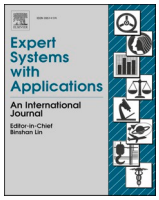
Expert Systems With Applications 255 (2024) 124753

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Understanding and predicting lapses in mortgage life insurance using a machine learning approach

Carlos Manteigas, Nuno António *

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

ARTICLE INFO

Keywords:
External data sources
Lapse risk
Machine learning
Mortgage life insurance

ABSTRACT

Mortgage Life Insurance (MLI) offers lucrative opportunities for insurers. However, customer retention has proven to be a daunting challenge, particularly following the regulatory changes of 2009 in Europe. New market entrants strategically employing low-premium tactics have reshaped the competitive landscape, leading established insurers and banks to grapple with retaining their MLI clientele. Consequently, increasing policy lapses hold critical implications for these financial entities. Responding to this intricate landscape, our research presents a predictive model that pinpoints the MLI policies at risk of lapse and disentangles the underlying factors propelling this risk. The objective is to provide insurers with a practical and strategic tool to improve customer retention, enabling them to identify specific actions to reduce customer attrition, improve financial stability, and strengthen customer loyalty. We used a dataset obtained from an insurance company and its partner bank to build the model. The effectiveness of four machine learning models, namely Logistic Regression, Random Forest, Neural Networks, and XGBoost, is investigated, with XGBoost outperforming the others. SHapley Additive exPlanations (SHAP) were utilized to bolster interpretability, thereby facilitating the conception and explication of the predictive model's most influential features. Underpinning the benefits of a nuanced exploration, the study's focus on a solitary insurance protection product and integrating bank data enabled us to apprehend the multifaceted drivers of lapse behavior. The study accentuates the merit of comprehensive data encapsulating a holistic perspective, with the four most influential features originating from bank data. From an insurer's standpoint, this research provides a strategic vantage point to proactively identify and engage with customers at risk of policy lapse and reformulate their policies to mitigate customer attrition.

1. Introduction

An insurance contract is represented by a policy between an individual or an entity (policyholder) and an insurance company. The policyholder pays out an amount (called the premium) in exchange for financial protection if certain events or circumstances occur, such as accidents, illnesses, or damages to property. In essence, it allows individuals or organizations to transfer the financial impact of certain types of risks to an insurance company.

Depending on the type of insurance policy, insurance can cover a wide range of risks. The most common types are auto, health, homeowners, and life.

Life insurance provides financial support to the policyholder's beneficiaries in case of their death. It is an important form of protection for individuals who have dependents or other financial obligations, providing peace of mind and ensuring that their loved ones are taken care of in case of their death.

While this basic idea of protection is simple, the technical workings of life insurance are a little more complex. The company must ensure that policyholders receive the appropriate amount of coverage at a fair price and that it can generate returns on their investments while maintaining sufficient reserves to pay out future claims. It is a difficult balance to maintain, considering that life insurance policies are typically long-term contracts. Thus, the company must forecast the likelihood of future events, such as the policyholder's death, by identifying the risk profile of policyholders and accurately predicting future claims.

One type of life insurance is mortgage life insurance (MLI), which banks or other lenders usually require when customers apply for mortgage credit. The mortgage lender traditionally sells it through an affiliated insurance company and guarantees the loan debt payment in case of the borrower's death or disability.

Banks tend to consider this insurance a precondition for mortgage

* Corresponding author.
E-mail addresses: m20210924@novaims.unl.pt (C. Manteigas), nantonio@novaims.unl.pt (N. António).

https://doi.org/10.1016/j.eswa.2024.124753
Received 8 March 2024; Received in revised form 3 July 2024; Accepted 7 July 2024
Available online 8 July 2024

loan approval and may worsen the customer's credit conditions if it needs to be removed. Although the client is not legally obliged to do so, they tend to accept the bank's insurance offer since, at this point, they want the bank to approve the mortgage. Customers primarily focus on negotiating the loan rate and tend to disregard the insurance package (Villeneuve, 2014).

The bank's concern is legitimate as it seeks to protect its interests and avoid possible future defaults. For the customer, despite the pressure exerted by the mortgage lender, this product is just as important, as it will ensure their family's financial security and avoid the financial burden of paying off the mortgage in the event of their death.

In 2009, due to the imbalance of power in the sale of this product, Portuguese regulators considered it necessary to guarantee that customers would not be subject to excessive conditions. This regulatory framework introduced new rules and clarified credit institutions' information obligations regarding MLI. In particular, the customer has the right to choose an insurance product from the company of their choice rather than the one offered by the bank (Decreto-Lei n.o 222/2009, de 11 de Setembro | DRE, 2009).

In Portugal, the MLI is typically sold through a term life insurance, which covers the risk of death (and sometimes also disability and serious illness) for a predetermined duration, with a variable or fixed premium for each year. In the event of the policyholder's death during the period of cover, the life insurance company will pay the insured sum to the designated beneficiaries, provided that the premiums have not been underpaid (Fang & Kung, 2021). MLI does, however, have particular characteristics: the duration of the policy and the insured capital correspond to those of the mortgage (in most cases, the insured capital is updated monthly with the repayment of the loan), the beneficiary is always the mortgage lender, and it is also possible to cover two individuals under the same policy if the mortgage involves two people (joint cover).

These characteristics make MLI the most attractive and profitable product for a life insurance company due to its long-term duration (sometimes more than 20 years) and the high sums insured due to the high average mortgage capital. Furthermore, MLI customers are usually young, typically in the early stages of their financial lives, so they are customers that businesses want to keep because of their enormous lifetime value potential.

The product's attractiveness for insurers, together with the new rules introduced by the regulatory framework in 2009, has led to the emergence of new players on the market, some specializing exclusively in this product and offering high price reductions compared to traditional insurers. In addition, customers are increasingly informed and can easily simulate and compare other products available on the market. Banks and traditional insurers face a significant challenge in attracting and retaining customers who opt for cheaper products, both when taking out a mortgage and afterward, by canceling the policy and transferring the insurance to another company.

This challenge is exceptionally high in customer retention, given the customer's ability to voluntarily terminate the policy at any time and the ease with which they can do so. Unlike the insurance company, the customer has the option to terminate the life insurance policy at any time (Fier & Liebenberg, 2013). They may also decide to stop paying the premiums, leading to the eventual cancellation of the policy (Hwang et al., 2022).

The term "voluntary termination" has been used in academic literature to account for both lapses and surrenders. But there are subtle differences between the two. A policy lapse occurs when the policyholder fails to pay the periodic premium, leading to the termination of the policy. A surrender, on the other hand, is the cancellation of a life insurance policy that has a cash value component, resulting in the policyholder receiving the cash value accrued when the cancellation occurs (Eling & Kiesenbauer, 2014; Kuo et al., 2003; Loisel et al., 2021; Shamsuddin et al., 2022).

It is crucial to differentiate between a surrender and a lapse, as their reasons for occurring, causes, and outcomes are different. A policy lapse may result from a lack of effort, knowledge, or motivation to make premium payments, while a surrender stems from the policyholder's evaluation and deliberate choice (Hwang et al., 2022). In this study, we will refer to lapses only since the early termination of the MLI does not imply the return of any cash value, which is usually applied to savings products.

Policy lapses in MLI obviously negatively impact the revenues and profitability of banks and insurers. In addition to losing the commission associated with the product, the bank also loses control over ensuring that the customer has an active policy associated with the loan.

For the insurance company, lapses can have a more profound impact. The insurance company may be unable to recoup all the expenses related to the policy, including those incurred during procurement, underwriting, and policy issuance. These expenses are paid by the insurance company prior to or at the time of policy issuance, but profits are earned throughout the contract. Lapsed policies can thus result in financial losses for the insurer (Kuo et al., 2003). The MLI policies are sold with the expectation of a long-term and, therefore, the expectation of profits over time. Early voluntary terminations (lapses) reduce those profit expectations (Hwang et al., 2022). Lapses can also cause threats to the insurance company's liquidity. The uncertainty of the macroeconomic environment, caused by the COVID-19 pandemic and the energy crisis aggravated by Russia's invasion of Ukraine and, more recently, the war in Israel, has brought new concerns about the potential liquidity risk faced by the financial sector, including life insurance companies (Arriola et al., 2023; Batten et al., 2023; Szczygielski et al., 2022).

Rare events with extreme consequences often deeply impact economies and their economic agents. For life insurance companies, a mass cancellation event – where a significant number or even a majority of policyholders suddenly cancel their life insurance policies – can be such an extreme event (Biagini et al., 2021). A high policy lapse rate can also damage the insurance company's reputation, potentially leading to more policy cancellations and harming new business acquisitions (Eling & Kochanski, 2013).

The supervisory authority also recognizes the importance of lapse risk. The European Insurance and Occupational Pensions Authority (EIOPA) confirmed in its fifth Quantitative Impact Study (QIS5), carried out in 2011 as part of the implementation of Solvency II, that policy lapse risk is the most considerable risk associated with life insurance, surpassing risks related to longevity, expenses, and catastrophes. Specifically, the lapse risk constitutes approximately 50 percent of life insurance underwriting risk (Biagini et al., 2021; Hwang et al., 2022; Reck et al., 2022).

The work presented in this paper aims to understand the main factors affecting lapses in MLI by developing a predictive model that can identify policyholders who are at risk of lapsing. This work is based on a project in a Portuguese life insurance company that sells its products exclusively through a major bank in Portugal.

The following section will analyze the existing literature on lapse risk in life insurance. The aim is to identify the main factors associated with lapse risk, such as policyholder demographics, underwriting, policy characteristics, and techniques used in previous studies. The present study will then be positioned in this research to identify the contributions to be made to the current literature.

2. Literature review

Over the years, the topic of lapse risk has drawn the attention of scholars to study what drives policyholders to lapse their policies. However, not until the 1990s did the number of publications examining lapsed policies significantly increase, emphasizing the analysis of the socio-demographic and economic factors involved (Shamsuddin et al., 2022). Eling and Kochanski (2013) reviewed 56 papers on the topic of life insurance lapse and found that only seven were published before the year 2000.

Early researchers on life insurance lapse based their studies on three theoretical hypotheses: the Emergency Fund Hypothesis (EFH), which posits that policyholders use the cash value component as a source of the emergency fund, the Interest Rate Hypothesis (IRH). IRH emphasizes the role of arbitrage in response to rising interest rates, and the Policy Replacement Hypothesis (PRH) suggests that policyholders tend to cancel their old policies when they get new policies that offer better terms and pricing (Shamsuddin et al., 2022).

Research on life insurance lapse can be divided into two main sets, depending on the data used: macro-oriented papers that used mainly environmental variables to explain the behavior of lapse rates and micro-oriented papers that used product and policyholder characteristics to assess the factors affecting lapse rates.

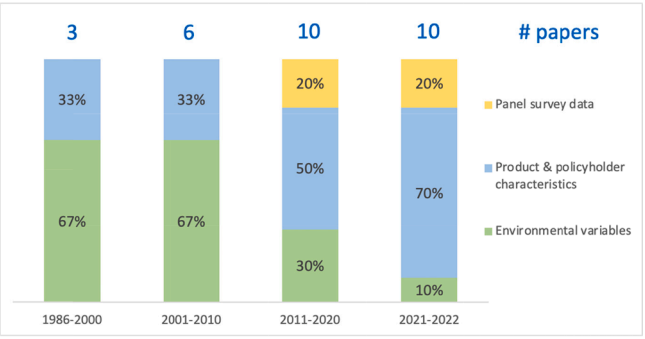
Some authors used panel survey data to identify which personal characteristics most affect lapses (Fang & Kung, 2021; Fier & Liebenberg, 2013), using information gathered from the University of Michigan Health and Retirement Study (HRS), or (Nolte & Schneider, 2017) that based their analysis on information obtained from the German SAVE study.

The Scopus database was used to search for previous literature on life insurance lapse. It was possible to analyze 29 scientific articles published between 1986 and 2022. Fig. 2.1 shows the evolution of the type of data used in these articles. While until 2010, most studies used environmental variables, the more recent literature mainly uses product and policyholder characteristics. Table 2.1 details the research on the latter component.

2.1. Macro-oriented research

Macro-oriented researchers mainly used environmental characteristics, including aggregate macroeconomic indicators and company data, to find evidence to support one of the theoretical hypotheses, particularly the EFH and IRH. Their emphasis was on examining the impact of various environmental factors, such as interest and unemployment rates, gross domestic product (GDP), interest returns in the capital markets, as well as company attributes like size and organizational structure, on the rates of policy termination (lapse) (Hwang et al., 2022). Some limited their studies on the impact of interest rates and unemployment (e.g., Dar & Dodds, 1989; Kuo et al., 2003; Outreville, 1990). Other included additional economic indicators such as GDP and capital markets development as well as specific company characteristics (e.g., Cox & Lin, 2006; Kiesenbauer, 2012; Kim et al., 2005). The methodologies used by these studies were mainly time-series analysis.

However, the empirical results from this research are inconsistent, which can be attributed to variations in the markets, types of products, methods, time frames, and specifications of the variables utilized (Eling & Kochanski, 2013). Some studies find evidence for the EFH theory (e.g., Dar & Dodds, 1989; Outreville, 1990). Kuo et al. (2003), though, discovered that the impact of interest rates on lapses is more significant from an economic perspective than the impact of unemployment rates. Ultimately, the IFH is more strongly supported than the EFH (Eling & Kochanski, 2013).



Source: Scopus database

Fig. 2.1. Set of 29 papers on life insurance lapse, published between 1986 and 2022. Source: Scopus database.

2.2. Micro-oriented research

This body of research primarily employs data from insurance companies on individual policies to evaluate the effect of product and policyholder characteristics on policy lapse behavior. Few studies have used these variables throughout the years, and only recently began to emerge with greater frequency. Perhaps because most of this data is not readily available due to its privacy characteristics and the limitations on its use due to General Data Protection Regulation (GDPR) rules. Fig. 2.1 shows that most papers using these variables are from 2021 and 2022.

Recently, Shamsuddin et al. (2022) conducted a study on the publication of scientific literature on life insurance lapse. Their study analyzed 178 documents published between 1971 and 2021. Regarding micro-oriented research, they discovered conflicting views on the role of sociodemographic factors, such as age, gender, marital status, number of dependents, and level of education, in influencing policy termination. Some studies found that the policyholder's age and number of dependents affect the likelihood of a policy lapsing (e.g., Ćurak et al., 2015; Gemmo & Götz, 2016). Others reported that the policyholder's age and number of dependents did not impact the rate of policy termination (e.g., Sirak, 2015; Yaakob et al., 2018), therefore suggesting not to consider these factors in future studies. These conflicting results present an intriguing and valuable opportunity for continued investigation of this subject.

Shamsuddin et al. (2022), through their bibliometric analysis, noticed that although the word "lapse" is an insurance topic that has been gaining increased attention by researchers, "life insurance" is still a trending theme to be explored by researchers. Through their trend topics analysis, the authors also found that the subject of analysis has evolved through time: from 1998 to 2012, scholars were more interested in the topic "lapse rate". The topic "policy surrender" gained interest from 2019 to 2021, the "lapse risk" from 2017 to 2020, and "machine learning" from 2021. Their analysis also found that research on policyholder behavior started in 2016 and lapsed behavior in 2020.

The methodologies used in the first studies were mainly Generalized Linear Models (GLM), including logistic regression (LR), binomial, negative binomial, and Poisson (e.g., Barucci et al., 2020; Cerchiara et al., 2008; Eling & Kiesenbauer, 2014; Kagraoka, 2005; Renshaw & Haberman, 1986). Other researchers used Classification and Regression Trees (CART) – (e.g., Groll et al., 2022; Loisel et al., 2021; Milhaud et al., 2010). Most recent papers used various machine learning (ML) classification methods, such as Naïve Bayes (NB), Random Forest (RF), Neural Networks (NN), k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), XGBoost (XGB) (e.g., Azzone et al., 2022; Babaoglu et al., 2017; Groll et al., 2022; Kiermayer, 2021; Loisel et al., 2021; Xong Lim et al., 2019). Reck et al. (2022) used the "Least Absolute Shrinkage and Selection Operator" (Lasso) method and its extension, the Fused Lasso.

Table 2.1 lists 15 micro-oriented research papers on life insurance lapse, detailing the variables, techniques used, and their main findings.

Despite the growing number of studies on policy lapse prediction, significant gaps remain. Firstly, there are inconsistencies in conclusions regarding sociodemographic factors, highlighting the need for more robust and nuanced analyses. Secondly, most research focuses broadly on life insurance rather than on specific products that may exhibit different characteristics and behaviors, indicating a gap in product-specific research. Additionally, most studies rely exclusively on data from insurance companies, suggesting a need to integrate more diverse data sources, such as bank data or broader financial behaviors. Furthermore, the application of advanced machine learning techniques in this context is relatively recent, and their full potential has yet to be fully exploited. Previous studies have produced inconsistent results, with no single model clearly standing out as superior. Another

Table 2.1
Micro-oriented research on life insurance lapse.

Author	Geography	Type of product	Variables	Techniques	Results
Renshaw & Haberman, 1986	Scotland	Endowment, whole-life, temporary insurance	Policyholder age, gender, contract age, product type, company	GLM – LR, binomial	Important factors (IF): age at entry, duration of the policy, office, type of policy
Kagraoka, 2005	Japan	Annuity-type personal accident	Policyholder age, gender, seasonality, unemployment rate, heterogeneity	GLM – Poisson model, Negative binomial model	IF: change of unemployment rates, time elapsed from contract date
Cerchiara et al., 2008	Italy	Savings	Policyholder age, contract age, product type, calendar year	GLM – Poisson model	IF: policy duration. Sensitivity of calendar year to product type and policyholder age
Milhaud et al., 2010	Spain	Endowment	Policyholder age, contract age, product type, sum insured, risk premium, savings premium	LR, Classification, and CART	IF: duration of the policy and profit benefit option
Eling & Kiesenbauer, 2014	Germany	Endowment, Annuity, Term Life	Policyholder age, gender, contract age, product type	GLM – Poisson model, Binomial model, Negative binomial model	IF: product type, contract age, policyholder age, and gender
Babaoglu et al., 2017	n.a.	General life insurance	Client profile (85 attributes, including product, gender, issue year, province, initial amount, age, payment mode, policy type, and macro-economic indicators)	LR, NB, RF	IF: gender, smoker code, product name, duration, underwriting class, blood test, marital status, payment mode, attained age, premium period, benefit period, face amount and funding level
Xong Lim et al., 2019	Malaysia	General life insurance	Policy Status, Sum Insured, Premium Frequency, Entry Age, Policy Term and Gender.	LR, k-NN, NN, SVMs	SVMs and NN outperformed LR and k-NN.
Barucci et al., 2020	Italy	Savings (traditional with participating, unit-linked)	Policyholder age, gender, contract age, contract size, product type, premium frequency, region, profession, inflation rate, growth rate of disposable income, growth rate of the European stock index	GLM and survival analysis	Some evidence supports EFH (positive correlation of lapse rates with personal financial difficulties). Determinants of lapses for the two types of contracts (traditional / unit-linked) are quite different
Hu et al., 2021	Ireland	General life insurance	Policyholder age, gender, contract duration, total number of policies, sum assured, number of lapsed policies, household composition, education level, employment status	Spatial analysis, LR	Adding census data does not improve the predictive model.
Loisel et al., 2021	Taiwan	Traditional (whole life, term life, endowment), interest-adjustable, investment-linked	Policyholder age, gender, occupation, physical examination, channel, premium payment, payment method, product type, participation type, currency	SVMs, XGB, CART, LR	XGB and SVMs perform better than LR and CART. Economic gains can be further enhanced when the optimization is done on a function linked to economic gains rather than on statistic accuracies
Kiermayer, 2021	n.a.	Endowment	(Simulated data using the previous literature)Current age, face amount, duration, duration (elapsed) , premium frequency, premium	XGB, RF, GLM, NN	XGB is the superior model
Hwang et al., 2022	Taiwan	Whole life, term life, endowment, life insurance with periodic living benefits	Policyholder age, gender, policy age, occupation, single premium, supplementary coverage, policy size, product type, channel, commission ratio, payment method, termination year	LR	The distinction between surrender and lapse propensities with respect to various determinants. Significant differences in the voluntary termination propensity models between the two major product types.
Groll et al., 2022	Germany	Endowment, private pensions	Annual premium, the sum insured, and the actuarial interest rate	Logit, CART, RF, XGB	IF: start of the contract, remaining duration, age of the first insured person, sum insured, annual premium, surplus system used, number of repayments. None of the classification methods outperformed the others.
Reck et al., 2022	n.a.	Traditional, unit-linked	Policyholder age, gender, country, contract duration, insurance type, payment frequency, payment method, nationality, premium increase percentage, the original term of the contract, premium payment duration, sum insured, premium	GLM, Lasso, Fused Lasso	The new Lasso model used outperforms GLM and regular Lasso.
Azzone et al., 2022	Italy	Savings (with guarantee and unit-linked)	Policyholder age, gender, region, company, time from the start, time to expiry, contract size, product type, macroeconomic variables (disposable income, inflation, Eurostoxx, interest rate)	LR, RF	IF: time passed from start, time to expiry, company, contract size, premium Behavioral and commercial reasons play important roles.RF performs better than LR.

Source: The Scopus database.

significant gap is the insufficient attention paid to the interpretability of models, with many studies failing to provide clear explanations of the factors influencing their predictions. Moreover, most existing models have been developed and validated in limited geographical contexts, raising concerns about their generalizability to different markets and regulatory environments. Addressing these gaps could provide more comprehensive insights into policy lapse behaviors and enhance the predictive accuracy and applicability of future models.

2.3. Expected contributions to the existing research

As described above, literature on life insurance lapse prediction has traditionally covered a broad spectrum of life insurance products, most often without considering the nuances and specific characteristics associated with those products. Some studies have made a valuable effort to distinguish between two main categories, savings and protection products, revealing that policyholders' motivations for lapsing can vary significantly.

Hwang et al. (2022) constructed individual surrender and lapse models for one type of protection product and one type of savings product, and the results of their study identified significant differences in the lapse propensity models between the two types of products. They concluded that individual models should be built for different types of products rather than relying on a "universal" model for all.

The initial expected contribution to the existing literature on lapse prediction is based on a fundamental observation: we need to make even deeper distinctions in the field of protection products. The introduction of credit-linked protection products, in this case, the MLI, represents a critical extension of this gap. The lapse of this product may be influenced by factors completely different from other life insurance protection products, such as macroeconomic factors, interest rate fluctuations, and real estate market conditions. In the context of MLI, there was also a change in the regulatory framework, which introduced new market dynamics and new players, generating other motivations for the customer to lapse the policy.

We believe this study represents a pioneering effort in the field of lapse prediction by narrowing its focus onto this specific insurance protection product, the MLI, which has specific characteristics and is recognized as one of the most important and profitable products for a life insurance company. In an increasingly aggressive competition environment, as described above, insurers and banks face significant challenges to retain MLI customers. This study recognizes the urgency and complexity of this challenge by focusing efforts on understanding and predicting lapses from this specific product. To the best of our knowledge, this is the first empirical study to do so.

Another distinctive feature of this work is the use of bank data, namely information on the mortgage loan (such as the amount of the installment, spread rate, and effort rate) and some sociodemographic features of customers (educational qualifications, employment, banking segment, wage domiciliation, other insurance products held by the customer).

In previous research, insurance companies have relied mainly on internal data and insurer-specific variables to develop lapse prediction models. However, while valuable, these models may not comprehensively understand the complex dynamics governing lapse behavior in a product like MLI. The introduction of mortgage loan variables and customer characteristics represents a critical departure from this convention, as it delves into external data sources, taking advantage of information that is usually beyond the insurer's reach.

By integrating external, bank-related data into the prediction framework, we recognize the multidimensionality of the MLI lapse problem and aim to present a more comprehensive and nuanced understanding of policyholder behavior.

This new approach offers insights into lapse prediction and highlights insurers' practical limitations. It demonstrates that in order to address the complex dynamics of MLI lapse, we need to tap into data that is not ordinarily available to the insurer. In doing so, this research seeks to empower insurers with a more holistic understanding of lapse behavior, enabling them to implement more effective retention strategies in the face of increasing competition and dynamic market conditions.

3. Methodology

The framework employed in this project, the "Cross-industry standard process for data mining" (CRISP-DM) (Chapman et al., 2000), is well known for its systematic and iterative approach to data mining and ML projects. CRISP-DM's structured methodology encompasses six phases (business understanding, data understanding, data preparation, modeling, evaluation, and deployment). It has proved to be invaluable for the development of the present study.

One of the key strengths of CRISP-DM is its emphasis on constant iteration across all six phases. This iterative nature allowed the maintenance of a flexible and adaptive mindset throughout the project, reacting to new insights, challenges, and data complexities in real time, optimizing the study's development as the project unfolded.

3.1. Business understanding

An insurance company has two strategies for managing the risk of lapse: a reactive approach, in which the company responds after the customer has already lapsed, and a proactive approach, which involves predicting which policyholders are likely to lapse and developing preventive retention strategies.

The company in focus for this project currently follows the reactive approach. Each month, the bank's sales representatives are provided with customer relationship management (CRM) contacts to communicate with customers who have already canceled their policies. Unfortunately, these contacts often produce low success rates since, at this stage, customers are usually determined to cancel their policies.

This project aims to help the company transition to a proactive approach, which is expected to improve customer retention for MLI.

As illustrated in Fig. 3.1, the cancellation ratio has steadily increased since 2015. The cancellation ratio measures the proportion of policies canceled during the year in relation to the total number of policies in force at the beginning of the year.

Despite this upward trend in cancellations, it should be noted that a significant proportion is due to the loan being repaid, as seen in Fig. 3.2, with the client no longer needing to maintain insurance. Lapses, which include cancellations due to non-payment and transfers to other companies, currently account for around 34 % of cancellations.

The main objective of this project is to optimize the retention rate of customers at risk of lapsing. This requirement implies the early identification and implementation of effective retention strategies.

At the same time, the company wants to minimize the costs associated with customer retention while maintaining a high retention rate. This objective requires the development of efficient and cost-effective retention measures.

In addition, one of the main focuses of this project is to identify the main factors contributing to MLI lapses. Identifying these crucial characteristics allows the company to take specific measures, proactively preventing lapses in the future.

In a broader context, the successful execution of this project can significantly improve the overall customer experience. Additionally, this project promotes greater customer loyalty, ultimately increasing customer lifetime value and strengthening the company's future profitability.

3.2. Data understanding

As mentioned, this project uses authentic data from a major Portuguese life insurance company and its associated bank. Both entities have

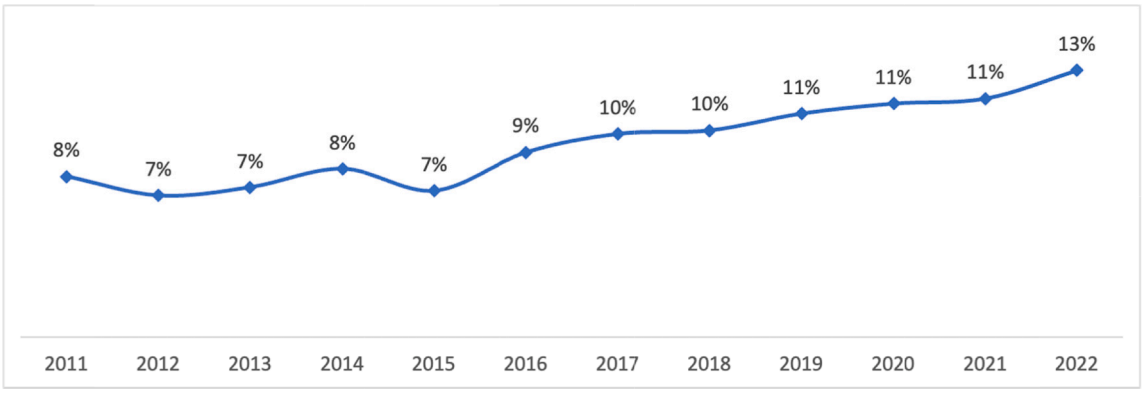


Fig. 3.1. Cancellation ratio evolution. Source: authors.

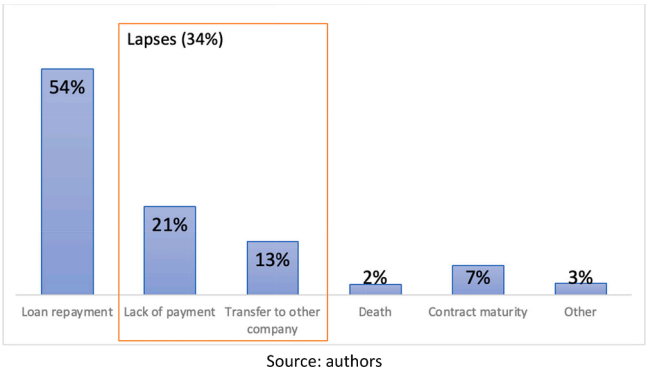


Fig. 3.2. Cancellation reasons (years 2021 and 2022). Source: authors.

chosen to remain anonymous in the context of this project.

The insurer's data was obtained directly from its operating system and represents the company's MLI policy portfolio at the close of 2022. The selection of variables for analysis was influenced by existing literature, covering factors such as date of birth, gender, premium, sum insured, marital status, and term. Additionally, some variables were chosen to align with the product's unique attributes, including the number of people insured, the percentage of capital coverage, and the initial and current insured capitals. It should be noted that payment frequency, often the subject of previous studies, was not included since all the policies considered have monthly payments. It is essential to stress that all data has been meticulously anonymized in order to comply with data protection regulations. For a complete list of the variables obtained from the insurance company, see Table 3.1.

The company's dataset has been enriched with the bank's information on the corresponding mortgage. Each policy corresponds to a single mortgage. Table 3.2 shows the information collected from the bank.

The project's dataset was also enriched with some socio-economic information about the client provided by the bank. Table 3.3 shows this information.

Based on the composition of the MLI portfolio and the unique characteristics of certain policy groups (see Fig. 3.3), it was deemed appropriate to exclude specific MLI policy segments from this study, which include:

- **Average Tariff (Rate) Policies:** This subset is made up of an older portfolio, with fixed rates that do not vary with the age of the client, which distinguishes it from the other products. It currently represents 14 % of the total portfolio.
- **Subsidized Loans:** These policies comprise a mere 1.7 % of the portfolio.

Table 3.1
Dataset – variables from the insurance company.

Variable name	Type	Description
POLICY_ID	Categorical	Policy / contract code (anonymized)
CLIENT_ID	Categorical	Client code (anonymized)
START_YEAR	Float	Policy start year
END_YEAR	Float	Policy end year
INSURED_SUM	Float	Current insured sum
INSURED_SUM_initial	Float	Initial insured sum
ANNUAL_PREMIUM	Float	Current annual premium
ANNUAL_PREMIUM_initial	Float	Initial annual premium
INSURED_PERSONS	Categorical	Number of insured persons (1 = 1 person; 2 = 2 persons; A = multi-persons)
BIRTH_DATE_PERS1	Date	Birthdate of the insured person 1
BIRTH_DATE_PERS2	Date	Birthdate of the insured person 2 (if exists)
GENDER_PERS1	Categorical	Gender of the insured person 1
GENDER_PERS2	Categorical	Gender of the insured person 2 (if exists)
%COVER	Float	Insured sum that covers the mortgage capital (3 options: 50 % for 2 persons; 100 % for 1 or 2 persons)
PORTFOLIO_regular	Float	Number of active policies of other products with regular payment
PORTFOLIO_PREMIUMS_regular	Float	Premiums of active policies of other products with regular payment
PORTFOLIO_single	Float	Number of active policies of other products with a single payment
PORTFOLIO_PREMIUMS_single	Float	Premiums of active policies of other products with single payment
PORTFOLIO_max_duration	Numeric	Maximum duration of a policy of the client
YEAR_FIRST_POLICY	Numeric	Year of the first policy of the client
YEAR_LAST_POLICY	Numeric	Year of the last policy of the client
PREVIOUS_CANCEL_POLICIES	Numeric	Number of policies canceled previously related to the same mortgage loan
CLIENT_CANCEL_LAST_5Y	Numeric	Number of policies lapsed by the customer in the last 5 years (excluding policies linked to mortgages)

- **Policies from other Banks:** Policies from other banks were excluded due to the unavailability of corresponding mortgage information. This category represents 6 % of the MLI portfolio.

Table 3.2
Dataset – variables from the bank (mortgage).

Variable name	Type	Description
CREDIT_COLLECTION_DAY	Numeric	Day of collection of the credit installment
CREDIT_DEED_DATE	Date	Date of the mortgage deed
CREDIT_TERM_DATE	Date	Mortgage term date
CREDIT_NEXT_DUE_DATE	Date	Due date of the next installment
CREDIT_LTV_INITIAL_BINS	Categorical	Initial mortgage Loan-to-Value (LTV) in bins – LTV corresponds to the ratio: "Loan Amount / Collateral Value"
CREDIT_AMOUNT_BINS	Categorical	Mortgage amount in bins
CREDIT_SPREAD_BINS	Categorical	Mortgage spread in bins
CREDIT_REPAYMENT_PERIOD_BINS	Categorical	Mortgage repayment period in bins
CREDIT_EFFORT_RATE_BINS	Categorical	Mortgage effort rate in bins – corresponds to the ratio: "Total credit installments / Net monthly income"
CREDIT_AMOUNT_INITIAL	Float	Mortgage initial amount
CREDIT_AMOUNT_OUTSTANDING	Float	Mortgage outstanding amount
CREDIT_REPAYMENT_PERIOD	Numeric	Mortgage repayment period
CREDIT_LAST_INSTALLMENT_AMOUNT	Float	Mortgage last installment amount

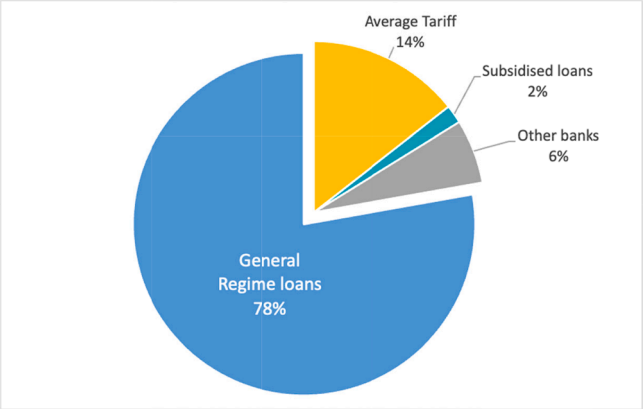
Table 3.3
Dataset – variables from the bank (customer).

Variable name	Type	Description
MARITAL_STATUS	Categorical	Code for the marital status of the first person
MARITAL_STATUS_PERSONS	Numeric	Number of persons equivalent to the marital status
ACADEMIC_QUALIFICATIONS	Categorical	Academic qualifications of the customer
_YEARS_EDUCATION	Numeric	Number of years of education corresponding to the academic qualifications
WORK_BOND	Categorical	Work bond of the customer
CUSTOMER_NATIONALITY	Categorical	Customer's nationality
NATIONALITY_GROUP	Categorical	Grouping the variable CUSTOMER_NATIONALITY into "Portugal" and "Other"
BANK_SEGMENT	Categorical	Bank's segment
ADDRESS_DISTRICT	Categorical	Portuguese district address of the customer
ADDRESS_COUNTRY	Categorical	Country address of the customer
WAGE_DOMICILIATION	Binary	Information on whether the client has the salary domiciled in the bank account (0 = no; 1 = yes)
INSURANCE1	Binary	Customer detention of car insurance in the bank
INSURANCE2	Binary	Customer detention of home insurance in the bank
INSURANCE3	Binary	Customer detention of health insurance in the bank

Notably, these three excluded policy groups also exhibit lower lapse rates, as evidenced by Fig. 3.4.

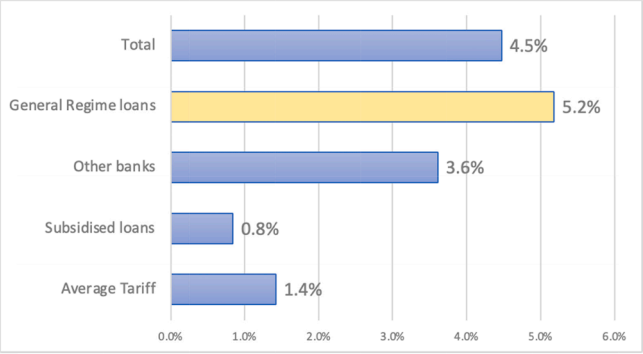
For the purposes of this study, the main focus is on the largest group of policies, which constitutes approximately 78 % of the company's MLI portfolio.

Initial data exploration was conducted using Python within a Jupyter Notebook environment. While the process uncovered a few minor data preparation issues, the overall quality of the data provided by the company and the bank is commendable.



Source: authors

Fig. 3.3. MLI portfolio composition. Source: authors.



Source: authors

Fig. 3.4. MLI lapse rate by type of product. Source: authors.

Descriptive statistics of the dataset are showcased in Table 3.4, providing an insightful overview of the dataset.

In Fig. 3.5 you can see boxplots representing the numeric variables, while Fig. 3.6 showcases the correlation matrix encompassing all variables.

The data exploration highlighted several patterns in the data:

- The policy start year spans from 2001 to 2022, while the end year ranges from 2022 to 2077.
- Cancellations (by lapse) occurred in 2022 and 2023. Given the infrequency of policy lapses, data from 2022 was included primarily to comprehend the data.
- The average insured sum is 64 thousand euros, and the annual premium is 365 euros. These variables exhibit a wide range of values, hinting at potential outliers requiring attention during data preparation.
- Several variables display missing values, most of which are expected, given their specific nature (particularly those related to cancellations and the option of a second insured person).
- The correlation matrix uncovers significant correlations between certain variables. Correlation is critical to address, as it can impact the predictive model and potentially lead to overfitting.

The correlation matrix of the original data reveals some highly correlated variables. As feature engineering created new features and eliminated some initial ones, the correlation analysis was repeated before modeling.

Table 3.4
Data descriptive statistics.

Variable	Count	mean	MIN	25 %	50 %	75 %	MAX	STD
START_YEAR	82,038	2012.8	2001	2007	2012	2019	2022	6.47
END_YEAR	82,038	2043.0	2022	2034	2043	2052	2077	10.59
INSURED_SUM	82,038	51,185	22.85	16,109	40,002	71,346	1,434,058	49,594
INSURED_SUM_initial	81,869	64,427	0.0	25,000	54,051	89,000	1,890,000	56,289
ANNUAL_PREMIUM	82,038	365.6	0.0	108.1	234.9	462.8	16047.1	445.29
ANNUAL_PREMIUM_initial	81,869	228.9	0.0	73.9	151.0	281.7	12359.3	300.67
BIRTH_DATE_PERS1	81,782	1976-05-24	1943-03-27	1970-10-08	1976-06-28	1982-03-11	2003-11-01	NaN
BIRTH_DATE_PERS2	45,059	1977-02-22	1946-02-22	1971-05-06	1977-02-10	1983-01-24	2004-01-23	NaN
%COVER	81,980	97.5	50.0	100.0	100.0	100.0	100.0	10.93
PORTFOLIO_regular	21,810	1.2	1.0	1.0	1.0	1.0	9.0	0.54
PORTFOLIO_PREMIUMS_regular	21,810	231.1	0.0	53.0	104.0	293.0	30000.0	408.78
PORTFOLIO_single	15,093	1.2	1.0	1.0	1.0	1.0	9.0	0.62
PORTFOLIO_PREMIUMS_single	15,093	2597.2	3.0	323.0	963.0	2000.0	338488.0	9146.3
PORTFOLIO_max_durantion	82,038	11.0	0.0	5.0	13.0	16.0	31.0	6.4
YEAR_FIRST_POLICY	82,038	2008.8	1987	2003	2007	2016	2022	7.5
YEAR_LAST_POLICY	82,038	2015.0	2001	2010	2017	2020	2023	5.9
PREVIOUS_CANCEL_POLICIES	5462	1.24	1.0	1.0	1.0	1.0	12.0	0.65
CLIENT_CANCEL_LAST_5Y	5317	1.28	1.0	1.0	1.0	1.0	6.0	0.50
CREDIT_COLLECTION_DAY	82,038	8.61	1.0	2.0	5.0	12.0	31.0	8.94
CREDIT_AMOUNT_INITIAL	82,038	72,576	542	31,400	64,500	99,000	1,625,000	58,201
CREDIT_AMOUNT_OUTSTANDING	82,038	51,898	0.0	16,648	41,121	72,365	143,111	49,373
CREDIT_REPAYMENT_PERIOD	82,038	34.20	1.0	30.0	33.0	40.0	55.0	8.53
CREDIT_LAST_INSTALLMENT_AMOUNT	82,038	253.85	0.0	120.8	227.3	329.6	10821.0	217.82
MARITAL_STATUS_PERSONS	82,038	1.51	1.0	1.0	2.0	2.0	2.0	0.50
_YEARS_EDUCATION	82,038	12.44	0.0	12.0	12.0	16.0	20.0	3.34
WAGE_DOMICILIATION	82,038	0.715	0.0	0.0	1.0	1.0	1.0	0.451
INSURANCE1	82,038	0.148	0.0	0.0	0.0	0.0	1.0	0.355
INSURANCE2	82,038	0.887	0.0	1.0	1.0	1.0	1.0	0.317
INSURANCE3	82,038	0.079	0.0	0.0	0.0	0.0	1.0	0.270
TARGET	82,038	0.083	0.0	0.0	0.0	0.0	1.0	0.275

3.3. Data preparation

In this phase, the data was transformed and prepared to create the final dataset used for modeling. Data preparation is a crucial step in ML because the quality of the data used to train the model directly impacts the Accuracy and effectiveness of that model (Kotsiantis et al., 2006).

It should be noted that missing values in the dataset are not missing due to any problems or errors in data collection. In fact, most of the missing values were expected due to variable specificity, namely, variables associated with the second insured person (an option of the MLI) and those associated with the policy status (canceled / in force). In these cases, the missing values were addressed in feature engineering.

In other cases, the missing or null values corresponded, in fact, to zero, namely in the variables that evaluate the customer portfolio or its previous cancellation behavior. In these cases, zero corresponds exactly to what was supposed to be obtained in the data collection. Missing value imputation with zero was applied in these cases.

It was decided to drop a small group of observations corresponding to policies with exceptionally more than two insured persons (due to the inclusion of the guarantor of the mortgage). These observations represented about 0.3 % of the dataset. Another small set of rows with incomplete information was eliminated (less than 0.1 % of the dataset).

To protect the privacy and confidentiality of the data, some of the variables obtained from the bank were provided in bin format. It was decided to convert these variables into a continuous value, imputing the average of each interval, allowing these variables to be used in the modeling stage. It was also decided to reduce the dimensionality of some of these variables to reduce noisy data.

Feature engineering was also performed, transforming and creating new features from the existing ones. Feature engineering is a fundamental step in ML that can improve models' performance and Accuracy as they can learn more relevant patterns and relationships between the features and the target variable. It can also make the model more interpretable by creating more meaningful and relevant features to the problem.

Table 3.5 lists the features created or transformed from the original variables.

Initial data exploration revealed some variables with data points significantly different from others in the dataset. These points are called outliers and can significantly impact the performance of a classification model. They can skew the data distribution and cause the model to make incorrect predictions. Outliers can be caused by measurement errors, natural variation, or other factors, and they can occur in any type of data.

A function to find and remove outliers based on percentile was applied. The function calculates each column's lower and upper percentile thresholds based on the specified percentiles. It then identifies all rows in the dataset where the column value falls outside the percentile range and returns a copy of the input data with outliers removed from the specified columns.

3.4. Modeling

All the modeling phases were performed using Python on a Jupyter Notebook environment.

Model selection involved careful consideration of the techniques employed in previous literature, as outlined in Table 2.1. Previous research has not highlighted a single model that significantly outperforms the others. Therefore, the approach prioritized the diversity of modeling techniques to cover a broad spectrum of strengths and capabilities. This covered factors such as handling unbalanced data, measuring the importance of features, balancing interpretability with recognizing non-linear patterns, and effectively detecting complex patterns.

The decision included four popular and well-established models: LR, RF, NN, and XGB. These models have a well-documented history of successful applications in various classification tasks. Table 3.6 provides a detailed comparison of each model's characteristics and advantages.

LR is a linear classification algorithm that models the relationship between input features and the binary outcome. It is known for its interpretability and ability to deal effectively with high-dimensional data.

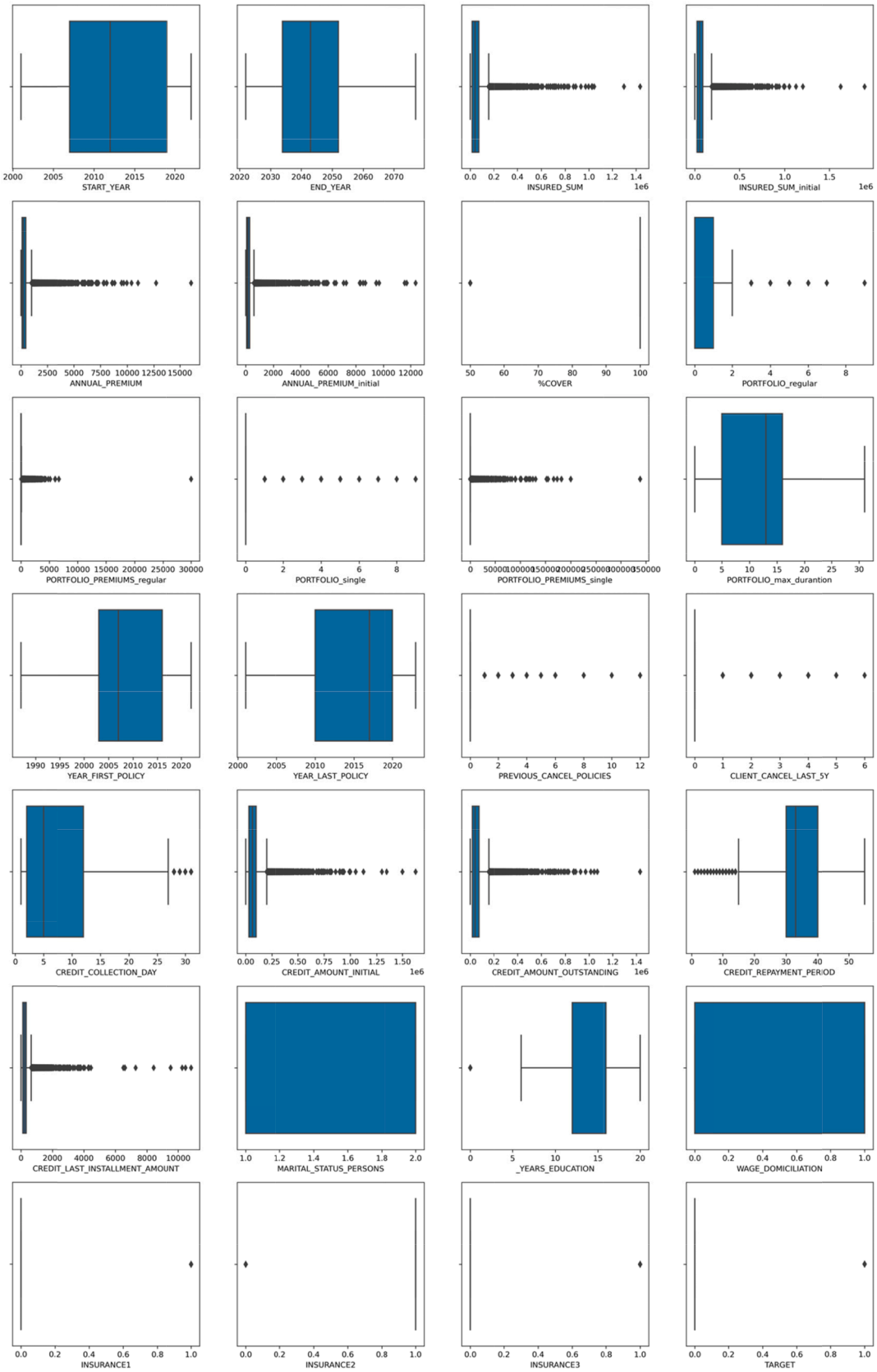


Fig. 3.5. Boxplots of the numeric variables.

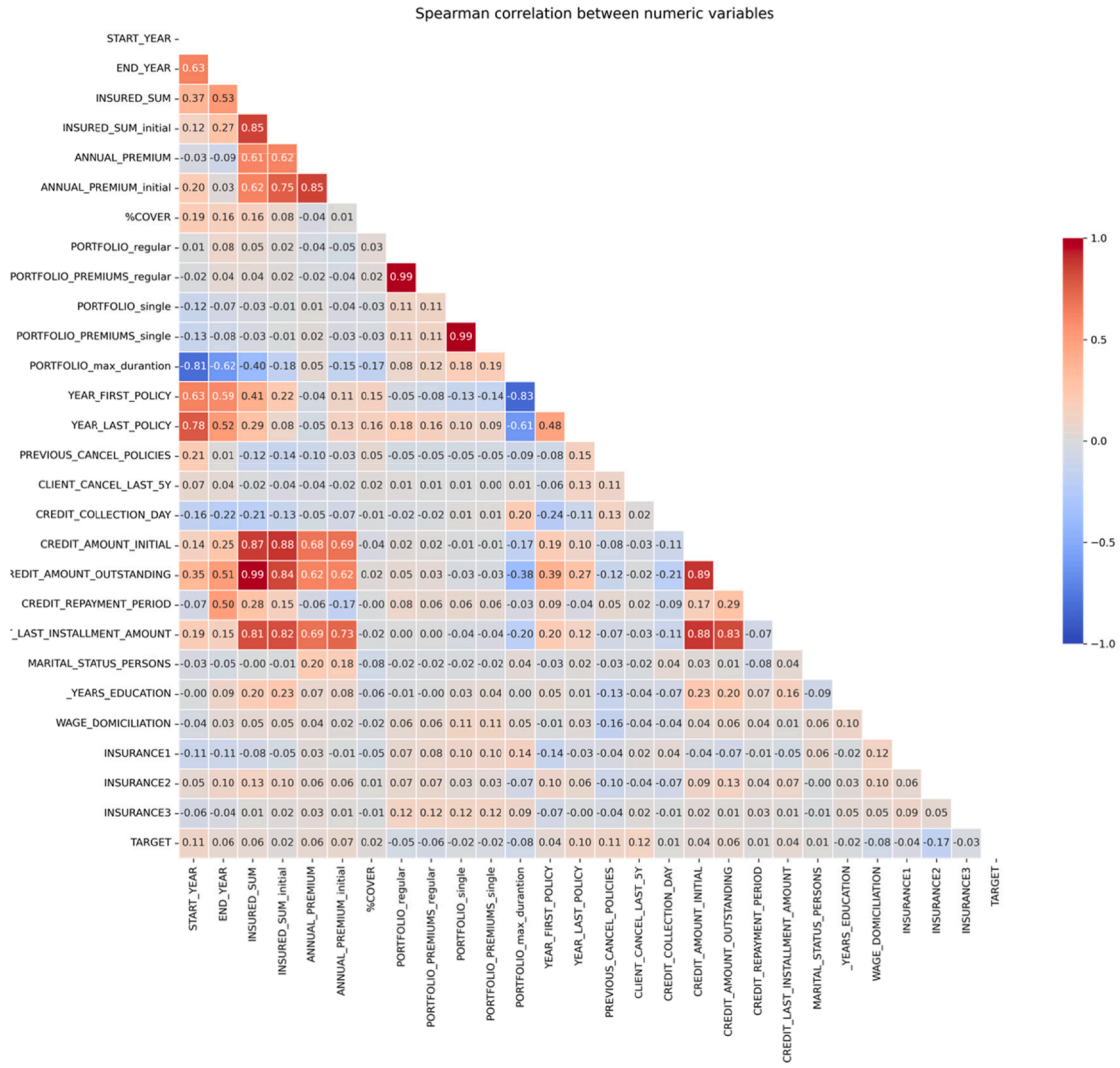


Fig. 3.6. Spearman correlation matrix.

RF is an ensemble learning method that combines several decision trees to make predictions. It leverages the power of decision trees and uses randomization to improve performance and mitigate overfitting. For further reading, see Breiman (2001).

NN is a deep learning algorithm inspired by the structure and function of the human brain. It can learn complex patterns and relationships in data, which makes it suitable for tasks with complex dependencies. See Bishop (2006) for further reading.

XGB is a gradient-boosting algorithm that sequentially adds decision trees to minimize a loss function. It excels at capturing non-linear relationships and interactions in data, leading to solid predictive performance (Chen & Guestrin, 2016).

Data splitting was conducted before initializing the models. This process allows us to train the model on one part of the data and evaluate its performance on another separate part, allowing the assessment of how well the model generalizes to new and unseen data.

For this purpose, 75 % of the data was considered for training and the remaining 25 % for testing. The "stratify" option was also applied during the data-splitting process to maintain consistent class proportions in the training and test sets. This step is crucial to ensure that the model is trained on a representative sample of both the minority and majority classes, ultimately enhancing its accuracy and ability to generalize effectively.

Imbalanced data presents a frequent challenge in ML, particularly in classification tasks where classes are not evenly distributed within the dataset. Predicting MLI lapses is no exception, given the infrequent occurrence of lapses, as evident in Fig. 3.4. As anticipated, the dataset used for this study exhibited a substantial imbalance, with the minority class accounting for roughly 9.5 % of the total data.

One common approach to address class imbalance is to use resampling techniques, such as oversampling or undersampling. Oversampling involves creating synthetic samples of the minority class to increase its representation in the dataset, while undersampling involves randomly removing samples from the majority class to reduce its representation.

Considering the severe class imbalance in the dataset, the main goal of this step was to enhance the representation of the minority class. Considering this objective, five techniques were considered:

- **Oversampling with SMOTE**: SMOTE (Synthetic Minority Oversampling Technique) creates synthetic minority class samples by interpolating between existing minority class samples. Specifically, it generates synthetic samples along the line segments that connect some or all the k minority class nearest neighbors of each minority class sample. A user-specified oversampling ratio determines the number of synthetic samples to generate.
- **Random Under Sampling (RUS) + Oversampling with SMOTE**: combines RUS and SMOTE. RUS reduces the number of samples in the majority class by randomly selecting a subset of the data. Conversely, SMOTE increases the number of samples in the minority

Table 3.5
Features transformation and new features.

Original variable	New / Derived variable	Type	Transformation / Calculation
CLIENT_ID	Number_Mortgages	Numeric	Number of mortgages of the customer
	Volume_Mortgages	Float	Volume (insured sum) of mortgages of the customer
	Number_Mortgages_group	Categorical	Grouping the number of mortgages to reduce dimensionality
START_YEAR	years_to_end-date	Float	Difference between the end year and the current year
	elapsed_years	Float	Difference between the current year and the start year
	Policy_Term_years	Float	Difference between the end year and the start year
	%years_to_the_end	Float	Percentage of the years to the end-date over the policy term years
END_YEAR			
INSURED_SUM	INSURED_SUM_bins	Categorical	Binning of the insured sum
	%capital_in_debt	Float	Percentage of the current insured sum over the initial insured sum
INSURED_SUM_initial			
ANNUAL_PREMIUM	monthly_premium	Float	Corresponds to the annual premium divided by 12
	current_price	Float	Corresponds to the annual premium over the insured Sum
ANNUAL_PREMIUM_initial	annual_premium_growth_group	Float	Corresponds to the "current price" over the "initial price" (that corresponds to the initial annual premium over the initial insured sum) – grouped
BIRTH_DATE_PERS1	age1	Numeric	Age for each insured person
	age2		
BIRTH_DATE_PERS2	age_mean	Float	Mean of the age of the two insured persons (or the age of the insured person if there is only one)
GENDER_PERS1	gender_group	Categorical	Gender of the insured person or its combination if there are two insured persons (grouped)
GENDER_PERS2			
YEAR_FIRST_POLICY	client_years	Float	Number of years as client: the difference between the year of the last contracted policy ("year last policy") and the first one ("year first policy")
YEAR_LAST_POLICY			
PREVIOUS_CANCEL_POLICIES	PREVIOUS_CANCEL_POLICIES_v2	Binary	Convert the variable into binary: 0 (zero cancellations), 1 (one or more cancellations)
CLIENT_CANCEL_LAST_5Y	CLIENT_CANCEL_LAST_5Y_v2	Binary	Convert the variable into binary: 0 (zero cancellations), 1 (one or more cancellations)
CREDIT_COLLECTION_DAY	collection_day_delta	Numeric	The difference in days between the company's premium collection day (that is fixed on the 23rd of each month) and the mortgage installment collection day
CREDIT_LTV_INITIAL_BINS	CREDIT_LTV_INITIAL_mid	Float	Convert to continuous value (average of each bin)
CREDIT_AMOUNT_BINS	CREDIT_AMOUNT_mid	Numeric	Convert to continuous value (average of each bin)
CREDIT_SPREAD_BINS	CREDIT_SPREAD_mid	Float	Convert to continuous value (average of each bin)
CREDIT_EFFORT_RATE_BINS	CREDIT_EFFORT_RATE_mid	Float	Convert to continuous value (average of each bin)
CREDIT_AMOUNT_OUTSTANDING	CREDIT_AMOUNT_OUTSTANDING_bins	Categorical	Binning of the variable
CREDIT_LAST_INSTALLMENT_AMOUNT	CREDIT_LAST_INSTALLMENT_AMOUNT_bins	Categorical	Bin the last installment amount
	%insurance_premium	Float	Percentage of the premium over the installment
BANK_SEGMENT	BANK_SEGMENT_GROUP	Categorical	Grouping the customer segment to reduce dimensionality
ADDRESS_COUNTRY	addr_country	Categorical	Group countries into two: Portugal and Other (that represent a small set of observations)
INSURANCE1	INSURANCE_TOTAL	Numeric	Total insurance equipment
INSURANCE2			
INSURANCE3			

Table 3.6
Model's comparative table.

Model	Interpretability	Nonlinearity handling	Robustness	Complex patterns	Imbalanced data handling	Feature importance
LR	Yes	No	Moderate	No	Yes (with weighting)	Yes
RF	No	Yes	High	No	Yes (through ensemble)	Yes
NN	No	Yes	Moderate	Yes	Yes (with techniques)	No
XGB	No	Yes	Moderate	Yes	Yes (with weighting)	Yes

class by generating synthetic samples based on the existing minority samples.

- **SMOTETomek**: is a combination of two techniques: SMOTE (explained above) and Tomek links. Tomek links are pairs of instances from different classes that are very close to each other. Removing the majority class instance from each Tomek link can help to increase the distance between the classes, making it easier for the classifier to differentiate between them. SMOTETomek combines these two techniques by first oversampling the minority class using SMOTE and then removing the majority class instances linked to the minority class instances using Tomek links. This combination helps to balance the class distribution while also reducing the overlap between the classes.
- **ADASYN**: ADASYN stands for "Adaptive Synthetic Sampling", a method for oversampling imbalanced data, similar to SMOTE. ADASYN also creates synthetic samples for the minority class, but it adapts the distribution of synthetic samples according to the density distribution of the input dataset. It generates more synthetic samples in regions of the feature space where the density of minority examples is low and fewer synthetic samples where the density is high. ADASYN can more effectively balance the class distribution in regions with complex and highly non-linear decision boundaries.
- **SMOTEENN**: short for "SMOTE with Edited Nearest Neighbors", combines SMOTE and ENN (Edited Nearest Neighbors). ENN is an under-sampling technique that removes noisy or redundant instances from the dataset. SMOTEENN is thus designed to improve the balance of class distributions while also cleaning the dataset.

It was decided to try these various techniques with different characteristics to determine which would work best on the dataset. Table 3.7 summarizes a comparison of the five techniques considered.

Min-max normalization (also known as min–max scaling) was applied to equalize the importance of the features due to the difference in scales in some variables. Normalization is essential to improve the performance of the LR and NN models. As RF and XGB can handle both scaled and unscaled data, applying min–max was unnecessary.

One hot encoding was also applied to transform all the categorical variables into a numerical format that could be used in the models. This technique has the advantage of not introducing any ordinality or hierarchy between the categories, which could lead to biased models. It can also improve the model's accuracy by reducing data noise and making the categories' relationships more explicit to the algorithm.

Data analysis revealed that the lapse rate is significantly higher in the first few years of the policy. This fact is mainly due to customer inertia and because, with increasing age, customers fear losing insurance protection or being penalized in a new insurance acceptance process. Fig. 3.7 shows the evolution of the lapse rate by the start year.

Given this fact, it was decided to subset the data to exclude the period before 2011, also helping to resolve the class imbalance and improve the model's performance.

From the set of variables available after the feature engineering process, a subset of 41 variables was selected. The selection process was primarily guided by domain expertise and insights acquired through previous data exploration, including visualizations and summary statistics to identify patterns and associations with policy lapse.

Derived features were created to capture more complex relationships within the data through transformations and aggregations of existing variables. Highly correlated variables were carefully reviewed to avoid redundancy and multicollinearity. In cases where multiple variables conveyed similar information, the most representative one was chosen to streamline the model and improve interpretability.

The final set of variables was chosen to ensure a diverse selection, incorporating multiple variables from each category and origin. This approach considered both the original and derived variables to maintain comprehensive coverage of relevant information. Table 3.8 provides a detailed overview of the selected variables, organized by origin and category.

Table 3.7
Class imbalance techniques comparison table.

Technique	Advantages	Disadvantages	Additional characteristics
SMOTE	Generates synthetic data for the minority class, preserving information	It may introduce noise if not used carefully	Effective in preventing overfitting
RUS + SMOTE	Combines strengths of undersampling and oversampling	Requires careful parameter tuning and validation	Balances class distribution while reducing the risk of overfitting
SMOTETomek	Focuses on cleaning and oversampling, improving class balance, and noise reduction	It may require parameter tuning for optimal performance	Effective in addressing both class imbalance and noise
ADASYN	Adapts to the distribution of the minority class, generating more synthetic data for sparse regions	Sensitive to noisy data; parameter tuning is important	This is particularly useful when the distribution of the minority class varies
SMOTEEN	Combines SMOTE with Tomek Links and focuses on oversampling and cleaning	May require parameter tuning for optimal results	Offers a balanced approach by simultaneously addressing class imbalance and noise

To improve the model's performance, hyperparameter tuning was performed using Optuna (Akiba et al., 2019), an automated hyperparameter tuning method. It uses a tree-structured Parzen estimator algorithm to explore the hyperparameter space efficiently and can be used with a wide range of ML frameworks.

3.5. Evaluation metrics

An appropriate evaluation metric is crucial in achieving the optimal classifier during classification training. It is, therefore, an important factor in discriminating between different classifiers and obtaining the optimal one for the given problem (Hossin & Sulaiman, 2015).

In a binary classification problem, such as the one we have in this study, the best or optimal solution assessment is usually based on the confusion matrix (see Table 3.9). This matrix shows the number of correct and incorrect predictions made by the model compared to the actual outcomes. It is composed of four elements:

- True positives (*tp*): number of correctly classified positive
- False positives (*fp*): number of incorrectly classified positive
- True negatives (*tn*): number of correctly classified negative
- False negatives (fn): number of incorrectly classified negative

The confusion matrix output also allows the calculation of commonly used metrics to assess the models' performance (see Table 3.10).

Another valuable metric to evaluate a binary classification model is the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The ROC curve is a plot of true positive rate (TPR) versus false positive rate (FPR) at various thresholds. The AUC measures the model's ability to distinguish between positive and negative classes across all possible thresholds. Maximizing the AUC can help the company achieve the best possible balance between true positives and false positives.

The choice of model evaluation metrics should align with the defined business objectives: maximizing customer retention while minimizing costs. To make a well-informed decision, one must consider various aspects, such as comparing the costs associated with false positives (predicting a customer will lapse when they do not) to the costs linked to false negatives (predicting a customer will not lapse when they do).

The costs of false positives can include administrative costs associated with trying to retain the customer, and reputational damage if the customer cancels their policy due to perceived aggressive tactics on the insurer's part.

The costs of false negatives can include the loss of revenue from the customer's premium payments, the costs associated with acquiring a new customer to replace the lost revenue, and the administrative costs associated with managing the policy cancellation process.

Table 3.11 illustrates a profit and loss confusion matrix, incorporating economic assumptions specific to this study. Note that this is a simplified assessment that only considers the costs associated with the retention contact and the retention proposal, as well as the average annual premium paid by the customer.

Considering these assumptions, we observe that the costs associated with failing to identify actual lapses (false negatives) significantly outweigh (by 80 %) the costs related to mistakenly retaining a customer (false positives). Given this scenario, the primary metric of concern should be Recall, as it quantifies the model's capability to detect genuine lapses accurately. Prioritizing Recall aligns seamlessly with this study's primary business goal of minimizing the expenses tied to customer retention.

However, it is crucial to strike a balance with Precision to ensure the model's overall effectiveness in achieving these objectives. In other words, while focusing on correctly identifying lapses, we must also be mindful of not overly aggressive retention tactics that could lead to unnecessary costs.

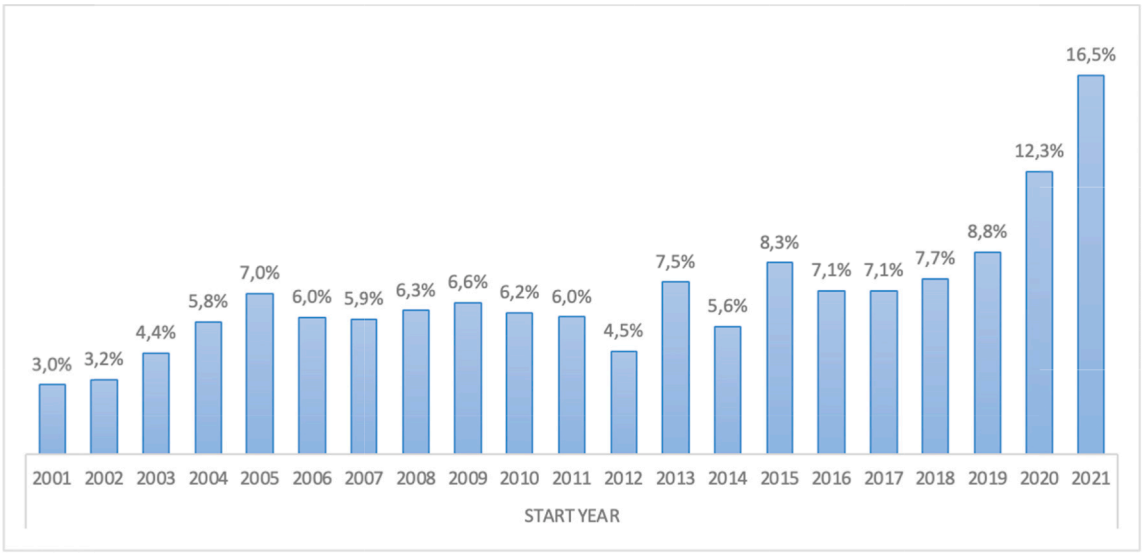


Fig. 3.7. Lapse rate by start year (in 2022).

Table 3.8
Variables selected for modeling.

Data origin	Category	Original variable	Derived variable 1	Derived variable 2
INSURANCE COMPANY	POLICY DETAILS	START_YEAR	years_to_end-date	elapsed_years
		END_YEAR	%years_to_the_end	Policy_Term_years
		INSURED_SUM	INSURED_SUM_bins	
		INSURED_SUM_initial	%capital_in_debt	
		INSURED_PERSONS		
		ANNUAL_PREMIUM	monthly_premium	current_price
		ANNUAL_PREMIUM_initial	annual_premium_growth_group	
		%COVER		
	CUSTOMER DATA	CLIENT_ID	Number_Mortgages	Volume_Mortgages
		BIRTH_DATE_PERS1	age1	age2
		BIRTH_DATE_PERS2	age_mean	
		GENDER_PERS1	gender_group	
		GENDER_PERS2		
	BEHAVIOURAL DATA	PORTFOLIO_regular		
		PORTFOLIO_single		
		PORTFOLIO_max_duration		
		YEAR_FIRST_POLICY	client_years	
		YEAR_LAST_POLICY		
		PREVIOUS_CANCEL_POLICIES	PREVIOUS_CANCEL_POLICIES_v2	
		CLIENT_CANCEL_LAST_5Y	CLIENT_CANCEL_LAST_5Y_v2	
BANK	MORTGAGE DATA	CREDIT_COLLECTION_DAY	collection_day_delta	
		CREDIT_LTV_INITIAL_BINS	CREDIT_LTV_INITIAL_mid	
		CREDIT_AMOUNT_BINS	CREDIT_AMOUNT_mid	
		CREDIT_SPREAD_BINS	CREDIT_SPREAD_mid	
		CREDIT_EFFORT_RATE_BINS	CREDIT_EFFORT_RATE_mid	
		CREDIT_AMOUNT_OUTSTANDING	CREDIT_AMOUNT_OUTSTANDING_bins	
		CREDIT_REPAYMENT_PERIOD		
		CREDIT_LAST_INSTALLMENT_AMOUNT	CREDIT_LAST_INSTALLMENT_AMOUNT_bins	%insurance_premium
	CUSTOMER DATA	MARITAL_STATUS_PERSONS		
		_YEARS_EDUCATION		
		BANK_SEGMENT	BANK_SEGMENT_GROUP	
		ADDRESS_COUNTRY	addr_country	
		WAGE_DOMICILIATION		
		INSURANCE1	INSURANCE_TOTAL	
		INSURANCE2		
		INSURANCE3		

Table 3.9
Confusion matrix.

	Predicted negative	Predicted positive
Actual negative	True negative (*tn*)	False positive (*fp*)
Actual positive	False negative (*fn*)	True positive (*tp*)

4. Results and discussion

The results of MLI lapse prediction models developed in the previous chapters are presented in this chapter. This stage was performed in Python on a Jupyter Notebook environment using Scikit-learn (Pedregosa et al., 2011) and Imbalanced-learn (Lemaître et al., 2017) libraries.

For an initial evaluation, a set of baseline models was employed using default settings before addressing class imbalance. Table 4.1

Table 3.10
Main data classification metrics.

Metric	Formula	Evaluation
Accuracy (*acc*)	$\frac{tp+tn}{tp+fp+tn+fn}$	Measures the percentage of correct predictions made by the model
Precision (*p*)	$\frac{tp}{tp+fp}$	Measures the proportion of true positives among all positive predictions. It is a useful metric when the cost of false positives is high.
Recall (*r*)	$\frac{tp}{tp+fn}$	Measures the proportion of true positives among all actual positives. It is a useful metric when the cost of false negatives is high.
F1-score (*F1*)	$\frac{2*p*r}{p+r}$	This is the harmonic mean of Precision and Recall. It is a good metric when you want to balance Precision and Recall.

Table 3.11
Profit and Loss confusion matrix.

Costs / Profits	True Negative		False Positive	
Costs	–	€ 0	Retention contact (€20) Retention proposal (€180)	– €200
Profits	–		–	
Costs / Profits	**False Negative**		**True Positive**	
Cost	Annual premium (€360)	– €360	Retention contact (€20) Retention proposal (€180)	+ €160
Profits	–		Annual premium (€360)	

summarizes the performance metrics obtained from these four baseline models.

The overall results of this first experiment are pretty revealing. All models achieve high Accuracy scores, indicating their ability to make correct predictions for both the positive and negative classes. However, Precision scores remain modest, suggesting that while the models make correct predictions, they generate many false positives.

Considering Recall, a crucial metric for this study, we notice that, except for XGB, all models have shallow scores, demonstrating that the models have difficulty detecting true positive cases.

XGB presents the highest scores in terms of Recall and AUC. However, the disparities between the training and test datasets results suggest that the model might be overfitting the data.

In summary, these primary results reveal promising predicting results but underscore areas requiring improvement.

In the second step, the five selected techniques to mitigate class imbalance were implemented to optimize the model's performance further.

Table 4.2 displays the metrics results on the test set for all models, using default settings and the five resampling techniques. These results are presented prior to cross-validation (CV) and hyperparameter tuning. The best score for each metric is highlighted for reference.

The implementation of class imbalance techniques led to a decrease in the Accuracy and Precision metrics for all the models. However, there was a substantial and consistent improvement in the Recall metric. Among the various techniques applied, SMOTEENN obtained the highest Recall score, while SMOTE and SMOTETomek showed a more balanced compromise between Accuracy and Recall.

Among the models evaluated, RF and XGB consistently obtained the highest scores, with one notable exception: LR outperformed them regarding the Recall metric in most cases. This difference in performance can be attributed to their distinct underlying algorithms and their approach to handling the data. Meanwhile, LR, as a linear model, has limitations in capturing complex non-linear relationships in the dataset. However, RF and XGB are ensemble methods known for effectively capturing complex non-linear patterns.

A CV was performed to improve understanding of the behavior of

Table 4.1
Results with default settings and without resampling.

Measure	LR		NN		RF		XGB	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Accuracy	0.9075	0.9074	0.9100	0.9071	0.9053	0.9051	0.8611	0.8082
Precision	0.5969	0.5929	0.6252	0.5517	1.0000	0.6666	0.3922	0.2625
Recall	0.0786	0.0814	0.1305	0.1166	0.0024	0.0024	0.8435	0.5637
F1 Score	0.1389	0.1431	0.2159	0.1925	0.0048	0.0048	0.5355	0.3583
AUC	0.5365	0.5377	0.5611	0.5533	0.5012	0.5011	0.8532	0.6988

Table 4.2
Results on the test set with default settings and resampling.

Sampling technique	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
ADASYN	LR	0.6859	0.1859	**0.6829**	0.2923	0.6845
	NN	0.6826	0.1833	0.6780	0.2886	0.6805
	RF	0.7608	0.2066	0.5346	0.2980	0.6596
	XGB	**0.7641**	**0.2281**	0.6221	**0.3338**	**0.7006**
RUS + SMOTE	LR	0.7299	0.2074	0.6537	0.3149	**0.6958**
	NN	0.7474	0.2089	0.5954	0.3092	0.6794
	RF	**0.8016**	**0.2457**	0.5261	**0.3350**	0.6783
	XGB	0.6318	0.1715	**0.7509**	0.2792	0.6851
SMOTE	LR	0.7274	0.2060	**0.6549**	0.3134	0.6950
	NN	0.7291	0.2059	0.6488	0.3126	0.6932
	RF	**0.7971**	**0.2407**	0.5273	0.3305	0.6764
	XGB	0.7736	0.2374	0.6258	**0.3443**	**0.7074**
SMOTEENN	LR	0.5929	0.1601	**0.7740**	0.2653	0.6739
	NN	0.6006	0.1597	0.7521	0.2635	0.6684
	RF	0.5607	0.1473	0.7570	0.2466	0.6485
	XGB	**0.6858**	**0.1949**	0.7375	**0.3084**	**0.7089**
SMOTETomek	LR	0.7294	0.2066	**0.6513**	0.3137	0.6944
	NN	0.7387	0.2112	0.6403	0.3177	0.6947
	RF	**0.8059**	**0.2493**	0.5188	0.3368	0.6774
	XGB	0.7731	0.2365	0.6233	**0.3429**	**0.7061**

these models and ensure their generalizability. This approach allows the models' performance to be assessed across various subsets of the data, providing insights into their consistency and robustness.

Table 4.3 presents the CV results, displaying all models' mean and standard deviation (SD) values under the various resampling techniques. The best score for each metric is highlighted for reference.

XGB consistently outperforms other models across all evaluation metrics and resampling methods, solidifying its position as the leading candidate for this predictive task. While RF delivers competitive performance and stands as a credible alternative, both LR and NN exhibit relatively lower performance when compared to XGB and RF.

Having identified XGB as the leading candidate with consistently strong performance across various metrics and resampling techniques, it was decided to focus the next modeling steps on improving the XGB model.

The next step involved hyperparameter tuning, using Optuna, to optimize the model's configuration to achieve the best possible predictive performance. This step includes an iterative process of systematically adjusting hyperparameters while cross-validating the model to find the optimal set of values.

First, we employed a set of XGB baseline models with default settings to provide a benchmark for evaluating performance after hyperparameter tuning. The results of these baseline models, using the different resampling techniques, can be seen in Table 4.4:

Among the five resampling techniques, SMOTETomek exhibited the highest scores across all metrics, except for Recall, where RUS + SMOTE scored higher. Furthermore, an economic evaluation was conducted based on the Profit and Loss confusion matrix, as detailed in Table 3.11. While all results indicated negative values, it is worth noting that SMOTETomek demonstrated the most favorable outcome in minimizing losses.

Another prominent observation from these results is the evident tendency of the models to overfit, as evidenced by their superior performance on the training data compared to the test data. This issue is most evident in the Precision, Recall, and AUC metrics. This outcome was somewhat anticipated, given the intrinsic characteristics of the dataset and the challenges observed in prior literature. Azzone et al. (2022) also identified this propensity for overfitting in their lapse prediction model.

Given the concern regarding overfitting, CV was conducted on the XGB baseline models to assess their performance and generalization ability to unseen data. The outcomes of this CV analysis for XGB, undertaken with various resampling techniques prior to hyperparameter tuning, are summarized in Table 4.5.

The results significantly enhance all performance metrics when contrasted with the baseline models. This noteworthy advancement in model performance, particularly regarding the Recall metric, underscores the effectiveness of resampling techniques in mitigating class imbalance. This improvement might appear counterintuitive, given the initial apprehension regarding overfitting.

Given this observation, hyperparameter tuning was performed as an additional measure to fine-tune the model's parameters for improved generalization on unseen data. This step alleviated potential concerns regarding the model's ability to generalize effectively and enhance its overall performance further.

A comprehensive series of experiments were conducted using

Table 4.3
CV results on all models with resampling (mean and SD).

Resampling Technique	Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
ADASYN	LR	0.659 (+/− 0.004)	0.672 (+/−− 0.006)	0.625 (+/− 0.003)	0.648 (+/− 0.004)	0.659 (+/− 0.004)
	NN	0.696 (+/− 0.006)	0.683 (+/− 0.014)	0.736 (+/− 0.026)	0.708 (+/− 0.007)	0.696 (+/− 0.006)
	RF	0.752 (+/− 0.006)	0.787 (+/− 0.009)	0.692 (+/− 0.008)	0.736 (+/− 0.006)	0.752 (+/− 0.006)
	XGB	**0.875 (+/− 0.002)**	**0.822 (+/− 0.004)**	**0.960 (+/− 0.004)**	**0.885 (+/− 0.002)**	**0.875 (+/− 0.002)**
RUS + SMOTE	LR	0.684 (+/− 0.006)	0.705 (+/− 0.009)	0.631 (+/− 0.008)	0.666 (+/− 0.006)	0.684 (+/− 0.006)
	NN	0.707 (+/− 0.006)	0.725 (+/− 0.006)	0.670 (+/− 0.017)	0.696 (+/− 0.009)	0.707 (+/− 0.006)
	RF	0.734 (+/− 0.006)	**0.794 (+/− 0.013)**	0.633 (+/− 0.005)	0.704 (+/− 0.005)	0.734 (+/− 0.006)
	XGB	**0.796 (+/− 0.005)**	0.726 (+/− 0.006)	**0.950 (+/− 0.003)**	**0.823 (+/− 0.004)**	**0.796 (+/− 0.005)**
SMOTE	LR	0.684 (+/− 0.006)	0.705 (+/− 0.008)	0.632 (+/− 0.005)	0.667 (+/− 0.006)	0.684 (+/− 0.006)
	NN	0.725 (+/− 0.006)	0.733 (+/− 0.009)	0.707 (+/− 0.007)	0.720 (+/− 0.005)	0.725 (+/− 0.006)
	RF	0.749 (+/− 0.008)	0.808 (+/− 0.011)	0.654 (+/− 0.008)	0.723 (+/− 0.008)	0.749 (+/− 0.008)
	XGB	**0.875 (+/− 0.003)**	**0.820 (+/− 0.003)**	**0.961 (+/− 0.003)**	**0.885 (+/− 0.003)**	**0.875 (+/− 0.003)**
SMOTEENN	LR	0.730 (+/− 0.004)	0.770 (+/− 0.004)	0.779 (+/− 0.007)	0.775 (+/− 0.004)	0.718 (+/− 0.004)
	NN	0.776 (+/− 0.006)	0.806 (+/− 0.005)	0.822 (+/− 0.010)	0.814 (+/− 0.005)	0.765 (+/− 0.006)
	RF	0.768 (+/− 0.006)	0.761 (+/− 0.005)	0.889 (+/− 0.004)	0.820 (+/− 0.004)	0.739 (+/− 0.007)
	XGB	**0.891 (+/− 0.003)**	**0.863 (+/− 0.004)**	**0.972 (+/− 0.002)**	**0.914 (+/− 0.002)**	**0.872 (+/− 0.004)**
SMOTETomek	LR	0.684 (+/− 0.006)	0.706 (+/− 0.006)	0.630 (+/− 0.010)	0.666 (+/− 0.007)	0.684 (+/− 0.006)
	NN	0.727 (+/− 0.003)	0.733 (+/− 0.007)	0.713 (+/− 0.008)	0.723 (+/− 0.003)	0.727 (+/− 0.003)
	RF	0.747 (+/− 0.002)	0.807 (+/− 0.003)	0.651 (+/− 0.007)	0.720 (+/− 0.004)	0.747 (+/− 0.002)
	XGB	**0.877 (+/− 0.003)**	**0.824 (+/− 0.005)**	**0.959 (+/− 0.004)**	**0.886 (+/− 0.003)**	**0.877 (+/− 0.003)**

Table 4.4
XGB results before hyperparameter tuning.

Measure	ADASYN		RUS + SMOTE		SMOTE		SMOTEENN		SMOTETomek	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Accuracy	0.9029	**0.7793**	0.8408	0.6515	0.8992	0.7791	0.9080	0.6958	0.8999	0.7777
Precision	0.8489	0.2395	0.7627	0.1803	0.8418	0.2406	0.8681	0.1881	0.8422	**0.2416**
Recall	0.9842	0.6087	0.9894	**0.7533**	0.9830	0.6148	0.9954	0.6646	0.9841	0.6269
F1 Score	0.9116	0.3438	0.8614	0.2910	0.9070	0.3458	0.9274	0.2932	0.9077	**0.3488**
AUC	0.9014	0.7030	0.8408	0.6970	0.8992	0.7056	0.8887	0.6818	0.8999	**0.7102**
Economic evaluation	− €353,760		− €537,280		− €352,560		− €483,840		**− €351,760**	

Table 4.5
CV results in XGB with resampling before hyperparameter tuning.

Resampling technique	Accuracy	Precision	recall	F1 Score	AUC
ADASYN	**0.882 (+/− 0.002)**	0.834 (+/− 0.003)	0.958 (+/− 0.003)	**0.880 (+/− 0.002)**	0.892 (+/− 0.001)
RUS + SMOTE	0.806 (+/− 0.005)	0.739 (+/− 0.006)	0.944 (+/− 0.004)	0.806 (+/− 0.005)	0.829 (+/− 0.004)
SMOTE	0.880 (+/− 0.002)	0.829 (+/− 0.003)	0.958 (+/− 0.003)	**0.880 (+/− 0.002)**	0.889 (+/− 0.002)
SMOTEENN	0.877 (+/− 0.004)	**0.848 (+/− 0.006)**	**0.965 (+/− 0.003)**	0.858 (+/− 0.005)	**0.903 (+/− 0.003)**
SMOTETomek	0.878 (+/− 0.003)	0.826 (+/− 0.005)	0.957 (+/− 0.002)	0.878 (+/− 0.003)	0.887 (+/− 0.003)

Optuna. The primary objective was to find the optimal set of hyperparameters that would produce the best model performance while concurrently minimizing the risk of overfitting. Parameters were meticulously adjusted to fine-tune the model's performance, ensuring that it could strike a delicate balance between delivering good predictive results and preventing overfitting to the training data.

Considering these objectives, the best trial found by Optuna selected the hyperparameters seen in Table 4.6.

The results following the Optuna hyperparameter tuning, as consolidated in Table 4.7, show notable improvements in mitigating the overfitting problem and increases in the Accuracy and Precision metrics compared to the "baseline" model. However, this improvement in the model came at a cost, as the Recall and AUC metrics experienced reductions, especially in Recall.

Within the five resampling techniques, SMOTETomek maintains its prominent standing by consistently delivering the best performance, particularly excelling in Accuracy, Precision, and F1 Score. Moreover, it distinguishes itself by yielding the most favorable economic evaluation, reflecting the lowest loss when examining the values within the confusion matrix. SMOTEENN emerges as the top performer regarding Recall and AUC scores but presents the least favorable economic outcome compared to the other resampling techniques.

Table 4.6
Optuna's best parameters.

Hyperparameter	Value
max_depth	4
learning_rate	0.00020906
min_child_weight	6
gamma	0.00306043
subsample	0.7
colsample_bytree	0.5
n_estimators	190
alpha	0.86942126
lambda	0.71728375
reg_lambda	10.7509136
reg_alpha	10.0095337

In the last stage of this analysis, CV was conducted on the finely tuned XGB models to rigorously assess their performance and ability to generalize effectively to previously unseen data. The results of this CV analysis for XGB, using the five resampling techniques and after hyperparameter tuning, are presented in Table 4.8.

Hyperparameter tuning and CV results showed that SMOTEENN stands out as the most effective resampling approach for this predictive model.

Fig. 4.1 presents the learning curve using the Accuracy metric on the XGB model with the SMOTEENN resampling technique. This visual representation of the model's performance on the training and validation sets is a function of the number of training examples.

In this case, the training and CV scores increase and converge as the training set size increases. This curve suggests that the model's performance improves and starts to generalize better with more data.

Understanding the most influential features contributing to the model's predictions is essential for interpretability and actionable insights. To achieve this, we employed SHAP (Lundberg et al., 2017), which stands for "SHapley Additive exPlanations," a powerful tool for explaining the output of ML models". By analyzing the importance of features through SHAP values, we can pinpoint the key drivers behind the model's predictions. This insight is invaluable for decision-makers and stakeholders seeking to understand the factors that predict customer lapses.

Fig. 4.2 shows the twenty top features that contribute the most to the predictive model. The first and most revealing observation is that the four most important features represent information collected from the Bank. Additionally, of the seven characteristics that have an incomparably greater weight than the others, three originate directly from the bank ("INSURANCE2″, "WAGE_DOMICILIATION" and "INSURANCE_TOTAL"), and two use information from the bank ("%insurance_premium" and "collection_day_delta").

Another observation is that the other top features are mostly sociodemographic variables, like gender, age, education, and geographic location. And variables associated with the insurance, like the monthly premium, insured persons, the capital in debt, and annual premium growth. The variables associated with income (customer bank segment) and those associated with previous cancellation behavior do not play an important role in predicting lapses.

The top two features, "INSURANCE2" and "WAGE_DOMICILIATION", notably influence the prediction of lapses. Their importance is evident when we observe the SHAP summary plot displaying the importance and impact of features on the model output (see Fig. 4.3).

The adverse influence of these two features, represented by the red marks on the left, indicates that higher values of these variables significantly raise the likelihood of lapses. Conversely, the blue markers on the right suggest that values surpassing the reference point positively reduce the probability of lapses. Another variable with a distinct impact on lapses is 'annual_premium_growth_group.' The SHAP plot reveals that lower values of this variable are associated with a negative effect on lapse likelihood.

In practical terms, this analysis implies that maintaining or improving these variables, namely preventing them from falling below

Table 4.7
XGB results after hyperparameter tuning.

Measure	ADASYN		RUS + SMOTE		SMOTE		SMOTEENN		SMOTETomek	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Accuracy	0.8457	0.8451	0.8457	0.8458	0.8494	0.8469	0.7983	0.8060	0.8511	**0.8512**
Precision	0.2826	0.2813	0.2829	0.2831	0.2887	0.2826	0.2288	0.2415	0.2911	**0.2927**
Recall	0.4069	0.4058	0.4077	0.4070	0.4008	0.3973	0.4746	**0.4872**	0.3960	0.3997
F1 Score	0.3336	0.3323	0.3341	0.3339	0.3357	0.3303	0.3088	0.3229	0.3355	**0.3379**
AUC	0.6493	0.6485	0.6497	0.6494	0.6486	0.6457	0.6535	**0.6633**	0.6474	0.6491
Economic evaluation	− €293,200		− €291,680		− €292,240		− €339,560		− **€284,200**	

Table 4.8
CV results in XGB with resampling after hyperparameter tuning.

Resampling technique	Accuracy	Precision	Recall	F1 Score	AUC
ADASYN	0.860 (+/− 0.001)	**0.892 (+/− 0.003)**	0.825 (+/− 0.002)	**0.861 (+/− 0.001)**	0.857 (+/− 0.001)
RUS + SMOTE	0.833 (+/− 0.004)	0.880 (+/− 0.007)	0.771 (+/− 0.010)	0.833 (+/− 0.004)	0.821 (+/− 0.005)
SMOTE	**0.861 (+/− 0.004)**	0.892 (+/− 0.004)	0.821 (+/− 0.007)	0.861 (+/− 0.004)	0.855 (+/− 0.004)
SMOTEENN	0.853 (+/− 0.004)	0.879 (+/− 0.003)	**0.872 (+/− 0.008)**	0.849 (+/− 0.003)	**0.875 (+/− 0.004)**
SMOTETomek	0.855 (+/− 0.005)	0.889 (+/− 0.005)	0.812 (+/− 0.007)	0.855 (+/− 0.005)	0.849 (+/− 0.005)

certain thresholds, is fundamental to minimizing the risk of policies lapsing. A global strategy focused on these variables can play a key role in strengthening the stability of MLI policies and building customer loyalty.

As consistent with findings in recent studies, RF and XGB performed globally better than LR and NN, while XGB proved to be more consistent after class imbalance mitigation. In their study, Loisel et al. (2021) concluded that XGB and SVMs performed better than LR and CART, while Kiermayer (2021) found that XGB was the superior model compared to RF, GLM, and NN. Azzone et al. (2022) stated that RF performed better than LR. Table 4.9 compares our XGB model performance (CV results in XGB with SMOTEENN after hyperparameter tuning) with other models reported in the literature.

Previous studies (Azzone et al., 2022; Babaoglu et al., 2017; Groll et al., 2022) concluded that the most important features for predicting life insurance lapses included mainly policy characteristics, such as variables associated with time (policy duration, start of the contract, age of the contract, remaining duration) or characteristics of the customer (gender, age). In this study, it was possible to prove the importance of other information associated with the customer's behavior and preferences by incorporating bank data. In their study, Azzone et al. (2022) also noted the important role behavioral and commercial reasons play in predicting lapses. This finding also validates the research gap identified earlier regarding the integration of external bank-related data to understand MLI lapses and the need for a more holistic approach to lapse prediction. The prominent influence of key bank-related variables such as "INSURANCE2″ and "WAGE_DOMICILIATION" underscores the importance of external data in refining predictive models and addressing the multidimensional challenges associated with lapses.

The results also confirmed the challenge of class imbalance as a persistent problem in predicting lapses. Previous research has acknowledged this (Groll et al., 2022), but often with limited success in striking a balance between Precision and Recall. The results of this study reflect these challenges, highlighting the continued need for differentiated approaches to class imbalance that avoid overfitting.

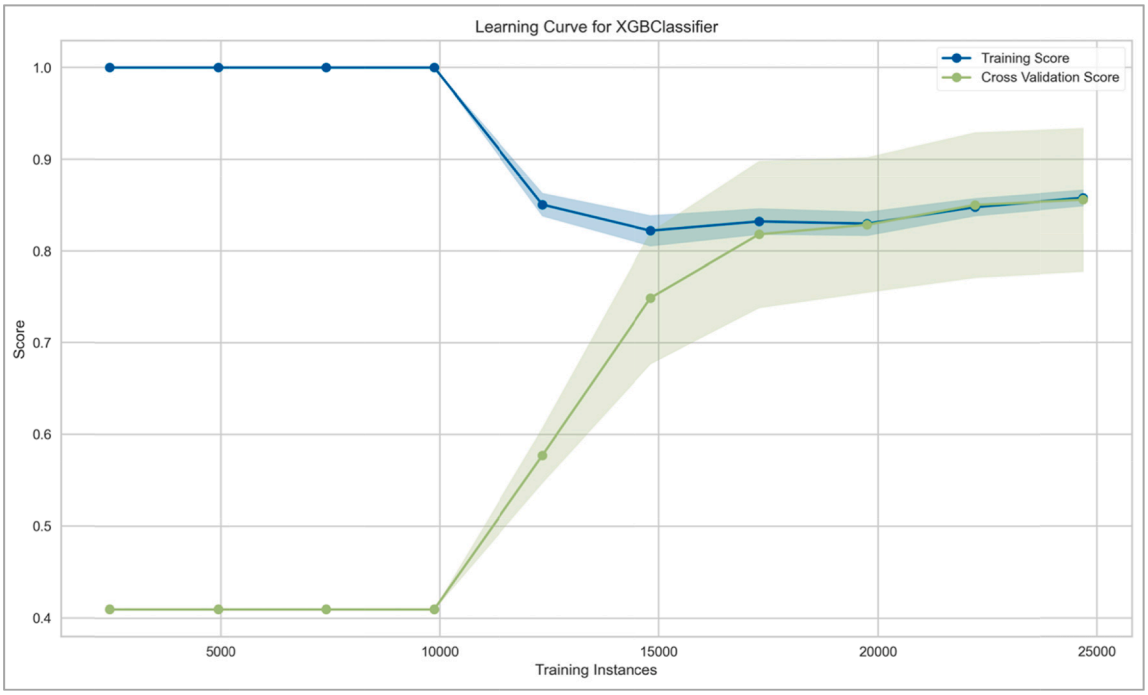


Fig. 4.1. Learning curve.

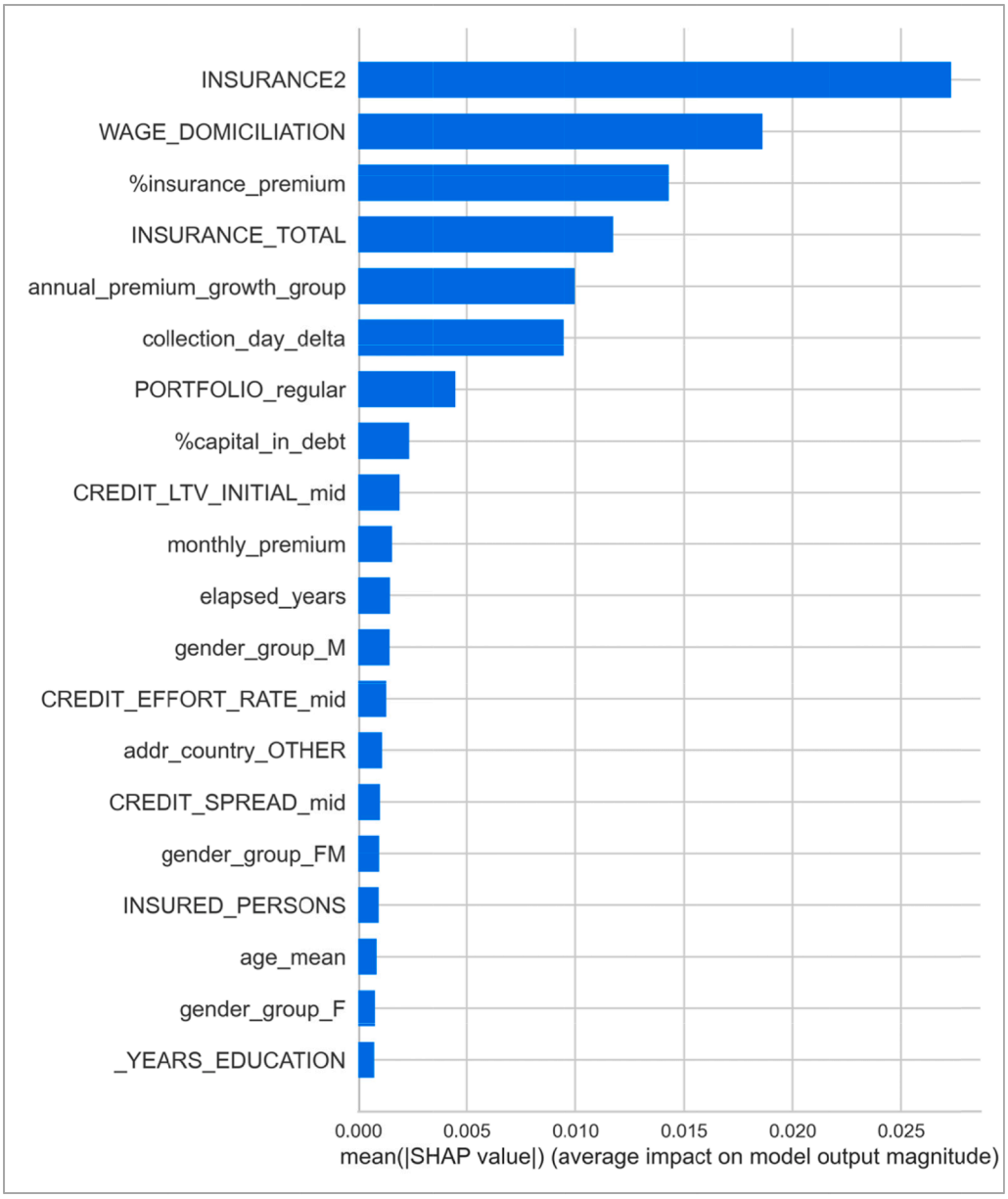


Fig. 4.2. Feature importance (SHAP summary plot).

5. Conclusions

This research focused on analyzing a specific life insurance product, MLI, which is distinguished by unique characteristics directly associated with a bank loan, making it the most appealing and profitable product for a life insurance company. Market dynamics and growing competitiveness have led to a progressive increase in cancellations of this product. The motivation behind this study came from addressing this context and developing a predictive model to identify customers prone to lapse MLI, thus developing a proactive approach to customer retention.

As anticipated from previous studies, this was a complex task with many challenges. A predictive model was built using ML and employing company and bank data, including policy data, the associated mortgage, and the customer's sociodemographic and behavioral information. Four models (LR, RF, NN, and XGB) were used to assess which would perform best. XGB proved to be more consistent.

Despite these challenges, this study has provided valuable information on the factors influencing MLI lapses. The application of SHAP analysis proved vital in identifying and interpreting the most important features of the predictive model. Although complex models such as the XGB are powerful in making accurate predictions, they often need more interpretability, a valuable feature for model evaluation and real-world application. Examining the SHAP values, we concluded that the four main features driving the performance of the XGB model originated from the banking data, with two features emerging as clearly influential in predicting lapses. This finding also proved that integrating bank data was a key aspect of this study's approach, emphasizing the value of external sources in strengthening lapse prediction models.

From the insurance company's perspective, this study introduces advanced ML techniques to enhance the accuracy of predicting policy lapses. An improvement that will allow the company to proactively identify and target customers at risk of lapsing, enabling timely intervention to retain policyholders. Identifying the key features that strongly influence lapse prediction allows the insurance company to prioritize specific customer attributes and behaviors for retention efforts, offering a more targeted approach.

Insurance companies can also use the knowledge gained from this

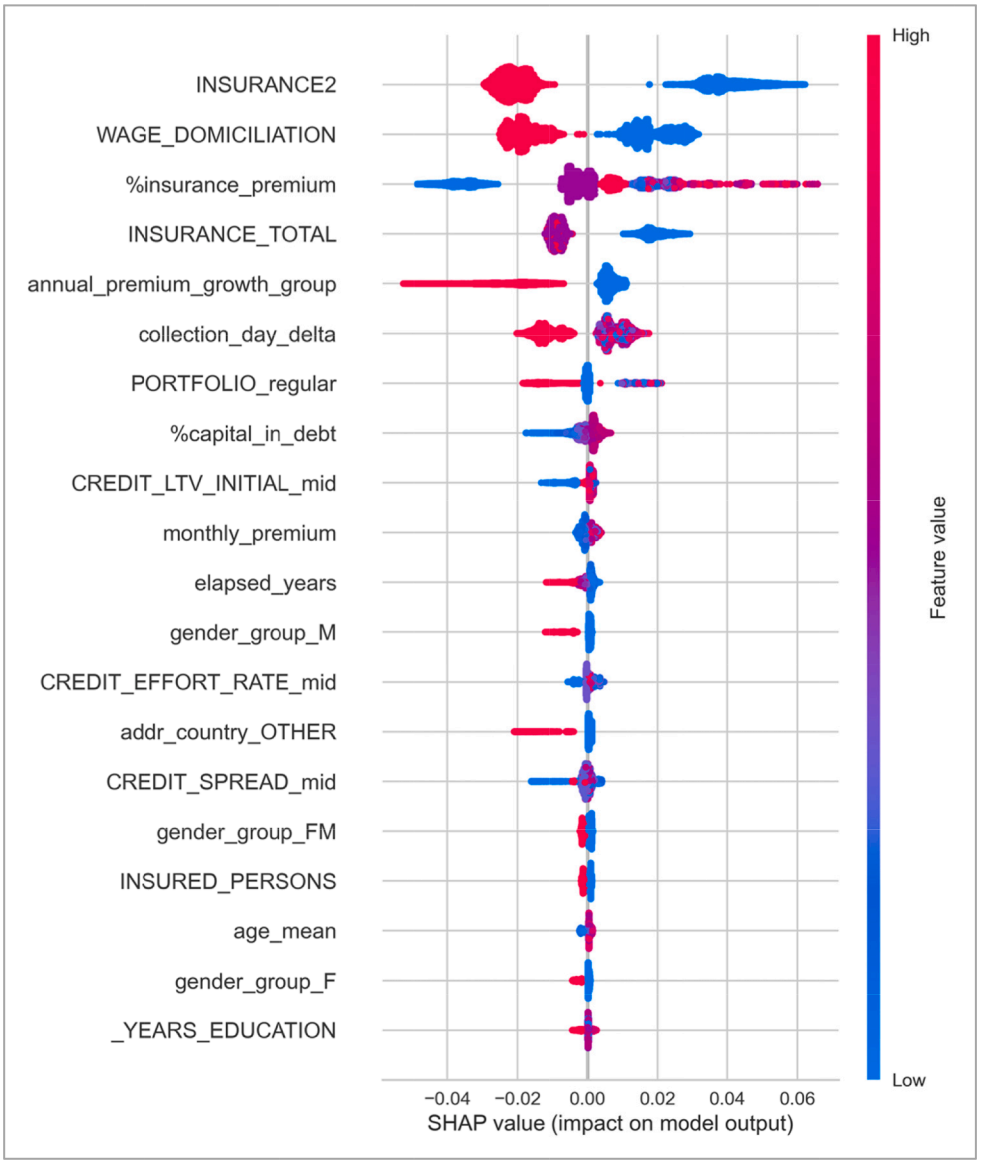


Fig. 4.3. Feature importance and its impact on model output (SHAP summary plot).

Table 4.9
Comparison of model performance with previous studies.

Study	Model	Accuracy	F1 Score	AUC
This study	XGB	0.853	0.849	0.875
Azzone et al., 2022	RF	0.880	N.A.	0.883
Groll et al., 2022	XGB	N.A.	0.079	0.734
Kiermayer, 2021	XGB	0.826	0.106	N.A.
Loisel et al., 2021	XGB	0.788	0.697	N.A.

Note: "N.A." indicates that the specific metric was unavailable in the referenced study.

study to manage better the risks associated with policy lapses. This knowledge can lead to more accurate pricing, risk assessment, and strategic decision-making.

Furthermore, this research makes a novel contribution to academia by examining lapse prediction dynamics specifically for MLI. Focusing on this unique insurance product adds needed granularity to the existing literature, filling a gap in the academic understanding of lapse behavior related to MLI. Additionally, incorporating external bank-related data into the predictive modeling framework acknowledges the multidimensionality of the MLI lapse problem. It offers a more comprehensive and nuanced understanding of policyholder behavior. It underlines the value of considering policy-related factors and external sociodemographic and behavioral data. It offers valuable guidance for future research about the merit of a more comprehensive understanding of the factors influencing policy lapse.

Moreover, this study enriches the debate on the effectiveness of ML algorithms in predicting lapses, particularly in the field of insurance. Evaluating and comparing four different algorithms reaffirms the superior performance of XGB. This result reinforces the existing literature with empirical evidence and could guide future researchers and industry professionals in carefully selecting algorithms.

5.1. *Limitations and future research*

Despite using various class imbalance techniques to mitigate class disproportion between lapsed and non-lapsed policies, striking a balance between model complexity and generalization while optimizing Precision and Recall proved challenging. To tackle this issue, future research should explore advanced strategies for class imbalance handling, extending beyond conventional methods. This could include cost-

sensitive learning approaches, which assign different misclassification costs to minority and majority classes, ensemble methods designed explicitly for imbalanced data, like BalancedRandomForest or EasyEnsemble, or other hybrid approaches that integrate multiple strategies.

Future studies should also delve into more advanced feature engineering techniques. Exploring non-linear feature interactions, innovative feature selection methods, and integrating new data sources may uncover latent patterns and relationships in the data, potentially leading to more robust predictive models.

Another limitation of this study was the dataset size, conditioned by the fact that it focused on a specific life insurance product. This comparatively small data set compared to those used in other studies and class imbalance brought additional challenges to mitigating overfitting. This limitation underscored the need for innovative solutions in future research to explore data augmentation approaches or investigate the viability of merging datasets from similar products or markets.

Considering the importance revealed by the variables provided by the bank, another significant limitation of this study was the understandably restricted access to other bank data that could be beneficial in improving the models' results. Privacy and confidentiality concerns restricted the range of variables that could be integrated into the predictive models. This limitation highlights the potential for future research to establish strong data-sharing partnerships or explore data sources with richer information, as these efforts could increase predictive accuracy.

Another inevitable constraint was the quality of external data, over which we had limited control. The data was sourced from an external provider, so ensuring its quality was challenging. Future research could investigate data quality assurance procedures, data cleansing, and validation techniques to bolster the accuracy and reliability of external data.

There is also significant potential for investigating advanced modeling approaches. Techniques such as deep learning and ensemble methods might offer enhanced predictive capabilities and alleviate the issue of overfitting.

Another promising research path involves customer risk profiling. By examining variables associated with customer assets, behaviors, and preferences, researchers can develop more granular risk assessments and predictive models.

Future research must also recognize the dynamic nature of the insurance sector. Changes in customer behavior, market competition, and the digital landscape introduce the need to focus on more dynamic forecasting models. While this study has made progress in this direction, the ever-evolving nature of the insurance sector paves the way for continued research and adaptation.

Funding

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS).

CRediT authorship contribution statement

Carlos Manteigas: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Nuno António:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Project administration.

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.

Data availability

The data that has been used is confidential.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701.

Arriola, C., Cadestin, C., Kowalski, P., José, J., Guilhoto, M., Miroudot, S., & Van Tongeren, F. (2023). *Challenges to International Trade and the Global Economy: Recovery from COVID-19 and Russia's War of Aggression Against Ukraine*.

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

Babaoglu, C., Ahmad, U., Durrani, A., & Bener, A. (2017). Predictive modeling of lapse risk: An international financial services case study. In *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017, 2017-January*, 16–21. https://doi.org/10.1109/SMC.2017.8122571.

Barucci, E., Colozza, T., Marazzina, D., & Rroji, E. (2020). The determinants of lapse rates in the Italian life insurance market. *European Actuarial Journal, 10*(1), 149–178. https://doi.org/10.1007/S13385-020-00227-0/TABLES/8

Batten, J. A., Boubaker, S., Kinateder, H., Choudhury, T., & Wagner, N. F. (2023). Volatility impacts on global banks: Insights from the GFC, COVID-19, and the Russia-Ukraine war. *Journal of Economic Behavior & Organization, 215*, 325–350. https://doi.org/10.1016/J.JEBO.2023.09.016

Biagini, F., Huber, T., Jaspersen, J. G., & Mazzon, A. (2021). Estimating extreme cancellation rates in life insurance. *Journal of Risk and Insurance, 88*(4), 971–1000. https://doi.org/10.1111/JORI.12336

Bishop, C. M. (2006). Pattern Recogniton and Machine Learning. *Information Science and Statistics, 738*. https://www.springer.com/gp/book/9780387310732.

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cerchiara, R. R., Edwards, M., Gambini, A., Wyatt, W., & Srl, I. (2008). Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II. In *18th International AFIR Colloquium*. https://www.actuaries.org/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.

Cox, S. H., & Lin, Y. (2006). Annuity Lapse Rate Modeling: Tobit or not tobit? *Society of Actuaries*. https://www.soa.org/493828/globalassets/assets/files/research/projects/cox-linn-paper-11-15-06.pdf.

Ćurak, M., Podrug, D., & Poposki, K. (2015). Policyholder and insurance policy features as determinants of life insurance lapse-evidence from Croatia 1. *Economics and Business Review, 1*(15), 58–77. https://doi.org/10.18559/ebr.2015.3.5

Dar, A., & Dodds, C. (1989). Interest Rates, the Emergency Fund Hypothesis and Saving through Endowment Policies: Some Empirical Evidence for the U.K. *The Journal of Risk and Insurance, 56*(3), 415. https://doi.org/10.2307/253166

Decreto-Lei n.º 222/2009, de 11 de setembro | DRE. (2009). https://dre.pt/dre/detalhe/decreto-lei/222-2009-489726.

Eling, M., & Kiesenbauer, D. (2014). What Policy Features Determine Life Insurance Lapse? An Analysis of the German Market. *Journal of Risk and Insurance, 81*(2), 241–269. https://doi.org/10.1111/J.1539-6975.2012.01504.X

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. https://doi.org/10.1108/JRF-12-2012-0088/FULL/XML

Fang, H., & Kung, E. (2021). Why do life insurance policyholders lapse? The roles of income, health, and bequest motive shocks. *Journal of Risk and Insurance, 88*(4), 937–970. https://doi.org/10.1111/JORI.12332

Fier, S. G., & Liebenberg, A. P. (2013). Life Insurance Lapse Behavior. *North American Actuarial Journal, 17*(2), 153–167. https://doi.org/10.1080/10920277.2013.803438

Gemmo, I., & Götz, M. (2016). Life insurance and demographic change: An empirical analysis of surrender decisions based on panel data. *SAFE Working Paper Series*. https://ideas.repec.org/p/zbw/safewp/240.html.

Groll, A., Wasserfuhr, C., & Zeldin, L. (2022). *Churn modeling of life insurance policies via statistical and machine learning methods – Analysis of important features*. http://arxiv.org/abs/2202.09182.

Hossin, M., & Sulaiman. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 5*(2). https://doi.org/10.5121/ijdkp.2015.5201

Hu, S., O'Hagan, A., Sweeney, J., & Ghahramani, M. (2021). A spatial machine learning model for analysing customers' lapse behaviour in life insurance. *Annals of Actuarial Science, 15*(2), 367–393. https://doi.org/10.1017/S1748499520000329

Hwang, Y., Chan, L. F. S., & Tsai, C. J. (2022). On Voluntary Terminations of Life Insurance: Differentiating Surrender Propensity From Lapse Propensity Across Product Types. *North American Actuarial Journal, 26*(2), 252–282. https://doi.org/10.1080/10920277.2021.1973507

Kagraoka, Y. (2005). *Modeling Insurance Surrenders by the Negative Binomial Model*. http://www.musashi.jp/.

Kiermayer, M. (2021). *Modeling surrender risk in life insurance: theoretical and experimental insight*. http://arxiv.org/abs/2101.11590.

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

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

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 111–117.

Kuo, W., Tsai, C., & Chen, W. K. (2003). An Empirical Study on the Lapse Rate: The Cointegration Approach. *Journal of Risk and Insurance, 70*(3), 489–508. https://doi.org/10.1111/1539-6975.T01-1-00061

Lemaître, G., Nogueira, F., & Aridas char, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research, 18*, 1–5. http://jmlr.org/papers/v18/16-365.html.

Loisel, S., Piette, P., & Tsai, C. H. J. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin: The Journal of the IAA, 51*(3), 839–871. https://doi.org/10.1017/ASB.2021.10

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems, 30*. https://github.com/slundberg/shap.

Milhaud, X., Loisel, S., & Maume-Deschamps, V. (2010). *Surrender triggers in life insurance: classification and risk predictions*.

Nolte, S., & Schneider, J. C. (2017). Don't lapse into temptation: A behavioral explanation for policy surrender. *Journal of Banking & Finance, 79*, 12–27. https://doi.org/10.1016/J.JBANKFIN.2017.02.011

Outreville, J. F. (1990). Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance: Mathematics and Economics, 9*(4), 249–255. https://doi.org/10.1016/0167-6687(90)90002-U

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos Pedregosa, Varoquaux, Gramfort et al. Matthieu Perrot. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Reck, L., Schupp, J., & Reuß, A. (2022). Identifying the determinants of lapse rates in life insurance: An automated Lasso approach. *European Actuarial Journal*. https://doi.org/10.1007/s13385-022-00325-1

Renshaw, A. E., & Haberman, S. (1986). Statistical analysis of life assurance lapses. *Journal of the Institute of Actuaries, 113*(3), 459–497. https://doi.org/10.1017/S0020268100042566

Shamsuddin, S. N., Ismail, N., & Roslan, N. F. (2022). What We Know about Research on Life Insurance Lapse: A Bibliometric Analysis. *Risks 2022, Vol. 10, Page 97*, *10*(5), 97. https://doi.org/10.3390/RISKS10050097.

Sirak, A. S. (2015). Income and Unemployment Effects on Life Insurance Lapse. *Retrieved September*, 18, 2020.

Szczygielski, J. J., Brzeszczyński, J., Charteris, A., & Bwanya, P. R. (2022). The COVID-19 storm and the energy sector: The impact and role of uncertainty. *Energy Economics, 109*, Article 105258. https://doi.org/10.1016/J.ENECO.2021.105258

Villeneuve, B. (2014). Mortgage life insurance: A rationale for a time limit in switching rights. *Mathematics and Financial Economics, 8*(4), 473–487. https://doi.org/10.1007/S11579-014-0124-2/FIGURES/7

Xong Lim, J., Jin Xong, L., & Ming Kang, H. (2019). A Comparison of Classification Models for Life Insurance Lapse Risk. *International Journal of Recent Technology and Engineering*, 2277–3878. https://www.researchgate.net/publication/333262288.

Yaakob, R., Abdullah, M. H. S. B., & Baharom, N. M. (2018). Analisis Polisi Luput Pelan Takaful Keluarga. *The Journal of Muamalat and Islamic Finance Research, 15*(1), 84–95. https://doi.org/10.33102/JMIFR.V15I1.103