Predictive Modeling for Policy Lapse Forecasting Using Machine Learning

DISSERTATION

Submitted in partial fulfillment of the requirements of the

Degree: **MTech in Data Science and Engineering**

By

Justin P Mathew

Under the supervision of

Venkata Girish Kumar Nidra
Assistant Consultant

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(July, 2024)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
SECOND SEMESTER 2023-24

DSECLZG628T DISSERTATION

Dissertation Title : Predictive Modeling for Policy Lapse Forecasting Using Machine Learning

Name of Supervisor : Venkata Girish Kumar Nidra

Name of Student : Justin P Mathew

ID No. of Student :

Abstract

This dissertation investigates the intersection of Public Policy Analysis, Predictive Analytics, and Policy Lifecycle Management, aiming to develop and apply predictive models to forecast policy lapses. Policy lapses—failures in policy renewal, enforcement, or compliance—can lead to considerable disruptions and inefficiencies. Traditionally, policy management has been reactive, addressing lapses only after they occur. However, the advent of predictive analytics offers a promising shift towards a proactive approach.

Predictive analytics leverages historical data and statistical algorithms to forecast future events with high accuracy. In policy management, these tools can analyze patterns and identify indicators preceding policy lapses. Factors such as policy age, compliance rates, economic conditions, and stakeholder engagement can be systematically evaluated to predict the likelihood of a policy failing to sustain its intended impact. This dissertation aims to bridge the gap between traditional policy analysis and modern predictive methodologies by developing robust predictive models. These models will provide policymakers with early warnings and actionable insights, enabling them to preemptively address potential lapses, thereby enhancing the efficiency and effectiveness of policy management and contributing to the overall stability and sustainability of public administration practices. The research employs a mixed-methods approach, combining quantitative data analysis with qualitative assessments to ensure a comprehensive understanding of the factors influencing policy lapses.

The objectives of this dissertation include developing a predictive model for policy lapses using advanced statistical techniques and machine learning algorithms, identifying key predictors of policy lapses, validating the predictive model with historical policy data, and assessing its practical implications for policymakers. The research also aims to enhance policy lifecycle management by providing recommendations for using predictive analytics to reduce policy lapses and contribute to the broader field of predictive analytics in public administration. Additionally, the dissertation will propose future research directions to refine predictive models and explore their applicability in different policy environments.

The scope of this dissertation encompasses several key areas like predictive analytics. It includes a comprehensive literature review on policy lifecycle management, policy lapses, and predictive analytics; data collection from historical policy records or open-source datasets; model development using advanced statistical techniques and machine learning algorithms; model validation with historical data; analysis of key predictors; and practical implications assessment. The dissertation will provide actionable recommendations for improving policy lifecycle management and contribute to the field by developing a scalable and adaptable predictive model.

The research is structured into several phases, beginning with a literature review and conceptual framework, then data collection, model development, model validation, analysis of key predictors, assessment of practical implications, and finally, writing and finalizing the dissertation. Each phase is designed to systematically build towards enhancing policy lifecycle management through predictive analytics.

In summary, this dissertation explores the potential of predictive analytics to transform policy lifecycle management by reducing the occurrence of policy lapses through informed, proactive decision-making. By developing and validating predictive models, the research aims to provide policymakers with the tools necessary to anticipate and mitigate policy lapses, thereby enhancing the sustainability and effectiveness of public policies.

Key Words:

- Machine Learning,
- Predictive Analytics
- Proactive Decision-Making
- Policy Lapses
- Policy Management

List of Symbols & Abbreviations used

ML : Machine Learning
EDA : Exploratory Data Analysis
RFC : Random Forest Classifier
FE : Feature Engineering
SVM: Support Vector Machine

List of Tables

Table 1: Column Names and their Data Types……………………………………………………15

List of Figures

Figure 1: Correlation Heatmap for Predicting Policy Lapse (is_lapse)………………………16
Figure 2: Original Class Distribution…………………………………………………………17
Figure 3: Resampled Class Distribution………………………………………………………18

Table of contents

Abstract..………………………………………………………………………………………02
List of Symbols and Abbreviations..…...............04
List of Tables..…..........04
List of Figures...04
Chapter 1...06
Introduction 06
Background and Motivation..06
Problem Statement...06
Objective...… 07
Scope and Significance of Study...….....07
Summary…………..……......08
Chapter 2..09
Introduction 09
Policy Lifecycle Management...09
Predictive Analytics and Objectives Met...10
Machine Learning Algorithms...11
Summary…………..…......……......12
Chapter 3...… 13
Data Description...13
Data Acquisition..14
Data Preprocessing ....…........................15
Correlation Analysis...16
Handling Class Imbalance……………………………………………… ………..17
Summary……………………………………………………………………………..18
Directions for future work after mid semester………………………………………………19
Bibliography / References………………………………………………………………20

Chapter 1

Introduction

1.1 Background and Motivation

Policy lapses, particularly within the insurance sector represents a significant challenge for effective policy management. A policy lapse occurs when a policyholder stops paying premiums, leading to the termination of the policy. This issue not only affects the financial stability of insurance companies but also undermines the trust of policyholders and the overall effectiveness of policy administration.

Traditional methods of managing policy lapses are often reactive, focusing on rectifying lapses after they occur rather than preventing them. These methods include sending reminder notices and contacting policyholders directly, which can be both time-consuming and costly. Additionally, they do not leverage the vast amounts of data available that could potentially predict and prevent lapses before they happen.

The motivation for this study stems from the need for a more proactive approach to policy management. By employing predictive modeling techniques, it is possible to anticipate policy lapses and take preemptive actions to mitigate their occurrence. This proactive approach can enhance the efficiency of policy management, reduce financial losses, and improve customer satisfaction by maintaining continuous policy coverage.

1.2 Problem Statement

The primary problem addressed in this study is the lack of predictive tools for anticipating policy lapses. Current reactive methods are inadequate for preventing lapses and do not leverage the available data effectively. This leads to inefficiencies in policy management and potential financial losses for insurance companies. By developing a predictive model, this research aims to provide a solution to this critical issue.

Specifically, this study seeks to answer the following questions:

- What are the primary predictors of policy lapses?
- How accurately can policy lapses be predicted using machine learning algorithms?
- What actionable insights can be derived from the predictive model to improve policy management?

1.3 Objectives of the Study

The main objectives of this research are as follows:

- **Develop a Predictive Model for Policy Lapse Forecasting:** Utilize machine learning techniques to create a model that can accurately predict policy lapses.
- **Identify Key Predictors of Policy Lapses:** Analyze the dataset to determine which factors significantly influence the likelihood of a policy lapse.
- **Evaluate the Performance of Different Predictive Algorithms:** Compare the accuracy and efficiency of various machine learning algorithms in forecasting policy lapses.
- **Provide Actionable Recommendations for Policymakers:** Offer insights based on the model's predictions to help policymakers take proactive steps to prevent lapses.

1.4 Scope of the Study

This study focuses on analyzing insurance policies, utilizing datasets obtained from Kaggle. The data covers includes various features such as Entry Age, Sex, Policy Type, Payment Mode, Policy Status and more.

The scope of this research is limited to the data provide and may not generalize to other companies or sectors. Additionally, the study will primarily explore machine learning techniques and may not delve deeply into other predictive methods.

1.5 Significance of the Study

This research makes significant contributions to both the field of policy management and predictive analytics. By introducing a novel approach to policy lapse forecasting, it advances the current understanding of how predictive modeling can be applied to policy management.

The practical implications of this study are substantial. By accurately predicting policy lapses, insurance companies can take preemptive actions to maintain continuous coverage, reduce financial losses, and improve customer satisfaction. Policymakers can use the insights derived from the predictive model to design more effective policy interventions and communication strategies.

1.6 Research Questions and Hypotheses

To guide this research, the following research questions have been formulated:

1. What are the primary predictors of policy lapses?
2. How accurately can policy lapses be predicted using machine learning algorithms?
3. What actionable insights can be derived from the predictive model to improve policy management?

Based on these questions, the study tests the following hypotheses:

1. Policies with certain characteristics (e.g., age, compliance rates) are more likely to lapse.
2. Machine learning models can predict policy lapses with high accuracy.

1.7 Methodology Overview

This study employs a combination of data collection, preprocessing, and machine learning techniques. The research methodology involves the following steps met till Midterm :

- **Data Collection:** Gathering relevant data from Kaggle, including policyholder demographics, payment history, and policy details.
- **Exploratory Data Analysis (EDA):** Conducting initial analysis to understand the data structure, identify patterns, and detect any anomalies.

1.8 Summary

This chapter introduced the research topic, outlined the problem statement, and presented the study's objectives, scope, and significance met till mid term. The next chapter will elaborate the objectives met till now on policy lapse management and predictive modeling, providing a theoretical framework for this study.

Chapter 2

Literature Review and Objectives Met

2.1 Introduction

The literature review serves as the foundation for understanding the current state of research in policy lapse management and predictive modeling. This chapter synthesizes existing studies, identifies key predictors of policy lapses, and evaluates the application of machine learning techniques in forecasting policy lapses. Additionally, this chapter highlights the objectives met so far till the midterm.

2.2 Policy Lifecycle Management

2.2.1 Definition and Importance

Policy lifecycle management involves overseeing the various stages of a policy's life, from inception to termination. Effective management is crucial for ensuring operational efficiency and customer satisfaction in the insurance sector.

2.2.2 Challenges in Policy Management

Managing policy lapses remains a significant challenge due to the reactive nature of traditional methods. These methods, such as sending reminder notices and direct contact, are often inefficient and costly.

2.2.3 Objectives Met

- **Literature Review Conducted**: A comprehensive review of existing literature on policy lifecycle management has been completed, identifying key challenges and gaps in current practices.

2.3 Policy Lapses

2.3.1 Understanding Policy Lapses

Policy lapses occur when policyholders discontinue premium payments, leading to the termination of their policies. This disrupts coverage and poses financial and operational challenges for insurance companies.

2.3.2 Traditional Methods of Addressing Policy Lapses

Conventional methods to manage lapses focus on reactive measures, which are often insufficient in preventing lapses.

2.3.3 Objectives Met

- **Problem Identification**: Clearly defined the issue of policy lapses and the inadequacies of traditional methods in managing them.
- **Data Collection**: Collected relevant data from Kaggle, covering various features such as entry age, sex, policy type, payment mode, and policy status.'

2.4 Predictive Analytics in Policy Management

2.4.1 Overview of Predictive Analytics

Predictive analytics involves using historical data to forecast future events. It offers a proactive approach to managing policy lapses by identifying patterns and predicting potential lapses before they occur.

2.4.2 Applications of Predictive Analytics

Predictive analytics has been successfully applied in various industries, including insurance, to address issues similar to policy lapses.

2.4.3 Objectives Met

- **Exploratory Data Analysis (EDA)**: Conducted EDA to understand the data structure, identify patterns, and detect any anomalies.
- **Initial Findings**: Identified key patterns in the data that could potentially predict policy lapses.

2.5 Machine Learning for Predictive Modeling

2.5.1 Introduction to Machine Learning

Machine learning involves training algorithms to identify patterns and make predictions based on data. It includes supervised and unsupervised learning techniques.

2.5.2 Machine Learning Algorithms

Common algorithms used in predictive modeling include decision trees, random forests, support vector machines, and neural networks. Each algorithm has its strengths and weaknesses.In this study, three machine learning algorithms have been selected for exploration due to their popularity and proven effectiveness in predictive modeling tasks. These algorithms are Decision Trees, Random Forests, and Support Vector Machines (SVMs).

- **Decision Trees**

  - **Description**: Decision trees are a type of supervised learning algorithm used for classification and regression tasks. They split the data into subsets based on the value of input features.
  - **Strengths**: Easy to interpret, visualize, and understand; can handle both numerical and categorical data.
  - **Weaknesses**: Prone to overfitting, especially with complex datasets.

- **Random Forests**

  - **Description**: Random forests are an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes or mean prediction of the individual trees.
  - **Strengths**: Reduces overfitting by averaging multiple trees; robust to noisy data and outliers.
  - **Weaknesses**: Can be computationally intensive and less interpretable than single decision trees.

- **Support Vector Machines (SVMs)**

  - **Description**: SVMs are supervised learning models that analyze data for classification and regression analysis by finding the hyperplane that best separates the data into classes.
  - **Strengths**: Effective in high-dimensional spaces; works well with clear margin of separation.
  - **Weaknesses**: Not suitable for large datasets due to high computational cost; less effective with noisy data.

2.5.3 Objectives Met

- **Algorithm Exploration**: Investigated the application of decision trees, random forests, and SVMs for predicting policy lapses.
- **Initial Findings**: Identified initial performance metrics indicating the strengths and weaknesses of each algorithm in the context of policy lapse prediction.

2.6 Key Predictors of Policy Lapses

2.6.1 Identifying Predictors

Previous research has identified several predictors of policy lapses, including demographic factors, policy-specific variables, and behavioral indicators.

2.6.2 Data Sources and Data Quality

The quality of data is crucial for building effective predictive models. Data from Kaggle has been preprocessed to ensure accuracy and reliability.

2.6.3 Objectives Met

- **Key Predictors Identification**: Analyzed the dataset to determine which factors significantly influence the likelihood of a policy lapse.
- **Data Preprocessing**: Completed initial data preprocessing steps, including cleaning and transforming the data.

2.7 Gaps in Existing Research

2.7.1 Identifying Research Gaps

Despite advancements in predictive modeling, gaps remain in effectively utilizing machine learning for policy lapse prediction.

2.7.2 Addressing Research Gaps

This study aims to fill these gaps by employing a comprehensive dataset and exploring various machine learning techniques.

2.7.3 Objectives Met

- **Research Gap Identification**: Identified specific gaps in current research that this study aims to address.
- **Proactive Approach**: Proposed a proactive approach to policy lapse management using predictive modeling.

2.8 Summary

This chapter reviewed the literature on policy lifecycle management, policy lapses, predictive analytics, and machine learning, emphasizing the objectives met so far. Significant progress has been made in understanding the problem, collecting and analyzing data, and exploring initial predictive models. The next chapter will delve deeper into the methodology and present detailed findings from the research conducted to date.

Chapter 3

Dataset, Data Acquisition, and Data Exploration

3.1 Introduction

This chapter provides a detailed overview of the dataset used in this study, the process of data acquisition, and the results of data exploration. It outlines the characteristics of the data, the steps taken to prepare it for analysis, and the initial insights gained through exploratory data analysis (EDA).

3.2 Dataset Description

3.2.1 Source of Data

The dataset used in this study was obtained from Kaggle, a well-known platform for data science and machine learning competitions. The specific dataset chosen is related to insurance policies and includes a variety of features that are relevant for predicting policy lapses.

3.2.2 Features and Attributes

The dataset comprises several features that provide information about the policyholders and the policies themselves. Key attributes include:

1. **CHANNEL1**: Represents the primary sales or distribution channel through which the policy was sold (e.g., direct, broker, online).
2. **CHANNEL2**: Represents a secondary sales or distribution channel, if applicable
3. **CHANNEL3**: Represents a tertiary sales or distribution channel, if applicable.
4. **ENTRY AGE**: The age of the policyholder at the time the policy was issued.
5. **SEX**: The gender of the policyholder.
6. **POLICY TYPE 1**: The first type or category of the insurance policy (e.g., term, whole life).
7. **POLICY TYPE 2**: The second type or category of the insurance policy, if applicable.
8. **POLICY TYPE 3**: The third type or category of the insurance policy, if applicable.
9. **PAYMENT MODE**: The frequency of premium payments (e.g., monthly, quarterly, annually).
10. **POLICY STATUS**: The current status of the policy (e.g., active, lapsed, terminated).
11. **BENEFIT**: The amount of benefit or coverage provided by the policy.
12. **NON LAPSE GUARANTEED**: Indicates whether the policy has a non-lapse guarantee feature, ensuring it doesn't lapse even if premium payments are missed, under certain conditions.
13. **SUBSTANDARD RISK**: Indicates if the policyholder is considered substandard risk (higher risk than standard due to health or other factors).
14. **NUMBER OF ADVANCE PREMIUM**: The number of premium payments made in advance.
15. **INITIAL BENEFIT**: The initial amount of benefit or coverage when the policy was issued.

16. **Full Benefit?**: Indicates whether the policy is providing the full benefit amount (yes/no).
17. **Policy Year (Decimal)**: The policy year represented in decimal format (e.g., 2.5 years).
18. **Policy Year**: The policy year as an integer (e.g., 2 years).
19. **Premium**: The amount of premium paid by the policyholder.
20. **Issue Date**: The date when the policy was issued.
21. **Unnamed: 20**: An unnamed or potentially irrelevant column that might need to be dropped or renamed.
22. **Unnamed: 21**: Another unnamed or potentially irrelevant column that might need to be dropped or renamed.

3.3 Data Acquisition

3.3.1 Data Collection Process

The data was collected from Kaggle, which allows for the downloading of datasets in a structured format. The dataset was provided as a CSV file, which is suitable for data analysis using various tools and programming languages.

3.3.2 Data Import and Initial Inspection

The dataset was imported into Python using the Pandas library. Initial inspection involved loading the data into a DataFrame and performing basic checks to understand its structure, such as:

- **Shape of the Data**: Number of rows: 185560, Number of columns: 22

- **Data Types**: Column name and their data types are given in the below Table 1.

Column Name	Data Type
CHANNEL1	int64
CHANNEL2	int64
CHANNEL3	int64
ENTRY AGE	int64
SEX	object
POLICY TYPE 1	int64
POLICY TYPE 2	int64
POLICY TYPE 3	object
PAYMENT MODE	object
POLICY STATUS	object
BENEFIT	object
NON LAPSE GUARANTEED	object
SUBSTANDARD RISK	float64
NUMBER OF ADVANCE PREMIUM	int64
INITIAL BENEFIT	float64
Full Benefit?	object
Policy Year (Decimal)	float64
Policy Year	int64
Premium	object
Issue Date	object
Unnamed: 20	float64
Unnamed: 21	float64

Table 1: Column Names and their Data Types

3.4 Data Cleaning and Preprocessing

3.4.1 Handling Missing Values

Two Columns i.e., Unnamed: 20 and Unnamed: 21 had 185560 values missing.We have dropped this two columns as it seemed to be a potentially irrelevant column.

3.4.2 Data Transformation

Two Columns Sex and Full Benefit were converted from Categorical Values to Numerical Values. Columns like Payment Mode and Non Lapse Guaranteed were encoded using one-hot encoding to convert them into a format suitable for machine learning algorithms.

3.4.3 Correlation Analysis

Heatmap of correlations between features to identify potential relationships. Key findings from the heatmap are:

- PAYMENT MODE_Monthly has a moderate positive correlation with is_lapse. This indicates that policies paid monthly might have a higher likelihood of lapsing.
- PAYMENT MODE_Quarterly and PAYMENT MODE_Semiannually have weak negative correlations with is_lapse, suggesting these payment modes might be less prone to lapsing compared to monthly payments.
- PAYMENT MODE_Single Premium has a weak negative correlation with is_lapse, indicating that single premium payments might also be less likely to lapse.
- NON LAPSE GUARANTEED_NLG Active has a strong negative correlation with is_lapse. This indicates that policies with active non-lapse guarantees are much less likely to lapse.
- Both Policy Year and Policy Year (Decimal) have moderate positive correlations with is_lapse. This suggests that as the policy year increases, the likelihood of lapsing might increase.
- ENTRY AGE, SEX, and SUBSTANDARD RISK have very weak correlations with is_lapse, suggesting these variables might have minimal impact on predicting lapses.

In summary, the variables with the most significant correlations to is_lapse are PAYMENT MODE_Monthly, NON LAPSE GUARANTEED_NLG Active, Policy Year, and Policy Year (Decimal).

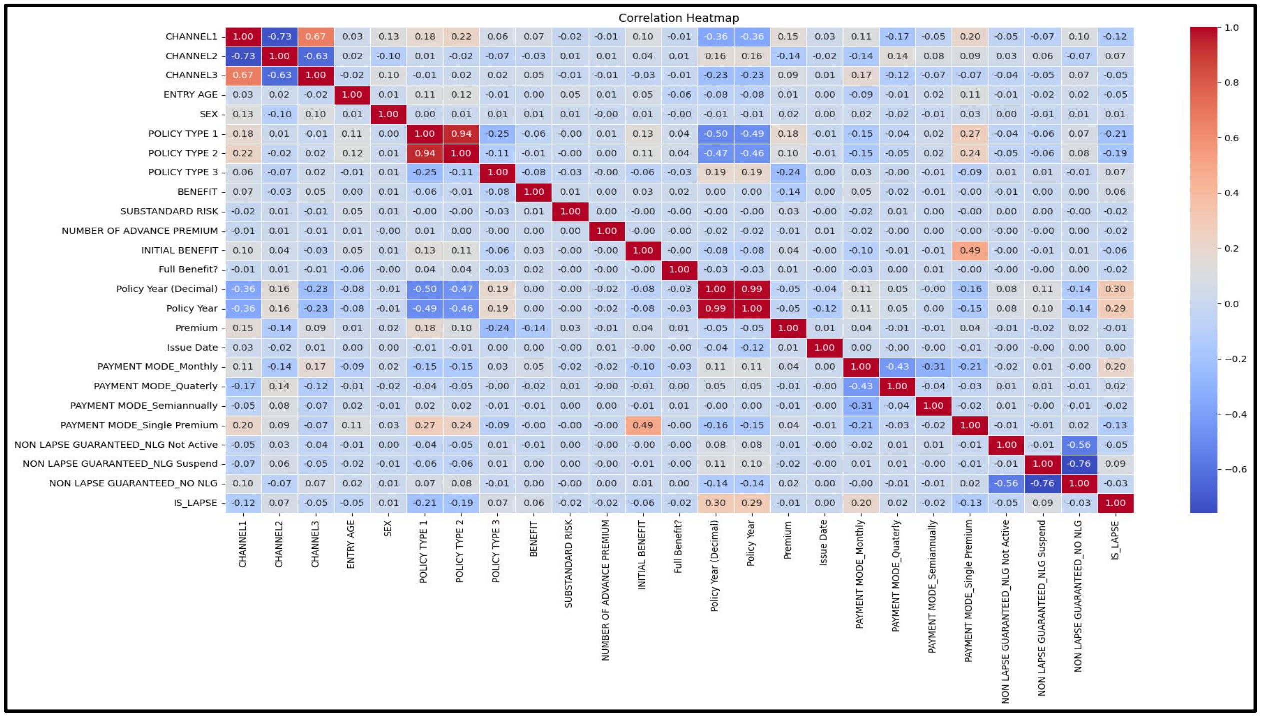

Figure 1: Correlation Heatmap for Predicting Policy Lapse (is_lapse)

3.4.4 Handling Class Imbalance

Class imbalance is a common issue in machine learning, especially in classification problems. In our dataset, the target variable is_lapse exhibits a significant imbalance, with 98,865 instances of policies that lapsed and 86,695 instances of policies that did not lapse. This imbalance can bias the model towards the majority class, leading to sub-optimal performance.

To understand the extent of the imbalance, we plotted the class distribution of is_lapse. The bar chart below (Figure 2) illustrates the original class distribution:

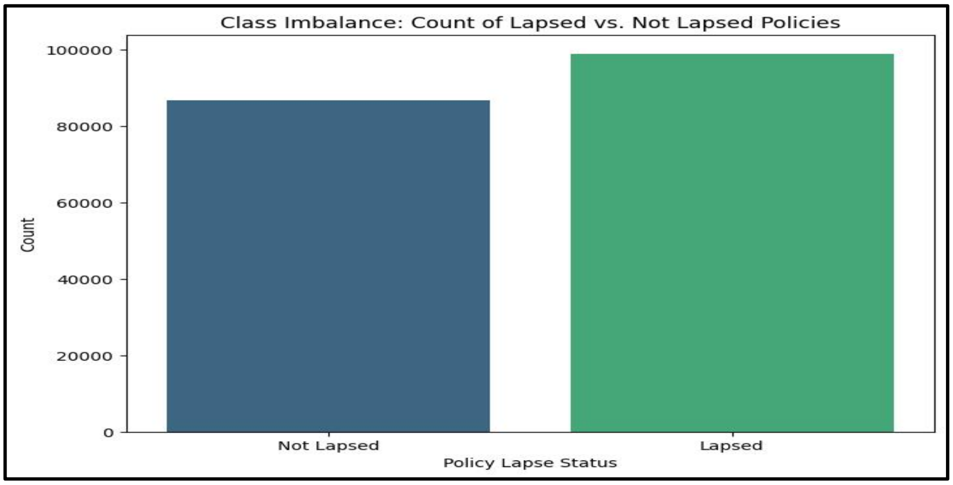

Figure 2: Original Class Distribution

As shown in Figure 2, the number of policies that lapsed is higher than those that did not, indicating a class imbalance that needs to be addressed.

3.4.5 Techniques for Handling Class Imbalance

We employed Synthetic Minority Over-sampling Technique (SMOTE) to address this imbalance. SMOTE is an oversampling technique that generates synthetic samples for the minority class (policies that did not lapse) to balance the dataset. This technique helps in creating a more balanced distribution without simply duplicating existing samples, which could lead to overfitting.

3.4.6 Implementation

The following steps were taken to apply SMOTE:

1. **Data Splitting:** The dataset was split into training and testing sets to ensure that the model is evaluated on unseen data.
2. **Applying SMOTE:** The SMOTE algorithm was applied to the training set to oversample the minority class. The `sampling_strategy='auto'` parameter was used to balance the minority class to the same number as the majority class.

After applying SMOTE, the class distribution became balanced, as shown in Figure 3 below:

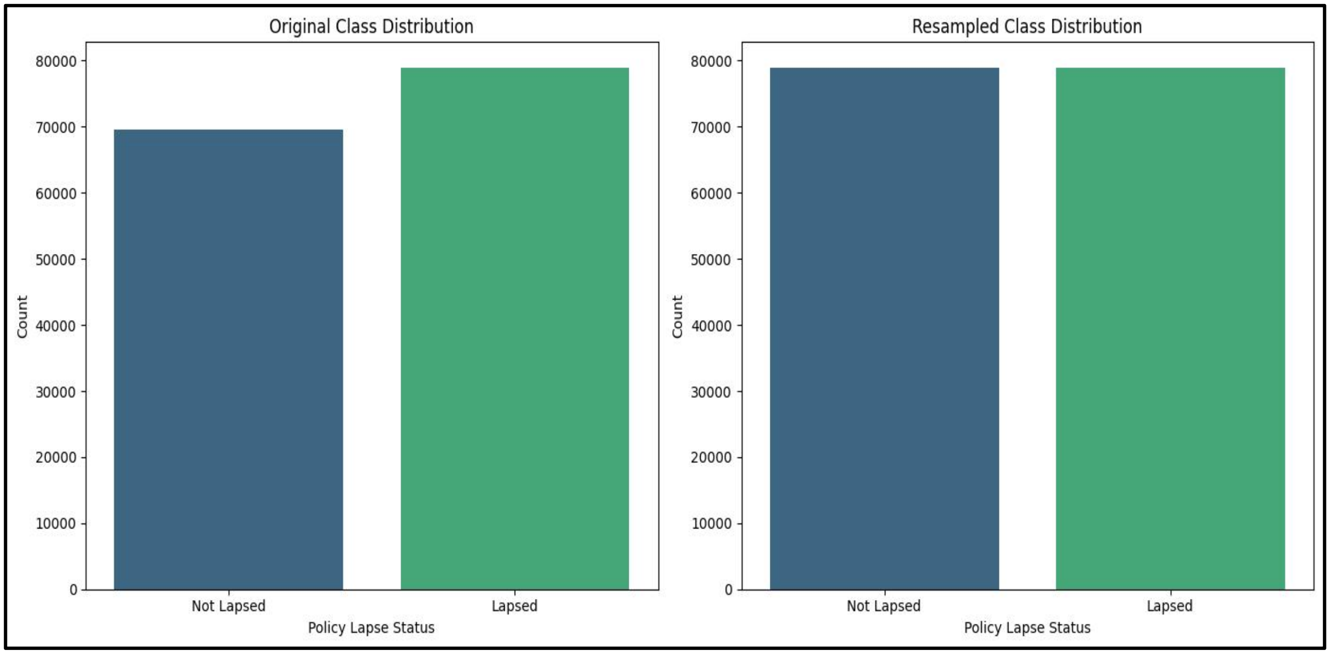

Figure 3: Resampled Class Distribution

This balanced dataset will be used to train our classification model. This approach helps in mitigating the bias towards the majority class and improves the model's ability to predict the minority class accurately.

3.5 Summary

This chapter provided a comprehensive overview of the dataset, the methods used for cleaning and preprocessing, the initial insights from exploratory data analysis and details regarding handling class imbalance. The next chapters will focus on the modeling techniques and their evaluation.

Directions for future work after mid semester

As we progress beyond the mid-semester, my focus will shift from exploratory data analysis (EDA) to the application and evaluation of predictive algorithms. The initial phase involved a comprehensive understanding of the dataset, including determining the number of rows and columns, examining the influence of various parameters on the target variable, and exploring several predictive algorithms. The future directions are as follows:

1. **Algorithm Implementation and Model Training:**

- **Selection of Algorithms**: We will apply three predictive algorithms to our dataset:

    1. Logistic Regression

    2. Random Forest

    3. Gradient Boosting

- **Model Training**: Train the selected algorithms on the training dataset. We will utilize cross-validation techniques to ensure the models are not overfitting and to get a reliable estimate of their performance.

- Evaluation Metrics: Use evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to compare the performance of the models.

- **Model Evaluation:**

- **Training and Validation Spli**t: Split the data into training and validation sets to evaluate the models. Ensure that the class imbalance is handled appropriately during this split.

- **Performance Comparison:** Compare the performance of the models using the evaluation metrics. Identify the model that performs the best on the validation dataset.

2. **Reporting and Documentation**:

- **Comprehensive Report**: Document the process, including the data preprocessing steps, model training, hyperparameter tuning, and evaluation. Summarize the findings and provide insights into the performance of each model.

- **Future Recommendations:** Based on the findings, provide recommendations for further improvement and potential areas of research.

The future work will focus on the practical application of machine learning algorithms to predict policy lapses. By systematically training, tuning, and evaluating the models, we aim to identify the best-performing algorithm. The end goal is to develop a reliable and accurate model that can be deployed for real-world use, providing valuable insights and predictions for policy management.

Bibliography / References

The following are referred journals from the preliminary literature review.

1. Barsotti, F., Milhaud, X. & Salhi, Y. (2016). Lapse risk in life insurance: Correlation and contagion effects among policyholders' behaviors.

2. Ćurak, M., Podrug, D. & Poposki, K. (2015). Policyholder and Insurance Policy Features as Determinants of Life Insurance Lapse - Evidence from Croatia