European Research on Management and Business Economics 31 (2025) 100286

Contents lists available at ScienceDirect

European Research on Management and Business Economics

journal homepage: www.elsevier.com/locate/iedeen

Leveraging xAI for enhanced surrender risk management in life insurance products

Lluís Bermúdez [a],*, David Anaya [a], Jaume Belles-Sampera [b]

[a] Riskcenter-IREA, Universitat de Barcelona, Spain
[b] Riskcenter-IREA and Grupo Catalana Occidente S.A., Spain

ARTICLE INFO

JEL classification:
C35
C38
G22
G32

Keywords:
Machine learning
Shapley values
Kohonen neural network

ABSTRACT

Explainable Artificial Intelligence (xAI) plays a crucial role in enhancing our understanding of decision-making processes within black-box Machine Learning models. Our objective is to introduce various xAI methodologies, providing risk managers with accessible approaches to model interpretation. To exemplify this, we present a case study focused on mitigating surrender risk in insurance savings products. We begin by using real data from universal life policies to build logistic regression and tree-based models. Using a range of xAI techniques, we gain valuable insight into the inner workings of tree-based models. We then propose a novel supervised clustering approach that integrates Shapley values with a Kohonen neural network (KNN). The process involves three main steps: computing Shapley values from a supervised tree-based model; clustering individuals into homogeneous profiles using an unsupervised KNN; and interpreting these profiles with a supervised decision tree model. Finally, we present several key findings derived from the application of xAI techniques, which have the potential to enhance surrender risk management practices.

1. Introduction

Universal life insurance products include a surrender option that allows policyholders to exchange their existing contracts for the cash surrender value at any time during the contract's term. This flexibility makes these products appealing in the insurance market but also introduces surrender risk, i.e., the possibility of policyholders prematurely surrendering their policies, leading to financial losses for insurers. This risk extends beyond insurance to other financial products, such as savings accounts and certificates of deposit. These banking products, like insurance policies, are vulnerable to premature withdrawals or surrenders, including external or internal transfers, and partial or full withdrawals. The reasons for surrendering these products often overlap with those for surrendering insurance policies, such as changes in financial needs, new investment opportunities, or dissatisfaction with product performance.

Surrender risk is recognized as a major challenge in the life insurance industry (Burkhart, 2018; Campbell et al., 2014; Kling et al., 2014) and has therefore received considerable attention in the literature. Recent studies specifically analyze the risk behavior associated with total withdrawals in savings products (Chang & Schmeiser, 2021; Huang et al., 2021), while others focus on quantifying the potential economic impact of this risk on insurance companies (Chunli & Jing, 2018; Hwang et al., 2021; Vincenzo et al., 2017).

In this context, both risk managers and academics have recognized the importance of quantifying risks and identifying their determinants, leading to a growing reliance on statistical models (Eling & Kochanski, 2013; Kiermayer, 2022). Statistical models systematically analyze historical data, identify patterns, and predict events, helping organizations understand the likelihood and impact of risk for better decision-making and resource allocation. For surrender risk, logistic regression has long been the most widely used method (Kiesenbauer, 2012; Kim, 2005). However, the growing demand for more effective surrender risk management has recently driven the adoption of Artificial Intelligence (AI) and Machine Learning (ML) techniques to more precisely analyze the likelihood of surrender events (Jia et al., 2024; Loisel et al., 2021).

In this new context, risk managers face the challenge of understanding, trusting, and effectively managing the outcomes produced by these advanced techniques. This paper aims to improve surrender risk management by introducing risk managers to eXplainable Artificial Intelligence (xAI) techniques, which offer valuable insights into surrender risk behavior. Owens et al. (2022), in a systematic review of xAI applications in insurance, highlight that only a small fraction of the studies reviewed address the use of xAI techniques in risk management (Azzone et al., 2022). These techniques may contribute to better decision-making by (1) facilitating critical knowledge extraction

* Correspondence to: Department of Economics, Financial and Actuarial Mathematics, Av. Diagonal 690, 08034 Barcelona, Spain.
E-mail addresses: lbermudez@ub.edu (L. Bermúdez), davanlu@gmail.com (D. Anaya), jaume.belles@gco.com (J. Belles-Sampera).

https://doi.org/10.1016/j.iedeen.2025.100286
Received 5 March 2025; Received in revised form 13 June 2025; Accepted 23 June 2025
Available online 22 July 2025

and rule identification, and (2) boosting confidence in the predictive accuracy of black-box ML models.

More specifically, this paper has a twofold objective: first, to advance comprehension of existing xAI techniques for analyzing surrender risk; and, second, to propose a new methodology (building on the approach by Bermúdez et al., 2023) for identifying clusters of policyholders with either high or low probability of surrendering their policies. Finally, using a real dataset, we discuss ways of integrating the insights gained from this analysis into risk management strategies.

As noted in Mensah et al. (2024), while AI and ML offer substantial benefits in enhancing risk management within the banking and insurance industries, thus contributing to greater financial stability, they also bring significant risks and challenges. These include cybersecurity threats, systemic vulnerabilities, the privacy of personal data (Abdulbaqi et al., 2023) and other regulatory, ethical, and social concerns -all of which require attention. Enhancing the transparency and interpretability of ML models enables risk managers to better understand their limitations and biases, resulting in improved decision-making and better results for organizations and stakeholders.

The remainder of this paper is structured as follows. The “Data and Methods” section describes the real dataset, outlines the fundamental concepts of ML models and xAI techniques, and provides justification for the development of novel xAI approaches, such as the one proposed in this paper. The sections “Results” and “Discussion” present the findings and explore their implications for risk management, respectively. Finally, the “Conclusions” section summarizes the key points and suggests areas for future research.

2. Data and methods

The dataset consists of policies issued by a life and non-life insurance company operating in the Spanish market during 2018 and 2019. It focuses on universal life insurance products, specifically those active as of December 31, 2018. In total, the dataset includes 49,810 policies, with 9.02% (4494 policies) surrendered by December 31, 2019. Table 1 provides a detailed overview of policy features, including the characteristics of the policyholder and the product, and the status of the policy.

Our analysis adopts a short-term perspective (one year ahead). This approach is intentional, as it enables a focused examination of the inherent policy underwriting features, excluding external factors such as market interest rates and the broader economic environment. Although these external elements may influence the probability of surrender risk, analyzing their impact would require additional external data sources and the application of dynamic modeling techniques, both of which are beyond the scope of the present study.

Although the adoption of a short-term perspective may be questioned due to the long-term nature of universal life contracts, this assumption aligns with the objectives outlined in the Introduction. Specifically, the focus on internal factors, which are typically stable over time and controllable by the insurer, justifies this approach.

Table 1 includes the variable *fee*, which indicates that certain life insurance products require an initial deposit that serves as a surrender fee in cases of early termination. This deposit aims to mitigate the risk of surrendering. Essentially, it represents a charge that policyholders agree to pay if they cancel the policy within a specified period, typically during the early years of the contract (up to 10 years). The table also includes the variable *res*, which represents the fund value for the current period. This value corresponds to the surrender value, calculated as the cash value minus any applicable surrender fees, which policyholders receive if they choose to terminate the policy before its maturity. The remaining variables are related to the characteristics of the policyholder, such as age and gender of the insured, as well as characteristics of the policy itself, such as the time since inception, annual premium, years remaining until the last premium to be paid, additional sum insured in case of death, and product-related characteristics such as tax advantages or a guaranteed interest rate.

Table 1
Definition of selected variables.

Variable	Definition
sur	Policy status (1: Surrendered, 0: Active)
res	Current value of the fund (€)
prem	Total annual premium (€)
age	Current age of the insured
loy	Number of years the policy has been in force
rem	Years remaining until final premium as per contract
gen	Gender of the insured (1: Female, 0: Male)
cap	Additional sum insured in case of death (1: High; 0: Low)
freq	Premium payment frequency (1: Other; 0: Monthly)
incr	Annual premium increment (1: Variable; 0: Constant)
pay	Active premium payment (1: Yes, 0: No)
unl	Unit-linked product (1: Yes, 0: No)
tax	Product with tax advantages (1: Yes, 0: No)
fee	Product with active surrender fee (1: Yes, 0: No)
rate	Product with fixed guaranteed interest rate (1: Yes, 0: No)

Tables 2 and 3 provide descriptive statistics for numerical and categorical explanatory variables, respectively, categorized according to whether they belong to the subset of surrendered policies or the complementary subset of policies that remain in force at the end of the 2018–2019 period. Both subsets exhibit similar behavior, making it difficult to discern a clear pattern in the data. However, policies in force generally show higher premium and reserve values, a larger proportion of high capital amounts, and a slightly greater presence of policies with active surrender fees.

2.1. Predictive models

To analyze the likelihood of surrender events, risk managers model a conditional Bernoulli random variable, such as the policy status (*sur*) of policy i for period $t+1$, given that the policy is active in period t. Within this framework, the state of policy i at time t is classified as 0 (Active or negative case) or 1 (Surrendered or positive case). The objective is to predict the state of policy i for the next period $t+1$, based on its active state in the current period t. Although the probability of surrender, $\pi(t)$, is unknown, it is assumed to depend on the specific characteristics of each policy i at time t. This modeling approach enables an exploration of the dynamics of policy surrender and the factors that influence the likelihood of a policy transitioning from an active to a surrendered state.

In this study, we evaluate a range of predictive models, including logistic regressions (LR) and tree-based models such as decision trees (DT), random forests (RF), and extreme gradient boosting (XGB). While logistic regression and decision trees offer traditional approaches with easily interpretable explanations of feature importance, the other ML models employ more advanced methodologies, each with distinct strengths and applications. For example, RF models use ensemble learning by constructing multiple decision trees during training and aggregating their outputs (e.g., by taking the mode for classification tasks). XGB models are based on a similar concept, but construct weak learners sequentially, optimizing a differentiable loss function through gradient descent. These iterative processes often improve predictive performance but come at the expense of interpretability, making these models commonly referred to as “black-box” ML models. As detailed in the following subsections, xAI techniques can help address these interpretability challenges, shedding light on the decision-making processes and results of these complex models.

To assess the performance of the classification models and validate the results, we use various metrics derived from the confusion matrix for each model, including accuracy, sensitivity, and specificity. In addition, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used for further evaluation. For all

Table 2
Descriptive statistics for the numerical explanatory variables.

Surrendered policies						
Variable	Mean	Min.	25th Pctl.	Median	75th Pctl.	Max.
rem	15.94	0	6.00	10.00	21.00	77.00
res	5,298.95	36.75	1,035.88	2,368.80	5,754.39	298,228.28
prem	804.20	0	532.70	565.20	1,065.50	10,654.80
age	46.39	15.00	38.00	45.00	55.00	83.00
loy	4.66	1.00	2.00	3.00	6.00	29.00
Active policies						
Variable	Mean	Min.	25th Pctl.	Median	75th Pctl.	Max.
rem	14.36	0	6.00	10.00	19.00	80.00
res	10,882.20	6.50	1,834.60	4,388.50	10,707.20	2,523,190.80
prem	881.20	0	532.70	639.30	1,065.50	31,964.40
age	48.14	16.00	41.00	48.00	56.00	90.00
loy	6.21	1.00	2.00	5.00	8.00	35.00

Table 3
Descriptive statistics for the categorical explanatory variables.

Surrendered policies	
Variable	Levels
gen	Male (1,939–46.97%); Female (2,555–53.03%)
incr	Constant (1,712–38.10%); Variable (2,782–61.90%)
freq	Monthly (289–6.43%); Other (4,205–93.57%)
cap	Low (2,211–49.20%); High (2,283–50.80%)
pay	Yes (3,660–81.44%); No (834–18.56%)
unl	Yes (923–20.54%); No (3,571–79.46%)
tax	Yes (1,729–38.47%); No (2,765–61.53%)
fee	Yes (2,204–49.04%); No (2,290–50.96%)
rate	Yes (85–1.90%); No (4,409–98.11%)
Active policies	
Variable	Levels
gen	Male (20,444–45.18%); Female (24,872–54.82%)
incr	Constant (14,024–30.95%); Variable (31,292–69.05%)
freq	Monthly (4,364–9.63%); Other (40,952–90.37%)
cap	Low (16,544–36.51%); High (28,772–63.49%)
pay	Yes (38,954–85.96%); No (6,362–14.04%)
unl	Yes (7,407–16.35%); No (37,909–83.65%)
tax	Yes (11,607–25.61%); No (33,709–74.39%)
fee	Yes (25,595–56.48%); No (19,721–43.52%)
rate	Yes (2,983–6.58%); No (42,333–93.42%)

models, 80% of the data is used for training (through 5 repeated 10-fold cross-validation), with the remaining 20% reserved for testing. The *caret* R package is used to fit each of the predictive models.

When it comes to handling data, the number of policies with a surrendered state is much smaller than the number of in-force policies, leading to an imbalanced dataset. This imbalance arises from limited prior exposure or information about the relevant events, often linked to rare or previously unobserved occurrences. As a result, building predictive models for binary classification with such imbalanced datasets can lead to unreliable outcomes (Japkowicz & Stephen, 2002). To address this issue, we explore resampling techniques (Chawla et al., 2002) to determine the most suitable method for our data in terms of performance, using metrics such as AUC.

2.2. xAI methods

As noted earlier, risk managers can incorporate xAI methods into their analyses to gain a more comprehensive understanding of model behavior. This deeper insight enables them to translate the model's findings into meaningful policyholder profiles, thereby improving risk management strategies.

Here, we focus specifically on model-agnostic xAI methods, which are versatile and can be applied to any ML algorithm. These methods are generally post-hoc, meaning they are used after the model has made its predictions. They can offer either global insight, explaining the behavior of the entire model, or local insight, focusing on individual predictions (Adadi & Berrada, 2018).

To support risk managers in their decision-making process, we begin by introducing the most widely used xAI techniques. For global xAI methods, we emphasize feature importance (FI) techniques, which identify the most influential features, as well as visualization tools like partial dependence plots (PDP) and accumulated local effects (ALE), which illustrate how predictions change with variations in specific features while others are held constant. For local xAI techniques, we focus on Shapley values, which use game theory to assign a value to each input feature based on its contribution to prediction (Lundberg & Lee, 2017), providing insight into the reasoning behind a model's decision for a specific data point.

Throughout the study, we use the *iml* (interpretable machine learning) R package to apply the most common xAI techniques.

2.3. Clustering model's predictions with xAI techniques

Using the xAI techniques mentioned above, risk managers can pinpoint the most relevant features (e.g., using FI and ALE) that any ML model employs to predict the probability of surrender. In addition, they can determine the exact contribution of each feature to the surrender likelihood prediction for every individual policyholder (e.g., using Shapley values).

However, risk managers may also be interested in identifying clusters of policyholders with similar surrender likelihood predictions or comparable predictive capabilities. Risks within the same cluster are likely to exhibit shared properties, potentially requiring similar risk response strategies. This approach enables risk managers to allocate resources more effectively, prioritizing the most critical clusters or those with greater predictive confidence.

As a key contribution of this paper, we propose adapting the Shapley values technique to a global xAI tool. Inspired by Bussmann et al. (2020), Gramegna and Giudici (2020) and Cooper et al. (2021), and expanding (Bermúdez et al., 2023), our supervised clustering approach combines Shapley values with a Kohonen neural network (KNN) to group similar policyholders based on their Shapley values, thus clustering them according to shared predictive outputs and characteristics.

In summary, we adopt a three-step process: (1) applying the xAI technique (Shapley values) to a supervised ML model; (2) identifying homogeneous profiles using an unsupervised model (KNN); and (3) using a supervised model (decision tree) to gain a deeper understanding of these profiles.

In the first step, after fitting a ML model to the data and obtaining surrender likelihood predictions for all policyholders, we calculate the Shapley values of the features for all policyholders. Beyond providing local interpretations for each policyholder (i.e., indicating the contribution of each feature to their individual prediction), these Shapley values can also serve as a global interpretability tool, as outlined in the second step of this approach.

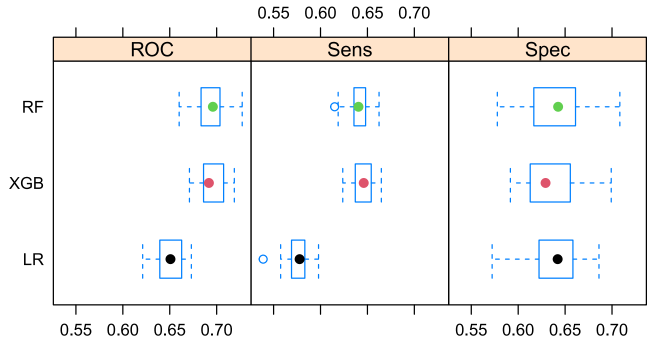

Fig. 1. Performance metrics box-plots (ROC, sensitivity, and specificity) of logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB) models.

In the second step, we apply a clustering technique to group policyholders with similar Shapley values, which reflect similar predictions and characteristics. For visualization purposes, we propose using a Kohonen neural network (KNN), also known as a self-organizing map (SOM), trained with the Shapley values to identify different policyholder profiles. The KNN is an unsupervised learning algorithm widely used for data visualization, clustering, and pattern recognition (Huysmans et al., 2006). This step offers a complementary global perspective on the various profiles leveraged by the ML model in its predictions.

Finally, we introduce a third step to gain a deeper understanding of these profiles. Using the cluster assigned to each policyholder from the SOM map as the response variable, and the original features of each policyholder as covariates, we construct a decision tree without growth constraints. This approach reveals the rules governing each branch and final nodes, providing detailed information about profiles with similar Shapley values and the characteristics of the policyholders within them. This step serves to approximate the initial ML model with an interpretable framework, giving clearer insight into the factors driving the predictions and the different levels of confidence associated with them.

For this approach, we used the *kohonen* and *caret* R packages.

3. Results

3.1. Fitting models

The preparatory phase includes analyzing alternatives to address dataset imbalance, normalizing and scaling continuous variables, and partitioning the data for cross-validation, with 80% allocated for training and the remaining 20% reserved for testing.

As previously discussed, the dataset is heavily skewed towards 0 (Active or negative; 90.98%) compared to 1 (Surrendered or positive; 9.02%). Without proper data treatment before building predictive models, these models would likely underestimate the probability of surrender events. To address this, multiple runs of each model were conducted and evaluated using the resampling techniques outlined in Section 2.1 (e.g., undersampling, oversampling, SMOTE and ROSE). Using repeated k-fold cross-validation (with k = 10, repeated five times) and standard performance metrics, the undersampling technique was identified as the most suitable approach.

After refitting the LR, RF and XGB models using the undersampling technique, their predictive accuracies were compared using standard performance metrics (see Fig. 1). In this comparison, the black-box models (RF and XGB) demonstrated higher predictive accuracy than the LR model.

In the next section, we enhance managerial skills by emphasizing the usability of global xAI techniques to interpret, understand, and build trust in the fitted RF model.

3.2. xAI: interpreting the RF model

Feature importance (FI) techniques identify the factors that most significantly influence model predictions. The importance of each explanatory variable is measured by removing it from the model and evaluating the resulting reduction in the model performance. As shown in Fig. 2, FI scores highlight the key features driving the prediction of surrender events in our model. The numerical features, such as *res* and *prem* (indicators of the commitment of the policyholder), emerge as critical factors. In contrast, categorical features related to the characteristics of the policyholder and the product, such as *freq* and *gen*, appear to have minimal impact on the predictions.

However, FI does not offer insight into how the most important variables affect the predictions of a model. Accumulated local effects (ALE) plots address this by illustrating how features impact the model predictions on average, with the most influential features exhibiting the highest ALE values. Furthermore, the nature of ALE plots accounts for nonlinear effects of individual features and interactions between features.

Fig. 3 shows the ALE plots for all the features of the RF model. The ALE value can be interpreted as the main effect of the feature at a certain value of the feature compared to the average prediction of the data (an ALE value of 0 indicates a neutral effect of the feature). For example, in the case of the feature *rem*, the average prediction decreases as the number of remaining years to the final premium increases. Specifically, *rem* has a positive effect up to 2–3 years, a neutral effect between 3 and 25 years, and a negative effect beyond 25 years.

Focusing on the most influencing features, we see that premium payments of up to 500€reduce the surrender risk, whereas payments exceeding this amount increase the risk, with a positive effect on the average prediction beyond 1200€. In contrast, the surrender risk gradually decreases as policyholders accumulate higher fund values, except for very low fund values (typically associated with recent policies) that exhibit a low average prediction.

To a lesser extent, influential categorical features impact the surrender risk: policies with tax benefits (*tax*), constant premium (*incr*), or inactive premium payments (*pay*) exhibit an increased surrender risk while, in contrast, products with an active surrender charge (*fee*) mitigate this risk.

It is worth noting that the FI score for a feature in Fig. 2 considers both the main effect (its direct impact in isolation) and the interaction effects (influence arising from interactions with other features) on model performance. Second-order ALE plots, which exclude main effects, estimate the combined impact of feature pairs on predictions, allowing managers to interpret feature interactions more effectively.

We start by identifying the interacting features within this RF model using the H-statistic from (Friedman & Popescu, 2008) to measure their strength. For purposes of illustration, we focus on the most significant interaction observed: between *res* and *prem* (see Fig. 4). Specifically, for funds below 10,000€, the effect on predictions varies according to annual premium payments. When *prem* is below 300€, the effect on surrender prediction is negative, mitigating surrender risk. In contrast, for *prem* above 300€, the effect is neutral or positive, increasing surrender risk. For funds exceeding 10,000€, the second-order ALE plot does not reveal a clear pattern.

After introducing the most common global xAI techniques, we turn our focus to understanding the prediction output for a particular individual by considering its specific input values. In this regard, Shapley values are a commonly used local xAI method to understand the importance of characteristics in individual predictions. Each Shapley value quantifies the contribution of a feature to the disparity between the model prediction for a single instance and the average prediction (approximately 45% in our analysis).

Fig. 5 displays the individual Shapley values for two policyholders (one with a low surrender prediction and the other with a high surrender prediction). Policy 1, held by a middle-aged policyholder (46)

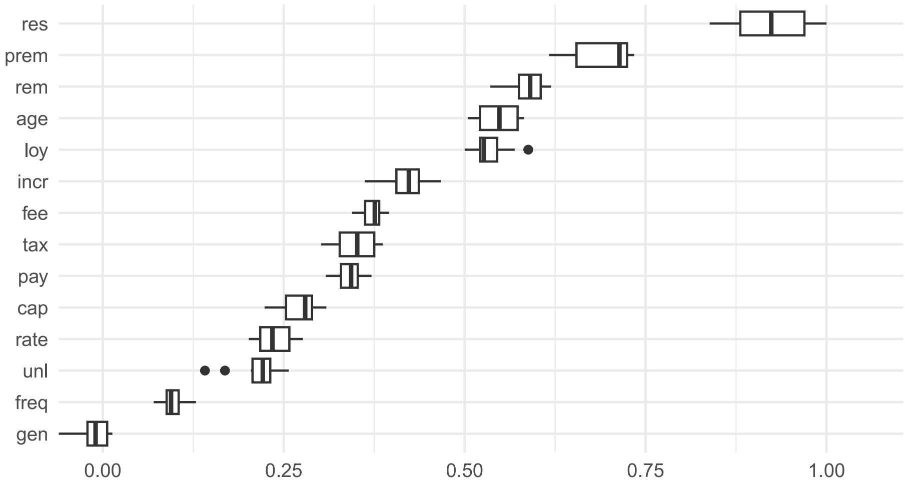

Fig. 2. Feature importance (FI) of the RF model.

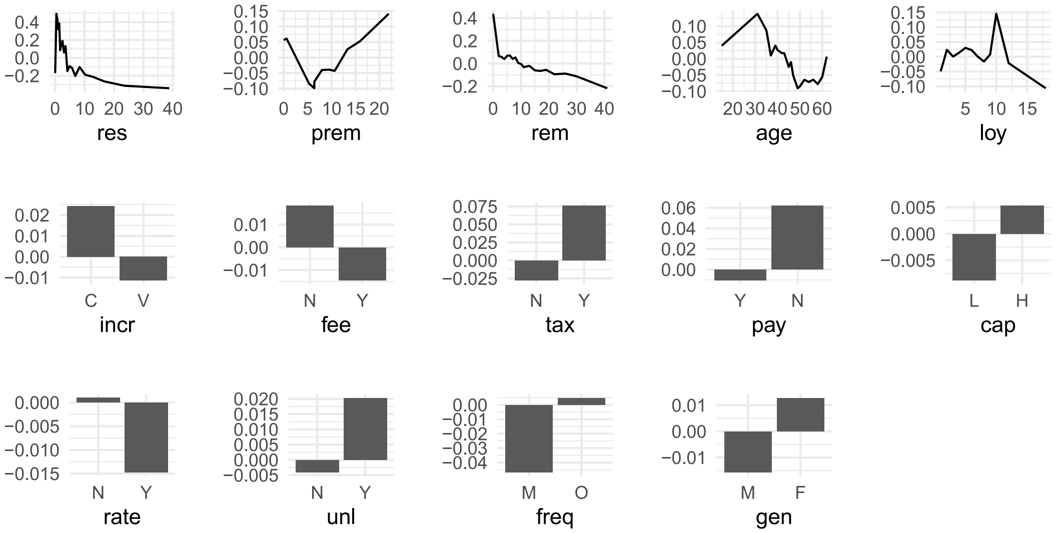

Fig. 3. Accumulated local effects (ALE) for all features of the RF model. The *res* feature is expressed in thousands of €, while the *prem* feature is expressed in hundreds of €.

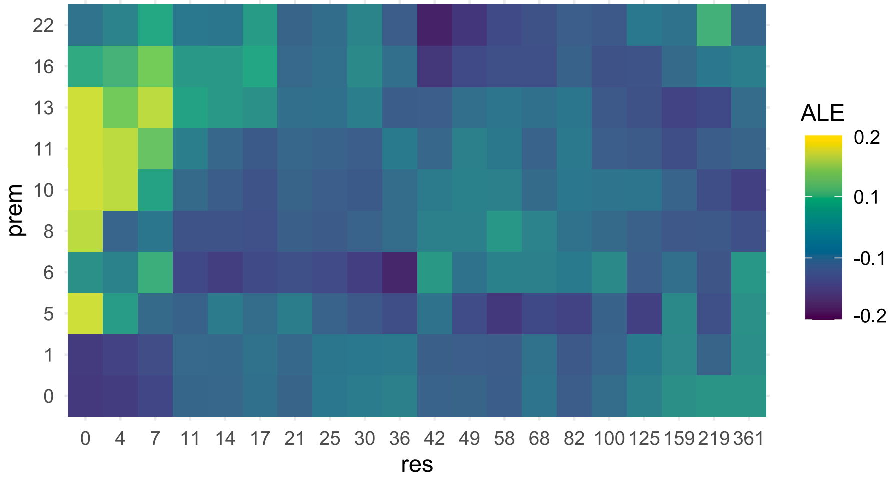

Fig. 4. Second-order ALE for *prem* and *res*. The *res* feature is expressed in thousands of €, while the *prem* feature is expressed in hundreds of €.

for 13 years, features a very high reserve fund and annual premiums. The RF model accurately predicts it as an active policy, assigning a surrender probability of 0.09. In contrast, Policy 2, owned by a young policyholder (25) for just one year, has a very low reserve fund but significant annual premiums. The model correctly classifies it as a surrendered policy with a surrender probability of 0.98. In both cases,

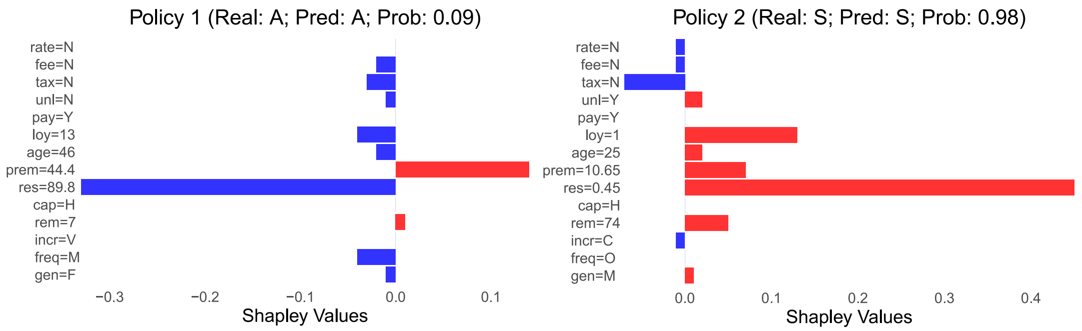

Fig. 5. On the left: Shapley values for a policy with low likelihood of total surrender. On the right: Shapley values for a policy with high likelihood of total surrender. The *res* feature is expressed in thousands of €, while the *prem* feature is expressed in hundreds of €.

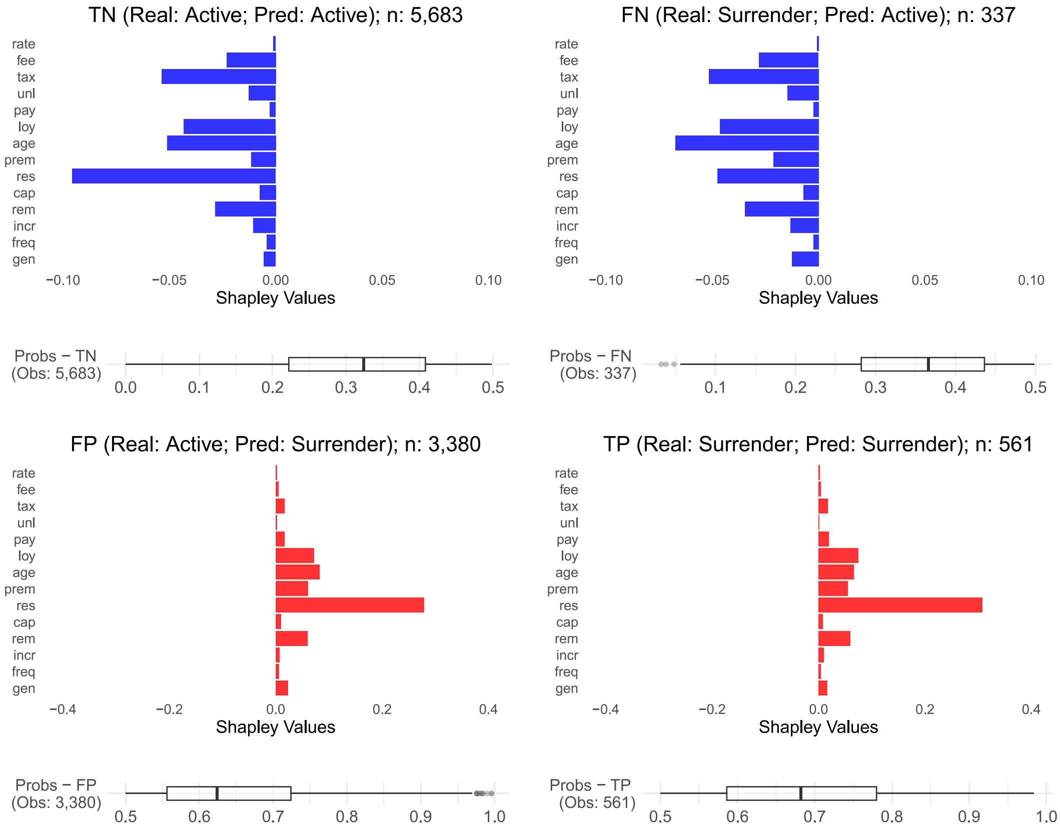

Fig. 6. Average Shapley values for all observations in the testing dataset, grouped by each cell of the confusion matrix, alongside the corresponding boxplots of the predicted surrender probabilities.

the reserve fund is the most influential feature, exerting a negative effect in Policy 1 and a positive effect in Policy 2. Note that the high surrender probability assigned to Policy 2 is largely influenced by the interaction effect depicted in Fig. 4.

3.3. Clustering model's predictions with Shapley values

As presented in Section 2.3, we propose adapting the Shapley values technique to a global xAI tool by clustering policyholders with similar Shapley values, that is, grouping them according to shared predictive outputs and features of importance.

Following the methodology outlined in Section 2.3, we start by computing the Shapley values of the features for all policyholders. As a first approach to using Shapley values for global analysis, Fig. 6 presents the average Shapley values for all 9961 observations in the testing dataset, categorized by each cell of the confusion matrix, along with the corresponding boxplots of the predicted surrender probabilities.

Fig. 6 illustrates that, as expected, policyholders predicted by the model to have Surrendered -both true positives (TP) and false positives (FP)- exhibit similar Shapley values, predominantly positive. Likewise, policyholders predicted to remain Active -true negatives (TN) and false negatives (FN)- demonstrate a comparable degree of similarity, though with negative values. Additionally, the boxplots of predicted surrender probabilities indicate that FP and FN (i.e., misclassifications) tend to occur when the predicted values are closer to the 0.5 threshold. However, there are no significant differences that enable the identification of patterns to better understand which characteristics affect the model's accuracy or errors in predicting whether a policy will remain in force within a year.

To address this issue, we proceed to the second step by applying a Kohonen neural network (KNN), also known as a self-organizing map (SOM). This network is trained using the Shapley values, for all observations in the testing dataset, to identify distinct policyholder profiles that share similar Shapley values, and consequently exhibit

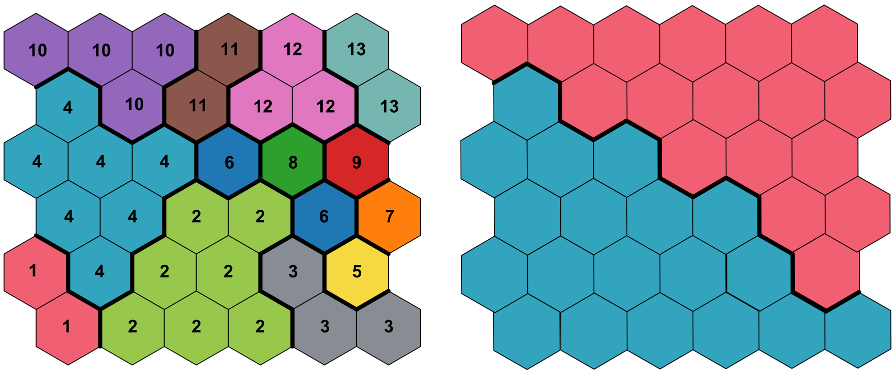

Fig. 7. On the left, the SOM map for all observations in the testing dataset. On the right, the same SOM map, categorized by Active (blue) and Surrendered (red) predicted statuses.

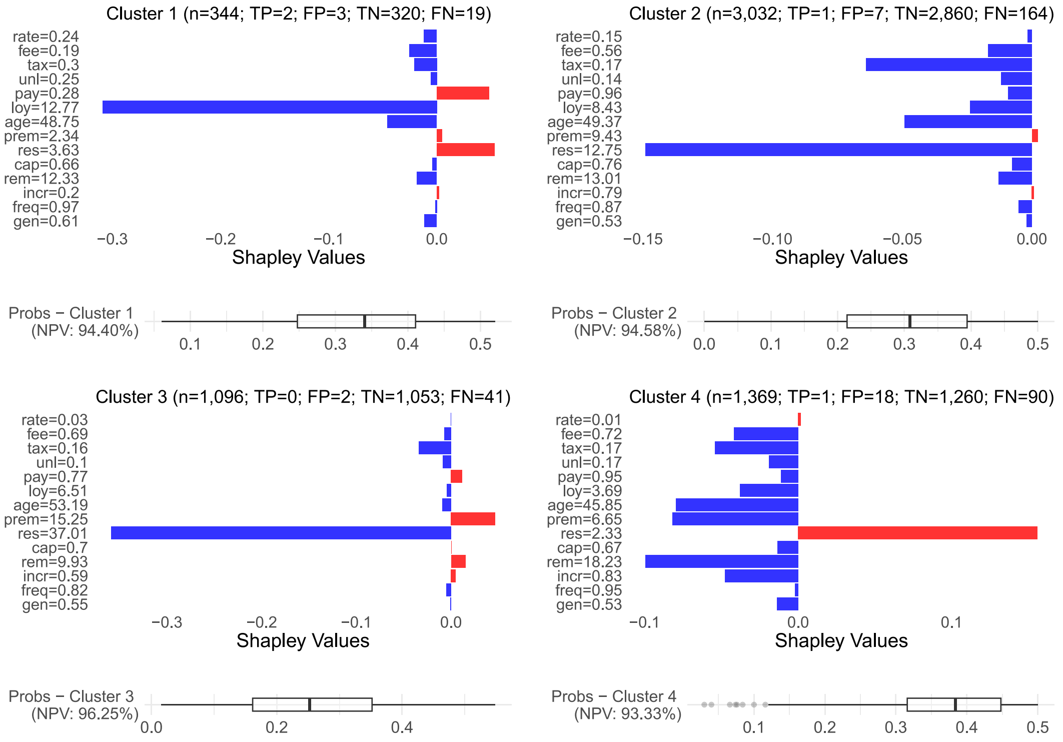

Fig. 8. Average Shapley values for all observations within each negative cluster, alongside the corresponding boxplots of the predicted surrender probabilities. The left column presents the mean feature values for each cluster. The *res* feature is expressed in thousands of €, while the *prem* feature is expressed in hundreds of €.

comparable predictions and feature importance characteristics. The inherent properties of the Shapley values, used as input features, allow the SOM model to function effectively. After performing several checks, we choose a SOM structure with a 6 × 6 hexagonal grid.

Fig. 7 presents the resulting SOM map for this case, which features 13 distinct clusters of policyholders. These clusters can be grouped into two main areas: those shown in blue, where the majority of policies are predicted as Active by the RF model, and those in red, where the model predominantly predicts Surrendered policies. For the clusters in the lower triangle (1–4), the average predicted surrender probability is 31%, while for those in the upper triangle (5–13), the probability is 62%. This contrast highlights the ability of the SOM maps to differentiate between the cases predicted by the model as negative and those predicted as positive.

To gain a deeper understanding of each cluster, Figs. 8 and 9 display the average Shapley values for all observations within each cluster, analogous to the approach used in Fig. 6 for each cell of the confusion matrix.

For illustration purposes, we focus on the cluster with the highest number of observations (Cluster 2). As shown in Fig. 8, the model predicted 3024 of 3302 policies as Active, achieving an overall accuracy of 94.4%. According to Shapley values, the most influential feature of this cluster is *res*, followed by *tax*, *age*, *loy*, and *fee*, all of which have a negative impact on prediction, thus reducing surrender risk. On average, this cluster comprises middle-aged policyholders (*age* = 49.37) with substantial fund values (*res* = 12,750€), holding policies without tax benefits (*tax* = 0.17) and approximately halfway through their duration (*loy* = 8.43, *rem* = 13).

Finally, after distributing the policies into clusters, we move to the third step by constructing a simple decision tree to visually illustrate the classification rules for all clusters and gain insight into the behavior of this specific RF model. As described in Section 2.3, we train a classification tree using the original variables of each policyholder as input features and the cluster assigned from the SOM map as the target variable. To visually distinguish the rules for the two target variable labels (Active and Surrendered), we also incorporate the output prediction of the RF model as an additional input feature. We do not

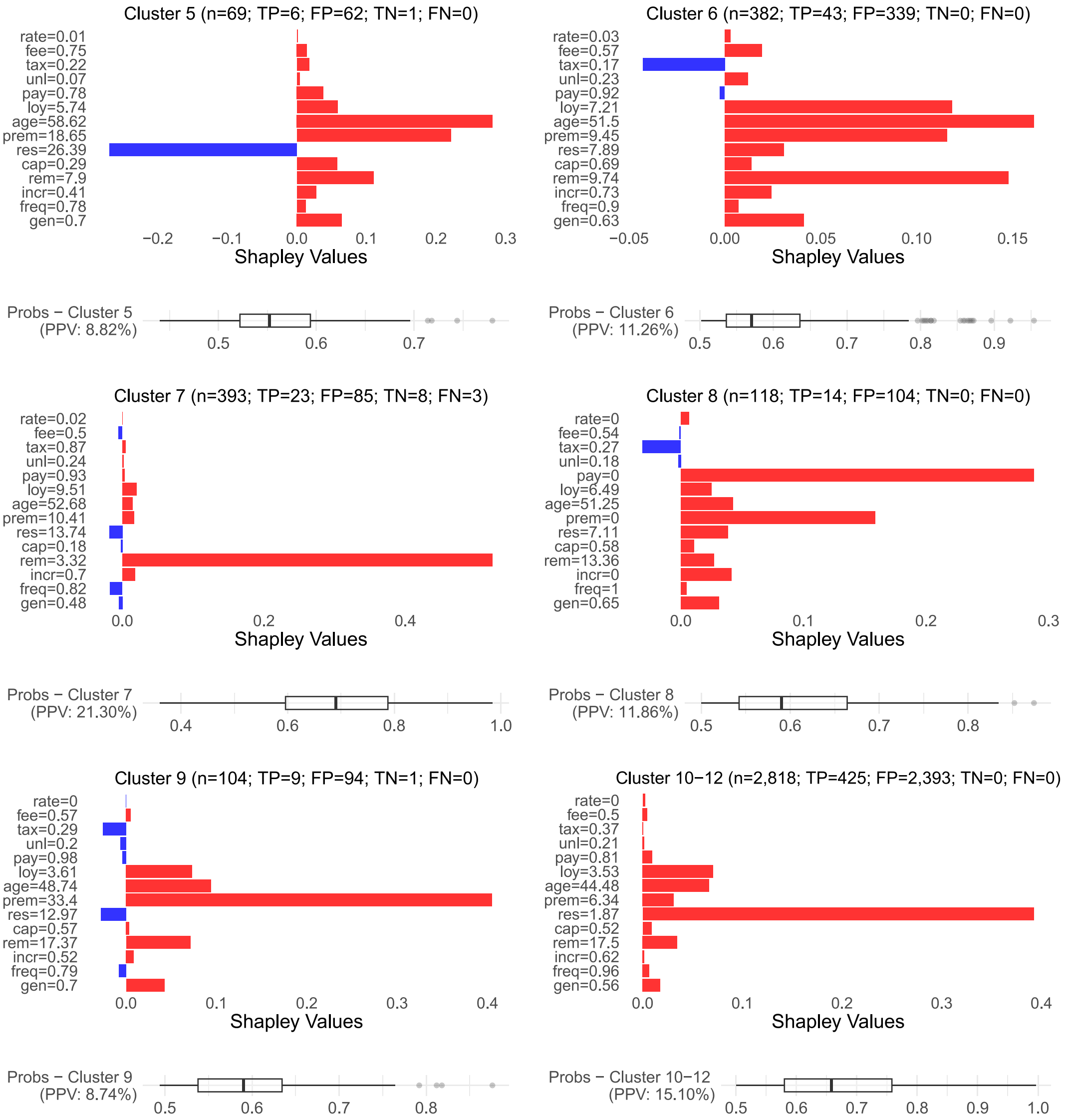

Fig. 9. Average Shapley values for all observations within each positive cluster, alongside the corresponding boxplots of the predicted surrender probabilities. The left column presents the mean feature values for each cluster. The *res* feature is expressed in thousands of €, while the *prem* feature is expressed in hundreds of €.

impose growth constraints since our goal is not prediction, but rather visualizing rules, making overfitting irrelevant.

Fig. 10 presents the resulting decision tree, highlighting the branches and nodes that influence the process. This visualization enhances the insights gained from each cluster and, more broadly, enables the identification of decision rules that explain the model's behavior. In addition, the most relevant terminal nodes include different metrics to evaluate predictive performance, providing risk managers with information on various confidence levels.

In Fig. 10, once again focusing on Cluster 2 for illustrative purposes, we identify three final nodes where the majority of cases belong to this cluster. Excluding the node with fewer observations (80 cases), we first examine the node containing 2545 cases, of which 2314 belong to Cluster 2, with a misclassification rate of only 6% (1–2174/2314). The decision rule for this node corresponds to policies with medium fund values (between 3823€and 16,000€). In parallel, the final node with 486 cases, 405 of which belong to Cluster 2, exhibits an even lower misclassification rate of 3%. The decision rule for this node applies to policies with large fund values (exceeding 16,000€) and a duration of less than 12 years.

Finally, it is important to highlight that Fig. 8 (or Fig. 9) and Fig. 10 provide complementary insights into the decision-making process of the RF model. As shown for Cluster 2, both options enable a more comprehensive analysis by presenting different perspectives on how the model arrives at its predictions.

4. Discussion

The xAI techniques introduced in this paper provide risk managers with valuable insights into how ML models generate their predictions, allowing them to uncover hidden trends that might not be immediately evident from raw data alone. In addition, xAI enhances managerial confidence in the predictive accuracy of models. This enhanced understanding enables managers to make more informed data-driven decisions and improve risk assessment strategies.

In this section, our goal is to summarize the key insights extracted by the RF model from the previous section, highlighting their relevance for risk managers and how these data-driven insights can be used to mitigate surrender risk.

To provide context for risk management, we outline some of the most commonly used strategies to enhance policyholder retention and reduce the likelihood of early withdrawals. The first set of mitigation measures focuses on product design, including customized and flexible

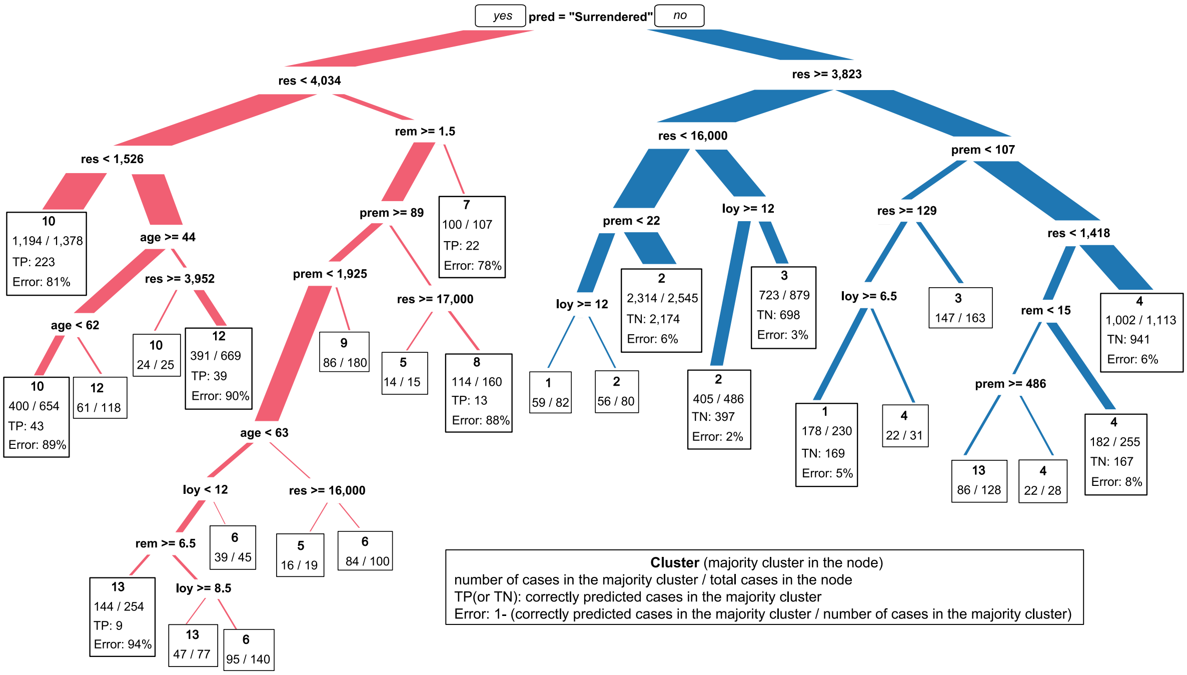

Fig. 10. Decision tree for classifying clusters of predicted policies, with the most relevant terminal nodes displaying various performance metrics.

policies (with regard to payment premium conditions, term conditions, or financial options), and implementing surrender charges and/or loyalty benefits. The second group involves risk control measures, such as implementing Key Risk Indicators (KRIs) to identify at-risk policyholders and deploying targeted retention strategies through enhanced customer engagement efforts. Finally, other strategies follow risk-based pricing and underwriting approaches, adjusting them based on identified risk profiles while considering financial market trends and competitor offerings.

In terms of the level of confidence that a risk manager may have in the RF model, Fig. 1 indicates that the model has difficulty recognizing enough patterns to accurately classify the two possible states of the policy, with sensitivity and specificity rates both around 63%. This underscores the inherent complexity of the risk and the difficulty of capturing it accurately through predictive models.

However, Fig. 6 reveals a sharp contrast between the proportion of correctly classified negative cases among predicted negative cases (Negative Predictive Value, NPV) and the corresponding proportion for positive cases (Positive Predictive Value, PPV). Specifically, when the RF model predicts that a policy remains Active, it is correct 94.4% of the time (NPV). Conversely, when it predicts a policy as Surrendered, it is incorrect in 85.8% of cases (PPV).

Given that both true and false cases (whether for positive or negative predicted policies) exhibit similar Shapley values on average, the following considerations can be made. On the one hand, the risk manager can be highly confident that the policies predicted as Active by the RF model (primarily those within the profiles categorized as Active in Fig. 7) will indeed remain active. The small error margin, which accounts for just 5.6% of the classifications, can likely be attributed to unobserved characteristics that the model was unable to capture. On the other hand, the risk manager knows that most of the policies predicted as Surrendered by the RF model (mainly those within the profiles categorized as Surrendered in Fig. 7) will actually remain active. The risk manager may consider true cases as the proportion of policyholders who have exercised the surrender option this year, while the rest, sharing the same profile characteristics, can be considered at-risk. These at-risk policies may potentially surrender in subsequent years and therefore warrant the implementation of appropriate monitoring strategies as part of the risk management framework.

In general, based on insights from global xAI techniques (see Figs. 2 and 3), the numerical factors that most significantly impact model predictions are those related to policyholder commitment, such as fund value (*res*) and annual premium (*prem*), as well as factors associated with policy seniority, including the policy age (*loy*), remaining years of premium payments (*rem*), and the insured's age (*age*). To a lesser extent, certain categorical features related to product characteristics, such as the presence of a surrender charge (*fee*), tax benefits (*tax*), incremental premiums (*incr*), or an active premium payment status (*pay*), also influence the model's predictions.

However, due to the presence of nonlinear effects and interactions between features, the risk manager should be careful when interpreting the impact of these influential factors on surrender likelihood predictions. For example, as shown in Fig. 4, for funds below 10,000€, the effect on predictions varies depending on annual premium payments. Another example, inferred from Fig. 3, is the nonlinear relationship between the insured's age and the model's predictions.

Given the limited accuracy of the current RF model and the lack of granularity in existing global xAI techniques to interpret it, new approaches are needed, such as the one proposed in Section 2.3. Our approach provides risk managers with profiles of policyholders who exhibit similar surrender likelihood predictions and individual characteristics, which may require similar risk response strategies. In the following paragraphs, based on Figs. 8–10, we analyze the characteristics that define each cluster of policyholders.

Starting with the four clusters categorized as Active (or negative), **Cluster 1** comprises policies with low premium and fund values, taken out many years ago by relatively young policyholders. Although these low values generally indicate a higher likelihood of surrender, reflected in the positive average Shapley values, this tendency is mitigated by the policyholders' loyalty. Even after surpassing the surrender charge period and suspending premium payments, they continue to maintain their policies. In fact, this cluster exemplifies the interaction effect between the premium and reserve values, as analyzed in Fig. 4, where the low premium and fund values help mitigate surrender risk.

Cluster 2 includes middle-aged policyholders with substantial fund values, holding policies without tax benefits, and approximately midway through their duration.

In **Cluster 3**, most policies have accumulated significant fund values in a relatively short period, with the majority still subject to an active surrender fee. However, as shown in Fig. 10, a small subset mirrors the low premiums and fund values seen in Cluster 1. This cluster exhibits the lowest surrender probabilities and the highest NPV.

The final negative cluster, **Cluster 4**, consists of very young policies, most of which remain subject to an active surrender fee. Despite relatively high premiums, policyholders in this group have yet to accumulate significant fund values. The ongoing annual payment of incremental premiums helps reduce surrender risk. However, this cluster has the highest surrender probability among the negative clusters and is predicted with slightly less accuracy.

In summary, we identify four distinct groups of policies that are predicted to remain active by the RF model. These clusters demonstrate a high level of accuracy, with approximately 95% of these policies remaining active. Next, let us examine the distinct groups of policies classified as Surrendered.

Cluster 5 consists mainly of female policyholders approaching retirement age, who have accumulated substantial fund values by making large annual premium payments over a relatively short period. Their clear intention to save for retirement, also reflected in a low sum insured in case of death, places them at potential risk of surrender. Although this cluster shares similar premium and fund values with Cluster 3, the key difference lies in policyholders' intentions: Cluster 3 policyholders, who are slightly younger, seek both savings and life insurance coverage.

Cluster 6 differs from Clusters 5 and 3, as the premiums and fund values are significantly lower. This group exemplifies the interaction effect between the premium and reserve values: for fund values below 10,000€, annual premiums exceeding 800€substantially increase surrender risk. Similarly to Cluster 3, these policies are held by younger policyholders and are approximately halfway through their duration.

Cluster 7 has the highest surrender probability and PPV. It represents policyholders who are still far from retirement, holding older policies with tax benefits that are nearing expiration. Having accumulated substantial fund values, they may opt to transfer these funds to another insurance company, for instance, through a pension plan exchange.

Cluster 8 consists entirely of policies with fixed monthly premiums, where payments were suspended long before their expiration. Based on the expertise of the risk manager, this cluster represents a clear example where, although only a few policies have actually been surrendered (true positive cases), the rest (false positive cases) are likely to be at risk of surrender in subsequent years.

Cluster 9 differs from Cluster 5 in that policyholders are significantly younger and pay much higher annual premiums.

Clusters 10, 11, and 12 are grouped together in Fig. 9 due to their similar Shapley values and feature characteristics. In particular, Cluster 11 does not appear as a final node in Fig. 10, since most cases were reassigned to Clusters 10 and 12. These clusters mainly include very young policyholders who recently acquired policies with moderate annual premiums of around 600€, resulting in small fund values. Once again, this exemplifies the increased surrender risk caused by the interaction between the premium and reserve values.

Finally, **Cluster 13** is excluded from Fig. 9 because it is not relevant for this analysis. The surrender probabilities given by this RF model are all around 0.5 and, hence, policies are predicted in almost equal proportions to be Active or Surrendered.

The RF model failed to accurately predict surrendered policies, with only about 15% of the predicted cases actually resulting in surrender. However, risk managers can still leverage insights from xAI techniques by considering Surrendered cluster cases as at-risk policies that may surrender in the future, thus prompting the implementation of targeted monitoring strategies.

In terms of surrender fee policy-making, the company's typical strategy includes a charge that policyholders agree to pay if they cancel their policy within the first 10 years. In general, this strategy has proven to be effective in reducing surrender risk across most clusters when policies have an active surrender fee.

However, in Clusters 2, 5, and 6, the surrender fee has not successfully mitigated the risk, prompting risk managers to reconsider the rules governing the fees for these profiles. For example, policyholders in Cluster 5, who are basically focused on saving for retirement, may choose early retirement and surrender their policies despite the associated fee. In such cases, offering a product tailored to their new needs could help improve retention.

Analyzing Clusters 9 to 12, which consist of very young policies, a significant number of policies without the standard surrender fee can be identified. Since this condition increases surrender risk, a thorough review of the surrender fee guidelines may be warranted.

The downside of any surrender fee strategy is the significant rise in policy surrenders once the fees expire, which in this case happens after 10 years. This pattern is evident in Clusters 7 and 8 and, to a lesser extent, in Clusters 1 and 3. In particular, in Cluster 8, where premium payments have been suspended, this trend is even more pronounced.

To address this issue, risk managers should consider implementing a Key Risk Indicator (KRI) as part of the overall surrender fee strategy. This KRI would signal when the surrender fee expiration date is approaching, allowing proactive retention strategies to be applied. For example, the policies in Cluster 7, which have accumulated substantial fund values and are at risk of being exchanged for a pension plan offering better conditions, can be revised with more attractive terms (based on financial market trends and competitor offerings) to mitigate the risk.

Another useful KRI to implement is related to the premium payment status. When this KRI signals that a policyholder has suspended premium payments, it can serve as an early indicator of potential surrender (e.g., Cluster 8). Taking early retention actions, such as offering new premium, term, or financial conditions can help mitigate the risk.

5. Conclusions

The study demonstrates the effectiveness of xAI techniques in improving the understanding of ML models used to predict surrender risk in life insurance products. These techniques provide valuable insight into the decision-making processes of black-box models, enhancing trust and interpretability.

The paper introduces a novel approach that combines Shapley values with Kohonen neural networks (KNN) to cluster policyholders based on their surrender risk profiles. This method offers a global perspective on policyholder behavior, facilitating the identification of groups with similar risk characteristics that can also require similar mitigation risk strategies.

We present several key findings, resulting from the effective application of the xAI techniques outlined in this manuscript, which should improve surrender risk management.

First, while the surrender fee strategy has generally been effective in reducing surrender risk, certain groups have not been sufficiently deterred by these fees. In such cases, offering tailored products that better align with policyholders' needs can enhance retention. Second, a significant number of young policies lack surrender fees, increasing surrender risk. Risk managers may consider reviewing the surrender fee guidelines to provide better protection against early cancellations. Lastly, key risk indicators (KRIs) can help identify policies at higher risk of surrender. Implementing KRIs to detect policies nearing their surrender fee expiration or those with suspended premium payments can enable proactive retention strategies and reduce cancellations.

By implementing these strategies and monitoring KRIs, risk managers can enhance retention efforts and mitigate surrender risk more effectively. Nonetheless, there remains significant scope for further research on surrender risk, particularly by relaxing current assumptions, such as transitioning to a long-term perspective, incorporating economic impacts, and considering exogenous factors alongside endogenous ones.

Finally, the paper advocates for further research to refine the modeling process, improve Shapley value calculations, and optimize the use of Kohonen networks. It also underscores the importance of developing more effective visualization techniques to enhance the interpretability of xAI results. In addition, conducting a bias and fairness evaluation of ML models (such as through disparate impact analysis) can help risk managers identify and mitigate potential discrimination against protected groups (e.g., based on race or gender), particularly in highly regulated sectors like finance and insurance.

CRediT authorship contribution statement

Lluís Bermúdez: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **David Anaya:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jaume Belles-Sampera:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Funding

This research has been partially supported by AGAUR, Spain (grants 2020DI9, 2021SGR00299 and 2023CLIMA00012).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdulbaqi, A. S., Salman, A. M., & Tambe, S. B. (2023). Privacy-Preserving Data Mining Techniques in Big Data: Balancing Security and Usability. *SHIFRA*, *2023*, 1–9. http://dx.doi.org/10.70470/SHIFRA/2023/001.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Azzone, M., Barucci, E., Moncayo, G. G., & Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, *191*, Article 116261. http://dx.doi.org/10.1016/j.eswa.2021.116261.

Bermúdez, L., Anaya, D., & Belles-Sampera, J. (2023). Explainable AI for paid-up risk management in life insurance products. *Finance Research Letters*, *57*, Article 104242. http://dx.doi.org/10.1016/j.frl.2023.104242.

Burkhart, T. (2018). Surrender Risk in the Context of the Quantitative Assessment of Participating Life Insurance Contracts under Solvency II. *Risks*, *6*(3), 66. http://dx.doi.org/10.3390/risks6030066.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, *3*, 26. http://dx.doi.org/10.3389/frai.2020.00026.

Campbell, J., Chan, M., Li, K., Lombardi, L., Purushotham, M., & Rao, A. (2014). Modeling of policyholder behavior for life insurance and annuity products: A survey and literature review. Society of Actuaries, https://www.soa.org/globalassets/assets/files/research/projects/research-2014-modeling-policy.pdf.

Chang, H., & Schmeiser, H. (2021). Life insurance surrender and liquidity risks. *Quantitative Finance*, *22*(4), 761–776. http://dx.doi.org/10.1080/14697688.2021.1998586.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. http://dx.doi.org/10.1613/jair.953.

Chunli, C., & Jing, L. (2018). Early default risk and surrender risk: Impacts on participating life insurance policies. *Insurance: Mathematics & Economics*, *78*, 30–43. http://dx.doi.org/10.1016/j.insmatheco.2017.11.001.

Cooper, A., Doyle, O., & Bourke, A. (2021). Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology. In *Machine learning and principles and practice of knowledge discovery in databases* (pp. 408–422). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-93733-1_29.

Eling, M., & Kochanski, M. (2013). Research on lapse in life insurance: what has been done and what needs to be done? *Journal of Risk Finance*, *14*(4), 392–413. http://dx.doi.org/10.1108/JRF-12-2012-0088.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954. http://dx.doi.org/10.1214/07-AOAS148.

Gramegna, A., & Giudici, P. (2020). Why to Buy Insurance? An Explainable Artificial Intelligence Approach. *Risks*, *8*(4), http://dx.doi.org/10.3390/risks8040137.

Huang, F. W., Chen, S., & Lin, J. H. (2021). Insurer Investment, Life Insurance Policy Choices, and Policy Surrender. *Emerging Markets Finance and Trade*, *58*(9), 2637–2651. http://dx.doi.org/10.1080/1540496X.2021.2007879.

Huysmans, J., Baesens, B., Vanthienen, J., & Van Gestel, T. (2006). Failure prediction with self organizing maps. *Expert Systems with Applications*, *30*(3), 479–487. http://dx.doi.org/10.1016/j.eswa.2005.10.005.

Hwang, Y., Chan, L. F. S., & Tsai, C. J. (2021). On Voluntary Terminations of Life Insurance: Differentiating Surrender Propensity From Lapse Propensity Across Product Types. *North American Actuarial Journal*, *26*(2), 252–282. http://dx.doi.org/10.1080/10920277.2021.1973507.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. http://dx.doi.org/10.3233/IDA-2002-6504.

Jia, B., Wang, L., & Wong, H. Y. (2024). Machine learning of surrender: Optimality and humanity. *Journal of Risk and Insurance*, *91*(4), 915–942. http://dx.doi.org/10.1111/jori.12428.

Kiermayer, M. (2022). Modeling surrender risk in life insurance: theoretical and experimental insight. *Scandinavian Actuarial Journal*, *7*, 627–658. http://dx.doi.org/10.1080/03461238.2021.2013308.

Kiesenbauer, Dieter (2012). Main Determinants of Lapse in the German Life Insurance Industry. *North American Actuarial Journal*, *16*(1), 52–73. http://dx.doi.org/10.1080/10920277.2012.10590632.

Kim, C. (2005). Modeling Surrender and Lapse Rates With Economic Variables. *North American Actuarial Journal*, *9*(4), 56–70. http://dx.doi.org/10.1080/10920277.2005.10596225.

Kling, A., Ruez, F., & Ruß, J. (2014). The impact of policyholder behavior on pricing, hedging, and hedge efficiency of withdrawal benefit guarantees in variable annuities. *European Actuarial Journal*, *4*(2), 281–314. http://dx.doi.org/10.1007/s13385-014-0093-0.

Loisel, S., Piette, P., & Tsai, C. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin*, *51*(3), 839–871. http://dx.doi.org/10.1017/asb.2021.10.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Red Hook, NY, USA: Curran Associates Inc., https://dl.acm.org/doi/10.5555/3295222.3295230,

Mensah, G. B., Mijwil, M. M., Abotaleb, M., Tawfeek, S. M., Ali, G., Dhoska, K., & Adamopoulos, I. (2024). The Era of AI: The Impact of Artificial Intelligence (AI) and Machine Learning (ML) on Financial Stability in the Banking Sector. *EDRAAK*, *2024*, 43–48. http://dx.doi.org/10.70470/EDRAAK/2024/007.

Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. (2022). Explainable Artificial Intelligence (XAI) in Insurance. *Risks*, *10*(12), 230. http://dx.doi.org/10.3390/risks10120230.

Vincenzo, R., Rosella, G., & Frank, J. F. (2017). Intensity-Based Framework for Surrender Modeling in Life Insurance. *Insurance: Mathematics & Economics*, *72*, 189–196. http://dx.doi.org/10.1016/j.insmatheco.2016.11.001.