Stockholms
universitet

Prediction of Non-life Cancellation Rates with Machine Learning

Lei Sun

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2024:11
Försäkringsmatematik
September 2024

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Acknowledgements

I would like to express my deepest gratitude to my family for their support and encouragement throughout the journey of writing this thesis. Their belief in me has been a constant source of motivation, and I couldn't have completed this work without their love and patience.

I am also profoundly grateful to my supervisor, Dr.Lindholm, for his invaluable guidance, insightful feedback, and continuous support. His expertise and dedication have greatly contributed to the successful completion of this thesis.

Furthermore, I wish to extend my heartfelt thanks to colleagues from my previous company, whose encouragement have been instrumental in keeping me motivated throughout this process.

Thank you all for being a part of this journey.

Contents

Chapter 1

Introduction

This thesis discusses the use of tree-based ML techniques to estimate cancellation rates. Currently, a multitude of insurance enterprises engage in the estimation of cancellation rates for non-life insurance products, including but not limited to property, home, and automobile insurances, utilizing empirical methodologies. Traditional approaches to estimating these rates, though widely used, often rely on a constrained set of variables, such as the policyholder's age and geographical location, particularly in the context of automobile insurance. However, they often overlook additional determinants that could influence cancellation rates, such as the duration of the insurance coverage, the pricing of the insurance policy, and the sales channel utilized for policy acquisition.

The advent of Machine Learning (ML) techniques presents an unprecedented opportunity to refine the accuracy of cancellation rate predictions by incorporating a comprehensive array of factors potentially affecting these rates. Furthermore, ML algorithms enable a detailed analysis of the relative impact of each factor on the probability of policy cancellation. Such insights can empower insurance companies to implement targeted strategies aimed at mitigating cancellation rates. Enhanced predictions of cancellation rates also contribute to more precise risk assessments for insurance premium calculations.

This thesis aims to illuminate the potential of Machine Learning to surpass current empirical methods in estimating cancellation rates for non-life insurance products accurately. The first step in the study is to use the Synthetic Minority Over-sampling Technique (SMOTE) to generate a simulated dataset based on existing empirical imbalanced datasets. By generating synthetic data points, SMOTE enables us to create a balanced dataset, thereby facilitating more effective model training and improved prediction accuracy. After having created SMOTE balanced datasets, two predictive models will be studied: the Generalized Linear Model (GLM) and the Gradient Boosting Machine (GBM) for cancellation rate estimation. Through comparative analysis of these models' predictive outcomes, this research endeavors to illustrate the potential improvements ML can introduce over traditional estimation methodologies.

In this thesis, Chapter 2 provides a comprehensive review of current empirical methodologies employed in cancellation rate estimation. Chapter 3 discusses the application of SMOTE for data simulation and delves into the methodology behind this approach. In Chapter 4, we explore the theory and application of GLM, presenting the results and comparing them with the outcomes generated using SMOTE-enhanced data. Chapter 5 shifts focus to GBM, outlining the theoretical framework and application of this method for predicting cancellation rates. Finally, Chapter 6 provides a comparative analysis of the predictive capabilities of the GLM and GBM models, discussing the potential advantages and limitations of integrating machine learning techniques into the estimation process.

Chapter 2

Traditional Cancellation Rate Estimation

2.1 An Empirical Overview

The cancellation rate, which reflects customer retention and turnover, is a crucial measure for insurance companies. The dynamics of customer loss and acquisition, particularly through brand switching, constitute a significant challenge within the insurance sector, as emphasized by Brockett et al. (2008). This phenomenon highlights the intense competition within the insurance industry, where companies are continually striving to attract and retain policyholders in their efforts to gain market share.

In the realm of insurance, retention rates have significant impact on financial outcomes. Günther et al. (2014) underscore this point, demonstrating that a slight increase in customer retention can translate into millions of dollars in additional premium revenue. This financial implication highlights the direct correlation between effective customer retention strategies and an insurance company's revenue-generating capacity.

Given this background, the present study seeks to investigate the potential of ML techniques to enhance the prediction accuracy of cancellation rates within the insurance industry. By doing so, this research aims to provide actionable insights that could enable insurance firms to devise more effective strategies for customer retention, thereby mitigating the adverse financial impacts of high cancellation rates.

At present, the predominant approach used by many insurance companies for predicting cancellation rates relies on empirical methodologies. This traditional methodological framework is characterized by its dependence on historical data and statistical analysis to forecast future cancellation trends.

$$\text{Cancellation Rate} = \left(\frac{\text{Number of Cancellations during a specific time period}}{\text{Total Number of insurances during a specific time period}}\right) \times 100 \tag{2.1}$$

Here, the "number of cancellations" is determined by deducting the total number of policies, where the policyholder has opted not to renew at the end of a given insurance period from the total number of policies that were active during the same period.

The data for these calculations come from historical information collected over previous years. Actuaries often need to determine the mean cancellation rate for a period of 3 to 5 years. This method allows actuaries to combine real data with professional judgment and ensure well-informed and hopefully more accurate predictions. However this technique might not be able to capture the complex, non-linear relationships present in real-world data. Due to its simplicity, the empirical method might be less sensitive to subtle patterns and trends in the data. It likely fails to recognize how different factors influence each other, missing valuable insights the data could provide. This can lead to predictions that are not as accurate as they could be. For insurance companies, it is crucial to understand the factors that affect cancellation rates, as it directly supports the strategic goal of reducing policy cancellations, a notably advantageous outcome for any insurer.

Building on this understanding, we are going to use GLMs and machine learning techniques for estimating cancellation rates. Our goal is to investigate the potential of integrating this approach into future analytical processes. In the subsequent sections, we will take a look at the use of the SMOTE algorithm for imbalanced dataset, followed by applying both GLM and GBM for cancellation rate estimations.

Chapter 3

SMOTE Introduction

3.1 Imbalanced data

SMOTE is an acronym for Synthetic Minority Oversampling Technique, introduced in Chawla et al (2002). This method is an innovative approach to addressing the challenges posed by imbalanced data in classification problems (Chawla et al. 2002).

Imbalanced data occurs when there is a significant discrepancy in the frequencies observed across the different categories of a categorical response variable. Within the insurance sectors, for instance, canceled policies usually constitute a small fraction of the total number of policies. This situation underscores the necessity for applying techniques like SMOTE to ensure more balanced representation and improved model performance.

In this thesis, we analyze data from a non-life insurance product over a one-year period, encompassing a total of 24,025 insurance policies. The graphical analysis distinctly reveals that more than 90% of policyholders opt to renew their insurance policies, while around 10% choose to cancel (see Figure 3.1).

Based on this data, we could initially conclude that the model is to predict ”renewal” for each policyholder with a quite high accuracy. However, this perceived accuracy is misleading due to the imbalanced nature of the data. Imbalanced datasets can mislead the accuracy of a model because a model might only predict the majority class for all inputs and still achieve high accuracy without truly learning to identify the minority class, often

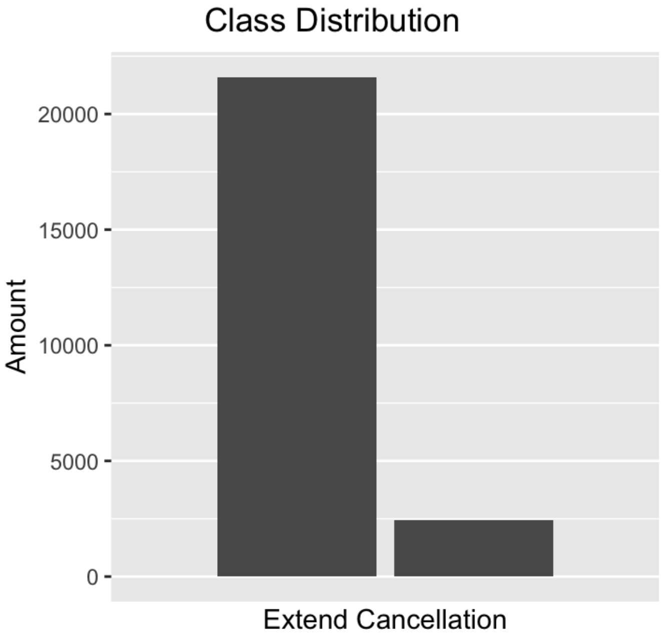


Figure 3.1: Imbalanced original data

the more critical to detect. To counteract this, techniques like resampling the data, using the Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic examples, or adjusting the decision threshold based on metrics other than accuracy are essential strategies.

In the upcoming section, we are going to discuss the possibility of using SMOTE for data resampling, with the goal of cultivating a more evenly distributed dataset. This effort aims to improve our model's effectiveness and insights, leading to stronger analysis results.

3.2 Implementing SMOTE

SMOTE is an over-sampling technique that enhances the representation of the minority class by generating "synthetic" examples instead of relying on traditional over-sampling methods, which often involve duplicating existing data points. This method was inspired by successful applications in fields like handwritten character recognition (T.M. and H. 1997). Extra training data is generated by performing certain operations to the existing real data (Chawla et al. 2002). This method generates synthetic examples rather than using over-sampling with replacement. It begins by selecting observations from the minority class, and for each of these observations, it finds its k nearest neighbors. Then, it randomly selects one of these neighbors and calculates the direction and distance to this neighbor to generate a new, synthetic data point along this path. This synthetic point is a combination of the feature space of the original sample and its randomly chosen neighbor, scaled by a random factor between 0 and 1. This factor is selected uniformly at random for each synthetic point, determining where along the line between the two points the new example will be placed. This process enriches the dataset with more diverse, yet plausible, minority class examples. This is pictorially presented below:

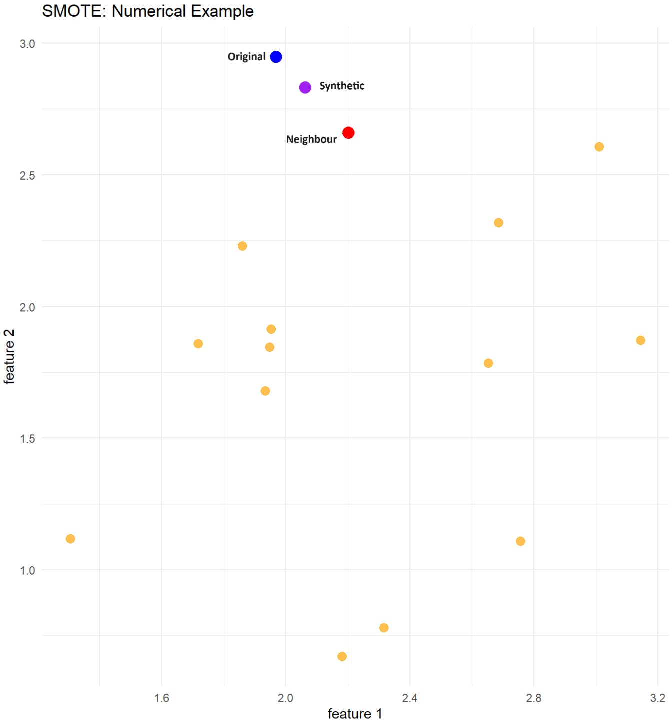


Figure 3.2: SMOTE: Numerical Example

In the above simple example, the original point $x_{orig} = (2.2, 2.66)$ and the nearest neighbour point is $x_{neig} = (1.97, 2.95)$. If the random factor is set to 0.6, according to the description,

new sample = original sample + factor $\times$(neighbour − original sample).

new sample = (2.2,2.66)+0.6*((1.97,2.95)-(2.2,2.66)) new sample = (2.2,2.66)+0.6*(-0.23,0,29)

new sample = (2.062, 2.834)

This process is repeated for each feature to generate a complete set of feature values for the new example, which is then added to the dataset as a synthetic example for the minority class.

This operation is actually very much like slightly moving the data point in the direction of its neighbor. This way, our synthetic data point is ensured that is not an exact copy of an existing data point while making sure that it is also not too different from the known observations in the minority class.

When dealing with categorical variables, SMOTE requires a slight modification to the process. It handles categorical variables by selecting the most frequent category among the k nearest neighbors of the original sample. Alternatively, for dataset with both categorical and continuous variables, a variant called SMOTE-NC (Nominal and Continuous) can be used, where categorical features are randomly chosen from the nearest neighbors. This approach ensures that the synthetic examples remain realistic and consistent with the distribution of categorical values in the minority class.

From the description of how SMOTE works we see that SMOTE has the advantage of not producing duplicate data points, but rather artificial data points that deviate marginally from the actual data points.

3.2.1 Data Preparation

Before using the SMOTE algorithm to re-balance the data, it is crucial to separate the dataset into training and test datasets. This allows us to evaluate the machine learning model's performance on data that was not used during the training phase.

To ensure that both training and test dataset has the same class distribution as the original one, we use stratified sampling to generate the these subsets. Stratified sampling randomly selects samples from each classification in proportion to that class' representation in the full dataset. This method could guarantee that the training and test sets reflect the class balance of the original data.

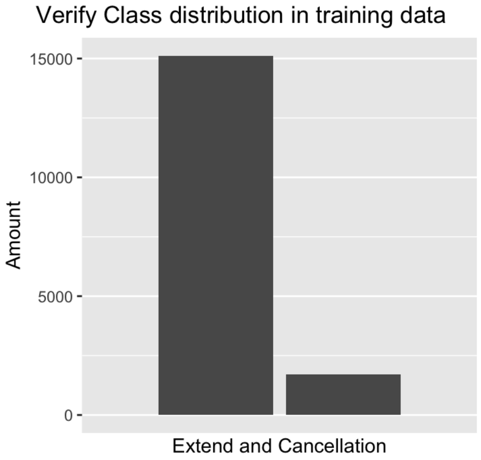


Figure 3.3: Imbalanced training dataset

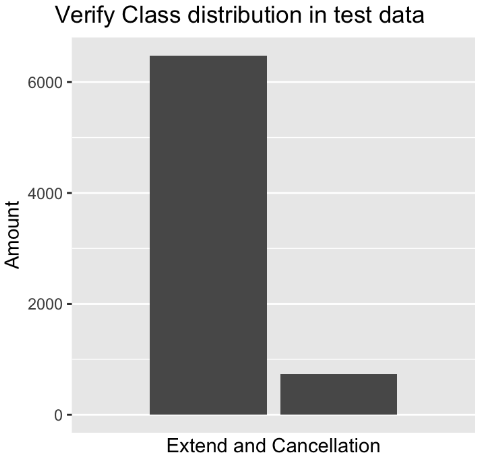


Figure 3.4: Imbalanced test dataset

By comparing the class distributions in Figure 3.3 and 3.4 with those of the original dataset, we can confirm that training and test datasets maintain the class proportions of original data. With this verification complete, we can proceed to implement the SMOTE technique to the training dataset.

Following the division of the dataset into training and testing subsets, we proceeded to define the binary response variable, y_i, which is assigned a value of 1 if policy i is cancelled, and 0 otherwise, such that

$$y_i = \begin{cases} 1 & \text{if policy } i \text{ is cancelled,} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

3.2.2 Applying SMOTE to Insurance Data

In the prior section, we discussed how SMOTE generates synthetic examples. In this thesis, we execute the resampling process using the statistical software **R**. The resampled dataset is generated using the **"SMOTE"** function from the **"DMwR"** package.

After executing the code, we examine the re-balance effects of SMOTE on our class by creating a bar graph similar to the one earlier constructed (see Figure 3.5). The purpose of this graph is to visually confirm that SMOTE has successfully increase the representation of the minority class.

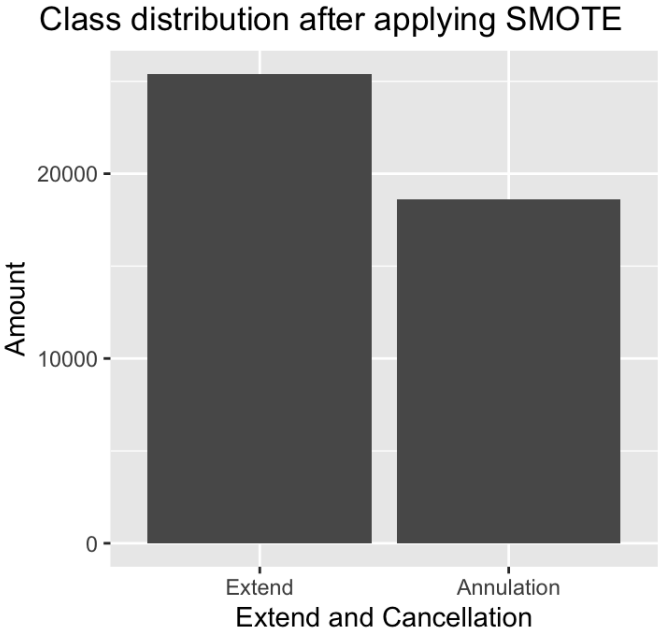


Figure 3.5: Balanced SMOTE data

The graph distinctly demonstrates a significant increase of the number of cancelled insurance policies, that we did not have previously. These additional entries are synthetic data points that have been created by SMOTE. The results indicate that, although the dataset still include the minority class, it has become considerably more balanced compared to the original dataset. Consequently, this enhanced balance allows for the use of machine learning to estimate the cancellation rate using the data augmented by SMOTE.

Chapter 4

Generalized Linear Model

4.1 Introduction of Generalized Linear Model

To assess the impact of SMOTE, we initially construct a Generalized Linear model using the original dataset. This step establishes a benchmark for evaluating the performance changed by SMOTE. Subsequently, we will compare the performance metrics pre- and post-SMOTE application.

Generalized linear models (GLM) were developed by John Nelder and Robert Wedderburn to unify different types of statistical models, such as linear regression, logistic regression and Poisson regression (J. Nelder and Wedderburn 1972).

In a GLM, each outcome Y of the explanatory variables is assumed to follow a distribution within the exponential family—a broad class of probability distributions that encompasses the normal, binomial, Poisson, gamma, and others. The conditional mean μ of the distribution is modeled as a function of the independent variables X, with the relationship defined by following link function:

$$\mathrm{E}(Y|X) = \mu = g^{-1}(X\beta) \tag{4.1}$$

In this equation, $\mathrm{E}(Y|X)$ represents the expected value of Y given X; $X\beta$ denotes the linear predictor, a linear combination of unknown parameters β, and g is the link function that connects the linear predictor to the mean of the distribution. The parameters β are typically estimated using maximum likelihood techniques (McCullagh and J. A. Nelder 1989).

As our response variable is binary (extension or cancellation), the Bernoulli distribution is chosen to be the distribution function and the interpretation of μ is the probability, p, of ”success” outcome Y.

$$g(p) = \mathrm{logit} p = \log(\frac{p}{1-p}) = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m. \tag{4.2}$$

The GLM is setup as a logistic regression model, in order to determine the corresponding log odds of the outcome which we then model as a linear combination of the explanatory variables.

In the expression, β_0 represents the intercept and $\beta_1, \beta_2, ..., \beta_m$ are the coefficients corresponding to explanatory variables $x_1, x_2, ..., x_m$.

The logit function can be written as

$$\log(\frac{p}{1-p}) = t, \tag{4.3}$$

where t is the linear function of the explanatory variables, which means that the probability p that an insurance policyholder will not renew their policy, is modeled through the logistic function below:

$$p = \frac{1}{1+e^{-t}}. \tag{4.4}$$

and the general logistic function $p : R \to (0, 1)$ can now be written as:

$$p = \sigma(t) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_m x_m)}}. \tag{4.5}$$

4.2 SMOTE's Enhancement on Model Performance

Considering the abundance of variables within our dataset, we selected those explanatory variables believed to impact the cancellation rate. These include the age of the customer, the sales channel through which the insurance was purchased, the amount of the premium, and the policy duration — the latter reflecting the length of time a customer has had their insurance policy.

Prior to estimation, it is beneficial to analyze the relationship between the cancellation rate and each explanatory variable.This analysis helps in understanding how individual variables influence the probability of insurance cancellation. For this purpose, we utilize a global method: Partial Dependence Plots (PDPs). PDPs evaluate the influence of specific features across all observations in the dataset, illustrating how these features affect the predicted outcome of a machine learning model (Friedman 1999a). The partial dependence function $\hat{f}_S$ based on the predictor $\hat{f}$ for the set of covariates S is estimated by calculating averages in the training data according to

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{4.6}$$

The partial dependence function shows the average marginal effect of given value(s) of features S on the prediction. In this formula, the x_S represents the features for which the partial dependence function is beinge plotted, while $X_C^{(i)}$ denotes the actual feature values from the dataset for the features not currently of interest. The parameter n is the number of observations in the dataset. It is important to note that the Partial Dependence Plot (PDP) assumes no correlation between the features in sets C and S. If this condition is not hold, the computed averages for the PDP may include data points that are either highly improbable or completely infeasible (Molnar 2022).

A partial dependence plot is particularly effective in revealing how individual features impact the model's predictions. It can illustrate whether the relationship between the target variable and a feature is linear, monotonic, or more complex. By providing these insights, PDPs help to identify the nature of the dependency between features and the target, thereby aiding in the interpretation of the model's behavior and in understanding the influence of each feature on the predicted outcomes (Molnar 2022).

In our logistic regression model, we consider several variables: the customer's age, the sales channel, the premium cost, and the duration for which the customer has held the policy. It is worth to mention that we do not employ the age variable directly. Instead, we apply a natural spline transformation of age with 10 degrees of freedom. Below, we present the PDP graphs for the explanatory variables (see Figure 3.6).

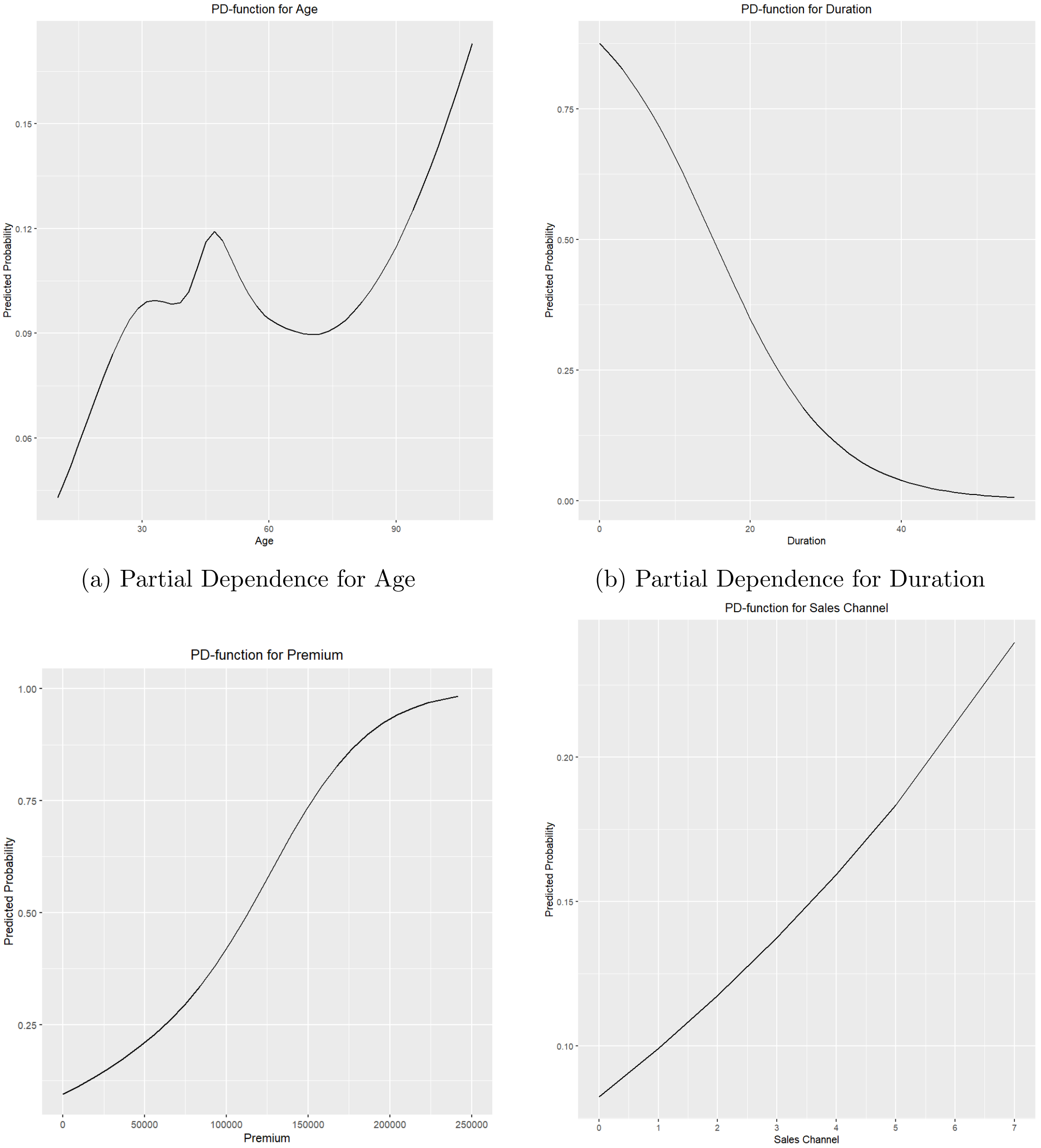
The first plot for age reveals a general trend where the probability of policy cancellation increases with age. However, there are slight decreases in the probability of cancellation around the ages of 50 and 70, indicating an increased loyalty for policyholders to keep the insurance during this age range. Beyond the age of 70, we observe a marked sharp increase in cancellation probability. A possible explanation for this pattern can be the cancellation of insurance policies due to the policyholder's death.

The analysis of the second plot, focusing on the length of time a customer has had their insurance policy, indicates that the longer a policyholder retains their insurance, the less likely they are to cancel the policy, suggesting a decreased propensity for cancellation among long-term policyholders.

The third plot explains that the probability of policy cancellation escalates with an increase in premium amount. There exists a specific premium threshold, beyond which the probability of policy non-renewal approaches almost 100%. This observation highlights a critical premium level, above which policyholders demonstrate a pronounced tendency to terminate their insurance coverage.

The last plot reveals a progressive increase in the predicted probability of insurance policy cancellation as the sales channel index, ranging from 0 to 7, rises. This suggests that the sales channel has influence on the probability of an insurance policy being canceled.

After having established the theoretical framework, we now turn to practical application by deploying the **"cv.glmnet"** function from the **"glmnet"** package, setting a



(a) Partial Dependence for Age

(b) Partial Dependence for Duration

(c) Partial Dependence for Premium

(d) Partial Dependence for Sales Channel

Figure 4.1: Partial Dependence Plots for GLM model

binomial distribution and applying a lasso penalty to train our model. The Least Absolute Shrinkage and Selection Operator (lasso) penalty is used to prevent over-fitting by adding a penalty equal to the sum of absolute value of the coefficients, all multiplied by a value λ. We use a example to explain it.

Let us assume we are aiming to predict a variable Y from three predictors, X_1, X_2 and X_3 and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, where "$\beta_0$" is our intercept and "$\beta_1$", "$\beta_2$", "$\beta_3$" are the coefficients. In a typical linear regression, we are simply aiming to find the values of β_0, β_1, β_2 and β_3 where the sum of squared residuals (SSR) as small as possible. With lasso penalty, we minimize instead

$$SSR + \lambda \cdot (|\beta_1| + |\beta_2| + |\beta_3|) \tag{4.7}$$

This can result in some coefficients reducing to zero. Consequently, lasso facilitates feature selection, which is useful when dealing with datasets that have a large number of features, some of which may be irrelevant or collinear. This not only enhances the model's interpretability but also its overall performance.

As SMOTE helps in balancing class distribution by generating synthetic observations, the amount of cancellations increases, and the overall class distribution in the training dataset no longer reflects the true distribution, which can affect the probability predicted by GLM. To correct this, we perform a calibration of the model's intercept β_0 to ensure that the predicted cancellation probabilities align more closely with the observed cancellation rate in the training data.

The calibration process involves adjusting the model's intercept β_0 to β_0^* such that the sum of the predicted probabilities matches the sum of the observed outcomes. Specifically, we calculate the calibrated intercept β_0^* as follows:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{p_i}^* = \sum_{i=1}^{n} \frac{1}{1 + e^{-(\beta_0^* + x_i'\hat{\beta})}} \tag{4.8}$$

where, y_i represents the observed outcome for observation i, with $\sum_{i=1}^{n} y_i$ denoting the sum of the observed outcomes (where $y_i = 1$ if cancellation, and $y_i = 0$ otherwise). $\hat{p_i}^*$ represents the adjusted probability that insurance policyholder will cancel their policy for observation i, and $\sum_{i=1}^{n} \hat{p_i}^*$ represents the sum of the observed outcomes. The variable n indicates the total number of observations in the dataset and $x_i'\hat{\beta}$ represents the linear predictors (without the intercept) for the same observation.

To solve for β_0^*, we need to calculate the sum of observed outcomes:

$$S_y = \sum_{i=1}^{n} y_i \tag{4.9}$$

and find β_0^* such that

$$S_y = \sum_{i=1}^{n} \frac{1}{1 + e^{-(\beta_0^* + x_i' \hat{\beta})}} \tag{4.10}$$

As this equation can not be solved analytically, so we used a numerical method to find the value of β_0^* that satisfies this equation. This method ensures that the model's predictions remain well-calibrated and aligned with real-world conditions. We use the **"uniroot"** function in **R** to executive the calculation. This calibration ensures that the predicted probability of cancellation matches the observed cancellation rate in the training data, leading to a more accurate and reliable model.

Following the calibration, the cancellation rate predicted by the model becomes more aligned with the actual observed rate. The cancellation rate with the original dataset is 10.07%. After balanced the dataset with SMOTE method, the cancellation rate becomes 9.79% after calibration.

Building on the previous discussion, we now turn our attention to the performance measures as precision, recall and F1-score to understand the model's performance in identifying customers who are likely to cancel their policies. Precision and recall values can be derived from the confusion matrix, which compares the model's predictions to the actual outcomes.

The confusion matrix provides the following numbers:

- True Positives (TP) - Correctly identified instances of the target class.
- True Negative (TN) - Correctly identified instances that not of the target class.
- False Positive (FP) - Incorrectly identified instances as the target class.
- False Negative (FN) - Incorrectly identified instances as not being of the target class.

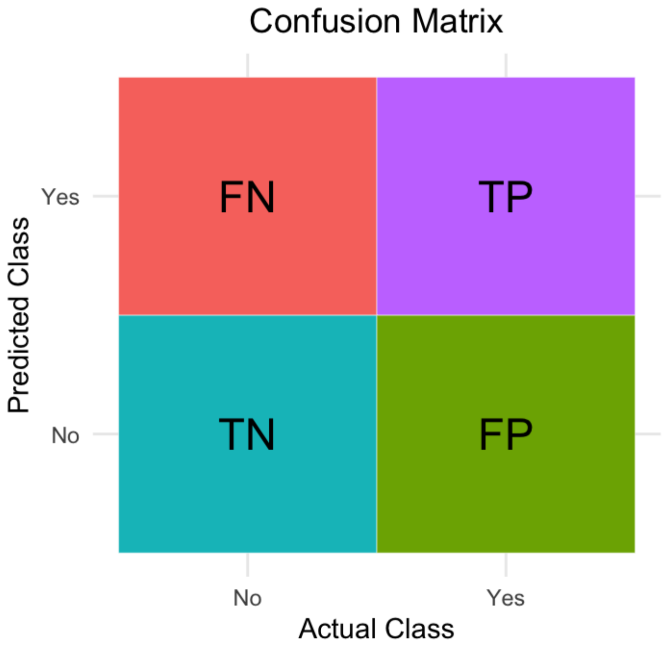


Figure 4.2: Confusion Matrix

Precision, which ranges from 0 to 1, refers to the model's ability of identifying positive predictions that are correct. A higher precision value suggests that the model is more accurate in its positive predictions (Powers 2011).

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{4.11}$$

Recall, also ranging from 0 to 1, measures the model's ability to correctly identify actual positive cases. A higher recall indicates that the model successfully captures more of the true positive observations, over all the positive cases in the data (Powers 2011).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.12}$$

F1-score is a measure that combines both precision and recall into a single measure. It is especially valuable in scenarios where the class distribution is imbalanced. The F1-score is calculated as the harmonic mean of precision and recall, offering a balanced assessment of a model's performance across these two important metrics (Powers 2011). Like precision and recall, the F1 ranges ranges from 0 to 1, with a higher F1-score reflecting a good balance between precision and recall, and a lower score indicating a disparity between them. The calculating formula is as follows:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.13}$$

With original data, we achieved a precision of approximately 0.742, a recall of 0.742 and F1-score of 0.742.

Next, we apply logistic regression to the dataset augmented by SMOTE, using precision, recall, and F1-score as benchmarks to evaluate the model's performance. In contrast to the previous analysis using the original training data, this phase utilizes the enhanced dataset. The outputs obtained from this adjusted model are:

The precision decreases to 0.418, suggesting a reduction in the model's overall accuracy. Meanwhile recall increases to 0.802, highlighting an improvement in correctly identifying cases of cancellations. F1-score decreases to 0.550 indicates that the balance between precision and recall has worsened.

The changes suggest that utilizing SMOTE has enhanced the model's ability to capturing positive cases - as evidenced by the increase of recall. However, it has also made the model more susceptible to misclassifying negative cases as positive, which is indicated by the reduction of precision.

To further evaluate our model's performance, we turn to the ROC (Receiver Operating Characteristic) curve. The ROC curve visually represents the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) at various threshold settings, providing a clear indication of the model's discriminatory power (Fawcett 2006). The 45-degree diagonal line in the ROC plot serves as a baseline for assessing the performance of the model. If the ROC curve of a model is above this red line, the model is performing better than random guessing. As illustrated in the following example, this visualization is instrumental in assessing the model's ability to distinguish between positive and negative classes. Accompanying the ROC curve is the AUC (Area Under the Curve), a measure that quantifies the overall effectiveness of the model's predictions (Hanley and McNeil 1982). The AUC value ranges from 0 to 1, where an AUC of 1 indicates perfect classification, and an AUC of 0.5 suggests performance no better than random guessing. The closer the AUC value is to 1, the better the model fits the data, effectively balancing sensitivity and specificity. We explore both aspects in Figure 3.8 and Figure 3.9.

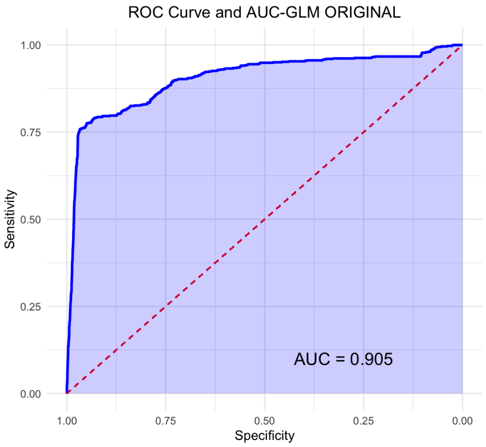


Figure 4.3: ROC Curve GLM with original data

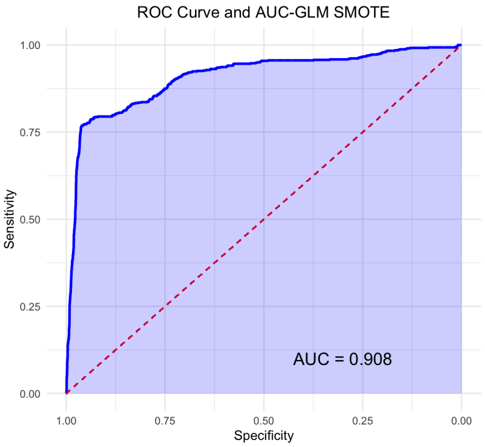


Figure 4.4: ROC Curve GLM with SMOTE data

In the example, the AUC value for GLM model with original data is 0.905, shaded in blue, indicating a strong ability to distinguish between the positive and negative classes. After applying SMOTE, the AUC value increases marginally to 0.908. This slight increase in AUC suggests that the SMOTE-enhanced model has a slightly better ability to discriminate between policy cancellations and non-cancellations. However, the small difference in AUC values also indicates that while SMOTE may contribute to a more balanced representation of the classes, the original GLM model was already performing well in terms of its predictive power. Therefore, we should consider whether it is necessary to calibrate the cancellation rate calculated by GLM with SMOTE-enhanced data. To address this, we introduce the Brier Score, which measures the mean squared difference between the predicted probability assigned to the possible outcomes and the actual outcome (Brier

1950). The Brier score is calculated as follows:

$$\text{BrierScore} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \tag{4.14}$$

where:

- n is the number of predictions.
- f_i is the predicted probability of the positive class (e.g., cancellation) for observation i.
- o_i is the actual outcome for observation i, which is 1 if the event happened (e.g., policy was canceled) and 0 if it did not.

The Brier score ranges from 0 to 1. A lower Brier score indicates better predictive accuracy. The Brier score of the GLM model with original data was 0.049 and increases to 0.052 with SMOTE-enhanced data, suggesting that, while SMOTE helps balance the class distribution, it introduce amount of noise or variance that slightly affects the model's overall accuracy.

Despite the SMOTE-enhanced model shows an improvement in identifying cancellations and discriminating between cancellations and non-cancellations, as evidenced by the increase in precision, recall and AUC, the introduction of SMOTE appears to have also introduced some noise or variance, as reflected in the increase of Brier score.

In the forthcoming section, we will explore the potential of using Gradient Boosting Machines (GBM) as a predictive model. We will conduct a comparative analysis with the Generalized Linear Model (GLM) and investigate whether the GBM model provides a better fit for SMOTE enhanced data.

Chapter 5

Gradient Boosting Machines

5.1 Introduction of Gradient Boosting Machines

Leo Breiman first introduced the concept of gradient boosting in 1997, observing that boosting could be understood as an optimization algorithm applied to a specific cost function (Breiman 1997). Building on this foundational idea, Jerome H. Friedman developed explicit regression gradient boosting algorithms (Friedman 1999a; Friedman 1999b), while a more general functional gradient boosting framework was concurrently presented by Mason et al. in 1999 (Mason et al. 1999a; Mason et al. 1999b). Unlike Generalized Linear Models (GLM), which handle various types of response variables via a link function, Gradient Boosting Machines (GBM) are ensemble learning models that work by sequentially enhancing the predictive strength of multiple weak learners, typically decision trees, to build a highly accurate and generalizable predictive model. GBM specifically utilizes the gradient descent algorithm to minimize a loss function, progressively improving model performance.

5.1.1 Optimizing Prediction Through Sequential Learning

Gradient Boosting builds models incrementally, in an additive and sequential fashion. The core idea behind this algorithm is to develop new base-learners that are highly correlated with the negative gradient of the loss function associated with the entire ensemble. The loss function acts as a measure that indicates how well the model's parameters fit the underlying data (see Natekin and Knoll 2013).

To better understand the algorithm, let us begin from the foundational assumptions:

Let y be the vector of observed responses, where y_i represents the target for the i^{th} observation.

Let X represent the feature matrix, with each row X_i corresponding to the feature values for the i^{th} observation.

The dependent function is denoted as $f(x)$ and the ensemble model after m steps is denoted as $F_m(x)$, which makes predictions for feature vector x.

The algorithm starts with an initial model, $F_0(x)$, often a constant value:

$$F_0(x) = \arg\min_{h_m \in \mathcal{H}} \sum_{i=1}^{M} L(y_i, f(x)) \tag{5.1}$$

where $L(y_i, f(x))$ is the loss function, measuring the difference between the actual target values y_i and the predictions. For regression, this could be mean squared error (MSE), and for classification, it could be logistical loss. For $m \geq 1$, it is given by

$$F_m(x) = F_{m-1}(x) + \left(\arg\min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^{M} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right] \right)(x), \tag{5.2}$$

where $h_m \in \mathcal{H}$ is a base-learner function, such as decision trees or splines.

For each iteration $m = 1$ to M, the algorithm performs the following steps:

a. Compute the negative gradient of the loss function with respect to the predictions made by the current model, evaluated at each data point:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \tag{5.3}$$

These r_{im} values serve as pseudo-residuals, indicating the direction in which to adjust the predictions to reduce the loss.

b. Fit a new model $h_m(x)$ to these pseudo-residuals:

$$h_m(x) = \arg\min_{h_m \in \mathcal{H}} \sum_{i=1}^{M} (r_{im} - h_m(x_i))^2 \tag{5.4}$$

c. Determine the optimal step size (learning rate) γ_m that minimizes the loss when adding $h_m(x)$ to the ensemble:

$$\gamma_m = \arg\min_{h_m \in \mathcal{H}} \sum_{i=1}^{M} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{5.5}$$

d. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5.6}$$

The final model $F_m(x)$ is used to make predictions, where M is the total number of iterations (trees).

5.2 Application and Performance analysis with SMOTE-Enhanced Data

Consistent with the methodology applied in the preceding chapter, we initially fit the model using the original dataset, followed by a re-estimation with data augmented by SMOTE. We will then undertake a detailed comparative analysis of the results, examining the impact of SMOTE both pre- and post-application.

As with our prior approach, we separated the dataset into training and testing subsets and then fitted the GBM model to the data, using y_i as the response variable along with the covariates.

The model is given by the function:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{5.7}$$

where $F(x)$ is the final ensemble model, $h_m(x)$ represents the m-th decision tree, and γ_m

is the shrinkage rate of the m-th tree to the model.

Each tree's influence on the final model is adjusted by a factor known as the learning rate or shrinkage rate, represented as γ_m. This rate controls the contribution of the m-th tree, helping to prevent the model overfitting by moderating the model's complexity. This approach enhances the model's robustness and its ability to generalize effectively from the training data to new, unseen data (Friedman 1999a).

We selected the same relevant explanatory variables as in the previous analyses, which include customer ages, sales channel, premier amount and duration.

The distribution argument is set to "bernoulli" as a binary outcome and a logistic regression framwork for the GBM. Initially, the model was trained using 1,000 trees to enhance its predictive performance. However, by using the **"gbm.perf"** function, which estimates the optimal number of iterations, early stopping was triggered, and the optimal model was found to have 826 trees.

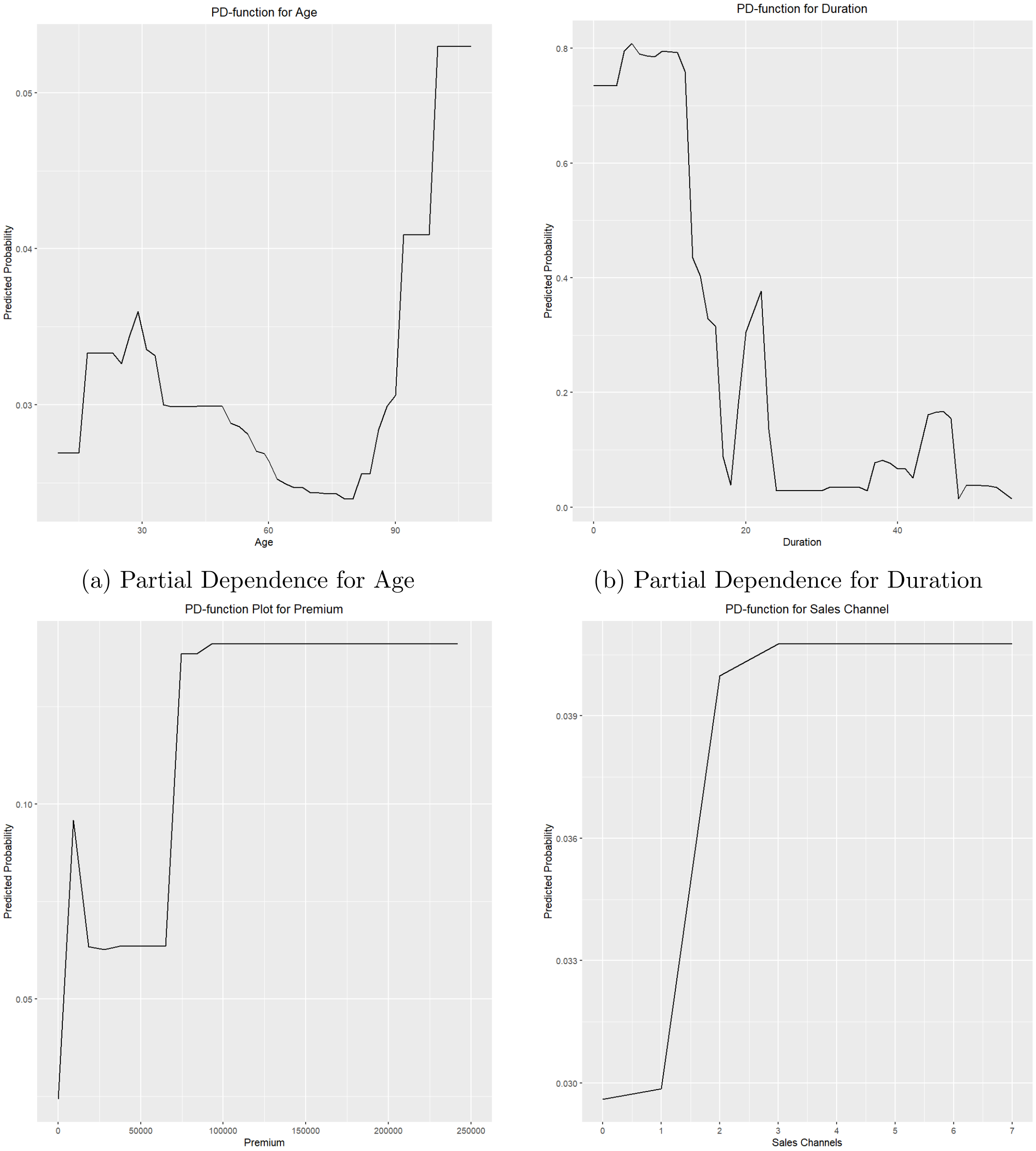
As earlier, we begin by examining the PDP for the GBM model, applying the same four variables used in the GLM. Unlike the GLM model, which uses a flexible spline transformation for age as predictor, the GBM model handles age as a linear predictor, and is capable of capturing complex interactions and non-linear relationships.

The first plot for age describes a trend similar to that observed in the GLM model: younger policyholders have a slightly higher probability of cancelling their policies, while policyholders between the ages of 50 and 70 show greater loyalty in maintaining their insurance.

The trend observed in the second plot for duration is consistent with the trend seen in the GLM model. It shows us that policies with shorter duration are more prone to cancellation.

The third plot is consistent with the trend in the GLM model as well, suggesting that higher premium amounts are associated with higher cancellation rates.

The last plot reveal that the effectiveness of sales channels varies, with some channels contributing to higher retention and others to higher cancellations. This could be due to



(a) Partial Dependence for Age

(b) Partial Dependence for Duration

(c) Partial Dependence for Premium

(d) Partial Dependence for Sales Channels

Figure 5.1: Partial Dependence Plots for GBM model

differences in customer service quality or the effectiveness of communication.

Next, we take a look at the values of precision and recall. The precision value was 0.760, while the recall was 0.741 and F1-score 0.750. The AUC value was 0.955 (see Figure 5.2) and the predicted cancellation rate was 9.82%. The Brier score was 0.040.

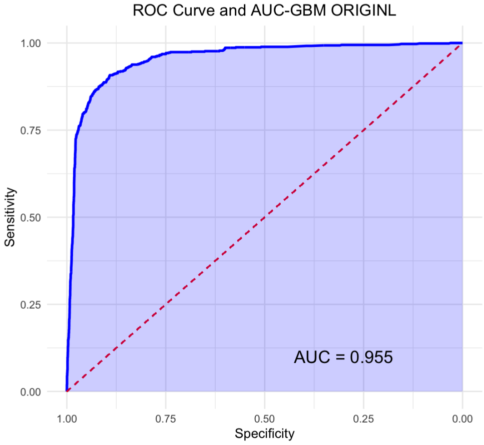


Figure 5.2: ROC Curve GBM with original data

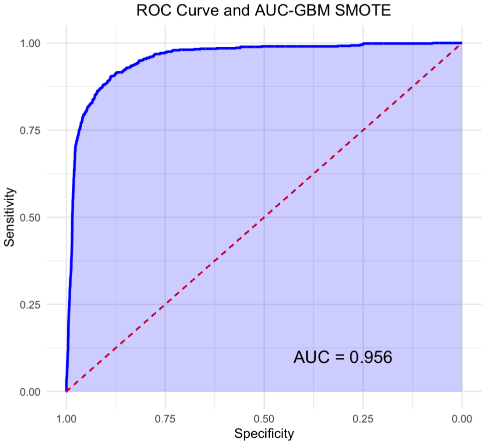


Figure 5.3: ROC Curve GBM with SMOTE data

Compared to GLM, the GBM model appears to be the better choice for predicting cancellations. It has a higher precision and F1-score, indicating better balance and performance in identifying cancellations. Most importantly, the GBM model's higher AUC value demonstrates superior discriminative ability compared to the GLM. Despite the GLM having a slightly higher recall, the overall performance, as reflected by the AUC, strongly favors the GBM.

Upon using SMOTE-augmented training data instead of the original training data in our model, we obtained the following outputs.

The precision value was 0.660, the recall was 0.796 and the F1-score was 0.722. The predicted cancellation rate was 12.15%, and the AUC value was 0.956 (see Figure 5.3). Notably, after applying SMOTE, the optimal number of iterations increased to 16,337, indicating that the model's ability to capture complex pattern is enhanced. The Brier score of the GBM model with SMOTE-enhanced data was 0.050, suggesting that on average, the predicted probabilities are close to the actual outcomes, reflecting strong model performance.

The SMOTE-enhanced model shows a trade-off between precision and recall. While it achieves a higher recall and slightly better AUC, it does so at the cost of lower precision. This suggests that while SMOTE helps the model capture more cancellations, it may also introduce some noise, leading to slightly less accurate and less well-calibrated predictions overall.

Chapter 6

Methodical Comparison

We implemented SMOTE on our imbalanced dataset and used two predictive models to estimate cancellation rates. It is evident that SMOTE strengthens our models' ability to identify cancellations by increasing the representation of the minority class.

Nonetheless, determining the suitable model for our specific needs remains a question of great interest to the insurance sector. Which model is more suitable to forecast cancellation rates? Based on the analysis in the previous chapters, the GBM model with original data exhibited a well-rounded performance, particularly in maintaining a strong balance between precision and recall, as indicated by its higher F1-Score. On the other hand, the GBM model with SMOTE-enhanced data demonstrated a greater ability to detect cancellations, evidenced by its higher recall and slightly improved AUC.

Model	**Data Type**	**Precision**	**Recall**	**F1-Score**	**AUC**	**Brier Score**
GLM	Original	0.742	0.742	0.742	0.905	0.049
GLM	SMOTE	0.418	0.802	0.550	0.908	0.052
GBM	Original	0.760	0.741	0.750	0.955	0.040
GBM	SMOTE	0.660	0.796	0.722	0.956	0.050

Table 6.1: Comparison of GLM and GBM models with Original and SMOTE-enhanced data

Further, we computed the Brier score from all models. The GBM model with original data achieved a Brier score of 0.040, whereas the GBM model with SMOTE-enhanced data resulted in a higher Brier score of 0.050. This suggests that the GBM model with original data provides better-calibrated probability predictions compared to the GBM model after SMOTE enhancement.

Analyzing these results leads us to conclude that each model offers distinct advantages, depending on the specific focus of the analysis. If the primary goal is to maximize the identification of cancellations, the GBM model with SMOTE-enhanced data might be preferred due to its higher recall and slightly improved AUC. However, if the focus is on maintaining higher precision and generating well-calibrated predictions, the GBM model with original data would be more appropriate.

The principal limitation of our empirical method is the weak correlation between the explanatory variables and the target variable, with the exception of duration. The correlation table below demonstrates that variables such as sales channel, age and premium show only weak linear relationship with cancellation rates. However, these empirical correlations do not fully capture the complex, non-linear relationships that can exist. In contrast, our application of the GLM and GBM models reveals a more substantial relationship between these explanatory variables and the probability of cancellation. Through these models, we have successfully accounted for the influence of each variable, including those with weak empirical correlations, on the cancellation probability. The models' ability to identify and incorporate these relationships highlights their strength in providing a more accurate and comprehensive understanding of the factors driving cancellations.

Variable	**Correlation**
Sales Channel	0.0643
Age	0.0185
Premium	-0.0484
Duration	-0.594

Table 6.2: Correlation Between Explanatory Variables and Cancellation

The Partial Dependence Plots (PDPs) further underscore this by providing a detailed exploration of how changes in each variable impact cancellation probabilities, revealing non-linear and interaction effects that may not be apparent from empirical correlations alone.

In our analysis we utilized SMOTE to synthesize new data, which effectively addresses the challenge of class imbalance. The advantage of the method is that it enhances the classifier's performance specifically on the minority class. By increasing the representation of the minority class, SMOTE ensures that the model has more balanced data to learn from, which reduces the risk of the model being biased towards the majority class. This, in turn, improves the model's ability to correctly identify observations of the minority class. Moreover, SMOTE contributes to better overall model generalization by providing a more representative dataset. This method enriches the dataset without requiring additional real-world data, which can often be challenging or costly to obtain.

As with other methods, the use of SMOTE comes with both advantages and disadvantages. The disadvantage of SMOTE is that can introduce noise and potential overfitting, particularly when synthetic samples generated from outliers or noisy data are not representative of the underlying class distribution. Additionally, the increased computational complexity with larger and higher-dimensional datasets raises practical challenges.

Building upon the understanding of SMOTE's strengths and limitations, we observe that SMOTE-enhanced models outperform traditional methods in managing imbalanced datasets. Through the application of machine learning techniques, it becomes possible to consider a comprehensive range of factors influencing predictions and thus enhance the predictive accuracy. However, it is important to recognize that current datasets may lack crucial future information necessary for refining cancellation rate forecasts. While SMOTE effectively addresses the issue of class imbalance, it cannot account for unknown future variables that could impact policy cancellations.

To overcome this limitation and further enhance the accuracy of our predictions, it is advisable to integrate model outcomes with anticipatory information. This approach ensures that predictions take into account potential future developments, providing actuaries with more accurate insights into fluctuations in policy amounts. Such precision is crucial for the calculation of premium risks, highlighting the importance of adaptive, data-driven decision-making in the actuarial field.

Bibliography

Breiman, L. (June 1997). *Arcing The Edge.* Technical Report 486. Statistics Department, University of California, Berkeley.

Brier, Glenn W. (1950). “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1, pp. 1–3.

Brockett, P.L. et al. (2008). “Survival analysis of a household portfolio of insurance policies: How much time do you have to stop total customer defection?” In: *Journal of Risk & Insurance* 75.3, pp. 713–737.

Chawla, N.V. et al. (2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Dobson, A.J. and Barnett A.G. (2018). *An Introduction to Generalized Linear Models.* New York: Chapman and Hall/CRC.

Fawcett, Tom (2006). “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8, pp. 861–874.

Friedman, J.H. (Feb. 1999a). *Greedy Function Approximation: A Gradient Boosting Machine.*

— (Mar. 1999b). *Stochastic Gradient Boosting.*

Günther, C.C. et al. (2014). “Modelling and predicting customer churn from an insurance company”. In: *Scandinavian Actuarial Journal* 2014.1, pp. 58–71.

Hanley, James A and Barbara J McNeil (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. In: *Radiology* 143.1, pp. 29–36.

Hastie, T, J Qian, and K Tay (2023). *An Introduction to glmnet.* Available at `https://glmnet.stanford.edu/articles/glmnet.html`.

Lalander, P and A Agresti (2013). *Categorical Data Analysis.* New Jersey: John Wiley Sons.

Leiria, M., N. Matos, and E. Rebelo (2021). “Non-life insurance cancellation: a systematic quantitative literature review”. In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 46.4, pp. 593–613.

Mason, L. et al. (1999a). *Boosting Algorithms as Gradient Descent.* Ed. by SA. Solla, TK. Leen, and K. Müller.

— (May 1999b). *Boosting Algorithms as Gradient Descent in Function Space.* Archived from the original on 2018-12-22.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models.* 2nd. Chapman and Hall/CRC.

Molnar, Christoph (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable.* Independently published.

Natekin, A. and A. Knoll (2013). "Gradient Boosting Machines, A Tutorial". In: *Frontiers in Neurorobotics* 7, p. 21. DOI: 10.3389/fnbot.2013.00021.

Nelder, J. and R. Wedderburn (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384. DOI: 10.2307/2344614.

Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation". In: *Journal of Machine Learning Technologies* 2.1, pp. 37–63.

T.M., Ha and Bunke H. (1997). "Off-line, Handwritten Numeral Recognition by Perturbation Method". In: *Pattern Analysis and Machine Intelligence* 19.5, pp. 535–539.