relevant source of risk. Specifically

> *Assumptions about the likelihood that policyholders will exercise contractual options should be based on analysis of past policyholder behavior. The analysis should take into account the following:*
>
> *a) how beneficial the exercise of the options was or would have been to the policyholders under past circumstances (whether the option is out of or barely in the money or is in the money),*
>
> *b) the influence of past economic conditions,*
>
> *c) the impact of past management actions,*
>
> *d) where relevant, how past projections compared to the actual outcome,*

> e) *any other circumstances that are likely to influence a decision whether to exercise the option.*

Solvency II introduces the lapse risk valuation in the Standard Formula as well as in any internal model. The paragraph SCR.7.44 in [9] gives a simple definition of such a risk:

> *Lapse risk is the risk of loss or change in liabilities due to a change in the expected exercise rates of policyholder options.*

In this paper, we will focus on the possible risk of loss due to unanticipated policyholder behavior for full (i.e. not partial) surrender. The adjective "unanticipated" refers to the exclusion of any dynamic policyholder behavior assumption from both pricing and reserving, just like it is still common in the actuarial practice. However, as pointed out at the paragraph TP.2.4 in [9] about the valuation technique of the best estimate of liabilities,

> *Cash-flow characteristics that should, in principle and where relevant, be taken into consideration in the application of the valuation technique include the following:*
>
> a) *Uncertainty in the timing, frequency and severity of claim events.*
>
> b) *Uncertainty in claims amounts and the period needed to settle claims.*
>
> c) *Uncertainty in the amount of expenses.*
>
> d) *Uncertainty in the value of an index/market values used to determine claim amounts.*
>
> e) *Uncertainty in both entity and portfolio-specific factors such as legal, social, or economic environmental factors, where practicable.* [...]
>
> ***f) Uncertainty in policyholder behavior.***
>
> g) *Path dependency.* [...]
>
> h) *Interdependency between two or more causes of uncertainty.* [...]

Anyway, we assume that the insurance company will allocate reserve based on the traditional actuarial criteria, which also exclude lapse rate. Thus, while dynamic policyholder behavior will be introduced in the profit valuation of a typical insurance product, we will consider neither surrender cash-flows nor policyholder behavior within the reserve calculation.

After a broad review of the past studies about policyholder behavior's risk factors in Section 2, we will start our analysis. We can distinguish three steps:

1. data preparation (Section 6)
2. lapse rate prediction (Sections 3, 4, and 6)
3. profit analysis (Sections 5-7).

So in short, we will firstly predict lapse rates in a dynamical way (that is, different lapse rates in different scenario simulations), and then use them as a contingency in a typical profit test for a specific insurance contract. Notice that policyholder behavior occurs *after* economic scenario simulation, because the former depends on the latter. In other terms, policyholder behavior will be treated as a deterministic function - represented by the machine learning algorithm - of the scenario simulation, among others. All in all, we will not build a stochastic model for the policyholder behavior, so that no further computational burden will be caused beyond that coming from the economic scenario simulation.

2. Drivers of the policyholder behavior

Because surrender activity can be so damaging to a single company or to the life insurance industry if it occurs *en masse*, research on widespread surrender activity and its possible determinants is especially important. In the last century, several researches and papers have been published about the very sources of the policyholder behavior. They referred on the financial and insurance market of various countries worldwide. Although this paper does not aim to detect the global sources of policyholder behavior, it is worth recalling the most relevant results about them, covering almost one century. This will provide us with good foundations to start analyzing a lapse rate database of an insurance company.

An extensive description of the lapse rate research's early stage can be found in [22], which focuses on North American markets. By earlier twenties, almost one hundred years ago, some results had already demonstrated correlation between lapse rate and economic conditions, curiously right before the great depression after 1929. Nonetheless, researchers were well aware that market variables could not completely explain the effective duration of any product. As a consequence, several studies started focusing on policyholder-related variables, finding out that lapse could be correlated to income, occupation, sex, age, family, premium frequency and amount, and others. Another relevant finding of those years regards the effects of global economic distress on lapse rates. Briefly, each policyholder seems to have a sort of tolerance threshold depending on its risk-propensity. It can be largely irrational, but it is also related to the actual economic condition: in time of economic distress, it is likely that most of policyholders feel beyond their threshold, which leads them to close their contracts. In other terms, lapse rate cannot be represented as a regular function of a global index market.

Without a doubt, three of the most comprehensive studies in that period were [5], [16], and [17]. Among others, they suggest that an insurance company can limit lapse from contract's inception by selecting quality business, recognizable from some objective indications of good persistency, especially age at issue, premium frequency, and plan. Surprisingly, the author of [5] concludes that the agent's ability in picking quality business can be even more crucial than the actual economic condition itself.

Later, in the 1960s and 1970s, a couple of interesting empirical contributions have been produced by the Institute and Faculty of Actuaries. Both [7] and [20] are based on Scottish data, and the first one is a sort of update of the second one. To some extent, we can say that [20] is based on data of 1960s as well as [7] is based on data of 1970s (this is the reason why both of them have been published at the end of the respective decade). As the titles suggest, such studies are exploratory in nature, but consider some of the variables which, on the base of the aforementioned sources, can drive surrender decisions. Sex, age at entry, occupation, purpose of assurance, calendar year (as a representation of variable economic conditions), sum assured, premium paying term, and premium payment frequency, distribution channel, policy duration - all of them have been included in the analysis, but only duration and age at entry showed a significant correlation with the lapse rate.

All in all, by the end of the 1970s, lapse rates were fairly steady, with increases occurring during recessions, and decreases occurring during expansions. However, in the late 1970s, some important aspects started changing. Markets began to experience the highest increase in interest rates and volatility ever, while new, more complex insurance products were introduced.

The contemporary improvement in financial literacy among policyholders led many of them to surrender their policies for more rational reasons such as interest rate arbitrage, preference for pure financial products, awareness about their own risk-propensity, and so on. Surrender was no longer the natural, though irrational, response to the need of money during time of distress (the so-called Emergency Fund Hypothesis, for instance, in [23]). In some cases, it turned to be the result of a precise, financial-oriented decision of the policyholder (the so-called Interest Rate Hypothesis, for instance, in [23]). Of course, it could only worsen the position of intermediaries. Insurers were forced to liquidate bonds to meet surrender requests at precisely the time when the values of bond portfolios were depressed by high interest rates.

The increased volatility in economic conditions and financial markets made the correlation between policyholder behavior and macroeconomic variables much more interesting than policy features. A number of studies published in the last thirty years focuses on the macroeconomic determinants of global lapse rates in various countries. Most of them succeed in proving the Emergency Fund Hypothesis, but not the Interest Rate Hypothesis. The authors of [8] explored the relationships between variables like interest rates and unemployment rate with surrender activity in the UK endowment life insurance market from the period 1952-1985. Something similar has been analyzed by the author of [19] using US and Canadian data of whole life policies from the period 1955-1979. In this study, the unemployment rate has a significantly positive effect on lapse rate, while policyholder's income has a significantly negative effect, while no significant relationship with interest rates was found. Further, the author of [10] concludes that the unemployment rate is the most significant variable in predicting surrender activity for universal life policies in US from the period 1982-1986. The authors of [15] used a dataset provided by the American Council of Life Insurance (ACLI) from the period 1951-1998 to confirm the strong correlation of the surrender activity with the unemployment rate. Nonetheless, they find a strong impact of the interest rate as well. In other words, [15] supports both the Emergency Fund Hypothesis and Interest Rate Hypothesis.

In more recent years, other researchers focused on some European and Asian countries, both for empirical study and regression-based prediction of lapse rates. The Italian insurance market of savings products has been analyzed in [6] by using surrender experience data of a large Italian bancassurer from the period 1991-2007. Explanatory variables included product type, calendar year, duration, and inception year. In [14], the author used logistic regression to model lapse rate of Korean interest indexed annuities. Explanatory variables included the difference between reference market rates and product crediting rates, policy duration, unemployment rate, economy growth rate, and some seasonal effects. One of the most comprehensive surrender analysis is represented by [13], which focuses on the German market. The study distinguishes five product categories (traditional endowment policies, annuities and long-term health contracts, term life insurance, group business, and unit-linked contracts), and includes both macroeconomic explanatory variables (e.g. current market yield, DAX performance, gross domestic product, and unemployment rate) and company-specific explanatory variables (e.g. company ages, distribution channel, company legal form, and company size). One of the most recent study is the working paper [11] on Taiwan data from the decade 1999-2009. Just like in [13], the paper considers a number of macroeconomic variables and company-specific variables, i.e. business line, premium income, company age, return-on-asset, domestic/foreign company, unemployment rate, home-ownership ratio, short-term interest rate, and economic growth rate.

Beyond the huge amount of empirical studies on surrender rates (only partially described so far in this section) trying to detect the most relevant predictors, a remarkable number of studies about surrender option's valuation also exists. Such papers deal with the surrender activity of policyholders as the exercise of an American option embedded in the insurance contract, and valuate it as a stand-alone option by using either analytic or numeric models. Given that our goal is not product pricing, the topic is beyond the scope of this paper, and it will not be analyzed further.

Obviously, persistency is a crucial factor in the pure financial market as well. For instance, the author of [25] values mortgage-backed securities in USA assuming that part of the prepayment decision of mortgage holders is rational and based on current economic conditions, while the remaining part is interpreted as irrational. Again, this is beyond the scope of this paper, and the topic will be not analyzed further.

To sum up, three categories of policyholder behavior's drivers can be empirically distinguished: macroeconomic factors, company-specific factors, and policy-specific factors. Of course, other factors could relate to life insurance surrender activity, although they might not have mentioned in past studies yet. At the same time, the set of relevant explanatory variables could change in time, for example, as target clients, product nature, or insurance purpose change. As a consequence, looking for a unique, stable set of explanatory variables seems to be the wrong way to go ahead. Therefore, we will focus on the dataset provided by a single insurance company, where most of the explanatory variables are policy-specific, while only one macroeconomic variable is included (obviously, there is no reason to include company-specific variables, given that data come from the same company).

3. A traditional approach: logistic regression (LR)

Regression-based models are by far the most used tools to fit and predict probabilities, including the surrender rates of the aforementioned papers. [6], [13], and [14] are only some examples. Among such models, the most commonly used is logistic regression, so we will start our analysis from it. To give a short, theoretical explanation to the model, we can simply start from the linear regression equation:

$$y = \alpha + \sum_{k=1}^{n} \beta_k x_k =: \mathbf{x}\beta^{\mathbf{T}} \tag{1}$$

where x_k denotes the k^{th} risk factor and β_k the related parameter. When it comes with predicting probabilities, that is, values in the interval $(0,1)$, the (1) faces its major drawback: it returns values in the entire domain $\mathbb{R}$. The solution is represented by the choice of a proper link function that takes values in $\mathbb{R}$ and returns values in $(0,1)$. For example

$$\begin{aligned} g \,:\, \mathbb{R} \;&\longrightarrow\; (0,1) && (2) \\ z \;&\longrightarrow\; \frac{1}{1+e^{-z}} && (3) \end{aligned}$$

which is the so-called logistic function. Also, the requirement of a non-decreasing function for cumulative distribution function is satisfied. As a consequence, the predicted probability equals

$$p = g(y) = g(\mathbf{x}\beta^{\mathbf{T}}) = \frac{1}{1+e^{-\mathbf{x}\beta^{\mathbf{T}}}} = \frac{1}{1+e^{-(\alpha+\sum_{k=1}^{n}\beta_k x_k)}} \tag{4}$$

or equivalently

$$\mathbf{x}\beta^{\mathbf{T}} = \ln\left(\frac{p}{1-p}\right) =: logit(p). \tag{5}$$

The choice of the logistic function comes from the best practice, but other link functions are available.

The framework is quite similar to the linear regression case, but it is not exactly the same. This is the reason why we cannot use the ordinary least square method to estimate the model parameters $\beta_1, \ldots, \beta_n$. Rather, we should use the maximum likelihood estimation. By definition, the likelihood function for N observations is

$$L(\mathbf{X}, \beta) = \prod_{i=1}^{N} g(\mathbf{x_i}\beta^{\mathbf{T}})^{y_i}[1 - g(\mathbf{x_i}\beta^{\mathbf{T}})]^{1-y_i} \tag{6}$$

so the log-likelihood function is

$$\begin{aligned} l(\mathbf{X}, \beta) &:= \ln L(\mathbf{X}, \beta) = \sum_{i=1}^{N} \Big\{ y_i \ln[g(\mathbf{x_i}\beta^{T})] + (1 - y_i) \ln[1 - g(\mathbf{x_i}\beta^{T})] \Big\} = \\ &= \sum_{i=1}^{N} \left\{ y_i \ln \left[\frac{1}{1 + e^{-\mathbf{x_i}\beta^{\mathbf{T}}}} \right] + (1 - y_i) \ln \left[1 - \frac{1}{1 + e^{-\mathbf{x_i}\beta^{\mathbf{T}}}} \right] \right\} = \\ &= \sum_{i=1}^{N} \Big\{ y_i(\mathbf{x_i}\beta^{\mathbf{T}}) - \ln[1 + e^{\mathbf{x_i}\beta^{\mathbf{T}}}] \Big\}. \end{aligned} \tag{7}$$

The maximum likelihood estimation for the vector of parameters β results from the maximization of $l(\mathbf{X}, \beta)$, or equivalently from the solution of the following system of equations:

$$\frac{\partial l}{\partial \beta_i} = 0, \quad \forall i = 1, \ldots, n. \tag{8}$$

Such a solution is indeed the estimation $\hat{\beta}$, which can be used to estimate p:

$$\hat{p}_i := g(\mathbf{x_i}\hat{\beta}^{\mathbf{T}}) = \frac{1}{1 + e^{-\mathbf{x_i}\hat{\beta}^{\mathbf{T}}}} = \frac{1}{1 + e^{-(\alpha + \sum_{k=1}^{n} \hat{\beta}_k x_{ik})}}. \tag{9}$$

At the beginning of the section, we implicitly assumed to know the explanatory variables $x_1, \ldots, x_n$. Of course, we know the explanatory variables in the dataset, but how should we select them as $x_1, \ldots, x_n$? Because of multicollinearity among potential explanatory variables, we cannot simply run the logistic regression on all of them, and then select only the most significant ones based on their p-values. Rather, we should somehow select different sets of explanatory variables and run the related logistic regressions: the model with the highest R^2 will be selected. The different sets of explanatory variables depend on the algorithm used to select them. There are mainly three popular iterative search algorithms.

In *forward selection*, we start with no predictors, and then add them one by one. Each added predictor is that (among all predictors) that has the larges contribution to R^2 on top of the predictors that are already in it. The algorithm stops when the contribution of additional predictors is not statistically significant.

In *backward selection*, we start with all predictors, and then eliminate the least useful one at each step according to statistical significance. The algorithm stops when all the remaining predictors have significant contributions.

Finally, *stepwise selection* is like forward selection except that at each step we consider dropping predictors that are not statistically significant, as in backward selection. As we will discuss in Section 6, our data will be regressed through stepwise selection.

4. A machine learning approach: bagging classification tree (BCT)

Decision trees were used as a machine learning tool in [3] for the first time to segment a population by splitting up the dataset through binary rules. The algorithm is now referred to as “classification and regression tree” (CART). Since our goal is the binary classification lapse vs. non-lapse, we will need the classification tree’s version of the algorithm (by contrast, if the independent variable were numeric, we would consider regression trees).

The classification tree ’s algorithm is based on recursive partitioning. It divides up the multidimensional space (that is, the dataset) of the explanatory variables into non-overlapping multidimensional rectangles. This division is accomplished recursively, i.e. operating on the results of the prior divisions. First, one of the explanatory variables is selected, say x_k (the first *node* of the tree, so-called *root*), and a value of x_k, say s_k, is chosen to split the n-dimensional space into two parts: one part contains all the points with $x_k \leq s_k$, while the other with all the points with $x_k > s_k$. Let’s consider on of the two sub-datasets: it could be either *pure* (i.e. it contains only records sharing the same value of the independent variable) or *impure*. In the first case, no further split is possible, so the sub-dataset will represent a *leaf* of the tree. In second case, other splits are possible, so the sub-datasets will represent another node of the tree. Unless both of the sub-datasets generated by the root node are pure, one of them (at least) will be divided in a similar manner by choosing a variable again (it could be x_k or another variable) and a split value for the variable. For example, if both of the sub-datasets are impure, the initial dataset is partitioned in four regions. Again, each of them will turn to be a leaf or a new node, and so on. This process is continued so that we get smaller and smaller rectangular regions. Sooner or later, we will have divided the whole space up into pure rectangles (of course, this is not always possible, as there may be records that belong to different classes but have exactly the same values for everyone of the predictor variables). In our case, the dataset will be partitioned into sub-datasets which contain either lapsed policyholders or retained policyholders. In fact, the classification tree resulting from recursive partitioning is a *pure* tree: lapses and non-lapses are perfectly separated (see Figure 9 for an example of classification tree on our dataset).

The main problem of recursive partitioning is the choice of the splitting rule node by node, that is, the choice of x_k and s_k at each step of the algorithm. Assume to define an *impurity function* $i(A)$ as an impurity measure of some rectangle A, or its related node. A specific splitting rule on A results in two sub-rectangles A_L and A_R, which are generally impure, that is, $i(A_L)$ and $i(A_R)$ are both nonzero. Intuitively, we want to choose the splitting rule in order to minimize some combination of $i(A_L)$ and $i(A_R)$. The most natural choice is the function

$$I(A_L, A_R) := \frac{|A_L|}{|A|} i(A_L) + \frac{|A_R|}{|A|} i(A_R) \tag{10}$$

which is the average of the two impurity measures, weighted by the number of observations in each rectangle. By comparing the reduction in $I(A_L, A_R)$ across all possible splits in all possible predictors, the next split is chosen.

What about the impurity function i? In our application and in most of them, one uses the *Gini index* (as defined in [24]):

$$i(A) := 1 - p_L^2(A) - p_{nL}^2(A) \tag{11}$$

where p_L (respectively: p_{nL}) is the proportion of records in rectangle A that did lapse (respectively: did not lapse). However, other impurity measures are also widely used, for example the *entropy index* (as defined in [24]):

$$E(A) := -p_L(A) \log_2(p_L) - p_{nL}(A) \log_2(p_{nL}). \tag{12}$$

All in all, so far the algorithm is quite intuitive as well as its application in classifying new records. For instance a new observation, whose explanatory values are known, will be dropped down the tree until it reaches a leaf. Since the tree is pure, the leaf will include either lapses or non-lapses, so the new observation will be simply classified on the base of the specific leaf's classification.

To evaluate the predictive performance of a classification method, whether classification tree or logistic regression, we generally use three measures:

$$\begin{aligned}
\text{sensitivity} &:= \frac{TP}{FN+TP} \qquad (13)\\
\text{specificity} &:= \frac{TN}{FP+TN} \qquad (14)\\
\text{misclassification error} &:= \frac{FN+FP}{n} \qquad (15)
\end{aligned}$$

where, in our case, TP (*true positives*) is the number of lapses correctly classified as lapses, FN (*false negatives*) is the number of lapses incorrectly classified as non-lapses, TN (*true negatives*) is the number of non-lapses correctly classified as non-lapses, FP (*false positives*) is the number of non-lapses incorrectly classified as lapses, and n is the total number of records in the dataset (i.e. $TP + FN + TN + FP$). Sensitivity and specificity are especially used to build the *Receiver operating characteristic* curve (or ROC curve), while the misclassification error is an overall measure of the predictive performance (both of them will be used in Section 6 to select the best algorithm between logistic regression and bagging classification tree).

By definition, recursive partitioning produces trees which classify the records without errors, i.e. zero misclassification error). Actually, we used a dataset to *train* the classification tree, which perfectly predict lapses on that dataset. This is the reason why we call it *training dataset*. But what if we use the same tree on a new dataset, say a *validation dataset*? In general, the predicted values on the validation dataset will result in a positive misclassification error, which is obvious. The misclassification error cannot be zero on datasets other than the training dataset itself. However, there is a major drawback of our classification tree. In fact, a possible comparison between misclassification error in the training dataset and in the validation dataset using the full tree is shown in Figure 1. As it usually happens through the first splits on the validation dataset, the full tree can still guarantee comparable misclassification errors on the two datasets. However, as the number of splits increases, the full tree starts *overfitting* the validation data: since it fully reflects the training dataset without distinguishing between "signal" and "noise", the noisy

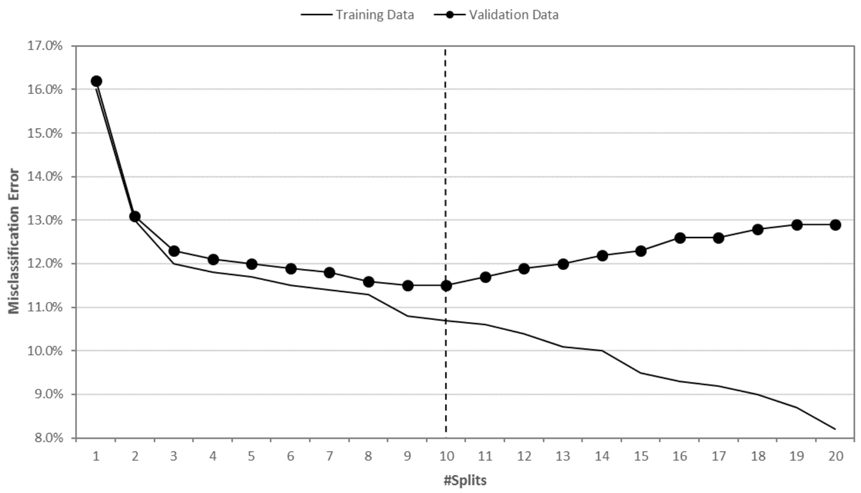


Figure 1: Training Misclassification Error vs. Validation Misclassification Error

component cause too high misclassification error in the validation dataset. Indeed, the typical consequence of overfitting is that, after some number of splits, the misclassification error on the validation dataset stops decreasing and starts increasing (in Figure 1, it occurs after ten splits). In the first ten splits both the training and validation misclassification errors decrease, but thereafter the full tree overfits the validation data.

Overfitting prevent us from using the full tree for predicting purposes, so we need to choose another tree. The most natural choice is suggested in Figure 1 itself, i.e. the classification tree built by the first n splits that do not induce overfitting ($n = 10$ in Figure 1). In other word, we choose the full tree's subtree leading to the lowest validation misclassification error: it is called *best pruned tree*. If we have new observations to classify, they will be dropped down the best pruned tree until they reach a leaf. So the full tree is useless for classification purposes, rather it is simply the formal result of recursive partitioning. What is really useful for classification is the best pruned tree only (or any other subtree built to somehow minimize some error measure on validation data).

So far we described the fundamentals of CARTs. One of the reasons for their popularity is that they are adaptable to a wide variety of applications, and have been successfully used in many situations. In particular, if there is a highly non-linear and complex relationship to describe, decision trees may outperform regression models. Furthermore, CARTs do not require massive data preparation, that is, they can handle non-standardized data, categorical data, missing data, outliers, and so on. By contrast, we should standardize the variables and take the natural logarithm of some numerical variables before running the logistic regression. Finally, trees provide easily understandable classification rules (at least if they are not too large), even easier than in regression.

An important advantage of CARTs is that no further selection algorithm is necessary. As opposed to logistic regression (see Section 3), the process itself selects the most relevant explanatory variables. We simply let the machine learning tool run on the whole dataset, and the resulting tree will include only some of the explanatory variables, which are the most significant on the base of the impurity measure used to split the dataset.

Unfortunately, CARTs do not have the same level of predictive accuracy and robustness as regression models. While the latter are characterized by low variance, decision trees tend to be relatively unstable, in the sense that little changes in the dataset may lead to completely different trees. Moreover, the goodness of each split is extremely dependent on the goodness of the previous splits. In fact, even if the algorithm picks the best split at each level, such a split is the less impure on *that* level, but we will never known whether a more impure split would have resulted in better predictions on the validation dataset.

Nonetheless, the predictive performance of CARTs can be dramatically improved by aggregating many decision trees from the same dataset. If we had n training datasets, we would use them to build n different classification trees, use them to predict the same validation dataset, and finally take the average of the n predictions. This approach would certainly reduce the variance of the estimation, but we do not have access to multiple training datasets. Instead, we can bootstrap by taking repeated samples from the single training dataset. The prediction $\hat{f}$ of a validation record x will be simply

$$\hat{f}(x) := \frac{1}{n}\sum_{k=1}^{n}\hat{f}_k(x) \tag{16}$$

where $\hat{f}_k(x)$ represents the prediction returned by the k^{th} classification tree (built on the k^{th} training sample) for the validation record x. This is called *bagging*. Notice that the n classification trees are all full tree, i.e. share very high variance, but also very low bias. Nonetheless, this is not relevant since the bagging process will reduce the variance. From a certain perspective, a bagging classification tree is an example of *ensamble*, that is, a machine learning tool resulting from the combination of several simpler machine learning methods.

While one of the main advantages of decision trees is interpretability, this feature is lost after bagging a high number of trees, because representing the prediction path on a single tree is no longer possible. However, an overall summary of the importance of each predictor is provided by the Gini index, as we will see in Section 6.

5. Segregated fund modeling and parametrization

In this section, we will build the whole actuarial model to test our lapse prediction, both asset and liability side. This section is based on [1].

The most difficult aspect of Italian segregated fund modeling is the simulation of the credited rate. In our case, it is not only used to reevaluate the sum assured (and the reserve), but even to predict the lapse rates themselves through the independent variable delta return. For the economic scenario generation, We will use a Gaussian two-factor model (like in [21] and [1]), which guarantees a number of useful properties. First, it embeds an instantaneous linear correlation between rates at different maturities, while single-factor models (e.g. CIR, Vasicek, etc.) implicitly assume correlation 1 (see [4]): effectively, each simulation leads to a rigid movement of the interest rate curve. Another reason relates to the fitting of the actual interest curve: while single-factor models, indeed, fit it, two-factor models like the Gaussian one can adapt to it perfectly.

The short rate under the Gaussian model is defined by the following equation:

$$i_t := X_t + Y_t + \phi(t) \tag{17}$$

where

$$dX_t = -\mu_x X_t dt + \sigma_x dZ_t^x \tag{18}$$

$$dY_t = -\mu_y Y_t dt + \sigma_y dZ_t^y \tag{19}$$

and initial conditions $X_0 = 0$ e $Y_0 = 0$. The instantaneous linear correlation between X and Y is represented by the parameter ρ:

$$dZ_t^x dZ_t^y = \rho dt. \tag{20}$$

The parameters in the equations (18), (19), and (20) are the same as in [21] and [1], based on the market data at 25/11/2016.
Assuming that the actual ZCB-price curve is interpolated by some polynomial function $\Pi(T)$, it can be proved (see [4]) that, if $f(T)$ denotes the instantaneous forward rate in T, that is,

$$f(T) := -\frac{d \ln \Pi(T)}{dT} \tag{21}$$

then the deterministic function $\phi(T)$ defined by

$$\begin{aligned} \phi(T) &:= f(T) + \frac{\sigma_x^2}{2\mu_x^2}(1 - e^{-\mu_x T})^2 + \frac{\sigma_y^2}{2\mu_y^2}(1 - e^{-\mu_y T})^2 + \\ &+ \frac{\rho\sigma_x\sigma_y}{\mu_x\mu_y}(1 - e^{-\mu_x T})(1 - e^{-\mu_y T}) \end{aligned} \tag{22}$$

guarantees a perfect fitting of the actual interest rate curve. However, notice that the choice of the polynomial function still affects the results. An example is given in [1]:

$$f(T) := \alpha + \beta_1 e^{-\frac{T}{\tau_1}} + \beta_2 e^{-\frac{T}{\tau_2}} + \gamma_1 \frac{T}{\tau_1} e^{-\frac{T}{\tau_1}} + \gamma_2 \frac{T}{\tau_2} e^{-\frac{T}{\tau_2}}. \tag{23}$$

which is calibrated on the Eurirs curve available at 25/11/2016. The related parameters are $\alpha = 1.203$, $\beta_1 = -2.733$, $\beta_2 = 1.594$, $\gamma_1 = -1.529$, $\gamma_2 = -4.093$, $\tau_1 = 1.059$ e $\tau_2 = 3.267$, which lead to the curve in Figure 2. The stochastic processes X_t and Y are defined by the parameters available in [21], i.e. $\mu_x = 40.1\%$, $\mu_y = 17.8\%$, $\sigma_x = 3.78\%$, $\sigma_y = 3.72\%$, and $\rho = -99.6\%$.
What really makes us prefer the Gaussian model with respect to others, more complex two-factor models is its analytical tractability, just like much simpler models such as the Vasicek model. For example, the authors of [4] prove that, at time t, the price of a ZCB maturing in $T > t$ is equal to

$$P(t,T) = e^{-\int_t^T \phi(\tau)d\tau - \frac{1-e^{-\mu_x(T-t)}}{\mu_x}X_t - \frac{1-e^{-\mu_y(T-t)}}{\mu_y}Y_t + V(t,T)} \tag{24}$$

where

$$V(t,T) = V_x(t,T) + V_y(t,T) + V_{xy}(t,T) \tag{25}$$

$$V_x(t,T) := \frac{\sigma_x^2}{2\mu_x^2}\left[T - t + \frac{2e^{-\mu_x(T-t)}}{\mu_x} - \frac{e^{-2\mu_x(T-t)}}{2\mu_x} - \frac{3}{2\mu_x}\right] \tag{26}$$

$$V_y(t,T) := \frac{\sigma_y^2}{2\mu_y^2}\left[T - t + \frac{2e^{-\mu_y(T-t)}}{\mu_y} - \frac{e^{-2\mu_y(T-t)}}{2\mu_y} - \frac{3}{2\mu_y}\right] \tag{27}$$

$$V_{xy}(t,T) := \frac{\rho\sigma_x\sigma_y}{\mu_x\mu_y}\left[T - t + \frac{e^{-\mu_x(T-t)} - 1}{\mu_x} + \frac{e^{-\mu_y(T-t)} - 1}{\mu_y} - \frac{e^{-(\mu_x+\mu_y)(T-t)} - 1}{\mu_x + \mu_y}\right]. \tag{28}$$

The equation (24) is extremely important since it provides us with a straightforward way to calculate stochastic deflators as a function of X and Y.
The short rate defined by the (17) should be calibrated from the actual risk-free curve. However, we also need a stochastic model for the future bond yields since the segregated fund invests in risky bonds.
Therefore, we adjust the (22) as follows:

$$\begin{aligned} \phi^*(T) &:= f(T) + \left(\frac{\sigma_x^2}{2\mu_x^2} + d_x\right)(1 - e^{-\mu_x T})^2 + \left(\frac{\sigma_y^2}{2\mu_y^2} + d_y\right)(1 - e^{-\mu_y T})^2 + \\ &+ \frac{\rho\sigma_x\sigma_y}{\mu_x\mu_y}(1 - e^{-\mu_x T})(1 - e^{-\mu_y T}) \end{aligned} \tag{29}$$

by using two deterministic factors which tends to the parameters d_x and d_y over time. In other words, the bond yield is simulated by the stochastic process

$$r_t := i_t + d_x(1 - e^{-\mu_x t})^2 + d_y(1 - e^{-\mu_y t})^2 \tag{30}$$

or equivalently

$$X_t^* := X_t + d_x(1 - e^{-\mu_x t})^2 \tag{31}$$

$$Y_t^* := Y_t + d_y(1 - e^{-\mu_y t})^2. \tag{32}$$

The parameters d_x e d_y, which represent the spread between the bonds in the segregated fund and the Eurirs curve, have been already calibrated in [1], in order to match an average 10-year spread approximately equal to the actual 10-year spread, i.e. 141 bps at 25/11/2016. Specifically, $d_x = d_y = 1.34\%$.
Generally, a minor part of the segregated fund is equity-based, so we need a stochastic model for it as well, for example a classical geometric Brownian motion defined by the risk-free component r_t, the risk premium parameter μ_S, and the non-systematic risk parameter σ_S:

$$S_t = S_0 e^{\int_0^t r_\tau d\tau + \left(\mu_S - \frac{\sigma_S^2}{2}\right)t + \sigma_S Z_t^S} \tag{33}$$

Market Index	μ_S	σ_S
FTSE MIB	-1.96%	20.22%
EURO STOXX 50	2.46%	12.15%
S&P 500	9.35%	10.83%

Table 1: Parameters of the equity component from the period 2010-2016

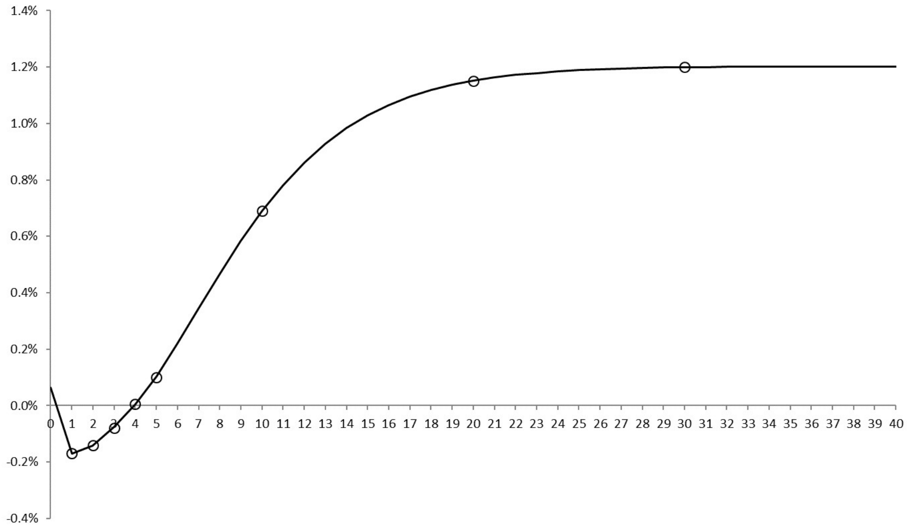


Figure 2: Polynomial function $f(T)$ interpolating Eurirs curve

where Z_t^S is uncorrelated with both Z_t^x and Z_t^y. For our application, we will consider three types of equity securities, calibrated on the performances from the last seven years. Assume that the segregated fund invests $b\%$ asset in n held-to-maturity coupon bonds bought at par and $1-b\%$ in equity. In $t=0$, when a new insurance contract is underwritten, each bond in portfolio has a different maturity, say $t_1, \dots, t_n$. In other words, the ith bond pays a known coupon c_i for the next t_i years. Since each bond is bought at par, its annual yield equals the coupon rate itself. More specifically, if F_i denotes the face value of the ith bond, its contribution to the segregated fund's credited rate is

$$C_i := \frac{c_i F_i}{\sum_{k=1}^{n} F_k}. \tag{34}$$

In fact, as long as $t \leq \min\{t_1, \dots, t_n\}$, that is, no bond has matured yet, the average yield of the bond component is

$$R_C := \sum_{i=1}^{n} C_i \equiv \sum_{i=1}^{n} \frac{c_i F_i}{\sum_{k=1}^{n} F_k}, \quad \forall t \leq \min\{t_1, \dots, t_n\} \tag{35}$$

which is known in $t = 0$ and constant. As soon as the ith bond matures after T_i years, it will be probably replaced by a comparable security, say a new coupon bond with same maturity in T_i years (for practical reasons, assume that T_i is greater than the duration of the insurance contract, in order to replace each bond at most one time). The new par bond yields exactly the stochastic forward rate $f(t_i + 1, T_i)$. Using the dummy function $\chi_{t \leq t_i}$, the contribution of the ith bonds to the segregated fund credited rate in t can be written as follows:

$$C_i(t) := \frac{[\chi_{t \leq t_i} c_i + (1 - \chi_{t \leq t_i}) f(t_i + 1, T_i)] F_i}{\sum_{k=1}^{n} F_k} \tag{36}$$

and finally the stochastic return of the whole bond component in the segregated fund:

$$R_C(t) := \sum_{i=1}^{n} C_i(t) = \sum_{i=1}^{n} \frac{[\chi_{t \leq t_i} c_i + (1 - \chi_{t \leq t_i}) f(t_i + 1, T_i)] F_i}{\sum_{k=1}^{n} F_k}, \quad \forall t. \tag{37}$$

Further, the (33) provides the return of the equity component:

$$R_S(t) := \frac{S_t}{S_{t-1}} - 1 = e^{r_{\tau-1} + \left(\mu_S - \frac{\sigma_S^2}{2}\right) + \sigma_S Z} - 1 \tag{38}$$

where Z denotes a standard normal distribution.
Given the yield contributions $R_C(t)$ and $R_S(t)$ for the bond and equity contributions respectively, the segregated fund stochastic credited rate is equal to

$$g(t) := b R_C(t) + (1 - b) R_S(t). \tag{39}$$

Nonetheless, sum assured and reserve reevaluation takes into account other contract parameters such as profit sharing η, minimum guaranteed rate R_{min}, and minimum management fee k. Therefore, the stochastic revaluation rate is

$$R(t) := \max\{\min\{\eta g(t), g(t) - k\}, R_{min}\} \tag{40}$$

	weight	coupon	duration in force	maturity
BTP1	33.3%	1.0%	0 years	10 years
BTP2	33.3%	3.0%	10 years	15 years
BTP3	33.3%	5.0%	23 years	30 years

Table 2: Initial bond component of the segregated fund

assuming no technical rate (which is common in Italian insurance contracts including a minimum rate guarantee).

To define $g(t)$ and hence $R(t)$, we still need to set the weight b and the initial bond portfolio. Given that the equity component of an Italian segregated fund is generally residual, we set $b = 90\%$. Moreover, the initial bond component will be represented by three BTPs, that is, Italian coupon bonds, as shown in the Table 2: Therefore, BTP1 has been just bought, BTP2 was bought 10 years ago, and BTP3 was bought 23 years ago. As soon as one of them matures, it will be replaced by a new bond with same maturity and yield equal to the related forward rate from the stochastic scenario.

The insurance contract we will analyze is a 20-year deferred capital without death benefit and terminal bonus at maturity. In the traditional case, we will assume that an average lapse rate of $\lambda = 5.61\%$ (as calculated from our dataset) is applied from the fourth policy year into the profit valuation, together with the mortality rate from the SIM2001. As a consequence, the average policy number in t is

$$N_t = \begin{cases} 1 & t = 0 \\ N_{t-1}(1 - q_{x+t}) & \forall t = 1, 2, 3 \\ N_{t-1}(1 - q_{x+t})(1 - \lambda) & \forall t = 4, \ldots, n-1 \\ N_{n-1}(1 - q_{x+n}) & t = n \end{cases} \tag{41}$$

where $x = 30$, and $n = 20$.

The annual premium P is fixed and affected by predetermined alpha cost α (2% per year entirely paid in at inception as a percentage of the premium), beta cost β (yearly 3% as a percentage of the premium), and gamma cost γ (yearly 0.5% as a percentage of the premium). Since P is also increased by a loading $l = 15\%$, the premium paid by the policyholder is

$$P = \frac{P_F(1 + l)}{1 - \alpha - \beta - \gamma} \tag{42}$$

where P_F is the theoretical fair premium.

The mathematical reserve in t is calculated in the following, straightforward way:

$$V_t = V_{t-1}[1 + R(t)] + P_F \qquad \forall t = 1, \dots, n \tag{43}$$

where $V_0 = P_F$.

The initial sum assured S_0 is function of P_F:

$$S_0 = P_F \frac{\ddot{a}_{\overline{x:n}|}}{{}_nE_x} \tag{44}$$

using the unloaded mortality rate from the SIM2001. For any $t > 0$, the sum assured is affected by the increase in reserve due to $R(t)$:

$$S_t = \frac{V_{t-1} + P_F \ddot{a}_{\overline{t+x:n-t}|}}{{}_{n-t}E_{x+t}} \qquad \forall t = 1, \dots, n. \tag{45}$$

As a consequence, the rate credited to the sum assured is not $R(t)$, but a much lower rate.

The aforementioned costs are assumed to offset the related expenses generated by the maintenance of the contract, so the initial expense is

$$E_0 = \alpha P \tag{46}$$

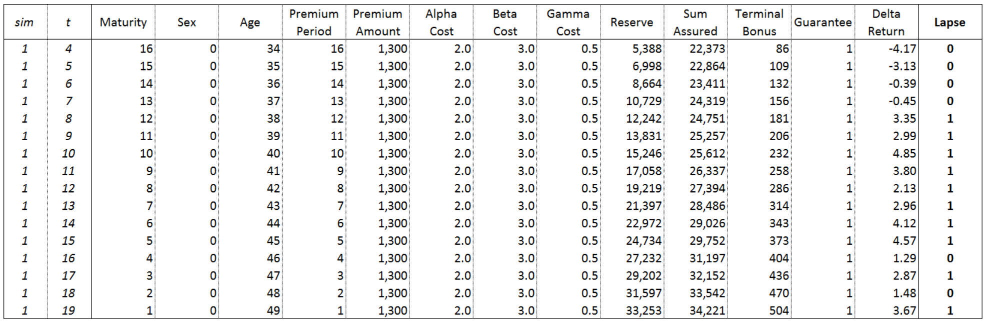
while the expense in any subsequent t is

$$E_t = (\beta + \gamma) P \qquad \forall t = 1, \dots, n. \tag{47}$$

Since no Zillmer reserve reduction for acquisition costs is permitted, and no timing gap between maintenance costs and premium payments occurs, no expense reserve will be allocated.

We can now calculate the annual profit generated by the contract. In $t = 0$, the company receives the first premium and immediately pays the acquisition costs. After that, the company still receives premiums and the related reserve begins to credit the rate $g(t)$; however, maintenance expenses and surrender benefits (from the fourth year) are paid. In the last policy year, the company receives the last return from the reserve, but pays the sum assured. In formulas

$$P\&L_t = \begin{cases} P - E_0 & t = 0 \\ \overline{P}_{t-1} + \overline{V}_{t-1} R(t) - \overline{E}_t - (\overline{V}_t - \overline{V}_{t-1}) & \forall t = 1, 2, 3 \\ \overline{P}_{t-1} + \overline{V}_{t-1} R(t) - \overline{E}_t - (\overline{V}_t - \overline{V}_{t-1}) - \lambda \overline{V}_t & \forall t = 4, \dots, n-1 \\ \overline{V}_{n-1} R(n) - \overline{E}_n - (\overline{S}_n - \overline{V}_{n-1}) & t = n \end{cases} \tag{48}$$



sim	*t*	Maturity	Sex	Age	Premium Period	Premium Amount	Alpha Cost	Beta Cost	Gamma Cost	Reserve	Sum Assured	Terminal Bonus	Guarantee	Delta Return	**Lapse**
1	*4*	16	0	34	16	1,300	2.0	3.0	0.5	5,388	22,373	86	1	-4.17	**0**
1	*5*	15	0	35	15	1,300	2.0	3.0	0.5	6,998	22,864	109	1	-3.13	**0**
1	*6*	14	0	36	14	1,300	2.0	3.0	0.5	8,664	23,411	132	1	-0.39	**0**
1	*7*	13	0	37	13	1,300	2.0	3.0	0.5	10,729	24,319	156	1	-0.45	**0**
1	*8*	12	0	38	12	1,300	2.0	3.0	0.5	12,242	24,751	181	1	3.35	**1**
1	*9*	11	0	39	11	1,300	2.0	3.0	0.5	13,831	25,257	206	1	2.99	**1**
1	*10*	10	0	40	10	1,300	2.0	3.0	0.5	15,246	25,612	232	1	4.85	**1**
1	*11*	9	0	41	9	1,300	2.0	3.0	0.5	17,058	26,337	258	1	3.80	**1**
1	*12*	8	0	42	8	1,300	2.0	3.0	0.5	19,219	27,394	286	1	2.13	**1**
1	*13*	7	0	43	7	1,300	2.0	3.0	0.5	21,397	28,486	314	1	2.96	**1**
1	*14*	6	0	44	6	1,300	2.0	3.0	0.5	22,972	29,026	343	1	4.12	**1**
1	*15*	5	0	45	5	1,300	2.0	3.0	0.5	24,734	29,752	373	1	4.57	**1**
1	*16*	4	0	46	4	1,300	2.0	3.0	0.5	27,232	31,197	404	1	1.29	**0**
1	*17*	3	0	47	3	1,300	2.0	3.0	0.5	29,202	32,152	436	1	2.87	**1**
1	*18*	2	0	48	2	1,300	2.0	3.0	0.5	31,597	33,542	470	1	1.48	**0**
1	*19*	1	0	49	1	1,300	2.0	3.0	0.5	33,253	34,221	504	1	3.67	**1**

Figure 3: Lapse predictions from simulation 1

where $\overline{P}_t := N_tP$, $\overline{V}_t := N_tV_t$, $\overline{E}_t := N_tE_t$, and $\overline{S}_t := N_tS_t$. Therefore, the total discounted profit is

$$D_n := \sum_{t=1}^{n} P\&L_tP(0,t) \tag{49}$$

where $P(0,t)$ denotes the stochastic deflator from (24).

So far we described the mathematics behind a typical profit testing of a simple insurance contract with a constant lapse rate. However, as soon as we introduce our lapse predictions, such a lapse rate loses meaning, given that the bagging classification tree predicts the exact lapse year. For example, the first interest rate scenario leads to the predictions in Figure 3. First, we assumed lapse is not permitted during the first three years, so predictions are needed from the fourth year. Similarly, no lapse is assumed to occur in the last policy year. In the specific case shown in Figure 3, the policyholder should lapse from the eighth policy year, but there are still a couple of future years - the sixteenth and the eighteenth - where non-lapse is predicted. Anyway, assuming that the policyholder will lapse in this scenario seems to be reasonable. In the simplest case, we can let him lapse in the first year with a lapse prediction (the eighth year for the simulation 1), but we can also assume that he/she will lapse at the second year with a lapse prediction, or even at the third year, and so on.

In this framework, λ should be set to zero, so that

$$N_t = \begin{cases} 1 & t = 0 \\ N_{t-1}(1 - q_{x+t}) & \forall t = 1, \ldots, n. \end{cases} \tag{50}$$

and

$$P\&L_t = \begin{cases} P - E_0 & t = 0 \\ \overline{P}_{t-1} + \overline{V}_{t-1}R(t) - \overline{E}_t - (\overline{V}_t - \overline{V}_{t-1}) & \forall t = 1, \dots, n-1 \\ \overline{V}_{n-1}R(n) - \overline{E}_n - (\overline{S}_n - \overline{V}_{n-1}) & t = n. \end{cases} \tag{51}$$

Correspondingly, assuming that the policyholder will lapse at time T in some specific scenario, the total discounted profit is

$$D^*_T := \sum_{t=1}^{T} P\&L^*_t P(0,t). \tag{52}$$

The following results will be based on D_n and D^*_T as well as other relevant measures derived from them. In particular, we will consider D_n as calculated in the certainty equivalent scenario, which is a sort of best estimate scenario including the traditional average lapse and the actual economic scenario:

$$D_{CE} := \sum_{t=1}^{n} P\&L_{CE,t} P_{CE}(0,t) \tag{53}$$

where P_{CE} denotes the deterministic deflator derived from the actual interest rate curve. Nonetheless, we will also consider D_n as the random variable in (49), which is function of the simulation k:

$$D_{k,n} := \sum_{t=1}^{n} P\&L_{k,t} P_k(0,t). \tag{54}$$

For the same reason, D^*_T is also function of the simulation k:

$$D^*_{k,T_k} := \sum_{t=1}^{T_k} P\&L^*_{k,t} P_k(0,t). \tag{55}$$

In conclusion, K simulations will provide us with the average λ-based total profit:

$$\overline{D}_n := E[D_{k,n}] = \frac{1}{K} \sum_{k=1}^{K} D_{k,n} \tag{56}$$

and the average T-based total profit:

$$\overline{D}_T := E[D^*_{k,T_k}] = \frac{1}{K} \sum_{k=1}^{K} D^*_{k,T_k}. \tag{57}$$

Although the average total profit - whether D_{CE}, $\overline{D}_n$, or $\overline{D}_T$ - is unquestionably the reference measure to value an insurance contract, there are other relevant measures as well. In particular, we will also consider the so-called time value of options and guarantees (TVOG).

When it comes with profit valuation, TVOG represents the fundamental difference between the traditional embedded value (TEV) framework and the more recent market consistent embedded value (MCEV). While the former is based on the mere certainty equivalent profit valuation (D_{CE}), the latter is based on the stochastic profit valuation (either $\overline{D}_n$ or $\overline{D}_T$). This means that MCEV is generally lower than TEV since

$$D_{CE} > \overline{D}_n \quad \text{and} \quad D_{CE} > \overline{D}_T \tag{58}$$

should hold because of the embedded option represented by the minimum guaranteed rate. The results in Section 6 will also show that

$$D_{CE} > \overline{D}_n > \overline{D}_T \tag{59}$$

although we cannot prove it always holds.

Notice that both TEV and MCEV are affected by other portfolio items beyond profit, so their reduction to the sole profit has no relevance in the actuarial practice, but it is still a useful and acceptable simplification for our purposes.

In the MCEV regulation, companies should not report the single stochastic profit, rather its split between certainty equivalent profit and TVOG:

$$\overline{D}_n = D_{CE} - (D_{CE} - \overline{D}_n) =: D_{CE} - TVOG_n \tag{60}$$

or

$$\overline{D}_T = D_{CE} - (D_{CE} - \overline{D}_T) =: D_{CE} - TVOG_T. \tag{61}$$

TVOG is crucial because it measures the value of any options and guarantees embedded in the contract. In our case, D_{CE} is a favorable scenario where the minimum guaranteed rate plays no relevant role. This is the reason why D_{CE} is very stable or even constant in most of the plots of Section 6. The full effect of the guarantee is only evident in the stochastic profit, that is, in the TVOG component.

6. Data preparation, lapse prediction, and impact analysis

The results we will show are based on the surrender data provided by a large Italian insurer from the period 2005-2015 for its endowment business. The whole dataset has been standardized, and each variable expressed in currency (i.e. premium amount, sum assured, mathematical reserve, and terminal bonus) has been replaced by its natural logarithm. It will make the related distributions comparable to a standard normal one, in particular with approximately zero asymmetry and kurtosis.
A large number of policy-specific features were available, while only one macroeconomic variable has been added, i.e. the difference between a reference market rate and the product crediting rate, just like in several published studies (for example, see [14]) and in the common actuarial practice. Naturally, we could add other relevant macroeconomic variables among those aforementioned in Section 2 (e.g. unemployment rate and gross domestic product), but such variables typically change very smoothly, in relation to the actual economic condition. Given that our analysis focuses on data of a specific company, and the historical horizon is not so long, we excluded any pure macroeconomic variable from the analysis.
Policy-specific explanatory variables include

- policyholder-related variables, i.e. sex and age
- contract-related variables, i.e. maturity, premium payment period, premium amount, premium loadings (alpha costs, beta costs, and gamma costs), and guaranteed rate
- path-dependent variables, i.e. sum assured, mathematical reserve, and terminal bonus.

The distinction between contract-related variables and path-dependent variables is fundamental to understand how the dataset has been structured. As confirmed in several empirical studies (for example, see [25]), surrender activity of a single policyholder - whether rational or irrational - tends to depend on a limited period of time. In other words, it is unlikely that policyholders will base their decisions on what has happened many years ago, or what is going to happen in many years. This is mainly the reason why any policy maturing in n years has been converted to n single-year policies. This leads to a policyholder behavior depending on the current year's condition only (i.e. current age, current number of premium payments, current duration, current reserve, current sum assured, and difference between the current reference market rate and the current crediting rate). Such "new" policies still share some common features, i.e. sex, premium amount, premium loadings, and technical rate. Among the n "new" policies, at most one can show surrender activity, i.e. we consider only surrender of the whole policy.

In data mining, datasets are usually partitioned in two sub-datasets, that is, a training dataset and a validation dataset (sometimes, a test dataset is also used, but it is not necessary for our purposes). The former "trains" the model, that is, the algorithm learns from its records and is tailored on them. By contrast, the latter is used to check how close the independent variable's predicted values are to its actual values by using such an algorithm on data that was NOT used to built it. Hopefully, the error on the validation dataset ("unknown" to the algorithm) will be as close as possible to the error on the training dataset ("known" to the algorithm).

Nonetheless, we will partition a dataset including few lapse occurrences. As a consequence, if we partition it randomly, the machine learning algorithm will be trained on few lapse occurrences. On the other term, it has little chance to get relevant information from such few records. This is the reason why *oversampling* is common practice in data mining: we build the training dataset in such a way that lapse occurrences are as likely as non-lapse occurrences (both 50%), and let the validation dataset include all the other records. Obviously, we will have fewer lapse occurrences to validate the algorithm, which is however trained on much more relevant information.

Using this dataset coupled with a proper prediction algorithm and a simplified ALM model (see Section 5), we will be able to predict which single-year policy of an hypothetical new contract will lapse, that is, the surrender year. Notice that more than one can show surrender activity (it is new data, so it is possible), but we assume that the policyholder lapses in the first possible surrender year.

We do not intend to generalize our absolute results. They are specific for one particular portfolio, whereas they may turn to be completely different for other portfolios. Nonetheless, we discussed how heterogeneous explanatory variables may be, varying by country, business line, or even policyholder. After a so huge number of empirical studies, we do not intend to discuss further the fundamental sources of policyholder behavior. By contrast, we will focus our attention on the impact that unanticipated surrender activity can have on the profit of a product. About that, much less has been written, especially because of the difficulties in embedding a comprehensive policyholder

	ρ
Maturity	-4.43%
Sex	-2.27%
Age	5.37%
Premium Period	-4.08%
Premium Amount	11.63%
Alpha Cost	1.86%
Beta Cost	7.48%
Gamma Cost	1.51%
Reserve	9.54%
Sum Assured	9.28%
Terminal Bonus	8.13%
Guarantee	1.48%
Delta Return	8.75%

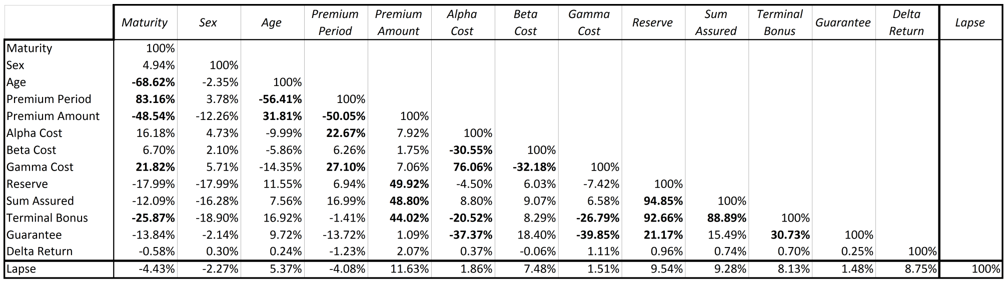
Table 3: Correlations with surrender activity

behavior model into profit valuation.

6.1. LR versus BCT

Obviously, traditional linear correlations between the independent variable and y and the explanatory variables $x_1, \dots, x_n$ are formally useless in a logistic regression. However, that is still based on a linear regression model. To some extent, we can still assume that the higher the correlations between explanatory variables and independent variable (in our case, a binary variable equal to 1 in case of surrender), the better the fitting of a logistic regression model. As just discussed, it is formally wrong, but linear correlations could still provide us with a useful indication about dependencies. The Table 3 summarizes the correlations between surrender activity and each explanatory variable. Remember that our dataset has more than 11.000 records and thus all correlations higher than about 4% can be considered significant.

Some of the relationships outlined in various empirical papers are somehow confirmed. Maturity is inversely correlated, i.e. policyholders tend to lapse in the first years, rather than in the last ones (the premium payment period is often close to the maturity, so its correlation is similar). Premium amount is significantly correlated, although it seems partially due to the loading components, especially beta costs. Similarly, both sum assured and reserve are



	Maturity	*Sex*	*Age*	*Premium Period*	*Premium Amount*	*Alpha Cost*	*Beta Cost*	*Gamma Cost*	*Reserve*	*Sum Assured*	*Terminal Bonus*	*Guarantee*	*Delta Return*	*Lapse*
Maturity	100%													
Sex	4.94%	100%												
Age	**-68.62%**	-2.35%	100%											
Premium Period	**83.16%**	3.78%	**-56.41%**	100%										
Premium Amount	**-48.54%**	-12.26%	**31.81%**	**-50.05%**	100%									
Alpha Cost	16.18%	4.73%	-9.99%	**22.67%**	7.92%	100%								
Beta Cost	6.70%	2.10%	-5.86%	6.26%	1.75%	**-30.55%**	100%							
Gamma Cost	**21.82%**	5.71%	-14.35%	**27.10%**	7.06%	**76.06%**	**-32.18%**	100%						
Reserve	-17.99%	-17.99%	11.55%	6.94%	**49.92%**	-4.50%	6.03%	-7.42%	100%					
Sum Assured	-12.09%	-16.28%	7.56%	16.99%	**48.80%**	8.80%	9.07%	6.58%	**94.85%**	100%				
Terminal Bonus	**-25.87%**	-18.90%	16.92%	-1.41%	**44.02%**	**-20.52%**	8.29%	**-26.79%**	**92.66%**	**88.89%**	100%			
Guarantee	-13.84%	-2.14%	9.72%	-13.72%	1.09%	**-37.37%**	18.40%	**-39.85%**	**21.17%**	15.49%	**30.73%**	100%		
Delta Return	-0.58%	0.30%	0.24%	-1.23%	2.07%	0.37%	-0.06%	1.11%	0.96%	0.74%	0.70%	0.25%	100%	
Lapse	-4.43%	-2.27%	5.37%	-4.08%	11.63%	1.86%	7.48%	1.51%	9.54%	9.28%	8.13%	1.48%	8.75%	100%

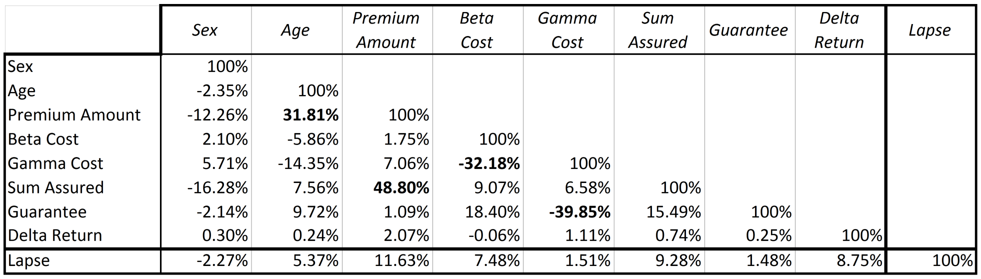
Table 4: Correlation matrix

positively correlated with the surrender activity (the two correlations are also very close, given that the reserve is a function of the sum assured): policyholders lapses only when it is really worth it. Finally, policyholders seem to be sensitive to the difference between what the financial market yields and what their policy effectively returns. It is worth noting that the delta return also embeds signals of a more volatile market as well as higher domestic funding cost: both of them stimulate surrender activity of policyholders, which are typically quite risk-adverse.

Those are the most significant correlations, but some non-significant correlations are interesting as well. For example, the technical rate guaranteed by the insurer has no significant impact on the surrender activity: it indirectly confirms the policyholder's short-term sight, which ignores the future benefits represented by a higher guaranteed rate.

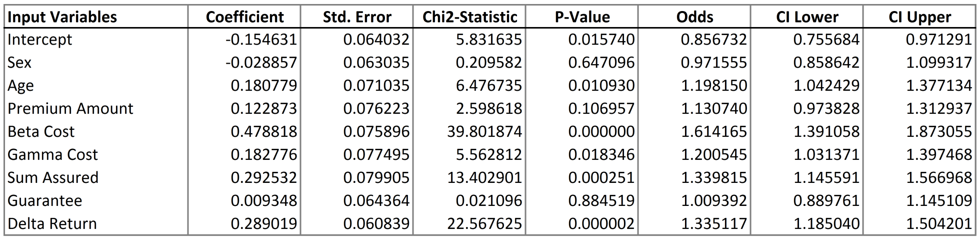
However, we should now consider the first limitation of the regression-based models, that is, we must reduce multicollinearity as much as possible, and, while observing the nature of the variables, we already know that some of them are correlated with each other very significantly. This is shown in the correlation matrix (Table 4), where we have highlighted in bold font all the correlations higher than 20% or lower than -20%. While some are well justifiable by the definition of the related variables (e.g. reserve, sum assured, and terminal bonus are all very correlated with each other), others are less expected. For example, the higher the guaranteed rate, the lower the gamma loading.

Of course, we must exclude the most redundant variables, especially those less correlated with the lapse variable. This variable reduction and the remaining correlations are shown in Table 5. Our goal was the exclusion of any correlation greater than 50% or lower than -50%. Some very significant



	Sex	*Age*	*Premium Amount*	*Beta Cost*	*Gamma Cost*	*Sum Assured*	*Guarantee*	*Delta Return*	*Lapse*
Sex	100%								
Age	-2.35%	100%							
Premium Amount	-12.26%	**31.81%**	100%						
Beta Cost	2.10%	-5.86%	1.75%	100%					
Gamma Cost	5.71%	-14.35%	7.06%	**-32.18%**	100%				
Sum Assured	-16.28%	7.56%	**48.80%**	9.07%	6.58%	100%			
Guarantee	-2.14%	9.72%	1.09%	18.40%	**-39.85%**	15.49%	100%		
Delta Return	0.30%	0.24%	2.07%	-0.06%	1.11%	0.74%	0.25%	100%	
Lapse	-2.27%	5.37%	11.63%	7.48%	1.51%	9.28%	1.48%	8.75%	100%

Table 5: Correlation matrix after variable reduction

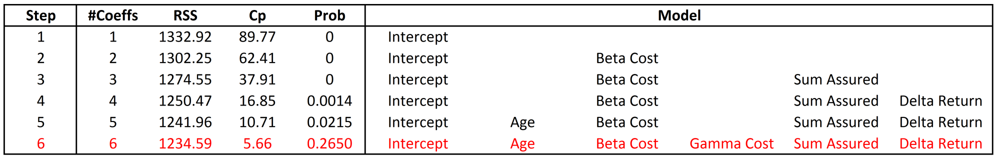


Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	-0.154631	0.064032	5.831635	0.015740	0.856732	0.755684	0.971291
Sex	-0.028857	0.063035	0.209582	0.647096	0.971555	0.858642	1.099317
Age	0.180779	0.071035	6.476735	0.010930	1.198150	1.042429	1.377134
Premium Amount	0.122873	0.076223	2.598618	0.106957	1.130740	0.973828	1.312937
Beta Cost	0.478818	0.075896	39.801874	0.000000	1.614165	1.391058	1.873055
Gamma Cost	0.182776	0.077495	5.562812	0.018346	1.200545	1.031371	1.397468
Sum Assured	0.292532	0.079905	13.402901	0.000251	1.339815	1.145591	1.566968
Guarantee	0.009348	0.064364	0.021096	0.884519	1.009392	0.889761	1.145109
Delta Return	0.289019	0.060839	22.567625	0.000002	1.335117	1.185040	1.504201

Table 6: Results of the stepwise logistic regression

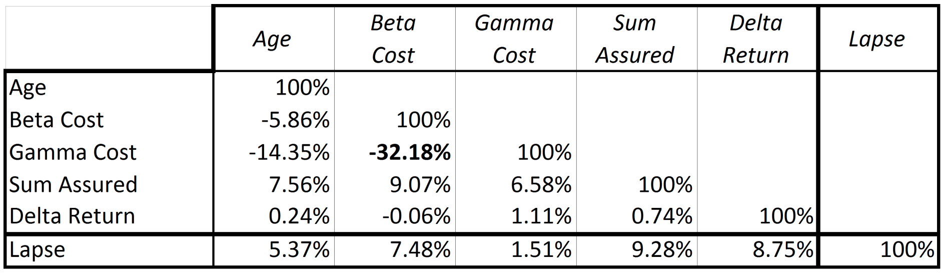
correlations are still there, but we have a chance that they will be automatically excluded by the variable selection process during the regression. Therefore, the stepwise logistic regression (see Section 3 for theoretical details) we run provided us with the results in Table 6, and the selection process in Table 7. Such a table highlights in red the chosen model in the last row, which is not so surprising since the stepwise regression has picked the most significant variables up, as evident from the Table 6. Moreover, most of the multicollinearity is now excluded from the model, as the final correlation matrix in Table 8 shows.

However, the resulting logistic regression model is quite poor, which could



Step	#Coeffs	RSS	Cp	Prob	Model					
1	1	1332.92	89.77	0	Intercept					
2	2	1302.25	62.41	0	Intercept		Beta Cost			
3	3	1274.55	37.91	0	Intercept		Beta Cost		Sum Assured	
4	4	1250.47	16.85	0.0014	Intercept		Beta Cost		Sum Assured	Delta Return
5	5	1241.96	10.71	0.0215	Intercept	Age	Beta Cost		Sum Assured	Delta Return
6	6	1234.59	5.66	0.2650	Intercept	Age	Beta Cost	Gamma Cost	Sum Assured	Delta Return

Table 7: Steps of the stepwise logistic regression



	Age	*Beta Cost*	*Gamma Cost*	*Sum Assured*	*Delta Return*	*Lapse*
Age	100%					
Beta Cost	-5.86%	100%				
Gamma Cost	-14.35%	**-32.18%**	100%			
Sum Assured	7.56%	9.07%	6.58%	100%		
Delta Return	0.24%	-0.06%	1.11%	0.74%	100%	
Lapse	5.37%	7.48%	1.51%	9.28%	8.75%	100%

Table 8: Correlation matrix after stepwise regression

be anticipated by the low correlations between the explanatory variables and the independent variable. ROC curves on training dataset (Figure 4) and validation dataset (Figure 5) provide us with a measure of the fitting and predictive power. They represent an area under the curve (AUC) of 69% and 64% respectively.

Even if we wanted to use this model, one more issue should be considered. A fundamental assumption of any logistic regression model is the normal distribution of the Pearson residuals. This can be quite hard to meet, so it is not obvious at all that we may use the logistic regression. In our analysis, we can prove the normality of residuals through a simple QQ-plot like in Figure 8. Unfortunately, such a plot emphasizes a remarkable problem to the normality assumption. In the negative branch, the residuals are not normally distributed at all, whereas we notice a relevant number of outliers in the positive one. In fact, it prevents us from using the logistic regression on this dataset, and it leads us to consider different models, say data-based models, which may return more powerful prediction, still avoiding the typical assumptions of a regression-based model.

So let's draw attention to the bagging classification tree method described in Section 4. Figure 9 describes one of the many possible classification trees built on our dataset. The blue circles are nodes, i.e. they represent the splitting rule within the dataset. For example, the first node bases its splitting rule on the sum assured, that is, the records with a sum assured lower than 0.63 are separated from the records with a sum assured greater than 0.63 (remember that the data are standardized). The splitting generates two sub-datasets: the first one includes 8082 records as represented by the left node generated by the split, while the second one includes 2500 records as represented by the right node generated by the split. The process is then

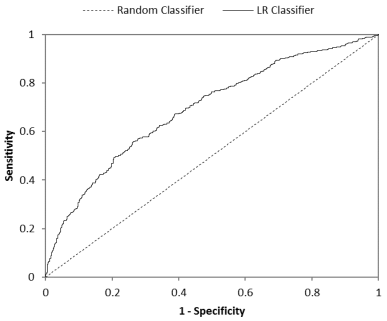


Figure 4: LR Training ROC

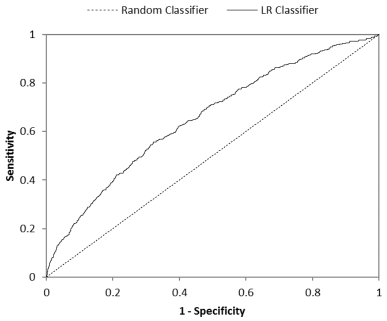


Figure 5: LR Validation ROC

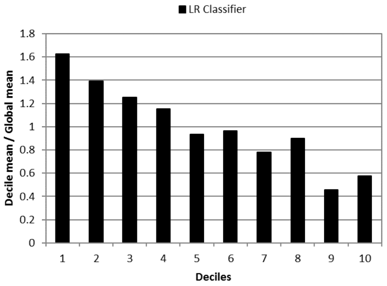


Figure 6: LR Training Decile

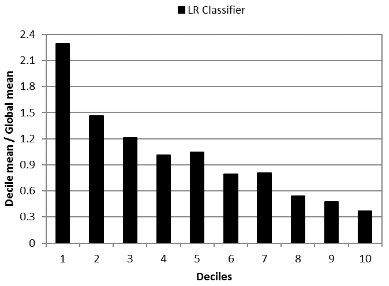


Figure 7: LR Validation Decile

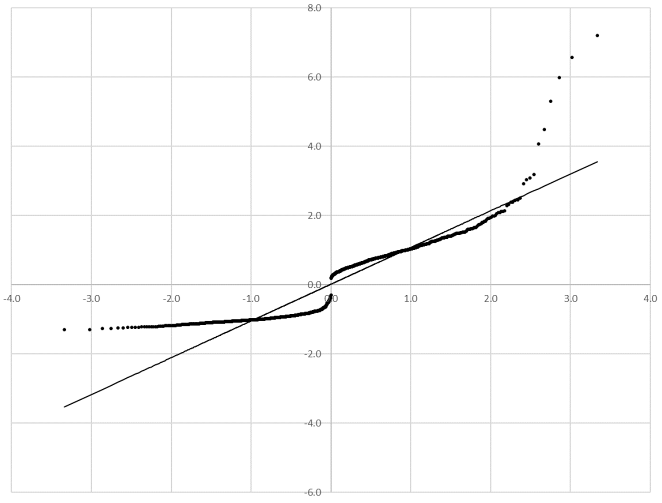


Figure 8: QQ-Plot of LR Pearson Residuals

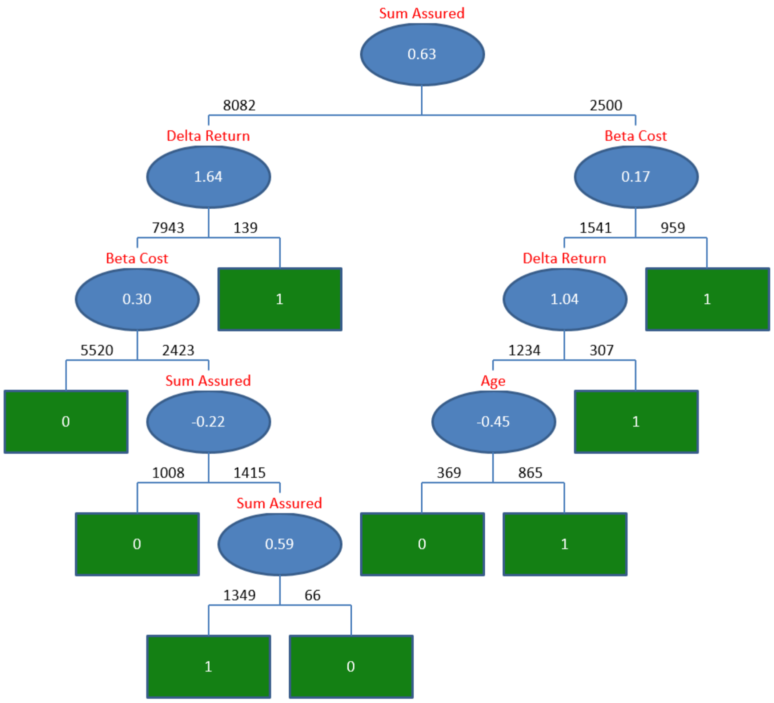


Figure 9: An Example of Classification Tree

iterated by using the new sub-datasets generated by the last splitting rule. When no more discrimination is possible, the records included in some sub-dataset can be all classified as either lapse (1) or non-lapse (0). In Figure 9 it is represented by the green squares, i.e. the leaves of the tree. After each step, the misclassification error is reduced since some more records are correctly classified. When such a recursive partitioning process is over, i.e. when no further split is possible, we have built the *full tree*: each record is classified exactly as it should, and the misclassification error is zero.

As discussed in Section 4, the full tree is of scarce utility. Of course, it guarantees no misclassification error in the dataset which generated it (i.e. the training dataset), but it also leads to relevant misclassification error in any other dataset, including the validation dataset. Using the bagging classification tree to predict lapses in our dataset (as already suggested in [18] and [12]), we get the ROC curves in Figure 10 and Figure 11 for training and validation respectively (the lighter curves refer to the logistic regression

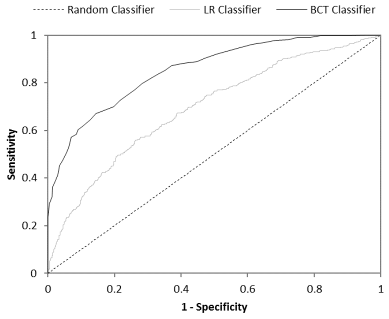


Figure 10: BCT Training ROC

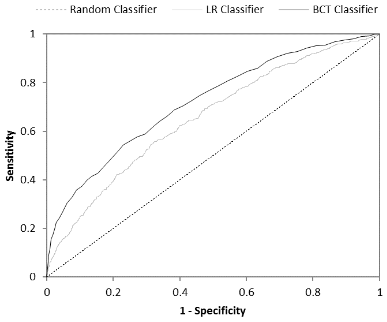


Figure 11: BCT Validation ROC

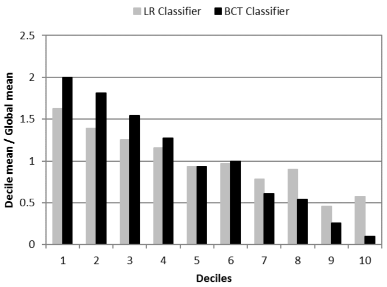


Figure 12: BCT Training Decile

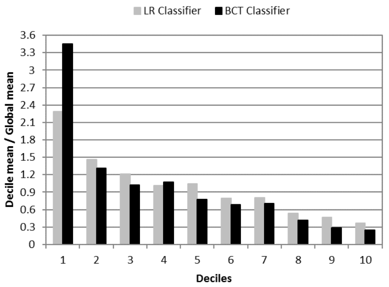


Figure 13: BCT Validation Decile

ROCs, the same as in Figure 4 and Figure 5). The corresponding AUCs are 85% and 72%, which indicate significant improvements with respect to 69% and 64% from logistic regression. However, we should notice the increase in overfitting when moving from a regression model to a machine learning tool: the difference between the training AUC and the validation AUC has more than doubled. In other word, the goodness-of-fit seems to improve much more than the predictive power. This is a typical effect in machine learning. What especially matters is the increase in validation AUC.

Figure 12 and Figure 13 show the training and validation decile histograms (the lighter bars refer to the logistic regression decile histograms, the same as in Figure 6 and Figure 7). Although they look different, they are basically delivering the same information. Remember that the training dataset was oversampled to 50% proportion of lapses, so the strong improvement in fitting is visible in the first four deciles, where we still have lapses. By contrast, in the validation dataset, lapses are all massed in the first decile, and the

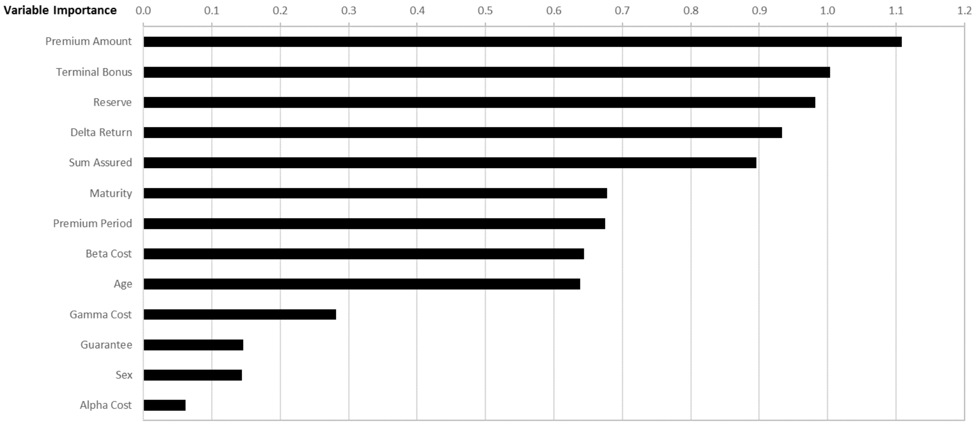


Figure 14: Bagging Classification Tree's Variable Importance

difference in predictive power is marked especially there. This is certainly the most important result: while the logistic regression predicts lapses about 2.3 times more often than a random method, the bagging regression tree predicts lapses nearly 3.5 times more often than a random method.

As we anticipated before, the bagging classification tree also provides us with a sort of variable importance. This is shown in the Figure 14. Unsurprisingly, the most important variable is the premium amount, which is also the most correlated one with the response (see Table 3). However, remember that it was excluded by the stepwise logistic regression (see Table 7). Beyond premium amount, the reserve-related variables (i.e. terminal bonus, reserve, and sum assured) are all quite important, which is understandable considering that all of them are highly correlated with each other. Furthermore, delta return seems to have a comparable importance as well. Less important variables include, for example, maturity, premium period, beta cost, and age. One more time, consider how much this is in line with the correlations in Table 3, even more in line than in the logistic regression model.

Finally, it is worth comparing the misclassification errors of logistic regression and bagging regression tree, although it is basically a more practical way to express the results of the ROC curves. Since we are more interested in the predictive power of the two models, let's focus our attention on the validation misclassification errors. Obviously, both the models have been run with the same cutoff probability, i.e. 50%. Overall, the misclassification error from

	Training		*Validation*	
	Pred. 1	Pred. 0	Pred. 1	Pred. 0
Act. 1	397	197	374	220
Act. 0	230	364	4151	5837

Table 9: LR Training and Validation Classification Matrix

	Training			*Validation*		
	#Cases	#Errors	%Errors	#Cases	#Errors	%Errors
1	594	197	33.2%	594	220	37.0%
0	594	230	38.7%	9988	4151	41.6%
Tot	1188	291	35.9%	10582	4371	41.3%

Table 10: LR Training and Validation Misclassification Error

	Training		*Validation*	
	Pred. 1	Pred. 0	Pred. 1	Pred. 0
Act. 1	432	162	381	213
Act. 0	129	465	3384	6604

Table 11: BCT Training and Validation Classification Matrix

	Training			*Validation*		
	#Cases	#Errors	%Errors	#Cases	#Errors	%Errors
1	594	162	27.3%	594	213	35.8%
0	594	129	21.7%	9988	3384	33.9%
Tot	1188	291	24.5%	10582	3597	34.0%

Table 12: BCT Training and Validation Misclassification Error

the logistic regression is 41.3% against 34.0% from the bagging classification tree. In fact, the latter can predict about two outcomes out of three policyholders. The biggest difference between the two models does not regard lapse prediction, rather non-lapse prediction. While they share a similar error on the lapses (37.0% and 35.8%), the error on the non-lapses is significantly different: among about 10,000 non-lapses, bagging regression tree can predict almost 800 non-lapses more than logistic regression. Of course, we are most interested in lapses than non-lapses (false negatives), but too many false positives can impact the profitability of a portfolio in a remarkable way since, generally, the more the lapses the lower the profit estimation.

6.2. Profit and TVOG

A wide range of results can be analyzed by integrating the prediction of the bagging classification tree and the liability model described in Section 5. First, we should clarify that lapse activity impacts the profit calculation only, that is, the data used for the lapse prediction (see Table 3) are completely independent from the lapse activity itself. Although it could seem trivial, it allows us for the following process split:

1. calculation of the independent variables from each scenario simulation
2. lapse prediction through bagging classification tree
3. calculation of the profit from each scenario simulation

which are indeed independent with each other.
Results come from some specific sets of parameters, and $K = 1000$ economic scenarios based on the stochastic model in (17), together with the related bond and equity scenarios. The initial set of parameters includes the assumptions in Table 2 and Table 1 for the EURO STOXX 50, $b = 90\%$ asset allocated in debt securities, profit share $\eta = 90\%$, minimum guaranteed $R_{min} = 1\%$, minimum management fee $k = 0.2\%$, fair premium $P_F = 1000$, loading $l = 15\%$, alpha costs $\alpha = 2\%$, beta costs $\beta = 3\%$, and gamma costs $\gamma = 0.5\%$. No terminal bonus is assumed.
Of course, a relevant analysis should involve the policyholder reaction to an increase in premium, which is the most important surrender risk factor according to the Table 14. Remember that the reserve is proportional to the sum assured, which is proportional to the premium, that is, an increase in premium will really impact three of the five most important variables (terminal bonus would be also impacted, if it were not zero) in Table 14. Figure 15 compares D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ when annual premium varies from the initial

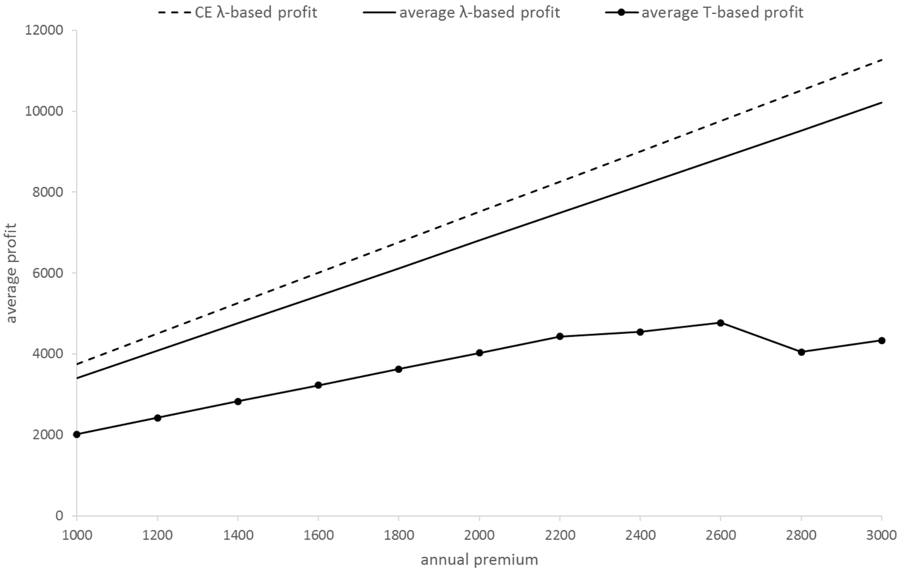


Figure 15: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by annual (fair) premium

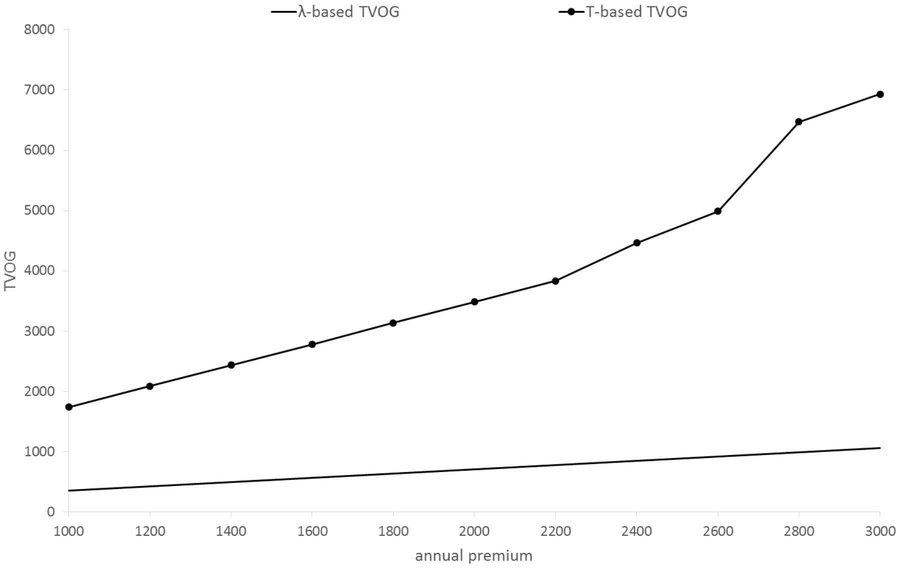


Figure 16: $TVOG_n$, and $TVOG_T$ varying by annual (fair) premium

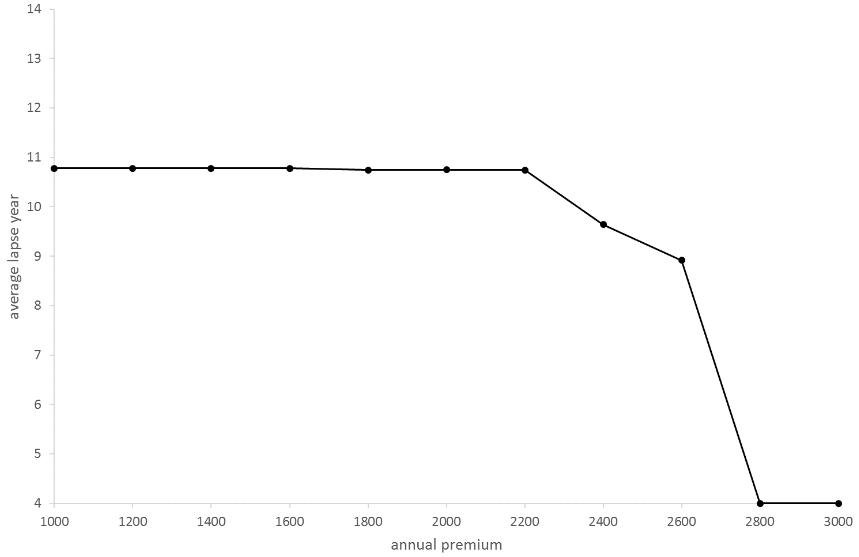


Figure 17: Average lapse year varying by annual (fair) premium

1000 to 3000. Take into account that the increase primarily affects the fair premium, so the final impact on the annual premium payments, which are impacted by loading and costs, is even greater. Also, the increase in fair premium leads a proportional increase in initial sum assured and thus reserve. Basically, an increase in premium should increase profit, and this is exactly what happens. However, while D_{CE} and $\overline{D}_n$ are relatively closed with each other (especially by low premiums) during their linear increment, $\overline{D}_T$ is strongly impacted by the dynamic policyholder behavior at any premium amount. Its growth is much more limited, and it even stops by the highest premium amounts. Correspondingly, $TVOG_T$ grows much faster than $TVOG_n$, and its shape seems somewhat exponential for high premium amounts.

In fact, these effects are well explained by Figure 17. Even in the basic scenario with annual premium 1000, the average lapse year is approximately the eleventh, and this is probably due to the actual market conditions. Unfortunately, this is enough to reduce the average profit from about $\overline{D}_n = 3500$ to about $\overline{D}_T = 2000$. As long as the premium increase has almost no effect on the average surrender activity - approximately by a premium of 2200 - $\overline{D}_T$ increases, but after that threshold the policyholder reaction is so immediate

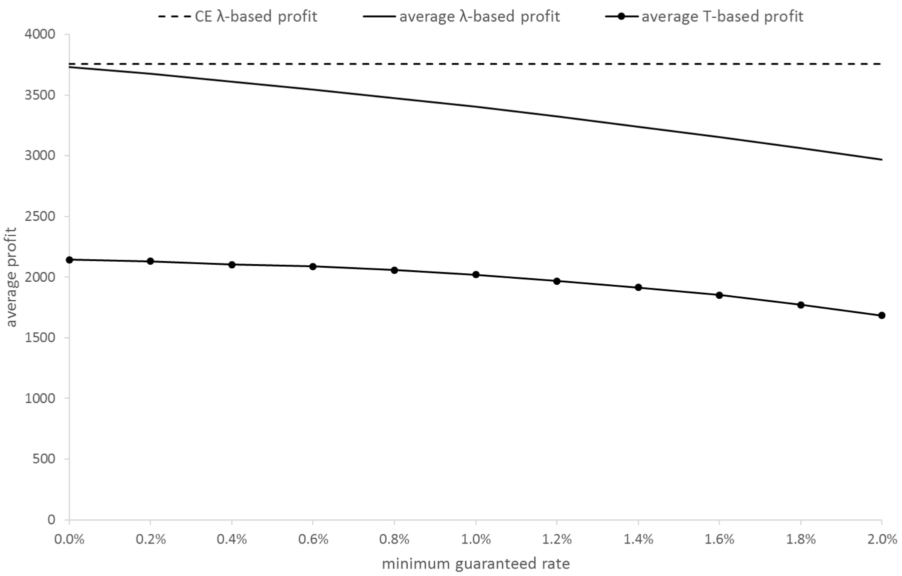


Figure 18: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by minimum guaranteed rate

to offset the profit of the contract. Probably, in case of better economic conditions, the policyholder tolerance would be higher, or premium amount would not be so important in lapse prediction, but in this particular case it could really lead to surrender *en masse*.

Even if the minimum guaranteed rate does not seem to be an important according to the Table 14, it may be worth noting how it can impact profit. It is shown in Figure 18. Even at guaranteed rate 0%, the gap between $\overline{D}_n$ and $\overline{D}_T$ is very large, but it should be caused by the actual market conditions, as we also discussed for the premium effect. Beyond this remark, $\overline{D}_n$ and $\overline{D}_T$ decrease somewhat parallelly as guaranteed rate increases, and this is also reflected in TVOG (19). Nonetheless, this similar shape is due to different reasons. While $\overline{D}_n$ solely decreases as a direct consequence of a higher rate to guarantee yearly (even in disadvantageous scenarios), $\overline{D}_n$ partly decreases because of profits lost from policyholder's surrender.

Given the regularity of the decrease in $\overline{D}_T$, we expect a regular increase in average lapse year. Looking at Figure 20, it approximately grows in a linear fashion from about 9 years to about 12 years.

So far we focused our attention on contract-based variables, which cannot

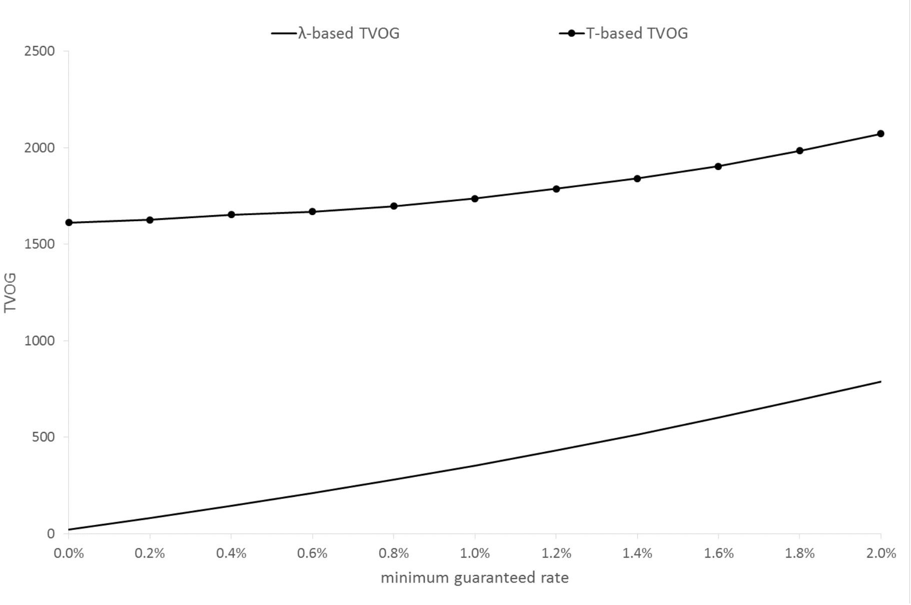


Figure 19: $TVOG_n$ and $TVOG_T$ varying by minimum guaranteed rate

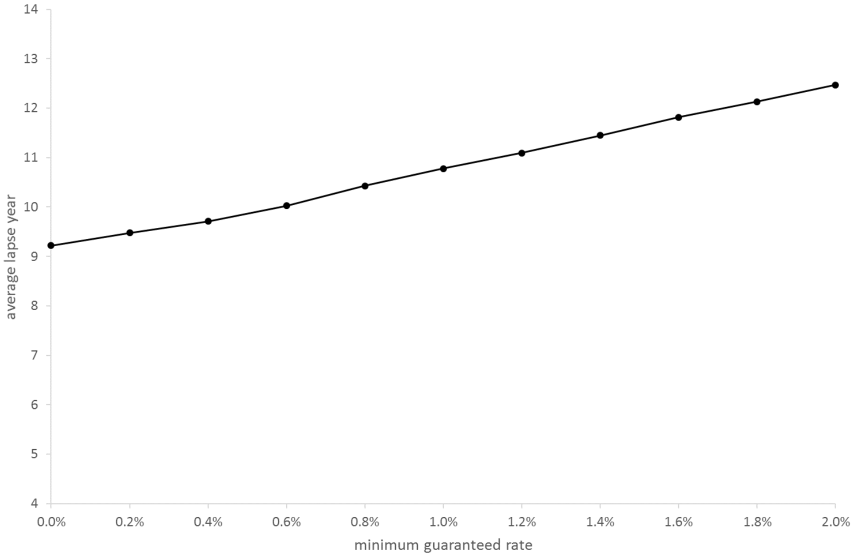


Figure 20: Average lapse year varying by minimum guaranteed rate

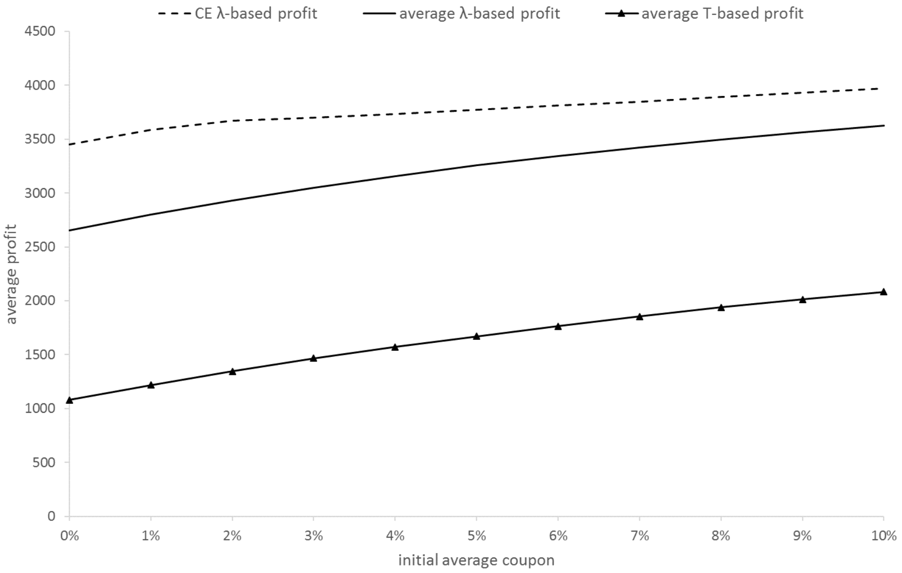


Figure 21: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by initial average coupon - FTSE MIB case

change during the life of the contract. However, the credited rate is especially function of the fund performance. Given that the so-called ”delta return” is among the most important variables according to the Table 14 in predicting surrender activity, we will consider variations in the fund-related parameters, in order to observe how lapse year is impacted via delta return.

On the base of the Table 2, the initial average performance of the bond component is 3%, but now we let it change between 0% to 10% adjusting each performance by ±1%. Notice that these parameters are used in the first policy years only, since new bonds will be bought as soon as the initial bonds mature. And new bonds will return the new (stochastic) forward rate. As a consequence, Figure 21, Figure 22, and Figure 23 all show an increase in profit as the initial coupon rate increases. At the same time, both $\overline{D}_n$ and $\overline{D}_T$ tend to D_{CE}, and this especially evident in Figure 23. In this case, the minimum guaranteed rate has low impact on profit because of the higher returns guaranteed by S&P 500 (see Table 1) as well as the growing initial coupon rate. As we can see in Figure 26, the higher the initial coupon rate, the later the policyholder's lapse activity, and this also contributes to the increase in $\overline{D}_T$ towards D_{CE}.

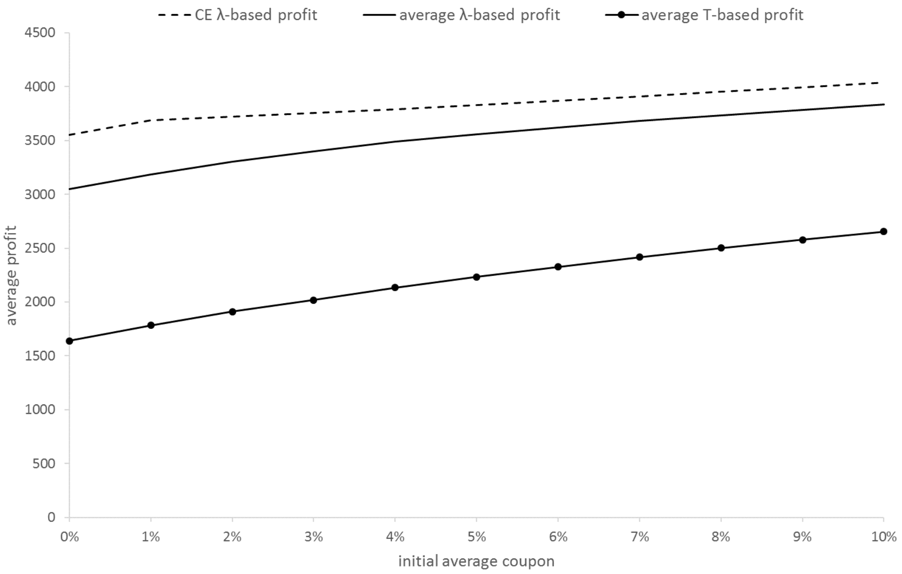


Figure 22: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by initial average coupon - EURO STOXX 50 case

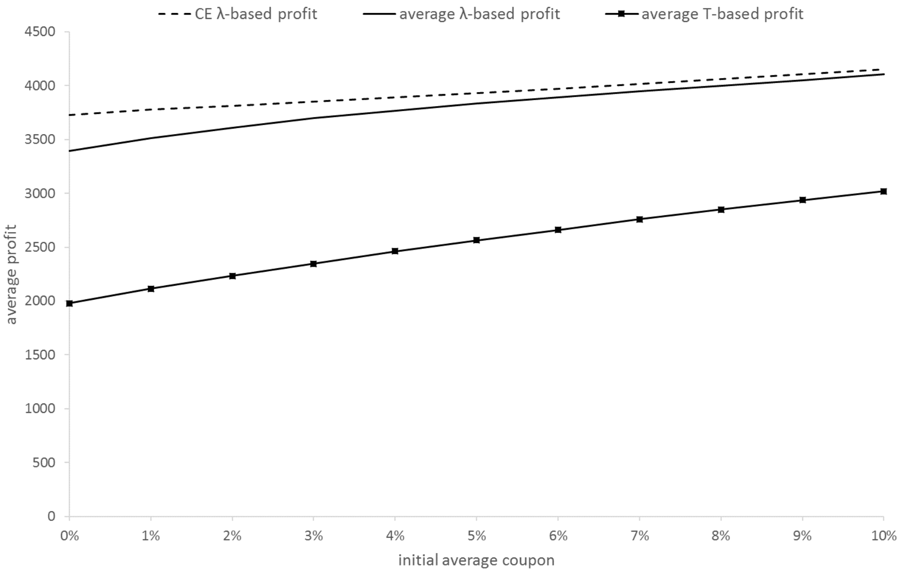


Figure 23: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by initial average coupon - S&P 500 case

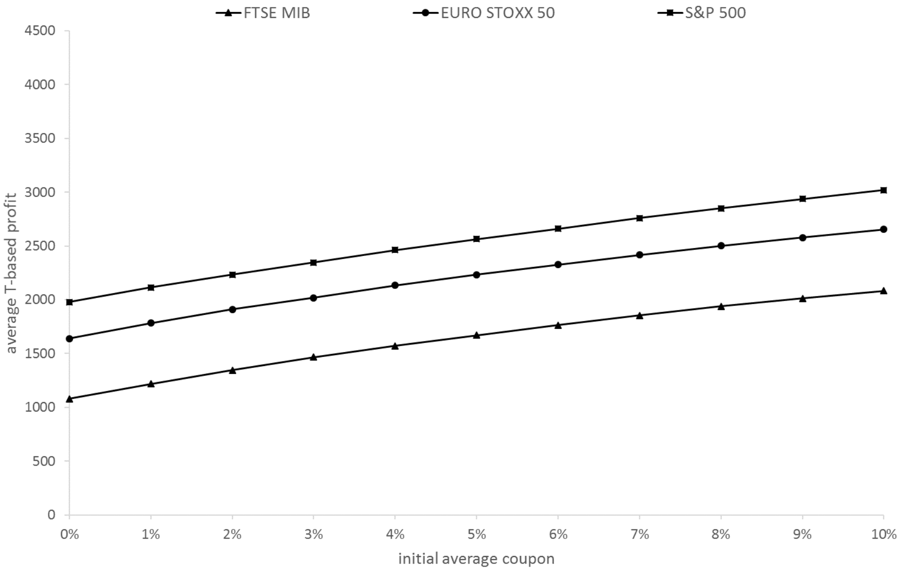


Figure 24: $\overline{D}_T$ varying by initial average coupon and equity investment

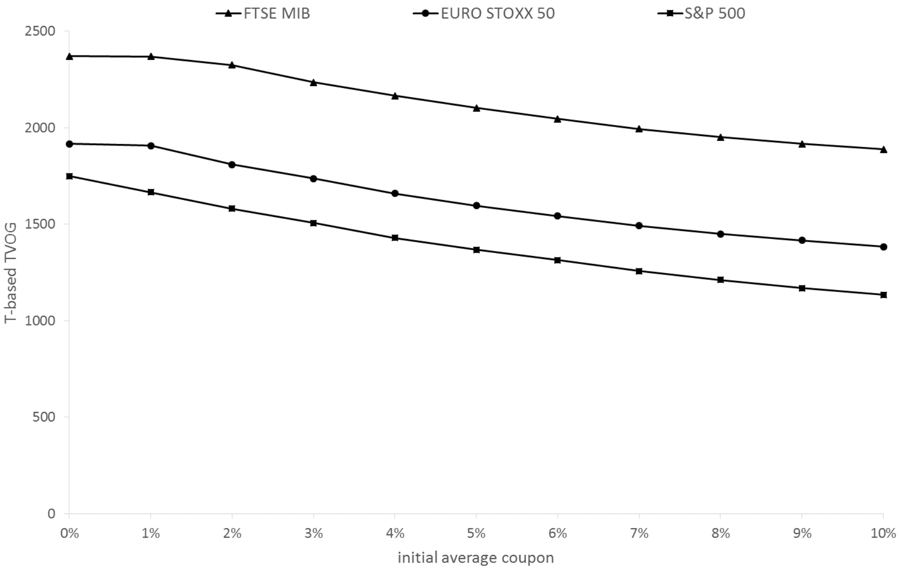


Figure 25: $TVOG_T$ varying by initial average coupon and equity investment

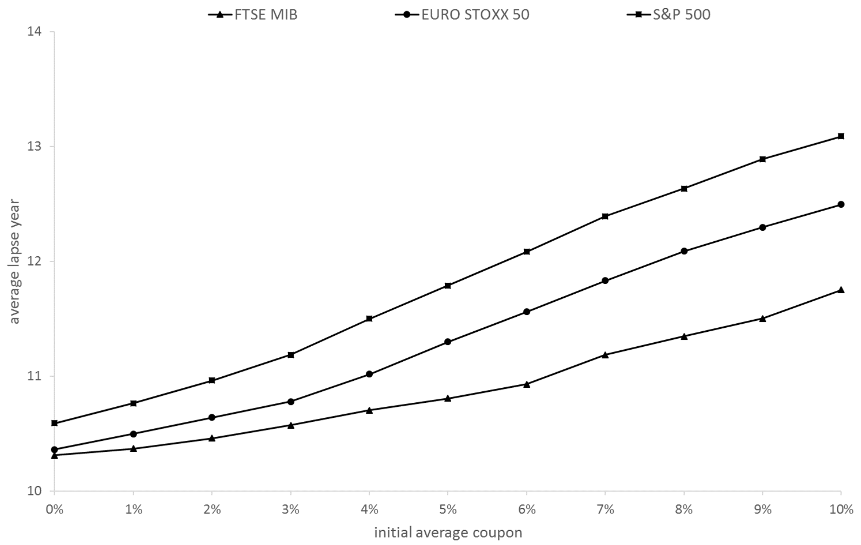


Figure 26: Average lapse year varying by initial average coupon and equity investment

However, Figure 26 reveals some difference between the average lapse year among the three different equity investment. While the curves for EURO STOXX 50 and S&P 500 appear approximately parallel, the curve for FTSE MIB grows more irregularly. Effectively, it shares the same average lapse year of the EURO STOXX 50 case when the coupon rate is zero, but the gap steadily increases proportionally to the coupon rate. Indeed, when the initial bond yield component is quite low, the policyholder lapses relatively early, as soon as he/she can find higher yields on the market. And this especially happens for low-return equity investments like FTSE MIB and EURO STOXX 50 (see Table 1) since bond yields have the major impact on the fund performance.
The last set of parameters considered in our analysis involves the equity allocation percentage of the fund. Since Italian segregated funds may incorporate just a minor equity investment (and real estate, which we do not consider here), typically upper-bounded at 20%-30%, we assume $1 - b \in [0\%, 20\%]$. We also repeat the analysis for the three different equity index in Table 1.
A first, strange effect we should explain is well evident in Figure 27, Figure 28, and Figure 29. When the equity percentage is low, $\overline{D}_n$ is even greater than D_{CE}. Theoretically, this is not admissible at all because $\overline{D}_n$ is affected

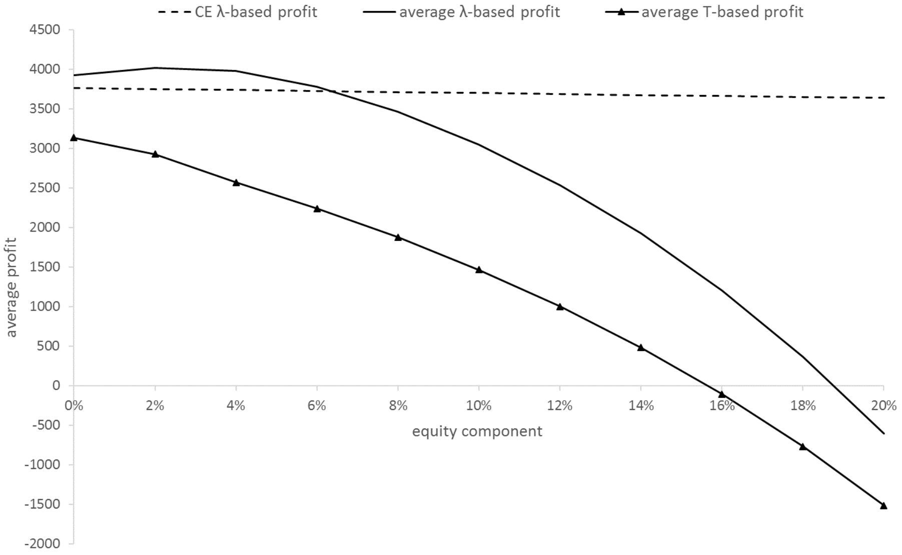


Figure 27: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by equity percentage - FTSE MIB case

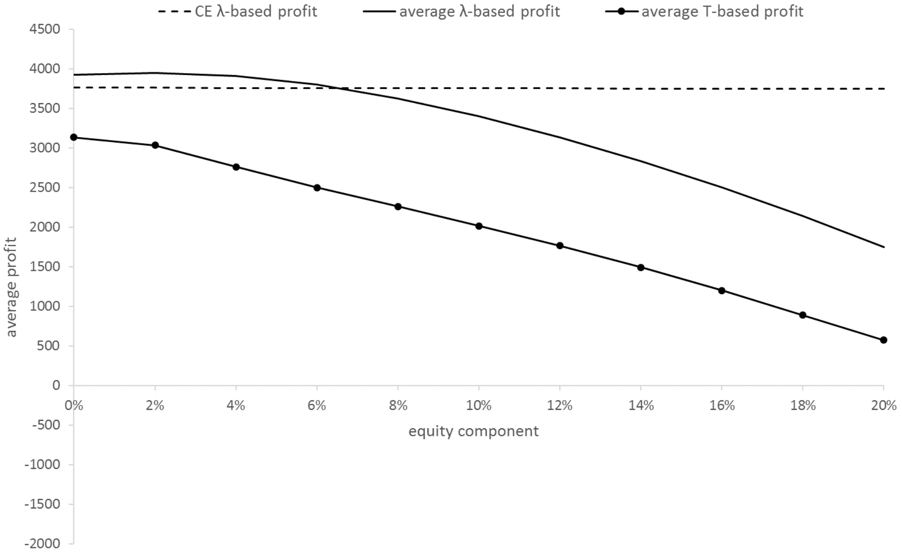


Figure 28: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by equity percentage - EURO STOXX 50 case

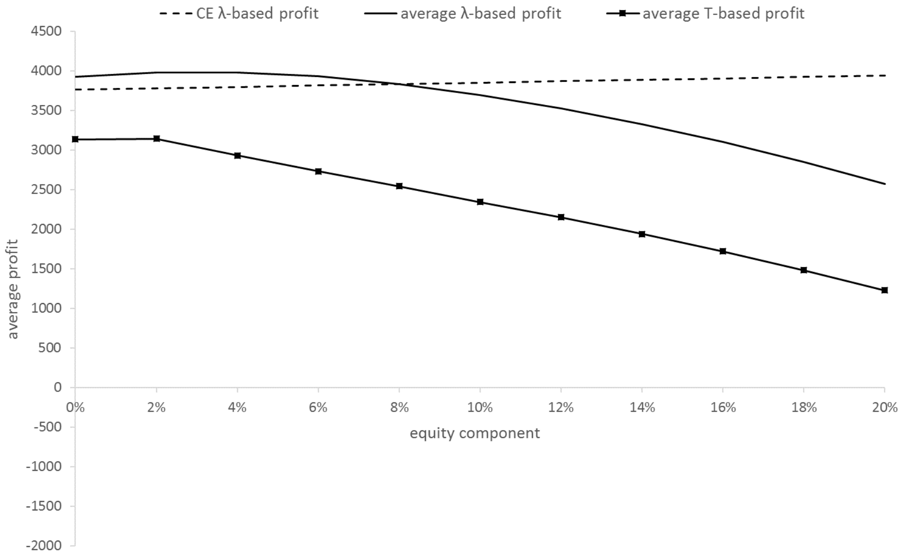


Figure 29: D_{CE}, $\overline{D}_n$, and $\overline{D}_T$ varying by equity percentage - S&P 500 case

by the rate guarantee in the disadvantageous scenarios, while D_{CE} is affected by the certainty equivalent scenario only (this directly impact the TVOG, as we will outline in Section 7). However, this anomalous gap is quite slight, probably due to the frequent negative-rate scenarios, so we will accept it as it is.

Without any equity component, the three profits are obviously equal (Figure 30), just greater than 3000. Notice that it is exactly equal to the total loading from the contract, i.e. 1000 × 15% × 20, since the credited rate matches the deflator whereas no investment profit comes from any equity component.

Reasonably, the only average loss occurs when investing in FTSE MIB, which yields a negative return on average (see Table 1). Nonetheless, average profit exponentially decreases as the equity percentage grows regardless to the market index. The main reason is well represented by the Figure 32. Whereas the policyholder tends to lapse later during the life of the contract if the equity component is residual, when it is approximately above 5% the average lapse year stabilizes at about 10.5-11.5 years (depending on the market index) since he/she starts profiting from high-yield equity scenarios, while being protected by the guarantee in low-yield scenarios. And such as

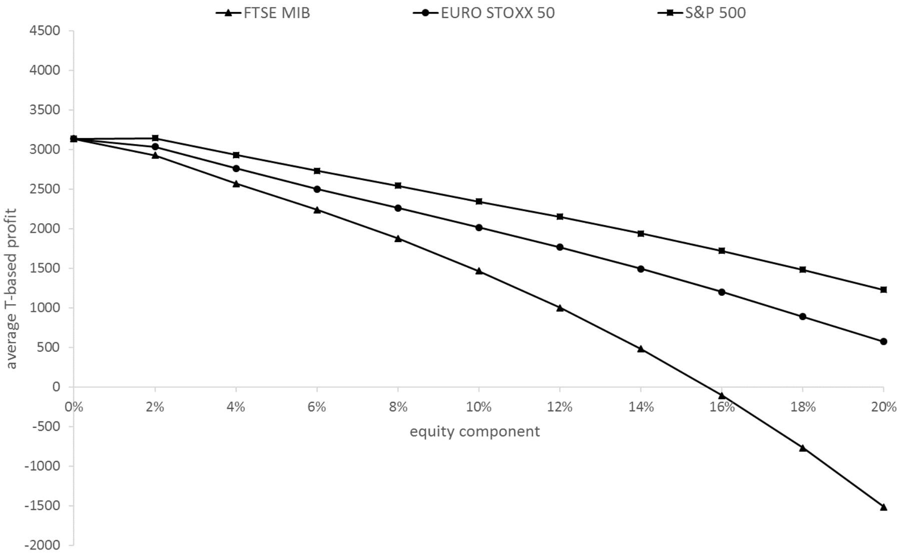


Figure 30: $\overline{D}_T$ varying by equity percentage and equity investment

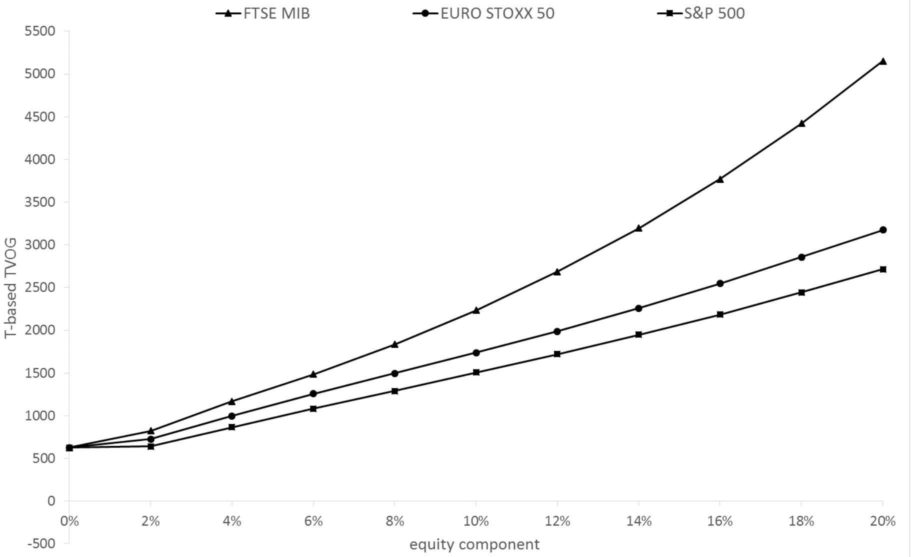


Figure 31: $TVOG_T$ varying by equity percentage and equity investment

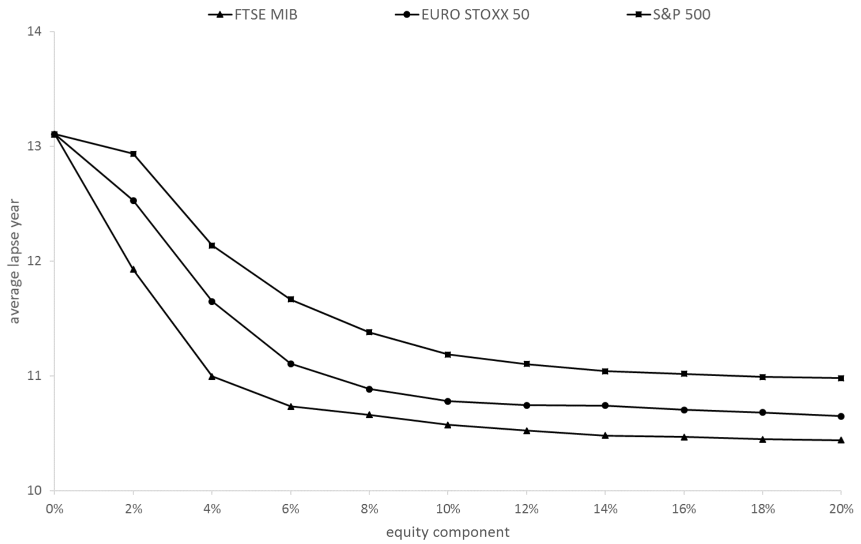


Figure 32: Average lapse year varying by equity percentage and equity investment

persistence behavior seems worth more than the lapse itself if the equity component is, indeed, not residual. Naturally, the policyholder's persistence in downward scenarios means a more and more significant loss for the company.

7. TVOG decomposition and policyholder behavior impact

A last result to mention regards the relative impact that an active policyholder behavior may have on an insurance contract. In effect, it could be negligible in some situations, whereas it could require some pricing adjustment in other situations.
Remember that $TVOG_n$ includes the effect of the guarantees, but not the effect of a dynamic policyholder behavior, where the surrender option is really exercised at some point in time. By contrast, $TVOG_T$ includes both the effects. This is basically the reason why we can assume that $\overline{D}_T < \overline{D}_n$, or equivalently $TVOG_T > TVOG_n$, and this is exactly what the plots in Section 6 have shown.
The gap between $TVOG_T$ and $TVOG_n$ is hence explained by the sole policyholder behavior, and it is a sort of measure for the surrender option's value,

say V_{PHB}. In formula

$$TVOG_T = TVOG_n + V_{PHB} \tag{62}$$

and it is interesting to study the impact of each component in several scenarios and parametrization, for example as a percentage of $TVOG_T$:

$$1 = \frac{TVOG_n}{TVOG_T} + \frac{V_{PHB}}{TVOG_T}. \tag{63}$$

This is represented in Figures 33-40 for each of the parametrization we have already considered in Section 6.
First, the PHB component seems to be relatively stable as a percentage of $TVOG_T$ if the annual premium increases (see Figure 33). Nonetheless is about 80% at least. By minimum guaranteed rate 0%, the PHB impact is nearly 100%, but it is not surprising at all since the guarantee effectively plays no role in the TVOG determination (see Figure 34). And naturally, the greater the guaranteed rate, the more significant its impact on TVOG, the less significant PHB impact on TVOG. By very high guarantees, the policyholder has much fewer reasons to lapse, and the TVOG will be entirely function of the guarantee itself.

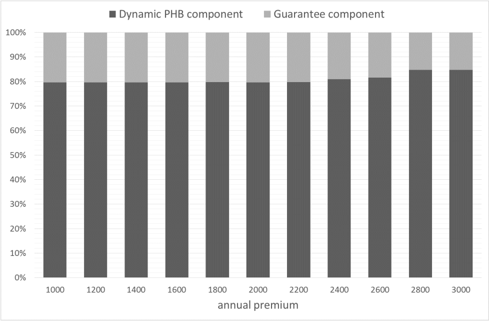


Figure 33: PHB-guarantee impact varying by annual (fair) premium - EURO STOXX 50

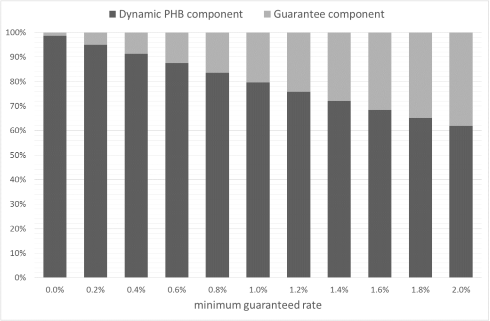


Figure 34: PHB-guarantee impact varying by min. guaranteed rate - EURO STOXX 50

Figure 35, Figure 37, and Figure 39 look relatively similar. As the initial average coupon from the bond in the segregated fund increases, the PHB impact slightly grows in each of the three cases. This is not so easy to understand. On a hand, higher coupons make the guarantee less significant, but on the other hand it should convince the policyholder to keep the contract.

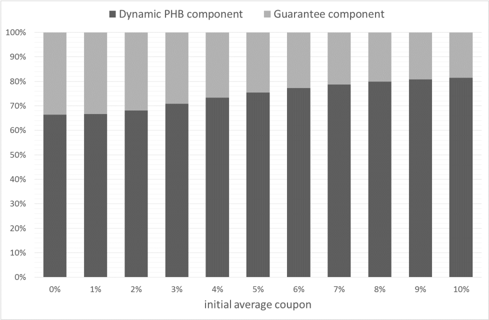


Figure 35: PHB-guarantee impact varying by initial average coupon - FTSE MIB

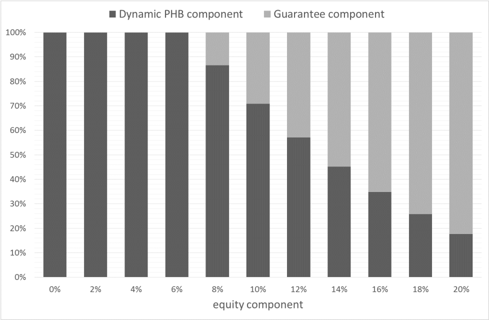


Figure 36: PHB-guarantee impact varying by equity percentage - FTSE MIB

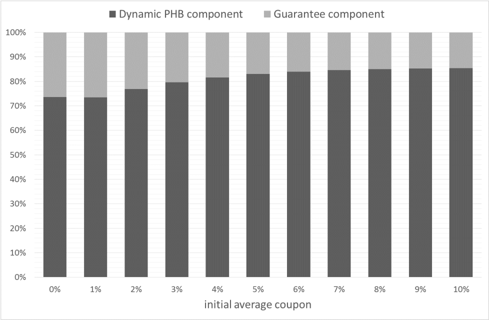


Figure 37: PHB-guarantee impact varying by initial average coupon - EURO STOXX 50

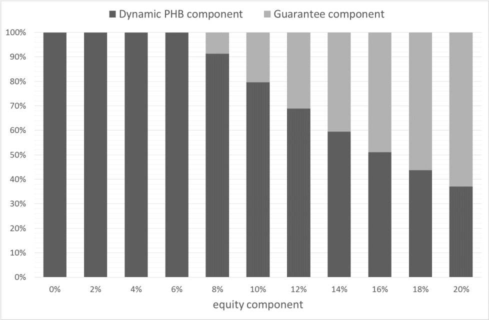


Figure 38: PHB-guarantee impact varying by equity percentage - EURO STOXX 50

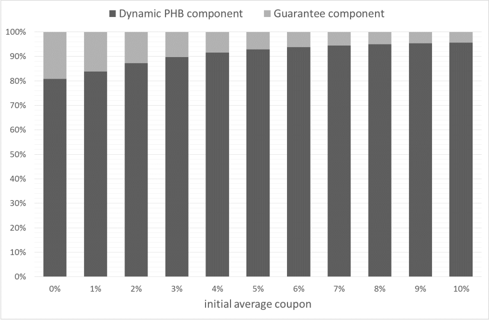


Figure 39: PHB-guarantee impact varying by initial average coupon - S&P 500

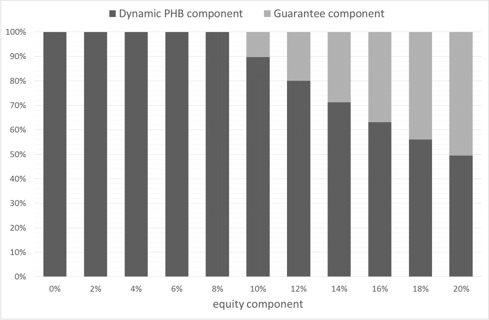


Figure 40: PHB-guarantee impact varying by equity percentage - S&P 500

However, the former seems to overcome the latter, that is, the policyholder still has some other reasons to lapse - and this is plausible, given that market conditions and fund performances do not represent the only PHB driver. Furthermore, this is also the reason why the PHB impact grows when investing in more performing equity indexes (for example, compare Figure 35 and Figure 39).

Figure 36, Figure 38, and Figure 40 look very similar as well. When the equity component is residual, TVOG solely comes from policyholder behavior, but remember we used an initial average coupon of 3% against a minimum guarantee rate of only 1%, that is, there is no impact from the guarantee in the first policy years (almost) regardless the economic scenario. Then, the introduction of a significant equity component increases the investment risk of the fund, and anticipates surrender activity. In the hypothetical case of 100% investment in equity, the policyholder has no motivation to lapse: while he/she benefits from any upside in any equity scenario - more favorable than the correspondent government bond scenario by definition - he/she is also covered by the guarantee in any downside scenario.

The results reported so far were derived from the assumption that the policyholder lapses as soon as a policy year is predicted as a lapse year by the bagging classification tree. However, this could be a too aggressive assumption. Indeed, remember that the machine learning algorithm succeeds in predicting about two correct behaviors - whether lapse or no lapse - each third. In other words, each lapse prediction is about 33% likely to be an actual non-lapse. At the same time, it sounds reasonable to imagine that the policyholder does not react immediately to concrete lapse conditions, rather

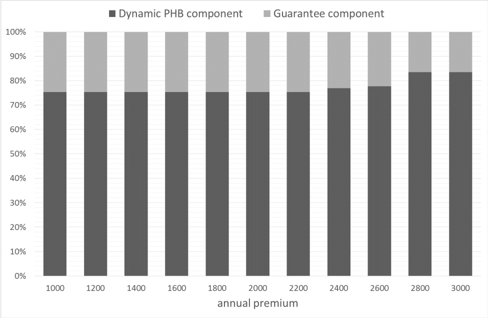


Figure 41: PHB-guarantee impact varying by annual (fair) premium - EURO STOXX 50 and delayed lapse

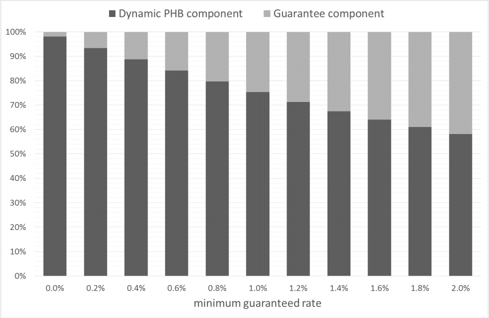


Figure 42: PHB-guarantee impact varying by min. guaranteed rate - EURO STOXX 50 and delayed lapse

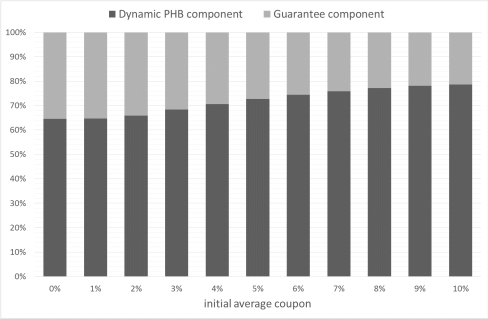


Figure 43: PHB-guarantee impact varying by initial average coupon - FTSE MIB and delayed lapse

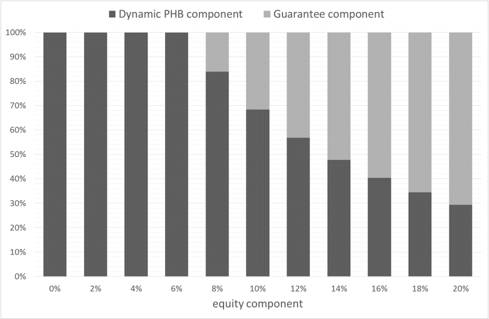


Figure 44: PHB-guarantee impact varying by equity percentage - FTSE MIB and delayed lapse

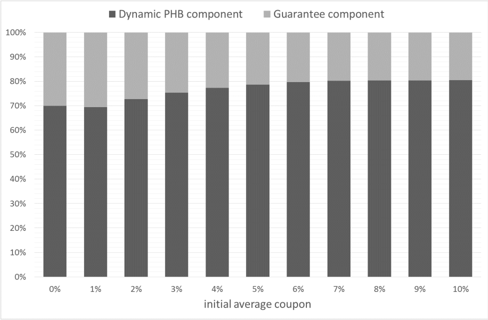


Figure 45: PHB-guarantee impact varying by initial average coupon - EURO STOXX 50 and delayed lapse

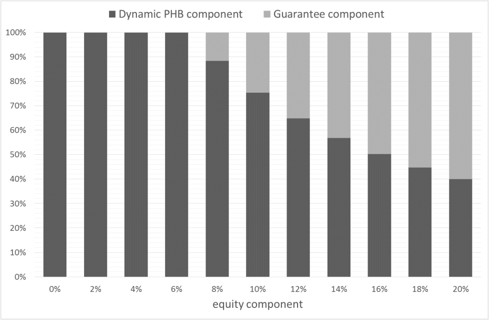


Figure 46: PHB-guarantee impact varying by equity percentage - EURO STOXX 50 and delayed lapse

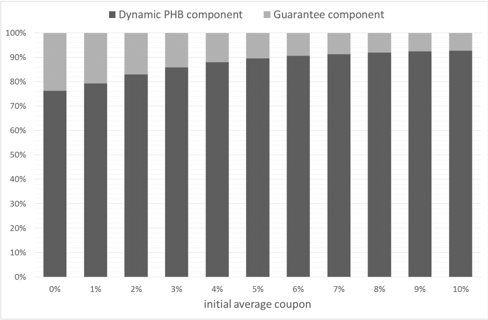


Figure 47: PHB-guarantee impact varying by initial average coupon - S&P 500 and delayed lapse

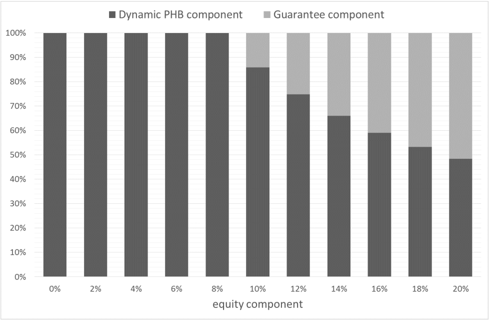


Figure 48: PHB-guarantee impact varying by equity percentage - S&P 500 and delayed lapse

he/she could delay the surrender.

For these reasons, we repeat the previous analysis while assuming that lapse does not occur in correspondence to the first lapse year prediction, rather in correspondence to the second one in each simulation. The Figures 41-48 refers to these new setting. As expected, the relative impact of the policyholder behavior slightly decreases for almost each parametrization, but unfortunately it is a quite marginal effect.

All in all, even if we take a look at the smallest policyholder behavior's effect on TVOG, i.e. in Figure 36 by 20% equity investment, it is above 10%, which is not negligible at all. On the other hand, some scenarios reveal an extremely great impact from surrender activity. In such cases, TVOG will be small because of favorable market conditions, but no dynamic policyholder behavior modeling is practically meaning that our TVOG estimation is zero since TVOG would come from guarantees only. And this cannot be an acceptable simplification.

8. Limitations and conclusions

Although the whole paper is based on a specific, proprietary dataset, the analysis provides us with a realistic idea of the relative impact of dynamic policyholder behavior.

In Section 5, we referred to a sort of simplified ALM model, the same as in [1]. Without a doubt, a real ALM model would be much more accurate and somewhat less questionable, because it would allow (at least theoretically) for a correct calculation of the segregated fund return as the ratio of revenue to average reserve. Actually, our model implicitly assumes that the market-value-based asset total return coincides with the balance-sheet-based segregated fund return: although it is a good proxy, they are not equivalent at all. In particular, we do not account for available-for-sale securities in portfolio, that is, all those securities (whether bonds or stocks) that the company may trade at its own discretion to realize gains or losses. In fact, we considered held-to-maturity securities only, although insurance companies generally hold available-for-sale securities as well. On the other hand, segregated fund allocations tend to be quite stable among insurance companies, in order to guarantee stable returns. This is the reason why we can accept our simplifications to the extent of this paper, avoiding a number of complications that are beyond the scope of the research.

Even if we recognize that some important variables have not been included

among the explanatory variables of the policyholder behavior (e.g. unemployment rate and policyholder's salary), our results are globally in line with those of other similar studies. Our analysis has brought out some typical risky profiles. Positive correlation between age and lapse tendency is confirmed: oldest people surrender more than younger people. At the same time, higher premiums make policyholders more prone to lapse, which is quite plausible. The performance of the contract plays an important role as well. Indeed, when the contract cannot return a yield comparable to the actual market yields, lapse is more likely. To some extent, it proves a sort of rational behavior of the policyholders since higher market yields can be either due to poor segregated fund returns (leading to an arbitrage-oriented behavior) or higher bond spreads (leading to a crisis-oriented behavior).
In the second part of the paper, we focused on the effective impact of dynamic policyholder behavior modeling on profit and TVOG estimation. In particular, the histograms in Section 7 reveal how significant the unique behavior of a policyholder can be on the profit valuation. Most of the time, indeed, the TVOG due to policyholder behavior covers the majority of the total TVOG. In other terms, TVOG calculated without dynamic policyholder behavior assumptions can extremely underestimate the total TVOG. And we should be aware of the fact that it is due to current economic conditions as much as intrinsic features of the specific policyholder. Both of them should be taken into account for a comprehensive and prudential dynamic policyholder behavior modeling.

Acknowledgements. The author wishes to thank his promoter and supervisor Prof. Susanna Levantesi for her continuous support and precious feedbacks since the very first draft of this paper.

References

[1] M. Aleandri, *Valutazione del rischio di mercato di prodotti tradizionali in gestione separata in base al nuovo regime informativo per i PRIIPs*, Dept. of Statistical Sciences at Univ. “La Sapienza”, Rapporto Tecnico n. 7, 2016.

[2] D. F. Babbel, *Asset-Liability Matching in the Life Insurance Industry* in E. Altman and I. Vanderhoof (eds.), *The Financial Dynamics of the Insurance Industry*, Irwin Professional Publishing, 1995.

[3] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[4] D. Brigo, F. Mercurio, *Interest Rate Models: Theory and Practice*, Springer, 2001.

[5] G. E. Cannon, *A Study of Persistency*, RAIA XXXVII, 1948.

[6] R. R. Cerchiara, A. Gambini, M. Edwards, *Generalized Linear Models in Life Insurance: Decrements and Risk factor analysis under Solvency II*, Giornale dell'Istituto Italiano degli Attuari, 2009.

[7] J. G. R. Crombie, K. G. Forman, P. R. Gibbens, D. C. Mason, M. D. Paterson, P. C. Shaw, J. M. G. Smart, H. Smith, C. G. Thomson, R. G. Thomson, *An Investigation into the Withdrawal Experience of Ordinary Life Business*, Transactions of Faculty of Actuaries 36, 1979.

[8] A. Dar, C. Dodds, *Interest Rates, the Emergency Fund Hypothesis and Saving through Endowment Policies: Some Empirical Evidence for the UK*, Journal of Risk and Insurance, 1989.

[9] European Commission, *QIS5 Technical Specifications*, 2010.

[10] R. E. Hoyt, *Modeling Insurance Cash Flows for Universal Life Policies*, Journal of Actuarial Practice, 1994.

[11] Y. Hwang, P. Lu, *Empirical Analysis of Surrender in the Taiwan Life Insurance Companies*, http://www.wriec.net/wp-content/uploads/2015/07/5B2_Hwang.pdf, 2014.

[12] S. Jamal, *Non-Life methodologies applied to lapse rate modeling*, Astin Colloquium, 2016.

[13] D. Kiesenbauer, *Main Determinants of Lapse in the German Life Insurance Industry*, North American Actuarial Journal, 2012

[14] C. Kim, *Policyholder Surrender Behaviors under Extreme Financial Conditions*, Korean Journal of Applied Statistics, 2010.

[15] W. Kuo, C. Tsai, W. Chen, *An Empirical Study of the Lapse Rate: The Cointegration Approach*, Journal of Risk and Insurance, 2003.

[16] Life Insurance Agency Management Association, *Factors Affecting Persistency of Orphan Business*, 1948.

[17] Life Insurance Agency Management Association, *The Persistency Raters*, 1949.

[18] X. Milhaud, S. Loisel, V. Maume-Deschamps, *Surrender triggers in life insurance: what main features affect the surrender behavior in a classical economic context?*, Bulletin Français d'Actuariat (Institute des Actuaires), 11 (22), 2011.

[19] J. F. Outreville, *Whole Life Lapse Rates and the Emergency Fund Hypothesis*, Insurance: Mathematics and Economics, 1990.

[20] F. D. Patrick, A. Scobbie, *Some Aspects of Withdrawals in Ordinary Life Business*, Transactions of Faculty of Actuaries 31, 1969.

[21] Produktinformationsstelle Altersvorsorge, *Basismodell der Produktinformationsstelle Altersvorsorge (PIA)*, http://www.produktinformationsstelle.de/assets/PIA-Kapitalmarktmodell-Basisprozesse.pdf.

[22] C. F. B. Richardson, J. M. Hartwell, *Lapse Rates*, Transactions of Society of Actuaries Vol. 3 No. 7, 1951.

[23] D. T. Russell, S. G. Fier, J. M. Carson, R. E. Dumm, *An Empirical Analysis of Life Insurance Policy Surrender Activity*, Journal of Insurance Issues, 2013.

[24] G. Shmueli, N. R. Patel, P. C. Bruce, *Data Mining for Business Intelligence*, Wiley, 2010.

[25] R. Stanton, *Rational Prepayment and the Valuation of Mortgage-Backed Securities*, Review of Financial Studies 8, 1995.