Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

ACTUARIAL SCIENCES

MASTER'S FINAL WORK

Internship Report

PREDICTIVE MODELS FOR LAPSE AND MIDTERM CANCELATIONS FOR INDIVIDUAL HEALTH INSURANCE: A MACHINE LEARNING APPROACH

PEDRO DAVID CANTANTE DOMINGUES

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

ACTUARIAL SCIENCES

PREDICTIVE MODELS FOR LAPSE AND MIDTERM CANCELATIONS FOR INDIVIDUAL HEALTH INSURANCE: A MACHINE LEARNING APPROACH

PEDRO DAVID CANTANTE DOMINGUES

SUPERVISION:
JOÃO AFONSO BASTOS
HENRIQUE MARQUES

JUNE 2026


ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

GLOSSARY

AP – Actual premium

ASF- Autoridade de Supervisão de Seguros e Fundos de Pensões

GWP- Gross written premium

LP- Lapse rate

SNS- Serviço nacional de saúde

TP- Technical premium

DF- Data frame

LR- Logistic regression

DT- Decision tree

RF- Random forest

MLP- Multi-layer perceptron

OSS- Observatório dos seguros de saúde

ABSTRACT

Accurately estimating the probability of policyholder behavior events, specifically midterm cancellations and lapses, has become a key operational priority for insurers. Midterm cancellation models support portfolio management by assessing the likelihood that a policyholder will exit before the end of the policy year following renewal. Lapse models play a central role in pricing optimization, enabling the simulation of renewal outcomes under alternative contract conditions. Despite their practical importance, most insurers continue to rely on logistic regression for these tasks, while more advanced machine learning approaches remain relatively underexplored.

Motivated by a concrete business need at Allianz Portugal, this study evaluates six classification algorithms: Logistic Regression, Decision Trees, Random Forest, gradient boosting methods including Extreme Gradient Boosting (XGBoost) and LightGBM, and a Multi-Layer Perceptron. Model performance is assessed along two key dimensions, discrimination and calibration. In addition to standard evaluation criteria, calibration quality is examined using the Brier score and its decomposition, as well as Expected Calibration Error and Maximum Calibration Error. In this setting, accurate probability estimation is more important than simply distinguishing between classes. To enhance interpretability, SHAP values and feature importance measures are employed, with SHAP also informing the feature selection process.

The results indicate that gradient boosting methods, particularly LightGBM, achieve superior performance relative to logistic regression in both discrimination and calibration. These findings suggest that gradient boosting models provide a robust and practical alternative to the current industry standard. Furthermore, this study offers a structured analysis of the balance between model complexity, interpretability, and probability calibration, providing useful guidance for practitioners facing similar modeling challenges.

KEYWORDS: Lapse, Midterm Cancelation, Neural Networks, Binary Classification, Reliability, Isotonic Regression

JEL CODES: C45, C55, C53, C52

DISCLAIMER

This master's internship report was prepared in full compliance with the academic integrity standards of ISEG, Universidade de Lisboa. All content reflects my own research, analysis, and authorship, except where appropriate references are provided.

To ensure transparency, I acknowledge that certain artificial intelligence tools were used in the development of this project. GitHub Copilot was used to assist with coding and LaTeX formatting of the formulas. In addition, text assistance tools were used to support the improvement of clarity and refinement of the written content.

To protect sensitive corporate information, the names of variables and certain data elements have been omitted or anonymized.

Due to size constrains a GitHub repository was created, so that the reader can further consult model results and other analysis (https://github.com/pedrodavid99/MFW-rep).

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACKNOWLEDGMENTS

I dedicate this work to my grandfather, Mr. Manuel Gomes Cantante, for the irreplaceable role he has played in my life. From raising me to teaching me most of what I know, he has always been and continues to be my greatest tutor and role model.

To my parents, sister, and brother, thank you for always being there when I needed you and for never doubting me or my work.

To my cousins in Lisbon, who welcomed me into their home and treated me as one of their own throughout this master's without you, none of this would have been possible.

I would also like to extend my gratitude to the Allianz pricing team for their wonderful support and for welcoming me as a true member of the team. A special thank you to Henrique Marques for consistently motivating me, sharing his knowledge, and offering guidance whenever I needed it and to Leonardo Lopez, who was instrumental in helping me get this project off the ground.

To my dear friend Dr. Ana Maria for helping me with the small details regarding thesis construction.

And to my supervisor Dr. João Afonso Bastos, for the ideas and teachings that culminate in this study.

1. INTRODUCTION

Lapse is defined as the non-renewal of a policy at its renewal date. The policyholder may stop paying the premium or cancel the direct debit, leading to automatic non-renewal (Xu et al, 2015). Alternatively, the insured may formally notify the insurer of the intention not to renew, while continuing to benefit from coverage until the contract expires. In rare cases, the insurer may decide not to renew the policy. Such situations are uncommon and typically arise in cases where fraud is suspected but cannot be definitively proven, a practice sometimes referred to internally as "pruning" (SCOR, 2021).

Midterm cancellation, by contrast, refers to the termination of a policy prior to its scheduled renewal date. As health insurance contracts typically have a one-year coverage period, early cancellation most commonly occurs through the interruption of premium payments, which is only possible when premiums are paid in installments. When the premium is paid upfront as a lump sum, cancellation before expiry is significantly more restricted. In such cases, policyholders are generally required to demonstrate just cause, and the legally recognized circumstances for early termination are limited.

The objective of this Master's Final Work is to model the probabilities of lapse and midterm cancellation for individual health insurance policies, with the aim of generating estimates that can be effectively used for price optimization, portfolio management, and other key components of the insurer's operational cycle. Given their exceptional nature and the difficulty in reliably identifying these cases, insurer-initiated non-renewals are excluded from the modeling approach adopted in this study.

The central research question is whether modern machine learning algorithms are better suited than logistic regression, the current industry standard and if the outputs of the models are well calibrated probabilities, not just a binary class. Accordingly, model performance is evaluated in terms of discriminative ability and with respect to calibration quality. Probability calibration is performed using isotonic regression.

The models analyzed in this study include Logistic Regression, several tree-based algorithms, namely Decision Trees, Random Forest, Extreme Gradient Boosting and LightGBM, which are widely used in classification tasks and recognized for their strong predictive performance. In addition, a Multi-Layer Perceptron is included as a complementary approach to evaluate the potential of neural networks in this setting.

The main research questions addressed in this work are:

1. Do modern machine learning techniques produce reliable and well-calibrated probability estimates for lapse and midterm cancellation in health insurance?
2. Which factors have the greatest influence on lapse and midterm cancellation?
3. Which models are best suited to these tasks, and how should their performance be evaluated and compared?
4. What are the main limitations of the models and the available data, and what opportunities exist for further improvement?
5. Is logistic regression, as the current industry standard, still appropriate in this context?
6. To what extent does probability calibration improve predictive performance?

In addition, a concise theoretical overview of the methods employed is provided to support the reader's understanding of the underlying concepts. This includes the fundamental principles of the classification models considered, as well as explanations of SHAP values, performance metrics, isotonic regression, and the Optuna optimization framework.

2. BUSINESS UNDERSTANDING

2.1. The Portuguese Health System: A Historical View

In Portugal, following the 1974 revolution, the universal right to healthcare was recognized, although it was only formally enshrined in the Constitution in 1976 under Article 64º, which states "*Everyone has the right to health protection and the duty to defend and promote it*" (Assembleia da República, 1976).

Three years later, in 1979, the first law governing the National Health Service was enacted. This legislation gave concrete form to the constitutional principle of universal access, establishing a system characterized by universal and predominantly free access to healthcare. It also defined the fundamental principles, structure, and responsibilities of the state in this domain. The framework has been revised several times, most recently through Law No. 95/2019, which remains in force (Assembleia da República, 1979).

Despite the legal guarantee of universal access to healthcare on a predominantly free basis, economic theory suggests that publicly provided goods offered at low or no direct cost may be subject to inefficiencies in usage (Pauly, 1968). This phenomenon is commonly associated with the concepts of "tragedy of the commons" and the "free rider problem." In the context of healthcare, such dynamics can contribute to excessive demand and strain on available resources (Arrow, 1963).

As a result, the public healthcare system can become simultaneously overstretched and costly, potentially falling short of the standards established in legislation. Evidence of this pressure can be observed in reported user dissatisfaction. For example, the Portuguese Health Regulatory Authority received 43,664 complaints in the first semester of 2025, the majority of which were directed at public providers (Entidade Reguladora da Saúde, 2025). This figure may underestimate overall dissatisfaction, as individuals may be less inclined to report issues when services are provided at little or no direct cost, a pattern consistent with Expectation Disconfirmation Theory (Oliver, 1980).

This perceived gap in service quality within the SNS has created favorable conditions for the expansion of private healthcare providers and health insurance markets. According to Observatório dos Seguros de Saúde (OSS) data, dissatisfaction with the

quality of public healthcare is one of the main factors influencing the decision to purchase private health insurance (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2024).

2.1.1. Market Characteristics

Demand for health insurance in Portugal has been increasing steadily over recent years, largely driven by the well-documented deficiencies of the National Health Service (SNS). In 2024, 27 insurance companies offered health insurance products in the Portuguese market, of which 19 were domestic insurers, 2 operated under the Freedom to Provide Services regime, and 6 were branches of foreign insurers, according to Observatório dos Seguros de Saúde (OSS) data. Despite this relatively broad supply, market share remains highly concentrated among three main players, Fidelidade, Ageas, and Generali, although this concentration has gradually decreased over the available observation period (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2024).

Health insurance products can be broadly divided into two segments: individual and collective. The individual segment targets private policyholders and families, while the collective segment consists of business-to-business contracts offered to companies. Given the structural differences between these segments, particularly in pricing mechanisms and risk pooling, this study focuses exclusively on individual health insurance (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2021).

A survey conducted by the Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF) indicates that the primary reason for not purchasing health insurance is the perceived lack of need, with price identified as the second most important barrier. Interestingly, internal industry surveys suggest that the premium paid is the main driver of dissatisfaction among existing policyholders. At first glance, this may appear inconsistent with the structure of the market. Health insurance operates under conditions like monopolistic competition, in which firms offer differentiated but broadly comparable products, particularly in terms of coverage and service quality, suggesting that price

should not be the sole competitive dimension (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2024).

However, regulatory developments have increasingly emphasized price transparency. Circular 6/2025 introduced a standardized health insurance product that defines a baseline set of coverages, allowing for direct comparisons across insurers. Although the objective is to enhance consumer information and promote competition, insurers are not legally required to offer this standardized product. As a result, product differentiation remains relevant, even as prices become more salient (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2025).

From 2019 to 2024, both premiums and the number of contracts has increased. When these variables are plotted over time, as shown in Figure 1, a positive relationship between price and quantity emerges, with each observation corresponding to a different year (Autoridade de Supervisão de Seguros e Fundos de Pensões, 2024).

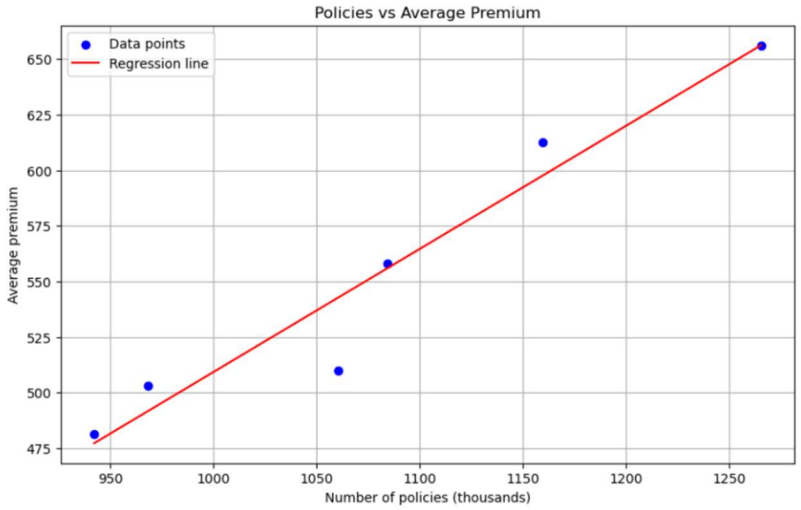

Figure 1 - Evolution of Price and Quantity in the Portuguese Health Insurance Market (2019–2024).

It is important to note that this pattern does not represent a standard demand curve or imply a positive price elasticity of demand. Instead, it reflects a sequence of shifting market equilibria over time. The simultaneous increase in both premiums and coverage suggests that the demand curve has been progressively shifting outward.

Several factors likely contribute to this dynamic, including the declining perceived quality of public healthcare, increasing waiting times, and growing awareness of private healthcare alternatives. As a result, each successive market equilibrium is characterized by both higher prices and higher quantities.

This interpretation supports the view that health insurance in Portugal is increasingly perceived not as a discretionary purchase, but as a necessity. Consequently, the observed growth in both premiums and coverage is better explained by structural increases in demand than by changes in price sensitivity alone. Importantly, this does not imply that health insurance behaves as a Veblen good. Rather, it reflects evolving demand conditions driven by changes in the broader healthcare environment.

3. Data Understanding and preparation

Data quality has been widely identified as a primary determinant of machine learning model performance, with cascading data quality issues shown to systematically degrade downstream model outputs (Sambasivan et al., 2021). Accordingly, a substantial portion of this project was dedicated to data cleaning and preparation.

The modeling process was based on a database provided by Allianz Portugal containing event-level data on individual health insurance policies, covering the period from 2022 to January 2026. Each row corresponded to a distinct event, including claims, renewals, cancellations, and other policy-related occurrences, resulting in a raw dataset comprising 1,845,223 rows and 497 columns. Many of these variables were redundant, auxiliary, or not directly suitable as model inputs, although several proved valuable for feature engineering.

The central objective of this phase was to transform this event-based and highly granular dataset into a clean, person-level data frame, in which each row corresponded to a single insured individual and the features captured a structured summary of both personal characteristics and policy history.

Two separate datasets were constructed: one for lapse modeling and another for midterm cancellation modeling. For the lapse model, the target variable was defined as the acceptance or rejection of the proposed renewal. The observation window was defined

as the full year of 2025. For each insured individual, the data where truncated at the corresponding renewal date, ensuring that the feature vector included only information available at the time of the renewal decision. This is a critical methodological constraint, as failure to enforce such a temporal boundary introduces data leakage, whereby future information contaminates the model inputs, leading to artificially inflated in-sample performance and poor generalization.

For the midterm cancellation model, the objective was to estimate the probability of cancellation within one year following the most recent renewal. The reference renewal year was 2024, and, similarly, the data for each insured were truncated at the renewal date. The outcome variable was defined over an observation window of one year minus one day, with the one-day adjustment implemented to exclude renewal-date lapses, which constitute a distinct event type and are modeled separately.

As an initial step, all variables with constant values across the dataset were removed, reducing the feature space to 419 variables. At this stage, the dataset contained 69,884 unique policies. Despite this reduction, the dataset remained highly redundant. For instance, geographic information was duplicated across multiple fields, including postal codes, municipalities, and parishes often originating from different data sources and encoded inconsistently.

The available features can be broadly categorized into five groups: premium-related variables, individual and policy characteristics, socio-demographic attributes of the insured's geographic area, healthcare supply indicators, and datetime-related variables.

The main data preparation steps included the removal of erroneous entries, missing value treatment, deduplication, limited outlier handling, and feature engineering. Given the constraints of this report, the description focuses primarily on the lapse dataset, as the midterm cancellation dataset underwent an analogous preparation process with minor adjustments reflecting differences in the modeling objectives.

Observations with missing premium information, specifically both the previously paid premium and the proposed renewal premium, were removed, as these variables are central to the modeling task. Additionally, policies with less than nine months of exposure were excluded to ensure a sufficient observation window. For categorical variables with

inconsistent encoding, mapping dictionaries were constructed to consolidate equivalent categories into a standardized format.

3.1. Missing value treatment

Different imputation strategies were applied depending on the type of variable and the extent of missingness. Variables with more than 80% missing values were removed entirely. For variables with only a small proportion of missing entries, the corresponding observations were excluded.

For socio-demographic variables, which are geographically determined, a hierarchical postcode-based mode imputation strategy was employed. Missing values were imputed using the modal value among observations sharing the same postcode. When insufficient data were available at the most granular level, the postcode was progressively aggregated by truncating digits, expanding the geographic scope from highly local to more regional levels until a valid imputation was obtained. This approach ensures that imputed values remain as locally representative as possible, while providing a systematic fallback mechanism.

Observations with missing or invalid postal codes, such as entries recorded as "0000-000" or "-", were removed, as they represented a negligible share of the dataset. For a limited number of remaining missing socio-demographic variables, manual validation was performed. In these cases, clear associations between locations and missing values allowed the correct information to be retrieved and inserted directly.

For continuous variables such as weight and height, missing values were imputed using an age-group segmentation approach. The population was divided into age bands, and group-specific means were computed using physiologically valid records. Observations containing invalid values, such as zero entries or implausible measurements (e.g., weight above 150 kg or height above 2.2 meters), were treated as missing and replaced with the corresponding group mean. This approach preserves the underlying demographic structure of the data and supports the reliable computation of derived metrics such as Body Mass Index.

3.2. Outlier treatment

Outlier handling was not the primary focus of this project. In the context of insurance data, aggressive removal of extreme observations may distort the tails of the distribution, which often contain economically significant cases. Furthermore, tree-based models, which constitute the primary modeling framework in this study, are inherently robust to outliers, as they rely on threshold-based partitioning rather than distance-based measures.

3.3. Feature engineering

Given the event-based nature of raw data, feature engineering was primarily aimed at aggregating individual histories into a single observation per insured. Group-by operations and pivot tables were extensively used to compute summary statistics and derive meaningful ratios from historical records.

This step proved particularly important, as engineered features consistently ranked among the most informative predictors in subsequent model evaluations, frequently appearing within the top 20 variables by importance.

Following feature construction, variables unsuitable for modeling were removed, including date fields, surrogate identifiers, variables referencing outdated periods, and intermediate variables used solely for feature derivation.

For categorical variables with sparse categories, low-frequency levels were merged into broader categories to ensure a minimum number of observations per level. This improves the statistical reliability of category-level estimates and reduces the dimensionality introduced by one-hot encoding.

3.4. Correlation-based feature reduction

After all data preparation steps, the dataset comprised 60,700 individual-level observations and 182 variables. Prior to modeling, a correlation analysis was conducted to identify and remove redundant numerical features.

Given the dimensionality of the dataset, computing a full correlation matrix across all variables was not feasible. Instead, variables were grouped into thematic subsets, including premium-related variables, claims-related features, socio-demographic attributes, and personal characteristics. Correlations were computed within each group, and redundant variables were removed iteratively.

For premium, claims, and personal variables, highly correlated pairs were assessed manually. In many cases, these reflected the same information recorded at different aggregation levels, such as policy-level versus individual-level variables. In such cases, policy-level variables were removed, both to reduce redundancy and to maintain consistency with the individual-level prediction objective.

For socio-demographic variables, a more automated approach was adopted. Variables exhibiting an absolute Pearson correlation coefficient above 0.80 were subjected to random elimination, ensuring that only one variable from each highly correlated pair was retained. This approach preserves the overall informational content while reducing dimensionality. It is also supported by existing literature, which suggests that socio-demographic variables often contribute with limited incremental predictive power in insurance churn models (Eling & Kiesenbauer, 2014).

Additionally, certain variables with deterministic relationships were removed *a priori*. For example, the proportion of males and the proportion of females within a group are perfectly negatively correlated. In such cases, one variable was removed without explicit correlation computation, as the redundancy is mathematically guaranteed.

This correlation-based reduction constituted the second stage of the feature selection process. A third, model-driven feature selection stage was conducted after the initial modeling phase and is discussed in the subsequent chapter.

4. Machine Learning

"*Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed*" (Samuel, 1959).

Machine learning is a subfield of artificial intelligence concerned with the development of algorithms capable of learning patterns from data and improving their

performance on a given task through experience. In practice, a model is trained on historical data to infer a function that can generate predictions or support decision-making on unseen observations. This learning process is typically guided by an objective measure, such as prediction error or a cost function, and implemented through optimization procedures that adjust model parameters to minimize a predefined loss function (Samuel, 1959). When successful, this enables the model to generalize beyond the training data.

A more formal articulation of this idea is offered by Mitchell: "*A computer program (P) is said to learn from experience (E) with respect to some Task (T) and some performance measurement (Pe), if its performance on (T), as measured by (Pe) improves with experience (E)"* (Mitchell, 1997).

Machine learning problems are commonly classified into supervised learning, unsupervised learning, and reinforcement learning. The present work falls within the supervised learning paradigm, where models are trained on labeled data, that is, input variables and corresponding outputs, with the objective of learning a mapping that generalizes unseen data (Feurer et al, 2019).

Within supervised learning, algorithms can be applied to several tasks, including regression, classification, time-series forecasting, anomaly detection, and ranking (Mitchell, 1997). This project is formulated as a binary classification problem, where the objective is to estimate the probability that a policyholder will lapse or cancel their health insurance. However, the primary focus is not merely class assignment, but the estimation of well-calibrated probabilities. Accordingly, the analysis focuses on classification algorithms capable of producing probabilistic outputs.

4.1. Models Used

To address the research objectives of modeling lapse and midterm cancellation probabilities, six classification algorithms were employed: Logistic Regression, Decision Trees, Random Forest, Extreme Gradient Boosting, LightGBM, and a Multi-Layer Perceptron.

Throughout this section, $\mathbf{X}^{(i)}$ denotes the feature vector of individual i, y_i denotes the observed binary outcome, and p_i denotes the model-estimated probability that individual i belongs to the positive class.

4.1.1. Logistic Regression

Logistic regression is a supervised learning algorithm commonly used for binary classification, where the objective is to estimate the probability that an observation belongs to the positive class (Hastie et al., 2009). In this study, the objective is to model:

$$(1) \qquad p_i = P\left(y_i = 1 \mid \mathbf{X}^{(i)}\right), \qquad y_i \in \{0,1\}.$$

Logistic regression models the log-odds of the positive outcome as a linear function of the explanatory variables:

$$(2) \quad z_i = \boldsymbol{\beta}^T \mathbf{X}^{(i)} = \log\left(\frac{p_i}{1-p_i}\right), \quad z^{(i)} \in R,$$

where $\boldsymbol{\beta}$ is the parameters vector. This value is mapped to the probability space using the sigmoid function in Equation (3), ensuring that predicted values lie in]0,1[and can be interpreted as probabilities

$$(3) \qquad \sigma(z_i) = \frac{1}{1+e^{-z_i}} \ .$$

Model parameters are estimated by iteratively minimizing the negative log-likelihood function, also known as the cross-entropy loss:

$$(4) \quad J(\boldsymbol{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(z_i)\right) + (1-y_i)\log\left(1-\sigma(z_i)\right)\right],$$

where N is the number of observations. Logistic regression is therefore a natural benchmark in this study, as it directly produces probability estimates and remains widely used in insurance applications.

4.1.2. Decision Tree

Decision Trees are non-parametric supervised learning algorithms used for classification and regression tasks (Breiman et al., 1984). They recursively split the

feature space into regions based on decision rules, where each internal node represents a split and each leaf node provides the final prediction.

In classification, splits are selected to reduce node impurity. Since the Decision Tree model in this study uses the Gini criterion, impurity is defined as:

$$(5) \quad G(\mathrm{D}) = 1 - \sum_{c} p(c)^2,$$

where $G(\mathrm{D})$, represents the Gini impurity of dataset D, and $p(c)$, represents the proportion of observations belonging to class c.

Decision Trees are intuitive and interpretable, but they are prone to overfitting specially if hyper parameters are not properly tuned. Their predicted probabilities are based on class frequencies within terminal leaves, which can make them unstable when leaves contain few observations.

4.1.3. Ensemble Techniques

No machine learning algorithm achieves perfect predictive performance. Part of this limitation is due to irreducible error, which arises from noise in the data and cannot be eliminated regardless of the model used. However, another part of the prediction error is reducible and is commonly associated with bias, which reflects underfitting, and variance, which reflects sensitivity to the training data (Hastie et al., 2009). Ensemble methods aim to improve predictive performance by combining multiple models instead of relying on a single learner.

Ensemble techniques are commonly divided into several strategies. Bagging reduces variance by training multiple models on different bootstrap samples and averaging their predictions (Breiman, 1996). Random Forests extend this idea by combining bagging with random feature selection at each split (Breiman, 2001). Boosting builds models sequentially, where each new learner attempts to correct the errors of the existing ensemble. This approach is the basis of gradient boosting methods such as XGBoost and LightGBM. Stacking, another ensemble strategy, combines the predictions of several base models using a higher-level meta-learner. In this project, the ensemble models considered are Random Forest, XGBoost, and LightGBM (Ke et al., 2017).

4.1.3.1. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce variance relative to a single tree. It is based on bagging and random feature selection. During training, several bootstrap samples are drawn from the original dataset, and a decision tree is trained on each sample. At each split, only a random subset of features is considered, reducing correlation between trees and improving generalization (Breiman, 2001).

In the original paper, for classification tasks, class labels are obtained through majority voting, while class probabilities come as the frequency of positive class across all trees. Although in this paper we use the scikit learn implementation, thus probabilities defined in Equation (1) are estimated by averaging the predicted probabilities across all trees:

$$(6) \quad p_i = \frac{1}{B}\sum_{b=1}^{B} p_{b,i} \,,$$

where B, represents the number of trees, and $p_{b,i}$ represents the probability estimated by tree b for individual i.

Random Forest performs well on high-dimensional tabular data, captures non-linear relationships, and is relatively robust to noise and multicollinearity. However, Random Forest is less transparent than a single Decision Tree and can become computationally expensive when the number of trees is large. Although Random Forest generally produces more stable probabilities than individual trees, these probabilities may still require calibration (Breiman, 2001).

4.1.3.2. Gradient Boosting Framework

Gradient boosting is a sequential ensemble learning framework in which models are built additively. At each iteration, a new tree is fitted to the negative gradient of a differentiable loss function, also known as pseudo-residuals, allowing the model to improve the current ensemble step by step. Unlike Random Forest, where trees are trained

independently and then aggregated, gradient boosting constructs trees sequentially, with each new tree aiming to correct the errors of the existing ensemble (Friedman, 2001).

Let $F_m(\boldsymbol{X}^{(i)})$ denote the ensemble prediction for individual i after m boosting iterations. In binary classification, this prediction is defined on the log-odds scale. The model is updated as:

$$(7) \quad F_m(\boldsymbol{X}^{(i)}) = F_{m-1}(\boldsymbol{X}^{(i)}) + \nu f_m(\boldsymbol{X}^{(i)}),$$

where $f_m(\boldsymbol{X}^{(i)})$ represents the regression tree added at iteration m, and ν represents the learning rate.

$F_m(\boldsymbol{X}^{(i)})$ is then passed to the sigmoid function in Equation (1) to get a probabilistic result.

4.1.3.3. Extreme Gradient Boosting

Extreme Gradient Boosting, commonly known as XGBoost, is an optimized implementation of the gradient boosting framework designed to improve predictive performance, scalability, and regularization (Chen & Guestrin, 2016). Like standard gradient boosting, XGBoost builds trees sequentially, with each tree contributing a correction to the current raw prediction score.

A central feature of XGBoost is its regularized objective function:

$$(8) \quad \mathcal{L} = \sum_{i=1}^{N} l\left(y_i, F(\boldsymbol{X}^{(i)})\right) + \sum_{m=1}^{M} \Omega(f_m),$$

where $l\left(y_i, F(\boldsymbol{X}^{(i)})\right)$, represents the loss function, $F(\boldsymbol{X}^{i})$, represents the raw model score for individual i, and $\Omega(f_m)$, represents the regularization term associated with tree m.

The regularization term penalizes tree complexity, encouraging simpler trees and reducing the risk of overfitting. XGBoost also includes computational improvements such as shrinkage, column subsampling, efficient split finding, and built-in handling of missing values. These features make it suitable for complex tabular datasets (Chen & Guestrin, 2016). In this study, XGBoost is relevant because it can model non-linear relationships and generate probability estimates for lapse and midterm cancellation. Nevertheless, these

probabilities may still require post-hoc calibration when probability accuracy is the main objective.

4.1.3.4. LightGBM

LightGBM is another optimized implementation of gradient boosting, designed for efficiency and scalability on large and high-dimensional datasets (Ke et al., 2017). Like XGBoost, LightGBM builds trees sequentially and produces raw scores that are transformed into probabilities for classification tasks.

The main distinction of LightGBM is its leaf-wise tree growth strategy. Instead of growing trees level by level, LightGBM expands the leaf that produces the largest reduction in loss (Ke et al., 2017). This can improve predictive performance and speed up convergence, but it may also increase overfitting risk if parameters such as the number of leaves, the maximum tree depth, and the minimum number of instances per leaf are not properly controlled.

LightGBM also uses histogram-based splitting, where continuous variables are discretized into bins to reduce memory usage and computational cost. Additional efficiency techniques include Gradient-based One-Side Sampling and Exclusive Feature Bundling, which improve training speed in large and sparse datasets (Ke et al, 2017). In this project, LightGBM is particularly relevant because it combines strong predictive performance with computational efficiency.

4.1.4. Multi-Layer Perceptron

A Multi-Layer Perceptron is a feedforward neural network composed of an input layer, one or more hidden layers, and an output layer. Each neuron in a layer is connected to neurons in the following layer, making the architecture fully connected. For binary classification, the output layer produces a value between 0 and 1, interpreted as the predicted probability of belonging to the positive class.

The transformation performed at layer l can be written as:

$$\boldsymbol{z}^{(l)} = \boldsymbol{W}^{(l)} \cdot \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}, \quad (9)$$

where $\boldsymbol{z}^{(l)}$, represents the pre-activation vector at layer l, $\boldsymbol{W}^{(l)}$, represents the weight matrix, $\boldsymbol{a}^{(l-1)}$, represents the activation vector from the previous layer, and $\boldsymbol{b}^{(l)}$, represents the bias vector. A non-linear activation function is then applied:

$$(10) \quad \boldsymbol{a}^{(l)} = \phi\left(\boldsymbol{z}^{(l)}\right),$$

where $\phi(\cdot)$ represents the activation function. In this study, the hyperbolic tangent function was used in the hidden layers:

$$(11) \quad \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} .$$

The sigmoid function in Equation (1) is then applied to the output layer to produce a probability estimate. The network is trained by minimizing binary cross-entropy using backpropagation (Rumelhart et al., 1986). The Adam optimizer was used to update model parameters, as Adam adaptively scales learning rates based on past gradients (Kingma & Ba, 2015).

Although Multi-Layer Perceptron can capture complex non-linear relationships, they are sensitive to hyperparameter choices and prone to overfitting, particularly in structured tabular datasets. For this reason, regularization techniques such as dropout or other regularization techniques are often used (Grinsztajn et al., 2022).

5. Performance Metrics

This chapter introduces the performance metrics used to evaluate the binary classification models. Since the objective of this study is not only to distinguish between policyholders who lapse or cancel and those who do not, but also to obtain reliable probability estimates, model performance is assessed along two dimensions: discrimination and calibration.

Discrimination refers to the model's ability to rank positive cases above negative cases. Calibration, by contrast, evaluates whether predicted probabilities correspond to observed event frequencies. This distinction is particularly important in this study, since the predicted probabilities may be used for pricing optimization, portfolio management, and retention decisions (Grinsztajn et al, 2022).

5.1. Threshold-Dependent Classification Metrics

Threshold-dependent metrics require converting predicted probabilities into class labels using a classification threshold τ. For example, an observation will be classified as positive if $p_i \geq \tau$.

The confusion matrix summarizes the relationship between predicted and observed classes (Powers, 2011). In this study, the positive class corresponds to lapse or midterm cancellation, while the negative class corresponds to no cancellation event.

The four elements of the confusion matrix are:

- **True Positives (TP):** observations correctly predicted as positive.
- **True Negatives (TN):** observations correctly predicted as negative.
- **False Positives (FP):** observations incorrectly predicted as positive.
- **False Negatives (FN):** observations incorrectly predicted as negative.

Accuracy

Accuracy measures the proportion of observations that are correctly classified:

$$(12) \quad Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \ .$$

Precision

Precision measures the proportion of predicted positive cases that are truly positive:

$$(13) \quad Precision = \frac{TP}{TP + FP} \ .$$

Recall

Recall, also known as the true positive rate, measures the proportion of actual positive cases correctly identified by the model

$$(14) \quad \text{Recall} = \frac{TP}{TP + FN} \ .$$

F1-Score

The F1-score is the harmonic mean of precision and recall:

$$(15) \quad F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \ .$$

5.2. Threshold-Independent Discrimination Metrics

Threshold-independent discrimination metrics evaluate the ranking ability of the model without requiring a fixed classification threshold. These metrics are useful because they assess whether the model assigns higher scores to positive observations than to negative observations (Fawcett, 2006).

ROC Curve and ROC-AUC

The Receiver Operating Characteristic curve plots the recall, Equation (14) against the false positive rate across all possible classification thresholds and the area under the ROC curve converts it into a single scalar that can be interpreted as the probability that the predicted probability assigned to a positive observation is higher than the predicted probability of a negative observation:

$$(16) \quad \mathrm{AUC}_{\mathrm{ROC}} = P\left(p_i^+ > p_j^-\right) = \int_0^1 TPR(FPR)\ dFPR \ ,$$

Where p_i^+ denotes the predicted probability assigned to a positive observation, and p_j^- denotes the predicted probability assigned to a negative observation.

A ROC-AUC of 1 indicates perfect discrimination, while a value of 0.5 indicates no discrimination beyond random ranking. Values below 0.5 indicate that the model ranks observations worse than random ordering. However, ROC-AUC evaluates discrimination, not calibration; a model may have high ROC-AUC while still producing poorly calibrated probabilities.

Precision-Recall Curve and Average Precision

Similar to the ROC and ROC-AUC but instead of plotting the TPR against the FPR, it plots the precision against the recall, and the AP summarizes it into a single scalar (Fawcett, 2006). This curve is particularly informative in imbalanced datasets because it focuses on performance for the positive class.

$$(17) \quad AP = \int_0^1 Precision(r)\ dr\ ,$$

where r denotes recall. Higher values of Average Precision indicate that the model maintains high precision over a wide range of recall levels. This metric is particularly useful in this project because lapse and midterm cancellation events are relatively infrequent.

6. Probability Calibration

Probability calibration evaluates the agreement between predicted probabilities and observed event frequencies. A model is said to be well calibrated if, among all observations assigned a predicted probability close to p, approximately a proportion p experiences the event (Powers, 2011). Formally, perfect calibration can be expressed as:

$$(18) \quad P(\,y_i = 1 \mid p_i = p\,) = p, \quad \forall p \in [0,1]\,,$$

where y_i, represents the observed binary outcome and p_i, represents the probability predicted by the model for individual i. In practical terms, if a group of policyholders is assigned a predicted lapse probability of 20%, then approximately 20% of those policyholders should lapse if the model is well calibrated.

Although many classification algorithms produce scores that can be interpreted as probabilities, these outputs are not necessarily well calibrated. Some models are optimized primarily for discrimination, that is, for separating positive from negative cases, rather than for ensuring that predicted probabilities match observed frequencies (Fawcett, 2006). This issue is particularly relevant in imbalanced datasets, where the positive class is relatively rare and model training procedures may produce overly confident or distorted probability estimates (Powers, 2011).

Among the models considered in this study, logistic regression is generally expected to provide relatively well-calibrated probabilities when the model is correctly specified and estimated without strong distortions. This is because logistic regression directly minimizes the negative log-likelihood, which is closely related to probabilistic accuracy. However, in finite samples, calibration may still be affected by class imbalance,

regularization, omitted variables, non-linear effects, or model misspecification (Guo et al., 2017).

Decision Trees, by contrast, often produce poorly calibrated probabilities. In a Decision Tree, the predicted probability for a given observation is based on the empirical proportion of positive cases within the terminal leaf (Chen et al., 2016).

Although boosting methods often achieve strong discrimination, their predicted probabilities may be mis calibrated, particularly when the data is imbalanced or when class-weighting and other imbalance-correction techniques are used. These techniques can improve classification performance for the minority class, but they may also shift the probability scale away from the true event frequency.

For the Multi-Layer Perceptron, the final layer also applies a sigmoid function to produce a probability estimate. In theory, a neural network trained with binary cross-entropy can produce calibrated probabilities. In practice, however, calibration can be affected by finite sample size, overparameterization, class imbalance, regularization, and imperfect optimization. Neural networks may also become overconfident, especially when they fit the training data too closely (Guo et al., 2017). For these reasons, calibration is also relevant for the MLP model.

6.1. Isotonic Regression

To address miscalibration, isotonic regression was applied to the probability outputs of the models. Isotonic regression is a non-parametric calibration method that teaches monotonic mapping from raw predicted probabilities to calibrated probabilities (Ayer et al, 1955). The purpose of this mapping is to adjust predicted probabilities so that they better match observed event frequencies, while preserving the ranking order of the original model scores.

The calibration mapping can be represented as:

$$(19) \quad f: p_i^{raw} \longrightarrow p_i^{cal} \ ,$$

where p_i^{raw} represents the uncalibrated probability, and p_i^{cal} represents the calibrated probability. The monotonicity constraint imposed by isotonic regression is:

$$(20) \quad p_i^{\text{raw}} \leq p_j^{\text{raw}} \Rightarrow f(p_i^{\text{raw}}) \leq f\left(p_j^{\text{raw}}\right),$$

This condition ensures that the calibration function does not reverse the ordering produced by the original model. The isotonic regression problem can be written as:

$$(21) \quad \min_f \sum_{i=1}^{N} w_i (f(p_i^{raw}) - y_i)^2, \quad s.t. \quad f(p_1^{raw}) \leq f(p_2^{raw}) \leq \cdots \leq f(p_N^{raw}),$$

where w_i, represents the weight assigned to observation i, $f(p_i^{raw})$, represents the calibrated probability, and y_i, represents the observed binary outcome. In practice, the observations are ordered by their raw predicted probabilities before applying the monotonicity constraint.

The solution to this optimization problem can be obtained using the Pool Adjacent Violators algorithm (Ayer et al., 1955). The resulting calibration function is a monotonic step function that maps groups of similarly scored observations to their observed event frequencies.

In this study, isotonic regression was implemented using a separate calibration subset. After hyperparameter optimization, each base model was retrained using 80% of the training data, while the remaining 20% was used to estimate the isotonic calibration mapping. This separation is important because calibrating and evaluating on the same observations would introduce optimistic bias.

The calibrated model was then evaluated on the test set and compared with the corresponding uncalibrated model. Although the calibrated model is trained on a smaller portion of the training data, this comparison is operationally meaningful: if a calibrated model trained with fewer observations achieves better probability quality on unseen data, then calibration provides practical value despite the reduced training sample available for the base learner.

6.2. Calibration Metrics

To assess the quality of predicted probabilities, this study uses the Brier score, the Brier score decomposition, Expected Calibration Error, and Maximum Calibration Error.

These metrics complement discrimination measures such as ROC-AUC and Average Precision by focusing specifically on the reliability of predicted probabilities.

Brier Score

The Brier score measures the mean squared difference between predicted probabilities and observed binary outcomes (Brier, 1950):

$$(22) \quad BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2,$$

where p_i, represents the predicted probability for individual i, y_i, represents the observed binary outcome, and N, represents the number of observations. Lower Brier scores indicate better probabilistic predictions.

The Brier score captures both discrimination and calibration. Therefore, it is useful as an overall probabilistic performance metric, but it does not by itself indicate whether poor performance is due to weak ranking ability or poor calibration (Brier, 1950).

Brier Score Decomposition

To obtain a more detailed view of probabilistic performance, the Brier score can be decomposed into reliability, resolution, and uncertainty (Murphy, 1973):

$$(23) \quad BS = Reliability - Resolution + Uncertainty\,,$$

The reliability component measures calibration error. It is defined as:

$$(24) \quad \text{Reliability} = \frac{1}{N}\sum_{k=1}^{K} n_k \left(p_k - \overline{y}_k\right)^2,$$

where K, represents the number of bins, n_k, represents the number of observations in bin k, p_k, represents the average predicted probability in bin p_k, and $\overline{y}_k$, represents the observed event frequency in bin k. Lower reliability values indicate better calibration.

The resolution component measures how much the observed event frequencies vary across probability bins:

$$(25) \text{ Resolution} = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\overline{y}_k - \overline{y}\right)^2,$$

where $\overline{y}_k$, represents the observed event frequency in bin k, and y represents the overall event frequency in the sample. Higher resolution is desirable because it indicates that the model separates observations into groups with different observed risks.

The uncertainty component is defined as:

$$(26) \quad \text{Uncertainty} = \bar{y}(1 - \bar{y}),$$

Where y represents the overall event frequency. This term reflects the inherent uncertainty of the binary outcome and corresponds to the variance of a Bernoulli variable with mean y.

Expected Calibration Error

Expected Calibration Error measures the weighted average absolute difference between predicted probabilities and observed event frequencies across bins (Guo et al, 2017):

$$(27) \quad \text{ECE} = \sum_{k=1}^{K} \frac{n_k}{N} \left| p_k - \overline{y}_k \right| ,$$

where n_k represents the number of observations in bin k, p_k represents the average predicted probability in bin k, and $\overline{y}_k$ represents the observed event frequency in bin k. Lower ECE values indicate better calibration.

Unlike the reliability component of the Brier decomposition, ECE uses absolute deviations rather than squared deviations. As a result, ECE is easier to interpret in percentage-point terms, while reliability penalizes larger calibration errors.

Maximum Calibration Error

Maximum Calibration Error measures the largest absolute calibration gap across all bins (Guo et al, 2017):

$$(28) \quad \text{MCE} = \max_{k \in \{1,\ldots,K\}} \left| p_k - \overline{y}_k \right| ,$$

where p_k represents the average predicted probability in bin k, and $\overline{y}_k$ represents the observed event frequency in bin k. While ECE summarizes average calibration error, MCE identifies the worst-calibrated region of the probability distribution. This is relevant in insurance applications because large local calibration errors may lead to inappropriate

pricing or retention decisions for specific groups of policyholders, note the MCE value is bin dependent so its value will vary largely with the number of observations in each bin.

Probability calibration is central to this study because the model outputs are intended to support business decisions that depend on the reliability of predicted probabilities. While discrimination metrics assess whether the model ranks risky policyholders above less risky ones, calibration metrics assess whether the predicted probabilities correspond to observed event frequencies.

For this reason, the final model assessment considers not only classification and discrimination metrics, but also the Brier score, its decomposition, Expected Calibration Error, and Maximum Calibration Error. Isotonic regression is used as a post-hoc calibration method to improve the reliability of probability estimates while preserving the ranking structure of the original model outputs.

7. SHAP-Based Model Interpretation

Because the predicted probabilities produced in this study are intended to support business decisions in pricing optimization, portfolio management, and retention strategy, model transparency is an important component of the analysis. Since the final recommended models are based on LightGBM, their predictions are not as directly interpretable as those of simpler linear models. To improve transparency, this study uses SHAP values to quantify the contribution of each feature to the predicted probability of lapse or midterm cancellation.

SHAP, or SHapley Additive exPlanations, is an interpretability method based on Shapley values from cooperative game theory (Shapley, 1953; Lundberg & Lee, 2017). The main idea is to decompose a model prediction into the contribution of each input feature relative to a baseline value, typically the average model prediction. This allows the model output to be interpreted both locally, by explaining individual predictions, and globally, by summarizing feature contributions across the full portfolio.

Let F denote the set of all features used by the model and let S be a subset of F that does not include feature i. The SHAP value of feature i for individual j is given by:

$$(29) \quad \phi_i^{(j)} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(\mathbf{X}^{(j)}\right) - f_S\left(\mathbf{X}^{(j)}\right) \right],$$

where $\phi_i^{(j)}$ represents the SHAP value of feature i for individual j, F represents the set of all model features, S represents a subset of features excluding feature i, and $\mathbf{X}^{(j)}$ represents the feature vector of individual j. The term $f_{S \cup \{i\}}\left(\mathbf{X}^{(j)}\right) - f_S\left(\mathbf{X}^{(j)}\right)$ measures the marginal contribution of adding feature i to coalition S, while the factorial term weights this contribution across all possible feature orderings. For each feature i, SHAP produces one value for each individual in the dataset.

Positive SHAP values indicate that a feature increases the model output relative to the baseline, while negative SHAP values indicate that a feature decreases the model output.

To obtain a global measure of feature importance, the mean absolute SHAP value can be computed for each feature:

$$(30) \quad I_i = \frac{1}{N} \sum_{j=1}^{N} \left| \phi_i^{(j)} \right|,$$

where I_i represents the global importance of feature i. And N represents the number of observations. This measure summarizes the average magnitude of each feature's contribution across all observations. Features with higher mean absolute SHAP values have a stronger influence on the model predictions.

However, because absolute values are used, this global importance measure does not preserve the direction of the effect. For this reason, SHAP beeswarm plots are also used. These plots show both the magnitude and direction of feature contributions. Each point represents the SHAP value of one observation for a given feature. Features are ordered according to their overall importance, while the horizontal position indicates whether the feature increases or decreases the model output. The color scale represents the value of the feature, allowing the relationship between feature values and predicted probabilities to be visually assessed.

In this study, SHAP values are used for two main purposes. First, they support model interpretation by identifying the variables that contribute most to the predicted

probabilities of lapse and midterm cancellation. Second, they support the feature selection process by helping identify variables with limited contribution to the model output. This is useful because a smaller set of informative variables can make the model easier to understand, maintain, and potentially implement in a business environment.

In addition, for feature interpretation, SHAP beeswarm plots are also used as a qualitative diagnostic tool. For both the lapse and midterm cancellation models, SHAP values were computed separately for the training and test samples. Comparing SHAP patterns across these samples helps assess whether the model relies on similar variables and their SHAP distributions remain broadly consistent between the training and test sets.

This provides evidence that the model has captured relationships that generalize beyond the training data, rather than patterns specific to the training sample. The resulting beeswarm plots for the best LightGBM models below demonstrate the impact of the features and the consistency between the train and test sets.

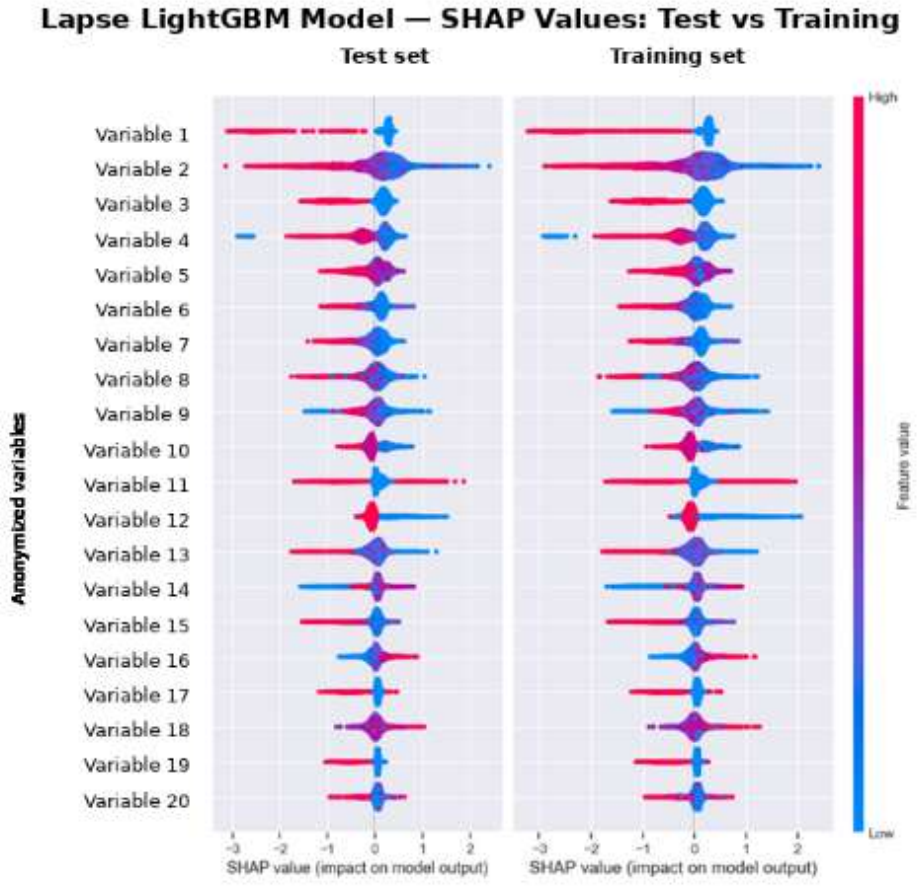

Figure 2 – Beeswarm plot LightGBM model for lapse

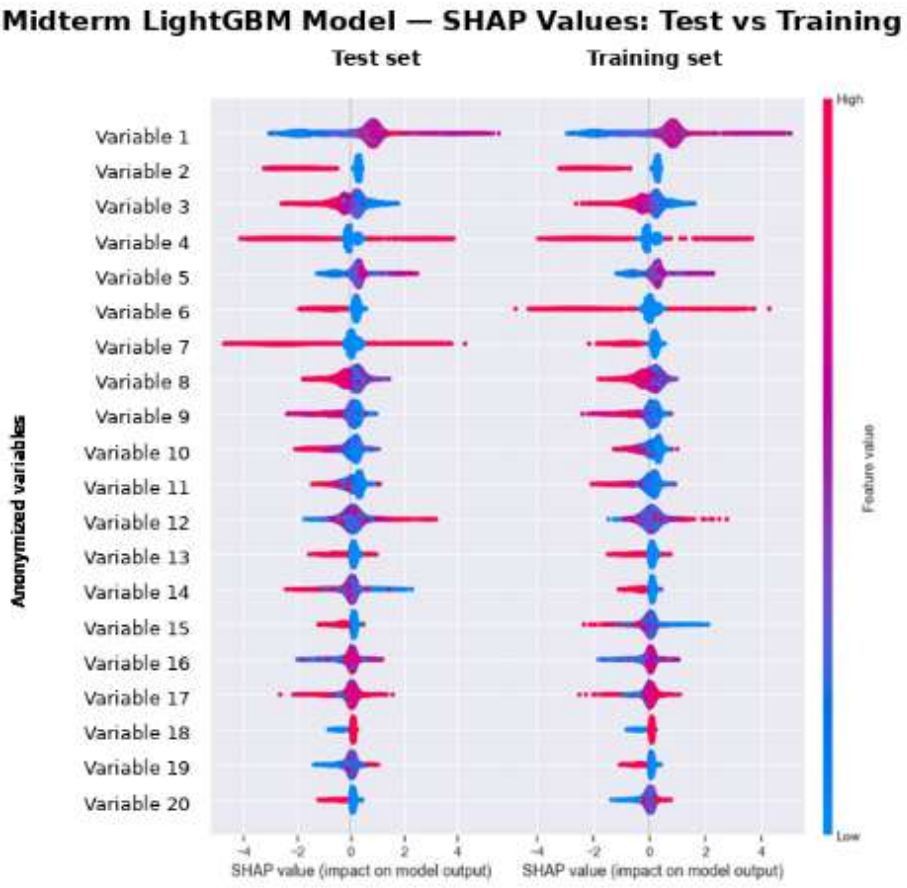

Figure 3 - Beeswarm plot LightGBM model for midterm

8. Modeling and Evaluation

The modeling and evaluation phase was designed to compare alternative classification algorithms under a common and reproducible framework. Since the objective of this study is to estimate reliable probabilities of lapse and midterm cancellation, rather than merely assign binary class labels, the modeling pipeline was built around three principles: avoiding data leakage, optimizing probabilistic performance, and evaluating both discrimination and calibration.

The dataset was split into training and test sets using an 80/20 stratified split. Stratification was applied to preserve the proportion of positive and negative cases in both subsets, which is particularly important given the class imbalance present in both prediction tasks. The training set was used for preprocessing, hyperparameter tuning,

model fitting, and calibration. The test set was kept fully separate and used only for final out-of-sample evaluation.

Categorical variables were encoded using one-hot encoding, while numerical variables were standardized using a standard scaler. Standardization was required for Logistic Regression and the Multi-Layer Perceptron, as both models are sensitive to the scale of the input variables. Tree-based models, including Decision Trees, Random Forest, XGBoost, and LightGBM, are generally insensitive to feature scaling because their splitting rules depend on feature thresholds rather than distances between observations. To prevent data leakage, both the encoder and the scaler were fitted exclusively on the training data and then applied to the test data.

8.1. Hyperparameter Tuning

Hyperparameters are model configuration choices that are not estimated directly from the data but can materially affect model performance. Examples include the maximum depth of a tree, the learning rate of a boosting algorithm, the number of estimators, or the regularization strength. Since different hyperparameter configurations can lead to substantially different results, hyperparameter tuning is a standard component of modern machine learning workflows (Feurer & Hutter, 2019).

The hyperparameter search can be formalized as an optimization problem. Given a machine learning model parameterized by a hyperparameter configuration $\lambda \in \Lambda$where Λ denotes the configuration space, the objective is to find the configuration that minimizes a chosen validation loss (Akiba et al, 2019):

$$(31) \quad \lambda^* = \arg\min_{\lambda \in \Lambda} \mathcal{L}(\lambda) \ ,$$

where λ^* represents the optimal hyperparameter configuration, λ represents a candidate configuration, Λ represents the configuration space, and $\mathcal{L}(\lambda)$, represents the validation loss associated with that configuration.

In this study, the validation loss was defined as the cross-validated Brier score. This choice is consistent with the main objective of the project, since the Brier score evaluates the quality of predicted probabilities rather than only the correctness of class labels.

Therefore, the hyperparameter search was explicitly aligned with the probabilistic nature of the business problem.

Hyperparameter optimization was performed using Optuna (Akiba et al., 2019), which implements efficient automated search strategies. In particular, the Tree-structured Parzen Estimator was used as the sampling algorithm. This Bayesian optimization approach uses information from previous trials to guide the search toward more promising regions of the hyperparameter space (Bergstra et al., 2011). In simplified terms, the method compares the kernel density estimation of configurations associated with good performance: $l(\lambda)$ with the kernel density estimation of configurations associated with weaker performance: $g(\lambda)$, and prioritizes configurations with high expected improvement:

$$(32) \quad EI(\lambda) \propto \frac{l(\lambda)}{g(\lambda)} ,$$

where $EI(\lambda)$, represents the expected improvement associated with configuration λ, $l(\lambda)$ represents the estimated density of good-performing configurations, and $g(\lambda)$ represents the estimated density of poorer-performing configurations.

For each model, a stratified five-fold cross-validation procedure was used during hyperparameter tuning. In each trial, the model was trained on four folds and evaluated on the remaining fold. This procedure was repeated until each fold had been used once as validation data, and the final objective value was computed as the average Brier score across the five folds.

The optimization process was limited to 20 trials per model. This restriction was imposed for computational reasons and to reduce the risk of overfitting the hyperparameter search to the validation procedure. As noted by (Cawley and Talbot, 2010), extensive model selection can itself become a source of overfitting when many configurations are evaluated. The use of an independent test set for final evaluation mitigates this risk.

8.2. Calibration Procedure

After hyperparameter tuning, each model was trained using the selected configuration and evaluated on the test set. To assess the effect of probability calibration, a calibrated version of each model was also estimated using isotonic regression.

The calibration procedure was implemented using only the training data. After hyperparameter optimization, a clone of each model was trained on 80% of the training set. The remaining 20% was used as a calibration subset to estimate the isotonic mapping from raw model probabilities to calibrated probabilities. This separation is important because fitting the calibration function on the same observations used to train the base model would lead to overly optimistic calibration estimates.

Both the uncalibrated model and the calibrated model were then evaluated on the same independent test set. It is important to note that the comparison is not perfectly symmetric, since the calibrated model is trained on a smaller effective training sample. Nevertheless, this comparison is operationally meaningful. If a calibrated model trained on less data produces more reliable probabilities on unseen observations, then calibration adds value despite the reduction in training data available to the base learner.

8.3. Initial Model Comparison

The initial comparison across all models shows a clear performance hierarchy. For both lapse and midterm cancellation, LightGBM provides the strongest overall results among the algorithms tested, as can be seen in the following pages.

Table I - Lapse Model Results for LR, DT and RF

Metric	LR	LR Calib	DT	DT Calib	RF	RF Calib
Accuracy	94.85%	94.87%	95.04%	94.93%	94.93%	94.94%
Precision	18.75%	36.00%	62.26%	50.00%	0.00%	51.22%
Recall	0.49%	1.46%	5.37%	1.95%	0.00%	3.41%
F1-Score	0.95%	2.81%	9.88%	3.76%	0.00%	6.40%
ROC-AUC	73.22%	72.93%	73.34%	71.85%	76.53%	75.41%
PR-AUC	15.40%	14.02%	20.17%	16.37%	18.08%	15.91%
Brier Score	0.0459	0.0460	0.0444	0.0450	0.0457	0.0454
Reliability	0.0004	0.0004	0.0004	0.0002	0.0005	0.0003
Resolution	0.0023	0.0023	0.0039	0.0031	0.0019	0.0023
Uncertainty	0.0481	0.0481	0.0481	0.0481	0.0481	0.0481
ECE	0.0031	0.0052	0.0086	0.0040	0.0086	0.0066
MCE	0.9187	0.6364	0.1630	0.2222	0.6929	0.5095

Table II - Lapse Model Results for XGBoost, LightGBM and MLP

Metric	XGBoost	XGBoost Calib	LightGBM	LightGBM Calib	MLP	MLP Calib
Accuracy	95.90%	95.64%	96.72%	96.46%	95.92%	95.94%
Precision	76.71%	75.90%	90.04%	82.46%	68.63%	69.06%
Recall	27.32%	20.49%	39.67%	38.21%	35.93%	35.93%
F1-Score	40.29%	32.27%	55.08%	52.22%	47.17%	47.27%
ROC-AUC	87.89%	85.83%	90.39%	88.20%	86.97%	84.39%
PR-AUC	51.26%	43.22%	68.22%	57.88%	51.53%	44.12%
Brier Score	0.0334	0.0353	0.0269	0.0292	0.0362	0.0349
Reliability	0.0006	0.0003	0.0018	0.0004	0.0022	0.0003
Resolution	0.0149	0.0125	0.0225	0.0190	0.0137	0.0132
Uncertainty	0.0481	0.0481	0.0481	0.0481	0.0481	0.0481
ECE	0.0151	0.0057	0.0205	0.0041	0.0322	0.0058
MCE	0.1410	0.6337	0.2855	0.1711	0.4532	0.4866

Table III - Midterm Model Results for LR, DT and RF

Metric	LR	LR Calib	DT	DT Calib	RF	RF Calib
Accuracy	90.54%	90.54%	91.49%	91.65%	90.35%	90.89%
Precision	65.62%	56.67%	71.83%	77.31%	0.00%	72.52%
Recall	3.99%	8.07%	19.37%	19.09%	0.00%	9.02%
F1-Score	7.52%	14.13%	30.52%	30.62%	0.00%	16.05%
ROC-AUC	75.62%	75.49%	77.35%	76.90%	80.30%	79.53%
PR-AUC	30.22%	28.25%	39.46%	37.03%	39.31%	34.63%
Brier Score	0.0777	0.0777	0.0710	0.0715	0.0785	0.0739
Reliability	0.0003	0.0003	0.0003	0.0001	0.0053	0.0002
Resolution	0.0091	0.0091	0.0157	0.0152	0.0127	0.0129
Uncertainty	0.0872	0.0872	0.0872	0.0872	0.0872	0.0872
ECE	0.0098	0.0121	0.0119	0.0061	0.0435	0.0058
MCE	0.2743	0.1405	0.1035	0.0916	0.6192	0.4000

Table IV- Midterm Model Results for XGBoost, LightGBM and MLP

Metric	XGBoost	XGBoost Calib	LightGBM	LightGBM Calib	MLP	MLP Calib
Accuracy	94.38%	94.60%	95.37%	95.23%	93.84%	93.87%
Precision	74.31%	88.14%	82.62%	89.35%	71.43%	86.78%
Recall	63.72%	50.81%	65.91%	57.36%	60.30%	43.02%
F1-Score	68.61%	64.46%	73.32%	69.87%	65.40%	57.52%
ROC-AUC	90.83%	89.37%	91.17%	90.36%	89.45%	87.33%
PR-AUC	74.80%	69.92%	77.97%	73.53%	69.10%	62.94%
Brier Score	0.0452	0.0453	0.0390	0.0405	0.0537	0.0513
Reliability	0.0036	0.0002	0.0017	0.0001	0.0062	0.0001
Resolution	0.0452	0.0417	0.0495	0.0464	0.0388	0.0354
Uncertainty	0.0872	0.0872	0.0872	0.0872	0.0872	0.0872
ECE	0.0280	0.0043	0.0232	0.0032	0.0502	0.0033
MCE	0.2862	0.2148	0.2841	0.0659	0.4926	0.1718

For the lapse task, the uncalibrated LightGBM model achieves the best overall performance, with a ROC-AUC of 90.39%, a PR-AUC of 68.22%, and the lowest Brier score, 0.0269. These results indicate that LightGBM is not only effective at ranking policyholders according to lapse risk but also produces comparatively accurate probabilities estimates before calibration.

For the midterm cancellation task, the same pattern is observed. The uncalibrated LightGBM model achieves a ROC-AUC of 91.17%, a PR-AUC of 77.97%, and the lowest Brier score, 0.0390. The higher PR-AUC and stronger recall observed in the midterm task suggest that this event is more easily distinguishable from non-cancellation cases than lapses, at least given the available features.

Across both prediction tasks, calibration substantially improves calibration-specific metrics, especially reliability, Expected Calibration Error, and Maximum Calibration Error. However, calibration does not uniformly improve every metric. In particular, isotonic calibration generally reduces ROC-AUC and PR-AUC and slightly worsens the Brier score for the full LightGBM models. This result is not contradictory. Calibration primarily adjusts probability reliability; it does not aim to improve ranking performance and may reduce resolution by compressing extreme probabilities toward observed frequencies.

This trade-off is central to the interpretation of the results. The uncalibrated LightGBM models deliver the strongest overall Brier scores and discrimination performance. The calibrated LightGBM models, by contrast, provide more reliable probabilities, with substantially lower calibration errors. Since the business objective of this study requires probabilities that can be used in pricing and portfolio decisions, calibration remains valuable even when it slightly reduces discrimination.

To quantify the improvement over a naive probabilistic baseline, the Brier Skill Score was computed, as it provides a measure of probabilistic accuracy by comparing predicted probabilities with the observed binary outcomes. It penalizes predictions that assign high probabilities to events that do not occur, as well as predictions that assign low probabilities to events that do occur. As a result, it captures the overall quality of the probability estimates rather than only the correctness of class assignments at a fixed threshold.

The Brier Skill Score is defined as (Brie, 1950):

$$\text{BSS} = 1 - \frac{\text{BS}_{\text{model}}}{\text{UNC}} \quad (33)$$

where, BS_{model}, represents the Brier score of the evaluated model, and UNC, represents the uncertainty component of the Brier decomposition.

In this context, the Brier Skill Score measures the proportional improvement of the model over a naive baseline that always predicts the average event frequency. A Brier Skill Score of 0.44 for the lapse model means that the model reduces the average squared probability error by approximately 44% relative to this base-rate benchmark. Similarly, the value of 0.55 for the midterm cancellation model indicates a 55% reduction in probabilistic error. This suggests that both models produce informative probability estimates, while the midterm cancellation model achieves stronger relative improvement over the naive forecast.

8.4. SHAP-Based Feature Selection

After identifying LightGBM as the strongest model family, a SHAP-based feature selection procedure was applied to assess whether a more parsimonious model could retain similar performance. This step was motivated by the high dimensionality of the modeling dataset, which contained approximately 300 candidate features. Although LightGBM can handle large feature sets, removing variables with limited contribution may improve interpretability, reduce noise, and simplify future implementation.

Feature importance was measured using the mean absolute SHAP value of each variable. Features were ranked by their average absolute contribution to the model output, and the cumulative share of total SHAP importance was computed.

Figure 4: Cumulative SHAP values midterm

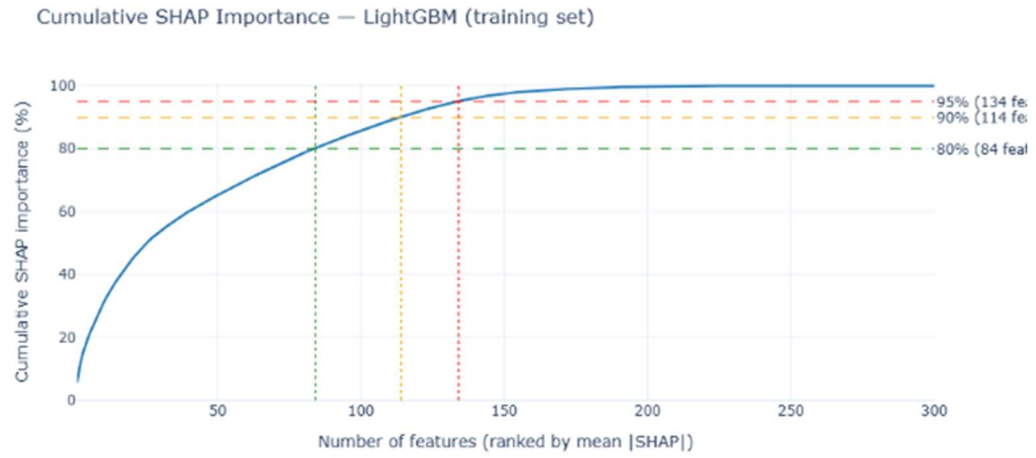

Figure 5: Cumulative SHAP values lapse

The cumulative SHAP curve shows that predictive signal is concentrated in a subset of variables: approximately 86 features explain 80% of total SHAP importance, 113 features explain 90%, and 136 features explain 95%.

Based on this analysis, three reduced feature sets were evaluated. The first retained features accounting for 98.2% of cumulative SHAP importance, representing a

conservative reduction. The second and third retained features accounting for 90% and 80% of cumulative SHAP importance, respectively. These thresholds were chosen to evaluate the effect of progressively stronger dimensionality reduction on model performance and calibration.

SHAP-based feature selection should not be interpreted as guaranteed performance improvement. Some variables with low individual importance may still contribute useful information through interactions with other variables. For this reason, feature selection was evaluated empirically by retraining and recalibrating the LightGBM model under each reduced feature set.

8.5. Lapse Model After Feature Selection

Table V - Truncated and full LightGBM model results for lapse

Metric	100%	100% Calib	98.2%	98.2% Calib	90%	90% Calib	80%	80% Calib
Accuracy	96.72%	96.46%	93.19%	95.90%	89.49%	95.36%	89.42%	95.22%
Precision	90.04%	82.46%	39.92%	67.77%	28.30%	62.38%	28.32%	55.38%
Recall	39.67%	38.21%	67.97%	36.59%	70.08%	21.30%	71.06%	29.27%
F1-Score	55.08%	52.22%	50.30%	47.52%	40.32%	31.76%	40.50%	38.30%
ROC-AUC	90.39%	88.20%	89.64%	87.02%	88.59%	86.40%	88.88%	86.73%
PR-AUC	68.22%	57.88%	52.84%	45.58%	43.94%	38.19%	43.96%	38.88%
Brier Score	0.0269	0.0292	0.0566	0.0340	0.0826	0.0374	0.0834	0.0373
REL	0.0018	0.0004	0.0211	0.0002	0.0463	0.0005	0.0475	0.0004
RES	0.0225	0.0190	0.0160	0.0140	0.0122	0.0110	0.0125	0.0109
UNC	0.0481	0.0481	0.0481	0.0481	0.0481	0.0481	0.0481	0.0481
ECE	0.0205	0.0041	0.1045	0.0035	0.1615	0.0058	0.1647	0.0059
MCE	0.2855	0.1711	0.3831	0.1353	0.4784	0.1857	0.4862	0.2411

For the lapse task, the full LightGBM model remains the strongest specification overall. The uncalibrated full model achieves a ROC-AUC of 90.39%, a PR-AUC of 68.22%, and the lowest Brier score, 0.0269. These results indicate strong discrimination and strong overall probabilistic performance.

However, the uncalibrated full model is not perfectly calibrated. Its ECE is 0.0205 and its reliability component is 0.0018. After isotonic calibration, ECE decreases to 0.0041 and reliability falls to 0.0004, showing a substantial improvement in probability reliability. The cost of this improvement is a modest deterioration in the Brier score, from

0.0269 to 0.0292, and a reduction in resolution from 0.0225 to 0.0190. This suggests that calibration improves the agreement between predicted and observed probabilities but slightly reduces the model's ability to separate policyholders into highly distinct risk groups.

The SHAP-reduced models produce different trade-off. Before calibration, the reduced models achieve considerably higher recall than the full LightGBM model. For example, recall increases from 39.67% in the full model to 67.97%, 70.08%, and 71.06% under 98.2%, 90%, and 80% SHAP thresholds, respectively. However, this improvement in recall is accompanied by a severe decline in precision and calibration quality. ECE rises to 0.1045 for the 98.2% SHAP model, 0.1615 for the 90% SHAP model, and 0.1647 for the 80% SHAP model.

Isotonic calibration corrects much of this miscalibration in the reduced models. Nevertheless, the reduced calibrated models continue to underperform the full calibrated LightGBM in terms of Brier score. Among the reduced specifications, 98.2% SHAP calibrated model is the most defensible alternative, with a Brier score of 0.0340, ECE of 0.0145 and MCE of 0.1353. However, this remains weaker than the full calibrated model, which achieves a Brier score of 0.0292 and ECE of 0.0041.

Therefore, for lapse prediction, the full calibrated LightGBM model is the preferred specification. It offers the best balance between discrimination, probability reliability, and overall probabilistic accuracy. The 98.2% SHAP calibrated model may be considered only if operational simplicity or interpretability constraints justify the loss in performance.

8.6. Midterm Cancellation Model After Feature Selection

Table VI- Truncated and full LightGBM model results for midterm

Metric	100%	100% Calib	98.2%	98.2% Calib	90%	90% Calib	80%	80% Calib
Accuracy	95.37%	95.23%	90.50%	94.15%	87.29%	93.35%	86.95%	93.40%
Precision	82.62%	89.35%	50.51%	86.70%	41.38%	77.66%	40.67%	76.47%
Recall	65.91%	57.36%	74.93%	46.44%	76.16%	43.59%	77.02%	45.68%
F1-Score	73.32%	69.87%	60.34%	60.48%	53.63%	55.84%	53.23%	57.19%
ROC-AUC	91.17%	90.36%	91.14%	90.00%	90.15%	89.00%	90.19%	89.38%
PR-AUC	77.97%	73.53%	72.08%	68.22%	67.05%	62.71%	67.33%	63.67%
Brier Score	0.0390	0.0405	0.0774	0.0473	0.0971	0.0531	0.0998	0.0523
REL	0.0017	0.0001	0.0320	0.0001	0.0465	0.0002	0.0489	0.0001
RES	0.0495	0.0464	0.0420	0.0397	0.0368	0.0337	0.0363	0.0347
UNC	0.0872	0.0872	0.0872	0.0872	0.0872	0.0872	0.0872	0.0872
ECE	0.0232	0.0032	0.1313	0.0038	0.1691	0.0057	0.1762	0.0055
MCE	0.2841	0.0659	0.3746	0.0440	0.4181	0.0921	0.3984	0.0820

For midterm cancellation, the results are even stronger. The full uncalibrated LightGBM model achieves a ROC-AUC of 91.7%, a PR-AUC of 77.97%, and a Brier score of 0.0390. These values indicate that the model is highly effective at distinguishing policyholders likely to cancel midterm from those who remain active.

As in lapse task, calibration improves probability reliability. For the full LightGBM model, ECE decreases from 0.0232 to 0.0032, reliability decreases from 0.0017 to 0.0001, and MCE decreases from 0.2841 to 0.0659. These changes indicate a substantial improvement in calibration. The Brier score increases slightly, from 0.0390 to 0.0405, again reflecting the trade-off between improved reliability and reduced resolution. The SHAP-reduced models again increase recall before calibration. Recall rises

from 65.91% in the full model to 74.93%, 76.16%, and 77.02% under the 98.2%, 90%, and 80% SHAP thresholds, respectively. However, this comes with large reductions in precision and severe deterioration in calibration. The ECE reaches 0.1313 for the 98.2% SHAP model, 0.1691 for the 90% SHAP model, and 0.1762 for the 80% SHAP model.

After isotonic calibration, the reduced models become much better calibrated. The 98.2% SHAP calibrated model achieves an ECE of 0.0038 and an MCE of 0.0440, both of which are strong calibration results. However, its Brier score is 0.0473, which remains

materially worse than the full calibrated LightGBM Brier score of 0.0405. The 90% and 80% SHAP calibrated models perform worse still in terms of overall probabilistic accuracy.

Consequently, the full calibrated LightGBM model is also the recommended specification for midterm cancellation. It provides the best overall balance between discrimination, calibration, and Brier score performance. If a reduced model is necessary, the 98.2% SHAP calibrated model is the preferred alternative, but the loss in probabilistic accuracy should be explicitly acknowledged.

8.7. Final Model Recommendation

The results support a clear conclusion. Across both lapse and midterm cancellation, LightGBM is the best-performing algorithm among those tested. It consistently outperforms Logistic Regression, Decision Trees, Random Forest, XGBoost, and the Multi-Layer Perceptron in terms of discrimination and Brier score performance.

However, the final recommendation depends on the intended use of the model outputs. If the objective were only to rank policyholders by risk, the uncalibrated LightGBM would be the strongest candidate. However, the purpose of this project is to generate probability estimates that can support pricing optimization and portfolio management.

The recommended specification is therefore the calibrated LightGBM model for both lapse and midterm cancellation. Although isotonic calibration slightly worsens the Brier score and reduces resolution, it substantially improves reliability, ECE, and MCE. This trade-off is acceptable in the present context because poorly calibrated probabilities could lead to inappropriate pricing or retention decisions.

SHAP-based feature selection did not improve overall model performance. While reduced models sometimes increased recall, this came at the cost of lower precision, poorer calibration, and higher Brier scores. The full feature set should therefore be retained unless operational constraints require a simpler model. In that case, the 98.2% SHAP calibrated model offers the most reasonable reduced alternative. This was a

expected result since tree based models perform internal feature selection by choosing the feature that reduces node impurity the most at each split.

Overall, the evidence indicates that gradient boosting methods, and LightGBM in particular, provide a strong and practical alternative to logistic regression for modeling lapse and midterm cancellation probabilities in health insurance.

9. Conclusion

This study set out to evaluate whether modern machine learning techniques can produce reliable probability estimates for lapse and midterm cancellation in individual health insurance, and to identify which models are most suitable for operational use by an insurer.

The results provide clear evidence that modern gradient boosting methods, particularly LightGBM, are effective for both prediction tasks. For lapse, the best LightGBM model achieved a Brier Skill Score of approximately 0.44, indicating a 44% improvement over a naive baseline that assigns the same probability to every policyholder. For midterm cancellation, the corresponding Brier Skill Score was approximately 0.55, showing an even stronger improvement. These results confirm that meaningful predictive signal exists in the available data and that machine learning models can extract this signal in a way that is operationally useful.

The midterm cancellation model consistently outperformed the lapse model across most evaluation metrics. This difference is likely related to the higher event rate in the midterm dataset, which provides the model with more positive examples during training. It may also suggest that midterm cancellation behavior is more strongly associated with observable policyholder and policy characteristics than lapse behavior at renewal.

Regarding model comparison, the evidence strongly favors LightGBM over the alternative algorithms considered. Logistic Regression, Decision Trees, Random Forest, XGBoost, and the Multi-Layer Perceptron all produced useful results in some dimensions, but none matched LightGBM's overall balance between discrimination, Brier score performance, and calibration behavior. Logistic Regression, while still a

valuable benchmark due to its interpretability and widespread use in insurance, performed poorly in threshold-dependent metrics, particularly recall and F1-score. This suggests that, in this specific context, the traditional logistic regression benchmark is not sufficient to capture the non-linear relationships, and interaction effects present in the data.

Because the objective of this work was not merely to classify policyholders, but to estimate reliable probabilities, performance assessment could not rely only on accuracy, precision, recall, or ROC-AUC. The Brier score and its decomposition were central to the evaluation, as they allowed probability quality to be assessed through reliability, resolution, and uncertainty. ROC-AUC and PR-AUC remained useful complementary measures of discrimination, but calibration metrics were essential for determining whether the predicted probabilities could be trusted in a business setting.

The results also confirm the importance of probability calibration. Raw LightGBM probabilities showed signs of miscalibration, especially in terms of Expected Calibration Error and reliability. Isotonic regression substantially improved calibration in both tasks: for lapse, ECE decreased from 0.0205 to 0.0041, while for midterm cancellation it decreased from 0.0232 to 0.0032. This improvement came with a trade-off, as calibration slightly increased the Brier score and reduced resolution. In practical terms, calibration made the probabilities more reliable but slightly less dispersed across risk groups. Given the intended use of the probabilities in pricing optimization, portfolio management, and retention decisions, this trade-off is acceptable and supports the use of calibrated LightGBM models for deployment.

The SHAP-based feature selection analysis showed that a relatively limited subset of variables accounts for most of the model's predictive signal. The cumulative SHAP importance curves indicated that approximately 86 features explain 80% of total SHAP importance, 113 features explain 90%, and 136 features explain 95%. This confirms that predictive importance is concentrated in a subset of variables rather than evenly distributed across the full feature space. However, the empirical results also show that reducing the feature set did not improve overall model performance. Although SHAP-reduced models sometimes increased recall, this generally came at the cost of lower precision, weaker Brier score performance, and poorer calibration before isotonic adjustment. Among the reduced models, the 98.2% SHAP calibrated specification was

the most defensible alternative, but the full calibrated LightGBM model remained the preferred choice.

The SHAP analysis also provided insight into the drivers of lapse and midterm cancellation without requiring disclosure of proprietary variable names. The results suggest that cancellation risk is not driven by a single dominant variable, but rather by a combination of interacting features. This finding is important because it helps explain why gradient boosting methods outperformed logistic regression. The relationships between the predictors and cancellation probabilities appear to be non-linear and interaction-dependent, making them difficult to capture with simpler linear models or traditional rule-based segmentation approaches.

From a practical perspective, the calibrated LightGBM models can be used to assign each policyholder a probability of lapse or midterm cancellation. These probabilities can support several operational decisions. For portfolio management, they allow the insurer to identify segments with elevated cancellation risk. For pricing optimization, they can be incorporated into expected value calculations at renewal. For retention strategy, they allow interventions to be prioritized toward policyholders with higher predicted cancellation probabilities. Importantly, because the outputs are probabilities rather than hard classifications, the insurer can adapt decision thresholds according to commercial objectives, budget constraints, and the expected value of intervention.

Several limitations should be acknowledged. First, the analysis was based on data from a single insurer, so the generalizability of the results to other portfolios, markets, or time periods remains untested. Second, model performance is constrained by the quality, completeness, and historical availability of the input data. Richer behavioral information, such as customer service interactions, complaints, digital engagement, or broker activity, could potentially improve model resolution and predictive performance. Third, the calibration procedure required reserving part of the training data for isotonic regression, reducing the sample available for fitting the base model. Future work could explore cross-validated calibration methods to reduce this loss. Finally, although SHAP values improve interpretability, they explain model behavior rather than causal relationships. Therefore, the results should not be interpreted as evidence that changing a specific feature would necessarily cause a change in cancellation behavior.

In summary, this study demonstrates that modern machine learning methods, when combined with appropriate calibration and evaluation procedures, can produce reliable and actionable probability estimates for health insurance cancellation risk. LightGBM, in particular, provides a strong practical alternative to the current logistic regression benchmark. The findings support the adoption of calibrated gradient boosting models as a more effective tool for lapse and midterm cancellation modeling, especially in applications where probability accuracy is essential for business decision-making.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Nextgeneration Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. ACM.

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. The American Economic Review, 53(5), 941–973.

Assembleia da República (1979). Diário da República n.º 214/1979, Série I de 1979-09-15.

Autoridade de Supervisão de Seguros e Fundos de Pensões. (2021). Contrato de seguro. ASF.

Autoridade de Supervisão de Seguros e Fundos de Pensões. (2024). Observatório dos Seguros de Saúde divulga resultados do segundo Inquérito à População. ASF.

Autoridade de Supervisão de Seguros e Fundos de Pensões. (2025, junho 3). Circular n.º 6/2025: Condições Padrão de Seguro de Saúde.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. The Annals of Mathematical Statistics, 26(4), 641–647.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. Advances in Neural Information Processing Systems, 24, 2546–2554.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3.

Cawley, G. C., & Talbot, N. L. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research, 11, 2079–2107.

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

Constituição da República Portuguesa. (1976). Artigo 64.º: Saúde. Diário da República.

Eling, M., & Kiesenbauer, D. (2014). What do we know about lapse in the German life insurance market? Journal of Risk and Insurance, 81(4), 849–884.

Entidade Reguladora da Saúde. (2025). Sistema de Gestão de Reclamações, Elogios e Sugestões: Relatório do 1.º semestre de 2025. ERS.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), Automated Machine Learning: Methods, Systems, Challenges (pp. 3–33). Cham: Springer.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232.

Ghotra, B., McIntosh, S., & Hassan, A. E. (2015). Revisiting the impact of classification techniques on the performance of defect prediction models. Proceedings of the 37th IEEE/ACM International Conference on Software Engineering, 1, 789–800. IEEE.

Grinsztajn, A., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? arXiv preprint arXiv:2207.08815.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 1321–1330.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed. New York: Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. Proceedings of the International Conference on Learning Representations.

Mitchell, T. M. (1997). Machine Learning. New York: McGraw-Hill.

Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595–600.

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. Journal of Marketing Research, 17(4), 460–469.

Pauly, M. V. (1968). The Economics of Moral Hazard: Comment. The American Economic Review, 58(3), 531–537.

Pendzialek, J. B., Simic, D., & Stock, S. (2016). Differences in price elasticities of demand for health insurance: A systematic review. The European Journal of Health Economics, 17(1), 5–21.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). Data cascades in high-stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 3(3), 210–229.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the Theory of Games II (pp. 307–317). Princeton: Princeton University Press.

Xu, R., Lai, D., Cao, M., Rushing, S., & Rozar, T. (2015). Lapse Modeling for the Post-Level Period: A Practical Application of Predictive Modeling. Society of Actuaries Research Report.

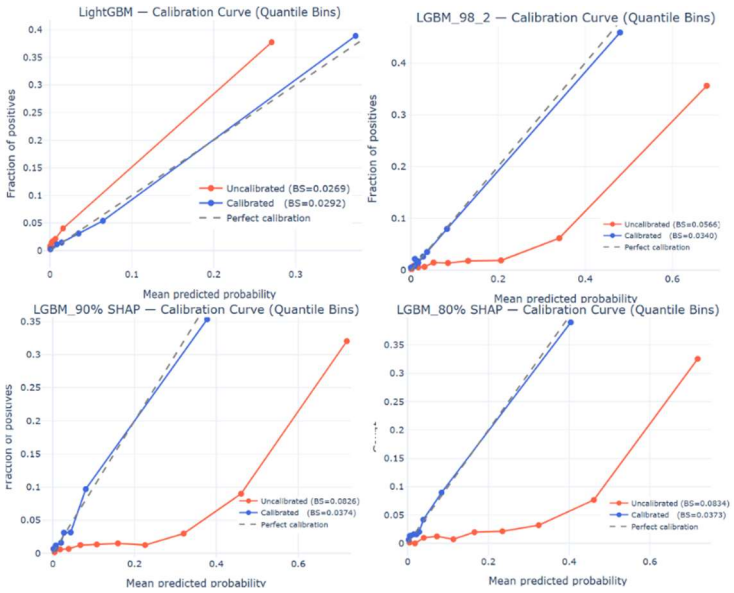

Figure 6: Calibration curve LightGBM's Lapse models

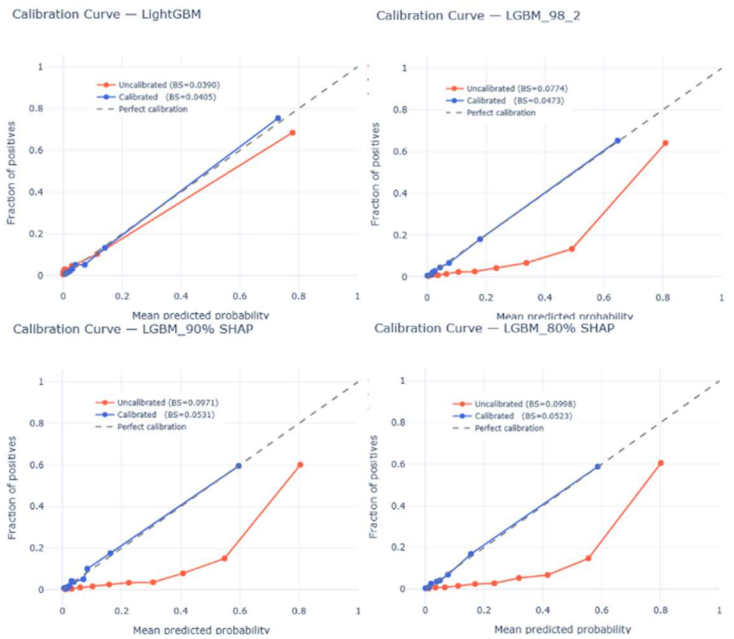

Figure 7: Calibration curve LightGBM's Midterm models