NOVA
IMS
Information
Management
School

MEGI

Mestrado em Estatística e Gestão da Informação

Master Program in Statistics and Information Management

Predicting Lapse Rate in Life Insurance: An Exploration of Machine Learning Techniques

Diogo da Cunha Alcaide

Dissertation submitted in fulfillment
of the requirements for the degree of

Master of Science in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão da Informação

Universidade Nova de Lisboa

NOVA INFORMATION MANAGEMENT SCHOOL

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTING LAPSE RATE IN LIFE INSURANCE:
AN EXPLORATION OF MACHINE LEARNING TECHNIQUES

by

Diogo da Cunha Alcaide

Dissertation presented as a requirement for obtaining the Master's degree in Information Management, with a specialization in Risk Analysis and Management

Advisor: Rui Alexandre Henriques Gonçalves

February 2023

Abstract

The implementation of machine learning techniques for the prediction of the lapse rate in life insurance is investigated in this study. The lapse rate, which refers to the rate of policy cancellations or expirations, plays a crucial role in the viability of life insurance companies as they determine pricing strategies, manage risk, and plan for the future.

Data was collected through a risk survey administered to policyholders, covering their characteristics, policy details, and historical lapse patterns. A variety of machine learning algorithms were then applied to the collected data to evaluate their performance in predicting the lapse rate.

The results of the study demonstrate the effectiveness of machine learning methods in forecasting the lapse rate in life insurance. The $Extreme\ Gradient\ Boosting$, $C5.0$, and $random\ forest$ algorithms produced the best results when applied to the dataset. Additionally, several key policy and customer characteristics were identified as having significant predictive power in regards to the lapse rate.

However, the limitations of the study must be taken into consideration. Further research is necessary to validate the results on larger and more diverse datasets and to examine the practical applications of the models in the life insurance industry.

In conclusion, this study makes a contribution to the existing body of knowledge on the use of machine learning in the insurance industry and holds the potential to inform the development of more efficient risk management practices in the life insurance sector.

Keywords: Life Insurance, Lapse Risk, Machine Learning, Classification Problem, Risk Management, Risk Assessment

Resumo

Os seguros de ramo vida são uma importante rede de segurança financeira para muitos indivíduos e famílias. Um fator-chave na viabilidade de uma seguradora é o risco de lapso, ou seja, a taxa de cancelamento ou expiração de apólices por parte dos segurados. A previsão precisa desta taxa de lapso é essencial para as seguradoras poderem preçar corretamente as apólices, gerir os riscos e planear o futuro estrategicamente.

Neste estudo, foi explorado o uso de métodos preditivos de Data Mining para prever a taxa de lapso em seguros de vida. Teve como base a análise e tratamento de dados, tendo em conta um questionário de risco com as características dos segurados, detalhes das suas apólices e padrões históricos de lapso. Com esta informação foi aplicada uma gama de métodos preditivos e feita uma avaliação de performance relativa à previsão da taxa de lapso.

Os nossos resultados mostraram que os métodos preditivos podem ser eficazes e coerentes na previsão da taxa de lapso em seguros de vida. Em particular, foi encontrada uma boa performance de resultados nos algoritmos *Extreme Gradient Boosting*, *C*5.0 e *Random Forest*. Além disso, com este estudo foi possivel identificar várias características importantes para conseguir prever as apólices e clientes em risco de lapso.

Embora os nossos resultados apontem para uma promessa no uso de metódos preditivos na antevisão da taxa de lapso, também existiram algumas limitações. É sugerido uma maior pesquisa para validar os resultos encontrados e aplicacões de modelos com um conjunto maior de dados e mais diversificados.

De modo geral, esta pesquisa contribui para o desenvolvimento do uso de métodos preditivos na indústria de seguros e grande potencial em informar e gerir riscos antecipados no setor segurador no ramo de Vida.

Palavras-chave: Ramo Vida, Seguros, Gestão de Risco, Métodos Preditivos de Data Mining, Problema de Classificação, Risco de Lapso, Classificação de Risco.

Contents

1 | Introduction

The insurance industry plays a crucial role in managing and mitigating risks for individuals and businesses. Within the insurance industry, life insurance companies are particularly interested in understanding and predicting the risk of policy lapse, as it can have a significant impact on their financial solvency. The ability to accurately predict which policyholders are at risk of lapsing can allow insurance companies to take proactive measures to retain these customers and avoid financial losses.

Since the 1980s, European Union regulators have identified lapse risk as one of the main risks faced by the life insurance industry, along with market risk and credit risk (see Insurance and Authority (2011)). Modelling lapse risk is critical for many actuarial tasks, such as product design, pricing, hedging, and risk management. Historically, lapses have posed problems for life insurers facing solvency issues. An increase in the lapse rate can significantly impact premiums and damage a company's reputation, leading to further policyholder lapses. This can result in the liquidation and insolvency of the company (see Eling and Kochanski (2012);Barsotti et al. (2016)). Therefore, it is important to improve the assessment and modelling of lapse risk exposure and to classify customers more accurately into different lapse risk groups (see Barsotti et al. (2016); Biagini et al. (2021)). This involves solving a classification problem.

The research questions for this study are:

1. What are the key factors that influence lapse rate in life insurance and how can they be predicted?

2. Can machine learning techniques be used effectively to predict lapse rate in life insurance?

3. How do different machine learning algorithms perform in predicting lapse rate, and which one is most accurate?

4. What is the impact of policyholder characteristics (e.g., age, occupation) on lapse rate, and can these characteristics be used to predict lapse risk?

5. How can lapse rate prediction models be used by insurance companies to improve their risk management and pricing strategies?

6. What are the limitations of current approaches to predicting lapse rate in life insurance, and what opportunities are there for further research in this area?

The goal of this study is to evaluate the effectiveness of machine learning (ML) classification models in predicting policy lapse risk in life insurance. The analysis will use a dataset of policyholder information, including demographics and coverage details, to train and test various statistical models. To ensure the reliability of the results, cross-validation and bootstrapping will be applied.

The study will also examine the historical context of policy lapse risk in the industry, as well as the impact of policyholder characteristics such as age and occupation on lapse rate. It's worth noting that this study is limited to a specific dataset and does not take into account external factors that could affect policy lapse risk.

However, the results of the study can still provide valuable insights for life insurance companies. By identifying key factors that contribute to policyholder lapses, the study aims to improve lapse rate prediction accuracy and help insurance companies retain customers and avoid financial losses. The findings of this study could also assist future research in this field.

2 | Literature Review

2.1 Exploring the Foundations: A Study of Background and Definitions

The main focus of this Literature Review is lapse risk in life insurance, also known as churn, and how machine learning can be used to predict lapse rate in life insurance. Furthermore, this chapter establishes the foundation for the research presented in the rest of the thesis, including the models studied and the challenges of using machine learning in the life insurance industry.

In particular, this section delves into the essential concepts and terminology related to life insurance and risk management, as well as their impact on insurance solvency.

2.1.1 Risk

The Oxford English Dictionary defines *risk* as "a chance or possibility of danger, loss, injury or other adverse consequences", and the *at risk* as "exposed to danger". According to these definitions, is the possibility that something bad happens and is related to uncertainty of outcome (see Promislow (2014):pp. 3–6).

The definitions of *risk* can be found from many sources. For instance, the Institute of Risk Management (IRM) defines risk as the combination of the probability of an event and its consequence, which consequences can range from positive to negative. And the Institute of Internal Auditors (IIA) defines *risk* as the uncertainty of an event happening with an impact on the achievement of objectives. The IIA adds that *risk* is measured in terms of consequences and likelihood (see Promislow (2014):pp. 15–21). A more general definition can be provided in mathematical terms. Considering a *risk* as a random function , X, whose actual outcome (or realization) is unknown. However, it is necessary to specify a set of possible outcomes, and assign the probabilities over this set (see Olivieri and Pitacco (2011):pp. 1–74).

The possibility of unfortunate events that can result in financial loss, including the loss of a primary breadwinner, exorbitant medical expenses, insurance policy cancellations, and devastating home fires, can have significant financial consequences. To protect against these potential losses, it is essential to take proactive measures. One of the most effective ways to do so is by purchasing insurance coverage, which can help mitigate the financial impact of these events and provide peace of mind (see Promislow (2014):pp. 3–6).

2.1.2 Risk Management

Risk management (RM) is a crucial practice for companies and organizations to identify and control the various risks that can impact their operations (see Promislow (2014):pp. 3–6). The RM process involves identifying and analysing risks, assessing their consequences, selecting and implementing appropriate risk mitigations, and monitoring these mitigations to ensure successful risk reduction (see Doerry and Sibley (2015); Pitacco (2007)). One of the key techniques used in RM is insurance, which aims to transfer potential losses with high frequency and high severity to an insurance company. Another common technique is investing in information to improve the accuracy of estimates and forecasts and reduce variability around expected values.

The RM process is constantly developing and progressing, with different areas such as project risk management, clinical/medical risk management, financial risk management, and Information technology (IT) risk management. The use of technology and data-driven approach can help automate and streamline the RM process, making it more efficient and cost-effective (see Olivieri and Pitacco (2011):pp. 14–23).

In the insurance industry, RM is essential in designing and pricing insurance products. It may involve risk rating in some markets or market segments, applying measures to control the level of risk, or a combination of approaches. The ability to segment using risk factors is important to ensure affordable and accessible contracts (see Patterson and Executive (2015); Stoneburner et al. (2002)).

Overall, RM is an ongoing process that requires continuous monitoring, assessment, and updating to ensure that the organization's goals and objectives are met. The use of sophisticated tools and techniques allows organizations to identify, evaluate, and mitigate risks in a timely and effective manner, enabling them to make informed decisions and achieve their business objectives.

2.1.3 Risk Assessment

Risk Assessment (RA) is a component of the Risk Management (RM) process that involves recognizing and rating risks. This process of risk identification and analysis aims to determine the probability and consequences of a given occurrence faced by an organization, project, or strategy (see Hopkin (2018):pp. 291–305).

In 2009, Iso et al. (2009) was published, providing a wide range of risk assessment techniques. One popular approach was the use of checklists and questionnaires, along with inspections and audits.

Checklists and questionnaires have the advantage of being relatively simple to complete and less time-consuming than other risk assessment techniques. However, this approach also has the disadvantage of potentially not addressing risks through appropriate questions or failing to recognize significant risks (see Hopkin (2018):pp. 291–305).

2.1.4 Risk Classification

The classification of risks faced by an organization is necessary for effective management of related or similar risks. Risks can be classified based on the nature of their impact, such as people, premises, processes, or products. Many risk management principles and frameworks recommend specific risk classification systems (see Hopkin (2018):pp. 132–140).

Risk Classification (RC) plays a vital role in actuarial practices in the insurance industry. The 2008 financial crisis showed the harm of disregarding risk factors, such as not considering age while offering life insurance, causing low use of the financial or personal security system, insufficient coverage for high-risk individuals and inadequate protection for low-risk individuals, putting the whole system at risk.

Classification can also bring unforeseen changes in risk distribution, for example, if an insurance company doesn't check for a certain risk factor, it could result in more people with that characteristic seeking coverage, causing higher overall expenses. Hence, RC is used to consistently manage individuals with similar risk characteristics, promoting better economic benefits, ensuring coverage availability and maintaining system stability (see Bykerk et al. (2005)).

2.1.5 Insurance

The insurance sector plays a crucial role in the financial services industry, as it impacts economic growth, promotes efficient resource allocation, reduces transaction costs, creates liquidity, simplifies economies of scale in investment, and spreads financial losses. Insurance is a contract between a policyholder and an insurer, in which the policyholder transfers the risk of potential financial loss to the insurer, who underwrites this uncertainty (see Haiss and Sumegi (2008)).

These contracts provide individuals and organizations with financial security or compensation in the event of specific losses. In exchange for these benefits, the policyholder agrees to pay a premium, which is a predetermined amount of money. However, the policy is limited by a maximum amount that the insurer will pay for a covered loss (see Kagan (2022)). As the insurer bears the policyholder's risks and provides coverage for unforeseen future events, the coverage represents the amount of risk, liability, or potential loss that is protected by the insurer. This coverage may have limitations, warranties, and exclusions (see Hopkin (2018)). When a situation arises where the insurer does not assume a specific risk at the given amount, the premium may increase for a greater risk, known as the aggravation of risk (see Cousy (2008)). To ensure sufficient income from premiums, insurance companies aim to accurately calculate the average expected loss and charge accordingly for insurance contracts (see Abachi (2018)). One method for anticipating these events may be determining their probability of occurrence.

2.1.6 Life Insurance

Life insurance is a contract between an insurer and a policyholder that provides coverage for individuals, groups, and pension plans. A life insurance contract is a legal agreement that aims to pay benefits depending on events related to the lifetime of one or more individuals. The parties involved in the contract are the insurer, the insured (whose lifetime determines the payment of benefits), the policyholder (who initiates the contract and pays the premium), and the beneficiary (who receives the benefits), (see Olivieri and Pitacco (2011):pp. 60–72). Normally, in a life insurance policy, the policyholder pays premiums during their lifetime and in exchange, the insurer agrees to pay a sum of money to named beneficiaries upon the policyholder's death (see Fontinelle (2021)). There are various types of life insurance products available to cover different needs.

The life insurance industry is constantly evolving and facing new challenges, one of which is the aggravation of risk. In the insurance industry, the aggravation of risk refers to any factor that increases the likelihood of a policyholder experiencing a loss or making a claim on their insurance policy. These factors, known as aggravation motives, can include pre-existing medical conditions, risky behaviours, or certain occupations or hobbies that increase the likelihood of accidents or illnesses. It is also important to note that aggravation of risk can also refer to increase of risk from external factors, such as natural disasters, political turmoil, or pandemics, that can cause an increase in the likelihood of a claim or impact the value of the claims. It is the responsibility of underwriters and actuaries to assess the overall risk of a policyholder, taking into account both inherent risks and aggravating factors (see Cousy (2008)).

2.1.7 Life Insurance Products

Life insurance differs from other types of insurance in that it is characterized by its long-term products. These products may provide benefits such as survival, death, or a combination of both.

Common types of life insurance products include pure endowment, life annuities, term insurance, whole life insurance, and endowment insurance. The monetary components of a life insurance contract include premiums, benefits, and expenses (see Olivieri and Pitacco (2011):pp. 60–72). For example, term life insurance, also known as temporary life insurance, lasts for a predetermined number of years. This type of insurance guarantees a payment to the specified beneficiaries equal to the face amount of the policy if the policyholder dies during the life of the policy. If the policyholder does not die during the policy term, no payments are made. The policyholder is required to make regular premium payments to the insurance company for the life of the policy or until the policyholder's death (see Hull (2018)).

Over time, insurance companies have attempted to improve the efficiency of selling products through the use of the underwriting processes (see MISHR (2016)).

2.1.7.1 Annual Renewable Term (ART) Insurance

Annual Renewable Term (ART) insurance is a form of temporary life insurance that enables policyholders to renew coverage annually without reapplying or undergoing a medical exam. These policies are underwritten using mortality tables and typically have monthly or yearly premiums for a one-year contract period. Over the years, renewing the insurance contract may increase the premiums due to the policyholder's aging. In case of the policyholder's death, the insurer pays a benefit that remains the same throughout the duration of the contract (see Beers (2021)). This research focus on Annual Renewable Term (ART) Insurance products.

2.1.8 Life Underwriting

Underwriting is a risk assessment process and division of an insurance company that helps evaluate policy proposals and prospective customer data provided by insurance agents. Insurance companies invest human and time resources to carry out underwriting. The manual assessment process to determine the appropriate policy premium, product, and associated risk typically can take around 30 to 60 days to issue a new insurance policy (see Hutagaol and Mauritsius (2020)).

During a life insurance application, the client must provide basic details and necessary information for evaluation, which typically includes filling out a survey containing several questions about their life, such as health, financial profile, and behaviour. This survey is important to determine whether a policy should be issued, whether changes need to be made based on the person's risk profile, and whether it is profitable to offer the insurance product and coverage (see Biddle et al. (2018); Black and Skipper (2000)).

The underwriting process can be labour-intensive, expensive, and time-consuming, but it is crucial to ensure the right policy and premium are issued (see Batty et al. (2010)). Therefore, the underwriting process should be made faster, more economical, more efficient, and more consistent (see Biddle et al. (2018)). As long as it is performed manually by human judgement, it is subject to inconsistency. As a result, traditional underwriting limits the degree to which an insurer can estimate risk from data and offer efficiently priced products. To mitigate these limitations, insurers use point systems developed by doctors and underwriters to compute risk by mapping medical and behavioural attributes, such as cholesterol, driving record, and family and personal medical history, to point values that either debit or credit an overall score (see Brackenridge et al. (2016)). Additionally, identifying the complex relationships between diverse knowledge areas and how they can be used to forecast risk can be challenging. One potential solution is the use of machine learning and pattern recognition tools, which may assist underwriters in increasing their knowledge base and characterizing these complex relationships.

In conclusion, the goal of the insurer is to accurately assess regularly the individual risk, where risk in life insurance is considered as the likelihood of an injury, sickness, disease, disability, or mortality. The use of technology and data-driven approach has been proposed as a solution to speed up the process, make it more cost-effective and consistent, while maintaining the accuracy of the assessment (see Biddle et al. (2018)).

2.1.9 Insurance Risk Classes

An insurance risk class is a categorization of individuals or companies with similar characteristics that allows insurance companies to determine the level of risk associated with underwriting a new policy and the corresponding premium that should be charged for coverage. The process of defining an insurance risk class is a crucial aspect of an insurer's underwriting procedure (see Kagan (2021)).

In a risk-rated market, the premium charged to an insured client is based on their unique risk profile. Clients with similar risk profiles will be charged the same premium. To determine these risk profiles, insurers segment their portfolio into different risk groups and calculate risk premiums for each group based on factors that affect the likelihood of a claim being made and the size of the claim (see Européen (2011)). These factors can include Age; Gender; Disability; Occupation; Leisure pursuits; Amount and duration of cover; Education, income, dwelling location; or Behaviour habits (like smoking, drinking, drugs).

The objective of this process is to accurately assess risk and determine appropriate premiums to ensure the financial stability of the insurance company.

2.2 An Analysis of the Factors Influencing Life Insurance

In this section, it is examined the various factors that impact the pricing and underwriting of life insurance policies. These explanatory variables, such as age, gender, occupation, medical history, and lifestyle habits, are used by insurance companies to assess the risk of insuring an individual and to predict the likelihood of their death. For this reason, it is explored how these factors may affect the cost of life insurance policies and the methods used to determine premiums. Understanding these explanatory factors is crucial for both individuals purchasing a policy and insurance companies pricing and underwriting them.

2.2.1 Demographics

Demographic factors, such as age, gender, occupation, and health, impact the risk profile of potential policyholders for life insurance and pensions. Insurance companies use different premium tables to set risk-reflective premiums based on these characteristics. Life insurers divide policyholders into risk groups, referred to as risk pooling, using risk-rating factors. It's essential that the rating factors are objective, verifiable, and low cost to prevent high expenses (see Cox et al. (2013); Hendren (2013)).

Two principles guide private insurance provision based on demographics: risk-based pricing and risk sharing. The risk-based pricing method determines the cost for each individual policyholder through medical and specialist tests and questions, resulting in additional underwriting costs. Risk classification can mitigate this issue (see Parsons (2015)). The second principle, risk sharing, involves premiums spreading risk among individuals in risk pools, allowing insurers to diversify risk (see Abachi (2018); Frees (2013)).

2.2.1.1 Age

Age is a commonly used risk factor in life insurance and pensions pricing, as it is a quantitative variable that is easy to validate and not expensive to collect (see Fong (2015)). Life expectancy decreases with age, and as a result, younger people are considered to have less health issues that may lead to death. Therefore, insurance companies charge higher premiums for older applicants for insurance policies and may not provide insurance policies past a certain age (see Abachi (2018)).

2.2.1.2 Gender

Gender is also used as a rating factor in the pricing structure, mainly for pension products, and combined with more detailed medical underwriting (see Fong (2015)). Studies have shown that men have a higher mortality rate than women, and as a result, insurers use this information to rate insurance products that cover dangers that vary between men and women (see Abachi (2018)).

However, it is important to note that in 2004, the European Court of Justice issued an EU Gender Directive which required the equivalent treatment between men and women in the access and supply of goods and services (see Hidalgo et al. (2013)).

2.2.2 Socio-Economic Factors

Socio-economic factors such as economic status, culture, health, lifestyle, and occupational risks impact insurance prices. To address these evolving risks, insurers offer new products.

For example, accidents are a frequent hazard in many jobs. The International Labour Organization (ILO) estimates that globally, 30% of all work-related deaths are due to accidents, with the remaining caused by diseases (see Pega et al. (2021)). Thus, individuals in high-risk occupations may pay an additional premium, even if medical and other factors are favorable (see Abachi (2018)).

Overall, demographic factors play a crucial role in determining the risk profile of policyholders and setting appropriately risk-reflective premiums.

2.2.2.1 Health and Lifestyle

Health and lifestyle information is utilized as a risk factor in determining insurance premiums, for this this reason policy applicants typically complete a questionnaire to verify their health and lifestyle habits (see Ngueng Feze and Joly (2014)). Information that can affect how an insurer sets its premium should consist of data to assist the risk assessment process (see Salman et al. (2016)). This information is verified through health facilities or third parties, or alternatively, the insurer may require the applicant to undergo a set of medical tests (see Abachi (2018)).

2.2.3 Regulation

Insurance companies are subject to several regulatory restrictions, such as risk-based capital requirements, controls on pricing and product design in order to guarantee a healthy financial situation (see Schlütter (2014)).

These regulations are designed to limit an insurer's default risk to a level deemed acceptable by regulators, and to control and incentive insurers' behaviour in product creation, premium pricing, or premium investment. It aims to create a market environment that balances premiums, claims, and expenses while generating sufficient profits (see Abachi (2018); Kwon (2013)).

Regulation also protects consumers by restricting insurers from engaging in certain types of activities and offering coverage for risks where there is no need for insurance (see Kwon (2013)). Additionally, insurers are legally obligated to contribute to a fund to protect their clients in the event of a possible bankruptcy (see Schmeiser and Wagner (2013)).

Similarly, regulations are put in place to protect the interests of the insurer as well. The insurer must safeguard its asset base while still providing a beneficial product to its customers. It may also control competition in the insurance industry, which affects prices. In large, competitive insurance markets, insurance companies competitively price their policies. An insurer with past performance of the risk assesses its own ability to undertake that particular risk entirely, with or without an agreement with re-insurers who underwrite on behalf of the insurer for a premium (see Abachi (2018)).

2.3 Evolving Regulations and Industry Development: A History of Life Insurance and Solvency Standards

2.3.1 Chronicle of Change: A Historical Overview of the Life Insurance Industry

Insurance in some form has been used by society for a long time in history. The origins can be traced to members of a community helping others who suffered loss in some form or another, for instance, by the Greeks and Romans (see Salman (2015)). It is evident that the purpose involves a sharing or pooling of risks among a large group of people (see Promislow (2014):pp. 196–203).

In the early 18th century, in London, the Amicable Society for a Perpetual Assurance Office was the first company to offer life insurance (see Salman (2015)). Over the centuries, Insurance has developed into a modern business with a significant positive impact on financial systems and consequently on people from protecting them several risks. The industry has been lucrative for many years and has been an essential aspect of private and public long-term finance (see Haiss and Sümegi (2008)).

2.3.2 Revision of Insurance Regulation in the European Union

Since the 1970s, the need for revising the existing regulations of the insurance market in the European Union (EU) has been acknowledged, as three generations of European Union Directives have established the basis for freedom of establishment and provision of services within the EU. However, as risks have evolved and become more diverse, it has become clear that the current regulatory framework was not sufficient. The lack of an economic, risk-based approach and differences in implementation across the EU have led to discussions among European institutions, regulators, and supervisors about the need to improve the prudential framework for the supervision of insurance and reinsurance undertakings. These discussions gained particular importance in 2007-2008, following the subprime crisis, which highlighted the need to review the EU supervisory model for the financial sector (see Hopkin (2018):pp. 206–216; Gatzert and Wesker (2012)).

The global financial crisis of 2008 further exposed the vulnerability of both insurance companies and banks, leading to the bankruptcy of numerous financial institutions around the world. The interconnectedness of insurance companies with banks further exacerbated the impact of the crisis. Larosière et al. (2009) report, emphasized the need for a harmonized understanding of risks and strengthened regulatory oversight in the insurance sector (see Risk and Soundness (2008)).

In the early 1970s, the European Commission introduced the first legislation for (re)insurance companies in the European Union that addressed rules on solvability requirements, known as Solvency I. This regulation required (re)insurers within their jurisdiction to reserve an amount of capital in case an extreme event should occur. In light of the subprime crisis of 2007-2008, the European Insurance and Occupational Pensions Authority (EIOPA, earlier CEIOPS) issued the Solvency II directive, replacing the previous EU insurance regulations, Solvency I. The increased complexity of the insurance industry is the reason why Solvency II was regulated in 2016 (see Larosière et al. (2009)).

2.3.2.1 Defining Solvency I and Solvency II

The Solvency I framework primarily focused on insurance liabilities and the amount at risk, but Solvency II takes a more comprehensive approach by including investment, financing, and operational risks in the calculation of capital requirements. The Solvency I framework was not sensitive to several key risks, such as market, credit, and operational risks (see Marano and Siri (2017)).

Solvency II, aims to enhance the protection of policyholders and beneficiaries, promote a risk-based culture within all aspects of the insurance company, and increase the sensitivity of the capital requirements to the actual risks the company is exposed to. It also seeks to promote convergence of practices between regulators and insurers, create a level playing field, and improve transparency and market discipline (see Solvency (2009)).

Similar to the Basel framework for banks, Solvency II consists of three interrelated pillars: quantitative requirements, governance, and disclosure. (see Marano and Siri (2017)). This framework takes an integrated approach to managing risks.

Table 2.1: The 3 Pillars of Solvency **Source:** Marano and Siri (2017)

PILLAR I	**Quantitative requirements** - Technical provisions - Solvency capital requirement (SCR) - Minimum capital requirement (MCR) - Investments - Own funds
PILLAR II	**Qualitative requirements** - System of governance - Risk management and Internal control - Supervisory review process
PILLAR III	**Disclosure and market discipline** - Disclosure of information to the public - Transparency - Harmonized reporting to supervisors

The primary focus of Pillar I is the determination of capital requirements for insurance companies, based on the Solvency Capital Requirement (SCR) and the Minimum Capital Requirement (MCR). These metrics are utilized to assess the availability of capital for an insurer. If there is a discrepancy between the MCR and SCR, supervisory intervention may be implemented. The standardized formula designed to calculate SCR follows at least risk modules, such as: non-life underwriting risk; life underwriting risk; health underwriting risk; market risk; counterparty default risk, Operational risk, and Intangible assets' risk (see Eikenhout (2015)).

Particularly, Life underwriting risk can be sub-divided into seven risks, such as Mortality Risk, Longevity Risk, Disability and Morbidity Risk, Life Expense risk, Revision risk, Life catastrophe risk, and Lapse risk (see Manolache (2019)). According to (see CEIOPS (2009)) definition, Lapse risk is the risk of loss, or of adverse change in the value of insurance liabilities, associated with the rates of policy lapses, surrenders, terminations, and renewals.

The majority of insurance companies that become insolvent is mainly because of poor risk and decision management. For this reason, regulators and risk managers must understand and identify lapse dynamics so that they can identify the real risks embedded in the life insurance contracts and exposure to massive lapses, surrenders, and cancellations (see Barsotti et al. (2016)). This research is focused on Lapse Risk, with the purpose of underlying lapses, surrenders, and cancellations.

2.3.2.2 Lapse Risk: Implications for Solvency and Reporting

According to European insurance and [CEIOPS] (2008), life underwriting risk is one of the most important components in the calculation of the Solvency Capital Requirement (SCR) and Minimum Capital Requirement (MCR). Lapse risk, in particular, can be a significant component of life underwriting risk before diversification effects. As a result, lapse rates have a significant impact on the capital requirement as part of SCR and MCR calculations. As part of these calculations, insurers must choose between using an external or internal model, which can lead to uncertainty and affect other calculations such as pricing, liquidity, and profitability (see Kuo (2003); Lim Jin Xong (2019)).

Therefore, it is crucial for insurers to accurately forecast and control lapse rates (see Lim Jin Xong (2019); Laurent (2016):pp. 5–6). This is because every insurer is required to submit quarterly reports to the national supervisory authority on lapses in the form of a file following the quantitative report template $S.41.01.11$. One of the log file requirements is the number of life contracts (policies) fully or partially lapsed or surrendered during the reporting period divided by the number of life contracts at the beginning of the period (see EIOPA (2020)). Hence, this research aims to improve the accuracy of lapse rate forecasting and help insurers meet regulatory requirements and make better decisions.

2.3.2.3 International Financial Reporting Standards (IFRS)

The International Financial Reporting Standards (IFRS) 17 is a new accounting standard for insurance contracts set to take effect in January 2023. It is intended to improve the efficiency and consistency of financial reporting for insurance companies, and facilitate comparisons between companies and industries. The standard replaces IFRS 4, which allowed for more flexibility in accounting practices. Compliance with IFRS 17 will require significant changes for insurance companies in terms of data management, financial presentation, actuarial calculations, product design, budgeting and forecasting (see Yeoh (2018); Laurent (2016):pp. 53–60). As a result, the standard also requires companies to recognize the impact of lapses on profit or loss, increasing the visibility of lapse risk in financial reporting. Insurance companies may need to implement new policies or strengthen existing ones to mitigate the impact of lapse risk on their financials.

2.4 Machine Learning (ML)

Machine learning (ML) is a subset of artificial intelligence (AI) that enables machines to learn and adapt through experience. When implemented effectively, ML can allow organizations to utilize data collection for business benefits (see Alzubi et al. (2018)).

ML techniques analyse potential relationships between independent and one or multiple dependent variables, and ultimately can identify a function that accurately predict a target attribute based on available input attributes (see Varian (2014)). There are a wide range of ML applications, including Naïve Bayesian Classifiers, Logistic Regression, Decision Trees (DT), k-Nearest-Neighbour (knn), and Neuronal Networks (NN).

Recent ML models have the advantage of being able to learn nonlinear transformations and interactions between variables from data without manual specification, and also offer various models for different types of feature formats (see Blier-Wong and Marceau (2021)). ML models have become increasingly important in the context of modelling insurance data, as they simplify various types of data sets, such as at: (see Burri and Buruga (April 2019); Denuit and Trufin (2021); Makariou and Chen (2021). These models can enhance actuaries' understanding of problems and data, by utilizing unstructured data directly. The field of ML is expanding and has great potential for use in actuarial science, but it is still recent and not neatly organized (see Blier-Wong and Marceau (2021)).

The determination whether one algorithm performs better than others depend on several factors, it is the field of application (see Singh et al. (2016)), including the dependencies of its inherent variables, data structure, data quality, parameter tuning or the performance measure. Until this date, there is not the best commonly accepted approach to solve a particular problem with the most suitable ML technique (see Kuhn and Johnson (2013a)). Therefore, become popular to apply several techniques to the same task and compare their performances (see Bärtl and Krummaker (2020)). ML algorithms are now easier to use with convenient packages in R that can fit several models (see Mullainathan and Spiess (2017)).

2.4.1 Machine Learning Approaches

The two main approaches to machine learning (ML) are supervised learning, and unsupervised learning (see Burri and Buruga (April 2019)). Supervised learning algorithms use labelled data to make predictions on unlabelled data. The labelled dataset, also known as the training set, consists of input variables (features) and an output variable. The choice of features will impact the accuracy of the prediction of the output variable (see MRM et al. (2015)).

The two main standard formulations of supervised learning are classification and regression. The distinction between these two is determined by the response variable, which can be either numeric or categorical. Classification aims to predict a categorical response by learning the behaviour of input-output examples, while regression predicts a continuous numeric response (see Ayodele (2021):pp. 19–20). One common issue in supervised learning is imbalanced data, where the class of interest has significantly fewer instances compared to other classes. This scenario is prevalent in many real-world applications and has therefore gained attention from researchers (see García et al. (2015):p. 8). This study focuses on optimizing binary classification problems.

In contrast to supervised learning, unsupervised algorithms analyse input data without predefined labels. Without guidance, it seeks to uncover patterns and structure within the data on its own. The goal of unsupervised learning is to detect regularities, irregularities, relationships, similarities, and associations in the input data, through the examination of unlabeled data. The algorithm examines the general correlation between data clusters (see Nasteski (2017); Ayodele (2021):pp. 19–20). Clustering and association rules are the two most well-known problems in unsupervised learning.

Additionally, related challenges such as Outlier and Anomaly Detection also exist. This process identifies data examples that deviate significantly from the expected behaviour and pattern, aiming to uncover exceptional cases that stand out from the norm (see García et al. (2015):p. 8).

Machine learning (ML) has the potential to improve accuracy in identifying risks, claims, and customer actions in the insurance industry. It's widely used for various purposes, including risk assessment, customer retention, fraud prevention, claim analysis, marketing analytics, risk analysis, sales forecasting, product development, and underwriting processing. In underwriting processing, it can act as a triage method by assessing cases that require further underwriting tests and analysis, reducing time and cost for underwriters (see Burri and Buruga (April 2019); Maier et al. (2019)).

Big data technologies have significantly impacted the insurance industry by allowing for more efficient collection, processing, analysis, and management of data. ML is a rapidly growing field with great potential for further use in actuarial science. The studies by Blier-Wong and Marceau (2021); Burri and Buruga (April 2019), and Maier et al. (2019), are examples of these transformations.

2.4.2 Machine Learning Models

There has been a significant amount of research conducted on the topic of predicting lapse rate in life insurance (see Shamsuddin et al. (2022)). This research has primarily focused on the use of machine learning techniques to analyse policyholder data and identify patterns and trends that may be indicative of a higher likelihood of policy lapse.

One common approach has been the use of regression models to predict lapse rate based on policy and customer characteristics, such as age, gender, policy type, and premium payment history (see Ferrario et al. (2018); Lee and Antonio (2015)). Other studies have employed classification models, such as decision trees, to classify policyholders as either “likely to lapse“ or “not likely to lapse“ (see Bolancé et al. (2016); Wang (2021)). Traditionally solved in the insurance practice with Generalised Linear Models (GLMs), mostly Logistic Regression (see Lim Jin Xong (2019); Loisel et al. (2021); Maynard et al. (2019); Wang (2021)).

Other research has focused on the use of ensemble models, which combine the predictions of multiple individual models in order to improve the overall accuracy of the prediction. This may be done through techniques such as boosting or bagging, which involve training multiple models on different subsets of the data and aggregating their predictions (see Diana et al. (2019); Fauzan and Murfi (2018); Grize et al. (2020); Lee and Antonio (2015); Loisel et al. (2021)).

There have also been studies that have explored the use of more advanced machine learning techniques, such as neural networks, for predicting lapse rate in life insurance. These methods have shown promising results, but they may be more complex and require more data and computational resources to implement (see Bolancé et al. (2016); Diana et al. (2019); Fauzan and Murfi (2018); Grize et al. (2020); Maynard et al. (2019)).

Also, many researches have been approaching several algorithmic techniques such as Regression and Classification Trees (like Grize et al. (2020); Groll et al. (2022); Loisel et al. (2021)); Elastic Net regularization method (like Groll et al. (2022); Loisel et al. (2021)), Naive Bayes (like Scriney et al. (2020)); Random Forests (like Bärtl and Krummaker (2020); Diana et al. (2019); Fauzan and Murfi (2018); Groll et al. (2022); Maynard et al. (2019); Wang (2021); Wuthrich and Buser (2019)), and K-Nearest-Neighbour (like Wang (2021)).

Several academic studies have focused on comparing the performance of various models, such as: Bärtl and Krummaker (2020); Diana et al. (2019); Fauzan and

Murfi (2018); Grize et al. (2020); Groll et al. (2022); Loisel et al. (2021); Maynard et al. (2019); Scriney et al. (2020); Wang (2021); Wuthrich and Buser (2019) and Lim Jin Xong (2019). For instance, Groll et al. (2022) compare logistic regression, with Classification and Regression Tree (CART), neural networks, Elastic-net, extreme gradient boosting, Support Vector Machine and random forest models on the lapse prediction problem that comes from churn management.

This research explores two main categories of algorithms: Supervised Models and Unsupervised Models (see 2.4.1). Based in this literature review, the emphasis will be on Supervised Models for classification, including Logistic Regression, Penalized Logistic Regression (Elastic Net), Non-Linear Classification Models like Naive Bayes Classifier, K-Nearest Neighbours, Neural networks, Tree-Based Approaches such as Regression and Classification Trees, and ensemble models like Bagged Classification Tree, C5.0, Adaptive Boosting, Extreme Gradient Boosting, and Random Forest. Additionally, the unsupervised algorithm K-means is employed for feature engineering purposes.

2.4.3 Generalized Linear Model: Logistic Regression (LR)

The logistic regression, also called *logit* is a special case of the generalized linear models (see Nelder and Wedderburn (1972)) obtained with the Bernoulli distribution. It is one of the most popular machine learning algorithms for binary classification. This modelling technique aims to predict the probability of a binary response based on one or more independent variables (see Kuhn and Johnson (2013b)).

This study will focus on the binary response of whether an insurance policy status becomes “Lapsed“ or “Not Lapsed“. The objective is to model the probability of a binary event, such as the lapse probability p_i of policyholder i. Given a training sample $(y_i, x_i)_{i=1}^{N}$ in which $x \in R^n$ and $y_i \in 0,1$ (see Loisel et al. (2021)).

The logistic regression is fitted by estimating the parameters using maximum likelihood, and the log-odds (see Kuhn and Johnson (2013b)). The model is written as:

$$\text{logit}(y_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i \tag{2.1}$$

where,

- y_i is the dichotomous target, defined as: $y_i = \begin{cases} 1, & \text{with probability } p_i \\ 0, & \text{with probability } 1-p_i \end{cases}$

- x_i are the predictor variables;

- β_i are the model coefficients, β_0 is the intercept. These parameters reflect the association between independent and dependent variables;

- p_i is the proportion of data with the Lapse target of policyholder i, and $1-p_i$ is the proportion of data with the Non-Lapse target of policyholder i. The entity

$p/(1-p)$ is called odds, it measures the level of the relationship between the predictor and response variables.

LR estimates probabilities, so the application of LR in a classification problem does not directly lead to labelled responses. In order to predict these responses, a threshold is set. Based upon this threshold, the obtained estimated probability is classified into classes. A common threshold used is 0.5 (see Handoyo et al. (2021); Kuhn and Johnson (2013b)).

The main challenge of modelling the LR can be the feature selection. A model with several features may result in increased multicollinearity, variables redundancy and overfitting (see Chowdhury and Turin (2020)). In the context of multiple regression, the term multicollinearity refers to the phenomenon of strong correlation among predictor variables. This means that one or more variables do not add any predictive value. According to (see Lieberman and Morris (2014)), multicollinearity do not affect the accuracy of LR predictions. This issue only would provide difficulties in assessing the best regression coefficients.

2.4.4 Regularized Generalized Linear Model: Elastic-Net (glmnet)

Regularizations are techniques often applied to increase performance of the *logit* approach. It is implemented to avoid overfitting of the data, especially when there is a large variance between train and test set performances. This method aims to penalize models with large weight vectors, which makes the resulting model less complex (see Araveeporn (2021)).

The estimates of the coefficient vector β in the *logit* model are often unstable and can lead to high variances. This can occur when there is a substantial correlation between covariates (see Fahrmeir et al. (2013)). In these situations, regularization techniques can be applied to obtain more appropriate estimates for the parameter β. Toward this end, Zou and Hastie (2005) proposed the elastic-net approach. This approach is a combination of the LASSO and ridge regression (see Hoerl and Kennard (1970)). Let

$$pen_{\alpha}(\beta) = \alpha||\beta||_1 + (1-\alpha)||\beta||_2/2 \tag{2.2}$$

be a penalty term with $\alpha \in [0,1]$, which measures the complexity of the parameter β, and $\lambda > 0$ a regularization parameter, which controls the trade-off between the reliability of the estimation of β and the influence of the penalty. An estimate of β is obtained by solving the optimization problem

$$\hat{\beta} = \operatorname{argmin}_{\beta \in R^{(p+1)}} \{l(\beta) + \lambda \cdot pen_{\alpha}(\beta)\} \tag{2.3}$$

where $l(\beta)$ is the log-likelihood of the *logit* model. For $\alpha = 0$, this is equal to ridge regularization, while for $\alpha = 1$ the LASSO is obtained (see Simon et al. (2011)).

2.4.5 Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier that have been successfully applied to a large number of real-world applications. This method is based on the Bayes theorem that assumes conditional independence between predictors (see Torgo (2011):p. 217).

The Bayes theorem is defined as $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. In addition, the Naive Bayes classifier theorem calculates the probability of each class c for a given test case as

$$P(c|X_1,\ldots,X_P) = \frac{P(c) \times P(X_1,\ldots,X_P|c)}{P(X_1,\ldots,X_P)} \tag{2.4}$$

where $X_1,\ldots,X_p$ represent the collection of predictor variables. The probability $P(c)$ is the prior probability of the class c and $P(X_1,\ldots,X_p)$ is the probability of the predictor values. $P(X_1,\ldots,X_p|c)$ is the conditional probability that represents the likelihood of the test case given the class c. The denominator is the probability of the observed evidence. This equation is calculated for all possible class values to determine the most probable class of the test case (see Torgo (2011):p. 217; Kuhn and Johnson (2013b)).

2.4.6 K-nearest neighbour (KNN)

The K-nearest neighbours (KNN) is one of the oldest and accurate algorithms for patterns classification and regression models. It is known to belong to the class of so-called lazy learner (see Cunningham and Delany (2021)). This model predicts each observation based on how similar it is to other observations. The k most similar training cases are used to obtain the prediction for a given test case (see Abu Alfeilat et al. (2019)).

In classification problems, KNN is an algorithm that uses distance metrics to classify different samples. These distances measures play an important role in determining the final classification target and can calculate a number representing the difference between samples (see Kuhn and Johnson (2013b)). Euclidean distance is the most commonly used metric that is defined as

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2.5}$$

where p is the number of predictors, and x_i and x_j are two observations. The Minkowski distance is a generalization of the Euclidean distance and is defined as $\left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q\right)^{\frac{1}{q}}$, where $q > 0$ (see Kuhn and Johnson (2013b)). When $q = 2$, the Minkowski distance represents the Euclidean distance. When $q = 1$, the Minkowski distance is equivalent to the Manhattan distance, which is a common metric used for samples with binary predictors.

According to (see Abu Alfeilat et al. (2019)), the basic KNN classifier steps can be described as

Table 2.2: Basic KNN algorithm **Source:** Abu Alfeilat et al. (2019).

```
Algorithm 1: Basic KNN algorithm
Input: Training samples D, Test sample d, K
Output: Class label of test sample
1.  Compute the distance between d and every sample in D
2.  Choose the K samples in D that are nearest to d; denote the set by P ∈ D
3.  Assign d the class it that is the most frequent class
```

2.4.7 Neural Networks (NN)

Neural Networks (NN) are a family of machine learning (ML) models that are inspired by the neural network system in the human brain. These models have been widely used in financial applications for tasks such as classification, particularly due to their ability to handle non-linear problems (see Agarwal et al. (2016); Boodhun and Jayabalan (2018)).

NN models consist of interconnected processing elements called neurons, each with a weight and a stimulation function that determine the output. The model learns from unseen variables through statistical and signal processing techniques (see Kuhn and Johnson (2013b); Torgo (2011):p. 123).

One of the NN models is the feed-forward network, which is considered the most basic (see Kuhn and Johnson (2013b)). In this model, neurons are organized in layers with an input layer that receives input, an output layer that gives predictions, and one or more hidden layers in between. The network's weights are optimized through backpropagation by iteratively updating to minimize prediction error (see Boodhun and Jayabalan (2018); Torgo (2011):p. 123).

In R, Feed-forward NNs with one hidden layer can be easily obtained using a function of the package *nnet* (see Venables and Ripley (2002)). However, NNs can be computationally intensive to train and may require additional pre-processing, such as centering and scaling the numerical data.

2.4.8 Classification Trees and Rule-Based Methods

Tree-based models and Rule-Based Methods are two common ML techniques for classification problems, each with its own strengths. Tree-based models use decision trees to split data into groups based on questions, while Rule-Based Methods classify data using predefined rules. Rule-Based Methods are efficient and quick, but may struggle in complex situations, while tree-based models are interpretable and flexible. The choice between tree-based models and Rule-Based Methods is dependent on the data and situation. Further information on both methods is available in the cited source Kuhn and Johnson (2013c).

2.4.8.1 Recursive Partitioning and Regression Trees (CART)

The classification and regression trees (CART) were developed by Breiman (1996) to segment a population by splitting up the input data into subsets according to binary rules (see Milhaud and Maume-Deschamps (2011)). Recursive partition is a type of decision tree classification technique which select the best or significant variables that are chosen for splits based on the target variable. This split can be determined based on the Gini index or the cross entropy, used as a measure of misclassification of the subsets. In a two-class classification problem of "Lapse" or "Non-Lapse", the Gini index can be defined as:

$$p_1(1-p_1)+p_2(1-p_2) \text{ or } 2p_1p_2 \tag{2.6}$$

where p_1 and p_2 are the two class probabilities (see Kuhn and Johnson (2013c)). The Partitioning algorithm then evaluate various splitting options and partitions of the data where minimize the probability of misclassification. Mostly, this algorithm is used as a fast learner, which creates decision trees based on the information gain and variance reduction (see Milhaud and Maume-Deschamps (2011)).

The main goals are to find the best possible segmentations, uncover the predictive structure of the problem, and produce an accurate classifier. The advantages of using decision trees are that have a simple structure, an easy visualization, and an easy interpretation. Otherwise, one of the weaknesses is that there is a tendency of overfitting the data (see Kuhn and Johnson (2013c)).

2.4.9 Ensemble Models

Ensemble learning is a technique in machine learning that combines multiple algorithms to improve the performance and reduce bias in predictions. Ensemble methods are commonly used because they are able to make better use of limited data. There are three main types of ensemble methods: Stacking, Bagging, and Boosting (see Zhang et al. (2022)).

Bagging, or bootstrap aggregation, is an ensemble technique that aims to increase the stability of classifiers and reduce variance. It involves generating multiple subsets of the original data, training a model on each subset, and taking the average of the predictions. Bagging works well with unstable classifiers such as decision trees (see Kuhn and Johnson (2013d)).

Boosting is another ensemble method that combines weak classifiers to form an ensemble classifier with higher performance. It is used to reduce variance and bias (see Kuhn and Johnson (2013d)). The Stacking method is a technique that combines multiple classifiers by training a model with different learning algorithms. It involves building multiple base-level classifiers in the first layer, using the results as new features to train a new classifier (meta-learner) in the second layer. In many cases, the second layer is a simpler model such as linear regression (see Li and Chen (2020);

Zhang et al. (2022)). One way to implement stacking in R is to use the “caretEnsemble“ package.

2.4.9.1 Bagged Classification Tree

Bagged Classification Tree is a type of ensemble model that combines the predictions of multiple decision trees to make a more accurate prediction. This is achieved by training several decision trees on different subsets of the training data, and then aggregating the predictions of all the trees to make a final prediction. The Bagged Classification Tree is commonly used when decision trees alone are not enough to achieve accurate predictions, as it helps to reduce overfitting and improves the generalization of the model. This technique is explained in more detail in the cited source Lemmens and Croux (2006).

2.4.9.2 Random Forest (RF)

Random forests (RF) are a type of ensemble learning part of the family known as decision trees. The basic idea of the random forest is to combine a set of simpler classification or regression trees. The main principle of this algorithm is the bagging method (see Kuhn and Johnson (2013c)).

The model divides the input data randomly into subset trees, searches for the best features amongst the subset of features and computes a prediction for each tree. Then the predictions of all computed trees are combined into a single prediction by a majority vote. The accuracy of RF is measured by the strength of each tree. One challenge of this method is that tend to overfit with too large trees. On other hand, it is not sensitive to outliers and missing data (see Breiman (2001)).

Random forests can be implemented with the package “randomForest“ in R.

2.4.9.3 C5.0

The C5.0 algorithm is a modified and updated version of Quinlan’s C4.5 classification tree model (see Quinlan (1996)). One of the key improvements of the C5.0 algorithm is the inclusion of a boosting method, which aims to increase the accuracy of the model (see Lemmens and Croux (2006)). Additionally, the C5.0 algorithm boasts improvements in terms of predictive performance, memory efficiency, and computation time (see Kuhn and Johnson (2013c)).

The algorithm builds decision trees by repeatedly splitting the samples based on the feature with the highest Information Gain (IG). IG is a measure of uncertainty from information theory that is based on the concept of entropy. The information entropy is defined as:

$$I(p) = \sum_{i=1}^{m} -p_i \log p_i \tag{2.7}$$

Where p_i refers to the probability of the outcome for each of the m possible classes for the target variable. The higher the information entropy, the more balanced the probabilities of the classes are. After the class with the highest information gain is selected, the algorithm continues the process of splitting the samples into smaller subsets. The algorithm also prunes the tree for branches that do not have a significant impact on the classification classes and replaces them with leaf nodes. Additionally, the C5.0 algorithm has an option to winnow or remove predictors that are not important to the outcome (see Kuhn and Johnson (2013c); Salman Saeed et al. (2020)).

2.4.9.4 Adaptive Boosting (AdaBoost)

The AdaBoost algorithm is a type of ensemble learning method that combines multiple weak classifiers to create a stronger, more accurate model. It uses an adaptive boosting technique, where each new classifier added to the ensemble focuses on the observations that were misclassified by the previous classifiers (see Kuhn and Johnson (2013c)). The weight of each observation is adjusted in each iteration, with incorrectly classified observations receiving a higher weight and correctly classified observations receiving a lower weight. This process is repeated until all observations are classified correctly. The final predictions are obtained by taking a weighted average of the predictions of the individual base models (see Torgo (2011):p. 217). AdaBoost is commonly used in conjunction with decision trees and is effective in reducing bias and variance in the model (see Quinlan (1996)).

The package "RWeka" in R can easily provide classification trees with a small number of nodes using the AdaBoost method (see Torgo (2011):p. 217).

2.4.9.5 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble learning method proposed by Chen and Guestrin (2016) that combines the power of Classification Tree (CART) and a specific implementation of Gradient Boosting (GB). XGBoost is considered to be faster and more accurate than traditional Gradient Boosting due to its sophisticated implementation and regularization control (see Bentéjac et al. (2020)).

The main principle of XGBoost is the boosting method, which aims to minimize the loss function by measuring the difference between the predicted value and the actual value. The algorithm creates a meta-model composed of many individual models (base learners) that combine to give a final prediction. The final model is non-linear and combines the predictions of the individual models, making it more powerful than a single model (see Fogelson (2022)).

There are two types of base learners used in XGBoost: Tree Base Learner and Linear Base Learner. The Tree Base Learner is a weighted sum of decision trees, which is the most commonly used base learner. The Linear Base Learner is a weighted sum of linear models (see Bentéjac et al. (2020); Fogelson (2022)). Both base learners are used to predict different parts of the dataset and their predictions are combined to obtain the final prediction.

2.4.10 Unsupervised learning method: K-means

K-means is a popular unsupervised learning algorithm that can be used for feature engineering. The algorithm groups similar data points together, known as clustering, and can be used to identify patterns or grouping within the data. By doing so, it gives an approximate separation of the data as a starting point and reduces the noise present in the dataset. In feature engineering, K-means can be used to identify and extract relevant features from the data that can be used to improve the performance of a supervised learning model. This can be done by using the cluster assignments as features or by using the cluster centers as a representation of the feature space. Additionally, K-means clustering can also be used as an initialization step for more computationally expensive algorithms (see Morissette and Chartier (2013)).

2.4.11 Machine Learning Challenges

In this subsection, we will explore some of the common challenges encountered in machine learning (see 2.4.2), and how to address them. We will discuss overfitting, class imbalance. Additionally, we will address the challenge of missing data and how to handle it in a machine learning model. Lastly, we will cover the specific challenges of text analysis and how to effectively use and analyse text data in machine learning models.

2.4.11.1 Overfitting

The term "overfitting" is used to determine whether a model is able to perform a particular task and make appropriate predictions on unseen data. It occurs when a model learns and performs the training data so well that affects the predictive capabilities on new unseen data. In other words, it happens when a model performs better on training data than on testing data. Also, it can happen because of inconsistencies in the dataset, limited training sets, and complexity of classifiers (see Ying (2019)), whereas underfitting occurs when the model is too simple and the accuracy is too low (see Guido et al. (2016)). A solution for overfitting can be by performing cross-validation on the dataset (see Santos et al. (2018)).

2.4.11.2 Class Imbalance

Dealing with imbalanced data is a common challenge in machine learning classification problems, where the distribution of observations across classes is uneven. Imbalance can negatively impact the performance of most ML algorithms, as they typically require a balanced representation of classes in order to make accurate predictions.

Real-world datasets often contain imbalanced class distributions, with one or more classes being under- or over-represented (see He and Garcia (2009)).

Imbalanced data is characterized by a majority class with more samples than the minority class. This can lead to a bias in the prediction model towards the majority class, at the expense of the minority class. To address this issue, several solutions have been proposed, such as adjusting the training data through data-sampling methods. These methods can restore balance by randomly under-sampling the majority class or by randomly over-sampling the minority class (see Madasamy and Ramaswami (2017); Tsai et al. (2022)).

One popular method for over-sampling is SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic samples for minority classes by interpolating between existing minority samples (see Xie et al. (2019)). This approach has been shown to be effective in a number of studies, such as (see Burez and Van den Poel (2009)) which compared the performance of random sampling on unbalanced datasets for predicting churn and found that under-sampling technique performed better than other sampling methods

2.4.11.3 Missing Data

Missing data can be a major issue in datasets, as some machine learning methods cannot handle missing values well. This can negatively impact the accuracy of the model. To address this problem, it is important to properly treat missing values as a step in preparing data for analysis. There are two common methods for dealing with missing data: eliminating features with missing data, or imputing the missing values. The method used should be selected with care, as it can have a significant impact on the model's conclusions (see Vieira et al. (2016)).

Large-scale surveys often have a high proportion of non-response samples, which can lead to missing data. A popular technique for handling non-response samples is imputation, where missing values are replaced with plausible values in order to create a complete dataset for analysis. However, before applying any imputation method, it is important to study the missing data structure and mechanism (see Andridge and Little (2010)).

There are three main mechanisms of missing data: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR indicates that the distribution of missing values does not show any relationship between the observed data and the missing data (see J et al. (2015)). In this situation, the Little's test can be used to determine if the data is MCAR, implementing the χ^2 test (see Li (2013); Little and Rubin. (1988)). MAR occurs when the missingness is dependent on other observed variables, but independent of any unobserved features. MNAR implies that the missing pattern relies on unobserved variables, and the observed part of the data cannot be explained by the missing values. This missing data mechanism is the most difficult to treat (see J et al. (2015)).

The deletion methodology involves removing all entries or variables with missing data. This method is suitable for cases where the data is large enough and missing completely at random (MCAR). However, in cases of small data, there is a risk of losing valuable data and introducing bias. Mean and mode imputation methods can also be used when the data is MCAR. These methods calculate the mean or mode of all non-missing values in a variable, and assign that value to the missing values (see Gelman (2010); Norazian (2013)).

There are also several advanced methods for dealing with missing values, such as Hot Deck imputation, Multivariate imputation by chained equations (MICE) and Random Forest imputation. Hot Deck imputation is a method where missing data is replaced with an observed response from a similar unit. The advantages of this method are that it imputes real values, avoids strong parametric assumptions, includes covariate information, and it can provide good inferences for linear and non-linear statistics. However, a disadvantage is that it requires covariate information (see Andridge and Little (2010)).

Multivariate imputation by chained equations (MICE) is a method of choice for complex incomplete data problems. It uses a combination of imputation techniques to estimate missing values in multiple variables. The method involves the imputation of each variable with missing data separately through predictive models, taking into account the dependence of the variables. This algorithm operates under the assumption that missing data are Missing At Random (MAR), (see Buuren and Groothuis-Oudshoorn (2011)). The advantages of this method are that it allows for the inclusion of complex relationships among continuous and categorical variables, and it provides realistic imputations for each variable. A disadvantage is that it can be computationally intensive, especially for large datasets. Additionally, highly correlated variables may cause problems due to collinearity, and it may not always produce the most accurate imputed values (see Azur et al. (2011); Shah et al. (2014)).

Random Forest imputation is a method that uses a Random Forest model to estimate missing values. The method trains a Random Forest model with the observed data and then uses it to predict the missing values. The advantages of this method are that it allows for the inclusion of non-linear relationships among variables and it can handle missing data in both categorical and continuous variables. The disadvantages are that it can be limited imputing continuous variables and computationally intensive (see Burgette and Reiter (2010); Stekhoven and B"uhlmann (2012)).

In conclusion, there are various methods to handle missing data, such as deletion, mean and mode imputation and advanced methods like Hot Deck imputation, Multivariate imputation by chained equations (MICE) and Random Forest imputation. The method to be used depends on the missing data mechanism, the number of missing values, and the nature of the data. It is important to study the missing data structure and mechanism and conduct the necessary tests, like the Little's test, to determine the appropriate method for treating missing values. Additionally, the choice of imputation method can also depend on the specific requirements of the ML algorithm being used and the nature of the problem being solved. It is important to consider the trade-offs between the different methods and choose the one that is most appropriate for the given situation. Ultimately, the goal is to minimize the impact of missing data on the model's performance and ensure that the analysis is based on a complete and accurate dataset.

In further research, this study will compare the performance of the MICE and Random Forest imputation methods on the sample under analysis.

2.4.11.4 Text Analysis

Text is a form of communication and, as such, it is considered to be both abundant and complex to learn within algorithms. Over the years, techniques have been developed to enhance the capacity of algorithms to utilize text data. These techniques provide many opportunities to better assess textual information and improve business processes, such as the underwriting process of an insurance company, which can enhance the monitoring of underwritten insurance risks and benefit both policyholders and insurers (see Ly et al. (2020)).

Natural Language Processing (NLP) aims to analyse text data by mimicking the human reading process and translating complexity of language into summarized information. The field of NLP is constantly evolving with numerous applications, including text classification, text summarization, and feature extraction. These are the focus areas of the thesis. The task of text classification involves categorizing text data into their respective classes, text summarization condenses the data into a smaller summary, and feature extraction obtains vector representations from the text (see Ly et al. (2020)).

In the life insurance industry, companies need to analyse a large number of policies at each renewal period. During this analysis period, medical surveys are commonly used to extract information and assist underwriters in focusing on the most difficult cases. For this reason, an insurance company usually conducts a survey containing information such as medical history and status updates to assess the risk of underwriting a policy for a client. The responses to the survey are often in the form of short sentences with abbreviations and technical terms. To make informed underwriting decisions, the company must accurately extract relevant information from the responses. This can be time-consuming and require manual effort, but NLP can help by automating the process and checking compliance within the survey responses (see Ly et al. (2020)).

This thesis aims to apply NLP techniques to extract relevant information from survey responses, categorize them, and simplify data analysis. This involves utilizing text classification, summarization, and feature extraction. The text information will be considered, similar to the approach in Wang (2021).

The first step in NLP pre-processing is tokenization, which splits a sequence of text into meaningful units called tokens. This process can be challenging due to difficulties in defining word boundaries and segmenting word tokens from sentences. For example, the presence of punctuation, accentuation, or language-specific characteristics can complicate tokenization.

After tokenizing the text, the next step is to clean it by removing irrelevant words such as stop words, which are frequently used and make the extraction of quality data harder. For example, in English the words “the”,” a”,” at”, or in Portuguese “por”,” na”, “tua”.

In addition, stemming removal is commonly used in text pre-processing. Stemming reduces words to their base form, reducing the dimensionality of text data and improving the efficiency of algorithms. The text is then normalized to standardize it by converting all words to the same format, such as lower or uppercase, so that the algorithm is not sensitive to the format of the text.

The Bag of Words (BOW) method is widely used in text classification and text mining, representing text as numerical feature vectors by counting the frequency of words in a document. However, BOW has the disadvantage of including irrelevant words and not capturing the context or meaning of words.

The “tm“ package in R is a popular tool for text pre-processing tasks, such as stop word removal and case folding. To summarize, text pre-processing is essential for NLP and machine learning as it formats text for algorithm comprehension and removes irrelevant information. The specific pre-processing steps depend on the task and language of the text, as explained in depth in (see Ly et al. (2020)).

2.5 Life Insurance Lapse Risk Management: A Literature Review Summary

Lapse risk refers to the possibility that a policyholder will prematurely end their insurance coverage by failing to pay the premium, resulting in the loss of coverage and termination of the contract. This risk is intertwined with the concept of “churn“, which is the frequency at which policyholders cancel or switch to a different insurance provider. These two factors have a significant impact on an insurance company’s financial well-being and future prospects, especially for life insurers where policy cancellations can negatively affect both finances and reputation (see Eling and Kochanski (2012); Kuo (2003)).

In some cases, the policyholder may receive a surrender value or cash value upon the termination of the contract. In Annual Renewable Term (ART) insurance (see 2.1.7.1), the grace period is a set amount of time after the premium due date during which the policyholder can pay their premium and maintain coverage. If the premium is not paid during the grace period, the policy may lapse, and the policyholder may lose their coverage (see Gatzert et al. (2009); Kuo (2003)).

As a result, a high lapse rate can have consequences for the insurance company. The company may have to pay out claims to policyholders who have lapsed, which can increase expenses and potentially decrease profits. Additionally, a high lapse rate may be seen as an indicator of financial instability or lack of customer satisfaction, which can negatively impact the company's reputation and ability to attract new clients (see Fang and Kung (2012)). Policyholders may decide to lapse for various reasons, such as death, changes in financial situation, dissatisfaction with the coverage or service provided by the insurance company, or the availability of more competitive options from other insurers (see Laurent (2016):pp. 240–241).

Therefore, insurance companies may adopt strategies to improve customer satisfaction and retention, as well as to identify and target potential policyholders at high risk of lapsing (see Laurent (2016):pp. 240–241). Predictive modelling techniques can be used to identify patterns and trends in customer behaviour that may indicate an increased likelihood of lapse (see 2.4). The use of data analytics, natural language processing, and machine learning can help to extract meaningful insights from unstructured data such as customer feedback and survey responses, which can be used to better understand customer needs and preferences to develop targeted retention strategies.

Additionally, risk managers should monitor lapse behaviour to prevent large financial losses. The possibility of unexpected cancellation has a significant impact on insurance companies' asset liability management and is one of the main drivers behind the large financial reserves required under the European risk management framework, Solvency II (see Section 2.3.2.1).

There are various factors that impact the management of life insurance policies, such as age, gender, occupation, medical history, lifestyle habits, economic status, culture, health status, social and lifestyle pressures, occupational risk. Policy cancellations based on these factors can impact an insurer's profitability. Unexpected changes in cancellation rates can result in liquidity issues, loss of expected profits, and imbalanced expenses (see Eling and Kochanski (2012); Kuo (2003)). Hence, it is important that insurance regulators require companies to improve their lapse risk management by monitoring quarterly reports and taking action if necessary, to maintain the stability of the system and ensure adequate capital requirements (see Section 2.3.2).

In addition, the lapse rate is also used as a key parameter in the supervisory framework for life insurance products (see Section 2.1.7). To design life insurance products, insurers may use data mining techniques to predict expected levels of lapsation in advance. This will allow them to accurately price the products and ensure that they have enough reserves to cover potential claims. In this way, proper management of lapse risk can help insurance companies to maintain financial stability and ensure that they are able to continue to provide coverage to their policyholders (see Milhaud and Maume-Deschamps (2011)).

To sum up, this literature review was based on the study by (see Shamsuddin et al. (2022)), which found that 84.40% of the 178 documents analysed were articles on the subject of life insurance lapse risk, with a growing trend in publication numbers over the years. The study also discovered a rising focus on modelling lapsation using techniques such as machine learning, policyholder behaviour analysis, pricing strategy, agent behaviour, stochastic interest rate modelling, logistic regression, and lapse risk assessment.

Furthermore, this literature review has answered the first set of research questions titled "What are the key factors that influence lapse rate in life insurance and how can they be predicted? ", and "How can lapse rate prediction models be used by insurance companies to improve their risk management and pricing strategies? ".

In particular, responding to the research question "What are the limitations of current approaches to predicting lapse rate in life insurance, and what opportunities are there for further research in this area? ", there are several gaps in the literature and opportunities for further research on the topic of predicting lapse rate in life insurance. Some potential areas for further research include:

- Developing more advanced predictive models: While current models for predicting lapse rate in life insurance are effective, there is always room for improvement. Further research could focus on developing more advanced models that incorporate new techniques and algorithms from the field of machine learning.
- Investigating the use of unstructured data: Despite the current trend of using structured data, like demographic information, to predict lapse rate in life insurance, many insurance companies collect large amounts of unstructured data, such as customer feedback, risk surveys with open-ended questions, and claims data. Further research could explore the use of this data to predict lapse rate in life insurance.
- Developing methods for handling class imbalance: Many real-world datasets used to predict lapse rate in life insurance suffer from class imbalance. Further research could focus on developing methods for handling this problem, such as resampling techniques and cost-sensitive learning.
- Exploring the use of natural language processing (NLP) techniques: Many of the data sources used in predicting lapse rate in life insurance are unstructured text data, such as survey responses. NLP techniques can be used to extract meaningful insights from this text data and improve the performance of prediction models.

Overall, the field of predicting lapse rate in life insurance is still in its early stages, and there is a lot of room for further research to improve the accuracy and robustness of prediction models, and to better understand the underlying factors that influence lapse rate.

3 | Methodology

In this chapter of the thesis, we will discuss the methodology used to tackle the binary classification problem of lapse risk in the life insurance industry. This problem involves predicting whether a policyholder is likely to lapse, or cancel, their life insurance policy. Accurately identifying policyholders who are at risk of lapsing is crucial for insurance companies as it allows them to take proactive measures to retain these customers and prevent financial losses.

The methodology for solving this problem will involve several key steps. First, we conducted an extensive literature review (chapter 2) to understand the current state of research in this area, identify relevant explanatory variables, and challenges that have been found to be associated with lapse risk. Next, we will collect and pre-process a large dataset of policyholder information, including demographic, financial, and policy-related variables. This dataset will be used to train and test machine learning models that will be used to predict lapse risk.

We will evaluate a variety of different binary classification algorithms, including logistic regression, decision trees, and random forests, to determine which model performs best on our dataset. We will use a number of metrics, including accuracy, precision, recall, and F1-score, to evaluate the performance of each model. Additionally, we will use techniques such as cross-validation and grid search to optimize the model's hyperparameters. Finally, we will interpret the results of the best-performing model and discuss the implications of our findings for the life insurance industry. We will also identify areas for future research that could improve the accuracy of lapse risk prediction models.

In summary, this chapter will provide a detailed overview of the methodology used to tackle the binary classification problem of lapse risk in the life insurance industry, including the techniques used to identify the relevant explanatory variables, the pre-processing of a large dataset, as well as the evaluation and optimization of machine learning models, and the interpretation of results.

3.1 Procedure of Machine Learning Application

The application of different predictive algorithms is a crucial aspect of machine learning. The process of predictive modelling involves utilizing various mathematical techniques on a dataset consisting of a response variable (target variable) and a set of predictors. The primary goal of this process is to identify the best model in terms of predictive performance using statistical methods. The performance of a model is evaluated based on its ability to make accurate predictions on unseen data (see Breiman (2001); Kuhn and Johnson (2013a)).

Based on Kuhn and Johnson (2013a), and Chapman et al. (2000), the procedure of applying machine learning to predict the lapse rate in life insurance can be broken down into the following key steps:

1. Data collection: The first step is to gather and organize relevant data. This may include policyholder characteristics, policy details, and historical lapse patterns;

2. Data pre-processing: Once the data is collected, it is typically necessary to pre-process it in order to prepare it for analysis. This may involve cleaning and formatting the data, as well as creating new variables or features that may be relevant to the prediction task;

3. Model training: Once the algorithms have been selected, they are then trained on the data. This involves adjusting the algorithm's internal parameters so that it can make accurate predictions based on the training data;

4. Model evaluation: After the models have been trained, it is important to evaluate their performance. This may involve using performance metrics such as accuracy, precision, and recall assessing the model's ability to predict the lapse rate;

5. Model improvement: If the model's performance is not satisfactory, it may be necessary to go back and iterate on the previous steps in order to improve the model. This may involve collecting additional data, selecting a different algorithm, or fine-tuning the model's parameters;

6. Model deployment: Once the models are performing well, it can be deployed and used to make predictions on new, unseen data. This may involve integrating the model into an insurance company's risk management processes or creating a user-friendly interface for policyholders to access the predictions.

7. Model selection: The next step is to select an appropriate machine learning algorithm used for the prediction task. There are many different types of algorithms available, and the choice will depend on the characteristics of the data and the specific goals of the analysis;

According to the International Business Machines Corporation (IBM), a professional spends around 80% of his effort preparing the data for data mining process and the remaining 20% on training and analysing models (see Patterson and Executive (2015)).

3.1.1 Software

In this research, the tasks were scripted using the RStudio programming language, version 4.2.2. The primary R package used was “caret“ version 6.0-93, which is a powerful train function that allows fitting different models with a single syntax. The “Classification and Regression Training“ package aims to simplify the process of creating and evaluating predictive models (see Kuhn et al. (2020); Kuhn and Johnson (2013a)).

3.1.2 Input Data

One important step to understanding what type of methods to use is the nature of the input data such as the type of the features, the relationship between different data points and, the dimensionality of the dataset. Features in a dataset represent specific characteristics of the data points and are represented by the columns in the dataset. Different features can have different types, but all values within the same feature must be of the same type. These types can include nominal/categorical, binary, ordinal, or numeric (see Kuhn and Johnson (2013a)).

3.2 Data Pre-Processing

Data Pre-processing is a crucial step in the Machine Learning (ML) process as it can greatly affect the outcome of the analysis. The format and scale of the input data plays a significant role in the performance of ML models and techniques. Real-world datasets often contain inconsistencies that can lead to poor model performance. To overcome these issues, pre-processing involves manipulating data through techniques such as addition, deletion, and transformation to improve the quality and suitability of the data for ML analysis (see García et al. (2015)). This includes cleaning the data, dealing with missing values (explained on 2.4.11.3), transforming variables, and scaling the data to a common range.

In this section, we will delve into commonly used pre-processing techniques in Machine Learning.

3.2.1 Feature Engineering

In this subsection of the thesis, we will delve into the process of feature engineering, which is an important step in any machine learning project. Specifically, we will focus on a few key techniques that are commonly used to enhance and optimize the features in a dataset before building a model. These techniques include Feature Scaling and Centering, Feature Encoding, Binning encoding and K-means clustering (2.4.10). Overall, this section will provide a comprehensive overview of the different feature engineering techniques that can be used to improve the performance of a machine learning model by optimizing the features in a dataset.

3.2.1.1 Feature Scaling and Centering

The measurement unit of the predictors used can affect the data analysis. Several modelling techniques require predictors to have a common scale of measure. Centre and scale are the most common transformations to improve the stability of numerical calculations and achieve a better objective. The most common used techniques for feature scaling are data normalisation and data standardisation. Normalization is a technique of uniformly scaling all the values in a dataset between 0 and 1 (see Hanafy and Ming (2022)). The normalizing formula is as follows:

$$Z = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

Standardisation transforms variable values in a way that it will have a mean of zero and a variance of one:

$$Z = \frac{x_{min} - u}{s} \tag{3.2}$$

where Z is new normalised/standardised value, x_i is the data point $(x_1, x_2, \ldots, x_n)$, μ is the sample mean, σ is the sample standard deviation, x_{min} is the sample minimum and x_{max} is the sample maximum. Both techniques are sensitive to outliers. However, Cao et al. (2016) shows that scaled models perform better than unscaled models.

3.2.1.2 Feature Encoding

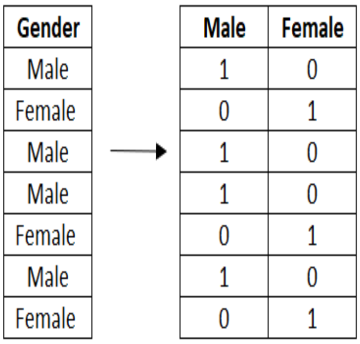
The majority of ML are built on mathematical models and techniques and only works with factors and continuous features. In case of categorical data, feature engineering is done using feature encoding techniques. It is used for the transformation of a categorical feature into a numerical variable.

One of the techniques is the one-hot encoding. This method is an important step in data manipulation as it enables categorical variables to be included in the analysis of statistical models. It increases the efficiency and facilitates the interpretation of the models.

One-hot encoding is a technique used to convert categorical variables into numerical variables by creating new binary columns (or “dummy variables“) for each category or attribute. It creates a separate binary variable for each category, with a value of 1 indicating the presence of that category and a value of 0 indicating its absence. This technique is widely used in machine learning and data analysis to handle categorical data, as it allows algorithms to work with numerical data, which is a requirement for most of them. For instance, in the case of gender, there are two characteristics: male and female. If a code of 0 refers to ‘not female“, a code of 1 refers to ‘female“. On the other hand, if a code of 1 refers to ‘male“, a code of 0 refers to ‘not male“ (see Kuhn and Johnson (2013e)).

One-hot encoding is a common method for converting categorical variables into numerical values for use in machine learning models. However, this technique can lead to the creation of highly correlated features, known as collinearity. This occurs because the same information is being represented multiple times through the new binary variables. Collinearity can make it difficult to interpret the individual effects of the predictor variables and can also affect the estimation of the model’s parameters. To address this issue, it may be necessary to remove some of the one-hot encoded features that are showing high correlation to mitigate the collinearity problem. For instance, by recording a 1 for male the information of whether the person is female is already known when the male column is 0. This double representation can lead to instability in the modelling process (see Kuhn and Johnson (2013e)).

Figure 3.1: One-Hot Encoding Example. **Source:** Hutcheson (2011)



Gender
Male
Female
Male
Male
Female
Male
Female

Male	Female
1	0
0	1
1	0
1	0
0	1
1	0
0	1

3.2.1.3 Binning encoding

Binning, also known as grouping or bucketing, is a widely used pre-processing technique in machine learning, data analysis, and as an algorithm to accelerate learning tasks. It is used to group, summarize, and simplify data by discretizing the features into bins (groups or buckets) in order to improve the understanding of the nonlinear dependence between a variable and a given target while reducing the model complexity (see Navas-Palencia (2020)). For example, Binning can be applied to group Body Mass Index (BMI) values into categories such as underweight, healthy weight, overweight, and obese.

There are several binning techniques, supervised and unsupervised (see Deckert and Kummerfeld (2019)). Supervised binning techniques optimize predictive information for an outcome of interest, such as entropy-based binning. Unsupervised techniques like equal-width, equal-size or equal-frequency interval binning, which don't require any specific outcome to be optimized, they are based on the distribution of the data.

These bins can be selected by using several methods, such as clustering. One commonly used method for learning the "optimal" binning is the K-means algorithm (see Deckert and Kummerfeld (2019); Gupta et al. (2010)). Overall, binning is a useful technique for simplifying and understanding complex data, but the choice of binning technique and the number of bins should be carefully considered to ensure the best results.

3.2.2 Anomaly Detection and Treatment

The proper detetion and analysis of anomalies is required to prevent form a global bias of the data. Anomalies are observations that appear to be statically different from the rest of the data. They are a minority of objects (observations, cases, or data points) that are inconsistent with the pattern suggested by most objects in the same dataset. In case of continuous data can be called Outliers. Those data points are not representative to train correctly, so it is necessary to remove or transform them (see Domingues et al. (2018)).

There are several methods to detect and remove anomalies. Most of the ways to deal with them are by deleting, transforming, binning, and imputing. Also, the anomalies differ in each type of data. The algorithms of anomaly detection are categorized according to whether the dataset has an outcome labelled data to build the detection model or not. These algorithms are named as Supervised, Unsupervised, or Semi-Supervised (see Goldstein and Uchida (2016)). In anomaly detection, unsupervised algorithms are used more frequently than supervised, because of the application without a previously labelled data set. Likewise, the supervised classes of this datasets used to be unbalanced, since the class of the anomalies would be much smaller than that of normal points and this can affect the efficiency of the supervised algorithms.

Anomaly Detection algorithms can report an anomaly using scores and/or labels. The Scores method assigns a score to each point according to their degree of anomaly. In the Label method, it is possible to directly output a label of the anomaly (see Chandola et al. (2009)). In this research, it is used the following two unsupervised methods:

The first method is a univariate unsupervised approach, called Tukey's method or Inter Quartile Range (IQR). In 1977, John Tukey published a method that displays information about continuous univariate data with a simple graph called boxplot (or box and whisker plot). It describes the spread of a distribution, using the plot's whiskers like the median, lower quartile, upper quartile, lower extreme, and upper extreme of a data set. Outliers can be identified as the points that lie beyond those plot's whiskers. This method it less sensitive to extreme values of the data than methods using the sample mean and standard variance because it uses quartiles which are resistant to extreme values. The IQR is the distance between the lower (Q_1) and upper (Q_3) quartiles ($IQR = Q_3 - Q_1$). And any data point that lies outside the range of 1.5 times the IQR below the first quartile (Q_1-($1.5 * IQR$)) or 1.5 times the IQR above the third quartile ($Q_3 + (1.5 * IQR)$) is considered an outlier (see Piegorsch (2015)). One possible way to avoid outliers is through binning and categorizing the data.

The second method, known as Isolation Forest is an unsupervised machine learning technique that is used to identify anomalies in a dataset. It's based on decision tree algorithms and works by randomly subsampling the input dataset, generating an isolation tree for each sample. The method starts by randomly selecting a feature and then randomly selecting a split value between the minimum and maximum values of that feature. It continues the process until each data point is completely isolated from the rest. The use of random partitions makes it likely for anomalies to be located near the root of the tree, as they are less frequent than regular data observations. The data points are scored based on the number of partitions required to isolate them, and a score greater than or equal to 0.6 is considered a potential anomaly. Isolation Forest is computationally efficient, it allows for the application to large datasets and it may produce better results with a smaller sample size (see Liu et al. (2008)).

3.2.3 Feature Selection

Dimensionality reduction is a crucial step in machine learning problems as it can greatly impact the performance of algorithms. One commonly used pre-processing method for achieving this is feature selection. This method aims to identify the most important dependencies or correlations between input and target features, with the goal of reducing the number of redundant features and decreasing the learning time while potentially improving classification accuracy and avoiding overfitting (see Haar et al. (2019)).

In supervised learning, where each feature is associated with a class label, it is particularly important to identify the features that are most important in predicting the target feature. Feature selection is applied to choose a subset of the original features based on certain evaluation criteria. The advantages of feature selection include data quality improvement, less computational time, improved predictive performance, and efficient data collection.

There are two main categories of feature selection techniques: unsupervised and supervised (see Haar et al. (2019)). Unsupervised feature selection techniques focus on selecting features based on their inherent characteristics, such as the correlation or similarity between features. Examples of unsupervised feature selection techniques include Factor Analysis of Mixed Data (FAMD), Low-Variance method, and Correlation Criteria.

Supervised feature selection techniques, on the other hand, take into account the target variable and select features based on their relationship with the target variable. These techniques are typically used when labelled data is available and the goal is to improve the performance of a supervised learning model. Examples of supervised feature selection techniques include chi-squared test and ANOVA (F-test).

Supervised feature selection techniques can be further divided into three types: filter methods, wrapper methods, and embedded methods. Filter methods, which are based on statistical measures, do not use learning algorithms to select features. They are considered efficient, computationally cheaper, and can avoid overfitting. However, they do not take into account the bias and heuristics of the learning algorithms. In this research, the unsupervised method FAMD will be compared against the supervised methods, filter methods in particular (see Kuhn and Johnson (2013f)).

3.2.3.1 Correlation Criteria

Feature correlation is a statistical method to calculate the relationships between various data features in a dataset. This method can be useful in determining dependencies between the data features and how each feature effects the target feature. In case of continuous variables is possible the use of Pearson´s correlation (see Grice (2013):pp. 70–89).

The method classifies the strength of association between two features with a correlation coefficient which takes a value between -1 and +1. The positive correlation means a strong association between the features. Whereas the negative correlation means a weak association between the features. If the correlation is zero, it means there is no association between the features. Supposing two features' observations $x_1, \dots, x_n$ and $y_1, \dots, y_n$, the correlation coefficient r is calculated as follows

$$r = r_{xy} = \frac{\sum(x_i y_i - n\bar{x}\bar{y})}{(n-1)s_x s_y} \tag{3.3}$$

where, n is the sample size, $\bar{x}$ and $\bar{y}$ are the means of x and y respectively. s_x and s_y are the standard deviations of x and y.

According to (see Cohen (1988)) an absolute value of r of 0.1 is classified as small correlation, an absolute value of 0.3 is classified as medium correlation and of 0.5 is classified as large correlation.

In datasets containing multiple features, the correlation values between the data features can be calculated with the covariance matrix. This matrix encodes the correlation coefficient of any linear combination of the entries (see Probability and Data Science (2020)).

3.2.3.2 Low-Variance method

Low variance method is used to remove features whose variance are below a predefined threshold. It diagnoses the features that have the same values for all instances, so the variance is 0 or the features that have few unique values relative to the number of samples. These features should be removed since it cannot help discriminating instances from different classes. For instance, supposing a dataset with only Boolean features, the features values are either 1 or 0 (see Li et al. (2017)). These Boolean features are Bernoulli random variables, so its variance value is computed as:

$$\mathrm{var}(f_i) = p(1-p) \tag{3.4}$$

Where p denotes the percentage of cases that take the feature value of 1. Therefore, the feature with variance score below a predefined threshold can be removed. he best advantage of this method is reducing the model-fitting time without reducing model accuracy (see Li et al. (2017)).

3.2.3.3 Factor Analysis of Mixed Data (FAMD)

Factor analysis of mixed data (FAMD) is a statistical technique used to analyse data sets that contain both quantitative and qualitative features. The FAMD algorithm combines elements of principal component analysis (PCA) and multiple correspondence analysis (MCA) to analyse the data. In other words, it uses PCA for quantitative variables and MCA for qualitative variables (see Visbal-Cadavid et al. (2020); Kassambara (2017):pp. 108–119).

Principal component analysis (PCA) is an unsupervised algorithm that is used to reduce the dimensionality of a dataset by transforming the data into a new set of variables, called principal components, that capture the most important information in the original features. This is done by finding the eigenvectors and eigenvalues of the covariance or correlation matrix of the data, and using them to create a new system of coordinates that is composed of the principal components in descending order of variance or eigenvalues. By excluding the principal components with lower eigenvalues, the dimensions of the original dataset are reduced. The Kaiser criterion is often used to determine which factors to preserve (see Kassambara (2017):pp. 12–50).

Multiple correspondence analysis (MCA) is used to analyse patterns of relationships among categorical dependent variables. It is similar to PCA, but is used for analysing categorical data instead of quantitative data. MCA is often used to analyse survey data, with the goal of identifying groups of individuals with similar profiles based on their answers to the survey questions (see Kassambara (2017):pp. 83–106).

3.2.3.4 Chi-Square (χ^2) test

The Chi-Squared filter method is a widely employed technique for feature selection in machine learning. It is based on the Chi-Squared statistical test, which is a measure of dependence between two categorical variables. In the context of feature selection, this test is used to determine the correlation between each input feature and the target variable in a classification problem. The test calculates the Chi-Squared statistic for each feature, and the features with the highest statistics are retained as the most informative and relevant for the classification task. This method is commonly used in combination with other feature selection techniques to achieve optimal results (see Singhal and Rana (2015)).

The following hypothesis are:

- Null Hypothesis (H_o): Two variables are independent.
- Alternate Hypothesis (H_1): Two variables are not independent.

The test statistic of Chi Square calculates the correlation strength of each feature individually by the following equation (see Oakes et al. (2001)):

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \tag{3.5}$$

where,

- k is the number of attributes
- n is the number of classes
- O_{ij} is the number of instances with value i for attribute and j for class
- E_{ij} is the expected number of instances for O_{ij}.

3.2.3.5 ANOVA test (F-test)

The technique known as Analysis of Variance (ANOVA) is a collection of parametric statistical tests used to determine whether there is a significant difference between the means of two or more samples (see Piegorsch (2015):pp. 242–248).

ANOVA utilizes F-tests, which calculate the ratio of variances, to statistically test the equality of means between the samples. This test is commonly applied in classification problems where the input features are numerical, and the target feature is categorical. It helps to determine whether there is a statistically significant relationship between the input features and the target feature, with the null hypothesis being that all group population means are the same. The null hypothesis is rejected when the p-value is less than or equal to the specified significance level α. The smaller p-value, the stronger the evidence to reject the null hypothesis. The ANOVA procedure is a powerful tool for identifying relationships between variables, and it is widely used in many fields such as psychology, sociology, and engineering.

The following hypothesis are:

- Null Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
- Alternate Hypothesis $H_1 : \exists(i, j)$ s.t. $(i \neq j) : \mu_i \neq \mu_j$

where, μ_k represents the mean value of the level k of the factor in the population, μ the mean value of the population, $(i = 1,..,k)$ and $(j = 1,..,n_i)$

3.3 Model Training

In this section, we will delve into the various techniques used in the process of training a machine learning model. One of the most important aspects of model training is the process of splitting the data into training and testing sets. This allows for evaluating the model's performance on unseen data, providing a more realistic estimate of its performance on new, unseen data. Additionally, we will explore the techniques used for Tuning and Improving Machine Learning Models, such as cross-validation, and grid search. These techniques help to optimize the model's performance by tweaking its parameters and features. Furthermore, we will discuss resampling techniques such as bootstrapping and k-fold cross-validation which can be used to improve the model's ability to generalize to new data by creating multiple versions of the training set and can help to reduce overfitting and improve the model's robustness. Overall, this section will provide a comprehensive understanding of the methods and techniques used to train a machine learning model for optimal performance.

3.3.1 Data Splitting Method

Data Splitting is a method commonly used in ML to split data into a train, test, or validation set. The training data is used to fit the model with a set of parameters, while the validation set is used to evaluate the performance of the model with different hyperparameter settings. Testing the model on the test set, which is a separate set of data that is not used in the training or validation process, is used to provide a final evaluation of the model's performance. This approach allows for the identification of an efficient set of model parameters without compromising the integrity of the test data.

The appropriate application of this method can be treated as a statistical sampling problem. One of the most used methods is Simple random sampling, which samples are selected randomly with a uniform distribution (see Kuhn and Johnson (2013g)).

3.3.2 Tuning and Improving Machine Learning Models

Hyperparameter tuning is a crucial step in machine learning that involves selecting the best set of hyperparameters for a learning algorithm. The purpose of this process is to avoid overfitting and underfitting, and to achieve the highest possible performance of the model.

One popular method for optimizing the parameters of a model is grid search. This approach involves creating a grid of hyperparameters and training the model based on each possible combination. Although this method can be time-consuming, it often leads to better performance (see Hossain and Timmer (2021)).

Another method for hyperparameter tuning is random search. This method involves randomly selecting combinations of hyperparameters from a grid to train and test the model. The goal of this method is to identify new combinations of parameters or to discover new hyperparameters that may not have been considered in a grid search. However, the randomly selected hyperparameters may not yield an optimal result (see Hossain and Timmer (2021)).

In R, the "caret" package is a well-known machine learning package, by default it uses default grid search as the method for optimizing tuning parameters during the training process (see Kuhn et al. (2020)). This method happens when the user does not specify a grid to use for the tuning parameters. This method creates a grid that covers a range of parameter values that are considered sensible for that specific model.

Hyperparameter tuning is a crucial aspect of machine learning as it involves selecting the optimal values for the parameters of a learning algorithm. Complex models, characterized by high values for their hyperparameters, can lead to overfitting, which is when a model performs well on the training data but poorly on unseen data. To mitigate this risk, it is common practice to divide the data into three sets (3.3.1). Recent studies have highlighted that the validation set alone is not enough to measure the model's performance (for example, Harrington (2018)). Cross-validation is a commonly used method to evaluate the true prediction error of models and to tune model parameters.

3.3.3 Resampling Techniques

In this subsection of the thesis, we will explore resampling techniques that are commonly used in machine learning and statistical modeling. These techniques are used to overcome issues such as small sample size and class imbalance, which can negatively impact the performance of a model. The two main techniques that we will focus on are Bootstrap and K-fold Cross Validation (CV).

3.3.4 Bootstrap

The bootstrap algorithm is a statistical method for resampling data, it is used to estimate statistics on a population by randomly selecting subsets of samples with replacement from the dataset. The bootstrap method allows for the estimation of the properties and statistics of a potential distribution without the need for knowledge of its true underlying distribution. The selected subset of samples is used to fit and train the model, while the remaining samples are used to validate the model. By repeating this process multiple times, the bootstrap algorithm provides a better representation of the sample population and a more robust estimate of the model's performance. The final estimation of the model's performance is computed as the average of the scores obtained from the validation set over the multiple iterations of the bootstrap process. Bootstrap is often used to estimate the variability of a model's performance and to construct confidence intervals for the model's performance (see Brownlee (2018)).

3.3.5 K-fold Cross Validation (CV)

The K-fold cross-validation algorithm is a technique used to assess the performance of a machine learning model by dividing the dataset into k smaller subsets, or folds. In this method, $k-1$ of the folds are used to train the model, while the remaining fold is used as a validation set. This process is repeated k times, with each fold being used as a validation set once. The performance metrics of the model are then averaged across the k estimates of each validated fold, and the best averaged predictive score is used as the optimal model performance. This technique is particularly useful in preventing overfitting, as the validation set is independent from the other k sets. This allows for a more robust evaluation of the model's performance as it is exposed to different data (see Santos et al. (2018)). Additionally, the repeated validation can provide a more consistent estimate of a model's performance on unseen data, as compared to an evaluation based on a single train/test split.

Ten-fold stratified cross-validation is a commonly applied variation of K-fold cross-validation, where the data is divided into 10 equal-sized folds, and the samples are chosen in a way that each fold contains roughly the same proportions of samples of each target class (see Berrar (2018)). Furthermore, by computing a confidence interval around the performance estimate, it is possible to evaluate its uncertainty and make more informed decisions about the model's performance.

In summary, bootstrap is a method to estimate the variability of a model's performance, while k-fold cross-validation is a method to estimate the model's generalization performance.

3.4 Model Evaluation

In this section of the thesis, we will discuss the various techniques and metrics used to evaluate the performance of machine learning models. The goal of model evaluation is to determine how well a model is able to make accurate predictions on new, unseen data. This is an essential step in the machine learning process as it allows us to identify the strengths and weaknesses of a model and make informed decisions about which model to use in a given application. In summary, the purpose of this section is to provide an understanding of how to evaluate the performance of different models and select the best one for a given problem.

3.4.1 Performance Metrics

In order to evaluate the performance of different models and methods applied in a predictive classification task with a binary target variable, various metrics can be used. One common approach is to rely on the confusion matrix, which provides information about the true positive, true negative, false positive, and false negative predictions made by the model. Metrics such as the area under the Receiver Operating Characteristic (ROC) curve and the accuracy of class probability estimates can also be used to assess model performance. These metrics provide insight into the trade-off between the true positive rate and the false positive rate, as well as the overall accuracy of the model's predictions. Additionally, other metric such as F1-score, precision, recall, and specificity are also helpful in evaluating the performance of a classification model.

3.4.1.1 Confusion Matrix

The assessment of predictive classification models with binary target variables can be achieved through the use of various performance metrics, such as those based on the confusion matrix. The Confusion matrix is a tabular representation of true and predicted class labels, and it is used to evaluate the model's accuracy, precision, recall, and other relevant statistics. Additionally, classification models can also be evaluated based on their predicted class probabilities, which provide a measure of the model's confidence in its predictions. For example, in the case of insurance Lapse, a classification model can predict the probability of lapsing, allowing the company to make more informed decisions. The selection of a specific target class is determined by a threshold, typically set at 0.5, and the model's performance is often evaluated through the use of graphical probabilistic performance measures such as ROC curves and precision-recall curves (see Kuhn and Johnson (2013b)). The function $confusionMatrix$ in R can be used to compute various summaries for classification models.

Table 3.1: Confusion matrix table **Source:** Authors.

CLASSIFICATION METHOD			**Actual Classes**		**Total**
			Lapse	**Non-Lapse**	
			Positive	**Negative**	
Predicted Classes	**Lapse**	**Positive**	TP	FP	TP+FP
	Non-Lapse	**Negative**	FN	TN	FN+TN
	Total		TP+FN	FP+TN	n

- True Positive (TP)

  - The predicted value matches the actual value
  - The model predicts correctly that the policyholder will not renew the policy (Lapse)

- True Negative (TN)
  - The predicted value matches the actual value
  - The model predicts correctly that the policyholder will renew the policy (Not Lapse)

- False Positive (FP): Type 1 error
  - The predicted value was falsely predicted
  - The model predicts falsely that the policyholder will not renew the policy

- False Negative (FN): Type 2 error
  - The predicted value was falsely predicted
  - The model predicts falsely that the policyholder will renew the policy

There are several performance metrics that can be calculated from a confusion matrix such as (detailed in Kuhn et al. (2015)):

- The simplest metric is the accuracy rate which is measured by the ratio of True Positive (TP) and True Negative (TN) on the total data set. The higher the accuracy the better the model. However, may not be the case with imbalanced classes where the frequency of the minority class is not equally represented. In this case, the Kappa statistic may be more relevant (see Vieira et al. (2010)). The classification accuracy can be defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.6}$$

- The Kappa statistic is used to evaluate interrater reliability. This means that it represents the level whether the data collected in the study are correct representations of the variables measured. The kappa can range from -1 to +1, where higher rate represents better agreement among the rates. In other words, higher the Kappa score means that there is a bigger difference between the accuracy and the null error rate (see McHugh (2012)).

- Precision measures the proportion of correctly positive events from all events identified as positive. Low precision indicates a high number of false positives. The precision can be defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.7}$$

- Sensitivity or recall measures the proportion of events identified as positive on all positive events. Low recall indicates a high number of false negatives positives, can be defined as:
$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.8}$$

- Specificity measures the proportion of events identified as negative on all negative events positives. A high Specificity reflects the model was good at identifying true negatives, can be defined as:
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.9}$$

- F1-Score is defined as the harmonic mean of Precision and Recall. This goodness-of-fit measure focus on the analysis of the minority class, $F1 \in [0,1]$ where a value of 1 is the best possible value. This means that a good F1-score represents low false positives and low false negatives. It is defined as:
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.10}$$

- Prevalence measures the proportion of all positive events on the total data set, defined as:
$$\text{Prevalence} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.11}$$

- Detection Prevalence measures the number of predicted positive events (both true positive and false positive) divided by the total number of predictions, defined as:
$$\text{Detection Prevalence} = \frac{\text{TN} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.12}$$

- Positive Predictive Value (PPV) is the proportion of positive test results, defined as:
$$\text{PPV} = \frac{\text{Sensitivity} \cdot \text{Prevalence}}{\text{Sensitivity} \cdot \text{Prevalence} + (1 - \text{Specificity}) \cdot (1 - \text{Prevalence})} \tag{3.13}$$

- Negative Predictive Value (NPV) estimates the proportion of subjects with a negative test results.
$$\text{NPV} = \frac{\text{Sensitivity} \cdot (1 - \text{Prevalence})}{(1 - \text{Sensitivity}) \cdot \text{Prevalence} + \text{Specificity} \cdot (1 - \text{Prevalence})} \tag{3.14}$$

- Balanced Accuracy is an adjusted development on the standard accuracy metric used to assess better the performance of a classification model with imbalanced target.
$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{3.15}$$

- Logarithmic loss is a measure that assesses the performance of a binary classification model. It is calculated by taking the natural logarithm of the probability that the model's prediction for a given sample is the actual class. This score is a continuous value ranging from 0 to 1, with 0 indicating a perfect prediction and 1 indicating a completely incorrect prediction. Log loss is particularly useful for evaluating models that output probability estimates for their predictions, such as logistic regression or neural networks. It is often used in machine learning competitions and is preferred over accuracy when the class distribution is imbalanced, as it penalizes incorrect predictions more heavily gives more details about this measure (see Nasteski (2017)).

3.4.1.2 Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC-ROC)

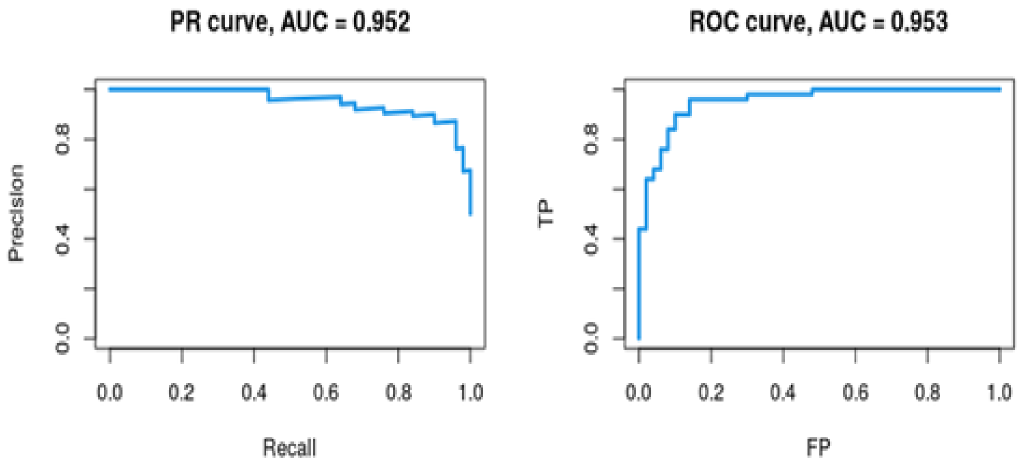
Receiver Operating Characteristic (ROC) is a performance measure of the binary classification problem. It evaluates the relationship between True Positive (TP) rate and False Positive (FP) rate. The ROC plots the trade-off between the true positive rate and the false positive rate at different thresholds. The ROC curve can be used to determine alternate cut-off values for class probabilities. The optimal cut-off point is the one that has the highest proportion of true positives and the lowest proportion of false positives, which is located on the upper left corner of the plot. The ROC can also be measured as a single metric by calculating the area under the ROC curve (AUC). The Area Under the Curve (AUC) of the ROC curve summarizes its overall performance, with a score of 1 representing a perfect classifier and 0 indicating a poor classifier (see Kuhn and Johnson (2013b)).

In summary, ROC curve and AUC are vital tools for assessing binary classifier performance and setting the optimal threshold for class probabilities. In imbalanced target variable scenarios, accuracy alone may not provide a sufficient evaluation. AUC-ROC, which considers both true positive and false positive rates, is a preferred metric in these cases. It provides a more comprehensive evaluation of classifier performance, especially when the positive class is rare or the class distribution is imbalanced. Thus, ROC is often favored as a performance metric in imbalanced data classification problems (see Agarwal et al. (2016)).

3.4.1.3 Precision-Recall Curve and Area Under Curve (AUC-PR)

Precision-Recall curves evaluate the performance of a classification model in identifying a particular type of data. They plot precision on the y-axis and recall on the x-axis, and a model with high scores in both is considered effective. The area under the Precision-Recall curve, obtained by integrating precision and recall values at various thresholds, is another evaluation metric for binary classifiers. The higher the area under the curve, the better the classifier's performance. This metric is often used in applications where both precision and recall are important, such as in fraud detection or medical diagnosis, as it balances both true positive and false positive rates (see Boyd et al. (2013)). It is particularly useful when the positive class is more frequent or the class distribution is balanced.

Figure 3.2: AUC-ROC & AUC-PR curves example. **Source:** Marku and Pancaldi (2022)



3.5 Model Selection

Model selection is the process of identifying the most appropriate machine learning model for a given problem. This involves training a variety of models using different configurations and evaluating their performance on a validation dataset. The ultimate goal is to select the model that performs best on unseen data, while also taking into account business costs and other constraints. The evaluation of models can be performed using a variety of metrics such as accuracy, precision, recall, F1 score, Logarithmic Loss, and others.

Cross validation is a widely used technique for estimating the performance of a model on unseen data. This method involves dividing the data into multiple subsets, training the model on one subset, and evaluating its performance on the remaining subsets. The average performance across all subsets is then used as an estimate of the model's performance on unseen data.

In model selection, the best model is typically chosen based on its predictive accuracy, as measured by metrics such as AUC-ROC, Kappa, F1-score, and Logarithmic Loss. These metrics provide a quantitative way to compare the performance of different models.

A box-and-whisker plot is a useful visualization tool for comparing the distribution of predictive accuracy across different models. This type of plot is commonly used in statistics and data science to display the distribution of a dataset (see Agarwal et al. (2016); Loisel et al. (2021)). The box-and-whisker plot can be used to compare the performance of different models by plotting their out-of-sample ROC scores. The model with the highest median ROC score is identified by the dotplot R-function, which is a visually simpler alternative to the bwplot function, available in the "lattice" package, version 0.20-45. This method can be particularly useful in the case of class imbalance, as well as the use of 10-fold cross validation for evaluating the performance of models in such scenarios (see Holloman (2021)).

It's worth noting that the above-mentioned approaches should be used with caution, as the selection of the best model depends on the specific characteristics of the data, the goal of the analysis and the business requirements. Furthermore, the use of a single metric such as accuracy may not be enough to evaluate the model's performance, especially in imbalanced datasets, where other metrics such as AUC-ROC, precision, recall, and F1-score should be considered as well.

3.6 Model Variable Importance

Over the years, the importance to quantify the power of the relationship between the predictors and the outcome have been growing. The results from ML models can be complex, so model-specific variable importance measures have been developed.

Model Variable Importance is a measure of the contribution of each variable to the performance of a machine learning model. It helps identify the most important variables in a model, which can then be given more weight or focus on the analysis. There are several methods for calculating model variable importance, such as permutation importance. These methods involve calculating the change in model performance when a specific variable is removed or modified. Variables that result in a significant change in model performance are considered the most important. Model variable importance is useful for identifying the key drivers of a model's prediction and for understanding the relationships between variables in a dataset. One example is costumer churn prediction which predictive importance is essential to help improve a service and product (see Groll et al. (2022)).

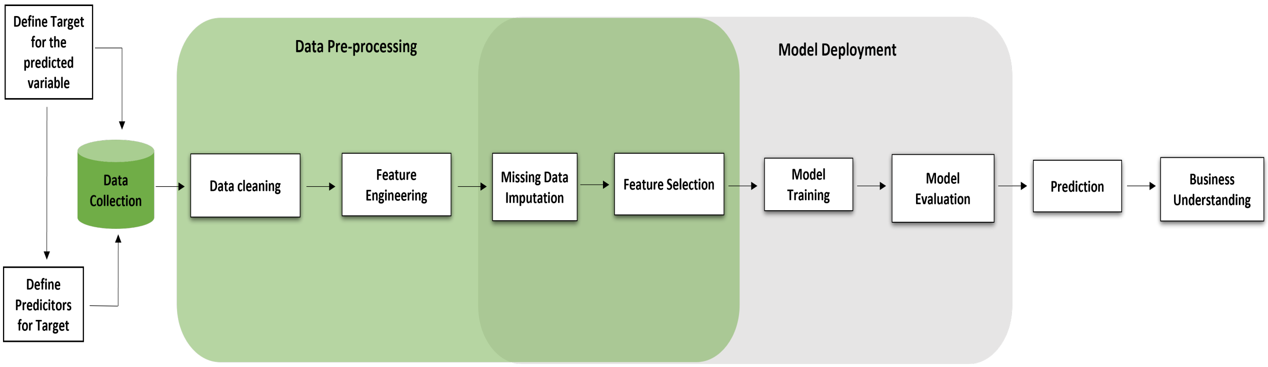
The generic function $varImp$ in R can be used to characterize the general effect of predictors on the model. As well as the $filterVarImp$ function can calculate the area under the ROC curve when the outcome variable is an R factor variable (see Kuhn and Johnson (2013h)).

4 Predicting Lapse Risk in Life Insurance: A Case Study

This chapter presents the experimental work of applying machine learning techniques from literature to predict Lapse risk. It examines the assumptions, choices and limitations in the methodology (see chapter 3), and literature (see chapter 2). It provides an overview of the data pre-processing techniques employed, including feature engineering and transformation, with a focus on extracting relevant features from the dataset. It also highlights the importance of collecting and structuring the data in the most optimal way to train machine learning models.

Additionally, the chapter emphasizes the significance of comparing various models in the context of re-underwriting decisions, as real-world data contains diverse formats such as text. Some techniques for addressing these issues, including natural language processing, are also discussed. Figure (4.1) outlines a simplification of the applied machine learning prediction process that was followed in the experimental study. The main titles of this chapter align with the steps outlined in the process figure.

Figure 4.1: Implementing applied machine learning methodology. **Source:** Authors.



4.1 Data Collection

The data used in this research was obtained from an insurance company's database and preprocessed in accordance with their guidelines. *SAS*, a powerful statistical software for data management and business intelligence, was used to structure the dataset, as the proper organization of data is essential for achieving the research's objectives.

The research relied on primary data, specifically a risk survey dataset of policyholders from the insurance company. The survey, conducted between February 2014 and December 2021, included 36,065 individual policies of twelve ART products (explained on Section 2.1.7.1) and was structured into six categories of questions designed to evaluate policyholders' medical, behavioural, routine, and physical underwriting risks. Conducted in Portuguese, the survey dataset is summarized in the accompanying table (4.1).

Table 4.1: Survey's Variables **Source:** Authors.

Variable name	Description
ID	Unique Policy (Contract) Identification Code (ID) per policyholder
ID_PRODUTO	Life Product Identification Code (ID)
GRUPOID	Category (Group) of questions identification code (ID)
TIPO	Question (QST), Answer (ANS) or Group (GRP)
DESCRICAO	Description of the variable TIPO
DATE_ANS	Policyholder's response Date

The variable $TIPO$ is coded to indicate the type of data represented by the variable $DESCRICAO$. The QST codes identify $DESCRICAO$ as a question, ANS codes indicate $DESCRICAO$ as an answer to a specific question, and GRP codes identify $DESCRICAO$ as the name of each category in the survey. Each policy (indicated by the ID variable) is linked to a unique policyholder who has purchased a specific product. For example, in the appendix (B.1), the policyholder with $ID = 1$ purchased product 10 and answered 13 questions.

The primary objective of this research was to analyse the responses of the policyholders who participated in the survey. To facilitate this analysis, the survey questions were encoded into codes (appendix A) and transformed into variables using the Transpose function in SAS. The variable $DESCRICAO$ was transposed, resulting in the creation of three new variables: QST, ANS, and GRP. The encoded questions (variable QST) were then further transposed and organized into new variables, with some variables removed.

In addition to the primary data, two supplementary databases were used to aid in the processing of the primary data. The first supplementary database, named Policy database, contained detailed policy information such as contract amounts, policy type, coverage details, etc. The second supplementary database, named Policyholder database, contained information about the policyholders such as age, gender, occupation, and other demographic information.

To merge these databases, a common key *ID* was used which is a unique policy identification code present in both the databases. Additionally, the *Entity_ID* variable serves as the policyholder identification code. The merging process was done using *SAS*, which several assumptions and decisions were made during the process to ensure accurate and complete data.

The policyholder characteristics could be found in the Policyholder database and the contract information in the Policy table. These databases were merged using the primary key *Entity_ID*. The variables were also encoded to ensure data consistency, and their descriptions were added to provide context. Only policyholders who participated in the survey were included in the merging process to ensure the data was accurate and relevant. To ensure that maximum information was obtained from each policyholder, missing information was filled using historical data.

The process of merging databases and preparing them for analysis involved several steps. The initial step was to eliminate duplicate records based on *Entity_ID* within the merged data, which was achieved by deduplicating the data.

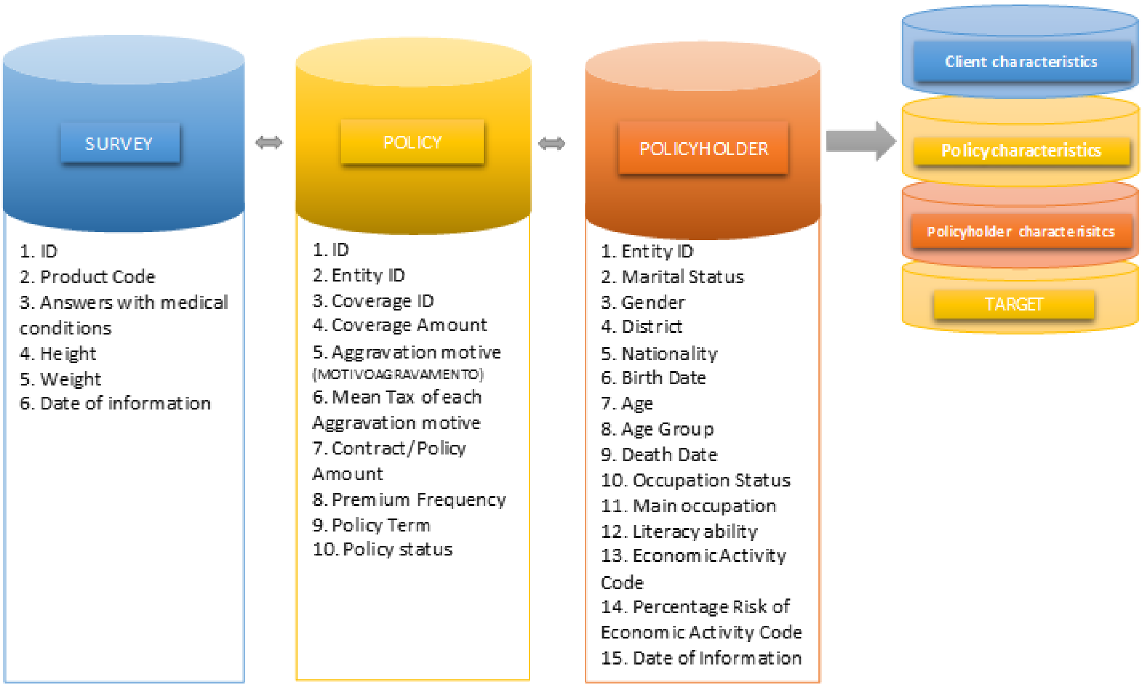
The next step was to group the dataset by *Entity_ID* and summarize the policy information for each policyholder. This resulted in a clean and well-structured dataset, where each record corresponded to a unique policyholder and contained all their policy information in one place.

After that, the aggregated dataset was merged with survey data using an inner join statement by the variable *ID*, ensuring only records with matches in both datasets were included in the final dataset. This resulted in a merged dataset that includes both policyholders who renewed and those who did not, with necessary variables for analysis such as policy details, policyholder characteristics and survey responses.

The final step was to construct the target feature, a variable that the prediction models will use to classify policyholders. In this case, the target feature was based on the policy status at 11/07/2022. It will be used to train the models and make predictions on which policyholders are likely to lapse in the future. Additional feature engineering is performed to make the models more robust, such as creating new variables or transforming existing ones. This process is illustrated in Figure (4.2).

At this point, the final dataset consists of 108 features, and is now prepared for further analysis and applications. The structure of the final dataset is outlined in the appendix (B.1) for reference.

Figure 4.2: Dataset Collection. **Source:** Authors.



4.2 Data Cleaning

4.2.1 Anomalies and Outliers

Before proceeding with the pre-processing of the variables, it was important to conduct a thorough analysis of the data quality in the dataset. This analysis, known as Data Cleaning, aims to address any anomalous data present in the dataset found in Appendix (B.1). This dataset is divided into two groups: Quantitative variables and Qualitative variables. For this analysis, the description variables and identification codes (ID) variables have been excluded.

To detect outliers in the Quantitative variables, the Inter Quartile Range (IQR) method and boxplots (or box and whisker plots) have been applied. Firstly, a boxplot for each quantitative variable has been created for the target variable. The boxplots reveal that all variables show a behaviour susceptible to the presence of outliers, which are data points that lie outside the overall distribution pattern.

To identify and quantify these points, the IQR method is applied in R. This method considers any data point that lies outside the range of 1.5 times the IQR below the first quartile ($Q1$˘$(1.5IQR)$) or 1.5 times the IQR above the third quartile ($Q3+(1.5IQR)$) as an outlier. The results of this method are as follows:

- The NUM_AGE (Age) variable has 61 outliers (about 0.17% of the total policies). These observations represent policyholders who are younger than 9 years old and older than 73 years old. Statistically, these points lie outside the Inter Quartile Range (see figure B.18);

- The $X111$ (Weight) variable had 277 outliers (about 0.77% of the total policies). These observations represent policyholders who weigh less than 32.5 kg and more than 108.5 kg. Statistically, these points lie outside the Inter Quartile Range (see figure B.20);

- The $X112$ (Height) variable had 125 outliers (about 0.35% of the total policies). These observations represent policyholders who are less than 145 cm and more than 193 cm tall. Statistically, these points lie outside the Inter Quartile Range (see figure B.19);

- The $VALORCOBERTURA$ (Coverage Amount) variable had 4093 outliers (about 11.35% of the total policies). These observations represent policies that have coverage amounts inferior to 2,500 euros and superior to 62,500 euros (see figure B.21). As a result, this variable was binned into categories;

- The $MONTANTECONTRATO$ (Contract Amount) variable had 143 outliers (about 0.40% of the total policies). These observations represent policies that have contract amounts superior to 275,000 euros (see figure B.22). As a result, this variable was binned into categories.

It should be noted that in this analysis, the outliers were not considered necessarily errors but assumed that they could be considered as exceptions or rare cases. Therefore, it was important to check if they were meaningful and decide whether to keep or remove them.

After conducting a thorough analysis of the data in the Appendix dataset (B.1), it was determined that the best approach for handling outliers in the quantitative variables was to use binning techniques. However, it was decided to remove the outliers of the NUM_AGE variable, in accordance with the business guidelines of the insurance company. As a result of these actions, there were 35,980 policyholders remaining in the data. This data removal was performed after processing the data on Section (4.3).

In terms of the qualitative variables, the method used for identifying and addressing anomalies was isolation forests. The $isolation.forest()$ R-function, available in the $isotree$ package version 0.5.17, was used to score and identify potential anomalous observations. A score of more than or equal to 0.6 was considered to indicate a potential anomaly. After analyzing the output, it was determined that there were no anomalous observations present in the data. The highest score observed was 0.59 and the lowest score was 0.31.

4.3 Data Pre-processing

In this section, we will delve into the methodologies employed in the handling of the data collected. This process, includes mainly feature engineering (3.2), which is crucial as it involves exploring and modifying the data to extract relevant features. This process included extracting new features from the data and defining derived features that can be reproduced from other existing features.

The first stage of this process involved exploring the data and transforming it based on quantitative assessments of the information. This included identifying and removing any irrelevant or redundant features, as well as encoding categorical variables so that they could be processed by Machine Learning (ML) models.

Once the relevant features have been extracted, they were further encoded to be used as input for the ML models. This involved converting categorical variables into numerical ones, as most ML models are only able to handle numerical inputs.

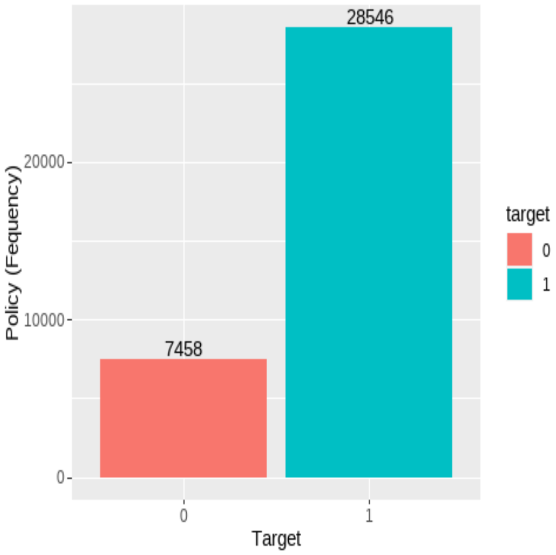
The analysis of the data was subdivided into four parts: the target variable, survey features, policy features, and, policyholder features. Each of these parts represented a different aspect of the data and required a unique approach to feature engineering in order to extract the relevant information.

Overall, feature engineering is a critical step in the process of building an ML application as it enables the identification and extraction of relevant information from the data, which in turn improves the performance of the ML models.

4.3.1 Target

The *target* variable was constructed as a binary variable, indicating whether an insurance policy status had lapsed (coded as 0) or not (coded as 1) at the date of analysis (11/07/2022). It represents whether the policy status was active (1) or inactive (0) at the time of analysis. However, the target variable is imbalanced, as only 20.71% of the policyholders had lapsed, while 79.29% of the policyholders had not lapsed at the time of analysis.

Figure 4.3: Target Variable. **Source:** Authors.



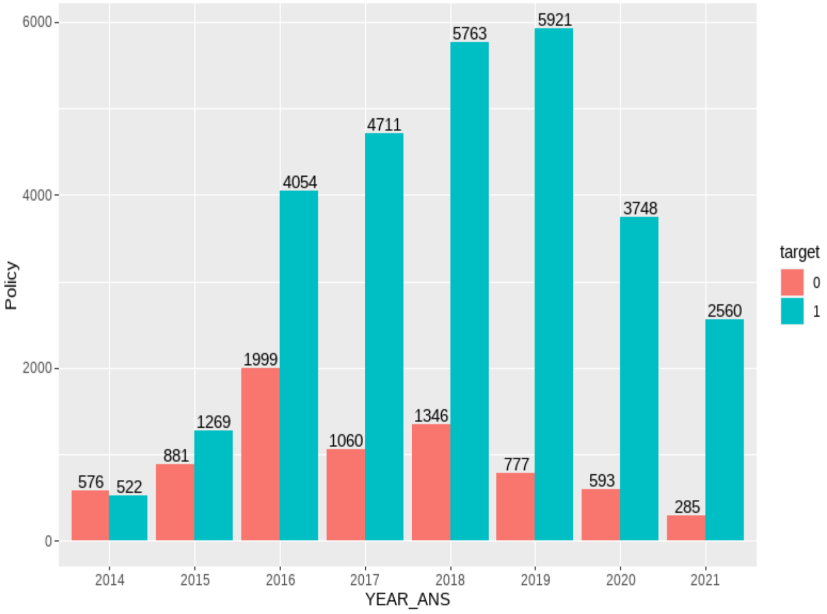
4.3.2 Survey Features

In this section, we discussed the process used to handle the dataset collected. The survey features were a crucial part of the dataset and are represented in Appendix (A). These features consisted of questions that were subdivided into 6 groups and their responses were an important part of the information used to create the model. However, since some of these responses were stored in long textual format, the algorithms were unable to process them directly. To overcome this limitation, was used natural language processing techniques to convert all words into their word origins through stemming, and used binning techniques to categorize the features into groups. This allowed us to extract relevant information from the survey responses and use it effectively in the model.

4.3.2.1 Date of Information ($DATE\_ANS$)

The feature $DATE\_ANS$ identified the date on which each policyholder responded to the survey, representing the initial date of the contract. However, the date format of this feature was $DD/MM/YYYY$. Therefore, to use this feature more effectively, a new feature $YEAR\_ANS$ was created by extracting just the year from the $DATE\_ANS$ feature. For example, from the date 22/02/2014, the year 2014 was extracted and used. Analysis of this feature revealed that the majority of the policyholders answered the survey in the year 2018, and the year with the least number of responses was 2014. This implies that the products of the type of ART (2.1.7.1) were more successful in 2018 and least successful in 2014. As shown in Figure 4.4, policyholders from 2019 had the highest renewal rate, while policyholders from 2016 had the highest lapse rate.

Figure 4.4: YEAR_ANS Variable. **Source:** Authors.



4.3.2.2 Question X101 (A.1.1)

The survey question $X101$ (A.1.1) was answered by 32,519 policyholders. Of those, 31,848 reported not being treated or having altered cholesterol and/or blood pressure values (coded as 0), 352 reported having altered cholesterol (coded as 1), and 319 reported having altered blood pressure values (coded as 2). Additionally, 3,546 policyholders did not respond to this question, which were represented as missing values in the dataset.

Table 4.2: Question X101 answers **Source:** Authors.

Code	Response	Count
		3546
0	NAO	31848
1	SIM COLESTEROL	352
2	SIM TENSAO ARTERIAL	319

4.3.2.3 Question X102 (A.1.2)

The survey question $X102$ (A.1.2) was answered by 392 policyholders, with 375 providing a response. The possible choices for the question were “NAO SEI“, “221-255“, “256-290“, “291-325“, “>345“, and “ATE 220“, which were grouped into two categories. The options “221-255“, “256-290“, “291-325“ were grouped as “>220“ and coded as 2, while “ATE 220“ was grouped as “<=220“ and coded as 1. “NAO SEI“ and non-responses were considered as missing values. 130 policyholders indicated having cholesterol values above 220 mg/dl, while 245 policyholders indicated having cholesterol values below or equal to 220 mg/dl.

Table 4.3: Question X102 (A.1.2) **Source:** Authors.

Code	**Category (mg/dl)**	**Response (mg/dl)**	**Count**	**Total**
			35673	35690
		Não Sei	17	
2	>220	221-255	98	130
		256-290	22	
		291-325	7	
		>345	3	
1	<=220	Até 220	245	245

4.3.2.4 Question X103 (A.1.3)

The survey question $X103$ (A.1.3) was answered by 378 policyholders, with 328 indicating that they have controlled blood pressure values (coded as 1). Conversely, 50 policyholders reported not having controlled blood pressure values (coded as 0). Additionally, 35,687 policyholders did not respond to this question.

Table 4.4: Question X103 (A.1.3) **Source:** Authors.

Code	**Response**	**Count**
		30028
0	NAO	5719
1	SIM	318

4.3.2.5 Question X104 (A.1.4)

In total, 32,433 policyholders answered question $X104$. Out of those, 539 policyholders reported being officially excduty for more than ten days due to an accident (coded as 1), 897 policyholders cited illness (coded as 2) as the reason for leave, and 30,997 policyholders reported never being on leave for more than ten days due to illness or accident (coded as 0). Additionally, 3,632 policyholders did not answer this question, resulting in missing values in the dataset.

Table 4.5: Question X104 (A.1.4) **Source:** Authors.

Code	Response	Count
		3632
0	NAO	30997
1	SIM POR ACIDENTE	539
2	SIM POR DOENCA	897

4.3.2.6 Question X105 (A.1.5)

Regarding question $X105$, 32,396 policyholders provided answers. Of those, 130 policyholders reported being pre-retired or having a process underway for disability (coded as 1), 88 policyholders reported being retired or having a process underway for old-age retirement (coded as 2), 32,178 policyholders reported not being retired (coded as 0), and 3,669 policyholders did not answer this question.

Table 4.6: Question X105 (A.1.5) **Source:** Authors.

Code	Response	Count
		3669
0	NAO	32178
1	SIM POR INVALIDEZ	130
2	SIM POR VELHICE	88

4.3.2.7 Question X106 (A.1.6)

In response to question $X106$, 32,311 policyholders provided answers. Among them, 321 policyholders reported being hospitalized or undergoing surgery due to an accident (coded as 1), 646 policyholders reported being hospitalized or undergoing surgery due to illness (coded as 2), and 1,090 policyholders reported being hospitalized or undergoing surgery due to both illness and accident (coded as 3). Additionally, 30,254 policyholders reported never being hospitalized or undergoing surgery (coded as 0) and 3,754 policyholders did not answer this question.

Table 4.7: Question X106 (A.1.6) **Source:** Authors

Code	Response	Count
		3754
0	NAO	30254
1	SIM HOSPITALIZACAO POR ACIDENTE	321
2	SIM HOSPITALIZACAO POR DOENCA	646
3	SIM INTERVENCAO CIRURGICA POR DOENCA OU ACIDENTE	1090

4.3.2.8 Question X107 (A.1.7)

Question $X107$ was answered by 1144 policyholders, with 12 reporting a surgery for Gastric Band (coded as 7), 7 for Cataracts (coded as 14), 36 for Breast Surgery (coded as 2), 75 for slipped disc (coded as 12), 29 for total or partial Hysterectomy (coded as 8), 49 for Meniscus (coded as 12), 16 for Myopia (coded as 14), 1 for Prostate (coded as 5), 72 for Cysts (coded as 15), 41 for Tendon and Ligament Tears (coded as 12), 49 for Thyroid (coded as 10), and 757 for other surgeries (coded as 1). However, 34,921 policyholders did not answer this question (see Table (B.2)).

According to the insurance company's standards and medical condition list, these options can be grouped into various categories of diseases. A surgery is often used to treat injuries, diseases, and other disorders. Therefore, the category "Diseases of the stomach, inflammation diseases of the intestine, pancreas or other" is related to Gastric Band, the category "Sense-Related Diseases" is related to Cataracts and Myopia, the category "Breast Diseases" is related to Breast Surgery, the category "Osteoarticular, spinal or rheumatological diseases" is related to Slipped Disc, Meniscus, and Ligament Tears, the category "Gynecologic diseases" is related to total or partial Hysterectomy, the category "Genitourinary diseases" is related to Prostate, the category "Diseases related to Tumors or any type of Cancer" is related to Cysts, and finally, the category "Metabolic or blood diseases" is related to Thyroid.

4.3.2.9 Question X108 (A.1.8)

The question $X108$ in the dataset consists of 711 different answers from policyholders regarding surgical interventions they have undergone or are waiting for. To effectively analyse this data, text pre-processing techniques such as tokenization, cleaning, normalization, and bag-of-words representation were applied (figure 2.4.11.4). The answers were first aggregated into a single text corpus, then converted to lowercase and cleaned of Portuguese stop words, punctuation, accentuation, numbers, and other anomalies. The resulting tokens were then stemmed and represented by a numerical feature vector. This vector was then used to group the tokens into categories based on their stemmed form (B.2). For example, the most common stemmed token, "apendic" represents the original word "apendice" (appendix) and the least common token, "cervical," represents the original word "cervical".

These original words could be grouped into specific diseases according to the insurance company's medical condition list. In total, the question X108 is represented in the dataset by 756 surgical interventions per policyholder, where 14 policyholders have undergone or are waiting for interventions related to "Breast Diseases" (encoded by 2), 9 policyholders related to "Diseases of the Cardiovascular system" (encoded by 4), 30 policyholders related to "Genitourinary diseases" (encoded by 5), and so on. Notably, 35,309 policyholders did not answer this question (table 4.8).

Table 4.8: Question X108 Transformed (A.1.8). **Source:** Authors.

Code	Category	Count
		35309
2	DOENCAS DA MAMA	14
4	DOENCAS DO APARELHO CARDIOVASCULAR	9
5	DOENCAS DO APARELHO GENITO-URINARIO	30
6	DOENCAS DO APARELHO RESPIRATORIO	2
7	DOENCAS DO ESTOMAGO, DOENÇAS INFLAMATORIAS DO INTESTINO, DO PANCREAS OU OUTRAS	64
9	DOENCAS INFECIOSAS	2
10	DOENCAS METABOLICAS OU DO SANGUE	3
11	DOENCAS NEUROLOGICAS	2
12	DOENCAS OSTEOARTICULARES, DA COLUNA VERTEBRAL OU REUMATOLOGICAS	376
14	DOENCAS RELACIONADAS COM OS SENTIDOS	42
15	DOENCAS RELACIONADAS COM TUMORES OU QUALQUER TIPO DE CANCRO	13
16	DOENCAS VASCULARES	44
1	OUTRAS	105
8	DOENCAS GINECOLOGICAS	50

4.3.2.10 Question X109 (A.1.9)

The question $X109$ was answered by 32,252 policyholders, with 167 policyholders reporting they have undergone or are awaiting results for a Biopsy exam (encoded as 2), 158 policyholders reporting they have undergone echocardiograms (encoded as 3), 229 policyholders reporting they have undergone Magnetic Resonance Imaging (MRI) (encoded as 4), 350 policyholders reporting they have undergone tomography scans (encoded as 5), and 202 policyholders reporting they have undergone other laboratory tests or medical examinations (encoded as 1). Additionally, 31,146 policyholders reported they have not undergone any of these exams, and 3,813 policyholders did not answer the question (table 4.9).

Table 4.9: Question X109 (A.1.9). **Source:** Authors.

Code	Response	Response pre-processed	Count
			3813
2	Biopsias	BIOPSIAS	167
3	Ecocardiogramas	ECOCARDIOGRAMAS	158
0	Não	NAO	31146
1	Outras	OUTRAS	202
4	Ressonância Magnética Ressonância magnética Ressoní¢ncia Magnética Ressoní¢ncia magnética	RESSONANCIA MAGNETICA	229
5	TAC	TAC	350

4.3.2.11 Question X110 (A.1.10)

Question $X110$ was answered differently by 274 policyholders, with the different responses being encoded as 1.

Table 4.10: Question X110 (A.1.10) **Source:**Authors.

Code	Count
	35791
1	274

4.3.2.12 Question X111 (A.1.11)

The question $X111$ was answered by 33,006 policyholders with their weight, with 3,059 policyholders not answering. The maximum weight reported in the survey was 300 kg, and the minimum was 10 kg. There were inconsistencies in some of the responses, such as “054“ and “075“ which were corrected to “54“ and “75“ respectively. Additionally, there were also weights between 10 and 15 kg reported by four policyholders between the ages of 39-55. The average weight reported was 71.22 kg, with a median of 70 kg and a mode of 70 kg. The weight data can be plotted in the histogram (B.3).

Table 4.11: Question X111 (A.1.11) **Source:** Authors.

	Response
MAX	300
MIN	10

4.3.2.13 Question X112 (A.1.12)

The question $X112$ was answered by 33,017 policyholders with their height, and 3,048 policyholders did not answer. The maximum height reported was 202 cm and the minimum was 100 cm (table (4.12)). Inconsistencies were detected in some responses, such as “0170“ and “50“ which were corrected to “170“ and “150“ respectively. The average height is 169 cm, the median is 170 cm, and the mode is 170 cm. Each height per policy can be plotted in a histogram. The height data is partitioned into 3 subsets using the k-means method. Heights between 100 cm and 165 cm are encoded as 1, heights above 165 cm and up to 173 cm are encoded as 2, and heights above 173 cm and up to 202 cm are encoded as 3 (figure 4.5). The height data can be plotted in the histogram (B.4).

Figure 4.5: Question X112 Grouped (A.1.12) **Source:** Authors.

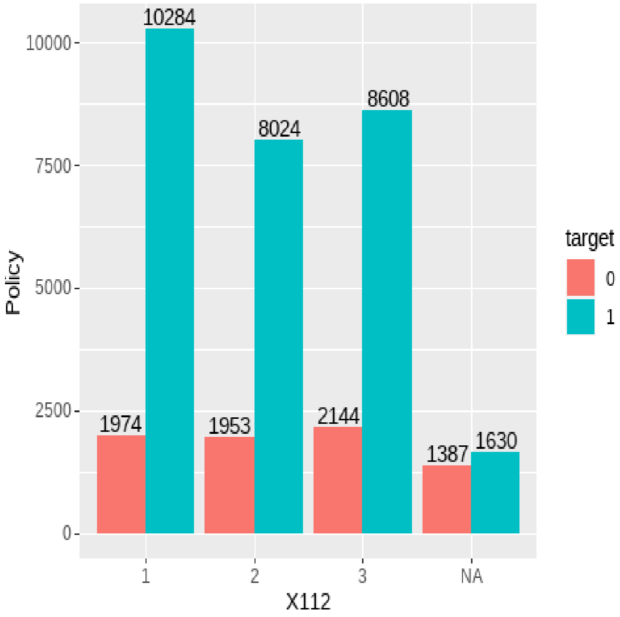


Table 4.12: Question X112 (A.1.12) **Source:** Authors.

	Response
MAX	202
MIN	100

4.3.2.14 Question X201 (A.2.1)

The question $X201$ aimed to identify the diseases or disorders of policyholders by considering the insurance company's medical condition list. Out of 32,177 policyholders, 3,888 did not answer and the remaining policyholders reported various diseases including metabolic or blood diseases, neurological diseases, osteoarticular, spinal or rheumatological diseases, psychiatric diseases, diseases related to the senses, tumors or cancer, vascular diseases, breast diseases, skin diseases, cardiovascular diseases, genitourinary diseases, respiratory diseases, stomach, inflammation, gynecological diseases and infectious diseases. Additionally, some policyholders reported other diseases that were not included in the list of options (figure 4.28).

Table 4.13: Question X201 (A.2.1). **Source:** Authors.

Code	Response	Count
		3888
0	NENHUMA DAS DOENCAS INDICADAS	30206
1	OUTRAS DOENCAS	318
10	DOENCAS METABOLICAS OU DO SANGUE	53
11	DOENCAS NEUROLOGICAS	54
12	DOENCAS OSTEOARTICULARES DA COLUNA VERTEBRAL OU REUMATOLOGICAS	207
13	DOENCAS PSIQUIATRICAS	112
14	DOENCAS RELACIONADAS COM OS SENTIDOS CONSIDERAR APENAS OUVIDOS OU OLHOS	298
15	DOENCAS RELACIONADAS COM TUMORES OU QUALQUER TIPO DE CANCRO	109
16	DOENCAS VASCULARES	70
2	DOENCAS DA MAMA	21
3	DOENCAS DA PELE	102
4	DOENCAS DO APARELHO CARDIOVASCULAR	118
5	DOENCAS DO APARELHO GENITO URINARIO	56
6	DOENCAS DO APARELHO RESPIRATORIO	282
7	DOENCAS DO ESTOMAGO DOENCAS INFLAMATORIAS DO INTESTINO DO PANCREAS OU OUTRAS	127
8	DOENCAS GINECOLOGICAS	41
9	DOENCAS INFECCIOSAS	3

4.3.2.15 Question X202 (A.2.2)

The question $X202$ aimed to identify specific cardiovascular diseases among policyholders. Out of 132 policyholders, 72 reported suffering from specific options of cardiovascular diseases and 60 reported suffering from other types of cardiovascular diseases. The question was not answered by 35,933 policyholders (table 4.14).

Table 4.14: Question X202 (A.2.2)

Code	Response	Count
		35933
4	ANGINA DE PEITO	5
4	ANGIOPLASTIA	3
4	ARRITMIAS	25
4	ENFARTE DO MIOCARDIO	16
4	OUTRAS	60
4	PACEMAKER	9
4	PROLAPSO DA VALVULA MITRAL	14

4.3.2.16 Question X203 (A.2.3)

The question $X203$ aimed to identify the prevalence of osteoarticular, spinal or rheumatological diseases among policyholders. Out of 249 policyholders who responded, 154 reported suffering from one of the seven options provided, while 95 reported suffering from other types of these diseases. The options were encoded with the number 12, representing “Osteoarticular, spinal or rheumatological diseases“. The question was not answered by 35,816 policyholders (table 4.15).

Table 4.15: Question X203 (A.2.3)

Code	Response	Count
		35816
12	ARTRITE REUMATOIDE	23
12	CIATICA	8
12	FIBROMIALGIA	9
12	HERNIA DISCAL	91
12	LOMBALGIA	15
12	LUPUS	2
12	OUTRAS	95
12	TRAUMATISMOS	6

4.3.2.17 Question X204 (A.2.4)

The question $X204$ aimed to identify the prevalence of respiratory diseases among policyholders. Out of 243 policyholders who responded, 149 reported suffering from one of the four options provided, while 94 reported suffering from other types of respiratory diseases. The options were encoded with the number 6, except for “Asthma“ which was encoded with 6.1, representing “Diseases of the Pulmonary System“. The question was not answered by 35,728 policyholders.

Table 4.16: Question X204 (A.2.4)

Code	Response	Count
		35728
6	BRONQUITE CRONICA	25
6	OUTRAS	94
6	PNEUMONIA	11
6	TUBERCULOSE	5
6.1	ASMA	202

4.3.2.18 Question X205 (A.2.5)

The question $X205$ aimed to identify the prevalence of genitourinary diseases among policyholders. Out of 72 policyholders who responded, 54 reported suffering from one of the seven options provided, while 18 reported suffering from other types of genitourinary diseases. The options were encoded with the number 5, representing “Diseases of the genitourinary system“. The question was not answered by 35,993 policyholders.

Table 4.17: Question X205 (A.2.5)

Code	Response	Count
		35993
5	BEXIGA	9
5	CALCULO RENAL	6
5	MAMA	1
5	OUTRAS	18
5	OUTRAS DOENCAS RENAIS	6
5	PROSTATA	8
5	TRANSPLANTE RENAL	4
5	UTERO	20

4.3.2.19 Question X206 (A.2.6)

The question $X206$ aimed to identify the prevalence of psychiatric diseases among policyholders. Out of 140 policyholders who responded, 131 reported suffering from one of the four options provided, while 9 reported suffering from other types of psychiatric diseases. The options were encoded with the number 13, representing "Psychiatric Diseases". The question was not answered by 35,916 policyholders.

Table 4.18: Question X206 (A.2.6)

Code	Response	Count
		35916
13	ANSIEDADE	46
13	DEPRESSAO	90
13	OUTRAS	9
13	PSICOSES	3
13	TENTATIVA DE SUICIDIO	1

4.3.2.20 Question X207 (A.2.7)

The question $X207$ aimed to identify the prevalence of neurological diseases among policyholders. Out of 73 policyholders who responded, 47 reported suffering from one of the six options provided, while 26 reported suffering from other types of neurological diseases. The options were encoded with the number 11, representing "neurological Diseases". The question was not answered by 35,992 policyholders.

Table 4.19: Question X207 (A.2.7)

Code	Response	Count
		35992
11	ACIDENTE ISQUEMICO TRANSITORIO	5
11	AVC	5
11	DOENCAS MUSCULARES	5
11	EPILEPSIA	21
11	ESCLEROSE MULTIPLA	9
11	OUTRAS	26
11	PARALISIAS	2

4.3.2.21 Question X208 (A.2.8)

The question *X*208 aimed to identify the prevalence of vascular diseases among policyholders. Out of 99 policyholders who responded, 86 reported suffering from one of the four options provided, while 13 reported suffering from other types of vascular diseases. The options were encoded with the number 16, representing "vascular diseases". The question was not answered by 35,966 policyholders.

Table 4.20: Question X208 (A.2.8)

Code	Response	Count
		35966
16	CLAUDICACAO	1
16	EMBOLIA	4
16	OUTRAS	13
16	TROMBOSE	13
16	VARIZES	68

4.3.2.22 Question X209 (A.2.9)

The question *X*209 aimed to identify the prevalence of infectious diseases among policyholders. Out of 9 policyholders who responded, 5 reported suffering from one of the two options provided, while 4 reported suffering from other types of infectious diseases. The options were encoded with the number 9, representing "Infectious Diseases". The question was not answered by 35,056 policyholders.

Table 4.21: Question X209 (A.2.9)

Code	Response	Count
		36056
9	HEPATITES	2
9	OUTRAS	4
9	VIRUS HIV	3

4.3.2.23 Question X210 (A.2.10)

The question $X210$ aimed to identify the prevalence of metabolic or blood diseases among policyholders. Out of 70 policyholders who responded, 55 reported suffering from one of the two options provided, while 15 reported suffering from other types of metabolic or blood diseases. The options were encoded with the number 10, except for "Diabetes" which was encoded with 10.1, representing "Metabolic or Blood Diseases". The question was not answered by 35,995 policyholders.

Table 4.22: Question X210 (A.2.10)

Code	Response	Count
		35995
10	ANEMIA	21
10	LINFOMAS	1
10	OUTRAS	15
10	TIROIDE COM CIRURGIA	2
10	TIROIDE SEM CIRURGIA	9
10.1	DIABETES	22

4.3.2.24 Question X211 (A.2.11)

The question $X211$ aimed to identify the prevalence of specific diseases of stomach, inflammatory diseases of the intestine, pancreas or other among policyholders. Out of 156 policyholders who responded, 143 reported suffering from one of the three options provided, while 13 reported suffering from other types of diseases of stomach, inflammatory diseases of the intestine, pancreas or other. The options were encoded with the number 7, representing "Diseases of the stomach, inflammation diseases of the intestine, pancreas or other". The question was not answered by 35,909 policyholders.

Table 4.23: Question X211 (A.2.11)

Code	Response	Count
		35909
7	ESTOMAGO	102
7	INTESTINO	28
7	OUTRAS	13
7	PANCREAS	13

4.3.2.25 Question X212 (A.2.12)

The question $X212$ was answered by 331 policyholders, with 296 of them reporting suffering from one of the three options of sense-related diseases, such as disorders related to eyes and ears. The remaining 35 policyholders reported suffering from a different type of sense-related disease. Each option was assigned a code of 14, representing "Sense-Related Diseases." The question went unanswered by 35,734 policyholders.

Table 4.24: Question X212 (A.2.12)

Code	Response	Count
		35734
14	OLHOS ASTIGMATISMO	66
14	OLHOS CATARATAS	8
14	OLHOS MIOPIA	183
14	OUTRAS	35
14	OUVIDOS	39

4.3.2.26 Question X213 (A.2.13)

For question $X213$, 196 policyholders provided different responses, which were all encoded with a code of 14 representing "Sense-Related Diseases.

Table 4.25: Question X213 (A.2.13)

Code	Response	Count
		35869
14	SIM	196

4.3.2.27 Question X214 (A.2.14)

Question $X214$ was answered by 121 policyholders, with 63 of them reporting suffering from one of the three specific options of skin diseases, and 58 policyholders reporting suffering from a different type of skin disease. Each option was assigned a code of 3, representing “Skin Diseases.“ The question went unanswered by 35,944 policyholders.

Table 4.26: Question X214 (A.2.14)

Code	Response	Count
		35944
3	ARTRITE PSORIATICA	5
3	CANCRO DA PELE	3
3	OUTRAS	58
3	PSORIASE	55

4.3.2.28 Question X215 (A.2.15)

Question $X215$ was answered by 122 policyholders, with all of them reporting suffering from one of the two types of tumours or cancer, benign or malignant. Each option was assigned a code of 15, representing “Diseases related to Tumors or any type of Cancer.“ The question went unanswered by 35,943 policyholders.

Table 4.27: Question X215 (A.2.15)

Code	Response	Count
		35943
15	BENIGNO	46
15	MALIGNO	76

4.3.2.29 Question X216 (A.2.16)

The question $X216$ consists of 277 different answers. To ensure the accuracy of this question, the same methodology and treatment as question $X108$ was used. A “bag of words“ visualization was created to show the most frequently used stemmed words in the answers, with “diabet“ being the most common and “autoimun“ being the least common. These stemmed words represent different diseases, such as “diabet“ representing diabetes and “tiroid“ representing thyroid (figure B.5).

Using this methodology, the final groups of diseases were as follows: 344 diseases per client, with 35 policyholders suffering from “Metabolic or blood diseases“ (encoded by 10), 51 policyholders suffering from “Diabetes“ (encoded by 10.1), 4 policyholders suffering from “Neurological Diseases“ (encoded by 11), 108 policyholders suffering from “Osteoarticular, spinal or rheumatological diseases“ (encoded by 12), 3 policyholders suffering from “Psychiatric diseases“ (encoded by 13), 27 policyholders suffering from “Sense-Related Diseases“ (encoded by 14), 14 policyholders suffering from “Diseases related to Tumors or any type of Cancer“ (encoded by 15), 3 policyholders suffering from “Vascular Diseases“ (encoded by 16), 2 policyholders suffering from

"Cholesterol" (encoded by 17), 4 policyholders suffering from "Breast diseases" (encoded by 2), 1 client related to 'Skin Diseases' (encoded by 3), 5 policyholders related to 'Diseases of the Cardiovascular system' (encoded by 4), 15 policyholders related to 'Diseases of the genitourinary system' (encoded by 5), 5 policyholders related to 'Respiratory Diseases' (encoded by 6), 3 policyholders related to 'Asthma' (encoded by 6.1), 14 policyholders related to 'Diseases of the stomach, inflammation diseases of the intestine, pancreas or other' (encoded by 7), 3 policyholders related to 'Gynecologic diseases' (encoded by 8), 11 policyholders related to 'Infectious diseases' (encoded by 9), and 36 policyholders have suffered some disease (encoded by 1). About 35 721 policyholders did not answer this question (figure 4.28).

Overall, the question $X216$ provides a detailed and comprehensive understanding of the different diseases suffered by the policyholders, and the use of the stemmed tokens and encoding system allows for easy aggregation and categorization of the data for further analysis.

Table 4.28: Question X216 (A.2.16). **Source:** Authors.

Code	Response	Count
		35721
1	OUTRAS	36
10	DOENCAS METABOLICAS OU DO SANGUE	35
10.1	DIABETES	51
11	DOENCAS NEUROLOGICAS	4
12	DOENCAS OSTEOARTICULARES, DA COLUNA VERTEBRAL OU REUMATOLOGICAS	108
13	DOENCAS PSIQUIATRICAS	3
14	DOENCAS RELACIONADAS COM OS SENTIDOS	27
15	DOENCAS RELACIONADAS COM TUMORES OU QUALQUER TIPO DE CANCRO	14
16	DOENCAS VASCULARES	3
17	COLESTEROL	2
2	DOENCAS DA MAMA	4
3	DOENCAS DA PELE	1
4	DOENCAS DO APARELHO CARDIOVASCULAR	5
5	DOENCAS DO APARELHO GENITO-URINARIO	15
6	DOENCAS DO APARELHO RESPIRATORIO	5
6.1	ASMA	3
7	DOENCAS DO ESTOMAGO, DOENÇAS INFLAMATORIAS DO INTESTINO, DO PANCREAS OU OUTRAS	14
8	DOENCAS GINECOLOGICAS	3
9	DOENCAS INFECIOSAS	11

4.3.2.30 Question X217 (A.2.17)

The question $X217$ was answered by 27 policyholders, where 21 policyholders specifically reported suffering from breast diseases. 6 policyholders reported suffering from other types of breast-related illnesses. Each option was encoded using the code 2, representing "Breast Diseases." A total of 36,038 policyholders did not answer this question.

Table 4.29: Question X217 (A.2.17)

Code	Response	Count
		36038
2	DOENCA FIBROQUISTICA	2
2	NODULOS	19
2	OUTRAS	6

4.3.2.31 Question X218 (A.2.18)

The question $X218$ was answered by 46 policyholders, with 34 policyholders reporting suffering from specific gynecological diseases. 12 policyholders reported suffering from other types of gynecological illnesses. Each option was encoded using the code 8, representing “Gynecological Diseases.“ A total of 36,019 policyholders did not answer this question.

Table 4.30: Question X218 (A.2.18)

Code	Response	Count
		36019
8	ALTERACOES DA CITOLOGIA DO COLO	4
8	OUTRAS	12
8	OVARIO	14
8	UTERO	16

4.3.2.32 Question X301 (A.3.1)

Medicines are commonly used to treat or prevent diseases, and are therefore closely related to them. In line with insurance company standards, some medicines are grouped into specific disease codes. The results of Question $X301$, answered by 32,106 policyholders, revealed that 30,346 of them did not regularly take any medication (coded as 0), 52 policyholders regularly took or had taken anticoagulants for metabolic or blood diseases (coded as 10), 42 policyholders took insulin for diabetes (coded as 10.1), 281 policyholders took antidepressants or tranquilizers for psychiatric diseases, 28 policyholders took medication for tumors or cancer (coded as 15), 495 policyholders took medication for heart and hypertension (coded as 4), and the remaining 862 policyholders were grouped as “other regular medications“ (coded as 1). Additionally, 3,959 policyholders did not answer the question.

Table 4.31: Question X301 (A.3.1)

Code	Response	Count
		3959
0	NENHUM	30346
1	OUTROS MEDICAMENTOS REGULARMENTE	798
10	ANTICOAGULANTES	52
10.1	INSULINA	42
13	ANTIDEPRESSIVOS OU TRANQUILIZANTES	281
15	PARA TUMORES OU CANCRO	28
4	PARA O CORACAO OU HIPERTENSAO	495
1	CORTICOIDES	64

4.3.2.33 Question X302 (A.3.2)

Question $X302$ was answered differently by 519 policyholders, with these varying responses encoded as 4, indicating “Diseases of the Cardiovascular system“.

Table 4.32: Question X302 (A.3.2)

Code	Response	Count
		35546
4	Sim	519

4.3.2.34 Question X303 (A.3.3)

Question $X303$ was answered differently by 68 policyholders, with these varying responses encoded as 10, indicating “Metabolic or blood diseases“.

Table 4.33: Question X303 (A.3.3)

Code	Response	Count
		35997
10	Sim	68

4.3.2.35 Question X304 (A.3.4)

Question $X304$ was answered differently by 309 policyholders, with these varying responses encoded as 13, indicating “Psychiatric Diseases“.

Table 4.34: Question X304 (A.3.4)

Code	Response	Count
		35753
13	Sim	309

4.3.2.36 Question X305 (A.3.5)

Question $X305$ was answered differently by 75 policyholders, with these varying responses encoded as 1.

Table 4.35: Question X305 (A.3.5)

Code	Response	Count
		35989
1	Sim	75

4.3.2.37 Question X306 (A.3.6)

Question $X306$ was answered differently by 867 policyholders, with these varying responses encoded as 1.

Table 4.36: Question X306 (A.3.6)

Code	Response	Count
		35197
1	Sim	867

4.3.2.38 Question X307 (A.3.7)

Question $X307$ was answered by 32,049 policyholders. Where 31,969 policyholders responded that they had never undergone any treatment options, 10 policyholders had undergone Detoxification (coded as 1), 38 policyholders underwent Chemotherapy, 32 policyholders underwent Radiotherapy. Additionally, 4016 policyholders did not answer this question. Both Chemotherapy and Radiotherapy are types of cancer treatment, so they were both encoded as 15, indicating that are related to "Diseases related to Tumors or any type of Cancer".

Table 4.37: Question X307 (A.3.7)

Code	Response	Count
		4016
0	NENHUM DOS TRATAMENTOS INDICADOS	31969
1	DESINTOXICACAO	10
15	QUIMIOTERAPIA	38
15	RADIOTERAPIA	32

4.3.2.39 Question X401 (A.4.1)

Question $X401$ was answered by 32,092 policyholders, who responded with one of the following options: “NAO“, “SIM ATE 25 CIGARROS“, “SIM ENTRE 26 E 30 CIGARROS“, “SIM ENTRE 31 E 35 CIGARROS“, “SIM ENTRE 36 E 40 CIGARROS“, and “SIM MAIS DE 40 CIGARROS“. These options were grouped into 3 groups. The options “SIM ENTRE 26 E 30 CIGARROS“, “SIM ENTRE 31 E 35 CIGARROS“, “SIM ENTRE 36 E 40 CIGARROS“, and “SIM MAIS DE 40 CIGARROS“ were grouped together as “> 25 CIGARROS“ and coded as 2, while the option “SIM ATE 25 CIGARROS“ was grouped as “<= 25 CIGARROS“ and coded as 1. As a result, 5,171 policyholders indicated that they are smokers, of which 5,135 smoke up to 25 cigarettes per day, while 36 smoke more than 25 cigarettes per day. On the other hand, 26,921 policyholders reported not being smokers, and 3,973 policyholders did not answer this question.

Table 4.38: Question X401 (A.4.1)

Code	Category	Response	Count
			3973
1	<= 25 CIGARROS	SIM ATE 25 CIGARROS	5135
2	> 25 CIGARROS	SIM ENTRE 26 E 30 CIGARROS	26
		SIM ENTRE 31 E 35 CIGARROS	2
		SIM ENTRE 36 E 40 CIGARROS	3
		SIM MAIS DE 40 CIGARROS	5
0	NÃO	NAO	26921

4.3.2.40 Question X402 (A.4.2)

Question $X402$ was answered by 32,007 policyholders, of which 217 policyholders reported consuming more than 15 units of alcohol per week (coded as 1), and 31,790 policyholders reported not consuming alcohol (coded as 0). About 4,058 policyholders did not answer this question.

Table 4.39: Question X402 (A.4.2)

Code	Response	Count
		4058
0	NAO	31790
1	SIM	217

4.3.2.41 Question X403 (A.4.3)

Question $X403$ was answered by 32,000 policyholders, of which 48 policyholders reported consuming or having consumed narcotics or drugs (coded as 1), and 31,952 policyholders reported not consuming them (coded as 0). About 4,065 policyholders did not answer this question.

Table 4.40: Question X403 (A.4.3)

Code	Response	Count
		4065
0	NAO	31952
1	SIM	48

4.3.2.42 Question X501 (A.5.1)

Question $X501$ was answered by 32,027 policyholders, of which 1,570 policyholders reported engaging in risky sports in an amateur way (coded as 2), 40 policyholders reported engaging in risky sports professionally (coded as 1), and 30,417 policyholders reported not engaging in any risky sports (coded as 0). About 4,038 policyholders did not answer this question.

Table 4.41: Question X501 (A.5.1)

Code	Response	Count
		4038
0	NAO	30417
2	SIM DE FORMA AMADORA	1570
1	SIM DE FORMA PROFISSIONAL	40

4.3.2.43 Question X502 (A.5.2)

The question $X502$ was answered by 1585 policyholders, where 182 policyholders reported participating in amateur risk sports (encoded as 1) and 1403 policyholders reported not participating in any of the listed options (encoded as 0). This question was not answered by 34,480 policyholders (table B.3).

4.3.2.44 Question X503 (A.5.3)

The question $X503$ was answered by 42 policyholders, where 25 policyholders reported participating in professional risk sports (encoded as 1) and 17 policyholders reported not participating in any of the listed options (encoded as 0). This question was not answered by 36,023 policyholders (table B.4).

4.3.2.45 Question X504 (A.5.4)

The question $X504$ was answered by 25,950 policyholders, where 158 policyholders reported plans to travel outside the European Union (encoded as 1) and 25,792 policyholders reported not having such plans (encoded as 0). About 10,115 policyholders did not answer this question.

Table 4.42: Question X504 (A.5.4)

Code	Response	Count
		10115
0	NAO	25792
1	SIM	158

4.3.2.46 Question X505 (A.5.5)

The question $X505$ was answered by 163 policyholders, where 74 policyholders reported plans to pursue predominantly manual labor (encoded as 1) and 89 policyholders reported plans to pursue predominantly non-manual labor (encoded as 0). About 35,902 policyholders did not answer this question.

Table 4.43: Question X505 (A.5.5)

Code	Response	Count
		35902
1	ATIVIDADE PREDOMINANTEMENTE MANUAL	74
0	ATIVIDADE PREDOMINANTEMENTE NAO MANUAL	89

4.3.2.47 Question X506 (A.5.6)

The question $X506$ was answered by 163 policyholders who reported plans to relocate to another country. Each country is represented in the data by ISO 3166-1 alpha-3 codes. The option “OUTRO PAIS“ represents other countries and is encoded as 1. About 35,902 policyholders did not answer this question (table B.5).

4.3.2.48 Question X507 (A.5.7)

The question $X507$ was answered by 30,028 policyholders, where 318 policyholders reported plans to travel to countries outside of the European Union (encoded as 1) and 5,719 policyholders reported not having such plans (encoded as 0). About 30,028 policyholders did not answer this question.

Table 4.44: Question X507 (A.5.7)

Code	Response	Count
		30028
0	NAO	5719
1	SIM	318

4.3.2.49 Question X6011 (A.6.1)

The question $X6011$ was answered by 38 policyholders, where all of them responded negatively (encoded by 0). The remaining 36,027 policyholders did not provide an answer to this question.

Table 4.45: Question X6011 (A.6.1)

Code	Response	Count
		36027
0	Não	38

4.3.3 Policy Features

4.3.3.1 Coverage ID

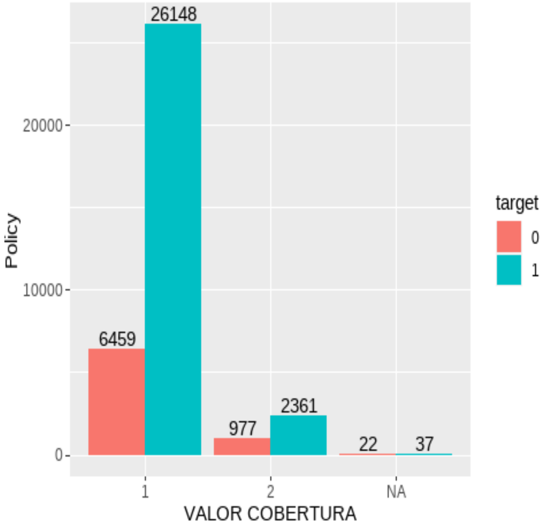
The $COBERTURAID$ feature identifies the coverage associated with a policy. In this study, the insurer assumes the risk of death for the policyholder. As a result, all policies include death coverage (COBERTURAID = 0).

4.3.3.2 Coverage Amount

The policy is limited by a maximum amount that an insurer will pay for a covered loss. This feature $VALORCOBERTURA$ refers to the amount of money that the insurer agrees to pay to the policy beneficiary in the event of the policyholder's death. This amount is based on the policyholder's age, health, and coverage needs. The coverage amount is in a range between 5 to 1 million euros, where the amounts of 5 to 62 euros correspond to the policies of 48 policyholders with less than fifteen years old and product nine. The other amounts are between ten thousand until one million euros.

The biggest concentration of policies have a coverage amount between 10,000 to 250,000 euros. About 48 policies don´t have this amount filled. This feature is partitioned into 2 groups with the application of the k-means method. The category of coverage amounts between 10 000 euros to less than 83,600 euros ("[10000,83600)") are encoded as 1, and coverage amounts between 83,600 euros until 1,000,000 euros ("[83600,1000000]") are encoded as 2 (figure 4.6).

Figure 4.6: Coverage Amount Grouped. **Source:** Authors.



4.3.3.3 Aggravation Motive 1-8

An aggravation motive refers to a medical condition or circumstance that increases the likelihood of a policyholder experiencing a loss or making a claim on their insurance policy.

In this research, there are eight features that represent aggravation motives, named $MOTIVOAGRAVAMENTO\_1$ to $MOTIVOAGRAVAMENTO\_8$. Each feature corresponds to a unique code that represents an aggravation motive associated with the policyholder's historical information. The features are ordered from $MOTIVOAGRA VAMENTO\_1$ as the first aggravation motive to $MOTIVOAGRAVAMENTO\_8$ as the eighth. For example, if a policyholder has only one aggravation motive, only $MOTIVOAGRAVAMENTO\_1$ will be filled. If $MOTIVOAGRAVAMENTO\_1$ is not filled, it means the policyholder has never had an aggravation motive. For this reason, If these features were not filled, they were encoded as 0.

In this study, the policyholders' aggravation motives are recorded in Table (B.6). 3788 policyholders had an aggravation to the initial premium, with the most common first or unique aggravation motives being "Weight/Height Ratio" and "Professional" (see figure B.6). 254 policyholders had a second aggravation to the initial premium (see figure B.7), 45 had a third aggravation (see figure B.8), 12 had a fourth aggravation and

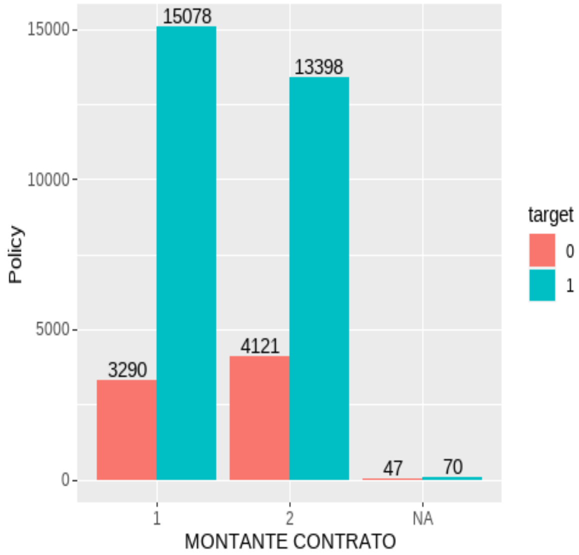
2 had up to eighth aggravation. From the third aggravation motive, the policyholders did not lapse (target=1). 32,277 policyholders never had an aggravation to the initial premium.

4.3.3.4 Contract Accumulation Amount (MONTANTECONTRATO)

The feature *MONTANTECONTRATO* refers to the total of premium payments that the policyholder has paid over the life of the policy and is used to determine the total savings or the total value of the policy. The sample's contract accumulation amount ranges from 10,000 to 1,100,000 euros. The biggest concentration of observations in the study is in a range of 10,000 to 300,000 euros.

This feature of contract amounts is categorised into 2 groups with the application of the k-means method. The group of contract amounts between 10,000 euros to 40,000 euros ("[10000,40000]“) are encoded as 1, and coverage amounts superior to 40,000 euros until 1,100,000 euros are encoded as 2 (see figure 4.7).

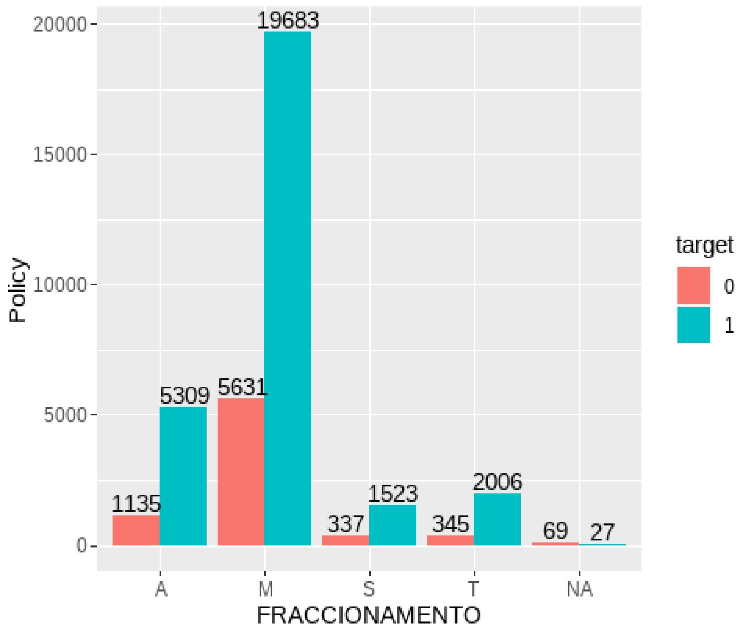
Figure 4.7: Contract Accumulation Amount. **Source:** Authors.



4.3.3.5 Premium Frequency (FRACCIONAMENTO)

The feature *FRACCIONAMENTO* characterizes the premium's payment frequency of the policy. This defines the number of times that a policyholder pays premiums during the policy year, where 'M' means monthly, 'A' means annually, 'S' means semi-annually, and 'T' means quarterly. This feature has 96 policies without this feature filled.

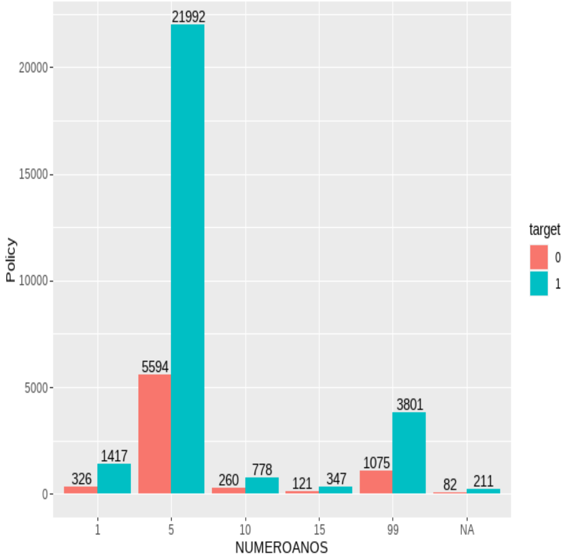
Figure 4.8: Premium Frequency. **Source:** Authors.



4.3.3.6 Policy Term (NUMEROANOS)

The feature *NUMEROANOS* identifies the initial term of the policy agreed between parts. It refers to the lifetime of an insurance policy. This feature have the terms of 1, 5, 10, 15, or 99 years. The most regular policy term is 5 years and the least one is 10 years. The feature is not filled to 96 policies.

Figure 4.9: Policy Term. **Source:** Authors.



4.3.4 Policyholder Features

4.3.4.1 Date of Birth

The *DT_BIRTH* feature represents the birth date of the policyholder. The range of dates in the dataset goes from January 20th, 1930 to September 24th, 2003. However, it should be noted that there are 49 policyholders with missing birth date information.

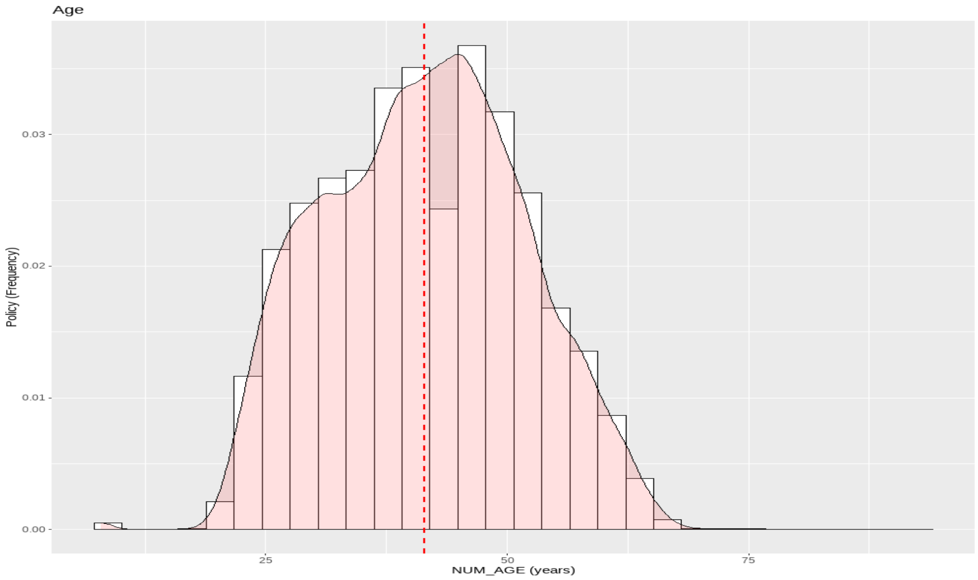
4.3.4.2 Date of Death

The *DT_DEATH* feature represents the death date of the policyholder. If a policyholder has passed away, the date of death is recorded in the dataset. The data is encoded as 1 if the policyholder is deceased and 0 if they are not (see figure (B.9)). In this research, only policyholders who are still alive are considered for modeling purposes. Therefore, the feature "DT_DEATH" and the 24 policyholders who have passed away are removed from the dataset.

4.3.4.3 Age

The *NUM_AGE* feature represents the age of the policyholder. The data shows that there are 50 policyholders who are under the age of 18. Despite this, the youngest policyholder in the dataset is 18 years old and the oldest is 92 years old.

Figure 4.10: Age Distribution. **Source:** Authors.

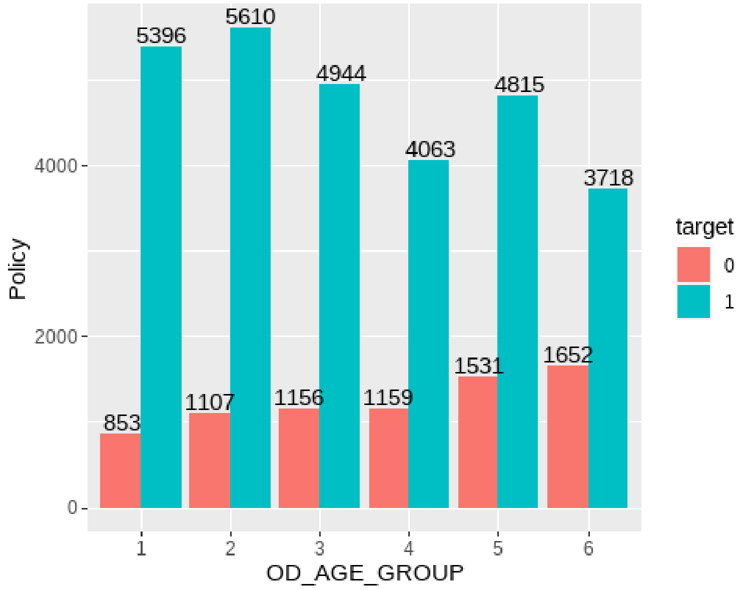


4.3.4.4 Age Group

The feature *OD_AGE_GROUP* in this study represents the age group of policyholders, which is a crucial element in understanding and analyzing the data. The original data from the database was divided into several categories based on age: policyholders less than 15 years old were categorized as “Menos 15 anos“, policyholders between 15 to 24 years old were categorized as “15/24 anos“, policyholders between 25 to 34 years old were categorized as “25/34 anos“, policyholders between 35 to 44 years old were categorized as “35/44 anos“, policyholders between 45 to 54 years old were categorized as “45/54 anos“, policyholders between 55 to 64 years old were categorized as “55/64 anos“, and policyholders older than 65 years old were categorized as “Mais 65 anos“. However, it was noticed that policyholders younger than 15 and older than 73 were not representative of the entire population, and thus were not included in this study, as mentioned in the section on anomalies (see figure B.10).

To further improve the data, the age group feature was recategorized into 6 groups using the unsupervised binning technique k-means. This technique enables to group similar observations together and identify patterns in the data by using clustering algorithms. The new categories are: policyholders between 18 to 30 years old (“[18,30]”) encoded as 1, policyholders between 30 to 37 years old (“(30,37]“) encoded as 2, policyholders between 37 to 42 years old “(37,42]“ encoded as 3, policyholders between 42 to 46 years old (“(42,46]“) encoded as 4, policyholders between 46 to 52 years old (“(46,52]“) encoded as 5, and policyholders between 52 to 73 years old encoded as 6 (figure 4.11).

Figure 4.11: Age Group. **Source:** Authors.



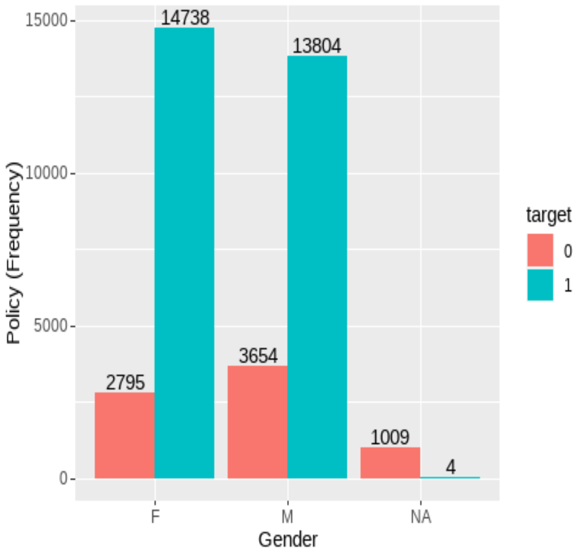
As a result of this recategorization, the feature *NUM_AGE* was not used in the modeling process, since the new feature *OD_AGE_GROUP* provides more informative and accurate data for analysis. The recategorization of the age group feature has improved the data and enabled more accurate analysis and modeling of the policyholder data.

4.3.4.5 Gender

The feature *OD_GENDER* is used to classify the gender of policyholders. The original data was divided into three groups: female policyholders were labeled as “F“, male policyholders were labeled as “M“, and policyholders representing a company were labeled as “Z“. However, upon further analysis, it was found that the group of 888 policies identified as representing a company were considered an information anomaly, as one of the conditions of the life insurance product in this study is that the policyholder must be an individual person. Therefore, it was decided that the group “Z“ should be considered as missing values. As a result, 888 policyholders were now marked as missing for this feature.

This recoding eliminated the anomaly and provided more accurate and informative data for analysis and modeling. As a result, the feature *OD_GENDER* now only classifies the gender of policyholders as “F“ or “M“ and 888 policies were considered as missing value. This cleaning improved the dataset and enables more accurate analysis and modeling of the policyholder data (figure 4.13).

Figure 4.12: Gender. **Source:** Authors.



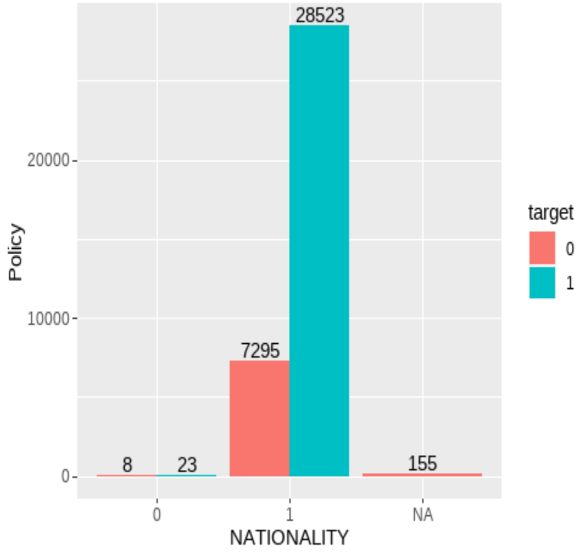
4.3.4.6 Nationality

The feature *OD_NATIONALITY* is used to identify the nationality of the policyholders. The nationalities of the policyholders were originally represented by codes such as “FRA“ for France, “CHE“ for Switzerland, “NLD“ for Netherlands, “GBR“ for United Kingdom, “LUX“ for Luxembourg, “KG“ for Kyrgystan, “BRA“ for Brazil, “ITA“ for Italy, and “POR“ for Portugal. The majority of the policyholders in the dataset were from Portugal, with a total of 35,879 policyholders. However, there were also 31 policyholders with other nationalities, and 155 policyholders who did not have this information recorded.

To improve the data, it was decided to partition this feature into two groups: Portuguese policyholders were encoded as 1, and policyholders with any other nationalities were encoded as 0. This recoding improved the dataset and enables more accurate analysis and modeling of the policyholder data. Additionally, the 155 policyholders without nationality information were considered as missing values and represented as "NA".

This recoding helped to better understand the dataset and improved the accuracy and informative of the data. The majority of the policyholders were Portuguese, which are now encoded as 1, and the other nationalities are encoded as 0 (figure 4.13).

Figure 4.13: Nationality. **Source:** Authors.



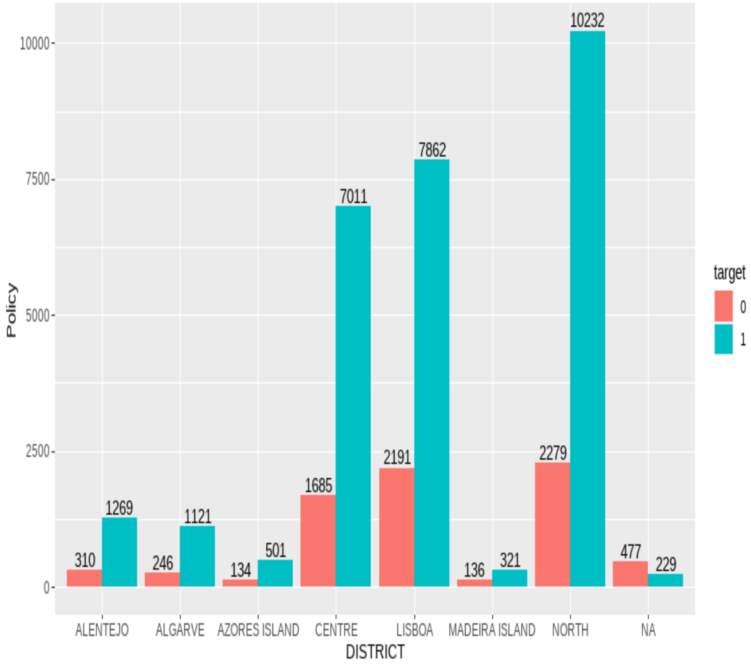
4.3.4.7 District

The feature *COD_DISTRICT* is the code associated with the district where the policyholder resides. Each code represents one Portuguese district (table B.7 and figure B.11).

The majority of policyholders live in Porto (encoded as 13) and Lisbon (encoded as 11). To better analyze the data, this feature was grouped into 7 regions, each representing a specific area of Portugal. The districts of Beja (encoded as 2), Évora (encoded as 7), and Portalegre (encoded as 12) were grouped together as the ALENTEJO region; the district of Faro (encoded as 8) was grouped as the ALGARVE region; the districts with codes between 41 and 49 were grouped as the "AZORES ISLAND" region; the districts of Aveiro (encoded as 1), Castelo Branco (encoded as 5), Coimbra (encoded as 6), Guarda (encoded as 9), Leiria (encoded as 10), and Viseu (encoded as 18) were grouped together as the CENTRE region; the districts of Lisboa (encoded as 11), Santarém (encoded as 14), and Setúbal (encoded as 15) were grouped together as the LISBOA region; and the districts of Braga (encoded as 3), Bragança (encoded as 4), Porto (encoded as 13), Viana do Castelo (encoded as 16), and Vila Real (encoded as 17) were grouped together as the NORTH region.

The majority of policyholders live in the North region, while the minority live in the Madeira Island region.

Figure 4.14: Portuguese Regions. **Source:** Authors.

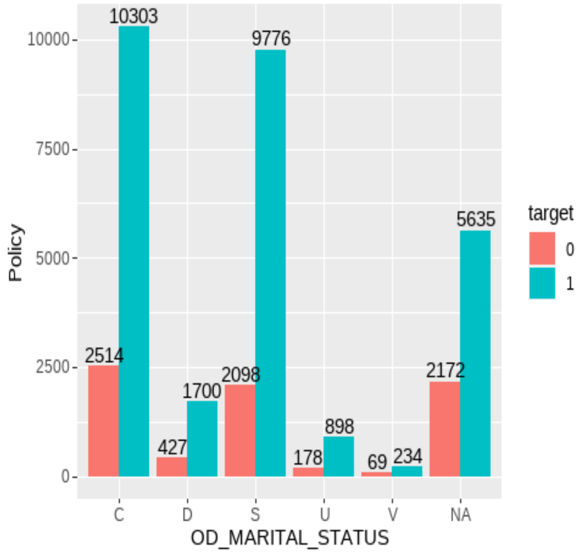


4.3.4.8 Maritial Status

The feature *OD_MARITAL_STATUS* indicates the marital status of the policyholder. It was divided into 7 groups, with codes representing married (“C”), divorced (“D”), separated (“P”), single (“S”), non-marital partnership (“U”), widowed (“V”) and company (“Z”), (table B.8). However, the group of 899 policies identified as company were considered an anomaly as the condition of the life insurance product in this study is that the policyholder is an individual person (figure 4.15).

Therefore, this group was considered as missing value (“NA”). As a result, 7823 policyholders do not have this information. Additionally, the insurance company considers the class “P” and “D” as the same information, resulting in a feature with only 5 possible classes (‘C’,’D’,’S’,’U’,’V’).

Figure 4.15: Maritial Status. **Source:** Authors.

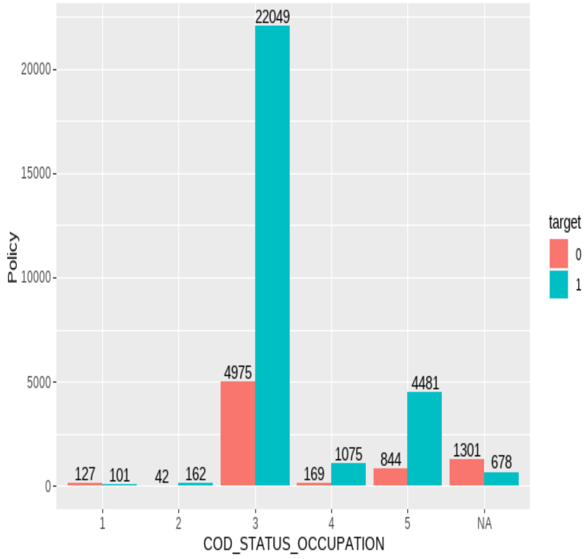


4.3.4.9 Status Occupation

The feature *COD_STATUS_OCCUPATION* describes the occupation status of the policyholder. This variable was originally divided into 6 categories, with codes representing active (19000), retired (16000), unemployed (17000), pre-retired (20000), student (21000), and self-employed (30000) individuals (figure B.12).

The categories were then grouped and encoded into 5 groups for analysis. Specifically, “retired“ and “pre-retired“ were combined and encoded as 1, “unemployed“ as 2, “active“ as 3, “student“ as 4, and “self-employed“ as 5. This allows for more detailed analysis of the occupation status of policyholders (figure 4.16).

Figure 4.16: Status Occupation Categorized. **Source:** Authors.



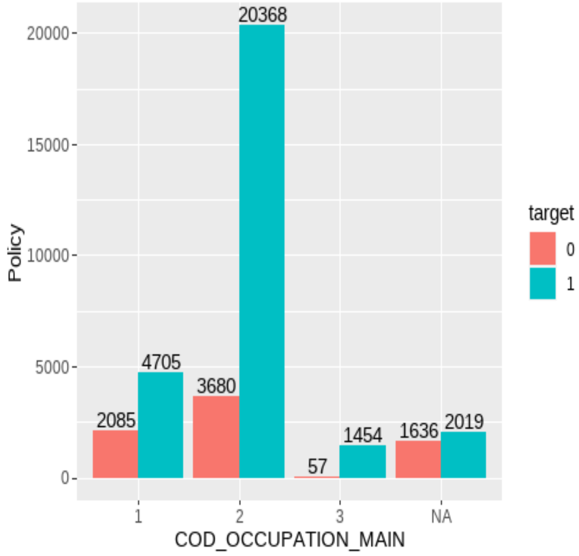
4.3.4.10 Main Occupation

The feature *COD_OCCUPATION_MAIN* identified the main occupation/profession of the policyholder. It was sourced from the insurance's Data Mart and classified into 464 codes, each representing the main occupation.

The goal of the study was to predict lapse contracts (target=0), hence the variable was binned into 5 groups. Frequency encoding was applied to this variable, taking into account the target class label equaled to one. The relative frequencies of each main occupation represented the proportion of policyholders associated with a lapsed policy. This means that if the frequency was equal to one, all policyholders with a specific occupation had lapsed. Conversely, if the frequency was equal to zero, no policyholders with a specific occupation had lapsed.

These frequencies were then binned into 3 groups using the unsupervised binning technique k-means. The group " [0,0.7857)" with 153 occupation codes was label encoded as 3, the group "[0.7857,0.9167)" with 154 occupation codes was encoded as 2, and the group "[0.9167,1]" with 155 occupation codes was encoded as 1. The class 2 was associated with the largest number of policyholders and about 3655 policies did not have this information (see figure 4.17). For example, the categories of "Automobile Mechanic" and "Carpenter" were assigned the code 1, "Medical Secretary" and "Insurance Mediator" were assigned the code 2, and "University and Higher Education Teacher" and "Expert Investigator" were assigned the code 3."

Figure 4.17: Main Occupation Categorized. **Source:** Authors.

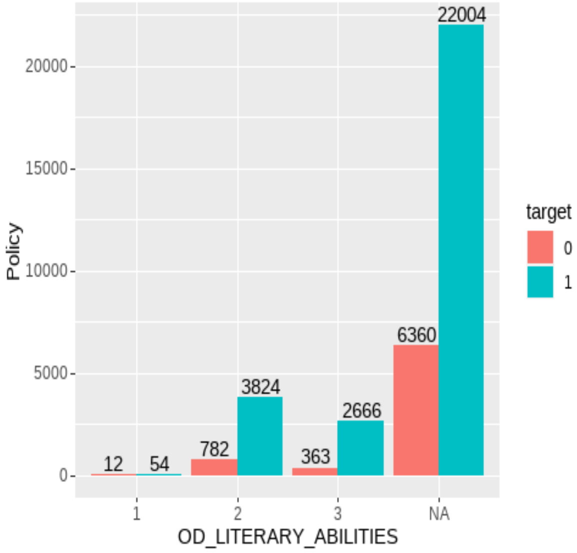


4.3.4.11 Literary Abilities

The feature *OD_LITERARY_ABILITIES* was used to classify the literacy ability of the policyholder. The feature's class 1 identified policyholders without schooling, class 2 identified those with incomplete basic education, class 3 identified those with basic education, class 4 identified those with secondary education, class 5 identified those with a bachelor's degree, class 6 identified those with an undergraduate degree, and class 7 identified those with a master's or doctoral degree (table B.9).

The feature was then grouped into 3 classes. Classes 1 and 2 were grouped and encoded as 1, classes 3 and 4 were encoded as 2, and classes 5, 6, and 7 were encoded as 3. The majority of the policies did not have this information, with approximately 79% of the policies being identified as missing value ("NA"), figure (4.18).

Figure 4.18: Literary Abilities. **Source:** Authors.



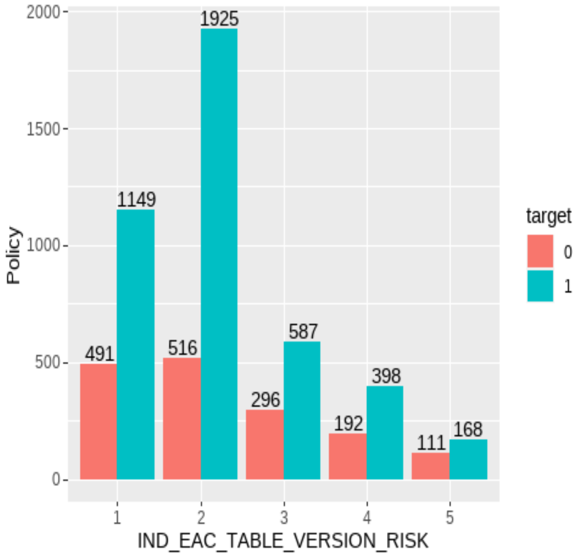
4.3.4.12 Economic Activity Code (EAC)

The Economy Activity Code (EAC) feature, *OD_EAC*, is used to classify the activities of companies associated with a policy for taxation purposes. This study found 319 different codes for the EAC feature. Additionally, the *IND_EAC_TABLE_VERSION* feature represents the first 2 or 3 digits of the EAC, with 144 different codes. The study found that 84% of the policies in the sample do not have EAC information.

Each EAC is associated with various risks for the insurance company, and the company assesses the risk level associated with each code. This assessment is represented as a percentage (feature *PCT_IND_EAC_TABLE_VERSION_RISK*) and is classified into 5 classes, with class 1 representing the least risky activities and class 5 representing the riskiest activities. For example, in this sample, the EAC for Maritime fishing is considered one of the riskiest economic activities (*IND_EAC_TABLE_VERSION_RISK*

= 4) while computer programming activities is considered one of the lowest risks ($IND\_EAC\_TABLE\_VERSION\_RISK = 1$). It's important to note that approximately 84% of the policies in the sample do not have this risk assessment information.

Figure 4.19: Economic Activity Code (EAC) Risk. **Source:** Authors.

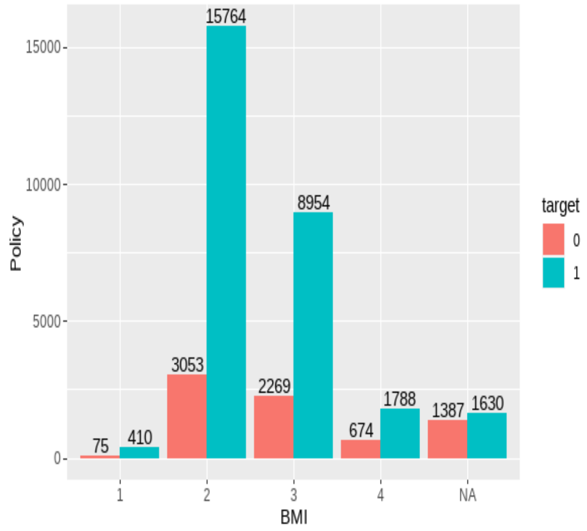


4.3.5 New Variables to Enhance Model Performance

4.3.5.1 Body Mass Index (BMI)

The Body Mass Index (BMI) is a measure of a person's weight in relation to their height. It was calculated by dividing a person's weight (as reported in question $X111$ (4.3.2.12)) by the square of their height (as reported in question $X112$ (4.3.2.13)), both of which are measured in meters (figure B.13). After the BMI was calculated, it was divided into four categories. Class 1 includes those with a BMI less than 18.5, Class 2 includes those with a BMI between 18.5 and 25, Class 3 includes those with a BMI between 25 and 30, and Class 4 includes those with a BMI greater than 30. The majority of policyholders in the sample are classified as Class 2 (normal weight) according to their BMI (see figure 4.20).

Figure 4.20: BMI Categories. **Source:** Authors.



4.3.5.2 Risk of Aggravation Motive 1-8

Aggravation motives are considered a risk to the insurance company. As such, the company assesses these motives by applying a tax to the initial premium. The tax varies depending on the specific event and policyholder. To evaluate each aggravation motive, the mean of different taxes was calculated and then classified on a scale from the lowest risk (class 1) to the highest risk (class 41) based on the average tax. This assessment takes into account all aggravation motives in the Database and the sample in this study are scored in table (B.6).

The lowest risk premium aggravation is "Weight/Height Ratio" (encoded as 1) and the highest risk premium aggravation is "Oncological Pathology" (encoded as 22). This assessment serves as a way to categorize the variable $MOTIVOAGRAVAMENTO$. The features of the Risk of aggravation motive are related to the features of aggravation motive. As such, there are eight features of the risk of aggravation motives represented as $RISK\_MOTIVOAGRAVAMENTO\_1$ to $RISK\_MOTIVOAGRAVAMENTO\_8$. For example, when MOTIVOAGRAVAMENTO = 1, then RISK_MOTIVOAGRAVAMENTO = 2. These features are filled according to the corresponding aggravation motive features.

4.3.5.3 Mean Risk Percentage of Aggravation Motives (mean_PCT_RISK_MTAGRAV)

The $mean\_PCT\_RISK\_MTAGRAV$ feature represents the average risk to the insurance company associated with each type of policy aggravation. It was calculated by averaging the tax associated with each aggravation motive, based on the number and type of aggravation motives associated with each policy. The range of values for this feature is from 0 to 2.54, with 0 indicating no aggravation and 2.54 indicating the highest level of risk.

The majority of policyholders in the sample have never had an aggravation associated with their policy, resulting in a mean value of 0 for 32,369 policyholders.

4.3.5.4 Disease 1-15 (DOENCA_1-15)

The features related to Diseases are derived from the responses provided by the policyholder in a survey about their personal medical history (see Appendix A.2), which were processed in Section 4.3.2 of the study. These features are based on the categorization of the responses to Group 2 questions in the survey, merged with $X107$ (4.3.2.8) and $X108$ (4.3.2.9). These features substituted the Group 2 questions features and were labelled from $DOENCA\_1$ to $DOENCA\_15$.

The feature $DOENCA\_1$ indicates whether the policyholder currently has or has had a disease, while the remaining features indicate the presence of additional diseases. For example, the variable $DOENCA\_2$ would be filled if the policyholder has a second disease, and the feature $DOENCA\_15$ would indicate if the policyholder has or had fifteen different diseases. The new variables are grouped into 19 different disease categories, based on the insurance company's medical condition list (see table B.10). The new variables are grouped into 19 different disease categories based on the insurance company's medical condition list, enabling a more comprehensive understanding of the policyholder's medical history.

This allows the insurance company to better understand the policyholder's medical history and assess their risk profile. Additionally, these features will help the insurance company to understand the policyholder's medical history in a more detailed way.

4.3.5.5 Mean Risk Percentage of Diseases (mean_PCT_RISK_DOENCA)

The feature $mean\_PCT\_RISK\_DOENCA$ represents the average risk percentage associated with each disease, as determined by the insurance company's guidelines. This value was calculated based on the number and types of diseases associated with each policyholder. The values of this variable range from 0 to 2.24, with the majority of policyholders having a value of zero and only a small number having the highest mean percentage of risk.

4.3.5.6 Medicine 1-7 (MEDICAMENTOS_1-7)

The features related to Medicines were constructed based on the policyholder's responses to Group 3 questions in the survey regarding their therapeutic history (see appendix A.3), which were processed in Section 4.3.2 of the study. These features replaced the Group 3 question features and were numbered from $Medicamentos\_1$ to $Medicamentos\_7$.

The feature $Medicamentos\_1$ indicates whether the policyholder has taken or is currently taking medication for a specific disease. The subsequent medicine features indicate whether the policyholder has taken or is currently taking medication for other diseases. The new variables are categorised into different diseases based on the information provided in the survey. This allows the insurance company to better understand the policyholder's therapeutic history, which can help to determine the risk profile of the policyholder.

Additionally, this information can be used to identify any potential drug interactions or contraindications that may need to be taken into consideration when assessing the policyholder's coverage or claims.

4.3.6 Dataset

The dataset for analysis can be found in the Appendix (see Table B.11). To ensure the relevance and accuracy of the analysis, the identification variables ID, $COBERTURA$ ID, and the description variables have been excluded. Furthermore, the variables $MOTIVO$ $AGRAVAMENTO$ have been removed and grouped into categories (represented by $RISK\_MOTIVOAGRAVAMENTO$ variables). The dataset now consists of 75 variables and 35,980 observations (see Section 4.2.1), with 72 being categorical and 3 being numerical (see table B.11).

The percentage features were considered categorical, as a significant number of cases had a high proportion of value zero. This was done in order to better model the data using machine learning algorithms. Furthermore, categorical factors were used to improve the classification performance.

4.4 One-Hot Encoding

Several models were used in this study, including logistic regression, which are most effective when the variables have been converted into a dummy encoded format. To achieve this, One-Hot encoding was applied to the dataset after previously performing data transformations such as binning techniques and anomaly elimination. One-Hot encoding is a method that converts categorical variables into a numeric format suitable for use in machine learning algorithms, which can enhance the classification performance of the models.

The One-Hot encoding technique was applied only to categorical variables that had already been divided into categories. As a result, the entire dataset was divided into several binary features, except for the $mean\_PCT\_RISK\_DOENCA$ and $mean\_PCT\_RISK\_MTAGRAV$ variables, which are already continuous. This data transformation was done using the "vtreat" package's $designTreatmentsZ()$ and $prepare()$ R-functions, version 1.6.3. The transformed dataset is described in the Tables (B.12) and (B.13).

As mentioned in the Section 3.2.1.2, to avoid issues with multicollinearity, variables that had only two categories were represented in the dataset by a single variable. Specifically, the variables OD_GENDER, $CONTRACTAMOUNT$, and $COVERAGEVALUE$. Additionally, the NUM_AGE feature has been removed as its information is categorized by the OD_AGE_GROUP variable.

4.5 Feature Scaling and Centering

The numerical features $mean\_PCT\_RISK\_DOENCA$ and $mean\_PCT\_RISK\_MT AGRAV$ underwent a data transformation through standard scaling and centering operations. This process of data standardization transformed the features into new, standardized features with a mean value of zero and a scale of unit variance. This standardization was implemented to ensure that the numerical features would have the same magnitude, which can help improve the performance of machine learning algorithms. It was performed using the $preProcess()$ R-functions available in the '*caret*' package, specifying the method as "center" and "scale" respectively. This was an important step, particularly for models like the Naive Bayes classifier and logistic regression, which greatly benefited from this transformation during the training process. The output of this process can be found on Table (4.46).

Table 4.46: Feature Scaling and Centring. **Source:** Authors.

Statistic	MEAN_PCT_DISEASE_RISK	MEAN_PCT_AggravationRisk
Min.	-0.4438	-0.676
1st Qu.	-0.4438	-0.2147
Median	-0.4438	-0.2147
Mean	0	0
3rd Qu.:	-0.4438	-0.2147
Max.	2.4708	8.08841

4.6 Handling missing data

A fundamental step in preparing data for analysis is treating the missing values. Firstly, the features with more than 30% of missing values were not considered in this study. As a result, the features related to Economic Activity Code (see Section 4.3.4.12), $X108$ (A.1.8), "X503" (A.5.3), "X505" (A.5.5), "X506" (A.5.6), "X110" (A.1.10), "X102" (A.1.2), "X103" (A.1.3), "X107" (A.1.7), "X502" (A.5.2), "OD_LITERARY_ABILITIES", "X507" (A.5.7), and "X6011" (A.6.1) were initially removed (see Table 4.47).

After the removal, the Hot-deck imputation method was applied first, where the data was sorted by variables that were correlated to the one being imputed, in order to improve the accuracy of the imputation. This method is based on the idea of using similar donors to fill the missing data.

The hot-deck imputation method was made more effective by sorting the data based on variables that are correlated to the one being imputed. In this case, as shown in the figure 4.21, there was a strong relationship between variables $X111$ (Weight) and $X112$ (Height). By exploiting this correlation, the data was ordered by $X112$ (Height) and imputed $X111$ (Weight) values were taken from donors with similar $X111$ (Weight) values. Similarly, imputed $X112$ (Height) values were taken from donors with similar $X112$ (Height) values. This resulted in imputation of 350 values of X112 (Height) and one value of $X111$ (Weight). As a result, the corresponding missing Body Mass Index (BMI) values were calculated using these imputed values. The hot-deck imputation was performed by using *hotdeck*() R-function available in the '*VIM*' package.

After undergoing the Hot-deck imputation and one-hot encoding process (detailed in Section 4.4), the total missing values were considered and consisted of 4.8% of the total data (Appendix B.14). There were 95 dummy variables that had anywhere between 0.13% to 21.68% missing values. Therefore, it was considered impute these missing values.

A Little's test was first conducted to determine if the data was Missing Completely At Random (MCAR). The result of this test showed that the data was not MCAR, and thus, alternative mechanisms such as Missing At Random (MAR) or Missing Not At Random (MNAR) were assumed (see Section 2.4.11.3). This test was applied by using *mcar_test*() R-function available in the '*naniar*' package version 0.6.1. As a result, it was obtained a p-value less than 0.05 implying that the data was not MCAR.

The missing values in the dataset were imputed using a combination of three different methods: Hot-deck imputation, MICE, and Random Forest imputation. The MICE and Random Forest imputation methods were independently applied to the data.

The MICE imputation method used logistic regression imputation ('*logreg*') for each incomplete binary variable. This method normally uses multiple imputation, where multiple sets of plausible values for the missing data are generated, based on the observed data, and then combined to produce a single imputed dataset. The MICE imputation method was performed by using mice() R-function available in the '*mice*' package.

The Random Forest imputation method used a different approach, where it calculated the Normalized Root Mean Squared Error (NRMSE) for continuous variables and the Proportion of Falsely Classified Entries (PFC) for categorical variables. These measures performed values close to zero, indicating that the imputed values were accurate. This method was performed using *missForest*() R-function available in the "*missForest*" package.

Both MICE and Random Forest imputation methods were successful in producing accurate imputed values. As a result, two different complete datasets were generated - one that was a combination of Hot-deck and MICE imputation and the other was a combination of Hot-deck and Random Forest imputation.

The use of multiple imputation methods, including Hot-deck, MICE and Random Forest, aimed a more robust imputation and improved the overall accuracy of the dataset. This approach not only may increase the power of statistical analysis but also allows for a more comprehensive understanding of the uncertainty and variability associated with the missing data.

After generating the two datasets, they were compared to evaluate the performance of each imputation method. Before this evaluation, both datasets were then used for further analysis to draw insights and conclusions.

Table 4.47: Percentage of missing data. (Top 14) **Source:** Authors.

Feature	Type	Percentage Data missing
X6011	Discrete	100.00%
X108	Discrete	99.98%
X503	Discrete	99.88%
X505	Discrete	99.55%
X506	Discrete	99.55%
X110	Discrete	99.24%
X102	Discrete	98.96%
X103	Discrete	98.95%
X107	Discrete	96.83%
X502	Discrete	95.60%
IND_EAC_TABLE_VERSION_RISK	Discrete	83.81%
OD_EAC	Discrete	83.81%
IND_EAC_TABLE_VERSION	Discrete	83.81%
PCT_IND_EAC_TABLE_VERSION	Discrete	83.81%
X507	Discrete	83.25%
OD_LITERARY_ABILITIES	Discrete	78.64%

4.7 Feature Selection

This study leverages a combination of multiple feature selection methods to identify the most critical features for accurately predicting the lapse rate in the given dataset. These methods include correlation-based filters, low-variance filters, statistical tests (such as the chi-squared filter method and F-test method), and Factor Analysis of Mixed Data (FAMD).

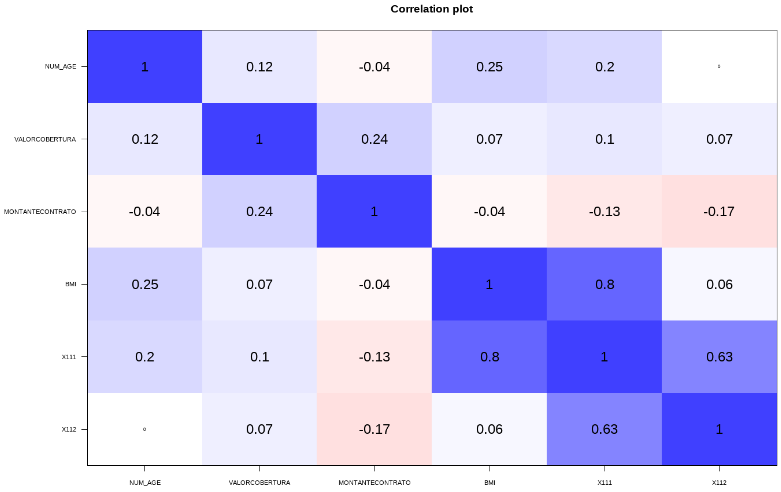
Initially, a correlation-based filter method was applied to eliminate features with constant values by analysing the covariance matrix of the quantitative variables in figure 4.21. It is visible that there were significant relationships between $X111$ (Weight) and $X112$ (Height), as well as between BMI and $X111$ (Weight), so these features are deemed redundant and candidates for removal. Consequently, it was decided to remove the $X111$ (Weight) feature from the data, resulting in a new covariance matrix (see Appendix B.15).

Subsequently, a low variance filter method was also applied, using the near zero-variance (*nzv*) method to identify predictors that may cause the models to crash. This method was performed by using *preProcess*() R-functions available in the *caret* package, specifying the type of processing with the method “*nzv*“, removing the feature *MEAN_PCT_ AggravationRisk*.

The study also compared supervised and unsupervised feature selection techniques. The Chi-Squared filter and F-test methods, which are supervised, were evaluated against Factor Analysis of Mixed Data (FAMD), an unsupervised method.

The first two steps were performed before imputing the missing data using MICE and Random Forest (see section 4.6). The techniques described in the third step were then applied to two datasets that had undergone imputation using MICE and Random Forest, respectively. This resulted in four datasets, which were evaluated using a prediction performance test. The best dataset was then used to conduct various prediction models.

Figure 4.21: Covariance Matrix I. **Source:** Authors.



	NUM_AGE	VALORCOBERTURA	MONTANTECONTRATO	BMI	X111	X112
NUM_AGE	1	0.12	-0.04	0.25	0.2	0
VALORCOBERTURA	0.12	1	0.24	0.07	0.1	0.07
MONTANTECONTRATO	-0.04	0.24	1	-0.04	-0.13	-0.17
BMI	0.25	0.07	-0.04	1	0.8	0.06
X111	0.2	0.1	-0.13	0.8	1	0.63
X112	0	0.07	-0.17	0.06	0.63	1

4.7.1 Supervised Feature Selection: Chi-squared and F-test selection methods

The selection method Chi-square and an F-test were applied to select only features that were relevant in predicting the target for categorical and continuous variables respectively. These methods were used to test the level of dependency of a variable on the target by analysing the relationship between the variables and the target. The variables were then scored based on their variable importance.

The *summary_factorlist*() R-function available in the “*finalfit*“ package version 1.0.5 was used to perform these methods, specifying the dependent variable as the target. From the output summary table, the variables with a p-value less than or equal to 0.01 were selected as significant in predicting the target. As a result of these tests, 26.16% of the features from the MICE imputed dataset and 29.07% of the features from the Random Forest imputed dataset were removed. This resulted in two different datasets that aim to improve accuracy, computational time, and reduce the risk of overfitting.

4.7.2 Unsupervised Feature Selection: Factor Analysis of Mixed Data (FAMD)

This analysis was conducted in R, in particular used *FactoMineR* package, version 2.7. And the results were extracted and visualized using the *Factoextra* package, version 1.0.7.

Before proceeding with the Factor Analysis of Mixed Data (FAMD), it was important to note that all numeric variables were standardized as previously described on Section 4.5. The principal dimensions (PDs) were calculated as linear combinations of the original variables in order to better account for the variance in the dataset. Scree plots were used to evaluate the eigenvalues and percentage variance explained by each PD, providing insight into the level of information represented by the original variables.

An eigenvalue greater than 1 indicates that the PD accounts for more variance than one of the original variables in the standardized data, and is commonly used as a cutoff point for determining the number of factors to retain in the analysis.

Additionally, an examination of the top contributing variables to the PDs can provide insight into which variables underlie variations in the dataset, and may aid in feature selection. This analysis was conducted on both the datasets obtained from the Multiple Imputations by Chained Equations (MICE) and Random Forest imputation methods.

The dataset resulting from random forest imputation was analysed to select the most relevant features using the *FactoMineR* package in R. The principal dimensions (PDs) were calculated, along with their associated eigenvalues, the variance explained by each factor, and the cumulative percentage of variance. It was found that only the first 72 PDs had an eigenvalue greater than 1, representing 73.68% of the original dataset's variability. These 72 PDs were used to select the features to retain in the analysis.

In addition, a scree plot was used to identify the first ten PDs that account for more variance than each of the original variables, with PD 1 accounting for only 8.9% of the total variance in the dataset (see figure (4.22)). The *Factoextra* package was then used to identify the top contributing variables to the 72 PDs, which were considered the most important variables in the dataset that met the cut-off represented by the red dashed line (see appendix (B.16)). This cut-off indicated the expected average contribution, and based on this cut-off the most important variables were selected, representing

55.81% of the total features. These 80 features were selected as the predictor features for posterior analysis, with key variables such as $X401\_lev\_x\_1$, $POLICYTERM\_5$, and $X401\_lev\_x\_0$ being identified as the most important predictors.

On other hand, the application of FAMD on the dataset resulting from the MICE imputation revealed that only 59 principal dimensions (PDs) had an eigenvalue greater than 1, representing 74.88% of the original dataset's variability. These 59 PDs were used to select the relevant features for the analysis.

The scree plot, presented in the figure (4.23), illustrates the first ten PDs that account for more variance than each of the original variables, with PD 1 accounting for 24.3% of the total variance in the dataset. The appendix (B.17) illustrates the variables that met the cut-off of a contribution value of 0.578%, which were considered important in contributing to the 59 PDs. As a result, 94 predictor features were selected, with the most important variables in this dataset being $POLICYTERM\_5$, $HEIGHT\_165\_173$ and $X401\_lev\_x\_0$.

Figure 4.22: Scree Plot of Random Forest imputation dataset. **Source:** Authors.

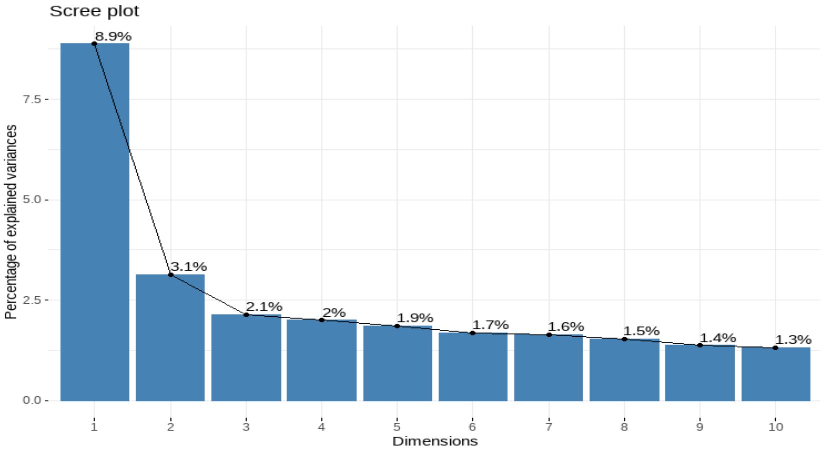
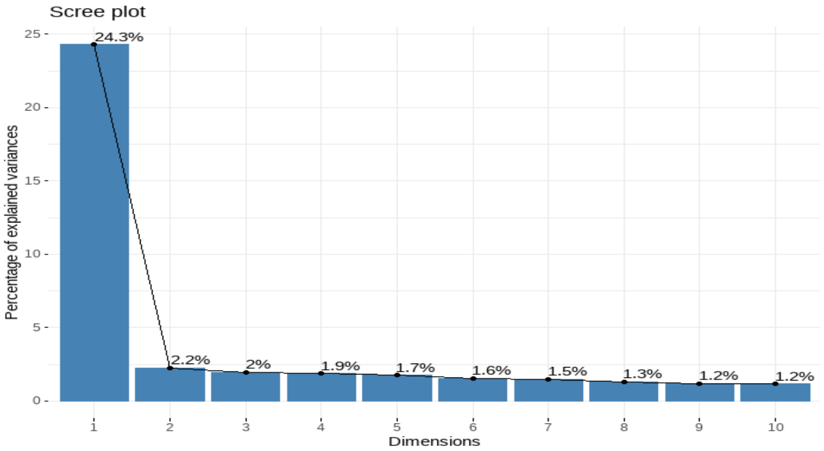


Figure 4.23: Scree Plot of MICE imputation dataset. **Source:** Authors.



4.8 Model Training

After cleaning the data, engineering features, and creating new variables to enrich it, we proceeded to develop predictive models to forecast future outcomes that would help the insurance company reduce their lapse rate. In this project, we trained and compared several models in order to select the one that is best suited for this particular problem. We then further inspected this model to understand the level of cancellation rate that it has extracted from the data.

4.8.1 Data Splitting

The models were trained and validated using a 75% training set and 25% testing set partition of the data.

Four different training datasets were produced using different combinations of missing values imputation and feature selection methods. Also, four corresponding testing sets were produced without applying feature selection methods. These training and testing sets were created using the *createDataPartition*() R-function from the “*caret*“ package (version 6.0-93).

From the dataset resulting from random forest imputation, the reduced training set produced with supervised feature selection was named “$training\_st\_rf$“ and consisted of 122 features. The reduced training set produced with FAMD was named “$training\_FAMD\_rf$“ and consisted of 80 features. Similarly, based on the dataset resulting from Mice imputation, the reduced training set produced with supervised feature selection was named “$training\_st\_mice$“ and consisted of 127 features, while the reduced training set produced with FAMD was named “$training\_FAMD\_mice$“ and consisted of 94 features. Each of these datasets consisted of 26,985 observations.

Corresponding testing sets were also produced, but without applying feature selection methods. From the dataset resulting from random forest imputation, the testing sets were named “$testing\_st\_rf$“ and “$testing\_FAMD\_rf$“, and from the dataset resulting from Mice imputation, the testing sets were named “$testing\_st\_mice$“ and “$testing\_FAMD\_mice$“. Each of these datasets consisted of 8,995 observations and 172 features.

4.8.2 Optimizing Training Data: Representative Sample Selection

Different imputation approaches and feature selection methods produce different data, and as a result, classification models perform differently with these different datasets.

Determining the optimum approach for achieving the best results can be challenging. To identify the best performing training set, a performance test was conducted using the Recursive Partitioning and Regression Trees model (”rpart“), and evaluated with a 10-fold cross-validation. The decision to choose this model was made based on its efficiency and simplicity, as evaluating all proposed models in this study would have been ideal but would have been excessively time-consuming and computationally demanding.

For each training set, the minimum, first quarter, median, mean, third quarter, and maximum AUC-ROC were calculated. The corresponding cross-validation procedure is displayed in Table (4.48) and Figure (4.24) for each training set. The training set (“*training_st_rf*“), resulting from the combination of random forest imputation and feature selection with supervised feature selection, ranked first. Additionally, obtained the smallest standard deviation of the AUC-ROC for this training set (1%), indicating that it is less prone to sample selection. This can be seen in the box plot in Figure (4.24).

Table (4.49) also shows that the training set (“*training_st_rf*“) obtained the best AUC-ROC on the corresponding testing set.

As a result, the training set (“*training_st_rf*“) was selected to train all the models. Overall, this evaluation showed that random forest imputation outperformed the mice imputation in our dataset, independently the feature selection method.

Table 4.48: Cross-Validated Statistic Accuracies (Test 1). **Source:** Authors

Training Sets	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Standard Deviation**	**NA's**
st_rf	0.73	0.75	0.76	0.76	0.76	0.77	0.01	0
FAMD_rf	0.68	0.70	0.74	0.73	0.75	0.77	0.03	0
FAMD_mice	0.66	0.68	0.71	0.70	0.72	0.74	0.03	0
st_mice	0.64	0.68	0.68	0.68	0.69	0.69	0.02	0

Figure 4.24: Box Plot of Traning Set Evaluation (AUC-ROC). **Source:** Authors.

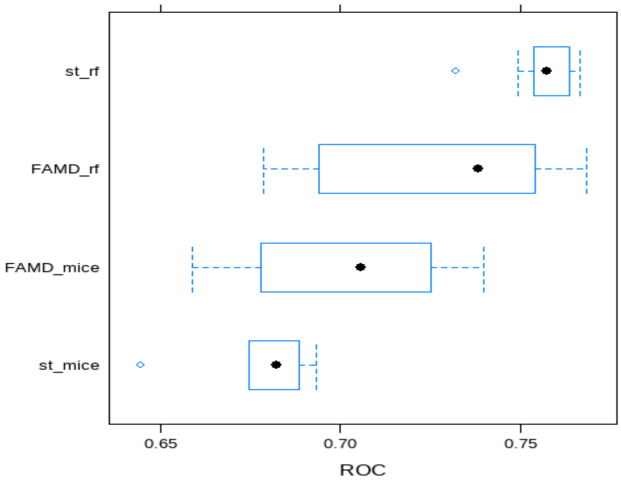


Table 4.49: Training sets' Area Under the Curve on testing set (Test 1)

Training Sets	AUC-ROC
st_rf	0.6874
FAMD_rf	0.6868
FAMD_mice	0.6818
st_mice	0.6396

4.8.2.1 Rebalancing the Selected Training Set

The analysis in Section (4.3.1) revealed an imbalanced class distribution in the dataset, with significantly fewer policyholders who had lapsed compared to those with current policies. This class imbalance was demonstrated in the Target variable.

To examine the impact of rebalancing techniques on the Recursive Partitioning and Regression Trees model, we applied 10-fold cross-validation to the training set "*training_st_rf*", which was derived from the random forest imputation and feature selection using statistical tests. For each rebalancing technique, we calculated the minimum, first quartile, median, mean, third quartile, and maximum accuracy, as well as the standard deviation.

The results, shown in Table (4.50) and Figure (4.25), indicate that the Cross-Validated Statistic area under the curve decreased when data was rebalanced using under-sampling (down), oversampling (up), or with no technique (orig). However, the SMOTE technique produced the highest AUC-ROC (see Table (4.51)). As a result, it was implemented SMOTE in the training process for all models studied in this research.

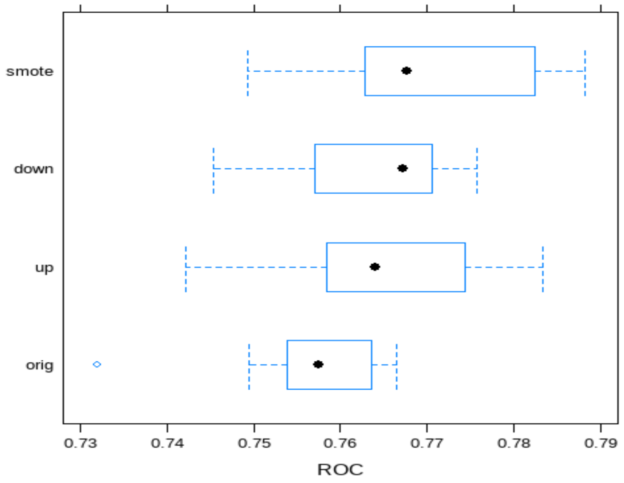
Table 4.50: Cross-Validated Statistic Accuracies (Test 2). **Source:** Authors

Training Sets	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard Deviation	NA's
st_rf (SMOTE)	0.749	0.763	0.768	0.771	0.782	0.788	0.013	0
st_rf (oversampled)	0.742	0.759	0.764	0.765	0.773	0.783	0.011	0
st_rf (undersampled)	0.745	0.759	0.767	0.764	0.770	0.776	0.009	0
st_rf (original)	0.732	0.754	0.758	0.756	0.763	0.766	0.010	0

Table 4.51: Training sets' Area Under the Curve on testing set (Test 2). **Source:** Authors.

Training Sets	AUC-ROC
st_rf (SMOTE)	0.792
st_rf (undersampled)	0.779
st_rf (oversampled)	0.768
st_rf (original)	0.687

Figure 4.25: Box Plot of Rebalancing Techniques Evaluation (AUC-ROC). **Source:** Authors.



4.9 Model Evaluation and Improvement

There were thirteen classification machine learning algorithms included: Random Forest, C5.0 model, Recursive Partitioning and Regression Trees model (*rpart*), Extreme Gradient tree boosting (*XGBoostTreeBooster*), Extreme Gradient Linear boosting (*XGBoostLinearBooster*), Elastic-Net Regularized Generalized Linear Model (*glmnet*), Adaptive Boosting (*AdaBoost*), Naïve Bayes (*nb*), K-nearest neighbours (*Knn*), Neural Networks, Bagged Classification Tree (*BaggingClassifier*), and logistic regression. In addition, it was built a new model with the application of the ensemble stacking method (see annex (II.1)).

4.9.1 Model Development

Before evaluating and comparing the models, it is important to optimize each model individually in terms of its hyperparameters, in order to achieve the best possible performance. During the hyperparameter tuning process, different parameter settings were tried, and the performance of each was measured. It was used a 10-fold cross-validation method to evaluate the models, which involved dividing the data into 10 equal parts and using 9 parts for training and 1 part for testing. This process was repeated 10 times, with each part being used once as the test set. Therefore, it was used the mean accuracy and standard deviation of the 10-fold validation as measures. The standard deviation is a measure of the distribution and is the square root of the variance. This means that a new prediction can be more reliable with a lower standard deviation.

The "*caret*" package in R was used to optimize the hyperparameters of the models. We first created a reusable *trainControl* object using the *trainControl*() function, which allowed us to use the same cross-validation folds for each model. The 10-fold cross-validation resampling method was specified in the *trainControl* object using the *createFolds*() function and setting the parameter k equal to 10.

In addition to the 10-fold cross-validation, we also used the grid-search method with cross-validation for parameter optimization. This involved specifying a range of values for each parameter and using cross-validation to evaluate the performance of each combination of parameter values. The goal of this process was to find the best set of parameter values for each model.

The process used to develop the 13 models was the same. It began with training and cross-validating a baseline model with minimal consideration of parameters, not specifying a grid for tuning. This served as a comparison to ensure that changes in parameters improved performance. Then, grid-search parameter optimization was used to define the "optimal" model. Grid-search was performed on one parameter at a time, using a wider range of values than the default grid for cross-validation. If the best-performing value was at the extreme of the range, the range was shifted and grid-search was run again. The model was evaluated with a test dataset and if there was a fixable problem with the results, the process returned to the baseline model.

Finally, the *trainControl* object was applied in the *train*() function from the caret package version 6.0-93. This function was used to build the specific model using resampling, and to evaluate the effect of tuning the model's parameters on performance. By default, the function automatically selects the tuning parameters associated with the best value and selects the model with the highest performance value. The best model was then selected among the top-performing classifiers in the sample. The conducted hyperparameter optimization is further described in the annex (II).

4.9.2 Model Evaluation

In this Section, it is reported the test results of the final models. These fitted models have been used to predict the Lapse probability on the test dataset. They have been ranked according to performance metrics and the better performing one is going to be selected to predict the Lapse rate on the test dataset.

Appendix (C) shows the model performances of the training dataset reduced by supervised feature selection and without missing values as a result of the random forest imputation method ("*training_st_rf*"). In addition, each model was re-balanced by the SMOTE technique.

First of all, it can be observed that the models achieved a remarkable predictive performance with an average AUC-ROC above 75%. This means that the supervised feature selection methods, the random forest imputation, and parameter tuning resulted to reduce the extent of overfitting as well as in performance.

Looking into a more detailed analysis based on Appendix (C), reveals differences between the different learning methods.

From the perspective of training and test sets' accuracies, the best performing algorithm based on the training accuracy (Train Accuracy) is XGBoostTreeBooster (83.19%) and the worst is Neural Networks (74.6%). However, the best performing algorithm based on the testing accuracy (Test Accuracy) is Random Forest (84.82%) and the worst is Neural Networks (77.18%).

Considering that the balanced accuracy on testing dataset (C.2) is used to assess better the performance of a classification model with imbalanced target, Neural Networks and Logistic Regression showed the best performance value (75.85% and 75.78%, respectively) and Naive Bayes shows the worst performance value (53.55%). Followed by XGBoostTreeBooster and XGBoostLinearBooster which showed a poor performance on balanced accuracy. Parallelly, the kappa is too a good performance measure of a classification model with imbalanced target which Random Forest method obtained the best value and Naive Bayes method the worst value.

Now comparing the performance of the algorithms by looking at the values of the area under the ROC for the training dataset (Train AUC-ROC) based on 10-fold cross validation (see table C.1). The Random Forest outperforms all the models with an average AUC-ROC of 82.22%, followed by XGBoostTreeBooster, C5.0, and glmnet. However, K-Neighbours (KNN) acquired the worst average AUC-ROC of 69.9%.

The smallest standard deviation of AUC-ROC values for the training dataset is obtained by the ensemble model ($KNN + nb$) followed by Random Forest model, with values of 0.11% and 0.25% respectively, which indicates that is less prone to sample selection. Conversely, the biggest standard deviation of area under the ROC curve was obtained by the K-nearest neighbour (KNN) method, 1.58%.

The algorithm with the best f1-score on testing data was the Random Forest method with a score of 61.2%, and with the worst f1-score was Naive Bayes (14.2%). The algorithm with best Precision, Specificity and Positive Predictive Value (PPV) is the Naive Bayes, by contrast Neural Networks is the worst algorithm with these specific measures.

Although Naive Bayes has the largest Specificity value on testing set of 99.2%, the figure C.2 shows that the $XGBoostTreeBooster$ obtained the best average Specificity value on training set, based on 10-fold cross-validation.

On the other hand, the algorithm with best Sensitivity and Negative Predictive Value (NPV) on testing set was Neural Networks and the algorithm with these specific worst scores is the Neural Networks model. The figure C.1 shows that the best average Sensitivity on training set is represented by the Neural Networks, and the worst is represented by Naïve Bayes. This implies that Neural Networks correctly identified the highest number of actual lapses. Moreover, Naive Bayes resulted in the largest specificity (99.2%) and lowest sensitivity (7.9%) which means that the model was good on identifying policyholders that will not lapse but have limitations when identifying policyholders that will lapse based on the threshold of 50% used across all the models.

In addition, considering the other results on testing set (C.2), the models Naive

Bayes (nb) and K-nearest neighbour (knn) obtained the largest Logarithmic Loss and Random Forest the smallest value of 0.345. The Detection Prevalence measures the number of predicted positive events, i.e., forecasts the number of policyholders that will lapse. Looking to the table, Detection Prevalence varies between 2.26% to 32.6%, which Naïve Bayes obtained the smallest number and Neural Networks predicted the largest number, respectively. These values demonstrate on testing dataset that between 2.26% to 32.6% of policyholders will Lapse.

Generally, Knn and Naive Bayes had the lowest performance compared to the other models. Therefore, it was proposed to combine the predictions of these two weaker models using the stacking technique. The ensemble model (knn+nb) was able to incorporate some of the strengths of each individual model and improve upon their original performance (see annex (II.1)).

4.9.2.1 An Overview of Receiver Operating Characteristic Curve, Precision-Recall Curve, and Area Under Curve

Next, it was compared the performance on testing dataset of the algorithms by looking at the values of the area under the ROC (AUC-ROC) and Precision-Recall curves (AUC-PR).

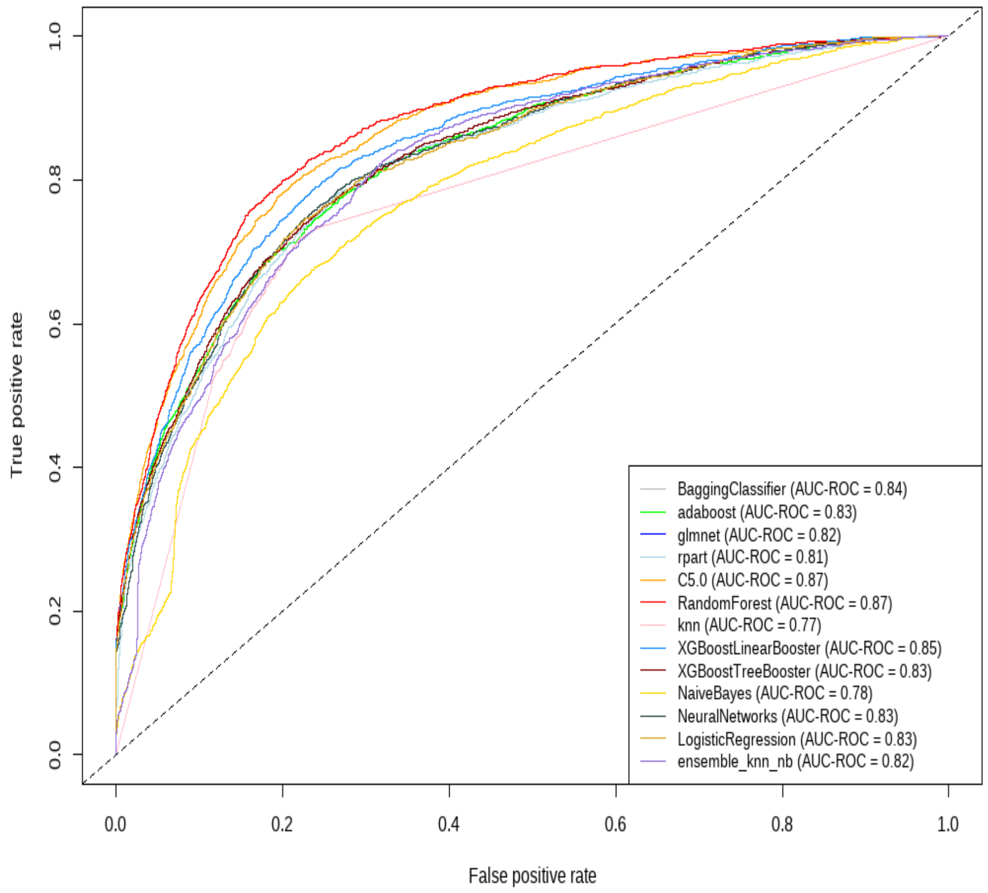
In terms of the area under the ROC Curve, all the algorithms have similar performance and behaviour as we can see in figure 4.26. The analysis of the ROC curves and the corresponding AUC values shows that the classifiers have a strong performance in distinguishing between positive and negative instances. The shapes of the ROC curves for all thresholds are relatively consistent and there is minimal variation in their behavior. The red diagonal line in each graph represents the random chance benchmark and it can be observed that all ROC curves stay above this line, indicating that the classifiers are performing better than a random classifier at all thresholds. This implies that the classifiers have a satisfactory balance of sensitivity and specificity, meaning that they are able to correctly identify most of the positive instances (Lapse instances) while minimizing the number of false negatives and false positives.

Nevertheless, the random forest, C5.0 and *XGBoost* methods achieved a slightly better performance with respect to the measures AUC-ROC on testing set. On the other hand, Naive Bayes and K-nearest neighbour show the worst performance amongst all AUC-ROC performance measures. In particular, Random Forest obtained the best AUC-ROC value on the testing set of 87.19% and K-Neighbours obtained the worst AUC-ROC value of 76.8%.

Now, in terms of the area under Precision-Recall curves (AUC-PR) on testing dataset (see figure (C.3)), *XGBoostLinearBooster* followed by Random Forest showed the best performances (66% and 64.25%, respectively). However, K-Neighbours (knn) obtained the worst AUC-PR performance of 19.64%.

In our specific case study, we placed greater emphasis on the AUC-ROC measure as it is a widely used and accepted metric in the literature for evaluating classifiers in imbalanced datasets. This is because AUC-ROC only considers the rank order of the predictions, whereas AUC-PR can be affected by imbalanced datasets and give more weight to the positive class, resulting in a less reliable measure.

Figure 4.26: AUC-ROC Curves. **Source:** Authors.



4.10 Model Selection

In general, Tree-based methodologies performed the best in predicting lapse rate in our life insurance portfolio. Among these, the Random Forest classifier algorithm showed the highest performance, correctly labelling 84.82% of policyholders in the test set as either lapsing or not lapsing. This accuracy rate of 84.82% on the test set is particularly promising for real-world application. The Random Forest algorithm also achieved an accuracy rate of 82.45% on the training dataset. The *C5.0*, *XGBoostLinearBooster*, and *XGBoostTreeBooster* algorithms also showed good performance, but were ranked lower than the Random Forest algorithm.

Considering the big picture, these results suggest that tree-based methodologies can be effective in predicting lapse rate in the life insurance industry. Among these, the Random Forest classifier demonstrated the most outstanding performance, with an average AUC-ROC of 87.19% on the testing set.

This conclusion is reinforced by the fact that the Random Forest model had the highest average AUC-ROC on both the training and testing dataset, the most favourable Kappa and F1-score, the lowest logarithmic loss and standard deviation, and the second-highest AUC-PR on test data. In contrast, the Naive Bayes model had the least favourable performance in this regard.

Additionally, the Bagged classification tree algorithm demonstrated superior performance on the testing set in terms of AUC-ROC, Balanced Accuracy, F1-Score, and Kappa compared to $XGBoostTreeBooster$ and $Adaboost$. However, when evaluating the performance on the training set, $XGBoostTreeBooster$ and $Adaboost$ showed a higher AUC-ROC and had lower standard deviation. Additionally, $XGBoostTreeBooster$ and $Adaboost$ outperformed Bagging in terms of AUC-PR and Logarithmic Loss on the testing set. On the whole, the results suggest that Bagging method performed better on the testing set and had a better generalization ability than Boosting method.

Overall, considering the threshold of 50% used across all the models, Random Forest predicted that 18.4% of the policyholders on testing dataset will lapse.

Bearing in mind the testing dataset and based on Random Forest's results, the prediction probabilities of Lapsation were divided into 3 bands from the highest to the lowest with the k-means algorithm. Table (4.52) reflect the distribution of total policies at each probability band. This means that RF predicted that 15% of total policies had a 57%-100% chance of lapsing, 25% of total policies had a 23%-57% chance of lapsing, and 60% of total policies had a 0%-23% chance of lapsing.

Table 4.52: Risk Assessment. **Source:** Authors.

	Probability of Lapse	**frequency**	**percentage**
low-risk	0%-23%	5411	60%
medium-risk	23%-57%	2272	25%
high-risk	57%-100%	1312	15%

Interesting business insights can be gleaned from analysing policyholders with a high probability of lapse, defined as those between 57-100%. The analysis provides several demographic and personal characteristics of this population within this high-risk group. It highlights that 26.4% have ages between 52-73, 64.7% are married, 30.6% have product two, 64.6% have category two of main occupation, 86.8% have taken or are currently taking medication for a specific disease, 51.9% have or have had some type of disease, with 41.9% related to the cardiovascular system. Additionally, approximately 98% had some type of aggravation motive, 30.2% are overweight, 46.3% have heights between 145-165cm, 76.1% have a policy term of five years, 77.6% have a monthly premium frequency and 94.3% are not active professionally.

4.11 Model Variable Importance

In this study, Lapse risk prediction models were created using risk survey questions as the main source. The results of these models were analysed to understand the impact of policyholder characteristics (such as age and occupation) on lapse rate and whether these characteristics can be used to predict lapse risk.

Overall, the findings on the interpretability properties of the models are presented in this section to answer the following questions: "What is the impact of policyholder characteristics (e.g., age, occupation) on the lapse rate, and can these characteristics be used to predict lapse risk? ".

The top 40 variables in terms of relevance for the model Random is presented in the figure 4.27. It is clear that the category one of occupation (*COD_OCCUPATION_MAIN_lev_x_1*) feature is the most important, followed by the Gender flag variable of missing values (*OD_GENDER_lev_NA*) and main occupation (*COD_STATUS_OCCUPATION_lev_NA*). Additionally, the figure (C.4) shows that the least relevant feature is *OD_PRODUCT_10*.

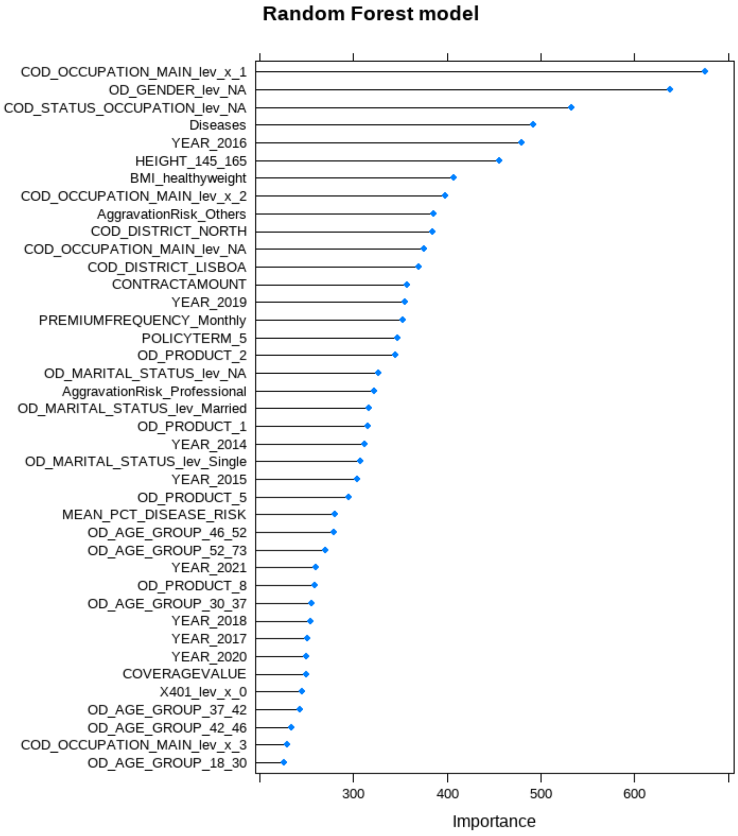
This suggests that some information captured in the application is not necessary for underwriting. And that the most important predictor identified by the model is the category one of the main occupation of the policyholder, which includes occupations such as geologists, actuaries, and chemical engineers. Generally, the predictors identified as most important can be grouped into four categories:

- features related to the contract content (such as occupation or the total policy cost amount) , for example, the features related to occupation, or the total policy cost amount contractually agreed (*CONTRACTAMOUNT*);
- health conditions, for example, the feature "Diseases" which defines whether the policyholder has or had a disease;
- time-related influences (such as age and remaining policy period). For instance, the feature *YEAR_2016*;
- the quality of the information used in the collection system (such as missing values). Represented by the flag features, for instance, *OD_GENDER_lev_NA*.

The survey questions were found to be important predictors of Lapse classification, particularly questions related to diseases, height, and district residence. Random Forest also ranked the category zero of the $X401$ question (see 4.3.2.39) as an important predictor, while C5.0 identified the category one of the $X105$ question (see 4.3.2.6) as the most important predictor. Moreover, an intriguing discovery in the dataset is that pulmonary system-related diseases have a stronger correlation with lapsation than other medical conditions, for instance cancer or cardiovascular diseases. This implies that further research is needed to fully comprehend the health and risk implications of these diseases.

Additionally, product two was found to be an important indicator of lapsation, followed by product one, while product ten was considered a weak indicator of lapse. The categories of the BMI feature were generally ranked as important by all feature selection methods, though several insurers do not include this feature in their application process. This warrants further investigation.

Figure 4.27: Random Forest model Variable Importance (Top 40). **Source:** Authors.



4.12 Extreme Business Scenario

A lapse extreme scenario in life insurance for the insurer would be a significant increase in the number of policyholders who cancel their policies or allow them to lapse within a short period of time. This could be caused by a number of factors, such as a change in economic conditions that affects the policyholders' ability to pay their premiums, or a lack of understanding of the benefits and value of the policies among policyholders. In this scenario, the insurer would experience a sharp decline in premium income and investment returns, as well as an increase in costs associated with the administration of the terminated policies. The financial impact on the insurer would be significant, and it could put a strain on the company's overall financial performance.

Additionally, this scenario would also mean that the insurer would lose out on the future premium income and investment returns that would have been earned on the lapsed policies, exacerbating the financial impact. Furthermore, in this scenario, if a large number of policyholders die within the grace period, the insurer would have to pay out a significant amount of death benefits to the beneficiaries, which would put even more pressure on the insurer's financials.

As a result, the insurer would need to take prompt action to mitigate the impact of the lapses, such as offering more flexible premium payment options or providing better customer service and education to policyholders to help them understand the value and benefits of their policies.

To determine the scenario financial impact of policy lapses on the insurer, we compared the coverage amount (potential loss) and contract accumulation amount (income) of 1863 lapsed policies, with a coverage gap of $76,877,723.00$ euros, using the testing dataset. This comparison accurately determined the financial impact on the insurer's financial performance by the coverage gap.

Table 4.53: Extreme financial scenario. **Source:** Authors.

	Actual LapseRate	Contract Amount(Total)	Coverage Amount(Total)	**Actual Coverage Gap**
Test Actual Values	20.71%	€ 163,130,172.00	€ 86,252,449.00	€ 76,877,723.00

The Table (4.54) shows and reforces that the results of this study indicate that all models performed well, however, the Random Forest model had the most accurate coverage gap compared to the other models when considering the financial impact on the insurance company. This is because it had the highest accuracy rate (84.82%) in predicting Lapse Risk. Thus, Random Forest is the most appropriate for predicting lapse rates in this study dataset. On the other hand, the Naive Bayes model performed poorly in this regard. Additionally, tree-based models generally performed better than other models, with the Bagging method having a better financial impact compared to boosting methods.

In conclusion, Random Forest can offer a valuable tool for insurance companies to anticipate policy lapse risk and coverage gaps, allowing them to proactively address potential issues and ensure policyholder financial security. This can benefit both the insurance provider and policyholders by reducing administrative costs and ensuring adequate coverage. By leveraging this technology, the industry can work towards providing better and more informed life insurance coverage for all.

Table 4.54: Prediction Extreme financial scenario. **Source:** Authors.

Model	**Predicted Lapse Rate**	Contract Amount (Total)	Coverage Amount (Total)	**Predict Coverage Gap**	Difference (Actual coverage gap - predicted coverage gap)
Random Forest (rf)	18.39%	€ 148,281,243	€ 74,503,545	€ 73,777,698	€ 3,100,025
BaggingClassifier	17.84%	€ 142,334,743	€ 70,627,545	€ 71,707,198	€ 5,170,525
C5.0	17.68%	€ 143,561,743	€ 72,482,045	€ 71,079,698	€ 5,798,025
adaboost	19.67%	€ 161,107,120	€ 78,349,120	€ 82,758,000	€ 5,880,277
XGBoostLinearBooster	14.65%	€ 120,921,788	€ 58,462,788	€ 62,459,000	€ 14,418,723
KNN	23.39%	€ 183,241,751	€ 91,764,158	€ 91,477,593	€ 14,599,870
Ensemble model(knn+nb)	25.44%	€ 199,899,751	€ 101,985,633	€ 97,914,118	€ 21,036,395
XGBoostTreeBooster	13.25%	€ 106,002,665	€ 53,337,665	€ 52,665,000	€ 24,212,723
rpart	29.85%	€ 230,565,628	€ 107,454,762	€ 123,110,866	€ 46,233,143
Logistic Regression (LR)	30.94%	€ 246,314,743	€ 118,587,285	€ 127,727,458	€ 50,849,735
Neural Networks (NN)	32.60%	€ 259,208,643	€ 126,901,752	€ 132,306,891	€ 55,429,168
glmnet	30.08%	€ 245,047,143	€ 112,594,118	€ 132,453,025	€ 55,575,302
Naive Bayes (nb)	2.26%	€ 18,869,700	€ 10,364,700	€ 8,505,000	€ 68,372,723

5 | Conclusions

The aim of this study was to investigate the factors that influence lapse rate in life insurance and to develop prediction models that can accurately predict lapse risk. To achieve this goal, a dataset consisting of policyholder information and survey responses was collected and analysed.

The study addressed the following research questions: ”What are the key factors that impact lapse rates in life insurance, and how can they be predicted?” and ”What is the effect of policyholder characteristics on lapse rates, and can these characteristics be used to predict lapse risk?”. The study findings indicated that several factors, including policyholder characteristics, occupation, contract content, health conditions, and time-related influences, can impact lapse rates. Additionally, the quality of information collected during the application process can influence lapse risk. By utilizing predictive modelling techniques, it is possible to identify patterns and trends in customer behaviour that may indicate a higher likelihood of lapse. The study also identified that certain insurance products were more likely to result in lapses than others, in particular product two.

To address the issue of lapse risk prediction, it was developed and compared thirteen machine learning algorithms utilizing various techniques such as imputation methods, feature selection, and resampling techniques. The study found that the Random Forest model was the most effective approach for predicting lapse risk, followed by Bagged Classification Tree, C5.0, and Boosting models. The study also found that the combination of random forest imputation and SMOTE resampling performed better than other methods, and that supervised feature selection was more effective than unsupervised feature selection in this case.

In addition to other objectives, the study addressed the following questions: ”Can machine learning techniques be used effectively to predict lapse rates in life insurance?” and ”How do different machine learning algorithms perform in predicting lapse rates, and which one is most accurate?”. The study found that tree-based models were the most effective approach for predicting lapse rates in the sample of life insurance policies. Furthermore, the study found that fine-tuning parameters, utilizing model bagging, and model boosting could enhance the accuracy of these predictions. Specifically, the study found that model bagging was more effective than model boosting in the sample studied.

The study's findings have several implications for both practice and future research. Firstly, the models developed in this study can be utilized by insurance companies to identify policyholders who are at high risk of lapsing and take proactive measures to retain them. Secondly, the study highlights the effectiveness of using predictive modelling techniques such as the Random Forest model to identify patterns and trends in customer behaviour that may indicate an increased likelihood of lapse or churn. Thirdly, the study emphasizes the importance of considering factors such as policyholder characteristics, contract content, health conditions, and time-related influences when developing prediction models for lapse risk. Fourthly, the results suggest that insurance companies should consider collecting more detailed information about policyholder characteristics, such as occupation and health conditions, to improve their risk assessment and underwriting processes. Lastly, future research could explore other factors that may influence lapse rates, such as customer satisfaction, and the availability of competitive options from other insurers.

Among its research objectives, the study included the investigation of the following questions: ”What are the limitations of current approaches to predicting lapse rates in life insurance, and what opportunities are there for further research in this area?” and ”How can lapse rate prediction models be used by insurance companies to improve their risk management and pricing strategies?”. The study found that current approaches to predicting lapse rates in life insurance have limitations, such as the limited data available from a single insurance company and the potential for changes in the underlying distribution of data over time to affect the model's accuracy. Nevertheless, the study provides valuable information for the life insurance industry and policyholders.

Insurance companies can use predictive models to identify policyholders at high risk of lapsing and take appropriate actions to retain them, such as targeted marketing efforts and personalized communications. By identifying policyholders who are at high risk of lapsing, insurers can design personalized communications that address the unique needs and concerns of those policyholders, increasing the likelihood that they will continue their coverage. This may include tailored offers or incentives that encourage policyholders to stay with their current insurer.

In addition to targeted marketing efforts and personalized communications, insurers can also use predictive models to design products and pricing strategies that reduce the likelihood of lapses and improve customer retention. For example, insurers can use data on policyholder characteristics, contract content, and health conditions to identify and address potential sources of dissatisfaction or non-renewal. They can also design products and pricing strategies that better align with the needs and preferences of their customers, such as more flexible coverage options or bundled discounts.

The use of predictive models for lapse rate prediction and risk management can also help insurers improve their financial performance by reducing the costs associated with customer churn and lapses. By identifying policyholders who are at high risk of lapsing and taking proactive measures to retain them, insurers can reduce the costs associated with acquiring new customers and building brand loyalty.

In conclusion, the insights from this study provide valuable information for insurers seeking to improve their risk management and pricing strategies. By leveraging predictive models and other data-driven tools, insurers can better understand and address the unique needs and concerns of their policyholders, leading to more stable and sustainable insurance markets. However, it will be important for insurers to continue to invest in data analytics and other technologies to stay ahead of changing customer needs and preferences, as well as new competitive threats from emerging players in the insurance industry.

Bibliography

Abachi, J. (2018). "FACTORS THAT INFLUENCE PRICING OF LIFE INSURANCE PRODUCTS: A CASE STUDY." Doctoral dissertation. United States International University-Africa.

Abu Alfeilat, H. A., A. B. A. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, and V. B. S. Prasath (2019). "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review." In: *Big Data* 7(4), pp. 221–248. DOI: 10.1089/big.2018.0175.

Agarwal, A., C. Baechle, R. S. Behara, and V. Rao (2016). "Multi method approach to wellness predictive modeling." In: *Journal of Big Data*. DOI: 10.1186/s40537-016-0049-0.

Alzubi, J., A. Nayyar, and A. Kumar (2018). "Machine learning from theory to algorithms: an overview." In: *Journal of physics: conference series*. Vol. 1142. 1. IOP Publishing, p. 012012.

Andridge, R. and R. Little (Apr. 2010). "A Review of Hot Deck Imputation for Survey Non-Response." In: *International statistical review = Revue internationale de statistique* 78, pp. 40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

Araveeporn, A. (2021). "The Higher-Order of Adaptive Lasso and Elastic Net Methods for Classification on High Dimensional Data." In: *Mathematics*. DOI: 10.3390/math9101091. URL: https://www.mdpi.com/2227-7390/9/10/1091.

Ayodele, T. O. (2021). *New Advances in machine learning*. University of Portsmouth.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011). "Multiple imputation by chained equations: what is it and how does it work?" In: *International Journal of Methods in Psychiatric Research* 20(1), pp. 40–49. DOI: 10.1002/mpr.329.

Barsotti, F., X. Milhaud, and Y. Salhi (2016). "Lapse risk in life insurance: correlation and contagion effects among policyholders' behaviors." In: *Insurance: Mathematics and Economics* 71, pp. 317–331. DOI: 10.1016/j.insmatheco.2016.09.008. URL: https://hal.archives-ouvertes.fr/hal-01282601v2/.

Batty, M., A. Tripathi, A. Kroll, C.-s. P. Wu, D. Moore, C. Stehno, L. Lau, J. Guszcza, and M. Katcher (2010). "Predictive modeling for life insurance, ways life insurers can participate in the business analytics revolution." In: *Deloitte Consulting LLP*.

Beers, B. (2021). "Annual Renewable Term – ART Insurance Definition." In: *Investopedia*. URL: https://www.investopedia.com/terms/a/annual_renewable_term.asp.

Bentéjac, C. et al. (2020). "A comparative analysis of gradient boosting algorithms." In: *Artificial Intelligence Review*, pp. 1–31.

Berrar, D. (Jan. 2018). "Cross-Validation." In: ISBN: 9780128096338. DOI: 10.1016/B978-0-12-809633-8.20349-X.

Biagini, F., T. Huber, J. Jaspersen, and A. Mazzon (2021). "Estimating extreme cancellation rates in life insurance." In: *Journal of Risk and Insurance* 88. DOI: 10.1111/jori.12336.

Biddle, R., S. Liu, P. Tilocca, and G. Xu (2018). "Automated underwriting in life insurance: Predictions and optimisation." In: *Australasian Database Conference*. Springer, pp. 135–146.

Black, K. and H. D. Skipper (2000). *Life and Health Insurance*. Pearson.

Blier-Wong C., C.-H. L. L. and E. Marceau (2021). "Machine Learning in PC Insurance: A Review for Pricing and Reserving." In: *Risks*.

Bolancé, C., M. Guillen, and A. E. Padilla-Barreto (2016). "Predicting probability of customer churn in insurance." In: *International Conference on Modeling and Simulation in Engineering, Economics and Management*. Springer, pp. 82–91.

Boodhun, N. and M. Jayabalan (2018). "Risk prediction in life insurance industry using supervised learning algorithms." In: *Complex Intell. Syst.* 4, pp. 145–154. DOI: 10.1007/s40747-018-0072-1.

Boyd, K., K. H. Eng, and C. D. Page (2013). "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals." In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by H. Blockeel, K. Kersting, S. Nijssen, and F. Železný. Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 451–466. ISBN: 978-3-642-40994-3. DOI: 10.1007/978-3-642-40994-3_29.

Brackenridge, R. D. C., R Croxson, and R. Mackenzie (2016). *Brackenridge's Medical Selection of Life Risks*. Springer.

Breiman, L. (1996). "Bagging predictors." In: *Machine Learning* 24, pp. 123–140. DOI: https://doi.org/10.1007/BF00058655.

Breiman, L. (2001). "Random Forests." In: *Machine Learning* 45(1), pp. 5–32.

Brownlee, J. (2018). *A Gentle Introduction to the Bootstrap Method*. A Gentle Introduction to the Bootstrap Method (machinelearningmastery.com). visited on 19/10/2022.

Burez, J. and D. Van den Poel (2009). "Handling class imbalance in customer churn prediction." In: *Expert Systems with Applications* 36(3, Part 1), pp. 4626–4636. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.05.027. URL: https://www.sciencedirect.com/science/article/pii/S0957417408002121.

Burgette, L. and J. Reiter (2010). "Multiple imputation for missing data via sequential regression trees." In: *American Journal of Epidemiology* 172(9), pp. 1070–1076.

Burri Rama Devi, B.-R. B. R. R. and S. R. Buruga (April 2019). "Insurance Claim Analysis Using Machine Learning Algorithms." In:

Buuren, S. van and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R." In: *Journal of Statistical Software* 45(3), pp. 1–67. URL: http://www.jstatsoft.org/.

Bykerk, C., W. Cutlip, L. Nathan, G. Perrott, W. Reimert, L. Sher, et al. (2005). "Risk Classification (for All Practice Areas)." In:

Bärtl, M. and S. Krummaker (2020). "Prediction of Claims in Export Credit Finance: A Comparison of Four Machine Learning Techniques." In: *MDPI*.

Cao, X. H., I. Stojkovic, and Z. Obradovic (2016). "A robust data scaling algorithm to improve classification accuracies in biomedical data." In: *BMC Bioinformatics* 17(1).

CEIOPS (2009). *CEIOPS' Advice for Level 2 Implementing Measures on Solvency II - Standard formula SCR - Article 109 c Life underwriting risk*. URL: https://register.eiopa.europa.eu/CEIOPS-Archive/Documents/Advices/CEIOPS-L2-Final-Advice-on-Standard-Formula-Life-underwriting-risk.pdf.

Chandola, V., A. Banerjee, and V. Kumar (2009). "Anomaly detection: A survey." In: *ACM computing surveys (CSUR)* 41(3), pp. 1–58.

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

Chatterjee, S., M. S. Chatterjee, I. Rcpp, and L. Rcpp (2016). *Package "fastAdaboost"*.

Chen, T. and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. DOI: 10.1145/2939672.2939785.

Chen, T., T. He, M. Benesty, and V. Khotilovich (2019). "Package 'xgboost'." In: *R version* 90, pp. 1–66.

Chowdhury, M. and T. Turin (2020). "Variable selection strategies and its importance in clinical prediction modelling." In: *Family Medicine and Community Health* 8(1). DOI: 10.1136/fmch-2019-000262.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Erlbaum: Hillsdale, NJ.

Cousy, H. A. (2008). "The Principles of European Insurance Contract Law: the duty of disclosure and the aggravation of risk." In: vol. 9. 1. Springer, pp. 119–132.

Cox, S. H., Y. Lin, R. Tian, and L. F. Zuluaga (2013). "Mortality Portfolio Risk Management." In: *The Journal of Risk and Insurance* 80(4), pp. 853–890.

Cunningham, P. and S. J. J. Delany (2021). "k-Nearest neighbour classifiers - A tutorial." In: *ACM Computing Surveys (CSUR)* 54(6), pp. 1–25.

Deckert, A. C. and E. Kummerfeld (2019). "Investigating the effect of binning on causal discovery." In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. DOI: 10.1109/bibm47256.2019.8983336. URL: https://doi.org/10.1109\%2Fbibm47256.2019.8983336.

Denuit M., C. A. and J. Trufin (2021). "Autocalibration and tweedie-dominance for insurance pricing with machine learning." In: *Insurance: Mathematics and Economics* 101, pp. 485–497.

Diana, A., J. E. Griffin, J. Oberoi, and J. Yao (2019). *Machine-Learning Methods for Insurance Applications: A Survey*. Society of Actuaries: Schaumburg.

Doerry, N. and M. Sibley (2015). "Monetizing Risk and Risk Mitigation." In: *Naval Engineers Journal* 127(3), pp. 35–46.

Domingues, M., P. Filippone, P. Michiardi, and J. Zouaoui (2018). "A comparative evaluation of outlier detection algorithms: Experiments and analyses." In: *Pattern Recognition*. Vol. 74, pp. 406–421.

Eikenhout, L. (2015). "Risk Management and Performance in Insurance Companies." Doctoral dissertation. University of Twente.

EIOPA (2020). *REPORTING AND DISCLOSURE: QUANTITATIVE REPORTING TEMPLATES*. URL: https://www.eiopa.europa.eu/sites/default/files/solvency_ii/eiopa-bos-20-754-quantitative-reporting-templates.pdf.

Eling and Kochanski (2012). "RESEARCH ON LAPSE IN LIFE INSURANCE – WHAT HAS BEEN DONE AND WHAT NEEDS TO BE DONE?" In:

European insurance, C. of and O. P. S. [CEIOPS] (2008). *CEIOPS' Report on its fourth Quantitative Impact Study (QIS4) for Solvency II*. URL: https://www.finanstilsynet.dk/upload/finanstilsynet/mediafiles/newmedia/solvens/solvens%20ii/ceiops_82_08_qis4report.pdf.

Européen, G. C. A. (2011). *Use of age disability as rating factors in insurance.*

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer-Verlag.

Fang, H. and E. Kung (2012). *Why Do Life Insurance Policyholders Lapse? The Roles of Income, Health and Bequest Motive Shocks*. Working Paper 17899. National Bureau of Economic Research. DOI: 10.3386/w17899. URL: http://www.nber.org/papers/w17899.

Fauzan, M. A. and H. Murfi (2018). "The accuracy of XGBoost for insurance claim prediction." In: *International Journal of Advances in Soft Computing Its Applications* 10, pp. 159–171.

Ferrario, A., A. Noll, and M. V. Wuthrich (2018). *Insights from Inside Neural Networks*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3226852.

Fogelson, S. (2022). "Extreme Gradient Boosting with XGBoost." In: Available: 23/10/2022. URL: https://app.datacamp.com/learn/courses/extreme-gradient-boosting-with-xgboost.

Fong, J. H. (2015). "Beyond Age and Sex: Enhancing Annuity Pricing." In: *The Geneva Risk and Insurance Review* 40(2), pp. 133–170.

Fontinelle, A. (2021). "Life Insurance Guide to Policies and Companies." In: *Investopedia*. URL: https://www.investopedia.com/terms/l/lifeinsurance.asp.

Frees, E. (2013). “Relative Importance of Risk Sources in Insurance Systems.” In: *North American Actuarial Journal*, pp. 34–49.

García, S., J. Luengo, and F. Herrera (2015). *Data Preprocessing in Data Mining*. Springer International Publishing Switzerland.

Gatzert, N., G. Hoermann, and H. Schmeiser (2009). “The impact of the secondary market on life insurers’ surrender profits.” In: *Journal of Risk and Insurance* 76(4), pp. 887–908.

Gatzert, N. and H. Wesker (2012). “A comparative assessment of Basel II/III and Solvency II.” In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 37(3), pp. 539–570.

Gelman, A. (2010). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, pp. 529–544. DOI: 10.1017/cbo9780511790942.031.

Goldstein, M. and S. Uchida (2016). “A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data.” In: *PloS one* 11(4), e0152173.

Grice, J. W. (2013). “Pearson’s correlation coefficient.” In: *Oklahoma State University*.

Grize, Y.-L., W. Fischer, and C. Lützelschwab (2020). “Machine learning applications in nonlife insurance.” In: *Applied Stochastic Models in Business and Industry*.

Groll, A., C. Wasserfuhr, and L. Zeldin (2022). “Churn modeling of life insurance policies via statistical and machine learning methods–Analysis of important features.” In: *arXiv preprint arXiv:2202.09182*.

Guido, A. et al. (2016). *Introduction to Machine Learning with Python*.

Gupta, A., K. G. Mehrotra, and C. Mohan (2010). “A clustering-based discretization for supervised learning.” In: *Statistics Probability Letters* 80(9-10), pp. 816–824. DOI: 10.1016/j.spl.2010.01.015. URL: https://www.sciencedirect.com/science/article/pii/S0167715210000271.

Haar, L., K. Anding, K. Trambitckii, and G. Notni (Jan. 2019). “Comparison between Supervised and Unsupervised Feature Selection Methods.” In: pp. 582–589. DOI: 10.5220/0007385305820589.

Haiss, P. and K. Sumegi (2008). “The relationship between insurance and economic growth in Europe: a theoretical and empirical analysis.” In: *Empirica* 35(4), pp. 405–431.

Haiss, P. and K. Sümegi (2008). “The Relationship of Insurance and Economic Growth - A Theoretical and Empirical Analysis.” In: *Empirica* 35(4), pp. 405–431.

Hanafy, M. and R. Ming (2022). “Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study.” In: *APPLIED ARTIFICIAL INTELLIGENCE* 36(1). DOI: 10.1080/08839514.2021.2020489.

Handoyo, S., Y. ping Chen, G. Irianto, and A. Widodo (2021). “The Varying Threshold Values of Logistic Regression and Linear Discriminant for Classifying Fraudulent Firm.” In: *Mathematics and Statistics*.

Harrington, P. d. B. (2018). "Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes." In: *Crit Rev Anal Chem* 48(1), pp. 33–46. ISSN: 1547-6510. DOI: 10.1080/10408347.2017.1361314.

He, H. and E. A. Garcia (2009). "Learning from imbalanced data." In: *IEEE Transactions on knowledge and data engineering* 21(9), pp. 1263–1284.

Hendren, N. (2013). "Private Information and Insurance Rejections." In: *Econometrica* 81(5), pp. 1713–1762.

Hidalgo, H., M. Chipulu, and U. Ojiako (2013). "Risk Segmentation in Chilean social health insurance." In: *International Journal of Health Care Quality Assurance*.

Hoerl, A. E. and R. W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems." In: *Technometrics* 12(1), pp. 55–67.

Holloman, A. (2021). *Machine Learning Toolbox*.

Hopkin, P. (2018). *Fundamentals of risk management: understanding, evaluating and implementing effective risk management*. 4th ed. Kogan Page Publishers.

Hossain, M. R. and D. Timmer (2021). "Machine Learning Model Optimization with Hyper Parameter Tuning Approach." In: *Global Journal of Computer Science and Technology* 21.

Hull, J. (2018). *Risk management and financial institutions*. John Wiley & Sons.

Hutagaol, B. J. and T. Mauritsius (2020). "Risk level prediction of life insurance applicant using machine learning." In: *International Journal of Advanced Trends in Computer Science and Engineering* 9(2).

Hutcheson, G. D. (2011). "Dummy Variable Coding." In: *The SAGE Dictionary of Quantitative Management Research*. Ed. by L. Moutinho and G. D. Hutcheson, pp. 98–102.

Insurance, E. and O. P. Authority (2011). *Report on the fifth Quantitative Impact Study (QIS5) for Solvency II*.

Iso, I et al. (2009). "Risk management–Principles and guidelines." In: *International Organization for Standardization, Geneva, Switzerland*.

J, N., D. P, and H. W (2015). "Attrition in developmental psychology: a review of modern missing data reporting and practices." In: *Int J Behav Dev* 41, pp. 143–153.

Kagan, J. (2021). "Insurance Risk Class Definition and Associated Premium Costs." In: *Investopedia*. URL: https://www.investopedia.com/terms/i/insurance-risk-class.asp.

Kagan, J. (2022). "Insurance: Definition, How It Works, and Main Types of Policies." In: *Investopedia*. URL: https://www.investopedia.com/terms/i/insurance.asp.

Kassambara, A. (2017). *Practical Guide to Principal Component Methods in R; Multivariate Analysis*.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, ..., and R. C. Team (2020). "Package 'caret'." In: *The R Journal* 223, p. 7.

Kuhn, M. and K. Johnson (2013a). *Applied Predictive Modeling*. Springer. DOI: 10.1007/978-1-4614-6849-3_1.

Kuhn, M. and K. Johnson (2013b). *Applied Predictive Modeling*. Springer Science+Business Media New York. DOI: 10.1007/978-1-4614-6849-3_11.

Kuhn, M. and K. Johnson (2013c). *Applied Predictive Modeling*. Springer. DOI: 10.1007/978-1-4614-6849-3_14.

Kuhn, M. and K. Johnson (2013d). *Applied Predictive Modeling*. Springer. DOI: 10.1007/978-1-4614-6849-3_8.

Kuhn, M. and K. Johnson (2013e). *Applied Predictive Modeling*. Springer Science+Business Media New York. DOI: 10.1007/978-1-4614-6849-3_3.

Kuhn, M. and K. Johnson (2013f). *Applied Predictive Modeling*. Springer Science+Business Media New York. DOI: 10.1007/978-1-4614-6849-3_19.

Kuhn, M. and K. Johnson (2013g). *Applied Predictive Modeling*. Springer Science+Business Media New York. DOI: 10.1007/978-1-4614-6849-3_4.

Kuhn, M. and K. Johnson (2013h). *Applied Predictive Modeling*. Springer Science+Business Media New York. DOI: 10.1007/978-1-4614-6849-3_18.

Kuhn, M., S. Weston, M. Culp, N. Coulter, R. Quinlan, et al. (2015). “Package ‘C50’.” In: *CRAN, UTC*.

Kuo, T. C. (2003). “An empirical study on the lapse rate: the cointegration approach.” In: *Journal of Risk and Insurance* 70(3), pp. 489–508.

Kwon, W. J. (2013). “The significance of regulatory orientation, political stability and culture on consumption and price adequacy in insurance markets.” In: *Journal of Risk Finance*.

Larosière, J. de, L. Balcerowicz, O. Issing, R. Masera, C. Mc Carthy, L. Nyberg, J. Pérez, and O. Ruding (2009). “The high level group on financial supervision in the EU.” In: *Report of the High-Level Group on Financial Supervision in the EU, Brussels* 25.

Laurent, J.-P. (2016). *Modelling in life insurance–a management perspective*. Springer, pp. 41–42.

Lee, S. and K. Antonio (2015). “Why high dimensional modeling in actuarial science?” In: *IACA Colloquia*. Sydney, Australia, pp. 23–27.

Lemmens, A. and C. Croux (2006). “Bagging and boosting classification trees to predict churn.” In: *Journal of Marketing Research* 43(2), pp. 276–286.

Li, C. (2013). “Little’s Test of Missing Completely at Random.” In: *The Stata Journal* 13(4), pp. 795–809. DOI: 10.1177/1536867X1301300407. URL: https://doi.org/10.1177/1536867X1301300407.

Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu (2017). “Feature Selection: A Data Perspective.” In: *ACM Comput. Surv.* 9(4), p. 39. DOI: 10.1145/0000000.0000000.

Li, Y. and W. Chen (2020). “A Comparative Performance Assessment of Ensemble Learning for Credit Scoring.” In: *Mathematics* 8(10). ISSN: 2227-7390. DOI: 10.3390/math8101756. URL: https://www.mdpi.com/2227-7390/8/10/1756.

Lieberman, M. and J. Morris (Jan. 2014). “The Precise Effect of Multicollinearity on Classification Prediction.” In: 40, pp. 5–10.

Lim Jin Xong, H. M. K. (2019). "A Comparison of Classification Models for Life Insurance Lapse Risk." In: *International Journal of Recent Technology and Engineering (IJRTE)*, pp. 245–250.

Little, R. J. A. and D. B. Rubin. (1988). *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). "Isolation forest." In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE, pp. 413–422.

Loisel, S., P. Piette, and C.-H. J. Tsai (2021). "Applying economic measures to lapse risk management with machine learning approaches." In: *ASTIN Bulletin: The Journal of the IAA* 51(3), pp. 839–871.

Ly, A., B. Uthayasooriyar, and T. Wang (2020). *A survey on natural language processing (nlp) and applications in insurance*. DOI: 10.48550/ARXIV.2010.00462. URL: https://arxiv.org/abs/2010.00462.

Madasamy, K. and M. Ramaswami (2017). "Data imbalance and classifiers: Impact and solutions from a big data perspective." In: *International Journal of Computational Intelligence Research* 13(9), pp. 2267–2281.

Maier, M., H. Carlotto, F. Sanchez, S. Balogun, and S. Merritt (2019). "Transforming Underwriting in the Life Insurance Industry." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01), pp. 9373–9380. DOI: 10.1609/aaai.v33i01.33019373. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4985.

Majka, M. and M. M. Majka (2020). *Package 'naivebayes'*.

Makariou D., B. P. and Y. Chen (2021). "A random forest based approach for predicting spreads in the primary catastrophe bond market." In: *Insurance: Mathematics and Economics* 101, pp. 140–162.

Manolache, A. (May 2019). "Stress and scenario tests in the context of a Romanian non-life insurance company." In: *Proceedings of the International Conference on Business Excellence* 13, pp. 149–161. DOI: 10.2478/picbe-2019-0014.

Marano, P. and M. Siri (2017). *Insurance Regulation in the European Union*. Palgrave Macmillan.

Marku, M. and V. Pancaldi (2022). *From time-series transcriptomics to gene regulatory networks: a review on inference methods*. DOI: 10.48550/ARXIV.2210.08542. URL: https://arxiv.org/abs/2210.08542.

Maynard, T., A. Bordon, J. B. Berry, D. B. Baxter, W. Skertic, B. T. Gotch, N. T. Shah, A. N. Wilkinson, S. H. Khare, K. B. Jones, and et al. (2019). *What Role for AI in Insurance Pricing?* URL: https://www.researchgate.net/publication/337110892_WHAT_ROLE_FOR_AI_IN_INSURANCE_PRICING_A_PREPRINT.

McHugh, M. (2012). "Interrater reliability: the kappa statistic." In: *Biochem Med (Zagreb)* 22(3), pp. 276–82.

Milhaud X., S. L. and V. Maume-Deschamps (2011). "Surrender Trigger in Life Insurance: What Main Features Affect the Surrender Behavious in an Economic Context." In: *Bullettin Francais d'Actuariat*(11).

MISHR, K. (2016). *Fundamentals of life insurance theories and applications.* 2nd ed. PHI Learning Pvt. Ltd.: Delhi.

Morissette, L. and S. Chartier (2013). "The k-means clustering technique: General considerations and implementation in Mathematica." In: *Tutorials in Quantitative Methods for Psychology* 9, pp. 15–24. DOI: 10.20982/tqmp.09.1.p015.

MRM, T., M. R, M. I, M. JL, and K. D (2015). *Information security analytics: finding security insights, patterns and anomalies in big data.* Syngress Books, Elsevier. DOI: 10.1016/B978-0-12-800207-0.00001-0.

Mullainathan, S. and J. Spiess (2017). "Machine Learning: An Applied Econometric Approach." In: *Journal of Economic Perspectives*, pp. 87–106.

Nasteski, V. (Dec. 2017). "An overview of the supervised machine learning methods." In: *HORIZONS.B* 4, pp. 51–62. DOI: 10.20544/HORIZONS.B.04.1.17.P05.

Navas-Palencia, G. (2020). *Optimal binning: mathematical programming formulation.* DOI: 10.48550/ARXIV.2001.08025. URL: https://arxiv.org/abs/2001.08025.

Nelder, J. and R. Wedderburn (1972). "Generalized Linear Models." In: *Journal of the Royal Statistical Society: Series A (General)* 135(3), pp. 370–384. DOI: 10.2307/2344614.

Ngueng Feze, I. and Y. Joly (2014). "Can't always get what you want? Try an indirect route you just might get what you need: A study on access to genetic data by Canadian life insurers." In: *Current Pharmacogenics and Personalized Medicine (Formerly Current Pharmacogenomics)* 12(1), pp. 56–64.

Norazian, M. (2013). "Roles of imputation methods for filling the missing values: A review." In: *Advances in Environmental Biology* 7(12), pp. 3861–3869.

Oakes, M., R. Gaizauskas, and H. Fowkes (2001). "A Method Based on the Chi-Square Test for Document Classification." In:

Olivieri, A. and E. Pitacco (2011). *Introduction to insurance mathematics: Technical and financial features of risk transfers.* 2nd ed. Springer. DOI: 10.1007/978-3-642-16029-5.

Parsons, C. (2015). "Insuring the unknown." In: *Human and Experimental Toxicology*, pp. 1238–1244.

Patterson, T. and C. C. S. Executive (2015). "The use of information technology in risk management." In: *Complex Solutions Executive IBM Corporation.*

Pega, F., B. Náfrádi, N. C. Momen, Y. Ujita, K. N. Streicher, A. M. Prüss-Üstün, T. A. Group, A. Descatha, T. Driscoll, F. M. Fischer, et al. (2021). "Global, regional, and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000–2016: A systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury." In: *Environment international* 154, p. 106595.

Piegorsch, W. W. (2015). *Statistical Data Analytics.* University of Arizona, USA, pp. 18–19.

Pitacco, E. (2007). *Mortality and Longevity: a Risk Management Perspective*. University of Trieste: Trieste, Italy.

Probability and S. for Data Science (2020). *Covariance matrix*.

Probst, P., M. N. Wright, and A. L. Boulesteix (2019). "Hyperparameters and tuning strategies for random forest." In: *WIREs Data Mining and Knowledge Discovery* 9, e1301.

Promislow, S. D. (2014). *Fundamentals of Actuarial Mathematics*. John Wiley & Sons Ltd: York University, Toronto, Canada.

Quinlan, J. R. (1996). "Bagging, Boosting, and C4.5." In: *AAAI/IAAI, Vol. 1*.

Ripley, B., W. Venables, and M. B. Ripley (2016). "Package 'nnet'." In: *R package version* 7(3-12), p. 700.

Risk, M. S. and R. F. Soundness (2008). "Global financial stability report." In: *International Monetary Fund, Washington*.

Salman, S., I. Ngueng Feze, and Y. Joly (2016). "Disclosure of insurability risks in research and clinical consent forms." In: *Global Bioethics* 27(1), pp. 38–49.

Salman, S. A. (2015). "Insurance as the backbone of risk Management." In: *International Business Management* 9(1), pp. 54–59.

Salman Saeed, M., M. W. Mustafa, U. U. Sheikh, T. A. Jumani, I. Khan, S. Atawneh, and N. N. Hamadneh (2020). "An Efficient Boosted C5.0 Decision-Tree-Based Classification Approach for Detecting Non-Technical Losses in Power Utilities." In: *Energies* 13(12). ISSN: 1996-1073. DOI: 10.3390/en13123242.

Santos, M., J. Soares, P. Abreu, H. Araujo, and J. Santos (2018). "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches." In: *IEEE Computational Intelligence Magazine* 13(4), pp. 59–76.

Schliep, K., K. Hechenbichler, and M. K. Schliep (2016). "Package 'kknn'." In: *CRAN R Project*.

Schlütter, S. (2014). "Capital Requirements or pricing constraints?: An economic analysis of measures for insurance regulation." In: *The Journal of Risk Finance*, pp. 533–554.

Schmeiser, H. and J. Wagner (2013). "The Impact of Introducing Insurance Guaranty Schemes on Pricing and Capital Structure." In: *Journal of Risk and Insurance* 80(2), pp. 273–308.

Scriney, M., D. Nie, and M. Roantree (Sept. 2020). "Predicting Customer Churn for Insurance Data." In: pp. 256–265. ISBN: 978-3-030-59064-2. DOI: 10.1007/978-3-030-59065-9_21.

Shah, A. D., J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway (2014). "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study." In: *American Journal of Epidemiology* 179(9), pp. 1353–1360. DOI: 10.1093/aje/kwt312.

Shamsuddin, S. N., N. Ismail, and N. F. Roslan (2022). "What We Know about Research on Life Insurance Lapse: A Bibliometric Analysis." In: *Risks* 10(5), p. 97.

Simon, N., J. H. Friedman, T. Hastie, and R. Tibshirani (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." In: *Journal of Statistical Software* 39, pp. 1–13. DOI: 10.18637/jss.v039.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v039i05.

Singh, A., N. Thakur, and A. Sharme (2016). "A Review of Supervised Machine Learning." In: URL: https://ieeexplore.ieee.org/abstract/document/7724478.

Singhal, R. and R. Rana (2015). "Chi-square test and its application in hypothesis testing." In: *Journal of the Practice of Cardiovascular Sciences*. DOI: 10.4103/2395-5414.157577.

Solvency, I. (2009). *Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II)*.

Stekhoven, D. and P. B"uhlmann (2012). "MissForest: non-parametric missing value imputation for mixed-type data." In: *Bioinformatics* 28(1), pp. 112–118.

Stoneburner, G., A. Goguen, A. Feringa, et al. (2002). "Risk management guide for information technology systems." In: *Nist special publication* 800(30), pp. 800–30.

Torgo, L. (2011). *Data Mining with R: Learning with Case Studies*. 1st. Chapman and Hall/CRC. DOI: 10.1201/9780429292859.

Tsai, H.-H., T.-W. Yang, W.-M. Wong, and C.-F. Chou (2022). "A Hybrid Approach for Binary Classification of Imbalanced Data." In: *Taiwan*.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." In: *Journal of Economic Perspectives*, pp. 3–28.

Venables, W. N. and B. D. Ripley (2002). *Modern applied statistics with S*. fourth. Springer.

Vieira, S., H. Proença, and C. Salgado (2016). *Missing data. In: Secondary analysis of electronic health records*. Springer, pp. 1–427. DOI: 10.1007/978-3-319-43742-2.

Vieira, S., U. Kaymak, and J. Sousa (2010). "Cohen's kappa coefficient as a performance measure for feature selection." In: pp. 1–8. DOI: 10.1109/FUZZY.2010.5584447.

Visbal-Cadavid, D., A. Mendoza-Mendoza, and E. J. De La Hoz (2020). "Use of Factorial Analysis of Mixed Data (FAMD) and Hierarchical Cluster Analysis on Principal Component (HCPC) for Multivariate Analysis of Academic Performance of Industrial Engineering Programs." In: *Journal of Southwest Jiaotong University* VL-55. DOI: 10.35741/issn.0258-2724.55.5.34.

Wang, Y. (2021). "Predictive machine learning for underwriting life and health insurance." In: *Actuarial Society of South Africa*.

Wuthrich, M. V. and C. Buser (2019). *Data analytics for non-life insurance pricing*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308.

Xie, W., G. Liang, Z. Dong, B. Tan, and B. Zhang (2019). "An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data." In: *Mathematical Problems in Engineering* 2019, pp. 1–13. DOI: 10.1155/2019/3526539.

Yeoh, J. (2018). "IFRS 17 insurance contracts: A brief history of IFRS 17." In: *IFRS 17 Workshop*.

Ying, X. (2019). "An overview of overfitting and its solutions." In: *Journal of Physics: Conference Series* 1168(2), pp. 1–7.

Zhang, Y., J. Liu, and W. Shen (2022). "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications." In: *Applied Sciences* 12(17). ISSN: 2076-3417. URL: https://www.mdpi.com/2076-3417/12/17/8654.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), pp. 301–320.

A | Appendix 1: Survey Guide

A.1 Group 1: Declaration of Health

A.1.1 Question X101

Is treated or has altered cholesterol (> 220 mg/dl) and/or blood pressure values (values other than 12 or 13/8)?

A.1.2 Question X102

Indicate your current cholesterol values.

A.1.3 Question X103

Do you currently have controlled blood pressure values (values equal to 12/7 or 13/8)?

A.1.4 Question X104

Are you on leave or have you had periods of sick leave or accident of more than 10 days?

A.1.5 Question X105

Are you retired or do you have a process underway for old-age retirement or disability?

A.1.6 Question X106

Have you ever been hospitalized or undergoing surgery due to illness or accident and/or are you waiting for hospitalization or surgery?

A.1.7 Question X107

Indicate the surgical intervention to which you have been subjected or are waiting for.

A.1.8 Question X108

Indicate which surgical interventions.

A.1.9 Question X109

You have already performed or are awaiting results of laboratory tests and/or other complementary diagnostic medical examinations.

A.1.10 Question X110

Indicate that other laboratory tests or complementary diagnostic medical tests have performed or are awaiting results.

A.1.11 Question X111

Indicate your weight in kg (kilogram)

A.1.12 Question X112

Indicate your height in cm (centimetre)

A.2 Group 2: Personal History

A.2.1 Question X201

Indicate whether you have or have suffered from any of the following disorders or diseases (do not consider cold conditions and colds):

Gynecological diseases; Breast diseases; Diseases of the cardiovascular system; Osteoarticular, spinal or rheumatological diseases; Diseases of the respiratory system; Diseases of the genitourinary system; Psychiatric diseases; Neurological diseases; vascular diseases; Infectious diseases; Metabolic or blood diseases; Stomach diseases; Inflammatory diseases; Diseases related to the senses considering only ears or eyes.

A.2.2 Question X202

Specify which cardiovascular diseases you suffer from or have already suffered.

A.2.3 Question X203

Specify which osteoarticular, spinal or rheumatological diseases you suffer from or have already suffered.

A.2.4 Question X204

Specify which respiratory diseases you suffer from or have already suffered.

A.2.5 Question X205

Specify which diseases of the genitourinary system you suffer from or have already suffered from.

A.2.6 Question X206

Specify which psychiatric illnesses you suffer from or have already suffered.

A.2.7 Question X207

Specify which neurological diseases you suffer from or have already suffered.

A.2.8 Question X208

Specify which vascular diseases you suffer from or have already suffered.

A.2.9 Question X209

Specify which infectious diseases you suffer from or have already suffered.

A.2.10 Question X210

Specify which metabolic or blood diseases you suffer from or have already suffered.

A.2.11 Question X211

Specify which diseases of the stomach, inflammatory diseases of the intestine, pancreas or others have already suffered.

A.2.12 Question X212

Specify which diseases are related to the senses you suffer from or have already suffered.

A.2.13 Question X213

Indicate dioptres.

A.2.14 Question X214

Specify which skin diseases you suffer from or have already suffered.

A.2.15 Question X215

Specify the type of tumour or cancer.

A.2.16 Question X216

Specify that other diseases.

A.2.17 Question X217

Specify which breast diseases you suffer from or have already suffered.

A.2.18 Question X218

Specify which Gynecological diseases you suffer from or have already suffered.

A.3 Group 3: Therapeutic

A.3.1 Question X301

Do you take medications regularly (do not consider birth control pill) or have you taken any of the following medications?

For heart or hypertension, Anticoagulants, Insulin, Antidepressants or tranquilizers, Corticosteroids, For tumours or cancer, other medicines regularly.

A.3.2 Question X302

Indicate which medicines for the heart or hypertension you have taken regularly or have already taken.

A.3.3 Question X303

Indicate which anticoagulant medicines you have taken regularly or have already taken.

A.3.4 Question X304

Indicate which antidepressant or tranquilizer medicines you regularly take or have taken.

A.3.5 Question X305

Indicate which corticosteroid medicinal products you take regularly or have already taken.

A.3.6 Question X306

Indicate which other medicines you take regularly.

A.3.7 Question X307

Indicate whether you have undergone any of the following Detoxification, Chemotherapy, Radiotherapy treatments.

A.4 Group 4: Habits

A.4.1 Question X401

Are you a smoker (indicate number of cigarettes/day)?

A.4.2 Question X402

Do you consume more than 15 units of alcohol a week? (1 unit = 1 cup)

A.4.3 Question X403

Do you consume, or have you ever consumed narcotics or drugs?

A.5 Group 5: Travel and Sports

A.5.1 Question X501

Do you play any risky sports?

A.5.2 Question X502

Indicate whether you practice any of the following amateur risk sports.

A.5.3 Question X503

Indicate whether you practice any of the following risky sports in a professional manner.

A.5.4 Question X504

Do you want to move outside the European Union?

A.5.5 Question X505

What kind of activity will you pursue?

A.5.6 Question X506

Indicate which country you want to change your residence to.

A.5.7 Question X507

X507: Will you travel or want to travel to countries outside the European Union other than Switzerland, Norway, USA, Canada, Argentina, Brazil, Japan, Australia, and New Zealand?

A.6 Group 6: Family Protection: Travel and Sports

A.6.1 Question X6011

Do you have some disability or physical defect, suffer from some chronic disease, suffer from sequelae, injuries, or residual symptoms of some disease, you have had labour losses and/or hospitalizations for more than 10 days because of illness or accident in the last 3 years, have you ever undergone any surgical intervention, receive or have in progress any process for receiving any disability pension?

B Appendix 2: Supporting Materials

Table B.1: Answers of the Policyholder one. **Source:** Authors.

ID	ID_PRODUTO	GRUPOID	TIPO	DESCRICAO	DATE_ANS
1	10	1	GRP	1- DECLARAÇÃO DE ESTADO DE SAÚDE	22/02/2014
1	10	1	QST	Indique o seu peso em Kg	22/02/2014
1	10	1	ANS	65	22/02/2014
1	10	1	QST	Indique a sua altura em cm	22/02/2014
1	10	1	ANS	170	22/02/2014
1	10	1	QST	Faz tratamento ou tem valores alterados de Colesterol (> 220 mg/dl)e/ou Tensão Arterial (valores diferentes de 12/7 ou 13/8)?	22/02/2014
1	10	1	ANS	Não	22/02/2014
1	10	1	QST	Está de baixa ou já teve períodos de baixa por doença ou acidente superiores a 10 dias?	22/02/2014
1	10	1	ANS	Não	22/02/2014
1	10	1	QST	É reformado ou tem em curso algum processo para atribuição de reforma por velhice ou invalidez?	22/02/2014
1	10	1	ANS	Não	22/02/2014
1	10	1	QST	Já foi hospitalizado ou sujeito a alguma intervenção cirúrgica por doença ou acidente e/ou aguarda hospitalização ou cirurgia?	22/02/2014
1	10	1	ANS	Não	22/02/2014
1	10	1	QST	Já efetuou ou está a aguardar resultados de testes laboratoriais e/ou outros exames médicos complementares de diagnóstico	22/02/2014
1	10	1	ANS	Não	22/02/2014
1	10	2	GRP	2 - ANTECEDENTES PESSOAIS	22/02/2014
1	10	2	QST	Indique se sofre ou já sofreu de alguma das seguintes perturbações ou doenças: (não considerar estados gripais e constipações)	22/02/2014
1	10	2	ANS	Nenhuma das doenças indicadas	22/02/2014
1	10	3	GRP	3 - TERAPÊUTICAS	22/02/2014
1	10	3	QST	Toma medicamentos regularmente (não considerar pílula anticoncecional) ou já tomou algum dos seguintes medicamentos?	22/02/2014
1	10	3	ANS	Nenhum	22/02/2014
1	10	3	QST	Indique se já foi submetido a algum dos seguintes tratamentos	22/02/2014
1	10	3	ANS	Nenhum dos tratamentos indicados	22/02/2014
1	10	4	GRP	4 - HÁBITOS	22/02/2014
1	10	4	QST	É fumador (indique nº de cigarros/dia)?	22/02/2014
1	10	4	ANS	Não	22/02/2014
1	10	4	QST	Consome mais do que 15 unidades de álcool por semana? (1 unidade = 1 copo)	22/02/2014
1	10	4	ANS	Não	22/02/2014
1	10	4	QST	Consome ou já consumiu estupefacientes ou drogas?	22/02/2014
1	10	4	ANS	Não	22/02/2014
1	10	5	GRP	5 - VIAGENS E DESPORTOS	22/02/2014
1	10	5	QST	Pratica algum desporto de risco?	22/02/2014
1	10	5	ANS	Não	22/02/2014
1	10	5	QST	Viaja ou pretende viajar para países fora da União Europeia, que não sejam a Suíça, Noruega, EUA, Canadá, Argentina, Brasil, Japão, Austrália e Nova Zelândia?	22/02/2014
1	10	5	ANS	Não	22/02/2014

Table B.2: Question X107 (A.1.7). **Source:** Authors.

Code	Category	Response	Count
			34921
7	DOENCAS DO ESTOMAGO, DOENÇAS INFLAMATORIAS DO INTESTINO, DO PANCREAS OU OUTRAS	BANDA GASTRICA	12
14	DOENCAS RELACIONADAS COM OS SENTIDOS	CATARATAS	7
2	DOENCAS DA MAMA	CIRURGIA DA MAMA	36
12	DOENCAS OSTEOARTICULARES, DA COLUNA VERTEBRAL OU REUMATOLOGICAS	HERNIAS DISCAIS	75
8	DOENCAS GINECOLOGICAS	HISTERECTOMIA TOTAL OU PARCIAL	29
12	DOENCAS OSTEOARTICULARES, DA COLUNA VERTEBRAL OU REUMATOLOGICAS	MENISCO	49
14	DOENCAS RELACIONADAS COM OS SENTIDOS	MIOPIA	16
1	OUTRAS	OUTRAS	757
5	DOENCAS DO APARELHO GENITO-URINARIO	PROSTATA	1
15	DOENCAS RELACIONADAS COM TUMORES OU QUALQUER TIPO DE CANCRO	QUISTOS	72
12	DOENCAS OSTEOARTICULARES, DA COLUNA VERTEBRAL OU REUMATOLOGICAS	ROTURA DE LIGAMENTOS	41
10	DOENCAS METABOLICAS OU DO SANGUE	TIROIDE	49

Table B.3: Question X502 (A.5.2)

Code	Response	Count
		34480
1	ACROBACIAS AEREAS A	78
1	ARTES MARCIAIS A	25
1	ASA DELTA PLANADOR COM MOTOR E PARAPENTE A	1
1	AUTOMOBILISMO A	4
1	AVIACAO PRIVADA A	2
1	BOXE OU KICHBOXING A	9
1	CANYONING A	2
1	HIPISMO A	7
1	JET SKI A	1
1	MERGULHO E CACA SUBMARINA A	5
1	MONTANHISMO ALPINISMO OU ESCALADA A	1
1	MOTOCICLISMO A	18
1	MOTONAUTICA A	1
1	PARAQUEDISMO A	1
1	PARKOUR A	1
1	VARIAS A	26
0	NENHUM DOS ASSINALADOS A	1403

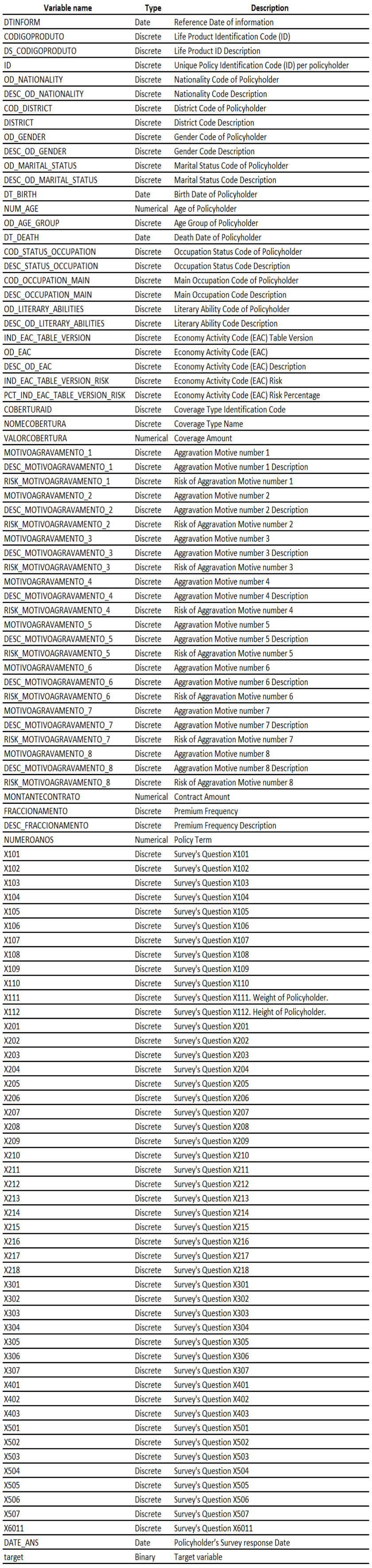
Table B.4: Question X503 (A.5.3)

Code	Response	Count
		36023
1	ACROBACIAS AEREAS P	6
1	ARTES MARCIAIS P	2
1	AUTOGIROS OU GIROPLANOS P	1
1	AUTOMOBILISMO P	7
1	BOXE OU KICHBOXING P	1
1	BUNGEE JUMPING P	2
1	CANYONING P	2
1	CICLISMO P	2
1	HOQUEI OU PATINAGEM NO GELO P	1
1	VARIAS P	1
0	NENHUM DOS ASSINALADOS P	17

Table B.5: Question X506 (A.5.6)

Code	Response	Count
		35902
ZAF	AFRICA DO SUL	3
AGO	ANGOLA	55
BRA	BRASIL	5
CHN	CHINA	4
ARE	EMIRATOS ARABES UNIDOS	4
USA	EUA	10
GNB	GUINE BISSAU	2
IND	INDIA	7
MOZ	MOCAMBIQUE	5
1	OUTRO PAIS	45
CHE	SUICA	19
VEM	VENEZUELA	4

Figure B.1: Dataset I. **Source:** Authors.



Variable name	Type	Description
DTINFORM	Date	Reference Date of information
CODIGOPRODUTO	Discrete	Life Product Identification Code (ID)
DS_CODIGOPRODUTO	Discrete	Life Product ID Description
ID	Discrete	Unique Policy Identification Code (ID) per policyholder
OD_NATIONALITY	Discrete	Nationality Code of Policyholder
DESC_OD_NATIONALITY	Discrete	Nationality Code Description
COD_DISTRICT	Discrete	District Code of Policyholder
DISTRICT	Discrete	District Code Description
OD_GENDER	Discrete	Gender Code of Policyholder
DESC_OD_GENDER	Discrete	Gender Code Description
OD_MARITAL_STATUS	Discrete	Marital Status Code of Policyholder
DESC_OD_MARITAL_STATUS	Discrete	Marital Status Code Description
DT_BIRTH	Date	Birth Date of Policyholder
NUM_AGE	Numerical	Age of Policyholder
OD_AGE_GROUP	Discrete	Age Group of Policyholder
DT_DEATH	Date	Death Date of Policyholder
COD_STATUS_OCCUPATION	Discrete	Occupation Status Code of Policyholder
DESC_STATUS_OCCUPATION	Discrete	Occupation Status Code Description
COD_OCCUPATION_MAIN	Discrete	Main Occupation Code of Policyholder
DESC_OCCUPATION_MAIN	Discrete	Main Occupation Code Description
OD_LITERARY_ABILITIES	Discrete	Literary Ability Code of Policyholder
DESC_OD_LITERARY_ABILITIES	Discrete	Literary Ability Code Description
IND_EAC_TABLE_VERSION	Discrete	Economy Activity Code (EAC) Table Version
OD_EAC	Discrete	Economy Activity Code (EAC)
DESC_OD_EAC	Discrete	Economy Activity Code (EAC) Description
IND_EAC_TABLE_VERSION_RISK	Discrete	Economy Activity Code (EAC) Risk
PCT_IND_EAC_TABLE_VERSION_RISK	Discrete	Economy Activity Code (EAC) Risk Percentage
COBERTURAID	Discrete	Coverage Type Identification Code
NOMECOBERTURA	Discrete	Coverage Type Name
VALORCOBERTURA	Numerical	Coverage Amount
MOTIVOAGRAVAMENTO_1	Discrete	Aggravation Motive number 1
DESC_MOTIVOAGRAVAMENTO_1	Discrete	Aggravation Motive number 1 Description
RISK_MOTIVOAGRAVAMENTO_1	Discrete	Risk of Aggravation Motive number 1
MOTIVOAGRAVAMENTO_2	Discrete	Aggravation Motive number 2
DESC_MOTIVOAGRAVAMENTO_2	Discrete	Aggravation Motive number 2 Description
RISK_MOTIVOAGRAVAMENTO_2	Discrete	Risk of Aggravation Motive number 2
MOTIVOAGRAVAMENTO_3	Discrete	Aggravation Motive number 3
DESC_MOTIVOAGRAVAMENTO_3	Discrete	Aggravation Motive number 3 Description
RISK_MOTIVOAGRAVAMENTO_3	Discrete	Risk of Aggravation Motive number 3
MOTIVOAGRAVAMENTO_4	Discrete	Aggravation Motive number 4
DESC_MOTIVOAGRAVAMENTO_4	Discrete	Aggravation Motive number 4 Description
RISK_MOTIVOAGRAVAMENTO_4	Discrete	Risk of Aggravation Motive number 4
MOTIVOAGRAVAMENTO_5	Discrete	Aggravation Motive number 5
DESC_MOTIVOAGRAVAMENTO_5	Discrete	Aggravation Motive number 5 Description
RISK_MOTIVOAGRAVAMENTO_5	Discrete	Risk of Aggravation Motive number 5
MOTIVOAGRAVAMENTO_6	Discrete	Aggravation Motive number 6
DESC_MOTIVOAGRAVAMENTO_6	Discrete	Aggravation Motive number 6 Description
RISK_MOTIVOAGRAVAMENTO_6	Discrete	Risk of Aggravation Motive number 6
MOTIVOAGRAVAMENTO_7	Discrete	Aggravation Motive number 7
DESC_MOTIVOAGRAVAMENTO_7	Discrete	Aggravation Motive number 7 Description
RISK_MOTIVOAGRAVAMENTO_7	Discrete	Risk of Aggravation Motive number 7
MOTIVOAGRAVAMENTO_8	Discrete	Aggravation Motive number 8
DESC_MOTIVOAGRAVAMENTO_8	Discrete	Aggravation Motive number 8 Description
RISK_MOTIVOAGRAVAMENTO_8	Discrete	Risk of Aggravation Motive number 8
MONTANTECONTRATO	Numerical	Contract Amount
FRACCIONAMENTO	Discrete	Premium Frequency
DESC_FRACCIONAMENTO	Discrete	Premium Frequency Description
NUMEROANOS	Numerical	Policy Term
X101	Discrete	Survey's Question X101
X102	Discrete	Survey's Question X102
X103	Discrete	Survey's Question X103
X104	Discrete	Survey's Question X104
X105	Discrete	Survey's Question X105
X106	Discrete	Survey's Question X106
X107	Discrete	Survey's Question X107
X108	Discrete	Survey's Question X108
X109	Discrete	Survey's Question X109
X110	Discrete	Survey's Question X110
X111	Discrete	Survey's Question X111. Weight of Policyholder.
X112	Discrete	Survey's Question X112. Height of Policyholder.
X201	Discrete	Survey's Question X201
X202	Discrete	Survey's Question X202
X203	Discrete	Survey's Question X203
X204	Discrete	Survey's Question X204
X205	Discrete	Survey's Question X205
X206	Discrete	Survey's Question X206
X207	Discrete	Survey's Question X207
X208	Discrete	Survey's Question X208
X209	Discrete	Survey's Question X209
X210	Discrete	Survey's Question X210
X211	Discrete	Survey's Question X211
X212	Discrete	Survey's Question X212
X213	Discrete	Survey's Question X213
X214	Discrete	Survey's Question X214
X215	Discrete	Survey's Question X215
X216	Discrete	Survey's Question X216
X217	Discrete	Survey's Question X217
X218	Discrete	Survey's Question X218
X301	Discrete	Survey's Question X301
X302	Discrete	Survey's Question X302
X303	Discrete	Survey's Question X303
X304	Discrete	Survey's Question X304
X305	Discrete	Survey's Question X305
X306	Discrete	Survey's Question X306
X307	Discrete	Survey's Question X307
X401	Discrete	Survey's Question X401
X402	Discrete	Survey's Question X402
X403	Discrete	Survey's Question X403
X501	Discrete	Survey's Question X501
X502	Discrete	Survey's Question X502
X503	Discrete	Survey's Question X503
X504	Discrete	Survey's Question X504
X505	Discrete	Survey's Question X505
X506	Discrete	Survey's Question X506
X507	Discrete	Survey's Question X507
X6011	Discrete	Survey's Question X6011
DATE_ANS	Date	Policyholder’s Survey response Date
target	Binary	Target variable

Figure B.2: BOW Analysis, Question X108 (A.1.8). **Source:** Authors.

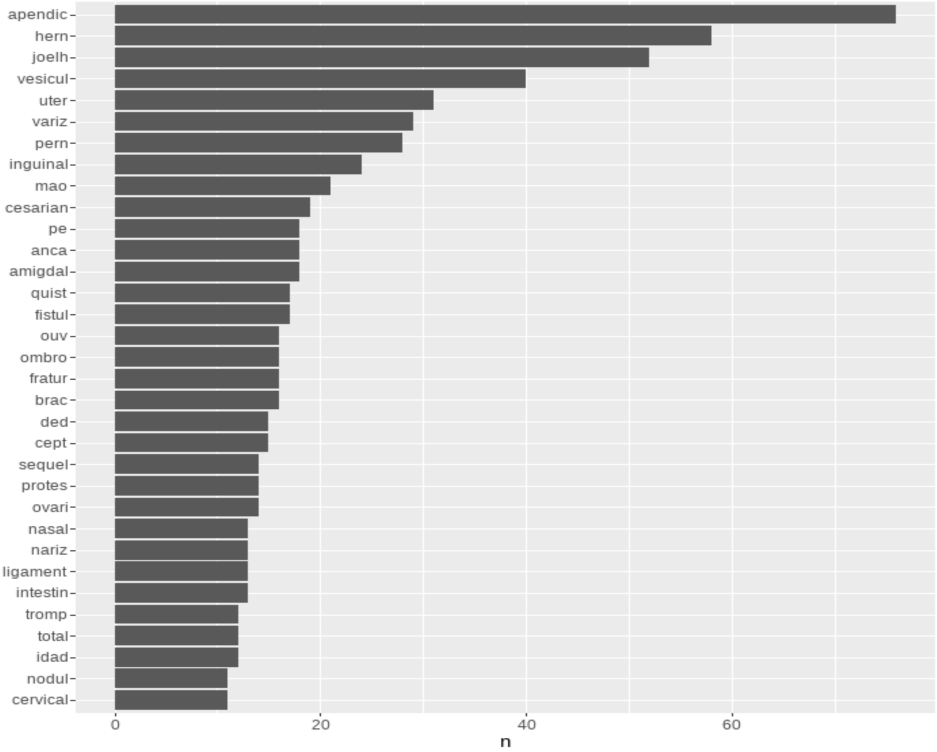


Figure B.3: Question X111 Distribution (A.1.11) **Source:** Authors.

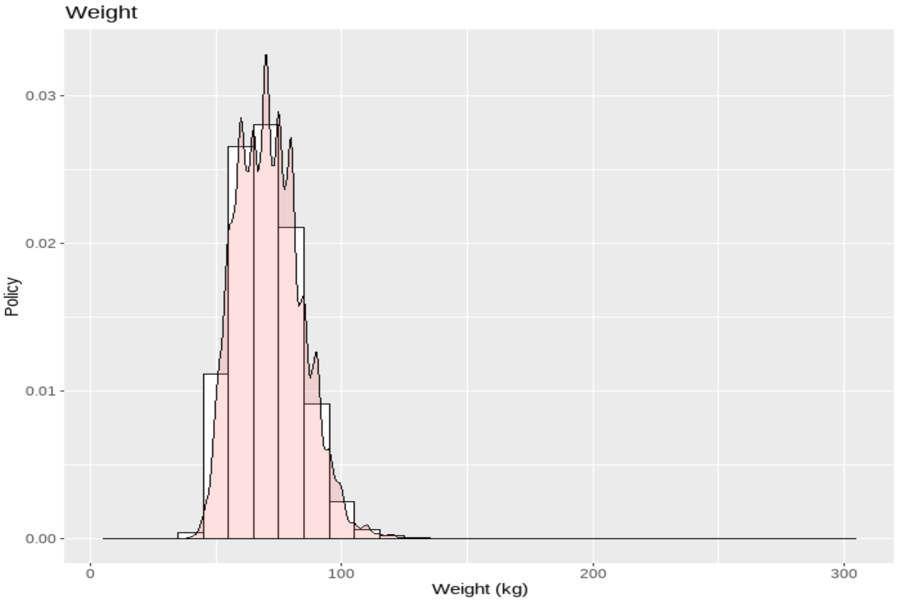


Figure B.4: Question X112 Distribution (A.1.12) **Source:** Authors.

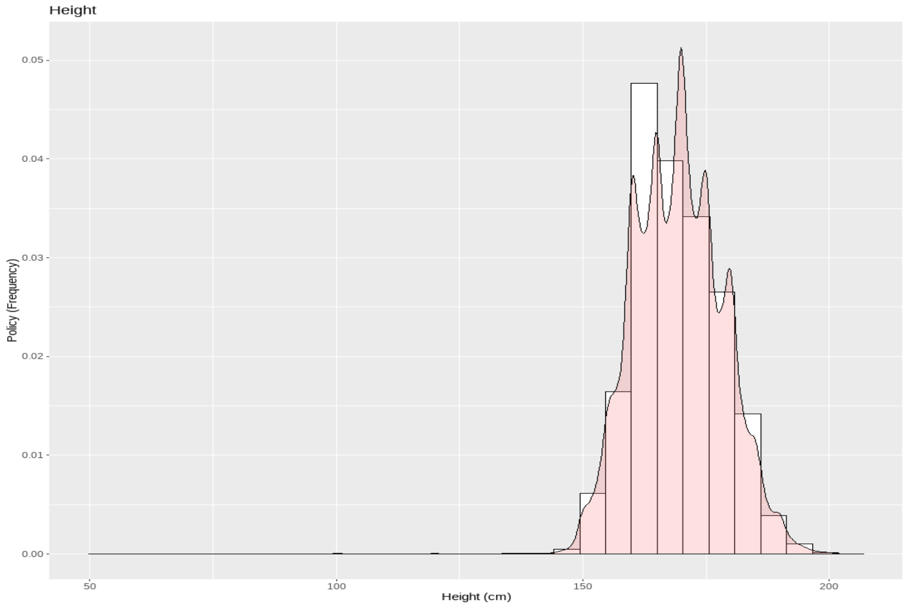


Figure B.5: BOW Analysis, Question X216 (A.2.16). **Source:** Authors.

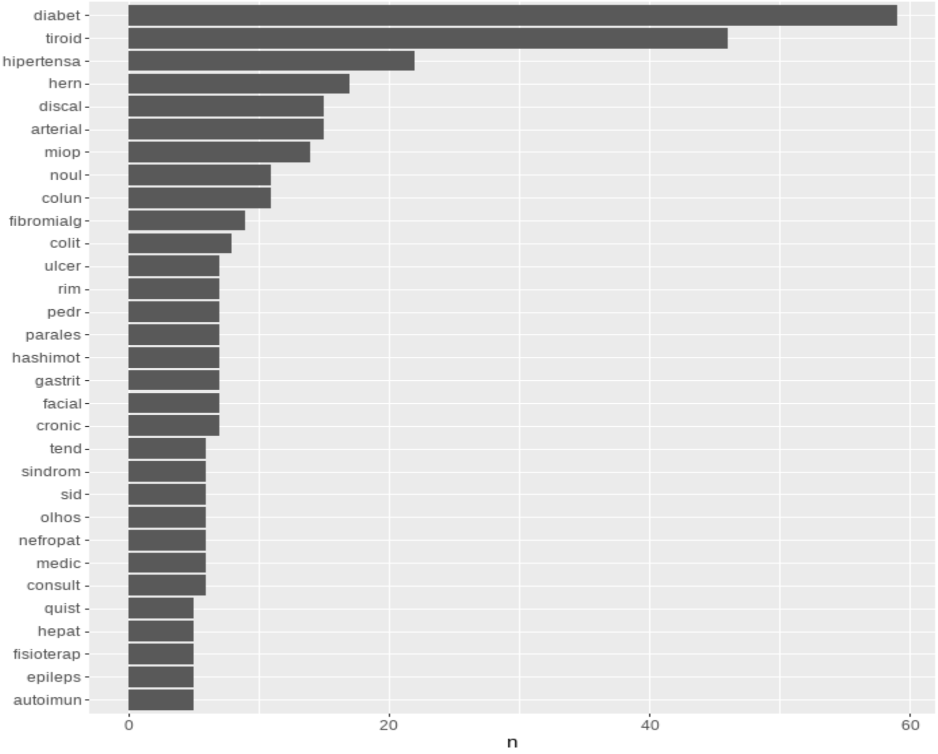


Table B.6: Aggravation Motive Codes. **Source:** Authors.

MOTIVOAGRAVAMENTO		**RISK_MOTIVOAGRAVAMENTO**
Code	**Description**	**Code**
1	Weight/Height Ratio	2
55	Sport Risk	3
28	Breast Pathology	4
2	Blood Pressure	6
53	Accommodation	7
24	Dermatological Pathology	9
12	Asthma	10
25	Digestive System Pathology	11
99	Non Coded Physical Aggravation	13
3	Factors of Cardiovascular Risk	14
51	Two Wheeled Vehicle Risk	15
26	Obesity Surgery	16
62	Clinical And Occupational	17
10	Intestine Pathology	18
60	Professional And Accommodation	19
54	Professional	20
21	Psychic Pathology	21
11	Pulmonary Pathology	23
56	Clinical And Professional	24
19	Rheumatism Pathology	25
7.1	Hepatitis B	27
27	Glucose Intolerance	27
96	Non Coded Special Aggravation	27
5	Dyslipidemia	28
59	Occupational	29
6	Diabetes	30
20	Neurological Pathology	32
15	Hematological Pathology	33
30	Infectious Diseases	34
14	Renal Pathology	36
9	Liver Pathology	37
M	Medical	38
4	Cardiovascular Pathology	39
22	Oncological Pathology	41

Figure B.6: Aggravation Motive One. **Source:** Authors.

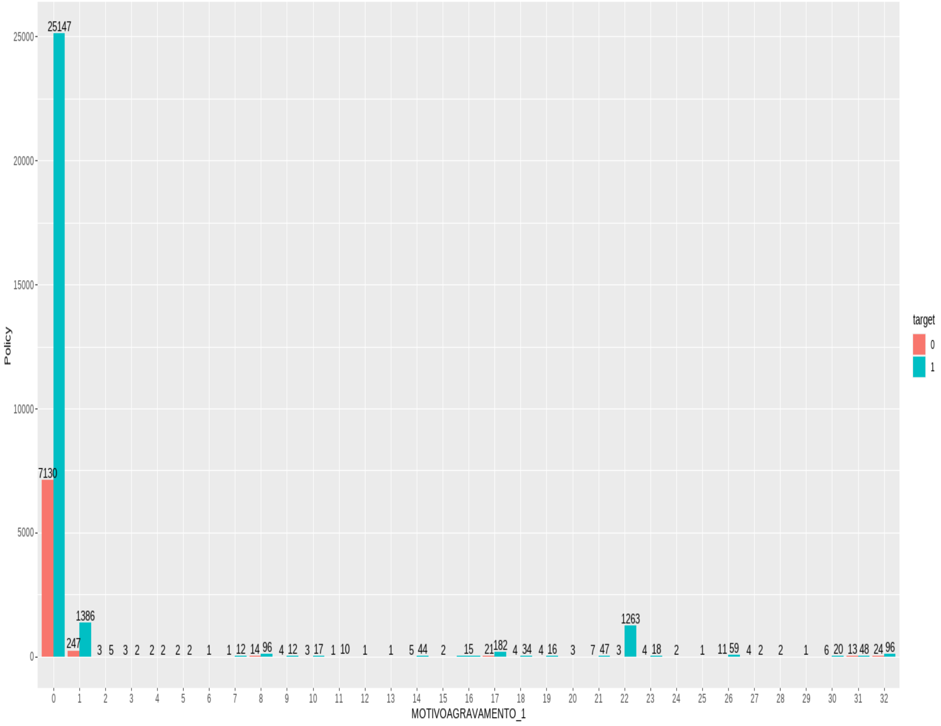


Figure B.7: Aggravation Motive Two. **Source:** Authors.

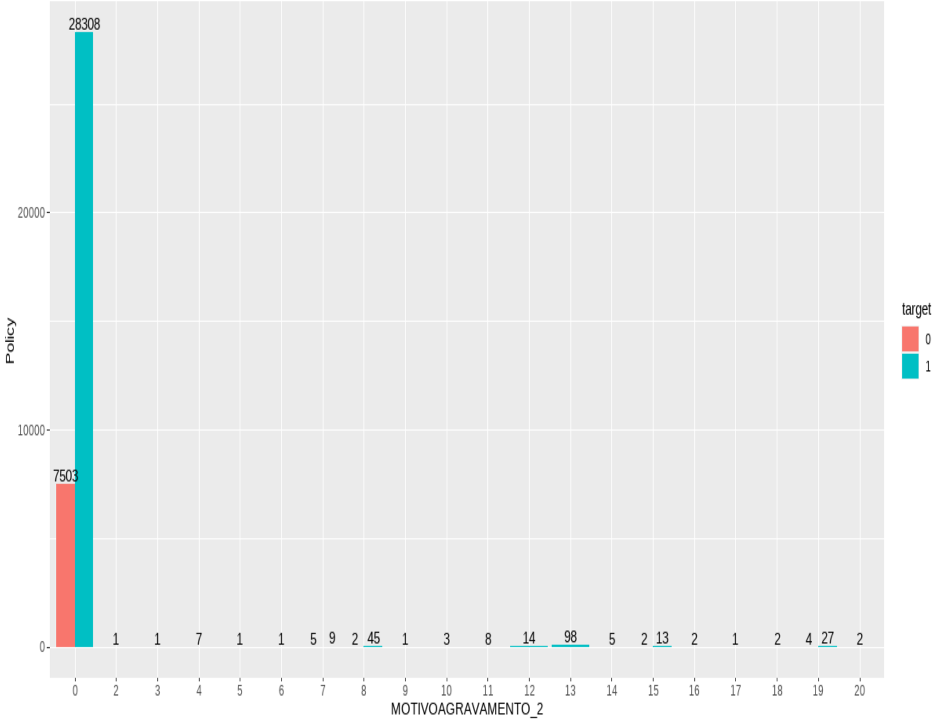


Figure B.8: Aggravation Motive Three. **Source:** Authors.

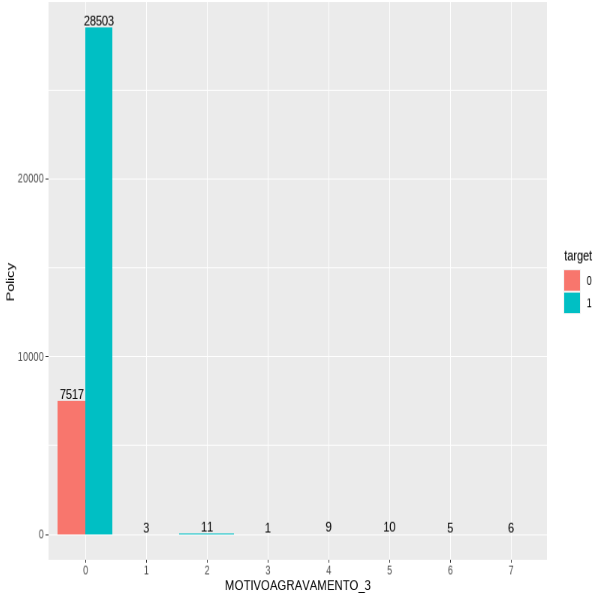


Figure B.9: Death Feature. **Source:** Authors.

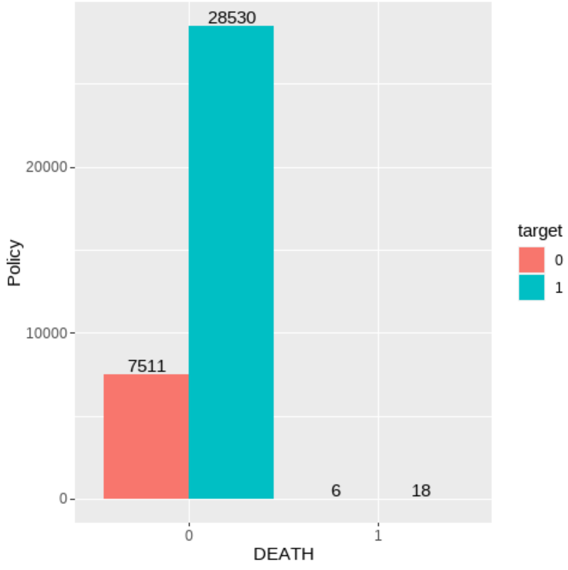


Figure B.10: Age Group. **Source:** Authors.

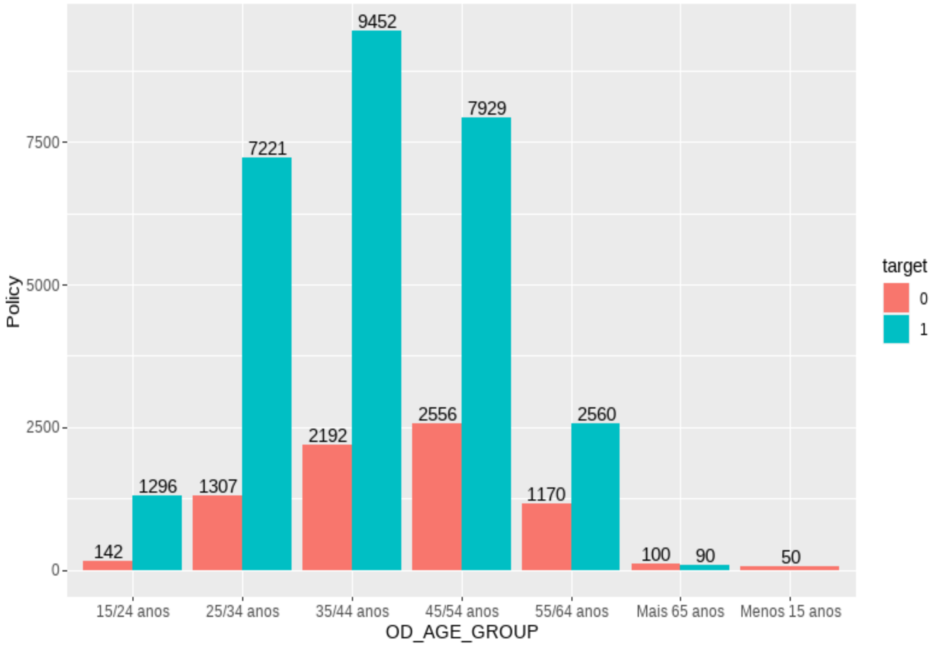


Figure B.11: Districts. **Source:** Authors.

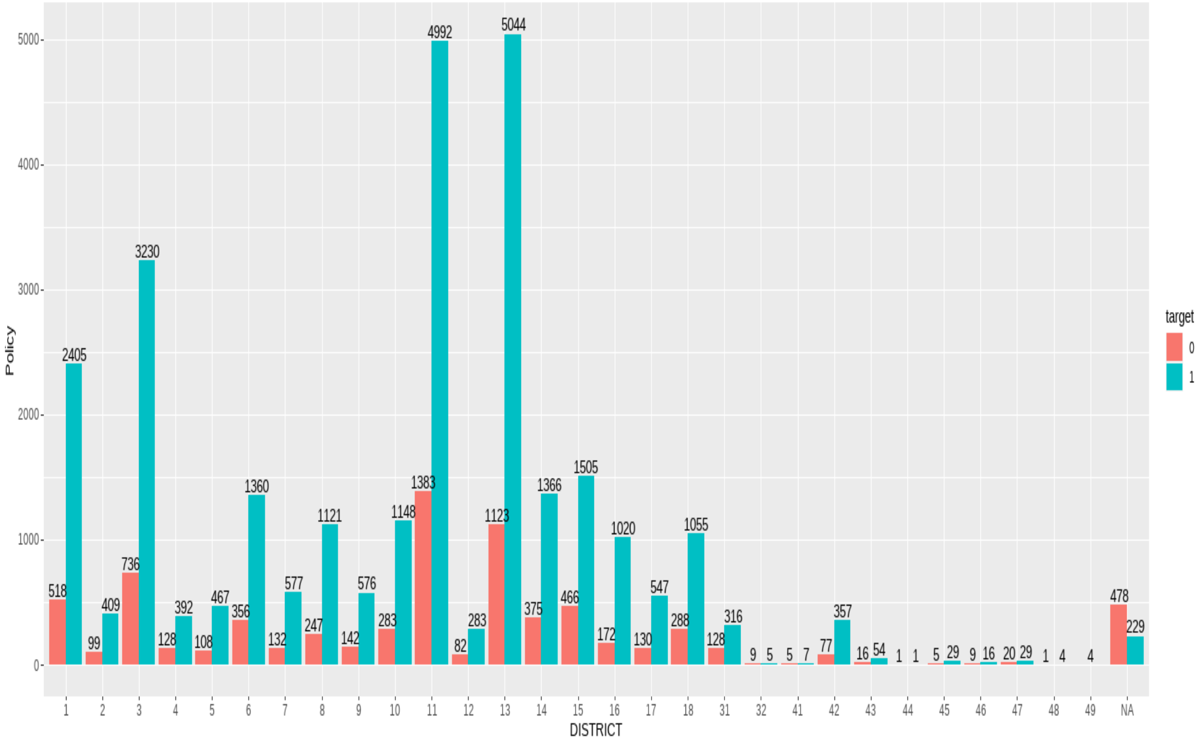


Table B.7: District Codes Lookup **Source:** Authors.

COD_DISTRICT	DISTRICT
1	AVEIRO
2	BEJA
3	BRAGA
4	BRAGANÇA
5	CASTELO BRANCO
6	COIMBRA
7	EVORA
8	FARO
9	GUARDA
10	LEIRIA
11	LISBOA
12	PORTALEGRE
13	PORTO
14	SANTARÉM
15	SETUBAL
16	VIANA DO CASTELO
17	VILA REAL
18	VISEU
31	ILHA DA MADEIRA
32	ILHA DE PORTO SANTO
41	ILHA DE SANTA MARIA
42	ILHA DE SÃO MIGUEL
43	ILHA TERCEIRA
44	ILHA DA GRACIOSA
45	ILHA DE SAO JORGE
46	ILHA DO PICO
47	ILHA DO FAIAL
48	ILHA DAS FLORES
49	ILHA DO CORVO

Table B.8: Marital Status Codes Lookup

OD_MARITAL_STATUS	
Code	**Description**
C	CASADO
D	DIVORCIADO
P	SEPARADO
S	SOLTEIRO
U	UNIAO DE FACTO
V	VIUVO
Z	EMPRESA

Figure B.12: Status Occupation. **Source:** Authors.

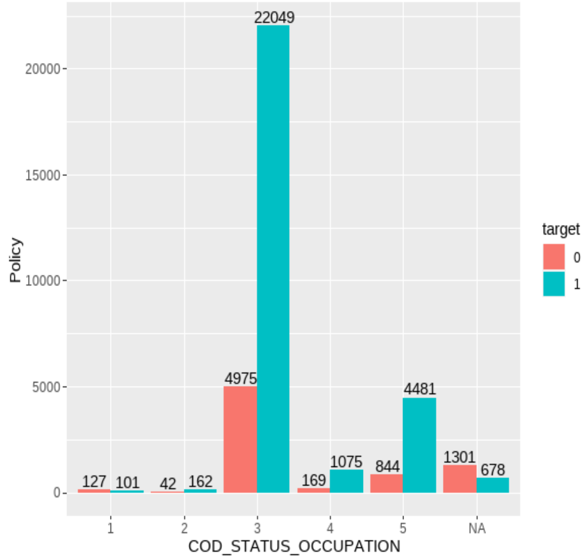


Table B.9: Literary Abilities Codes Lookup

OD_LITERARY_ABILITIES	
Code	**Description**
1	SEM ESCOLARIDADE
2	ENSINO BASICO INCOMPLETO
3	ENSINO BASICO
4	ENSINO SECUNDARIO
5	BACHARELATO
6	LICENCIATURA
7	MESTRADO OU DOUTORAMENTO

Figure B.13: BMI Distribution. **Source:** Authors.

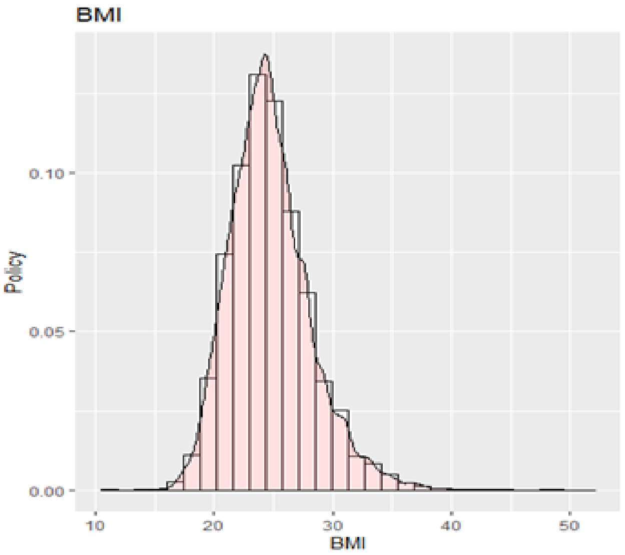


Table B.10: Disease Codes Lookup. **Source:** Authors.

Disease Code	Description
0	Without Diseases
1	Others
10	Metabolic or blood diseases
10.1	Diabetes
11	Neurological diseases
12	Osteoarticular, spinal or rheumatological diseases
13	Psychiatric diseases
14	Diseases related to the senses considering only ears or eyes
15	Diseases related to Tumors or any type of Cancer
16	Vascular diseases
17	Cholosterol
2	Breast diseases
3	Skin diseases
4	Diseases of the Cardiovascular system
5	Diseases of the genitourinary system
6	Diseases of the respiratory system
6.1	Asthma
7	Diseases of the stomach, inflammation diseases of the intestine, pancreas or other
8	Gynecological diseases
9	Infectious diseases

Table B.11: Dataset II **Source:** Authors.

Dataset Features	Type	Description
CODIGOPRODUTO	Discrete	Life Product Identification Code (ID)
OD_NATIONALITY	Discrete	Nationality of Policyholder
COD_DISTRICT	Discrete	District of Policyholder
OD_GENDER	Discrete	Gender of Policyholder
OD_MARITAL_STATUS	Discrete	Marital Status of Policyholder
NUM_AGE	Numerical	Age of Policyholder
OD_AGE_GROUP	Discrete	Age Group of Policyholder
COD_STATUS_OCCUPATION	Discrete	Occupation Status of Policyholder
COD_OCCUPATION_MAIN	Discrete	Main Occupation of Policyholder
OD_LITERARY_ABILITIES	Discrete	Literary Abilities of Policyholder
IND_EAC_TABLE_VERSION	Discrete	Economy Activity Code (EAC) Table Version
OD_EAC	Discrete	Economy Activity Code (EAC)
IND_EAC_TABLE_VERSION_RISK	Discrete	Economy Activity Code (EAC) Risk
PCT_IND_EAC_TABLE_VERSION_RISK	Discrete	Economy Activity Code (EAC) Risk Percentage
VALORCOBERTURA	Discrete	Coverage Amount
RISK_MOTIVOAGRAVAMENTO_1	Discrete	Risk of Aggravation Motive number 1
RISK_MOTIVOAGRAVAMENTO_2	Discrete	Risk of Aggravation Motive number 2
RISK_MOTIVOAGRAVAMENTO_3	Discrete	Risk of Aggravation Motive number 3
RISK_MOTIVOAGRAVAMENTO_4	Discrete	Risk of Aggravation Motive number 4
RISK_MOTIVOAGRAVAMENTO_5	Discrete	Risk of Aggravation Motive number 5
RISK_MOTIVOAGRAVAMENTO_6	Discrete	Risk of Aggravation Motive number 6
RISK_MOTIVOAGRAVAMENTO_7	Discrete	Risk of Aggravation Motive number 7
RISK_MOTIVOAGRAVAMENTO_8	Discrete	Risk of Aggravation Motive number 8
mean_PCT_RISK_MTAGRAV	Numerical	Mean of Aggravation Motive Risk Percentage
MONTANTECONTRATO	Discrete	Contract Amount
FRACCIONAMENTO	Discrete	Premium Frequency
NUMEROANOS	Discrete	Policy Term
X101	Discrete	Survey's Question X101
X102	Discrete	Survey's Question X102
X103	Discrete	Survey's Question X103
X104	Discrete	Survey's Question X104
X105	Discrete	Survey's Question X105
X106	Discrete	Survey's Question X106
X107	Discrete	Survey's Question X107
X108	Discrete	Survey's Question X108
X109	Discrete	Survey's Question X109
X110	Discrete	Survey's Question X110
X111	Discrete	Survey's Question X111. Weight of Policyholder.
X112	Discrete	Survey's Question X112. Height of Policyholder.
BMI	Discrete	Body Mass index of Policyholder
DOENCA_1	Discrete	Disease number 1 of Policyholder
DOENCA_2	Discrete	Disease number 2 of Policyholder
DOENCA_3	Discrete	Disease number 3 of Policyholder
DOENCA_4	Discrete	Disease number 4 of Policyholder
DOENCA_5	Discrete	Disease number 5 of Policyholder
DOENCA_6	Discrete	Disease number 6 of Policyholder
DOENCA_7	Discrete	Disease number 7 of Policyholder
DOENCA_8	Discrete	Disease number 8 of Policyholder
DOENCA_9	Discrete	Disease number 9 of Policyholder
DOENCA_10	Discrete	Disease number 10 of Policyholder
DOENCA_11	Discrete	Disease number 11 of Policyholder
DOENCA_12	Discrete	Disease number 12 of Policyholder
DOENCA_13	Discrete	Disease number 13 of Policyholder
DOENCA_14	Discrete	Disease number 14 of Policyholder
DOENCA_15	Discrete	Disease number 15 of Policyholder
mean_PCT_RISK_DOENCA	Numerical	Mean of Disease Risk Percentage
MEDICAMENTOS_1	Discrete	Medicine number 1 of Policyholder
MEDICAMENTOS_2	Discrete	Medicine number 2 of Policyholder
MEDICAMENTOS_3	Discrete	Medicine number 3 of Policyholder
MEDICAMENTOS_4	Discrete	Medicine number 4 of Policyholder
MEDICAMENTOS_5	Discrete	Medicine number 5 of Policyholder
MEDICAMENTOS_6	Discrete	Medicine number 6 of Policyholder
MEDICAMENTOS_7	Discrete	Medicine number 7 of Policyholder
X401	Discrete	Survey's Question X401
X402	Discrete	Survey's Question X402
X403	Discrete	Survey's Question X403
X501	Discrete	Survey's Question X501
X502	Discrete	Survey's Question X502
X503	Discrete	Survey's Question X503
X504	Discrete	Survey's Question X504
X505	Discrete	Survey's Question X505
X506	Discrete	Survey's Question X506
X507	Discrete	Survey's Question X507
X6011	Discrete	Survey's Question X6011
target	Discrete	Target variable

Figure B.14: Missing Data. **Source:** Authors.

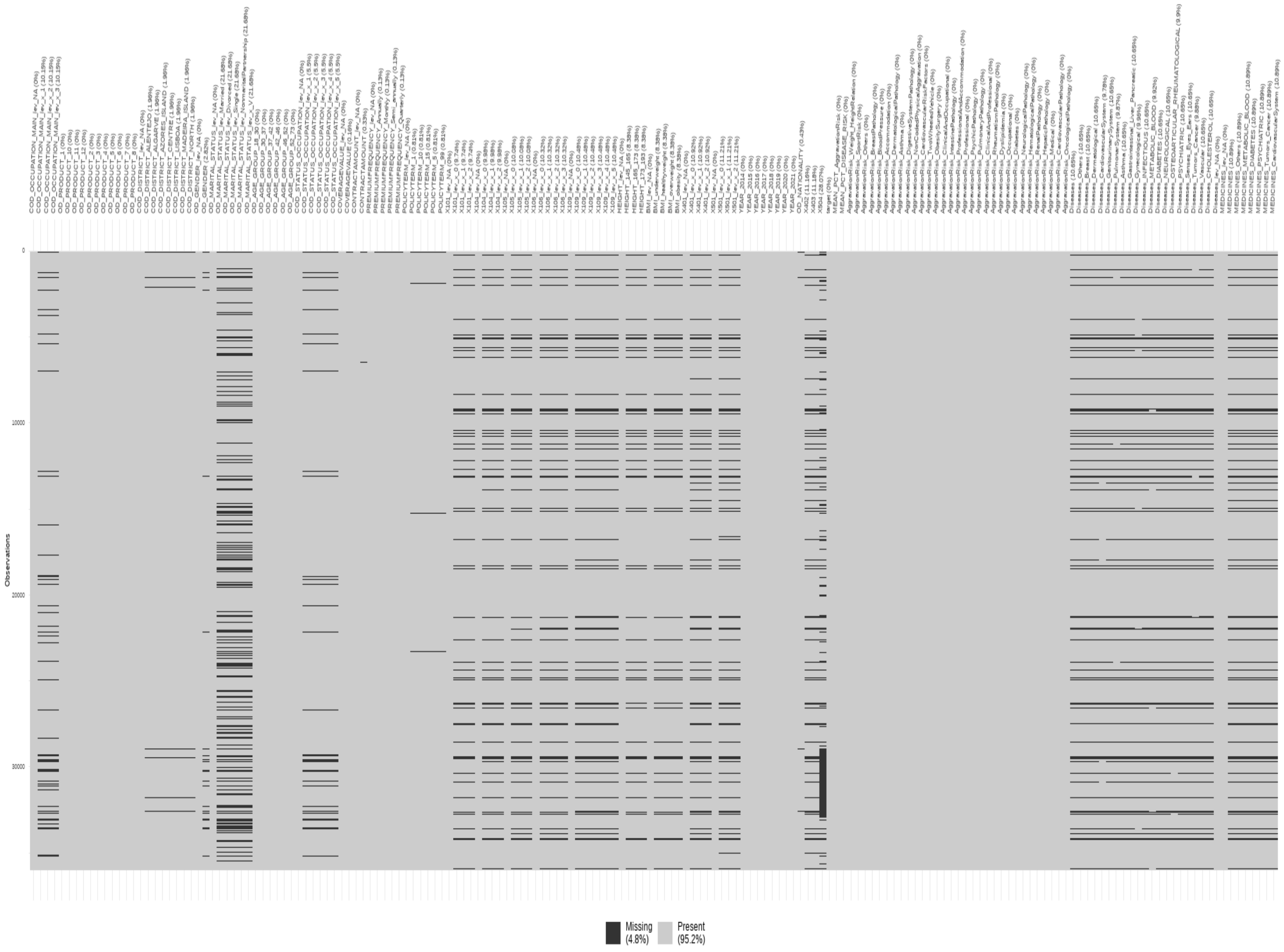
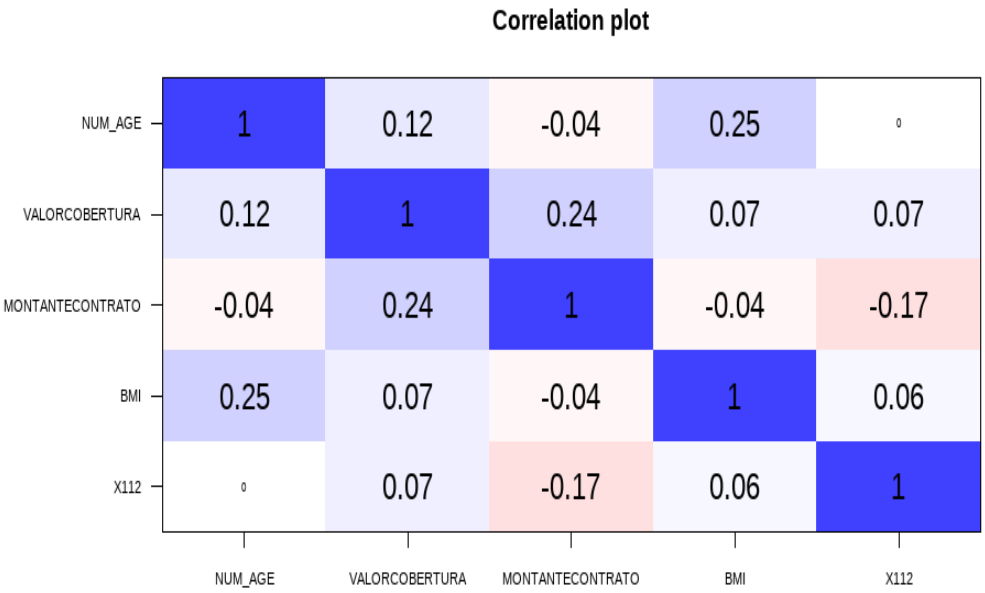


Figure B.15: Covariance Matrix II. **Source:** Authors.



Correlation plot

	NUM_AGE	VALORCOBERTURA	MONTANTECONTRATO	BMI	X112
NUM_AGE	1	0.12	-0.04	0.25	0
VALORCOBERTURA	0.12	1	0.24	0.07	0.07
MONTANTECONTRATO	-0.04	0.24	1	-0.04	-0.17
BMI	0.25	0.07	-0.04	1	0.06
X112	0	0.07	-0.17	0.06	1

Figure B.16: FAMD analysis of Random Forest imputation dataset. **Source:** Authors.

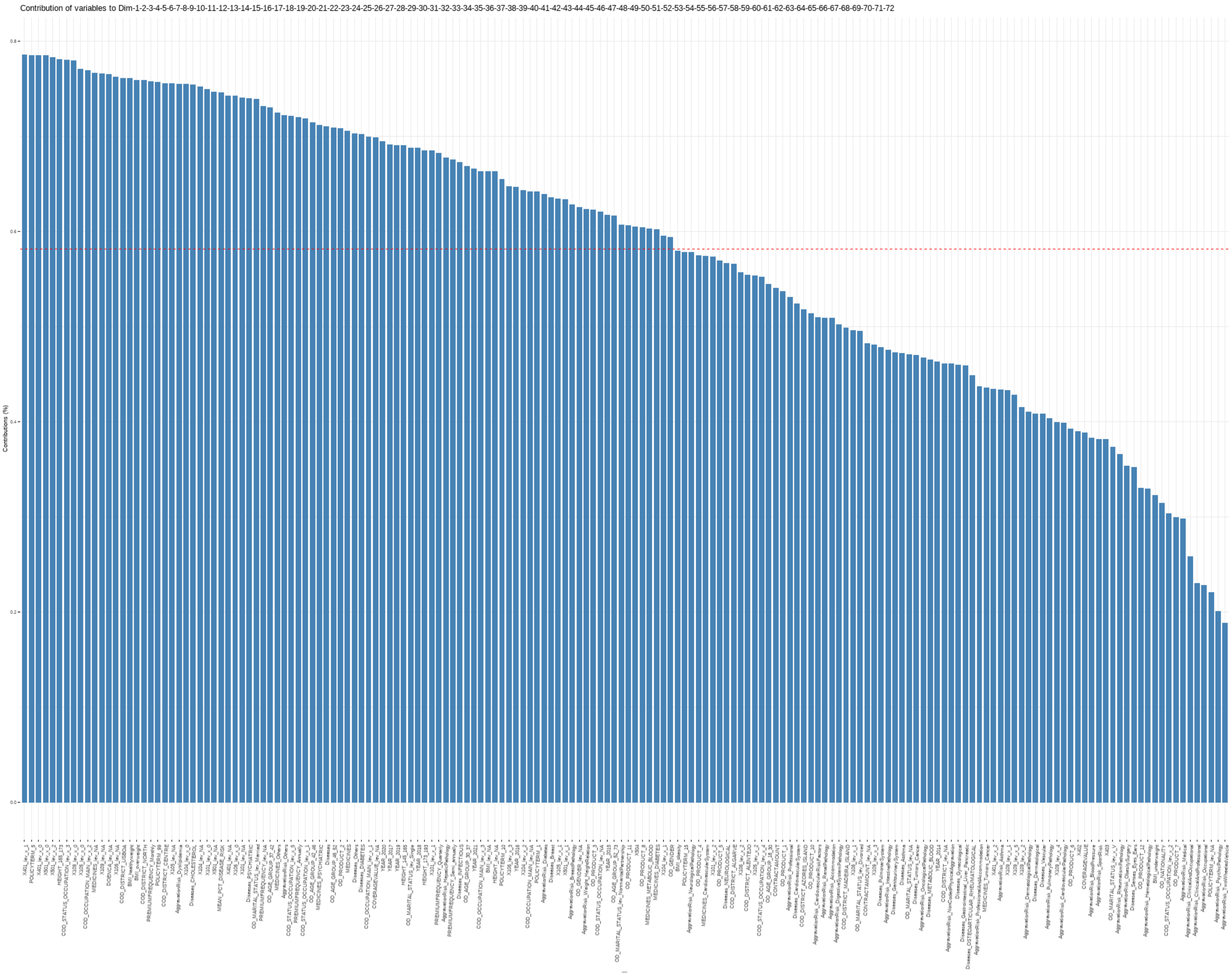


Figure B.17: FAMD analysis of MICE imputation dataset. **Source:** Authors.

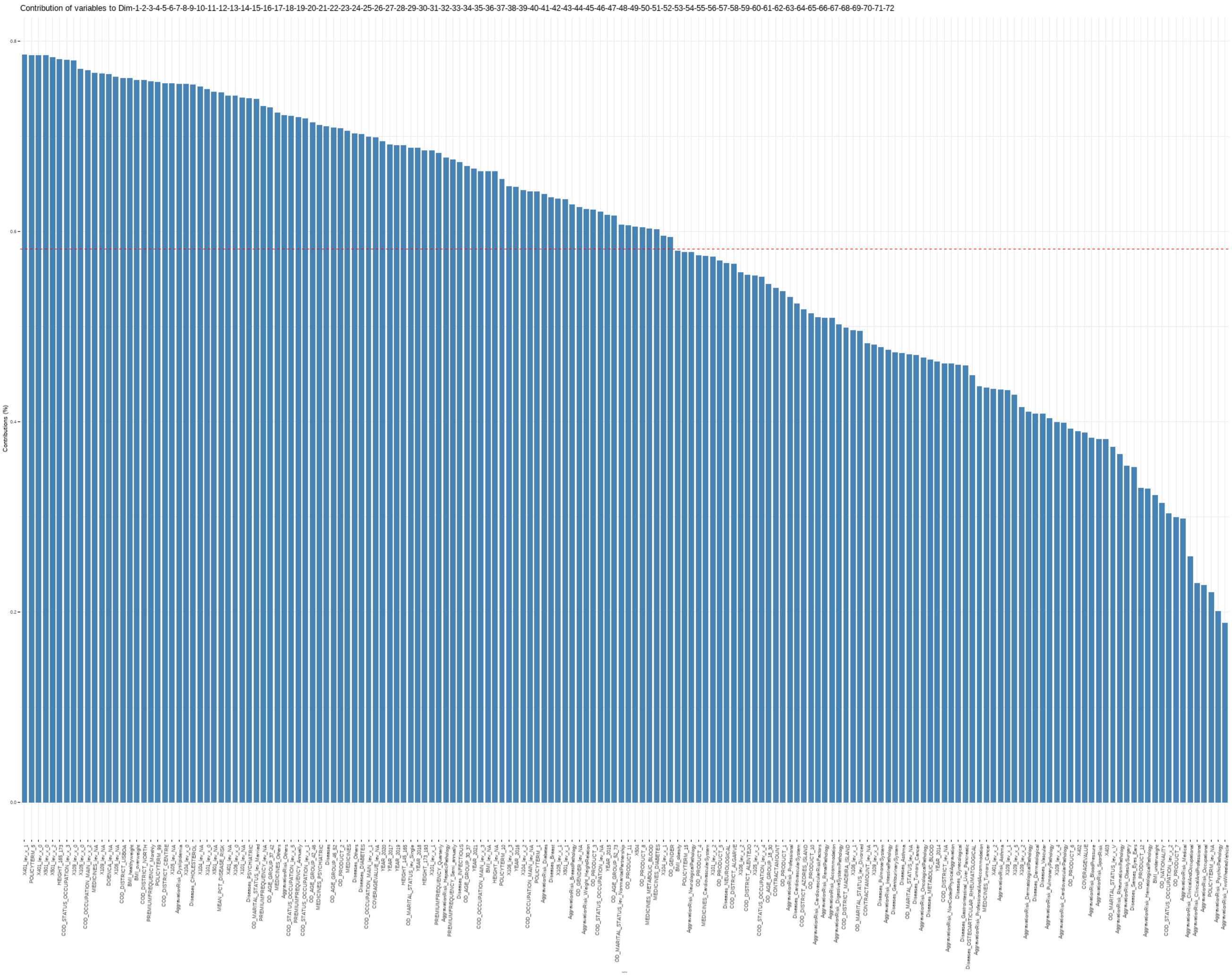


Table B.12: One-Hot Coding Dataset (Part I). **Source:** Authors.

New	New name	Type	Description
COD_OCCUPATION_MAIN_lev_NA	COD_OCCUPATION_MAIN_lev_NA	Binary	if ’COD_OCCUPATION_MAIN’ is missing (1) or not missing (0)
COD_OCCUPATION_MAIN_lev_x_1	COD_OCCUPATION_MAIN_lev_x_1	Binary	if ’COD_OCCUPATION_MAIN’ is the category one (1) or not (0)
COD_OCCUPATION_MAIN_lev_x_2	COD_OCCUPATION_MAIN_lev_x_2	Binary	if ’COD_OCCUPATION_MAIN’ is the category two (1) or not (0)
COD_OCCUPATION_MAIN_lev_x_3	COD_OCCUPATION_MAIN_lev_x_3	Binary	if ’COD_OCCUPATION_MAIN’ is the category three (1) or not (0)
CODIGOPRODUTO_lev_x_1	OD_PRODUCT_1	Binary	if is the policy is associated to product one (1) or not (0)
CODIGOPRODUTO_lev_x_10	OD_PRODUCT_10	Binary	if is the policy is associated to product two (1) or not (0)
CODIGOPRODUTO_lev_x_11	OD_PRODUCT_11	Binary	if is the policy is associated to product three (1) or not (0)
CODIGOPRODUTO_lev_x_12	OD_PRODUCT_12	Binary	if is the policy is associated to product four (1) or not (0)
CODIGOPRODUTO_lev_x_2	OD_PRODUCT_2	Binary	if is the policy is associated to product five (1) or not (0)
CODIGOPRODUTO_lev_x_3	OD_PRODUCT_3	Binary	if is the policy is associated to product six (1) or not (0)
CODIGOPRODUTO_lev_x_4	OD_PRODUCT_4	Binary	if is the policy is associated to product seven (1) or not (0)
CODIGOPRODUTO_lev_x_5	OD_PRODUCT_5	Binary	if is the policy is associated to product eight (1) or not (0)
CODIGOPRODUTO_lev_x_6	OD_PRODUCT_6	Binary	if is the policy is associated to product nine (1) or not (0)
CODIGOPRODUTO_lev_x_7	OD_PRODUCT_7	Binary	if is the policy is associated to product ten (1) or not (0)
CODIGOPRODUTO_lev_x_8	OD_PRODUCT_8	Binary	if is the policy is associated to product eleven (1) or not (0)
COD_DISTRICT_lev_NA	COD_DISTRICT_lev_NA	Binary	if ’COD_DISTRICT’ is missing (1) or not missing (0)
COD_DISTRICT_lev_x_ALENTEJO	COD_DISTRICT_ALENTEJO	Binary	if the District is from the region of Alentejo (1) or not (0)
COD_DISTRICT_lev_x_ALGARVE	COD_DISTRICT_ALGARVE	Binary	if the District is from the region of Algarve (1) or not (0)
COD_DISTRICT_lev_x_AZORES_ISLAND	COD_DISTRICT_AZORES_ISLAND	Binary	if the District is from the region of Azores Island (1) or not (0)
COD_DISTRICT_lev_x_CENTRE	COD_DISTRICT_CENTRE	Binary	if the District is from the region of Centre (1) or not (0)
COD_DISTRICT_lev_x_LISBOA	COD_DISTRICT_LISBOA	Binary	if the District is from the region of Lisbon (1) or not (0)
COD_DISTRICT_lev_x_MADEIRA_ISLAND	COD_DISTRICT_MADEIRA_ISLAND	Binary	if the District is from the region of Madeira Island (1) or not (0)
COD_DISTRICT_lev_x_NORTH	COD_DISTRICT_NORTH	Binary	if the District is from the region of North (1) or not (0)
OD_GENDER_lev_NA	OD_GENDER_lev_NA	Binary	if ’OD_GENDER’ is missing (1) or not missing (0)
OD_GENDER_lev_x_F	OD_GENDER	Binary	if the policyholder is Female (1) or Male (0)
OD_GENDER_lev_x_M	OD_GENDER_lev_x_M	Binary	if the policyholder is Male (1) or Female (0)
OD_MARITAL_STATUS_lev_NA	OD_MARITAL_STATUS_lev_NA	Binary	if ’OD_MARITAL_STATUS’ is missing (1) or not missing (0)
OD_MARITAL_STATUS_lev_x_C	OD_MARITAL_STATUS_lev_Married	Binary	if the policyholder is Married (1) or not (0)
OD_MARITAL_STATUS_lev_x_D	OD_MARITAL_STATUS_lev_Divorced	Binary	if the policyholder is Divorced (1) or not (0)
OD_MARITAL_STATUS_lev_x_S	OD_MARITAL_STATUS_lev_Single	Binary	if the policyholder is Single (1) or not (0)
OD_MARITAL_STATUS_lev_x_U	OD_MARITAL_STATUS_lev_NonmaritalPartnership	Binary	if the policyholder has a Non-marital Partnership (1) or not (0)
OD_MARITAL_STATUS_lev_x_V	OD_MARITAL_STATUS_lev_Widower	Binary	if the policyholder is Widower (1) or not (0)
OD_AGE_GROUP_lev_x_1	OD_AGE_GROUP_18_30	Binary	if the policyholder is between 18 to 30 (1) or not (0)
OD_AGE_GROUP_lev_x_2	OD_AGE_GROUP_30_37	Binary	if the policyholder is between 31 to 37 (1) or not (0)
OD_AGE_GROUP_lev_x_3	OD_AGE_GROUP_37_42	Binary	if the policyholder is between 38 to 42 (1) or not (0)
OD_AGE_GROUP_lev_x_4	OD_AGE_GROUP_42_46	Binary	if the policyholder is between 43 to 46 (1) or not (0)
OD_AGE_GROUP_lev_x_5	OD_AGE_GROUP_46_52	Binary	if the policyholder is between 47 to 52 (1) or not (0)
OD_AGE_GROUP_lev_x_6	OD_AGE_GROUP_52_73	Binary	if the policyholder is between 53 to 73 (1) or not (0)
COD_STATUS_OCCUPATION_lev_NA	COD_STATUS_OCCUPATION_lev_NA	Binary	if ’COD_STATUS_OCCUPATION’ is missing (1) or not missing (0)
COD_STATUS_OCCUPATION_lev_x_1	COD_STATUS_OCCUPATION_lev_x_1	Binary	if the policyholder is Retired (1) or not (0)
COD_STATUS_OCCUPATION_lev_x_2	COD_STATUS_OCCUPATION_lev_x_2	Binary	if the policyholder is Unemployed (1) or not (0)
COD_STATUS_OCCUPATION_lev_x_3	COD_STATUS_OCCUPATION_lev_x_3	Binary	if the policyholder is Active (1) or not (0)
COD_STATUS_OCCUPATION_lev_x_4	COD_STATUS_OCCUPATION_lev_x_4	Binary	if the policyholder is Student (1) or not (0)
COD_STATUS_OCCUPATION_lev_x_5	COD_STATUS_OCCUPATION_lev_x_5	Binary	if the policyholder is an individual business person (1) or not (0)
VALORCOBERTURA_lev_NA	COVERAGEVALUE_lev_NA	Binary	if ’VALORCOBERTURA’ is missing (1) or not missing (0)
VALORCOBERTURA_lev_x_1	COVERAGEVALUE	Binary	if ’VALORCOBERTURA’ is the category one (1) or two (0)
VALORCOBERTURA_lev_x_2	VALORCOBERTURA_lev_x_2	Binary	if ’VALORCOBERTURA’ is the category two (1) or one (0)
MONTANTECONTRATO_lev_NA	CONTRACTAMOUNT_lev_NA	Binary	if ’MONTANTECONTRATO’ is missing (1) or not missing (0)
MONTANTECONTRATO_lev_x_1	CONTRACTAMOUNT	Binary	if ’MONTANTECONTRATO’ is the category one (1) or two (0)
MONTANTECONTRATO_lev_x_2	MONTANTECONTRATO_lev_x_2	Binary	if ’MONTANTECONTRATO’ is the category two (1) or one (0)
FRACCIONAMENTO_lev_NA	PREMIUMFREQUENCY_lev_NA	Binary	if ’PREMIUMFREQUENCY’ is missing (1) or not missing (0)
FRACCIONAMENTO_lev_x_A	PREMIUMFREQUENCY_Annually	Binary	if the policy’s premium is paid annually (1) or not (0)
FRACCIONAMENTO_lev_x_M	PREMIUMFREQUENCY_Monthly	Binary	if the policy’s premium is paid monthly (1) or not (0)
FRACCIONAMENTO_lev_x_S	PREMIUMFREQUENCY_Semi_annually	Binary	if the policy’s premium is paid twice a year (1) or not (0)
FRACCIONAMENTO_lev_x_T	PREMIUMFREQUENCY_Quarterly	Binary	if the policy’s premium is paid every quarter of a year (1) or not (0)
NUMEROANOS_lev_NA	POLICYTERM_lev_NA	Binary	if ’POLICYTERM’ is missing (1) or not missing (0)
NUMEROANOS_lev_x_1	POLICYTERM_1	Binary	if the policy term is one year (1) or not (0)
NUMEROANOS_lev_x_10	POLICYTERM_10	Binary	if the policy term is 10 years (1) or not (0)
NUMEROANOS_lev_x_15	POLICYTERM_15	Binary	if the policy term is 15 years (1) or not (0)
NUMEROANOS_lev_x_5	POLICYTERM_5	Binary	if the policy term is 5 years (1) or not (0)
NUMEROANOS_lev_x_99	POLICYTERM_99	Binary	if the policy term is 99 years (1) or not (0)
X101_lev_NA	X101_lev_NA	Binary	if the question ’X101’ was answered (1) or not (0)
X101_lev_x_0	X101_lev_x_0	Binary	if the answer of the question ’X101’ is the category zero (1) or not (0)
X101_lev_x_1	X101_lev_x_1	Binary	if the answer of the question ’X101’ is the category one (1) or not (0)
X101_lev_x_2	X101_lev_x_2	Binary	if the answer of the question ’X101’ is the category two (1) or not (0)
X104_lev_NA	X104_lev_NA	Binary	if the question ’X104’ was answered (1) or not (0)
X104_lev_x_0	X104_lev_x_0	Binary	if the answer of the question ’X104’ is the category zero (1) or not (0)
X104_lev_x_1	X104_lev_x_1	Binary	if the answer of the question ’X104’ is the category one (1) or not (0)
X104_lev_x_2	X104_lev_x_2	Binary	if the answer of the question ’X104’ is the category two (1) or not (0)
X105_lev_NA	X105_lev_NA	Binary	if the question ’X105’ was answered (1) or not (0)
X105_lev_x_0	X105_lev_x_0	Binary	if the answer of the question ’X105’ is the category zero (1) or not (0)
X105_lev_x_1	X105_lev_x_1	Binary	if the answer of the question ’X105’ is the category one (1) or not (0)
X105_lev_x_2	X105_lev_x_2	Binary	if the answer of the question ’X105’ is the category two (1) or not (0)
X106_lev_NA	X106_lev_NA	Binary	if the question ’X106’ was answered (1) or not (0)
X106_lev_x_0	X106_lev_x_0	Binary	if the answer of the question ’X106’ is the category zero (1) or not (0)
X106_lev_x_1	X106_lev_x_1	Binary	if the answer of the question ’X106’ is the category one (1) or not (0)
X106_lev_x_2	X106_lev_x_2	Binary	if the answer of the question ’X106’ is the category two (1) or not (0)
X106_lev_x_3	X106_lev_x_3	Binary	if the answer of the question ’X106’ is the category two (1) or not (0)
X109_lev_NA	X109_lev_NA	Binary	if the question ’X109’ was answered (1) or not (0)
X109_lev_x_0	X109_lev_x_0	Binary	if the answer of the question ’X109’ is the category zero (1) or not (0)
X109_lev_x_1	X109_lev_x_1	Binary	if the answer of the question ’X109’ is the category one (1) or not (0)
X109_lev_x_2	X109_lev_x_2	Binary	if the answer of the question ’X109’ is the category two (1) or not (0)
X109_lev_x_3	X109_lev_x_3	Binary	if the answer of the question ’X109’ is the category two (1) or not (0)
X109_lev_x_4	X109_lev_x_4	Binary	if the answer of the question ’X109’ is the category four (1) or not (0)
X109_lev_x_5	X109_lev_x_5	Binary	if the answer of the question ’X109’ is the category five (1) or not (0)
X112_lev_NA	HEIGHT_lev_NA	Binary	if the answer about policyholder’s height (’X112’) is missing (1) or not missing (0)
X112_lev_x_1	HEIGHT_145_165	Binary	if the policyholder’s height is between 145 to 165 cm (1) or not (0)
X112_lev_x_2	HEIGHT_165_173	Binary	if the policyholder’s height is between 166 to 173 cm (1) or not (0)

Table B.13: One-Hot Coding Dataset (Part II). **Source:** Authors.

New	New name	Type	Description
X112_lev_x_3	HEIGHT_173_193	Binary	if the policyholder's height is between 174 to 193 cm (1) or not (0)
BMI_lev_NA	BMI_lev_NA	Binary	if 'BMI' is missing (1) or not missing (0)
BMI_lev_x_1	BMI_underweight	Binary	if the policyholder's Body Mass index is between the range of underweight (1) or not (0)
BMI_lev_x_2	BMI_healthyweight	Binary	if the policyholder's Body Mass index is between the range of healthy weight (1) or not (0)
BMI_lev_x_3	BMI_overweight	Binary	if the policyholder's Body Mass index is between the range of over weight (1) or not (0)
BMI_lev_x_4	BMI_obesity	Binary	if the policyholder's Body Mass index is between the range of obesity (1) or not (0)
X401_lev_NA	X401_lev_NA	Binary	if the question 'X401' was answered (1) or not (0)
X401_lev_x_0	X401_lev_x_0	Binary	if the answer of the question 'X401' is the category zero (1) or not (0)
X401_lev_x_1	X401_lev_x_1	Binary	if the answer of the question 'X401' is the category one (1) or not (0)
X401_lev_x_2	X401_lev_x_2	Binary	if the answer of the question 'X401' is the category two (1) or not (0)
X501_lev_NA	X501_lev_NA	Binary	if the question 'X501' was answered (1) or not (0)
X501_lev_x_0	X501_lev_x_0	Binary	if the answer of the question 'X501' is the category zero (1) or not (0)
X501_lev_x_1	X501_lev_x_1	Binary	if the answer of the question 'X501' is the category one (1) or not (0)
X501_lev_x_2	X501_lev_x_2	Binary	if the answer of the question 'X501' is the category two (1) or not (0)
YEAR_ANS_lev_x_2014	YEAR_2014	Binary	if the policyholder's answer to the survey was in the year 2014 (1) or not (0)
YEAR_ANS_lev_x_2015	YEAR_2015	Binary	if the policyholder's answer to the survey was in the year 2015 (1) or not (0)
YEAR_ANS_lev_x_2016	YEAR_2016	Binary	if the policyholder's answer to the survey was in the year 2016 (1) or not (0)
YEAR_ANS_lev_x_2017	YEAR_2017	Binary	if the policyholder's answer to the survey was in the year 2017 (1) or not (0)
YEAR_ANS_lev_x_2018	YEAR_2018	Binary	if the policyholder's answer to the survey was in the year 2018 (1) or not (0)
YEAR_ANS_lev_x_2019	YEAR_2019	Binary	if the policyholder's answer to the survey was in the year 2019 (1) or not (0)
YEAR_ANS_lev_x_2020	YEAR_2020	Binary	if the policyholder's answer to the survey was in the year 2020 (1) or not (0)
YEAR_ANS_lev_x_2021	YEAR_2021	Binary	if the policyholder's answer to the survey was in the year 2021 (1) or not (0)
OD_NATIONALITY	OD_NATIONALITY	Binary	if the policyholder's nationality is Portuguese (1) or not (0)
X402	X402	Binary	if the question 'X402' was answered (1) or not (0)
X403	X403	Binary	if the question 'X403' was answered (1) or not (0)
X504	X504	Binary	if the question 'X504' was answered (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_2	AggravationRisk_Weight_HeightRelation	Binary	if is the policy has associated Weight/Height Relation Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_3	AggravationRisk_SportRisk	Binary	if is the policy has associated Sports Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_5	AggravationRisk_Others	Binary	if is the policy has associated an Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_4	AggravationRisk_BreastPathology	Binary	if is the policy has associated Breast Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_6	AggravationRisk_BloodPressure	Binary	if is the policy has associated Blood Pressure Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_7	AggravationRisk_Accommodation	Binary	if is the policy has associated Accommodation Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_9	AggravationRisk_DermatologicalPathology	Binary	if is the policy has associated Dermatological Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_10	AggravationRisk_Asthma	Binary	if is the policy has associated Asthma Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_11	AggravationRisk_DigestiveSystemPathology	Binary	if is the policy has associated Digestive System Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_13	AggravationRisk_NonCodedPhysicalAggravation	Binary	if is the policy has associated Non-Coded Physical Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_14	AggravationRisk_CardiovascularRiskFactors	Binary	if is the policy has associated Factors of Cardiovascular Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_15	AggravationRisk_TwoWheeledVehicle	Binary	if is the policy has associated Two Wheeled Vehicle Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_16	AggravationRisk_ObesitySurgery	Binary	if is the policy has associated Obesity Surgery Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_17	AggravationRisk_ClinicalAndOccupational	Binary	if is the policy has associated Clinical And Occupational Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_18	AggravationRisk_IntestinePathology	Binary	if is the policy has associated Intestine Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_19	AggravationRisk_ProfessionalAndAccommodation	Binary	if is the policy has associated Professional And Accommodation Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_20	AggravationRisk_Professional	Binary	if is the policy has associated Professional Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_21	AggravationRisk_PsychicPathology	Binary	if is the policy has associated Psychic Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_23	AggravationRisk_PulmonaryPathology	Binary	if is the policy has associated Pulmonary Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_24	AggravationRisk_ClinicalAndProfessional	Binary	if is the policy has associated Clinical And Professional Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_25	AggravationRisk_RheumatismPathology	Binary	if is the policy has associated Rheumatism Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_28	AggravationRisk_Dyslipidemia	Binary	if is the policy has associated Dyslipidemia Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_29	AggravationRisk_Occupational	Binary	if is the policy has associated Occupational Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_30	AggravationRisk_Diabetes	Binary	if is the policy has associated Diabetes Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_32	AggravationRisk_NeurologicalPathology	Binary	if is the policy has associated Neurological Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_33	AggravationRisk_HematologicalPathology	Binary	if is the policy has associated Hematological Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_36	AggravationRisk_RenalPathology	Binary	if is the policy has associated Renal Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_37	AggravationRisk_HepaticPathology	Binary	if is the policy has associated Hepatic Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_38	AggravationRisk_Medical	Binary	if is the policy has associated Medical Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_39	AggravationRisk_CardiovascularPathology	Binary	if is the policy has associated Cardiovascular Pathology Aggravation Risk (1) or not (0)
RISK_MOTIVOAGRAVAMENTO_lev_x_41	AggravationRisk_OncologicalPathology	Binary	if is the policy has associated Oncological Pathology Aggravation Risk (1) or not (0)
mean_PCT_RISK_MTAGRAV	MEAN_PCT_AggravationRisk	Numerical	Mean Risk Percentage of Aggravation Motives
DOENCA_lev_x_0	Diseases	Binary	if the policyholder has/had a disease (1) or not (0)
DOENCA_lev_x_1	Diseases_Others	Binary	if the policyholder has/had a disease considered in the category of Others (1) or not (0)
DOENCA_lev_x_2	Diseases_Breast	Binary	if the policyholder has/had a breast disease (1) or not (0)
DOENCA_lev_x_3	Diseases_Dermatological	Binary	if the policyholder has/had a dermatological disease (1) or not (0)
DOENCA_lev_x_4	Diseases_CardiovascularSystem	Binary	if the policyholder has/had a Disease related to the Cardiovascular system (1) or not (0)
DOENCA_lev_x_5	Diseases_GenitourinarySystem	Binary	if the policyholder has/had a Disease related to the Genitourinary system (1) or not (0)
DOENCA_lev_x_6	Diseases_PulmonarSystem	Binary	if the policyholder has/had a Disease related to the Pulmonar system (1) or not (0)
DOENCA_lev_x_6_1	Diseases_Asthma	Binary	if the policyholder has/had a Disease related to Asthma (1) or not (0)
DOENCA_lev_x_7	Diseases_Gastrointestinal_Liver_Pancreatic	Binary	if the policyholder has/had a Disease related to stomach, inflammation diseases of the intestine, pancreas or other (1) or not (0)
DOENCA_lev_x_8	Diseases_Gynecological	Binary	if the policyholder has/had a Gynecological Disease (1) or not (0)
DOENCA_lev_x_9	Diseases_INFECTIOUS	Binary	if the policyholder has/had a Infectous Disease (1) or not (0)
DOENCA_lev_x_10	Diseases_METABOLIC_BLOOD	Binary	if the policyholder has/had a Metabolic or Blood Disease (1) or not (0)
DOENCA_lev_x_10_1	Diseases_DIABETES	Binary	if the policyholder has/had Diabetes (1) or not (0)
DOENCA_lev_x_11	Diseases_NEUROLOGICAL	Binary	if the policyholder has/had a Neurological Disease (1) or not (0)
DOENCA_lev_x_12	Diseases_OSTEOARTICULAR_RHEUMATOLOGICAL	Binary	if the policyholder has/had a Osteoarticular, spinal or rheumatological Disease (1) or not (0)
DOENCA_lev_x_13	Diseases_PSYCHIATRIC	Binary	if the policyholder has/had a Psychiatric Disease (1) or not (0)
DOENCA_lev_x_14	Diseases_Senses_Eyes_Ears	Binary	if the policyholder has/had a Disease related to the senses considering only ears or eyes (1) or not (0)
DOENCA_lev_x_15	Diseases_Tumors_Cancer	Binary	if the policyholder has/had a Disease related to Tumors or any type of Cancer (1) or not (0)
DOENCA_lev_x_16	Diseases_Vascular	Binary	if the policyholder has/had a Vascular Disease (1) or not (0)
DOENCA_lev_x_17	Diseases_CHOLESTEROL	Binary	if the policyholder has/had Cholesterol (1) or not (0)
DOENCA_lev_NA	Diseases_lev_NA	Binary	if the policyholder has disease history missing (1) or not (0)
mean_PCT_RISK_DOENCA	MEAN_PCT_DISEASE_RISK	Numerical	Mean Risk Percentage of Diseases
MEDICAMENTOS_lev_NA	MEDICINES_lev_NA	Binary	if the policyholder has the therapeutic history missing (1) or not (0)
MEDICAMENTOS_lev_x_0	MEDICINES	Binary	if the policyholder has taken or took a medicine (1) or not (0)
MEDICAMENTOS_lev_x_1	MEDICINES_Others	Binary	if the policyholder has taken or took a medicine considered in the category of Others (1) or not (0)
MEDICAMENTOS_lev_x_10	MEDICINES_METABOLIC_BLOOD	Binary	if the policyholder has taken or took a medicine related to Metabolic or Blood Disease (1) or not (0)
MEDICAMENTOS_lev_x_10_1	MEDICINES_DIABETES	Binary	if the policyholder has taken or took a medicine related to Diabetes (1) or not (0)
MEDICAMENTOS_lev_x_13	MEDICINES_PSYCHIATRIC	Binary	if the policyholder has taken or took a medicine related to a Psychiatric Disease (1) or not (0)
MEDICAMENTOS_lev_x_15	MEDICINES_Tumors_Cancer	Binary	if the policyholder has taken or took a medicine associated to Disease related to Tumors or any type of Cancer (1) or not (0)
MEDICAMENTOS_lev_x_4	MEDICINES_CardiovascularSystem	Binary	if the policyholder has taken or took a medicine related to Disease related to the Cardiovascular system (1) or not (0)
target	target	Binary	Target . if the Policy have not Lapsed (1) or have Lapsed (0)

Figure B.18: Box plot Age. **Source:** Authors.

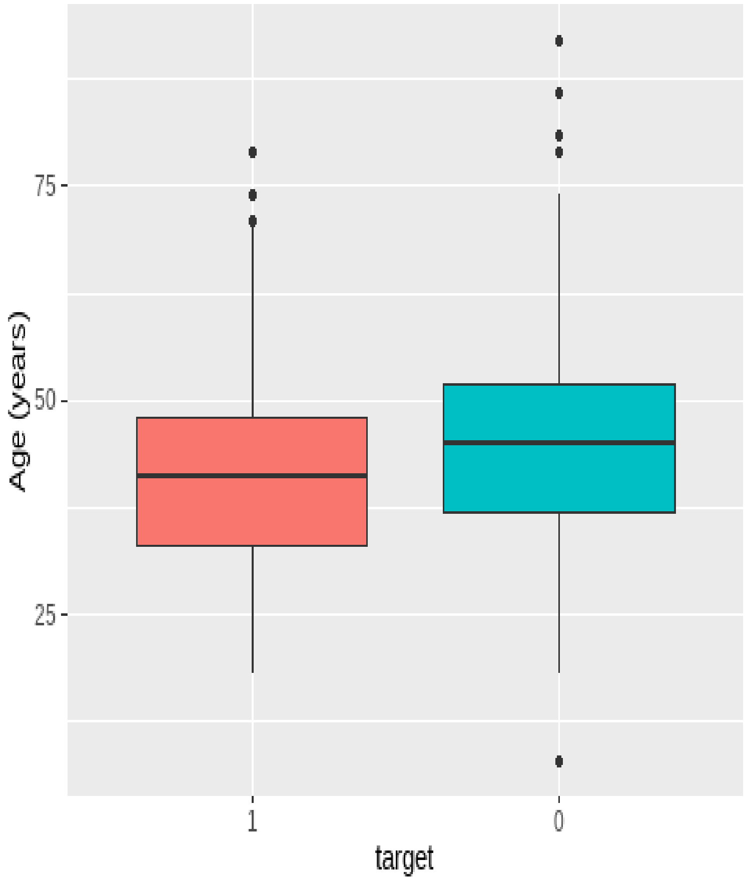


Figure B.19: Box plot Height. **Source:** Authors.

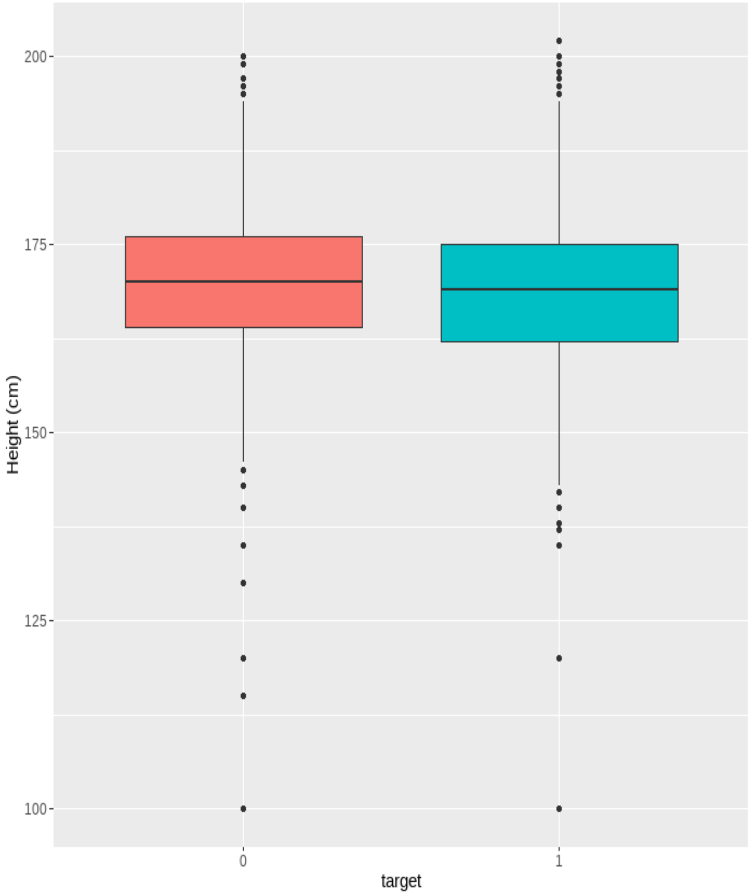


Figure B.20: Box plot Weight. **Source:** Authors.

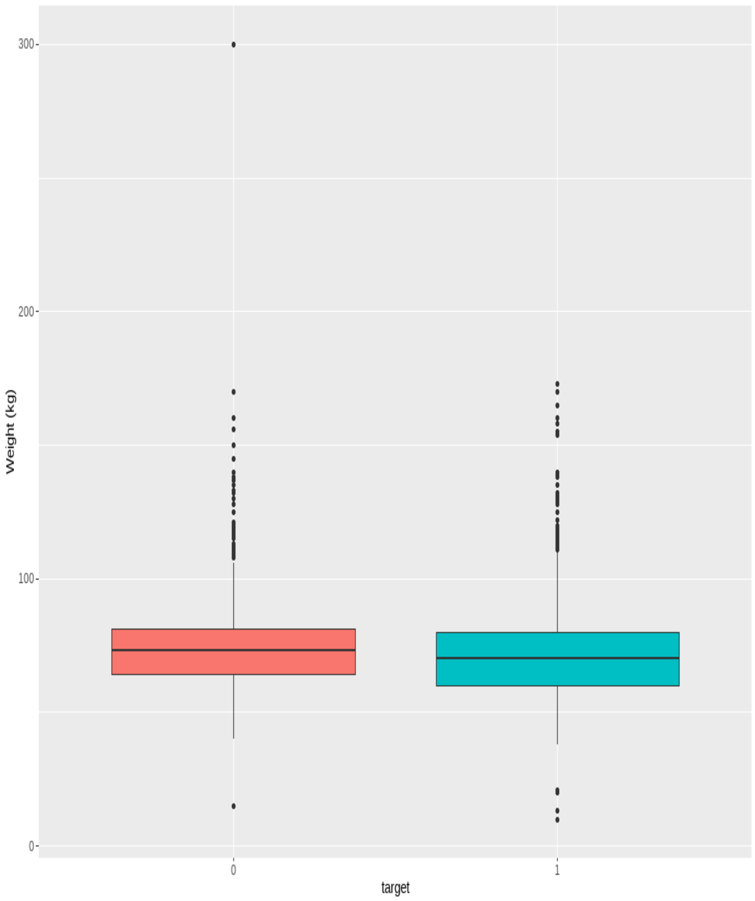


Figure B.21: Box plot Coverage Amount. **Source:** Authors.

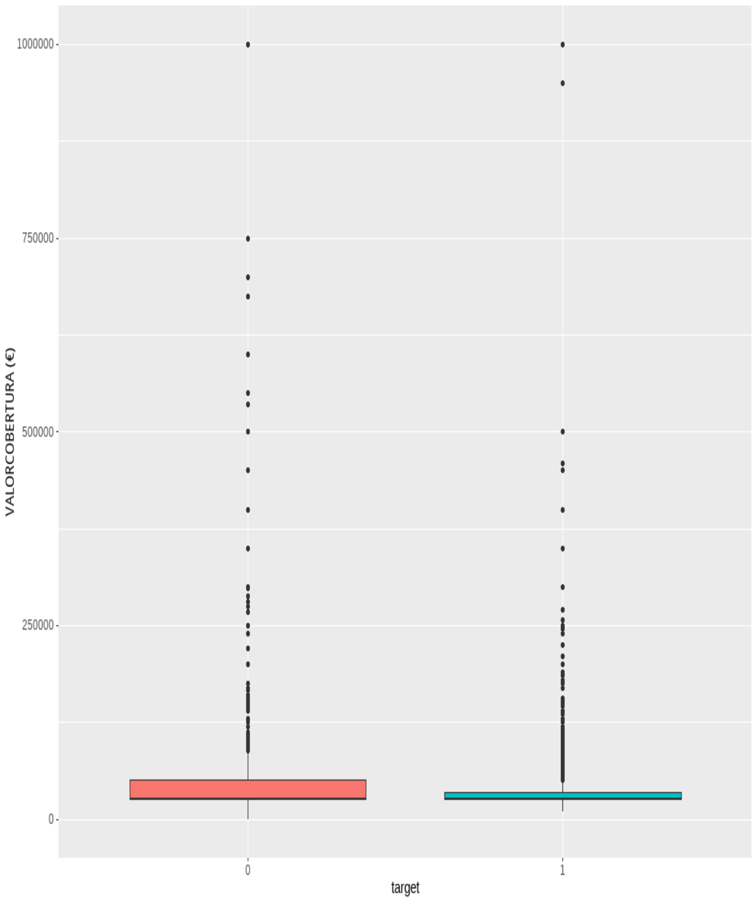
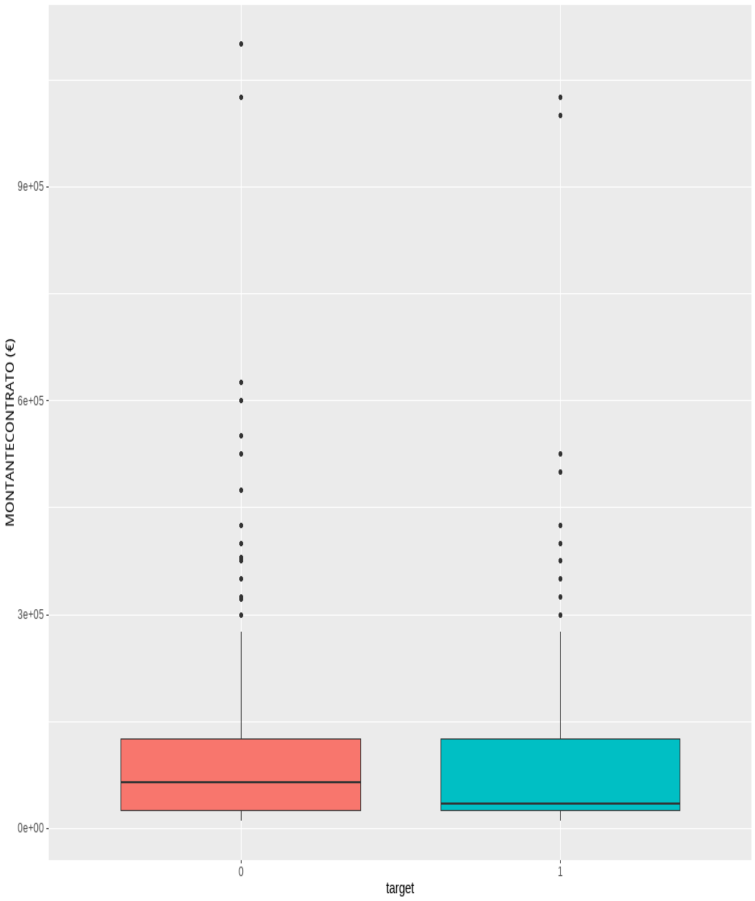


Figure B.22: Box plot Contract Amount. **Source:** Authors.



C Appendix 3: Performance Results Visualization

Table C.1: Performance Metrics (Training set). **Source:** Authors.

Model	Train Accuracy(Average)	TrainAUC-ROC(Average)	Standard Deviation(Train AUC-ROC)
Random Forest (rf)	0.8245	0.8222	0.25%
C5.0	0.8263	0.8180	0.36%
XGBoostLinearBooster	0.8265	0.8107	0.46%
BaggingClassifier	0.8077	0.7882	0.57%
XGBoostTreeBooster	0.8319	0.8218	0.26%
adaboost	0.8227	0.8131	0.39%
Logistic Regression (LR)	0.7607	0.7958	0.65%
Neural Networks (NN)	0.7460	0.7997	0.53%
glmnet	0.7741	0.8159	0.30%
Ensemble model (knn+nb)	0.7403	0.7508	0.11%
rpart	0.7881	0.7744	1.03%
Naive Bayes (nb)	0.7927	0.7613	0.63%
KNN	0.7579	0.6990	1.58%

Table C.2: Performance Metrics (Testing set). **Source:** Authors.

Model	Recall (Sensitivity)	**Specificity**	**PPV**	**NPV**	**Precision**	**F1-Score**	**Prevalence**	**Detection Prevalence**	Test Accuracy (Average)	Test Accuracy (95% Confidance Interval)	**Balanced Accuracy**	**Logarithmic Loss**	**Kappa**	Test AUC-ROC (Average)	**Test AUC-PR**
Random Forest (rf)	0.578	0.919	0.651	0.893	0.651	0.612	0.207	0.184	0.8482	(0.8407, 0.8556)	0.7483	0.345	0.518	0.8719	0.6425
C5.0	0.552	0.921	0.647	0.887	0.647	0.595	0.207	0.177	0.8447	(0.837, 0.8521)	0.7365	0.364	0.500	0.8658	0.6018
XGBoostLinearBooster	0.467	0.937	0.660	0.871	0.660	0.547	0.207	0.147	0.8398	(0.8321, 0.8473)	0.7021	0.359	0.453	0.8479	0.6601
BaggingClassifier	0.538	0.916	0.625	0.884	0.625	0.578	0.207	0.178	0.8375	(0.8297, 0.845)	0.7270	0.845	0.478	0.8441	0.5035
XGBoostTreeBooster	0.431	0.945	0.674	0.864	0.674	0.526	0.207	0.133	0.8389	(0.8311, 0.8465)	0.6882	0.375	0.434	0.8328	0.6405
adaboost	0.547	0.895	0.576	0.883	0.576	0.561	0.207	0.197	0.8228	(0.8147, 0.8306)	0.7209	0.592	0.450	0.8304	0.6378
Logistic Regression (LR)	0.718	0.797	0.481	0.916	0.481	0.576	0.207	0.309	0.7810	(0.7723, 0.7895)	0.7578	0.491	0.436	0.8298	0.6357
Neural Networks (NN)	0.736	0.781	0.468	0.919	0.468	0.572	0.207	0.326	0.7718	(0.7629, 0.7804)	0.7585	0.503	0.427	0.8297	0.4856
glmnet	0.696	0.802	0.479	0.910	0.479	0.567	0.207	0.301	0.7802	(0.7715, 0.7887)	0.7490	0.532	0.427	0.8238	0.6129
Ensemble model(knn+nb)	0.621	0.841	0.506	0.895	0.506	0.558	0.207	0.254	0.7958	(0.7873, 0.8041)	0.7312	0.554	0.427	0.8221	0.5600
rpart	0.697	0.806	0.483	0.911	0.483	0.571	0.207	0.299	0.7830	(0.7743, 0.7915)	0.7511	0.475	0.432	0.8094	0.4898
Naive Bayes (nb)	0.079	0.992	0.724	0.805	0.724	0.142	0.207	0.023	0.8030	(0.7946, 0.8112)	0.5355	3.401	0.106	0.7769	0.4833
KNN	0.579	0.856	0.512	0.886	0.512	0.544	0.207	0.234	0.7987	(0.7902, 0.8069)	0.7174	5.133	0.415	0.7680	0.1964

Figure C.1: Sensitivity Dotplot. **Source:** Authors.

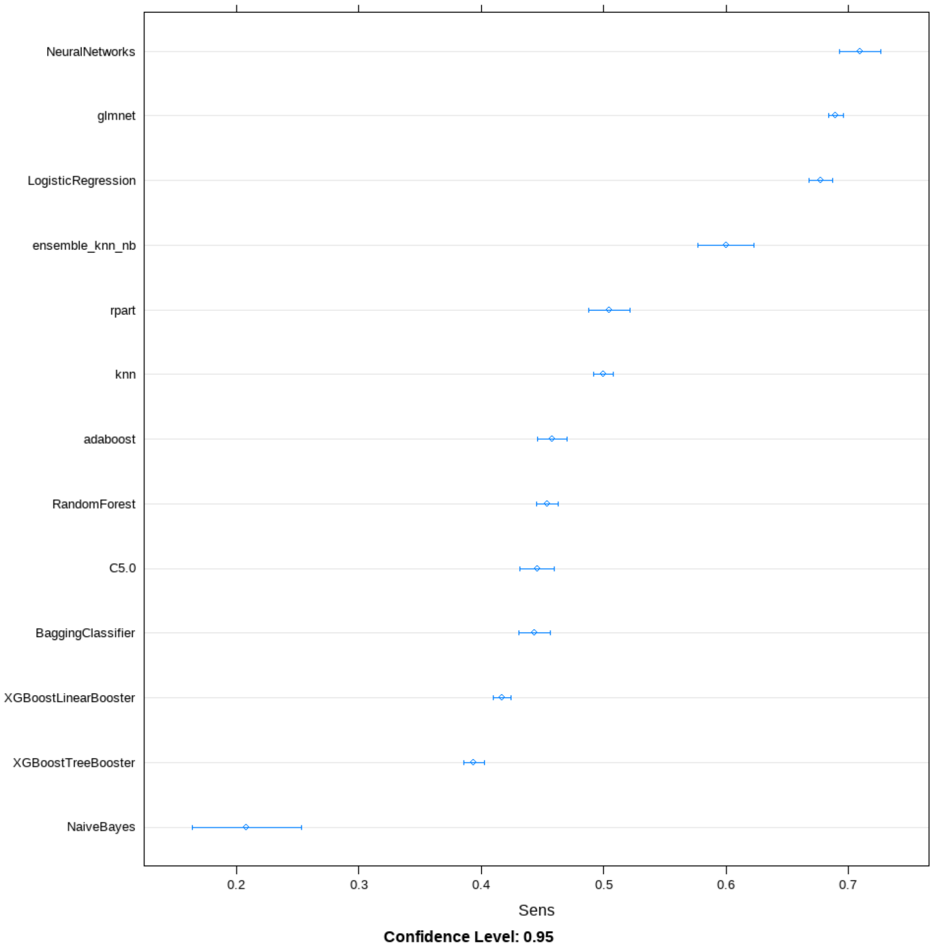


Figure C.2: Specificity Dotplot. **Source:** Authors.

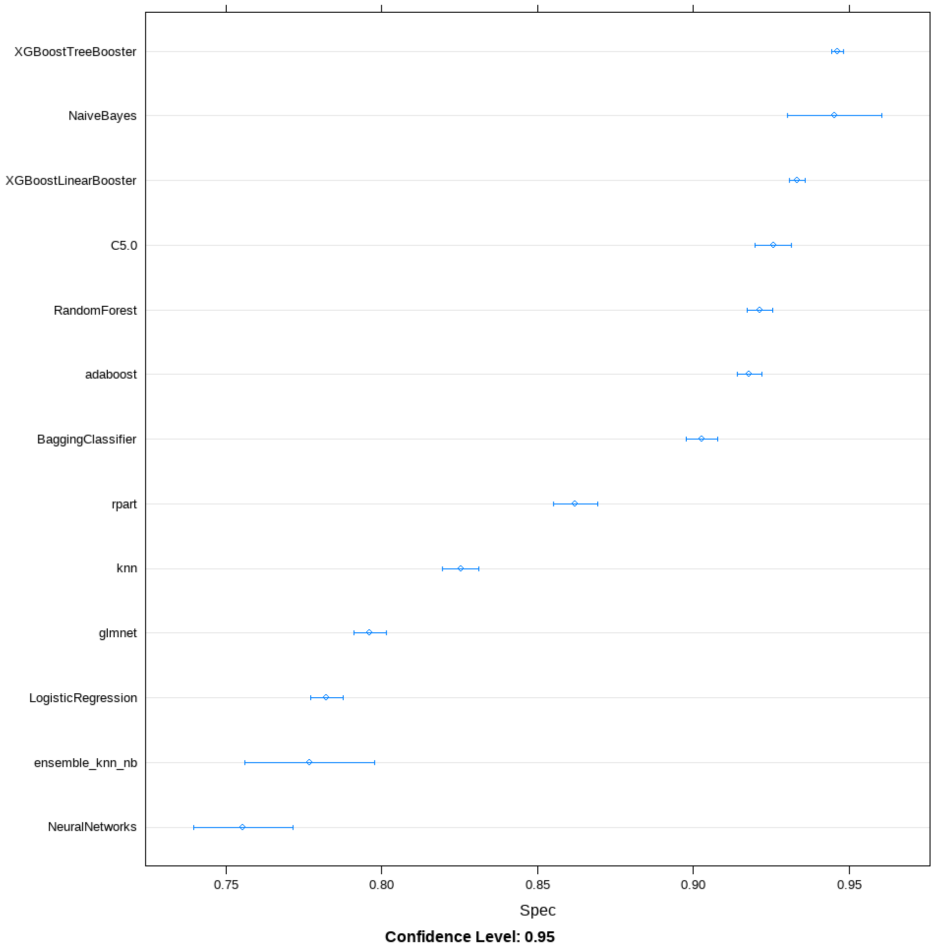


Figure C.3: AUC-PR Curves. **Source:** Authors.

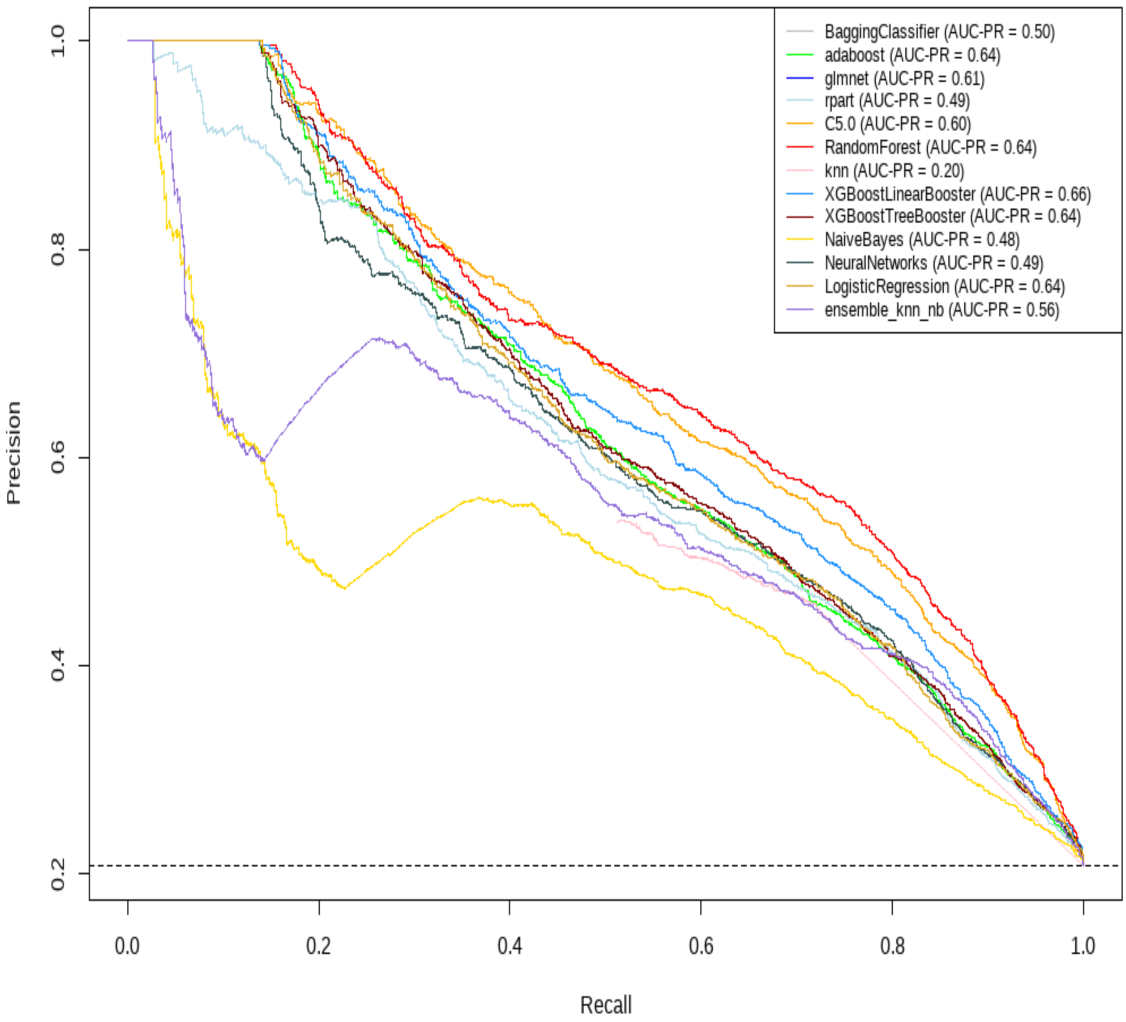
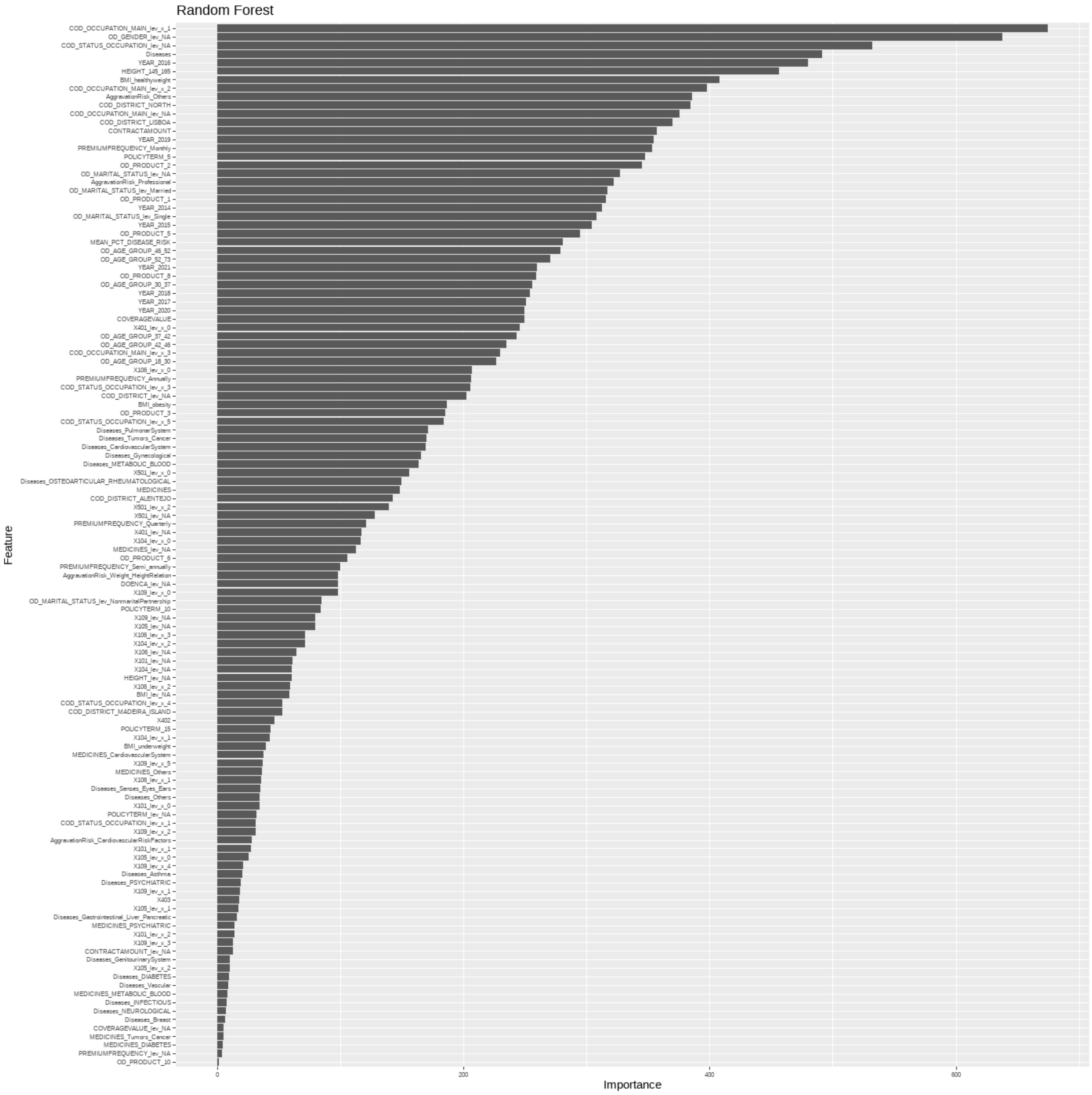


Figure C.4: Random Forest model Variable Importance. **Source:** Authors.



I Annex 1: Ensemble model (knn+nb)

Ensemble models are machine learning techniques that combine the predictions of multiple models to make a more accurate prediction. One type of ensemble model is the stacking model, which involves training multiple base models and then using a second level model to make a final prediction based on the predictions of the base models.

In this study, we used the *caretEnsemble* package in R version 2.0.1 to implement a stacking model with a combination of k-nearest neighbors (KNN) and Naive Bayes classifiers as the base models and a generalized linear model (GLM) as the second level model.

To evaluate the performance of this ensemble model, we split the data into a training set and a testing set and used 10-fold cross-validation to measure the model's accuracy. The results of the cross-validation showed that the ensemble model with KNN and naive Bayes base models and a GLM second level model had an average accuracy of 75.08% on the training set and 82.21% on the testing set.

Overall, the ensemble model with KNN and naive Bayes base models and a GLM second level model performed well in this study and provided valuable insights into the relationships between the features and the target variable.

II Annex 2: Hyperparameter Optimization

II.1 Elastic-Net Regularized Generalized Linear Model ('glmnet') Tuning

The glmnet model is capable of fitting multiple models simultaneously by exploring a wide range of lambda values, which control the amount of penalization in the model. The *train*() R-function is utilized to fit one model per alpha value and to simultaneously evaluate all lambda values.

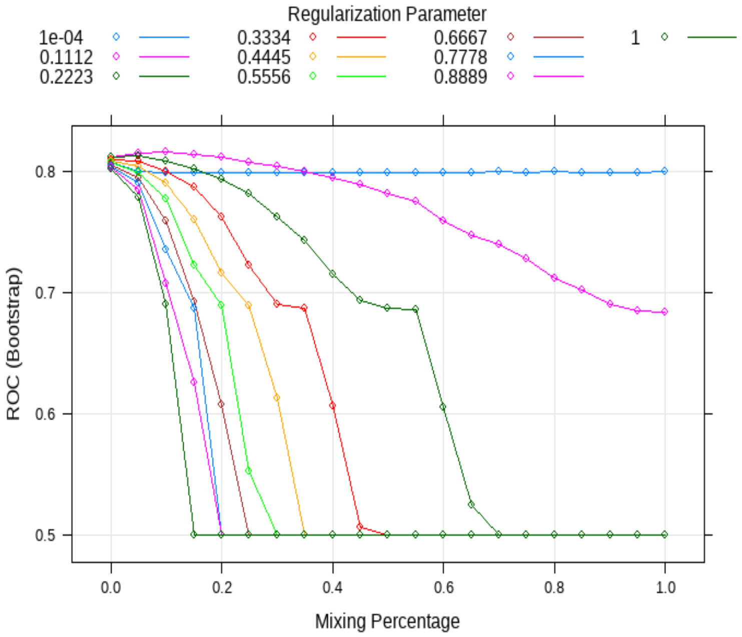
A manual grid search was performed to explore 10 randomly selected lambda values and 20 randomly selected alpha values, both ranging from 0 to 1, using the tuneGrid function in the train() R-function. This approach uses two forms of penalized models: ridge regression and lasso regression. Pure ridge regression is represented by alpha=0, while pure lasso regression is represented by alpha=1. Additionally, the mixture of these two models was also fitted to create an elastic-net model using an alpha value between 0 and 1.

The best hyperparameters for the glmnet model were determined to be alpha=0.1 and lambda=0.1112, as shown in Table II.1, resulting in an average AUC-ROC value of 0.816. The results of the model evaluation are presented in Figure II.1, which shows that the AUC-ROC values range from 0.5 to 0.816 based on cross-validation.

Table II.1: Glmnet Parameters Best Tuning **Source:** Authors.

parameter	Best Tune	class	label
alpha	0.1	numeric	Mixing Percentage
lambda	0.1112	numeric	Regularization Parameter

Figure II.1: ROC values per parameters (glmnet). **Source:** Authors.



II.2 Random Forest (RF) Tuning

The specific hyperparameters for the Random Forest (RF) model are presented in Table II.2, with a detailed explanation provided in Probst et al. (2019). The results of the model evaluation are shown in Figure II.2, which demonstrates that the splitting rule “extratrees“ outperforms other splitting rules in terms of AUC-ROC values. The figure also shows that the AUC-ROC values range from 0.798 to 0.822 based on 10-fold cross-validation. The optimal hyperparameter tuning is presented in Table II.3, which was automatically determined using the Caret R-package.

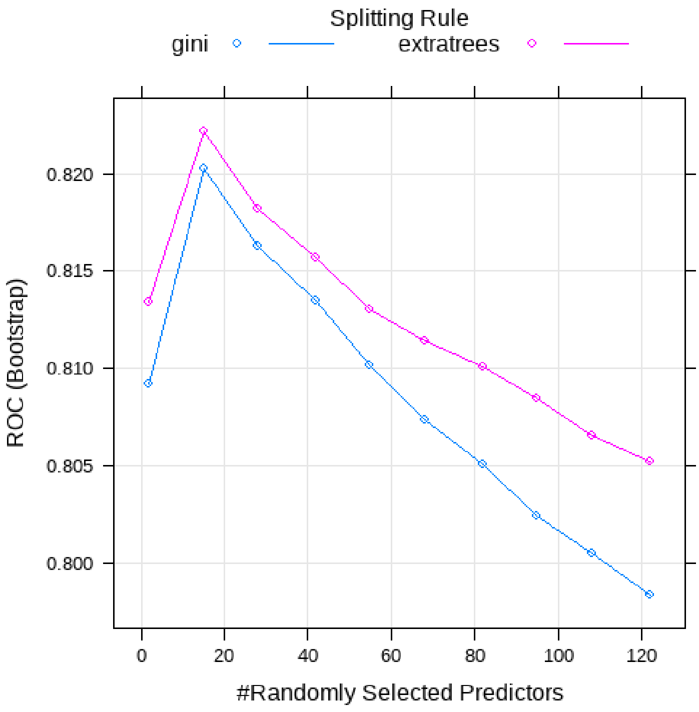
Table II.2: ROC values per parameters (RF)**Source:** Authors.

parameter	Trial Values	Best Tune	class	label
mtry	2, 15, 28, 42, 55, 68, 82, 95, 108, 122	15	numeric	#Randomly Selected Predictors
splitrule	gini, extratrees	extratrees	character	Splitting Rule
min.node.size	1	1	numeric	Minimal Node Size

Table II.3: Random Forest Best Tune**Source:** Authors.

Hyperparameter:	**Probability estimation**
Number of trees:	500
Sample size:	42792
Number of independent variables:	122
Mtry:	32
Target node size:	1
Variable importance mode:	impurity
Splitrule:	extratrees
Number of random splits:	1
OOB prediction error (Briers.):	0.0556

Figure II.2: ROC values per parameters (rf). **Source:** Authors.



II.3 Recursive Partitioning and Regression Trees (CART) Tuning

The optimal hyperparameters for the CART model are presented in Table II.4 below, which were determined by manually conducting a grid search of 20 randomly selected values of the complexity parameter (cp) ranging from 0.0001 to 0.001. The results of this search can be seen in Figure II.3, which shows how the AUC-ROC value changed from 0.77016 to 0.77118. As can be seen in Table II.5, the complexity parameter was tested on trees with varying numbers of splits, from 0 to 56. The best tree, with a cross-validated AUC-ROC of 0.77118, had 57 terminal nodes (56 splits). This tree represented the optimal balance between tree size and model performance.

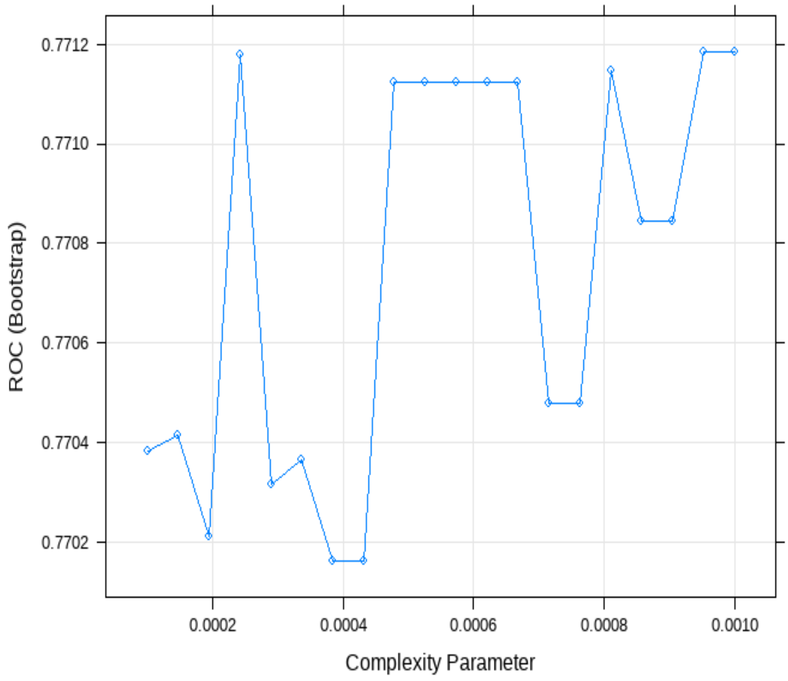
Table II.4: CART ("rpart") Best Tune **Source:** Authors.

parameter	Best_Tune	class	label
cp	0.001	numeric	Complexity Parameter

Table II.5: Complexity Parameters **Source:** Authors.

cp	nsplit	rel error
0.243	0	1
⋮	⋮	⋮
0.001	56	0.372

Figure II.3: ROC values per parameters (CART). **Source:** Authors.



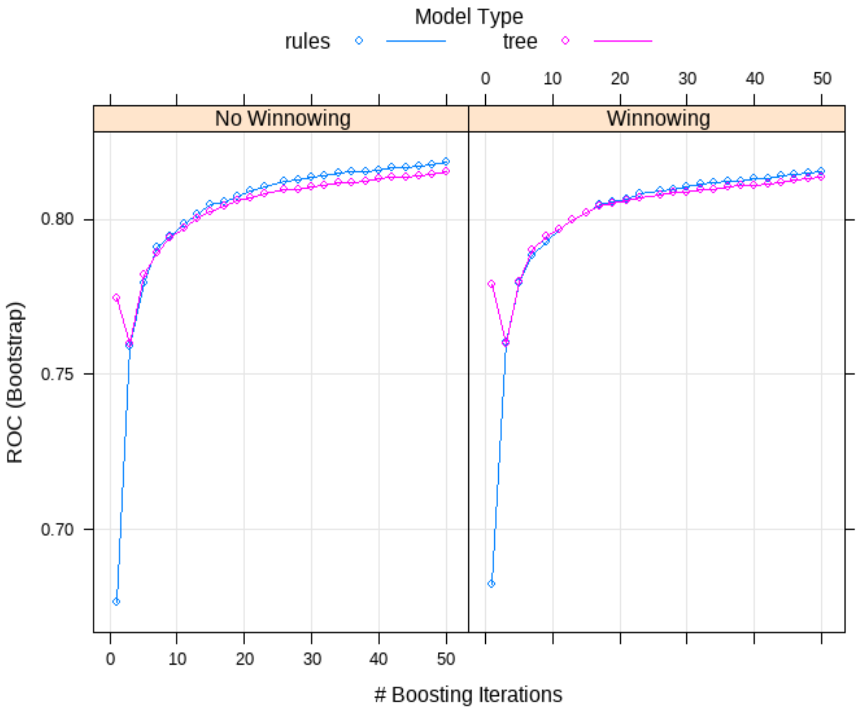
II.4 C5.0 Tuning

The best hyperparameters for the C5.0 model were carefully determined through a manual grid search process. The grid search considered several combinations of parameters such as the choice between tree-based or rule-based models, the use of winnowing for feature selection, and the number of random boosting trials ranging from zero to fifty. The results of the grid search are visible in Figure II.4, which shows that the AUC-ROC value varies from 0.6846 to 0.818 through cross-validation. The lowest AUC-ROC value was obtained when no boosting iterations were applied (i.e., trials=0). After thoroughly analyzing the results, the best hyperparameters tuning for the C5.0 model was found to be when the model was rule-based, winnowing was set to false, and 50 boosting trials were performed, resulting in an average AUC-ROC value of 0.818. These findings are detailed in Kuhn et al. (2015).

Table II.6: C5.0 Best Tune **Source:** Authors.

parameter	**Best Tune**	**class**	**label**
trials	50	numeric	Boosting Iterations
model	rules	character	Model Type
winnow	FALSE	logical	Winnow

Figure II.4: ROC values per parameters (C5.0). **Source:** Authors.



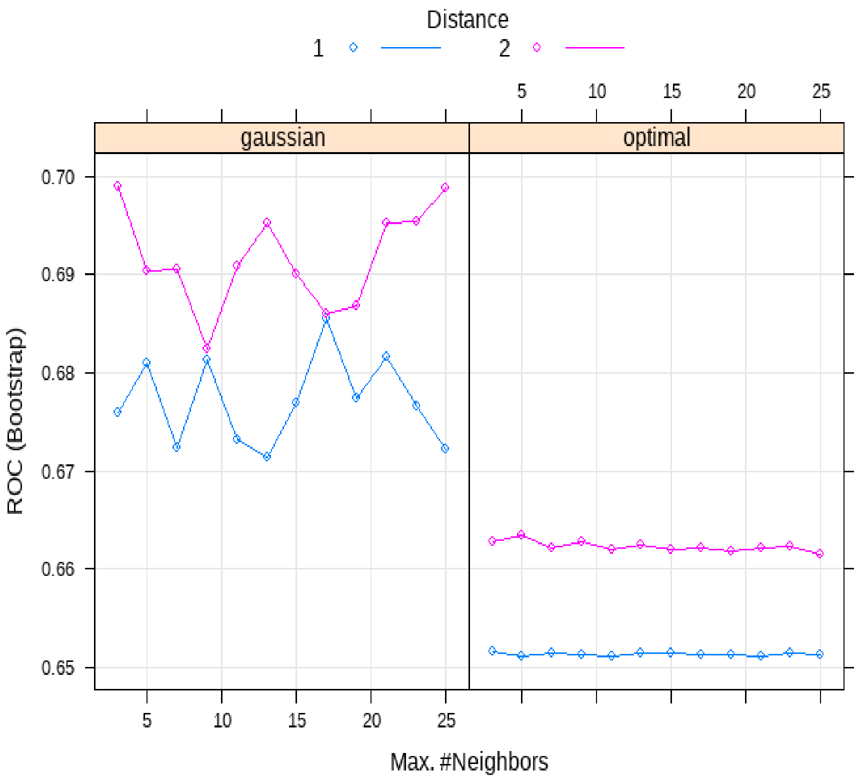
II.5 K-Nearest Neighbours (knn) Tuning

The best hyperparameters for the k-nearest neighbours (knn) model have been determined through a thorough manual tuning process, as detailed in Schliep et al. (2016). The grid search explored different combinations of parameters including the use of Euclidean distance (distance=2) or Manhattan distance (distance=1) as well as Gaussian kernel (gaussian), Optimal kernel (optimal), and twelve random values of k ranging from 3 to 25 (kmax). These results can be seen in the figure II.5, where the AUC-ROC value varied between 0.6615 and 0.699 based on cross-validation. The results show that in general, the Gaussian kernel achieved better performance. The best tuning parameters for the knn model were found to be kmax=3, distance=2, and kernel=gaussian, which had an average AUC-ROC of 0.699, as shown in table II.7.

Table II.7: KNN Best Tune **Source:** Authors.

parameter	Best Tune	class	label
kmax	3	numeric	Max. #Neighbors
distance	2	numeric	Distance
kernel	gaussian	character	Kernel

Figure II.5: ROC values per parameters (KNN). **Source:** Authors.



II.6 Extreme Gradient Boosting (XGBoost) tuning

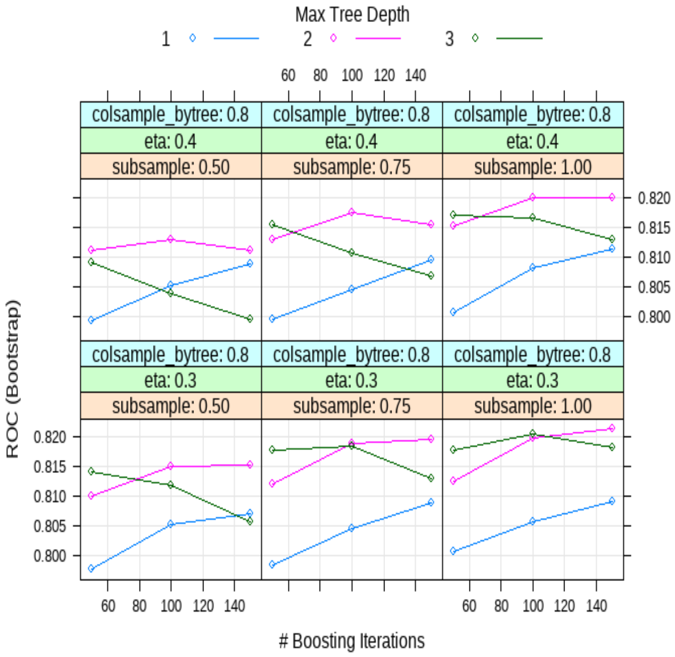
II.6.1 Tree Booster (XGBoostTreeBooster)

The performance of XGBoost with Tree Booster was evaluated using a grid search, as shown in table II.8 and explained in detail in Chen and Guestrin (2016). The parameters were automatically tuned by the Caret R-package, which produced the best performance of the combination of parameters with an AUC-ROC value of 0.82. The figure II.6 illustrates the influence of the parameter colsample_bytree on the AUC-ROC value, which varied from 0.7977 to 0.82 when the value of colsample_bytree was set to 0.8.

Table II.8: XGBoostTreeBooster Hyperparameters **Source:** Authors.

parameter	Trial _Values	Best_Tune	class	label
nrounds	50, 100, 150	150	numeric	# Boosting Iterations
max_depth	1, 2, 3	2	numeric	Max Tree Depth
eta	0.3, 0.4	0.3	numeric	Shrinkage
gamma	0	0	numeric	Minimum Loss Reduction
colsample_bytree	0.6, 0.8	0.8	numeric	Sub sample Ratio of Columns
min_child_weight	1	1	numeric	Minimum Sum of Instance Weight
subsample	0.5, 0.75, 1	1	numeric	Subsample Percentage

Figure II.6: ROC values per parameters (XGBoostTreeBooster). **Source:** Authors.



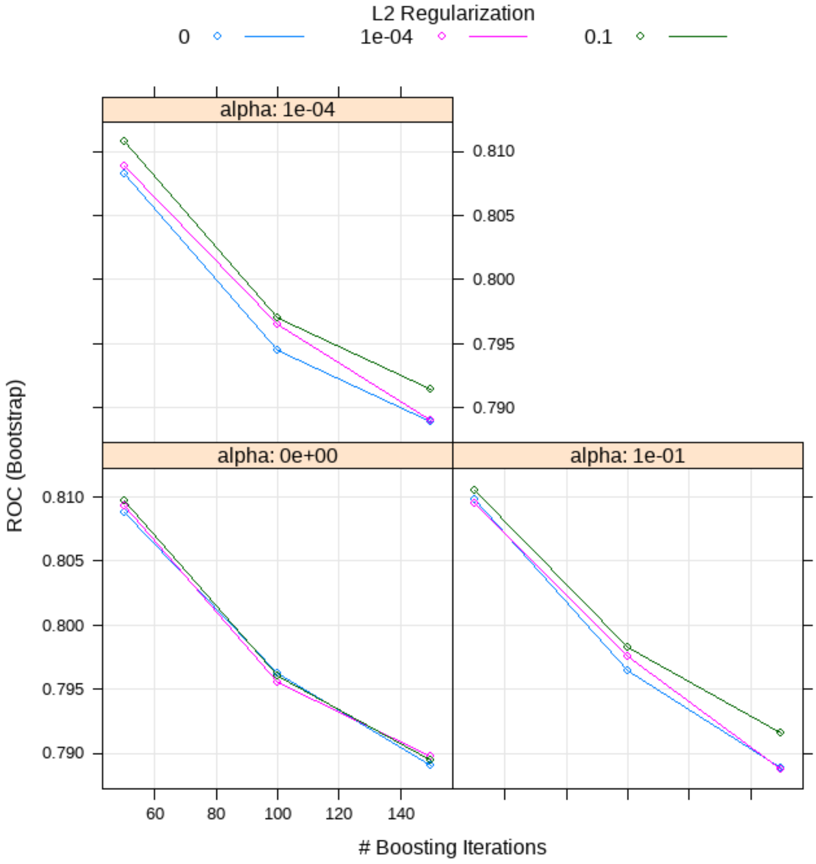
II.6.2 Linear Booster (XGBoostLinearBooster) Tuning

Similarly, the performance of XGBoost with Linear Booster was also evaluated using a grid search, as shown in table (see II.9) and explained in depth in Chen et al. (2019) and Chen and Guestrin (2016). The parameters were also automatically tuned by the Caret R-package. The best performance of the parameter combination (Best_Tune) achieved an AUC-ROC value of 0.8107, and the AUC-ROC value varied from 0.7888 to 0.8107 based on the different combinations of parameters, as illustrated in figure II.7.

Table II.9: XGBoostLinearBooster Hyperparameters **Source:** Authors.

parameter	Trial _Values	Best_Tune	class	label
nrounds	50, 100, 150	50	numeric	#Boosting Iterations
lambda	0, 0.0001, 0.1	0.1	numeric	L2 Regularization
alpha	0, 0.0001, 0.1	0.0001	numeric	L1 Regularization
eta	0.3	0.3	numeric	Learning Rate

Figure II.7: ROC values per parameters (XGBoostLinearBooster). **Source:** Authors.



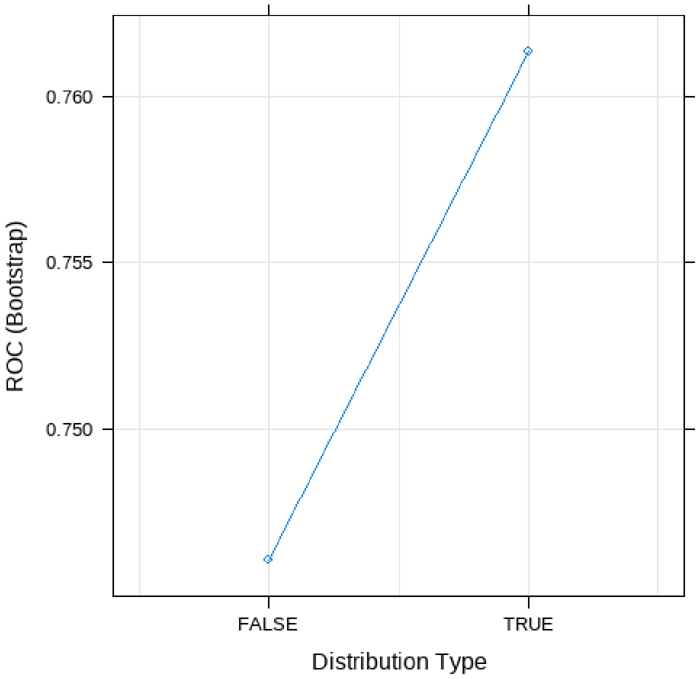
II.7 Naïve Bayes (nb) tuning

The hyperparameters for the Naive Bayes model were thoroughly tested in a grid search as shown in Table II.10 and explained in depth in reference Majka and Majka (2020). The Caret R-package was utilized to automatically tune the parameter values. The best combination of parameters (Best_Tune) achieved an AUC-ROC value of 0.76. The performance of two different parameter combinations can be seen in Figure II.8, where the AUC-ROC value varied based on the usekernel being set to either TRUE or FALSE.

Table II.10: Naïve Bayes (nb) Hyperparameters

parameter	Trial_Values	Best_Tune	class	label
laplace	0	0	numeric	Laplace Correction
usekernel	TRUE, FALSE	TRUE	logical	Distribution Type
adjust	1	1	numeric	Bandwidth Adjustment

Figure II.8: ROC values per parameters (nb). **Source:** Authors.



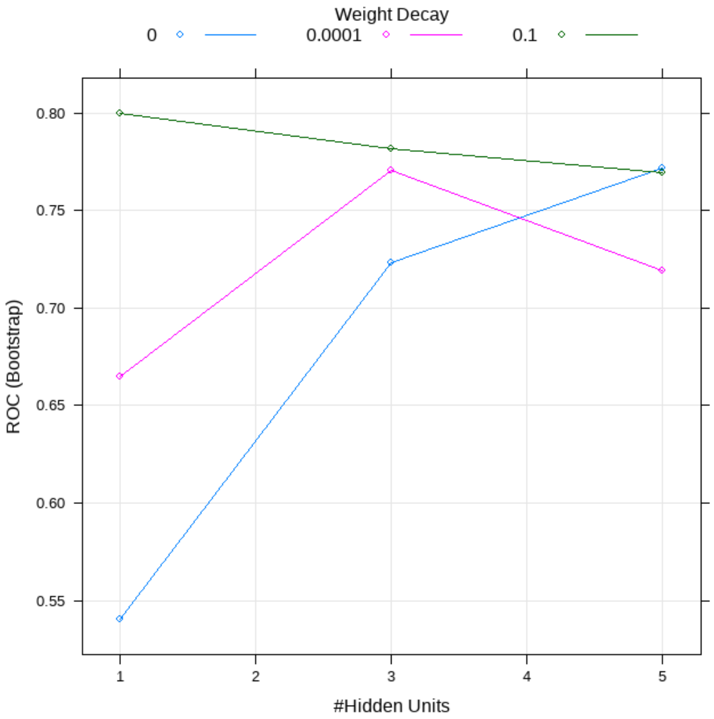
II.8 Neural Networks (NN) tuning

The hyperparameters for Neural Networks were extensively evaluated through a grid search as shown in Table II.11 and explained in detail in references Ripley et al. (2016). The Caret R-package was used to automate the tuning process. A 10-fold cross-validation was performed to train and evaluate neural networks with different combinations of hidden unit sizes and decay rates. The model with the highest mean AUC-ROC was selected as the best tuned model and was then evaluated on the test data to estimate its generalization performance. The best combination of parameters (Best_Tune) achieved an AUC-ROC value of 0.79977. Figure II.9 shows the performance of different parameter combinations.

Table II.11: Neural Networks (NN) Hyperparameters

parameter	Trial_Values	Best_Tune	class	label
size	1, 3, 5	1	numeric	#Hidden Units
decay	0, 0.0001, 0.1	0	numeric	Weight Decay

Figure II.9: ROC values per parameters (NN). **Source:** Authors.



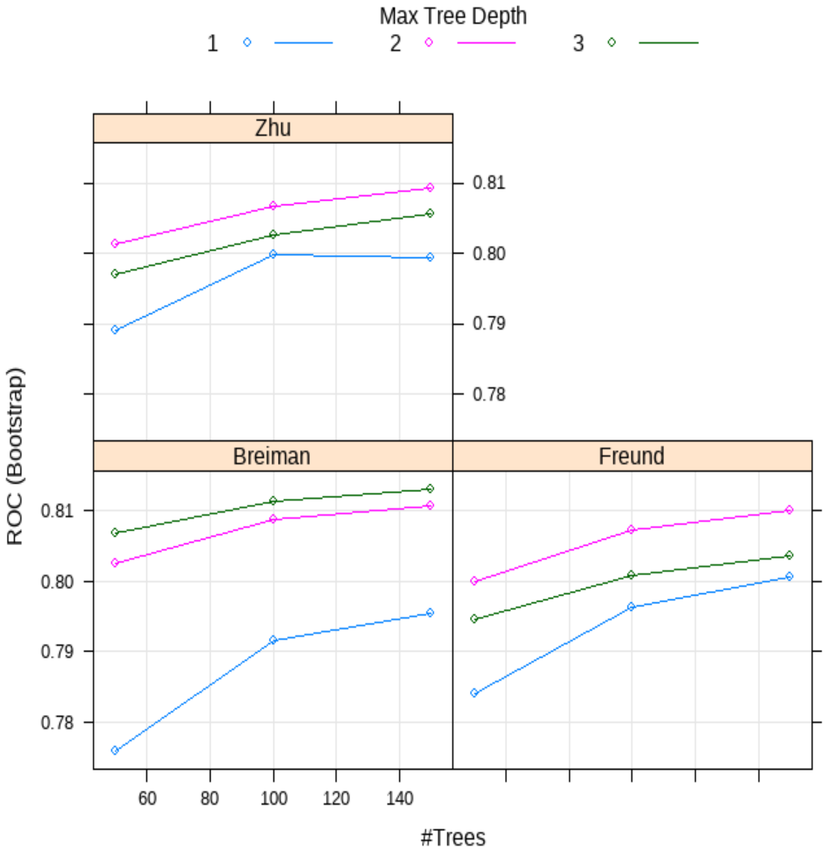
II.9 AdaBoost tuning

The hyperparameters for the AdaBoost model were analyzed through a grid search as shown in Table II.12 and explained in depth by reference Chatterjee et al. (2016). The Caret R-package was utilized to automatically tune the parameters. The best combination of parameters (Best_Tune) resulted in an AUC-ROC value of 0.813. The performance of different parameter combinations can be seen in Figure II.10, where the AUC-ROC value varied from 0.776 to 0.813.

Table II.12: AdaBoost Hyperparameters **Source:** Authors.

parameter	Trial_Values	Best_Tune	class	label
mfinal	50, 100, 150	150	numeric	#Trees
maxdepth	1, 2, 3	3	numeric	Max Tree Depth
coeflearn	Breiman, Freund, Zhu	Breiman	character	Coefficient Type

Figure II.10: ROC values per parameters (AdaBoost). **Source:** Authors.


Max Tree Depth
Zhu
Breiman
Freund
ROC (Bootstrap)
#Trees