Modeling paid-ups in life insurance products for risk management

David Anaya[1] · Lluís Bermúdez[2] · Jaume Belles-Sampera[3]

Life insurance companies are subject to various risks related to universal life products. One such risk -paid-up- arises when policyholders, at some point before maturity, exercise their option to stop paying the periodic premiums initially agreed to for the life of the policy. Here, several predictive models are applied, aimed at anticipating the future state of in-force premium payment policies. This is undertaken in conjunction with balancing techniques, designed to avoid misclassification errors caused by the scarcity of paid-up events in our data. Using the findings from our predictive modeling, we initially identify certain policyholder profiles that seem less likely to paid-up premiums and consequently may be considered as potential targets for underwriting. Additionally, we delve into an essential aspect of policy design: surrender fees. Our analysis highlights a pattern where surrender fees, intended to mitigate surrender risk, may actually exacerbate the risk of policies becoming paid-up under certain circumstances.

Keywords: classification, machine learning, universal life, rare events, risk mitigation.

[1] Faculty of Economics and Business, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain. danayalu7@alumnes.ub.edu

[2] Department of Economic, Financial and Actuarial Mathematics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain. lbermudez@ub.edu

[3] Riskcenter-IREA, Grupo Catalana Occidente, Av. Alcalde Barnils, 63, Sant Cugat del Vallès, 08174 Barcelona, Spain. belles.sampera@ub.edu

1. Introduction

The purpose of this article is two-fold: firstly, to advance comprehension of optimal machine learning techniques for performing a risk analysis of paid-ups; and secondly, to discuss ways of integrating the insights gained from this analysis into risk management strategies. The main sources of these risks lie in policyholder behavior, which, in an actuarial context can be understood as the decisions that policyholders (individuals, groups, or organizations) make in the selection and utilization of benefits and guarantees embedded in life insurance policies, Campbell et al. (2014). Other risks that need to be taken into consideration include the expenses risk – i.e. the risk of not being able to meet administrative and acquisition costs – and the death risk (albeit with a relatively low impact given that the amounts insured are appreciably smaller than wealth account values).

Because the short-term impact for insurance companies of paid-ups is not as great as that of surrenders, this risk has received little attention in the literature. Indeed, unlike significant increases in surrenders rates, paid-ups would not lead to a company's short-term financial instability (Russell et al. 2013; Fier and Liebenberg 2013; Gatzert, Hoermann, and Schmeiser 2009); they would not oblige the insurer to liquidate high-yielding investments to meet policyholder requests for surrender values (Smith 1982); and they should not, in theory, have a direct impact on an insurer's reputation (Eling and Kochanski 2013).

This, however, does not mean that the paid-up risk of universal life insurance is devoid of any research interest. Indeed, such products must be valued from a long-term perspective (i.e. the whole of a policy's life), given that, for valuation purposes, companies must account not only for major short-term impacts but also for smaller, long-term, cumulative impacts. Best estimate liabilities (BEL, a Solvency II requirement) or fulfilment cash-flows (FCF, one of the IFRS17 key elements of the balance sheet) summarize all the future liabilities of an insurer at valuation date, based on assumptions regarding the future behavior of its policyholders, including those related to the exercise of their paid-up options. Inaccurate assumptions regarding paid-up probabilities adversely affect business valuations (i.e. embedded values as proposed by the CFO Forum[1] or profit testing of a company's products); and the amount of regulatory capital to be set aside as a cushion against potential insolvencies (the SCR of the Solvency II regulation)[2]; and have direct impact on the bottom-line of the profit and loss account each year under the IFRS17 rules.

Concerning investment management, deviations from anticipated patterns in earned premiums may result in unfavorable scenarios, such as the inability to procure assets when required, such as when existing assets covering liabilities reach maturity. While such occurrences are rare and typically manifest in extreme circumstances, paid-up risk is intricately linked to asset-liability risk management. Furthermore, it is plausible that a rise in paid-up premiums may serve as a key risk indicator for an elevated likelihood of surrender, leading to more adverse economic consequences as previously mentioned.

Thus, it is evident that the study of paid-up probabilities and their impact on business valuations and the accounting statements of insurance companies is of some relevance.

In common with any risk, the probability and economic impact of an event associated with paid-ups needs to be understood. Despite universal life policies typically being long-term agreements, our analysis adopts a short-term (one-year-ahead) perspective. This choice is deliberate, as it allows us to concentrate solely on the inherent policy underwriting features, without considering external factors such as market interest rates and the overall economic climate. Although these external elements could potentially influence the probability of paid-up risk, exploring their impact requires additional external data sources and the implementation of dynamic models, both of which are beyond the scope of this paper. The effectiveness of this approach will be examined in the Results section.

In what follows, as a key objective of this article, we conduct an initial analysis of predictive models for deriving paid-up probabilities and apply these well-documented approaches – including, logistic regression, individual and ensemble tree-based methods and neural networks – to calculate a desirable level of risk. One of our objectives in taking such an approach is to make machine-learning predictive tools available to actuaries and risk analysts who are responsible for conducting valuations of life companies but that are not necessarily data science experts. For this reason, a step-by-step approach has been deemed the most adequate here. Indeed, this study of paid-up probabilities in universal life products should be of interest to other financial institutions, including banks, because clear parallels can be identified with the study of loan paid-up probabilities, among others.

After concluding the paid-up risk analysis, this article aims to explore ways for incorporating the acquired insights into risk management strategies as an additional objective. This will be done through two different approaches. First, by analyzing fitted predictive models, we seek to offer insights into the key factors influencing the prediction of a paid-up event. Armed with this knowledge, we can identify policyholder profiles less prone to paid-up premiums, making them an attractive underwriting target to mitigate this risk in our portfolio. Second, by leveraging the same knowledge source, we can explore the evaluation of the policy design for such products. Specifically, we aim to examine the relationship between paid-up risk and surrender fees. This analysis aims to inform product managers about potential counter-effects that components designed to mitigate surrender risk, like surrender fees, may exert on paid-up risk.

The rest of this paper is structured as follows. Section 2 provides definitions of interest and outlines basic notions underpinning the predictive models, the cost-sensitive and resampling techniques and the performance and validation tools used in the study. Section 3 describes the dataset, while Section 4 presents the results. In Section 5, we present our conclusions and identify areas for further research.

2. Models and techniques

We focus our attention on conditional Bernoulli random variables, such as the following:

$$Y_{i,t+1} \coloneqq X_{i,t+1} | \{X_{i,t} = 0\} = \begin{cases} 0 & with\, probability \;\; 1 - \pi(t) \\ 1 & with\, probability \;\; \pi(t) \end{cases}$$

where $X_{i,t}$ indicates the premium payment state of the policy i at time t, and where the feasible states are 0 (active or negative case) and 1 (paid-up or the positive case).

That is, we are interested in the random variable of the premium payment state of policy i for period $t + 1$, given that we know that in period t the policy has an active state. The paid-up probability $\pi(t)$ is unknown but can be assumed to depend on specific features of each policy i at time t. $Y_{i,t+1}$ is the response variable in all the models proposed hereinafter; however, note that references to t or $t + 1$ should be omitted unless necessary. Thus, Y_i rather than $Y_{i,t+1}$ is the usual notation for our response variable.

2.1. Approaches to predictive modeling

The following set of predictive models is tested in our study: logistic regressions, decision trees, random forests, extreme gradient boosting, and neural networks. The first two models can be considered more *classical* approaches than the rest, but they also facilitate interpretation of the importance of the explanatory variables. In what follows, we provide a brief description of each model and a few key references.

As an extension of classical linear regression models, logistic regression is a non-linear regression technique that assumes that the expected probabilities of our binary response variable Y_i are as follows,

$$\pi_i^{lr} = \frac{e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are the n −dimensional vectors of explanatory variables for the ith policy and of regression parameters, respectively (see Henriksen, Nielsen, and Steffensen 2014).

Decision trees are models that predict the class or value of the target (categorical) variable (in our case, the value of the random variable Y_i) by learning simple decision rules (a set of conditions or if-else statements) inferred from a training dataset (Breiman et al. 1984).

The random forest algorithm, an extension of the bagging method (Breiman 1996), relies on the creation of a random subset of features of a sample or samples to create an uncorrelated forest of decision trees (Breiman 2001). This means any possible correlation between decision trees can be reduced and the errors associated with the model minimized, because while individual decision trees consider all possible feature splits, random forests only select a subset of those features.

Extreme gradient boosting (Breiman et al. 1984) is an extension of the gradient boosting trees introduced by Friedman (2001). It is a set-based learning algorithm that gives a different weight to the distribution of the training datasets at each iteration. Each iteration of boosting adds a weight to the misclassified rate (the error rate) and subtracts a weight from the correctly classified sample, thereby effectively changing the distribution of the training data (Chen and Guestrin 2016; Ramraj et al. 2016).

Finally, the neural networks algorithms (Venables and Ripley 2002) can be represented as a group of nodes interconnected by edges. The output of one node plays an input role for another node, to carry on the process according to the interconnections of the model. The nodes are usually grouped according to the transformation they perform, in a matrix known as a layer.

None of the proposed models performs a statistical classification but rather provides probabilities of observing a paid-up event. If we denote these probabilities by π_i^m (where $m \in \{lr, dt, rf, xgb, nn\}$ refers to the source model), then, once a threshold $q \in (0,1)$ is fixed, the premium payment state $\hat{y}_i^m$ assigned by model m to the ith observation is as follows:

$$\hat{y}_i^m = \begin{cases} 0 & if\ \hat{y}_i^m < q \\ 1 & if\ \hat{y}_i^m \geq q \end{cases}$$

When the data present balanced classes, as is the case here following the arrangements explained in Section 4.1, a threshold $q = 50\%$ can be used.

In this study, the caret R package (Kuhn 2008) was used to fit each of the predictive models described above.

2.2. Dealing with imbalanced data sets (rare events)

Fortunately for life insurance companies in general, paid-up events can be considered as rare events: the number of policies with a paid-up state is proportionally very small with respect to the number of active premium payment policies, leading to an imbalanced dataset.

Using an imbalanced dataset of this kind to build predictive models for the binary classification problem under study is likely to give unreliable results because the classifiers may be biased towards predicting the majority class. Thus, any information we obtain about the quality of the model would be either inaccurate or incomplete (Maalouf and Siddiqi 2014; Wallace et al. 2011). In practical applications, imbalanced datasets are usually negatively biased towards the class representative of the concept to be learned, i.e. the minority class in the dataset tends to be the class of interest. Such a situation arises because of a lack of previous experience/information about the relevant events, given that they are related to exceptional or previously unobserved cases (Seiffert et al. 2007; Weiss 2004). In other words, relevant events can be considered rare events.

To address the impact of biased classifiers on imbalanced data and mitigate the possible negative impact of rare events, two main techniques are currently employed in the literature: cost-sensitive learning (Elkan 2001; Ling and Sheng 2008; Thai-Nghe, Gantner, and Schmidt-Thieme 2010; Khan et al. 2017) and resampling (Estabrooks, Jo, and Japkowicz 2004; Chawla et al. 2002).

Cost-sensitive learning techniques consider the costs of misclassification of all the different classes. When dealing with imbalanced problems, positive (scarce) instances have to be over-weighted, to some degree, vis-à-vis negative instances. Here, this is achieved by forcing the cost of misclassifying a positive instance to be higher than the cost of misclassifying a negative one. Given the cost matrix, an evaluated element is classified into the class with the lowest expected cost, in accordance with the minimum expected cost principle (Elkan 2001; Sun, Wong, and Kamel 2009; He and Garcia 2009; López et al. 2013).

Resampling techniques can be classified into three main groups: undersampling methods, which create a subset of the original dataset by eliminating majority class instances; oversampling methods, which create a superset of the original data set by replicating some minority class instances or creating new instances from existing ones; and, finally, hybrid methods – such as the SMOTE (Chawla et al. 2002) and ROSE (Menardi and Torelli 2014) algorithms. The SMOTE algorithm draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbours in the feature space. The ROSE algorithm uses smoothed bootstrapping to draw artificial samples from the feature space around the minority class.

2.3. Performance and validation metrics and tools

To measure the performance of the classification models and to validate the results obtained, a set of metrics are considered (see Table 2). Most of these metrics are based on the confusion matrix associated with each model, which for a binary classification model such as the paid-up model under study here can be represented as in Table 1. Recall that in our case the negative class is the class of policies with an active premium payment state and the positive class is linked to the set of policies with a paid-up state.

In addition to the metrics provided in Table 2, the receiver operating characteristic (ROC) curve – a graphical representation of the ability of the binary classification model to provide correctly classified positive cases – is also considered as a performance evaluation tool. The area under the ROC curve (AUC) is a summary statistic of the information provided by the ROC curve about the classification performance of the model under analysis. It is expressed on a scale of 0.5 to 1, where AUC = 0.5 represents the performance of a random classifier (no discriminatory ability) and AUC = 1 is a perfect classifier.

3. Data

All data related to this study come from a life and non-life insurance company operating in the Spanish insurance market but note that several transformations/modifications have been made to these raw data so as not to compromise confidential information.

The data refer to life insurance policies sold between 2018 and 2020. More specifically, the policies are universal life products, and our data comprise a detailed description of policy features – including premium payment status – and policyholder characteristics. All policies analyzed in the portfolio are periodic premium payment policies.

We set two timeframes (2018–2019 and 2019–2020), in order to predict the one-year premium payment probability. Policies associated with a particular year in our database correspond to those policies in force as of December 31st of that same year. For the purposes of this study, for the first timeframe, we select those active premium payment policies from 12/31/2018 that were in force as of 12/31/2018 and also as of 12/31/2019 (although, obviously, at this latter date the premium payment states of the policies are no longer restricted to being active). The total number of policies in the 2018–2019 timeframe is 45,227, while the subset of interest (i.e. policies with a paid-up premium as of 12/31/2019) represents 3.68% (1,665 policies). For the second timeframe, we follow the same rationale to filter the database: thus, we have active premium payment policies in force at the end of 2019, which are still in our insurance portfolio at the end of 2020.

The 2018–2019 timeframe was used to build the models and to analyze their performance and the accuracy of their results (see Section 4). Of these observations, 80% were employed for training and the remaining 20% for testing. The 2019–2020 timeframe was used as an additional test set, to check whether the predictive models generate accurate results for a different timeframe.

The Table 3 contains a brief definition of each of the proposed variables. A minimum-maximum scaling was applied to adjust the value of the continuous variables in Table 4 for each policy in the range of $[-1{,}1]$. This was done to scale both the training (main) and testing sets with the same mean and standard deviation and, by so doing, avoid problems related to the magnitudes of the variables, given that several of them present different ranges of values. In addition, it allows us to adjust the data for higher throughput for machine learning models, especially those based on neural networks or heavy workloads. In parallel, the categorical variables shown in Table 5 were also pre-processed before incorporation into the runs of each model by applying the one-hot encoding technique. This creates new binary columns composed of vectors with values 0 and 1 representing each of the categories of each variable, where if the value is true, it corresponds to 1, and if the value is false, it corresponds to 0, for the entire dataset.

It should be pointed out that the four values taken by the aggprod variable in our dataset (AG_1, AG_2, AG_3 and AG_4) were initially proposed by actuaries in the company of reference. Policies in the dataset were assigned to each group depending on just one characteristic: the overall penalty rate applied to paid premiums if the policyholder exercises a partial or total withdrawal option at any time before a pre-established term (typically 10 years). Specifically, AG_1 refers to products with an overall penalty rate of 25%, AG_2 to products with an overall penalty rate of 45%, AG_3 to products without a penalty rate, and AG_4 to products with an overall penalty rate of 10%. However, the AG_3 group present other particular characteristics: i.e. most unit-linked products and most products with tax advantages for the policyholder are included in this group. As we explain in Section 4.3, these characteristics are relevant when analyzing the results obtained.

Actual values of the target variable for the ith policy in the dabaset ($Y_{i,t+1}$) are added under the name bin_resp to the dataset. Thus, if we know that the ith policy has paid-up premiums in $t+1$, the bin_resp value is set at 1, and 0 otherwise. Tables 4 and 5 show the descriptive statistics for the numerical and categorical explanatory variables, both for the subset of paid-up policies (i.e. our positive cases) and the complementary subset of policies (those whose premium payment remains active) at the end of the 2018-2019 timeframe.

In Table 4, it seems that values of wealth funds (res) and values of total insured amounts to be paid in case of death (cap) are greater for active premium payment policies at the end of the 2018-2019 timeframe than the respective values for paid-up policies. In the case of the categorical variables, paid-up policies present two main characteristics: all of them have a monthly premium frequency and all of them have constant premiums. In contrast, active payment policies present more heterogeneous payment frequencies and types of premium increments. Additionally, a higher percentage of AG_3 policies are to be found in paid-up than in active payment policies (see Table 5).

4. Results

4.1. Balancing dataset

As discussed, the distribution of cases in our dataset is more biased towards 0 (active or negative; 96.32%) than towards 1 (paid-up or positive; 3.68%). Thus, unless the data are treated in some way before building predictive models, the latter will tend to underestimate the probability of the occurrence of paid-up events. Moreover, the default threshold of 50% used to classify predictions of active and paid-up policies will generate very few, if any, cases classified as paid-up.

Here, several runs of each model were performed and evaluated adopting the balancing methods discussed in Section 2.2 in what is known as a repeated k-fold cross-validation (where k=10 and we repeated the validation five times). The performance metrics considered for each validation were the AUC, the sensitivity and specificity rates, and the balanced accuracy rate (as defined in Table 2). Given that we obtained various values for these metrics for each model and balancing technique considered, we summarize them using a simple average.

For illustrative purposes, here, we report only those results for logistic regression and eXtreme Gradient Boosting (XGB) (see Tables 5 and 6).

When using the original imbalanced dataset with all models, the sensitivity values are high, while the specificity rates drop, ranging between 42.50% (with logistic regression) and 62.67% (with XGB). This indicates the need to consider some of the balancing techniques that might address this issue. As can be seen, for all models and balancing techniques, remarkable improvements were achieved in terms of specificity rates, albeit at the cost of slightly deteriorated sensitivity values. For all models considered here, a noticeable improvement was detected in the balanced accuracy *(bacc)* rates when using any of the balancing techniques described in Section 2.2.

In addition to comparing the relative performance metrics for each balancing technique applied to the predictive models, we identified the best balancing technique by ranking their overall performance from highest balanced accuracy rate to lowest (see, for example, final column in Tables 5 and 6). The technique with the best overall ranking among all the predictive models was then selected as the technique that best fits our purposes herein.

Thus, we selected the undersampling technique: ranked first for decision tree, random forest and XGB models, and second for logistic regression. Only in the case of neural networks did its accuracy rate fall (ranked 4th). In addition to these quantitative criteria, other rationale for using this technique is that it makes use of all the available information about positive cases in our training dataset and that it is relatively straightforward to explain.

4.2. Predictive capacity of the models

We next analyzed the predictive capacity of the models and selected the one that performed best. Again, we worked with a 10-fold cross-validation repeated five times, applied in this instance only to the training dataset after applying the undersampling technique. To assess the robustness of the models, we analyzed the data from the training set using the previously defined metrics (and not a single value as in Tables 6 and 7, but several statistics for each: percentiles and mean values).

Table 8 presents summary statistics for the set of AUC, sensitivity, and specificity metrics. The dispersion of the three metrics is low for all models, except for the decision tree. If we consider the AUC as the main decision metric, then the random forest model is optimal, followed by the XGB and neural network models. In the case of sensitivity, XGB seems to classify true positives more accurately than the rest, while in that of specificity, the logistic regression model has the best true negative classification rate.

Below, we comment in detail on the predictions and risk management implications of three of the five models. The random forest model is the best performing, with a mean AUC of 97.92%, a mean sensitivity rate of 88.82%, and a mean specificity rate of 98.06% (see 8), followed by the neural network model with means of 97.23%, 87.41%, and 97.61%, respectively; however, given the computational burden involved in performing the calculations, the latter model is discarded as an appropriate model for our purposes here.

We opted to replace it with the XGB with respective means of 97.15%, 89.50%, and 93.66%. Finally, the third model selected for further study is the logistic regression model, with means of 96.73%, 86.01%, and 99.67%, respectively. We consider this last model to be of interest based on the strength of its true negative classification rate, and the simplicity with which it can be executed, and its results interpreted. Moreover, by being a *classical* predictive model, it provides a good counterweight to what might be considered the more *trending* predictive models of random forest and XGB.

To assess the predictive capacity of these three models, we analyze the results obtained when contrasting the predictions obtained by each model with the actual values. Table 8 shows the confusion matrices of these three models when applied to the 2018–2019 timeframe testing dataset (with 9,045 observations), as well as several performance metrics for each of the models.

The confusion matrices shown in Table 8 are better visualized in Figure 1. Here, the model predictions are plotted as follows: predictions over the 50% threshold are positive cases for the models (predicted paid-up premiums) and those below are negative cases (active premium payment policies). If the prediction coincides with the actual premium payment status in the dataset, then the point is colored green (true prediction), otherwise, it is colored red (false prediction).

In summary, all three predictive models perform well; however, the confusion matrices point to significant levels of false positives in each, especially in the logistic regression model. In other words, a relevant number of policies are incorrectly predicted to be paid-up premiums. A careful inspection of the characteristics of these false positive policies shows that positive cases (both TP and FP) have constant monthly premiums for all models. In contrast, only a negligible percentage of negative cases (both TN and FN) feature constant monthly premiums.

To some extent, this reflects the fact that all the paid-up cases in the original dataset feature constant monthly premiums (see Table 5; Paid-up policies) and, hence, the models capture this reality by classifying the vast majority of constant monthly premium policies as paid-up cases. However, this does not coincide with the data, since there are a significant number of constant monthly premium policies with an active premium payment status (see Table 5; Policies paying premiums).

To address this and, in so doing, to enhance the balance between the models' sensitivity and specificity, we propose a two-step modeling approach. First, based on empirical evidence, we can predict that all policies without constant monthly premiums are active premium payment policies. This holds true for the 2018–2019 timeframe dataset and for all but 3 of 46,243 cases in the 2019–2020 timeframe dataset, which is bearable. Second, we repeat the above modeling procedure but, on this occasion, we limit it to the subset of constant monthly premium policies from the 2018–2019 timeframe dataset. Figure 2 highlights the marked reduction of false cases (red points) compared to those observed in Figure 1. Moreover, these false cases are now more balanced between positive and negative cases.

Similar conclusions can be drawn from Table 10, where the same performance metrics reported in Table 8 are now shown for the two-step approach. A comparison of these two tables shows that the two-step approach improves substantially the overall proportion of correctly classified predictions (accuracy rate) for all models. This enhanced predictive performance is explained by the improvement in the specificity rate, i.e. the proportion of active payment premium cases correctly classified by the models. The models no longer consider all constant monthly premiums policies as being paid-up and some of them are now correctly predicted as being active payment premiums. However, this enhancement in specificity is obtained at the cost of a slight decrease in sensitivity, with a larger number of paid-up cases being classified as active payment premiums.

Finally, we conducted an out-of-sample validation using the 2019–2020 timeframe as our testing dataset and fitting the three models with the two-step approach. Table 11 reports the same performance metrics as those employed previously. As is evident, the performance of the models when applied to this dataset (comprising 46,243 elements, none of which were used to calibrate the models) is very similar to that of the models when predicting the pure testing dataset (i.e. the 2018–2019 timeframe testing dataset).

These results serve to validate even further, were this possible, the models' predictive performance. In parallel, one might question the suitability of adopting a short-term perspective in this instance, given the long-term nature of universal life contracts. Nevertheless, as we will explore shortly, this assumption proves satisfactory for achieving the outlined objectives in the Introduction. This is because our emphasis is on internal factors, typically maintaining stability over time and deemed controllable by the insurer. By exploring these internal factors influencing paid-up events, we can enhance the customization of our risk management strategies.

4.3. Risk management implications

Our application leaves little doubt that incr (type of premium increments) and freq (frequency of premium payments) are the most relevant variables (without there being any need to employ a variable importance methodology). As discussed, in the 2018–2019 timeframe dataset, only policyholders with constant monthly payments exercised their paid-up option. Thus, to reduce paid-up risk, the underwriting of universal life products with increasing premiums and payments on at least a quarterly basis should be encouraged. On the one hand, policyholders with increasing premiums are much less likely to stop their premium payments given that when they take out a policy with increasing premiums, they show no indication of wanting to reduce them to their minimum expression. On the other hand, policyholders with monthly payments are more likely to exercise their paid-up option either because they are more aware of what they are paying (given they receive a monthly reminder) or because they do not have the financial capacity to meet a higher premium frequency.

From Figure 3 and Table 12, we can draw the following conclusions. First, the most important variable in tree-based models is res (current value of the fund), while its logistic regression coefficient indicates that the larger the fund value, the less likely a policyholder is to stop paying premiums. This is reasonable considering that policyholders with larger funds are likely to have a higher degree of confidence in the product. Second, gender does not seem to have an influence on paid-up probability. Third, the rest of the continuous variables (prem - initial premium, cap - death capital, finalpp - years until final scheduled premium, age -current insured's age, and loy -policy in force in years) in tree-based models appear to have a similar and considerable influence on the outcome. Here again, for each of these variables, the logistic regression coefficients indicate that a larger value is associated with a lower paid-up probability. Considering these findings, one might view customers aged 50 and above, providing a substantial initial premium and funds, as a preferred underwriting focus. This may include customers who transfer significant initial funds from another entity.

When evaluating our universal life policy design, our primary concern is the impact of incorporating a surrender fee into the paid-up probability. From a risk management perspective, it's crucial to note that measures aimed at mitigating one risk (like surrender risk) could potentially exacerbate another risk (like paid-up risk). According to the fitted models, savings products belonging to group AG_3 (without penalty rates for withdrawals) are associated with a marked deduction in the likelihood of a policyholder exercising their paid-up option. Meanwhile, savings products belonging to group AG_1, AG_2, or AG_4 (incorporating surrender fees) do not seem to have an influence on paid-up probability. This might, at first glance, appear counter-intuitive, and merits further examination and analysis.

To further our knowledge of policyholder profiles, we examined the main characteristics of the policies in the test set identified as TPs, TNs, FPs, and FNs by all the models (two-step approach) when applied to the 2018–2019 timeframe. This analysis provides better insights into the variables that determine whether a policyholder is identified as TP or TN, and a broader picture of what is happening when the models fail to predict positive or negative cases (i.e. FP vs. TP, or FN vs. TN).

First, contrary to previous comments, it is evident that the inclusion of surrender fees (such as in AG_1, AG_2, or AG_4) does indeed affect paid-up probabilities. When a product incorporates a surrender fee (to manage its surrender risk), it either amplifies the paid-up risk for policyholders with low premiums and reserves (TP cases) or diminishes it for those with high premiums and reserves (TN cases). This outcome might prompt risk managers to reconsider the rules pertaining to surrender fees. Second, for products lacking surrender fees (such as AG_3), TN instances predominantly involve policyholders making moderate to low premium payments over a lengthy period (e.g., a 40-year-old individual saving for retirement) while benefiting from tax advantages.

From the perspective of paid-up risk management, this profile (policyholders that aim to save wealth in the long-term, but who seek some tax advantages in so doing) can be considered a target for a company to address this risk. Conversely, TP cases within the AG_3 group typically involve policyholders nearing retirement age, making substantial premium payments, and holding significant reserves. In such scenarios, reconsidering surrender fee policies or introducing new paid-up fees could mitigate this risk.

5. Conclusions

Through the development of predictive models, the aim is to not only forecast these probabilities but also to investigate ways for integrating the resulting insights into risk management practices, such as underwriting or policy design for universal life products.

As paid-up events are rare in our data, balancing techniques were required as a prior step to the application of predictive modelling. Here, we have applied a set of cost-sensitive and resampling techniques precisely to avoid biased classifiers and a marked improvement in balanced accuracy rates was achieved as a result. More specifically, undersampling was identified as the optimum method because of both its simplicity and because it provided the best-balanced accuracy rates. Five predictive models – logistic regression, decision tree, random forest, XGB, and neural network – were then applied to the balanced dataset. However, only three were selected for further analysis, the decision tree model being discarded for presenting, by far and away, the worst performance metrics together with the neural network model, which despite having similar metrics to those selected, presented an excessive computational burden.

The three models selected performed well, especially when adopting a two-step approach to their application. This procedure was required because all the positive cases in our dataset presented the same two main characteristics: namely, a monthly payment frequency and the fact that their premiums are constant. Random forest and XGB models, with a balanced accuracy of around 85%, performed better than logistic regression (c. 78%). Of significance is the fact that largely the same performance metrics were recorded when the models were used to predict paid-up events in a different timeframe dataset from that from which they were derived.

A detailed analysis of the results provides a series of insights that can help improve paid-up risk management allowing companies to focus on those endogenous risk factors on which they can act. First, what emerges from a simple inspection of the data is that the underwriting of savings products with a non-constant monthly premium profile should be fostered. Second, as this recommendation is not always readily applied, companies can reduce the paid-up probability of constant monthly premium policies by fostering the underwriting of savings products with a long-term horizon combined with tax advantages (the case of most of the policies included in the AG_3 category). More specifically, according to the corresponding coefficient obtained in the logistic regression, the odds ratio (i.e. the probability that paid-up occurs divided by the probability that it does not occur) is reduced by 63% when the policyholder purchases a savings product with a long-term horizon and tax advantages.

We conducted a number of calculations to estimate the potential short-term economic benefit of boosting the underwriting of products with a long-term horizon and tax advantages. For each predictive model and for the constant monthly premium policies included in our 2018–2019 timeframe dataset, expected paid-up premiums are obtained in two distinct scenarios: the first, a base scenario in which all the policies remain as they are and, the second, a scenario that requires all policies to be of type AG_3. To compare the two types of expected paid-up premium, we made an approximation of the best improvement on earned premiums. Here, the improvement rates for active premiums were found to be +29% (LR), +10% (RF), and +14% (XGB). In other words, given that the total amount of active premiums in our dataset stands at around 8 million monetary units, it is estimated that at least 800 thousand additional premiums could be earned, one year ahead, by fostering the underwriting of AG_3 type products. If the net return that the company is able to obtain from each premium is about 3%, then a net income of 24 thousand monetary units before tax could be achieved by implementing these management actions. As was stressed in the introduction to this paper, the economic impact of paid-up risk in the short term is much less than that attributed to surrender risk; nonetheless, if these reductions in paid-up rates are maintained over time, then the economic benefit in the long-term could be increased considerably.

In addition to its application in revisiting the underwriting policy, conducting such an analysis can offer valuable insights for reviewing our policy design. Specifically, our analysis has centered on evaluating the consequences of integrating a surrender fee, intended to mitigate surrender risk, into the paid-up probability, thus impacting our paid-up risk management strategy. In essence, our findings indicate that the inclusion of a surrender fee tends to either exacerbate the paid-up risk for policyholders with low premiums and reserves or mitigate it for those with high premiums and reserves. This observation may urge risk managers to reevaluate the guidelines concerning surrender fees.

Finally, a collateral data management strategy, aimed at improving the predictive power of the models, requires a redefinition of the categories of the aggprod variable. For instance, here, we have detected at least three different subgroups (long-term products with tax advantages; short-/mid-term products with tax advantages; and unit-linked policies) among AG_3 type products that present different policyholder behavior with respect to exercising the paid-up option. Including all of them in the same category may well negatively affect the performance of the models.

There is obvious scope for further research on paid-up risk and this would involve, first and foremost, relaxing each of the restrictions we imposed herein: that is, changing the short-term to a long-term perspective with respect to paid-ups, considering economic impact and not only probabilities, and taking exogenous and not solely endogenous factors into account. This expanded approach would be particularly beneficial for assessing the economic repercussions on the income statement stemming from this risk, especially in the context of regulatory valuation.

6. Notes

1. See https://www.cfoforum.eu/embedded_value.html
2. As contributions to the SCR are usually the difference between a stressed BEL and the central one, although both depend on several valuation assumptions, such as paid-up probabilities.

7. References

Breiman, L. (1996). “Bagging Predictors.” In: *Machine Learning* 24: 123–40.

— (2001). “Random Forests.” In: *Machine Learning* 45 (1): 5–32.

Breiman, L. et al. (1984). “Classification and Regression Trees.” In: *Chapman and Hall/CRC*.

Campbell, J. et al. (2014). “Modelling of Policyholder Behaviour for Life Assurance and Annuity Products.” In: *Society of Actuaries*.

Chawla, N. V. et al. (2002). “SMOTE: Synthetic Minority over-Sampling Technique.” In: *Journal of Artificial Intelligent Research* 16: 321–57.

Chen, T. and C. Guestrin (2016). “Xgboost: A Scalable Tree Boosting System.” In: *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, no. 2, pp. 785–794.

Eling, M. and M. Kochanski (2013). “Research on Lapse in Life Insurance: What Has Been Done and What Needs to Be Done?.” In: *The Journal of Risk Finance*.

Elkan, C. (2001). “The Foundations of Cost-Sensitive Learning.” In: *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 973–78.

Estabrooks, A., T. Jo, and N. Japkowicz (2004). “A Multiple Resampling Method for Learning from Imbalanced Data Sets.” In: *Computational Intelligence* 1 (20), pp. 18–36.

Fier, S. G., and A. P. Liebenberg (2013). “Life Insurance Lapse Behavior.” In: *North American Actuarial Journal* 2 (17), pp. 153–167.

Friedman, J. H. (2001). “Greedy Function Approximation: A Gradient Boosting Machine.” In: *The Annals of Statistics* 29.5, pp. 1189–1232.

Gatzert, N., G. Hoermann, and H. Schmeiser (2009). “The Impact of the Secondary Market on Life Insurers’ Surrender Profits.” *Journal of Risk and Insurance* 4 (76), pp. 887–908.

He, H., and E. A. Garcia (2009). “Learning from Imbalanced Data.” In: *IEEE Transactions on Knowledge and Data Engineering* 9 (21), pp. 1263–84.

Henriksen, L. F. B., J. W. Nielsen, and M. Steffensen (2014). “Markov Chain Modeling of Policyholder Behavior in Life Insurance and Pension.” In: *European Actuarial Journey* (4), pp. 1–29.

Khan, S. H. et al. (2017). “Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data.” In: *IEEE Transactions on Neural Networks and Learning Systems* 8 (29), pp. 3573–3587.

Kuhn, M. (2008). “Building Predictive Models in R Using the Caret Package.” In: *Journal of Statistical Software* 28 (5), pp. 1–26.

Ling, C. X., and V. S. Sheng (2008). “Cost-Sensitive Learning and the Class Imbalance Problem.” In: *Encyclopedia of Machine Learning*, pp. 231–235.

López, V. et al. (2013). “An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.” In: *Information Sciences*, no. 250, pp. 113–141.

Maalouf, M., and M. Siddiqi (2014). “Weighted Logistic Regression for Large-Scale Imbalanced and Rare Events Data.” In: *Knowledge-Based Systems* (59), pp. 142–148.

Menardi, G., and N. Torelli (2014). “Training and Assessing Classification Rules with Imbalanced Data.” In: *Mining and Knowledge Discovery* 1 (28), pp. 92–122.

Ramraj, S. et al. (2016). “Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets.” In: *International Journal of Control Theory and Applications* 40 (9).

Russell, D. T. et al. (2013). “An Empirical Analysis of Life Insurance Policy Surrender Activity.” In: *Journal of Insurance Issues*, pp. 35–57.

Seiffert, C. et al. (2007). “Mining Data with Rare Events: A Case Study.” In: *19th International Conference on Tools with Artificial Intelligence*, pp. 132–139.

Smith, M. L. (1982). “The Life Insurance Policy as an Options Package.” In: *Journal of Risk and Insurance*, 583–601.

Sun, Y., A. K. Wong, and M. S. Kamel (2009). “Classification of Imbalanced Data: A Review.” In: *International Journal of Pattern Recognition and Artificial Intelligence* 4 (23), pp. 687–719.

Thai-Nghe, N., Z. Gantner, and L. Schmidt-Thieme (2010). “Cost-Sensitive Learning Methods for Imbalanced Data.” In: *In The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Venables, W. N., and B. D. Ripley (2002). “Modern Applied Statistics with S.” In: *Springer* 4.

Wallace, B. C. et al. (2011). “Class Imbalance, Redux.” In: *IEEE 11th International Conference on Data Mining*, pp. 754–763.

Weiss, M. (2004). “Mining with Rarity.” In: *ACM SIGKDD Explorations Newsletter* 6.1, pp. 7–19.

A. Tables

***Table 1**: Confusion matrix for the binary classification problem associated with paid-up risk.*

		Actual	
		Active	Paid-up
Predicted	Active	True Negative (TN)	False Negative (FN)
	Paid-up	False Positive (FP)	True Positive (TP)

***Table 2**: Performance and validation metrics considered for the paid-up risk models under analysis.*

Metric	Expression	Concept
Accuracy rate	$acc = \frac{TP + TN}{TP + TN + FP + FN}$	Overall proportion of correctly classified predictions.
Misclassification rate	$misc = \frac{FP + FN}{TP + TN + FP + FN}$	Overall proportion of misclassified predictions. Note that $misc = 1 - acc$.
Sensitivity rate	$r = \frac{TP}{TP + FN}$	Proportion of actual positive cases correctly classified by the model (also known as recall or recovery rate)
Specificity rate	$s = \frac{TN}{FP + TN}$	Proportion of actual negative cases correctly classified by the model.
Precision rate (PPV)	$p = \frac{TP}{TP + FP}$	Proportion of correctly classified positive cases among predicted positive cases.
Negative predicted rate (NPV)	$npv = \frac{TN}{FN + TN}$	Proportion of correctly classified negative cases among predicted negative cases.
Balanced accuracy rate	$bacc = \frac{r + s}{2}$	Simple average of sensitivity and specificity rates.
Youden's Index	$YI_q = r_q + s_q - 1$	Evaluation of the discriminative power of the test by combining sensitivity (r_q) and specificity (s_q). The q is the threshold

Metric	Expression	Concept
		considered.
F1 score	$F_\beta = (1+\beta^2)\frac{pr}{\beta^2 p + r}$	Evaluation of positive classification ability of the model. It combines sensitivity (r) and precision (p), and balances the relative importance of each by means of the β parameter.

***Table 3**: Definition of selected variables.*

Variable	Definition
bin_resp	Premium payment status for the period t+1
finalpp	Years remaining until final payment of the contract premium
cap	Total sum insured in case of death, in terms of monetary unit
res	Current value of the fund, in terms of monetary unit
prem	Initial premium paid at inception, in terms of monetary unit
age	Current age of the insured
loy	Number of years the policy has been in force
gender	Gender of the insured
aggprod	Grouping of the different savings products in the portfolio
incr	Type of premium increments (constant, geometric and arithmetic)
freq	Frequency of premium payment (monthly, yearly and other)

***Table 4**: Descriptive statistics of continuous variables at the end of the 2018-2019 time-frame.*

Paid-up policies

Variable	Mean	Min.	Pctl.25	Median	Pctl.75	Max.
finalpp	13.96	0	6.00	9.00	18.00	77.00
cap	3,789.90	0	887.90	2,663.70	5,327.40	76,843.90
res	7,514.70	1.23	663.39	2,705.90	7,143.05	253,611.19
prem	1,204.10	0	532.70	745.80	1,164.20	22,197.50
age	48.21	19.00	39.00	47.00	57.00	81.00
loy	5.03	1	2.00	3.00	7.00	35.00

Policies paying premiums

Variable	Mean	Min.	Pctl.25	Median	Pctl.75	Max.
finalpp	14.60	0	6.00	10.00	19.00	80.00
cap	4,906.50	0	887.90	5,327.40	5,488.80	155,382.50
res	11,154.00	23	2,104.00	4,719.00	11,240.00	720,660.00
prem	1,032.52	36.34	532.74	710.32	1,065.48	31,964.40
age	48.07	15.00	40.00	47.00	56.00	90.00
loy	5.90	1	2.00	4.00	8.00	35.00

Table 6*: Pre-processing imbalanced dataset by methods and metrics - Logistic regression.*

Resampling	**AUC**	**Sensitivity**	**Specificity**	***bacc***	**Rank by *bacc***
Original	96.61%	99.38%	42.50%	70.94%	6th
Weighting	96.82%	85.93%	99.62%	92.78%	3rd
Undersampling	96.73%	86.01%	99.67%	92.84%	2nd
Oversampling	96.82%	85.94%	99.58%	92.76%	4th
SMOTE	96.80%	87.34%	97.50%	92.42%	5th
ROSE	96.73%	85.86%	99.86%	92.86%	1st

Table 7*: Pre-processing imbalanced dataset by methods and metrics - XGB.*

Resampling	**AUC**	**Sensitivity**	**Specificity**	***bacc***	**Rank by *bacc***
Original	98.30%	99.50%	62.67%	81.09%	6th
Weighting	96.95%	86.70%	92.82%	89.76%	2nd
Undersampling	97.15%	89.50%	93.67%	91.59%	1st
Oversampling	98.41%	98.17%	78.05%	88.11%	3rd
SMOTE	98.39%	99.31%	65.56%	82.44%	5th
ROSE	95.85%	65.64%	100.00%	82.82%	4th

***Table 8**: Statistics for performance metrics in each model [Undersampling].*

AUC

Model	Min.	Pct25	Median	Mean	Pct75	Max
Logistic regression	96.06%	96.43%	96.73%	96.73%	96.98%	97.69%
Decision tree	92.44%	95.10%	96.29%	95.74%	96.81%	97.33%
Random forest	97.38%	97.71%	97.87%	97.92%	98.11%	98.58%
XGB	96.42%	96.90%	97.18%	97.15%	97.38%	98.01%
Neural network	96.64%	96.95%	97.15%	97.23%	97.46%	98.12%

Sensitivity

Model	Min.	Pct25	Median	Mean	Pct75	Max.
Logistic regression	84.71%	85.56%	86.03%	86.01%	86.31%	87.63%
Decision tree	84.68%	86.41%	87.66%	87.76%	89.22%	92.22%
Random forest	87.37%	88.31%	88.75%	88.82%	89.30%	90.36%
XGB	87.78%	89.20%	89.54%	89.50%	89.87%	91.19%
Neural network	86.08%	86.93%	87.36%	87.41%	87.83%	89.58%

Specificity

Model	Min.	Pct25	Median	Mean	Pct75	Max.
Logistic regression	97.74%	99.25%	100.00%	99.67%	100.00%	100.00%
Decision tree	89.55%	95.70%	98.13%	97.48%	99.25%	100.00%
Random forest	94.74%	97.18%	97.76%	98.06%	99.06%	100.00%

AUC						
XGB	89.47%	91.92%	94.01%	93.66%	94.78%	98.50%
Neural network	91.00%	96.45%	98.13%	97.61%	99.25%	100.00%

***Table 9**: Performance metrics of selected models. 2018-2019 timeframe testing dataset.*

	Logistic regression		**Random forest**		**XGB**	
	Actual(0)	Actual(1)	Actual(0)	Actual(1)	Actual(0)	Actual(1)
Predicted(0)	7,421	0	7,634	3	7,631	7
Predicted(1)	1,291	333	1,078	330	1,081	326
Sensitivity	100.00%		99.10%		97.90%	
Specificity	85.18%		87.63%		87.59%	
NPV	100.00%		99.96%		99.91%	
PPV	20.51%		23.44%		23.17%	
Youden’s Index	85.18%		86.73%		85.49%	
F1-Score	34.03%		37.97%		37.62%	
Accuracy	85.73%		88.05%		87.97%	
Balanced Accuracy	92.59%		93.36%		92.75%	

***Table 10**: Performance metrics of selected models. 2018-2019 timeframe testing dataset (Two-step approach).*

	Logistic regression		**Random forest**		**XGB**	
	Actual(0)	Actual(1)	Actual(0)	Actual(1)	Actual(0)	Actual(1)
Predicted(0)	8,450	139	8,544	95	8,507	80
Predicted(1)	262	194	168	238	205	253
Sensitivity	58.26%		71.47%		75.98%	
Specificity	96.99%		98.07%		97.65%	
NPV	88.10%		98.90%		99.07%	
PPV	42.54%		58.62%		55.24%	
Youden’s Index	55.25%		69.54%		73.62%	
F1-Score	59.40%		73.61%		70.93%	
Accuracy	95.57%		97.09%		96.85%	
Balanced Accuracy	77.63%		84.77%		86.81%	

***Table 11**: Performance metrics of selected models. 2019-2020 timeframe dataset (Two-step approach).*

	Logistic regression		**Random forest**		**XGB**	
	Actual(0)	Actual(1)	Actual(0)	Actual(1)	Actual(0)	Actual(1)
Predicted(0)	43,295	863	45,287	657	43,236	585
Predicted(1)	1,133	1,071	719	1,277	990	1,349
Sensitivity	55.38%		66.03%		69.75%	
Specificity	97.45%		98.44%		97.76%	
NPV	98.05%		98.57%		98.67%	
PPV	48.59%		63.98%		57.67%	
Youden’s Index	52.83%		64.47%		67.51%	
F1-Score	64.98%		77.59%		72.80%	
Accuracy	95.69%		97.13%		96.59%	
Balanced Accuracy	76.41%		82.23%		83.76%	

***Table 12**: Logistic regression coefficients (Constant monthly premium policies)*

Coefficient	Estimate	Std. Error	z value	Pr(>$\|z\|$)	
(Intercept)	0.04016	0.08852	0.454	0.650071	
finalpp	-0.35156	0.10264	-3.425	0.000614	***
cap	-0.12288	0.09995	-1.229	0.118916	
res	-0.17535	0.09477	-1.850	0.064275	.
prem	-0.13663	0.09334	-1.464	0.143272	
age	-0.16144	0.09422	-1.713	0.086630	.
loy	-0.20724	0.10853	-1.910	0.056187	.
gender_Female	-0.12960	0.08787	-1.475	0.140232	
aggprod_AG_2	0.04172	0.12822	0.325	0.744877	
aggprod_AG_3	-0.98607	0.14004	-7.041	1.91E-12	***
aggprod_AG_4	0.06962	0.12293	0.566	0.571160	

B. Figures

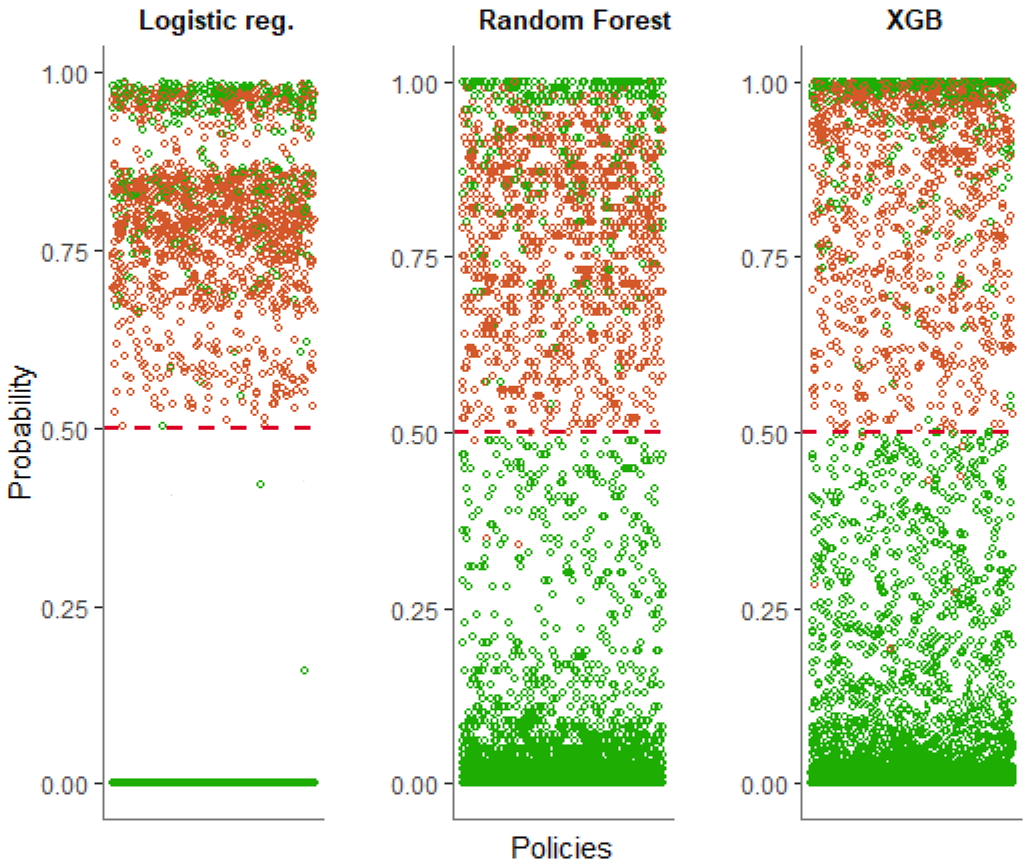


Figure 1*: Probabilities estimated by the three models on the 2018-2019 timeframe testing dataset. In green, policies correctly classified (TP above 50%, TN under 50%). In red, policies incorrectly classified (FP above 50%, FN under 50%).*

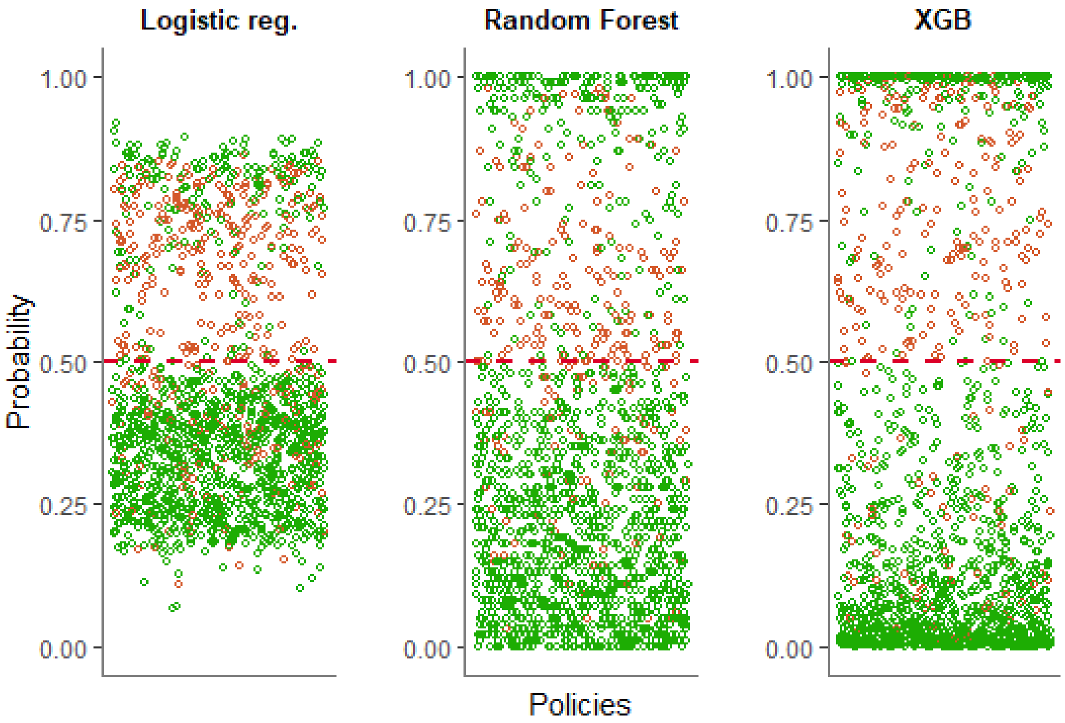
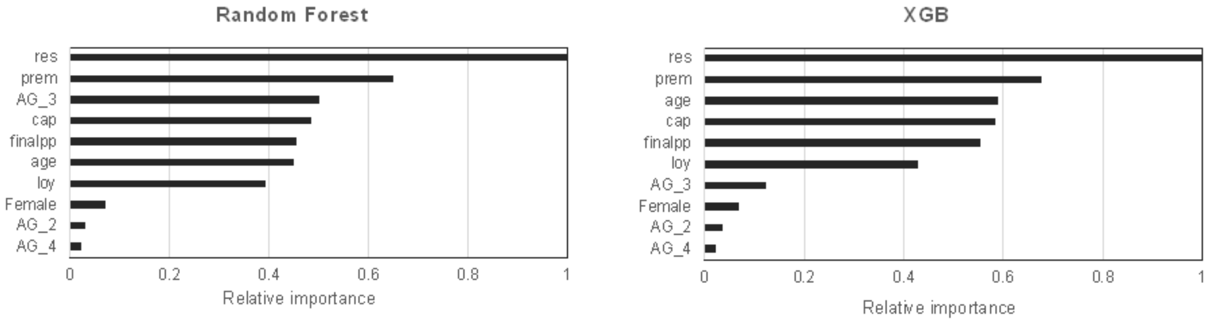


***Figure 2**: Probabilities estimated by the three models in the subset of constant monthly premium policies in the 2018-2019 timeframe testing dataset. In green, policies correctly classified (TP above 50%, TN under 50%). In red, policies incorrectly classified (FP above 50%, FN under 50%).*

***Figure 3**: The relative importance of the variables for the predictions of the two tree-based models (Constant monthly premium policies)*


Random Forest
res
prem
AG_3
cap
finalpp
age
loy
Female
AG_2
AG_4
Relative importance
XGB
res
prem
age
cap
finalpp
loy
AG_3
Female
AG_2
AG_4
Relative importance