KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

PREDICTIVE MODELING OF SURRENDER RATE IN LIFE INSURANCE

Acheampong Kyei Rabbi

Amlalo Barnabas Doe

Ibrahim Hakim

Dagbey Thea

(BSc. Actuarial Science)

In pursuit of the BSc degree in Actuarial Science, this thesis is submitted to the Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology.

AUGUST, 2023

Declaration

We certify that this paper is our own unique work for the BSc in Actuarial Science. To the best of our knowledge, it doesn't contain any material that has already been published by another person or that has been used to grant a degree from another university. Due credit has been attributed to all previously published works and online publications through proper referencing.

signature .... Date ....

Acheampong Kyei Rabbi

signature .... Date ....

Amlalo Barnabas Doe

signature .... Date ....

Ibrahim Hakim

signature .... Date ....

Dagbey Thea

This thesis has been submitted, and we, as university supervisors, affirm our approval of its submission.

signature .... Date ....

Prof. Nana Kena Frempong

Supervisor

signature .... Date ....

Prof. A. O. Adebanji

Head Of Department

Dedication

We dedicate this project to the Almighty God, who has guided and supported us through all the challenges we faced during this thesis period and led us to this point. Additionally, we dedicate this endeavor to our families and friends, whose unwavering love and prayers have accompanied us every step of the journey. Lastly, we extend our dedication to our supervisor, whose unwavering support and guidance have been invaluable throughout this process.

Acknowledgement

We express our deep gratitude to the Almighty God for granting us the strength and perseverance to undertake this task. Our heartfelt appreciation also goes to Prof. Nana Kena Frempong, our supervisor, who served as a guiding light as we embarked on this journey. His kindness, understanding, and patience will always be cherished. He provided valuable mentorship and guidance, correcting us when needed. We also want extend our thanks to all the instructors at Kwame Nkrumah University of Science and Technology (KNUST) who supported us in our studies. To everyone involved, we are truly grateful.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ML:	Machine learning
AI:	Artificial Intelligence
SML:	Supervised Machine Learning.
RF:	Random Forest
DT:	Decision Tree
NN:	Neural Network
NBC:	Naïve Bayes Classifier
CART:	Classification and Regression Trees
SVM:	Support Vector Machine
LR:	Logistic Regression
AUROC:	Area Under the Receiver Operating Characteristics
PCA:	Principal Component Analysis
EDA:	Exploratory Data Analysis
ROC:	Receiver Operating Characteristics
TP:	True Positive
FP:	False Positive
FN:	False Negative
TN:	True Negative
MSE:	Mean Square Error

Abstract

A high surrender rate among policyholders is a significant problem facing insurers. Most policyholders surrender their life insurance policies for various reasons, such as financial difficulties, changing life circumstances, or dissatisfaction with the policy. Policy surrender affects both the policyholder and the insurance company negatively. Some of the effects include; loss of revenue, reduced customer retention, negative impact on reputation, and loss of potential benefit on the side of the insured. The main objective of this study is to identify the key factors, at-risk policyholders, and accurately estimate the probability of policyholders surrendering their life insurance policies based on our data using Machine Learning algorithms. A dataset from Kaggle, an open online data science platform, was used for the study. The dataset was divided into 70% for training and 30% for testing. Standardization was done to ensure that the values were within the specified range of 0 and 1. The three supervised machine learning algorithms used for classifying policy status (surrender, in-force, and lapse) were Decision Tree, Random Forest, and Support Vector Machines. A confusion matrix and algorithm performance metrics including recall, Mean Square Error (MSE), Accuracy, Precision, and F1 Score were used to assess the predictive performance of the model. Based on these evaluation metrics, the Random Forest algorithm emerged as the most effective model in terms of accuracy and overall performance, thus it had the highest value for Recall, Accuracy, F1 Score, and the smallest Mean Square Error. The best model, Random Forest had 68.25% accuracy, 68.25% recall, 68.93% precision, and 64.81% F1 Score. On the other hand, the Support Vector Machine and Decision Tree had an accuracy of 68.06% and 67.74% respectively.

Chapter 1

Introduction

This chapter provides an overview of the life insurance industry, surrender rates, predictive modeling, Machine Learning (ML), the problem statement, the study's objectives, its significance, a general overview of recent life insurance implementations, and the structure of the remaining chapters.

1.1 Background Of Study

A life insurance policy is an agreement between an insured and an insurer whereby the insurer agrees to pay a sum of money in exchange for a premium upon the death of an insured person or after a certain period. The life insurance sector is crucial in helping policyholders and their families maintain financial stability, Kgare (2021). Depending on the terms of the agreement, payment may also be provided for other situations like severe illness or terminal disease. The policyholder usually pays a premium, which could be a one-time or regular payment. The policy may also cover other expenses such as burial costs. The contract specifies the events that are excluded from coverage under the law. The insurer's responsibility may be restricted by the insurance contract for certain exceptions like suicide, war, and fraud. Additionally, uncertainties associated with an incident may lead to complications. For example, if the insured was aware of the risks involved in an experimental medical procedure or taking medication that could cause harm or death.

Today's insurance plans offer policyholders a wide range of options, which can have a major impact on the insurer's exposure to liability. For instance, It is possible for policyholders to receive a surrender value by surrendering their policy or to stop paying premiums and let the policy lapse. Therefore, the life insurance industry must have a better understanding of the factors that can influence the surrender rate of the policy, Martin et al., (2013). Surrender rates refer to the rate at which policyholders decide to terminate their insurance policy, which can severely affect the overall profits of an insurer. Many studies have been conducted to understand the factors that contribute to the surrender rate and identify strategies for reducing it. The insurance industry's numerous aspects, including life insurance, health insurance, and casualty insurance, have been the subject of these studies. These studies frequently come to the following conclusions about the elements influencing policy surrender: Milhaud et al., (2010); Eling et al., (2014).

i. Policy features: Policyholders are more likely to surrender their policies if they find them to be inflexible, expensive, or poorly suited to their needs.

ii. Customer service: Policyholders are more likely to stay with their insurer if they have a positive experience with customer service representatives, including timely and accurate responses to inquiries and complaints.

iii. Economic conditions: Economic factors such as fluctuations in interest rates, inflation, and the performance stock market can influence the surrender rate as they can impact the financial situation and expectations of policyholders.

iv. Marketing and sales practices: Some insurers have been found to use deceptive marketing and sales practices that can lead to higher surrender rates.

To forecast the likelihood of policy surrender, and the factors that result in it, there is a need for predictive modeling. Predictive modeling refers to the process of analyzing a large dataset to draw conclusions or identify significant relationships and using these relationships for forecasting future occurrences, (Stehno et al., 2018). It uses statistical algorithms and Machine Learning (ML) to make predictions for the future. It has been growing in popularity in various fields, including finance, healthcare, marketing, and manufacturing, (Cranmer et al., 2017).

In insurance, predictive modeling is used to assess risk, determine premiums, and make underwriting decisions. It involves analyzing large datasets to identify patterns and correlations between various factors and insurance outcomes, such as claims frequency and severity, (Staudt et al., 2010). This information is then used to develop models that can predict future outcomes and inform pricing and underwriting decisions. In addition, predictive modeling is an essential method that can help identify the factors that are most correlated with increased surrender rates of life insurance policies and to also forecast customer surrender rates, and identify customers who are likely to surrender their policies,(Babaoglu et al., 2017). Surrendering a policy means that the policyholder cancels their coverage and receives the cash value of their policy before the maturity date. The models are based on the analysis of existing data and trends in the life insurance industry. A variety of variables such as policyholder age, gender, policy type, premium amount, and policy duration may be used by the model to estimate the likelihood of surrender. It may also consider macroeconomic factors like interest rates, market trends, and demographic shifts.

By leveraging predictive modeling, life insurance companies can better understand and anticipate the risks of policyholders surrendering their policies and use this information to design better investment options and improve customer service, (Tiwari et al., 2010). Understanding the policies that are more likely to be surrendered, insurers can adjust their pricing and marketing strategies to minimize the risk of policyholders canceling their policies prematurely. Predictive models can be developed using a variety of statistical and machine learning techniques, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVMs), and Neural Networks (NN). Historical data can be utilized to train models which can identify patterns and correlations. These can then be utilized to predict surrender rates in the future and formulate specific strategies to decrease the number of customers surrendering their policies, (Sen Hu et al., 2021; Loisel et al., 2021). This paper will discuss the application of predictive models to the life insurance industry, and the benefits of this approach for both policyholders and insurance providers.

1.2 Problem Statement

A high surrender rate among policyholders is a significant problem facing insurers and the insurance industry at large. Most policyholders surrender their life insurance policies for various reasons, such as financial difficulties, changing life circumstances, or dissatisfaction with the policy. However, policy surrender affects both the policyholder and the insurance company negatively (Russell et al.,2013). On the side of the insurer, there are;

- Lost revenue: Insurers rely on premiums paid by policyholders to generate revenue. When policyholders surrender their policies, the insurer loses the opportunity to collect future premiums, which can have a significant impact on its financial performance.

- Increased expenses: Surrendering a policy can be costly for insurers, as they may need to pay out surrender charges or penalties to policyholders. In addition, the process of replacing lost policies can be time-consuming and expensive.

- A negative impact on reputation: High surrender rates can damage the reputation of an insurer, as it can be seen as a sign of poor customer service, inflexible policies, or ineffective marketing and sales practices.

- Underwriting risk: When policyholders surrender their policies, it can create uncertainty and risk for insurers in terms of their underwriting practices. Surrender rates can be influenced by economic conditions, market trends, and other factors that can be difficult to predict.

- Customer retention: High surrender rates can indicate a lack of customer loyalty and satisfaction, which can be problematic for insurers looking to build long-term relationships with policyholders.

Some negative consequences on the side of policyholders surrendering their policies are;

- Loss of potential benefit: Policyholders may lose the potential benefits of the policy such as death benefits, long-term care benefits, or retirement income when they surrender the insurance policy before it matures.

- Limited options: In situations where policyholders encounter financial difficulties or find themselves no longer needing coverage, surrendering a policy may be the only option. In such cases, surrender charges or penalties can further reduce the amount of money that the policyholder can receive.
- Reduced coverage: Surrendering a policy can result in reduced or no coverage for the policyholder, which can leave them vulnerable to financial loss in the event of an unexpected event.
- A negative impact on credit score: Surrendering a policy can hurt the policyholder's credit score, as it can indicate a lack of financial stability or responsibility.

To address this challenge, there is a need for insurance companies to develop a model that can accurately predict the likelihood of policyholders surrendering their life insurance policies before their maturity date, (Martin et al., 2013). The model would need to be based on relevant data such as policyholder demographics, policy details, and economic factors, and would need to be able to identify key factors that contribute to policy surrender.

1.3 Objectives of the study

This study aims to identify and forecast surrender rates in life insurance using machine learning. The objectives of our study are as follows;

i. Accurately estimate the probability of policyholder surrender in life insurance policies based on our data using Machine Learning algorithms.

ii. Identify the key features that contribute to policy surrender.

iii. To identify at-risk policyholders who are likely to surrender their policies.

1.4 Research Questions

Some research questions associated with our study are as follows;

- What factors most accurately predict a policyholder's choice to surrender their life insurance policy?
- How accurately can surrender rates be predicted using machine learning algorithms?
- Can insurers identify at-risk policyholders and act before they surrender their policies using predictive modeling of surrender rates?

1.5 Significance of the study

This study will provide more insight into the need for predictive modeling using machine learning algorithms and the key variables that contribute to policy surrender. The ability for insurers to identify high-risk policyholders, the main factors associated with life policy surrender, and develop predictive models of surrender rates in life insurance is beneficial to insurers in so many ways. Some of the key benefits are as follows;

Improved decision-making: Predictive modeling can help insurers make better decisions about pricing, underwriting, and policy design. By analyzing historical data on surrender rates, insurers can identify patterns and trends that can inform their decision-making and help them better anticipate future surrender rates.

Better risk assessment: Predictive modeling can also help insurers assess the risk of policies more accurately. By analyzing factors that influence surrender rates, such as policy features, demographics, health status, and other factors to identify policyholders who are at a higher risk of surrendering their policies, insurers can take proactive measures including targeted marketing campaigns, personalized offers, and incentives, educating policyholders, offering flexible policies and improved customer service and support to retain those customers and minimize financial losses. Insurers can also base on the data to develop models that can predict the policies that are more likely to be surrendered and adjust their underwriting practices accordingly.

Increased profitability: By improving decision-making and risk assessment, predictive modeling can help insurers increase profitability. Insurers can use models to identify policies that are more likely to be profitable and adjust pricing and underwriting practices accordingly.

Enhanced customer experience: Predictive modeling can also help insurers improve the customer experience by developing more personalized policies and marketing strategies. Insurers can enhance customer satisfaction and retention rates by examining the factors that impact surrender rates and customizing policies accordingly to suit the preferences and requirements of policyholders.

1.6 Methodology

Machine learning can be used in several ways to analyze and forecast surrender rates. In our analysis, we'll make use of methods like Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVMs). Python 3.8.8, a program for machine learning, will also be used. We would employ statistical characteristics such as the mean, variance, correlation matrix, total variance, standard deviation, proportion of variance, and cumulative proportion to analyze and predict the surrender rate. Our dataset would be scaled, its dimension would be reduced with PCA, and predictions would be made using Random Forest (RF), the Decision Tree (DT), and Support Vector Machines (SVMs), and their accuracy would be compared. Additionally, we would gather data on surrender rates and machine-learning techniques from websites and research papers.

1.7 Limitations

1. The data used for this study was obtained from the Kaggle platform, where fewer descriptions of dataset features were provided. This limited our understanding of the features and therefore, limited analysis and conclusion drawn on the features.

2. The main focus of the study was to ascertain whether ML techniques can be

employed to predict surrender, due to this further techniques to improve the model's predictive performance and accuracy were not performed.

3. The target variable illustrated imbalances in its composition and very few analysis was presented in this study to determine the effect of such imbalances on the model's predictive power.

1.8 Organization of the study

There are five chapters in this thesis. The first chapter provides an overview and a brief history of the subject being studied. It addresses the study's problem statement, significance, goals, research questions, methodology, and definitions. The general literature review of the theories is included in Chapter Two. This chapter's main focus is on the work of researchers who used machine learning techniques to develop a model for forecasting surrender rates. The methodology of the study, mathematical and statistical estimations of the principal components, total variance, standard deviation, proportion of variance, and cumulative proportion of the surrender rate data are all covered in Chapter Three. The data gathered and an analysis of the findings are presented in Chapter Four along with data on the surrender rate. Chapter Five, the concluding chapter, examines the thesis' conclusions and recommendations.

Chapter 2

Literature Review

2.1 Introduction

Many researchers have conducted studies on the problems associated with life insurance policies, consequences, and detection of more nuance problems in life insurance since the profitability of life insurance is on the decline in recent years (McKinsey Global insurance pools, 2021). Several Machine Learning (ML) algorithms which include Neural Networks (NN), Logistic Regression (LR), Support Vector Machines (SVMs), Classification and Regression Trees (CART), Decision Trees (DT), Random Forest (RF), and others have been employed by different researchers to develop a predictive model on surrender rate and lapses. In this chapter, a literature review is made on Predicting surrender rates in life insurance policies. This chapter focuses on understanding the difference between lapse and surrender, the consequences of the associated problems in life insurance, and related works on using Machine learning algorithms to analyze lapse and surrender rates. The study will emphasize much on surrender rates and the need for predictive modeling using Machine Learning (ML) algorithms in life insurance.

2.2 Lapse and Surrender

The term **"lapse"** in the context of life insurance refers to the discontinuation of an insurance contract and the loss of benefits when the policyholder stops paying premiums (Eling & Kochanski, 2013). Lapsing is generally not a voluntary decision by the policyholder, but rather a result of missed payments. Policy lapse has several reasons (Russell et al., 2013). These include; the policyholder for some reason becoming unemployed and unable to afford to pay premiums, the dynamics of contracts, a firm's reputation, competition, and regulations, as well as economic risk factors like tax breaks and inflation, which may also have an impact on policy lapses (Barsotti et al., 2016).

Surrender in life insurance refers to a situation where the policyholder decides to terminate the policy and receive the policy's surrender value before the maturity date. The surrender value is the amount of money that the insurer will pay the policyholder upon surrendering the policy, which may be less than the total premiums paid, depending on the policy's terms and conditions. Usually, it is the policyholder who voluntarily surrenders the policy.

The main difference between surrender and lapse is that surrender is a voluntary decision by the policyholder to terminate the policy and receive the surrender value, while lapse occurs when the policyholder stops making premium payments and the policy terminates due to non-payment without any surrender value paid to the policyholder.

2.3 Related works on using Machine learning algorithms to analyze lapse and surrender rates

2.3.1 Introduction

Surrender rates are an important metric for insurance companies, as they provide insight into the financial health of the company. Predictive modeling for surrender rates in life insurance has gained increasing attention in recent years, with many researchers exploring different methods and techniques to develop accurate and reliable models. In this literature review, we will examine some of the key studies on predictive modeling for surrender rates in life insurance.

2.3.2 Diagnosing with unsupervised and supervised algorithms

An article by (Milhaud et al., 2010) presented a study of the surrender triggers in life insurance policies. The authors classified these triggers into different categories, such as financial or personal reasons, and examine their impact on the likelihood of policyholders surrendering their policies. They used statistical techniques, including logistic regression and decision trees, to analyze the data and predict the risk of surrender. They compared the predictive performance of these methods using standard measures such as the Area Under the Receiver Operating Characteristic curve (AUROC) and misclassification rates. The results suggest that the predictive performance of the different methods was relatively similar, with no method consistently outperforming the others. The authors found that the main factors influencing surrender behavior were financial, such as the level of premiums, interest rates, and the policy's cash value. They also found that personal reasons, such as health problems or changes in marital status, can also have a significant impact on the likelihood of surrender. The study enhances the comprehension of surrender patterns in life insurance policies, and it has significant implications for insurers in terms of recognizing and handling the risk of surrender.

In a study by (Tiwari et al., 2010), the authors present a new method and structure for anticipating customer churn by employing Neural Network. In their research, they distinguish the clients who are prone to churning in the future, contrasting with prior studies that centered on identifying customers who are likely to churn immediately.

(Russell et al., 2013) present an empirical analysis of the surrender activity of life insurance policies. The study is based on an empirical analysis of data from a sample of life insurance policies. The study uses statistical methods to analyze data from a sample of 54,000 policies and investigates the factors that influence policyholders' decisions to surrender their policies. The authors find that surrender rates are highest in the first few years of a policy's life and decline as the policy ages. Additionally, the authors discovered that policyholders are prone to surrendering their policies when they experience financial difficulties and when interest rates are low. Ultimately, the authors analyze the significance of these findings for insurers and recommend that insurers closely monitor policy surrenders as a precursor to possible financial difficulties among policyholders.

(Babaoglu et al., 2017) also developed a predictive model for lapse risk. The research compared Logistic Regression (LR), Random Forest (RF), and the Naïve Bayes Classifier (NBC) algorithm. The goal was to manage the insurance business using actuarial analysis and to identify the variables that affect lapse rates. The results highlighted that Random

Forest was most effective as compared to logistic regression and naive Bayes.

(Llave et al., 2019) explore how geographical factors affect the surrender prediction for an insurance company located in the urban area of Madrid. The authors used data on policyholder characteristics, premiums, and claims, as well as information on the location of policyholders' residences, to develop a predictive model for policy churn. They found that geographical factors, such as the distance between a policyholder's residence and the insurance company's office, have a significant impact on policy churn. According to the research, a customer's likelihood to churn increases if neighboring consumers churn as well.

(Loisel et al., 2021) utilized two machine-learning techniques, namely Support Vector Machine (SVM) and, Extreme Gradient Boosting (XGBoost) to develop lapse models. They compared their performance with two conventional statistical methods based on statistical accuracy and profitability measures. The authors also adopted a fresh perspective on the surrender prediction issue by transforming the classification problem into a regression query. They then conducted optimization, which is a novel lapse and surrender risk management approach. The outcomes of the four methods indicated that XGBoost and SVM outperformed Classification and Regression Tree (CART) and Logistic Regression (LR) in terms of statistical accuracy. However, LR was just as effective as XGBoost in terms of retention gains.

Another research paper (Kgare, 2021) predicted Lapse Rate using Machine Learning Algorithms in life insurance. This study compared nine machine learning classifier models with traditional algorithms for life insurance lapse prediction, using two different insurer datasets with different feature selection methods. The ensemble models (Gradient Boost and Random Forest) showed stronger prediction abilities than single classifiers, and parameter tuning and model boosting improved performance. Gradient Boosting was the best overall classifier. It had an accuracy of 92%, 76%, F-measure of 92%, and 84% for the two datasets. The study recommends using ensemble models for life insurance lapse prediction as they are more effective than single classifiers.

Research by (Michele et al., 2021), conducted a study that explored the prediction of policy lapses in life insurance using ML algorithms. They focused on the application of the Random Forest (RF) methodology to forecast whether policyholders will lapse on their life insurance contracts. The authors discovered that the Random Forest methodology outperforms logistic models, even when accounting for feature interactions. The study found that non-economic aspects, such as the length of time since the contract's inception, time to expiry, the insurance company, and its marketing approach, significantly impact the lapse decision. In contrast, economic or financial features have a limited effect, except for the growth rate of disposable income.

In a study published in the Annals of Actuarial Science by (Sen Hu et al., 2021), The researchers created a Machine Learning model that combines traditional methods like Logistic Regression and Decision Trees with spatial analysis to predict the likelihood of a policy lapse. Their goal was to determine if using spatial analysis would improve the accuracy of predicting lapses. They used data on policyholder attributes, premiums, claims, and location to build the model, which outperformed non-spatial models in predicting policy lapse. The authors also discussed the benefits of using machine learning in insurance analytics and provided insights into the factors that contribute to policy lapse.

In conclusion, these studies demonstrate the potential of predictive modeling for surrender rates in life insurance. By analyzing large datasets and using advanced statistical and machine learning techniques, insurers can develop more accurate and reliable models that can help them to identify high-risk policyholders, reduce surrender rates, and improve profitability. However, there is still a need for further research to explore the effectiveness of different predictive modeling techniques and to refine these models for specific policy types and markets.

Chapter 3

Methodology

3.1 Overview of the chapter

This chapter describes; the sources of data or data collection procedure, data pre-processing, model building, and model evaluation metrics. The models or algorithms used in the model creation include Random Forest (RF), Decision tree (DT), and Support Vector Machine (SVM). The ROC and Confusion matrix metric, such as sensitivity, Recall, precision, etc., was used to evaluate the model's performance in predicting the surrender rate in life insurance. Below is the methodology used for this study;

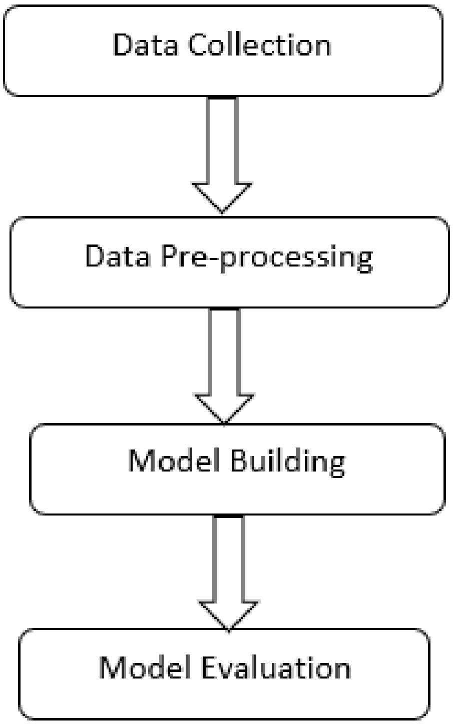

Figure 3.1: Methodology

3.2 Data collection

In this study, a dataset from Kaggle (2023), an open online data science platform, was used. The dataset consisted of 185,561 observations and 20 features, one of which was the class label indicating the Policy Status. The other 19 attributes included Channels, Policy types, Premium paid, Benefits, Policy description, and Policyholder's characteristics. To provide a more detailed description, the Policyholder's characteristics were represented by the entry age and sex, while the policy description was represented by various factors such as policy type, payment mode, policy status, policy year, benefit, premium, and more.

Table 3.1: Data description

NO	FEATURES	DESCRIPTION
1	Channel 1	
2	Channel 2	
3	Channel 3	
4	Entry Age	The age at which an individual enters or starts the insurance policy
5	Sex	The gender of the insured
6	Policy type 1	
7	Policy type 2	
8	Policy type 3	
9	Payment mode	The frequency of payments of premiums
10	Policy status	The state or condition of the policy as time elapses
11	Benefit	The contingent payment received by the policyholder
12	Non-Lapse guaranteed	An agreement that the death benefit of the insured is assured

NO	FEATURES	DESCRIPTION
13	Substandard risk	Individuals considered to be of higher risk
14	No. of advance premium	Payments made to an insurer for coverage that has not been provided
15	Initial benefit	The first benefit payable to the policyholder in accordance with the life insurance policy
16	Full benefit	This tells whether the policyholder has received full benefit or not
17	Policy year (decimal)	The actual number of years and months the insured is covered under the insurance policy
18	Policy year	The number of years the insured is covered under the insurance policy
19	Premium	The amount of money given to the insurer by the policyholder
20	Issue date	The date on which the policy was issued

3.3 Data Pre-processing

To improve the data's quality and create a refined dataset suitable for model development, data pre-processing was performed. Without pre-processing, various issues can arise such as inconsistencies, missing values, irrelevant features, model overfitting, and errors. To address these problems, data pre-processing was conducted, specifically to eliminate irrelevant features. In this case, the Issued Date column was deemed irrelevant for predicting whether the policy status inforce and surrender. For the relevant features such as Channel 1, Channel 2, Channel 3, Policy Type 1, Policy Type 2, Policy Year (Decimal), benefits, and premiums, extra cleaning of the dataset was done to remove the empty rows and to convert the objects to numeric values to ensure consistency in the dataset since inconsistency and null values influence the modeling of the data. Some categorical variables such as sex, policy status, non-lapse guaranteed, full benefit, and payment mode were converted to binary digits. This was done to ensure consistent numeric values throughout the data. Following the identification of non-normalized features, a normalization process was implemented.

3.3.1 Normalisation

Data normalization involves arranging data uniformly, ensuring consistency across all records and fields. This process enhances the coherence of entry types, leading to data cleaning, lead generation, segmentation, and overall improvement in data quality. To perform various analyses, including feature selection with Principal Component Analysis (PCA), the dataset (features) needs to be normalized. This means that the data should follow a standard normal distribution with a mean of zero and a variance of one $N(0,1)$. Normalization is necessary because features with high variance can have a disproportionate impact on the outcome of feature selection. The formula is given below;

$$Z = \frac{x-\mu}{\sigma} \tag{3.1}$$

x = feature value

μ = Mean

σ = Standard deviation

Following the process of standardization or normalization, each feature's mean is adjusted to zero, while the standard deviation is set to one. This ensures that all features are on a standardized scale and have a consistent distribution, allowing for fair comparison and analysis across different features.

- Mean

The term "mean" refers to the mathematical average obtained by summing up a set of two or more numbers and dividing it by the total count. There are various methods to calculate the mean, such as the arithmetic mean which involves summing up the numbers in the series, and the geometric mean which calculates the average of a group of products.

In this particular study, the arithmetic mean method was used. It is worth noting that, in most cases, the results obtained from different approaches to calculating the average are quite similar. To calculate the arithmetic mean or average of a dataset, two simple steps are followed: first, the values are summed up by adding them all together, and then the sum is divided by the total count of values in the dataset. The formula is given by;

$$\mu = \frac{\sum\limits_{i=1}^{n} x_i}{n} \tag{3.2}$$

x_i = Value of the ith feature

n = Number of observations (total count)

μ = Mean

- Variance

This refers to the statistical measure of the dispersion or variation among numbers within a dataset. It specifically evaluates how each number in the collection deviates from the mean (average) and one another. The symbol commonly used to represent variance is σ^2. To calculate the variance, the following steps are followed:

1. Compute the mean (average) of the data.
2. Determine the difference between each data point and the mean value.
3. Square each of these differences.
4. Sum up all the squared values.
5. Divide the sum of squares by either $n - 1$ (for a sample) or N (for a population), where n represents the number of data points in the sample

The formula for variance is given below;

$$\sigma^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \mu)^2}{n - 1} \tag{3.3}$$

x_i = Value of the ith feature

n = Number of observations (total count)

μ = Mean of features

σ^2 = Variance

- Standard deviation

The standard deviation is a statistic that quantifies the dispersion or spread of a dataset around its mean. It is computed by taking the square root of the variance. The standard deviation measures how far each data point deviates from the mean, indicating the extent of the dataset's distribution. The formula for calculating the standard deviation is as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \tag{3.4}$$

x_i = Value of the ith feature

n = Number of observations (total count)

μ = Mean of features

σ = Standard deviation of features

3.4 Feature selection

This involves identifying and choosing the relevant features from a dataset while eliminating the irrelevant ones, focusing on those that possess the highest predictive capability. The presence of irrelevant features can have a detrimental effect on the performance of machine learning models, particularly leading to overfitting when there are excessive features. By selecting the appropriate features, feature selection not only enhances prediction and analysis of results but also reduces processing time and storage requirements.

Initially, Exploratory Data Analysis (EDA) was done to understand the data components and to remove irrelevant features.

3.5 Machine learning

The field of Artificial Intelligence (AI) which deals with enabling computer systems to acquire knowledge and improve their performance over time without explicit programming is called machine learning. Supervised Machine learning (SML), unsupervised learning, and reinforcement learning are the three main categories of machine learning. This learning method starts with data or sets of examples, experiences, or instructions, which the system can use to recognize patterns and/or enhance its performance if required. Machine learning, a field of computer science and artificial intelligence, was given its name by Arthur Samuel, an IBM employee and a leading figure in computer gaming and AI, in 1959. During that period, the term ”self-teaching computers” was also used as a synonym for machine learning. In the following decades, research and interest in machine learning continued to grow, with books and reports focusing on pattern classification and teaching strategies for Neural Networks. (Mitchell, 1997) provided a formal definition of machine learning algorithms, emphasizing the improvement in performance through experience. Today, machine learning aims to classify data and make predictions based on developed models, such as using computer vision for cancerous mole classification or predicting stock trading outcomes.

3.5.1 Supervised and Unsupervised Algorithms

Supervised learning algorithms create a mathematical model using a dataset that includes both inputs and desired outputs. These algorithms iteratively optimize an objective function to learn a function that can predict outputs for new inputs.

On the other hand, unsupervised learning algorithms work with datasets that only have inputs and aim to find patterns or groups within the data. These algorithms learn from unlabeled data, without any specific feedback, classification, or categorization. Instead, they identify similarities or differences in the data and respond accordingly when new data is encountered.

3.6 Model building

In this section, the models were constructed by initially dividing the dataset into training and testing data using the cross-validation technique. Cross-validation is a statistical approach widely employed to evaluate classification models. It involves dividing the dataset into k folds, where one fold is designated as the test data while the remaining folds are used to train the models. This process is repeated until all folds have been utilized as training sets, and the average performance results are then computed from these iterations. Cross-validation is an effective method to achieve higher accuracy when working with limited data. In this study, the dataset is divided into a 70% training dataset and a 30% testing dataset. The training dataset is used to train the model and identify patterns within the data. The testing dataset is reserved for prediction, ensuring the model remains unbiased during the prediction phase. It is crucial to evaluate the model's performance on new, unseen data. The training dataset is typically larger than the testing dataset as a larger sample size allows the model to observe more patterns and improve accuracy. The Scikit-learn package plays a vital role in model building. The specific models used are discussed below;

3.6.1 Decision tree

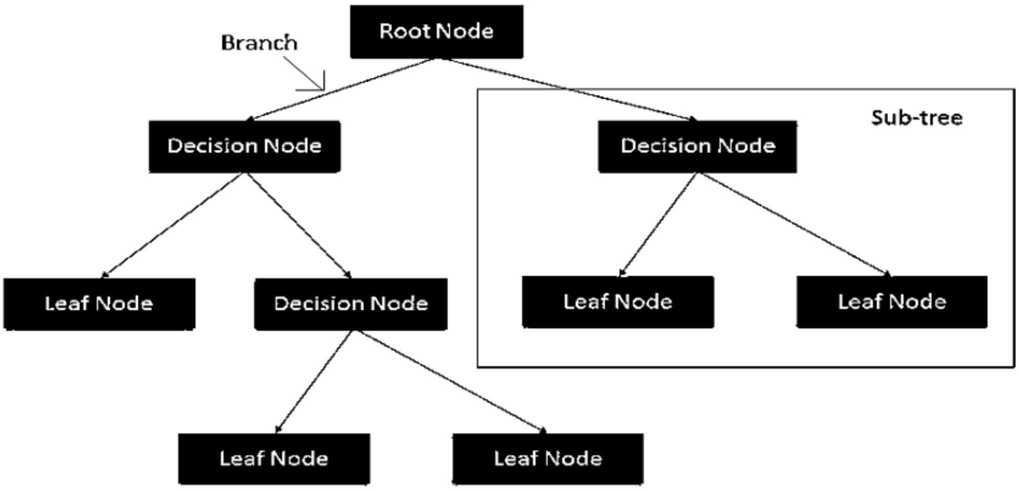

Figure 3.2: Decision Tree structure

The Decision Tree Algorithm is a supervised learning algorithm used in machine learning that learns from data to construct a tree-like model of decisions and their possible consequences. The algorithm recursively splits the data into smaller and smaller subsets based on the value of certain features, eventually arriving at a decision or prediction. The decision tree is made up of three nodes, Decision nodes, Chance nodes, and End nodes. Decision Nodes (splitting points of data), Decision paths (which are the rules), and Decision Leaves (the final results) are the three main parts of the decision tree algorithm. This algorithm operates through splitting, pruning, and tree selection techniques. It can construct decision trees that handle both numerical and categorical inputs. Decision tree algorithms are particularly useful for managing large datasets with reduced time complexity. The primary usage of this algorithm in business is to create customer groups and establish marketing strategies.

In a tree-based model, the dataset is divided recursively into groups based on a specific criterion until a predetermined stopping condition is met. The terminal nodes, also known as leaf nodes, are located at the bottom of the decision trees. Decision trees can be used for both classification tasks (categorical outcomes, such as logistic regression) and regression tasks, depending on how the splitting and stopping criteria are defined. The selection of predictor variables to split an internal node depends on specified criteria, which are considered optimization problems for both classification and regression operations. Entropy, which is a practical implementation of Shannon's (2001) source coding theory, is a commonly used criterion for dividing data into classification problems. At each internal decision node, entropy is calculated as,

$$E = -\sum_{i=1}^{c} P_i \times \log(P_i) \tag{3.5}$$

where P_i represents the prior probability for each specific class, and c denotes the total number of distinct classes. The entropy, a measure of impurity or uncertainty, is calculated at each decision tree split. The goal is to maximize the entropy value to extract the maximum amount of information at each split. In regression problems, the mean squared error is a commonly used criterion to determine the best splits at each internal node.

3.6.2 Random forest

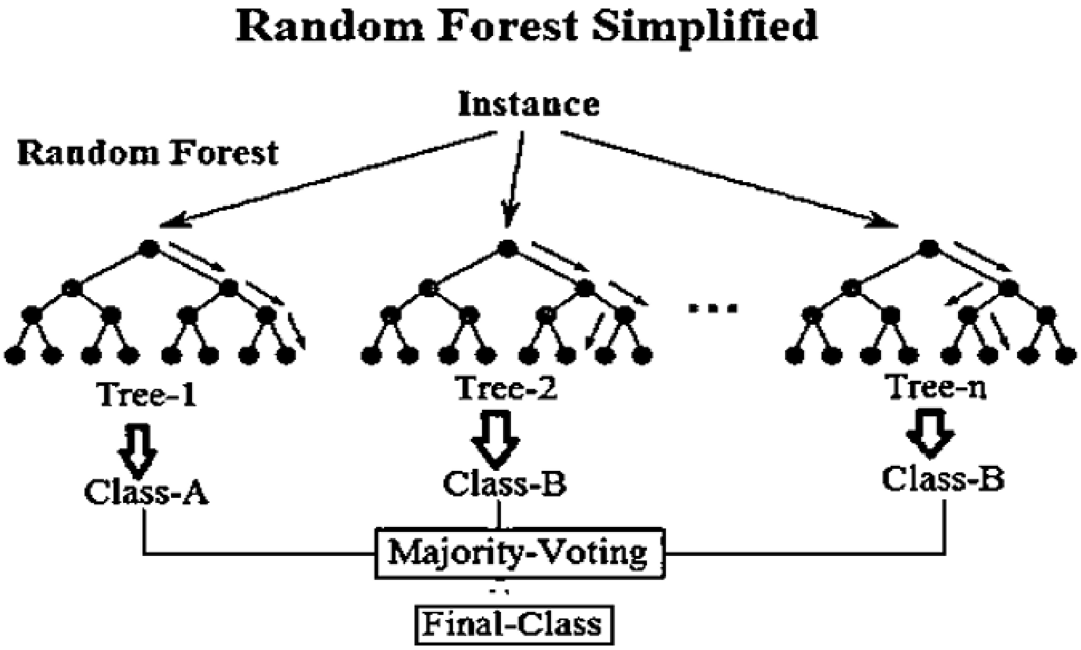

Figure 3.3: Random Forest structure

RF method employs ensemble learning techniques for both classification and regression to generate a considerable number of decision trees. It uses bagging and feature randomness to create these trees. One advantage of Random Forest (RF) over Decision Tree (DT) is that it avoids over-fitting the data. For classification tasks, the output of a Random Forest is determined by the majority class chosen by most of the individual trees. In regression tasks, the mean prediction of each tree is returned as the output. Random decision forests address the issue of decision trees over-fitting the training data. While random forests often outperform decision trees, gradient-boosted trees tend to be more accurate. Random forests are highly regarded for their accuracy and reliability due to the involvement of a large number of decision trees in the process. The construction of k distinct decision trees is attempted by randomly selecting a subset s of the training samples. These trees are built without pruning, following the principles of fully Iterative Dichotomiser 3 (ID3) trees. Random forests take the average of all predictions, effectively mitigating biases and making them resistant to overfitting. The final prediction is generated by taking the mean of all individual predictions.

- Principle of Random Forest

A random forest is a type of classifier that comprises a set of tree-based classifiers $h(x, \theta_k, k = 1, ...)$ where the θ_k is randomly generated and independent vectors. Each tree within the random forest independently assigns a vote to the most popular class for a given input x (Breiman, 2001). Therefore, Random Forest can be described as a collection of multiple classifiers that utilize tree structures. The growth of each tree in the ensemble is governed by the generation of random vectors, which contribute to the development of the overall model. In the work (Breiman, 1996), The ensemble model is built by random splits where at each node the split is selected at random from among the best $K - split$. The process involves generating a random vector θ_k for each kth tree, independent of the previous vectors $\theta_1 ... \theta_{k-1}$ but with the same distribution. Using this vector and the training set, a tree is built to create a classifier $h(x, \theta_k)$, where x represents an input vector. In the case of random split selection, θ is composed of multiple independent random integers between 1 and k. The specifics and dimensions of θ vary based on its role in constructing the tree. After numerous trees are created, they collectively vote for the most popular class predicted by each classifier $h(x, \theta_k)$ to determine the final prediction. The majority voting is determined by the formula or the decision function $H(x) = \arg\max_y \sum\limits_{i=1}^{k} I \times h_i(x) = Y$, where $H(x)$ is a combination of the classification model, h_i is a single decision tree model, Y is the output variable, $I \times h_i(x) = Y$ is the indicator function, this is the principle of classical random forest.

3.6.3 Support Vector Machines (SVMs)

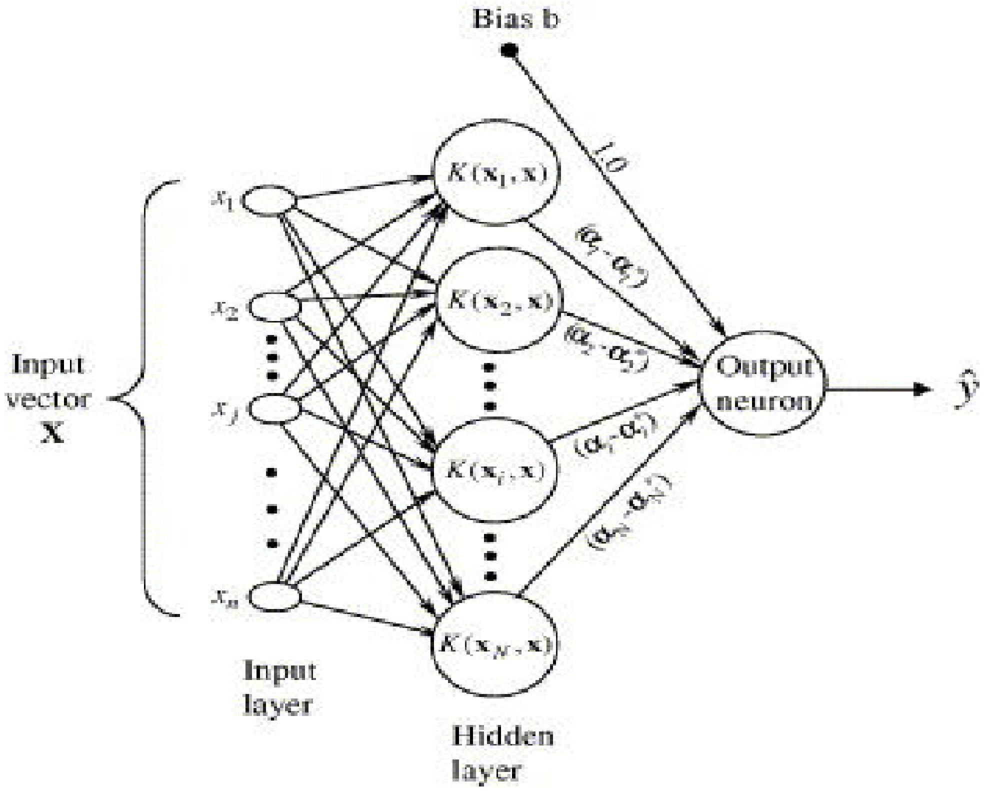

Figure 3.4: Support Vector Machines (SVMs) structure

SVM is a robust data classification tool that can effectively model complex problems, including text classification, handwriting recognition, and complex number classification. It can deal with both linear and nonlinear issues.

Support vector machines (SVMs) are supervised learning models and learning algorithms that evaluate data for regression and classification purposes. created by Vladimir Vapnik and associates at AT & T Bell Laboratories (Boser et al., 1992; Guyon et al., 1993; Cortes and Vapnik, 1995; Vapnik et al., 1997); SVMs, which are based on statistical learning frameworks or the VC theory put forward by Chervonenkis (1974) and Vapnik (1982, 1995), are among the most reliable prediction techniques. An SVM training algorithm constructs a model that categorizes new samples into two categories, acting as a binary linear classifier without probability estimates. By examining a set of training examples, the SVM maps the features of these examples to points in space, aiming to maximize the separation between the two categories. When new features are introduced, they are also mapped to the same space and classified based on which side of the separation they fall on.

3.7 Classification of data

Data classification is one of the important features of machine learning. In the case of support vector machines, the overview of the classification process is to decide which class a given data point belongs to. Suppose a data point consists of k numbers, we are to decide if it can be separated by a $(k-1)-dimensionalhyperplane$. Due to the numerous hyperplanes that can be classified, the best hyperplane is chosen such that the distance from it to the nearest data point is maximized. This hyperplane is called the maximum-margin hyperplane and the associated linear classifier is called the maximum-margin linear classifier. To make the separation easier, the desired hyperplane is achieved by defining the linear kernel function to suit the problem. Other kernel functions include nonlinear, polynomial, radial basis functions, and sigmoid.

3.7.1 The principle of linear support vector machines.

Consider a training dataset consisting of m points of the form $(x_1, y_1)...(x_m, y_m)$ where the y_i can take values either -1 or 1, each indicating the class to which the data point x_i belongs. Each x_i represents a vector in a $k-dimensionalspace$. We want to find the maximum-margin hyperplane that divides the group of points x_i, where $y_i = 1$ from the group where $y_i = -1$. The hyperplane is defined in a way that the distance from the hyperplane to the nearest point on either side is maximized. Mathematically, the basic equation of sets of points that satisfy any given hyperplane is $w^T x - b = 0$, where w is the normal vector to the hyperplane. The distance between two hyperplanes is called margin. The maximum margin hyperplane is the hyperplane that is positioned exactly halfway between the two hyperplanes. Geometrically, the distance between these two hyperplanes is $\frac{2}{||w||}$. Hence, we minimize the norm of w $(||w||)$ to maximize the margin between the hyperplanes. Once the optimal hyperplane is chosen, then we use it to make predictions as $\left\{\begin{array}{ll} w^T x_i - b \geq 0 & y_i = 1 \\ w^T x_i - b \leq 0 & y_i = 0 \end{array}\right\}$ where y_i is the output variable.

3.8 Model Evaluation

This section evaluates the performance of the developed model to determine how effective it is. A confusion matrix is generated, serving as the basis for computing various performance measures including accuracy, specificity, sensitivity, recall, $f1$ score, and many more. These measures provide valuable insights into the effectiveness and reliability of the model's performance.

3.8.1 Confusion Matrix

A confusion matrix is a concise tabular representation of the classifier's predictions summarizing both the number of correct and incorrect classifications. The confusion matrix is used to evaluate the performance of a classification model, allowing us to compare predicted outcomes with the actual observed results. By analyzing the confusion matrix, we gain insights into the accuracy of our predictions and the trade-offs involved. The Matrix also helps in comparing different classifiers by computing Accuracy, Specificity, Sensitivity, Area Under the Curve (AUC), and ROC curve. The confusion matrix gives us a comprehensive view of the model's overall performance, providing a detailed analysis of its effectiveness (Corporate Finance Institute, 2022).

Table 3.2: Confusion Matrix for Surrender Prediction

Diagnosis	Predicted as In-force	Predicted as Surrender	Predicted as Lapse
In-force	TP	FN	FN
Surrender	FP	TN	TN
Lapse	FP	TN	TN

In-force is a positive outcome.

Surrender and Lapse are negative outcomes.

TP: True Positive is the number correctly predicted as In-force

FP: False Positive is the number incorrectly predicted as In-force

FN: False Negative is the number incorrectly predicted as Surrender

TN: True Negative is the number correctly predicted as Surrender

3.8.2 Algorithm Evaluation Metric

Several metrics and techniques can be employed to summarize the information in the confusion matrix forming the final step of the prediction model. In this step, various evaluation metrics like classification accuracy, precision, specificity, recall, F1-score, and ROC metrics are used to assess and analyze the prediction results.

ROC: It is a performance measure specifically designed for binary classification problems.

The accuracy measure is the proportion of correct predictions to the total number of input samples illustrating the percentage of accurate predictions. It is given as

$$Accuracy = \frac{Correct\ prediction}{Total\ observations}$$

Accuracy can also be computed from the confusion matrix by averaging the values on the main diagonal, providing the overall accuracy of the predictions. It is given as

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

Sensitivity, also known as recall determines the percentage of the positive class accurately classified. It shows the proportion of the actual positive instances that were correctly identified. Mathematically,

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

Specificity determines the percentage of negative instances accurately classified. It describes the proportion of the actual negative cases that were correctly identified. Mathematically,

$$Specificity = \frac{TN}{TN + FN}$$

Precision measures the likelihood of a sample classified as positive being truly positive, providing insight into the quality of positive predictions. It addresses the question of how accurate the positive classifications are by calculating the proportion of correctly predicted positive instances out of all instances classified as positive. It is given by,

$$Precision = \frac{TP}{TP + FP}$$

The $F1 - measure$ evaluates the precision of a test and represents the harmonic mean of accuracy and recall. Ranging from 0 to 1, it assesses both the precision and robustness of the classifier, striking a balance between precision and recall. Mathematically, it is calculated as follows:

$$F1 - measure = 2 \times \frac{precision \times Recall}{Precision + Recall}$$

Chapter 4

Results and Discussion

4.1 Introduction

In this chapter, the study presents its findings and discusses the objectives. The exploratory analysis of multivariate data is displayed through tables and graphs. Results of three machine learning algorithms are presented, with the Receiver Operating Characteristics (ROC) serving as the metric for comparison. Various model evaluation metrics, such as precision, recall, the F1 measure, accuracy, and specificity, are used. Therefore, achieving higher accuracy in the outcomes of these models is crucial. These models will serve as essential diagnostic tools for decision-making by underwriters and other stakeholders. Their application aims to mitigate underwriting risks, reduce expenses, and prevent revenue loss resulting from an increased number of claims by policyholders with higher risks of surrendering. The predictive models for surrender status are of utmost importance due to the significant negative impact surrendering life policies can have on both insurers and policyholders.

4.2 Data Description and Summary

Exploring the life insurance policy dataset is vital for understanding the data.

Table 4.1: Profile of the continuous features.

FEATURES	MEAN	ST. DEVIATION	MINIMUM	MEDIAN	MAXIMUM
CHANNEL1	4.164	2.082	1	6	8
CHANNEL2	2.461	0.561	1	2	3
CHANNEL3	10.116	14.254	0	1	82
ENTRY AGE	34.841	10.223	16	33	70
POLICY TYPE 1	6.701	4.145	1	6	20
POLICY TYPE 2	29.453	22.402	1	34	84
BENEFIT	34725.486	81181.095	30	20000	5000000
SUBSTANDARD RISK	0.566	7.458	-50	0	250
NUMBER OF ADVANCE PREMIUM	0	0.029	0	0	5
INITIAL BENEFIT	384.963	4379.393	0	0	266017.24
POLICY YEAR (DECIMAL)	2.913	1.768	0	3	8
POLICY YEAR	3.453	1.770	1	4	9
PREMIUM	1289.998	5678.369	0	432	427053

Table 4.1 displays descriptive statistics for the continuous variables. It was evident that certain variables had significantly higher means than others. To facilitate our analysis and predictions, it is necessary to standardize the dataset.

Table 4.2: Frequency Table of Policy Status (target variable)

POLICY STATUS	**FREQUENCY**	**PERCENTAGES**
Lapse	43253	52.6%
Surrender	7780	9.5%
Inforce	30894	37.6%
Death	230	0.3%
Expired	93	0.1%
TOTAL	82250	100%

Table 4.2 illustrates the frequency of each policy status. The dataset used in this analysis comprises 82,250 observations and 20 variables, with the insured's policy status (target variable) included. Among the 82,250 observations, 43,253 (52.6% of the data) represent Lapses, 7,780 (9.5% of the data) represent Surrender, and 30,894 (37.6% of the data) represent In-force policies. A small fraction of the data, specifically 250 (0.3% of the data), represents deaths, and 93 (0.1% of the data) represents expired policies.

The pie chart below gives a clear view of the percentages of each policy status.

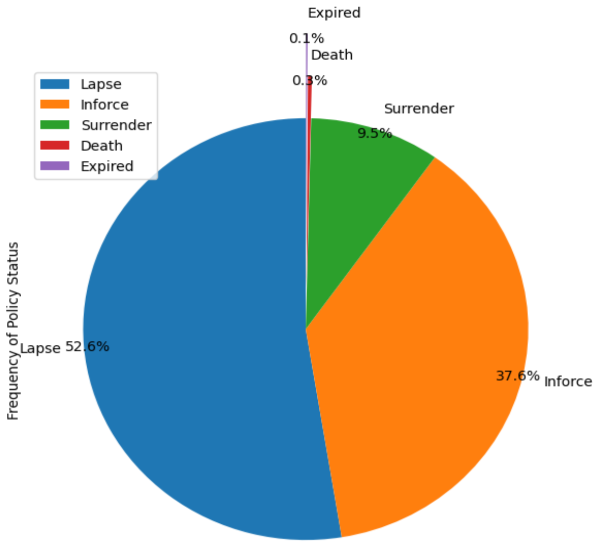

Figure 4.1: Pie chart of the frequency of Policy status

Due to imbalances in the data, it is imperative to drop death and expired observations since their proportions will affect accuracy woefully. This is because most standard supervised classification models assume that the categories in the target variable are balanced (Somasundaram et al, 2016).

See Table 5.1 in the Appendix. In Table 5.1, certain features exhibit larger mean values, minimums, standard deviations, and maximums compared to the other features. However, as observed in Table 4.1, after standardization, all the features were scaled to fall within a specific range. Moreover, all the features now have a mean of zero and a standard deviation of one, indicating that the assumption of standardization has been met, and the features follow a standard normal distribution ($N(0, 1)$).

4.3 Bivariate Analysis of Data

Based on the data, an empirical hypothesis can be formulated to determine which features influence the policy status of a policyholder through bivariate analysis of the dataset. This hypothesis will form the basis for obtaining the final results from the predictive model's feature importance using data mining techniques.

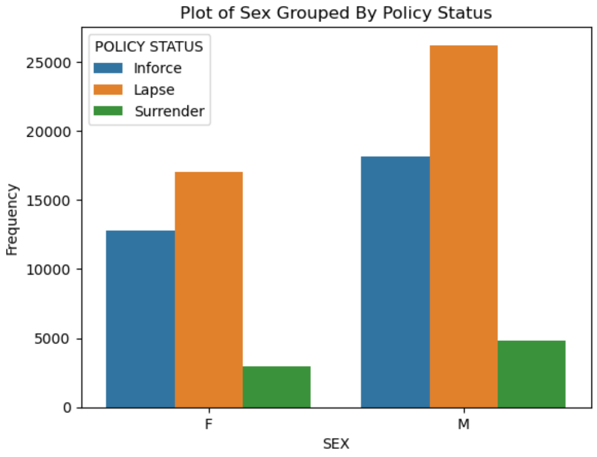

Figure 4.2: Policy Status categorized by Sex

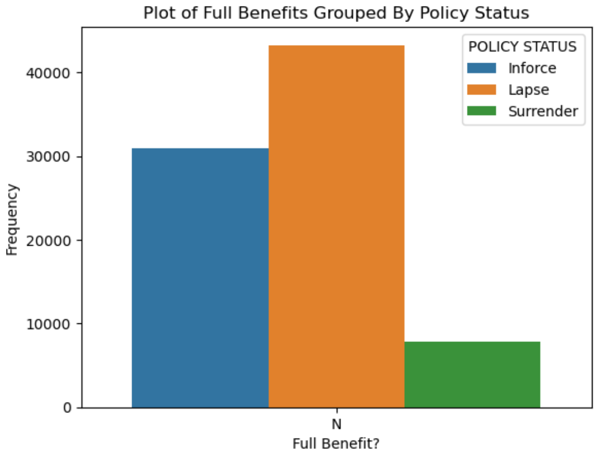

Figure 4.3: Policy Status grouped by Full Benefit?

Figure 4.2 shows the policy status of policyholders categorized by sex. It can be observed from the diagram that Inforce, Lapse, and Surrender are higher for males as compared to females. Hence sex can be considered a predictive factor in policyholders surrendering their life insurance policy.

Likewise, **Figure 4.3** shows the policy status of policyholders categorized by full benefit. It can be observed from the diagram that Inforce, Lapse, and Surrender are higher for policyholders not receiving full benefits as compared to policyholders receiving full benefits. Hence full benefits can be considered as a predictive factor in policyholders surrendering their life insurance policy.

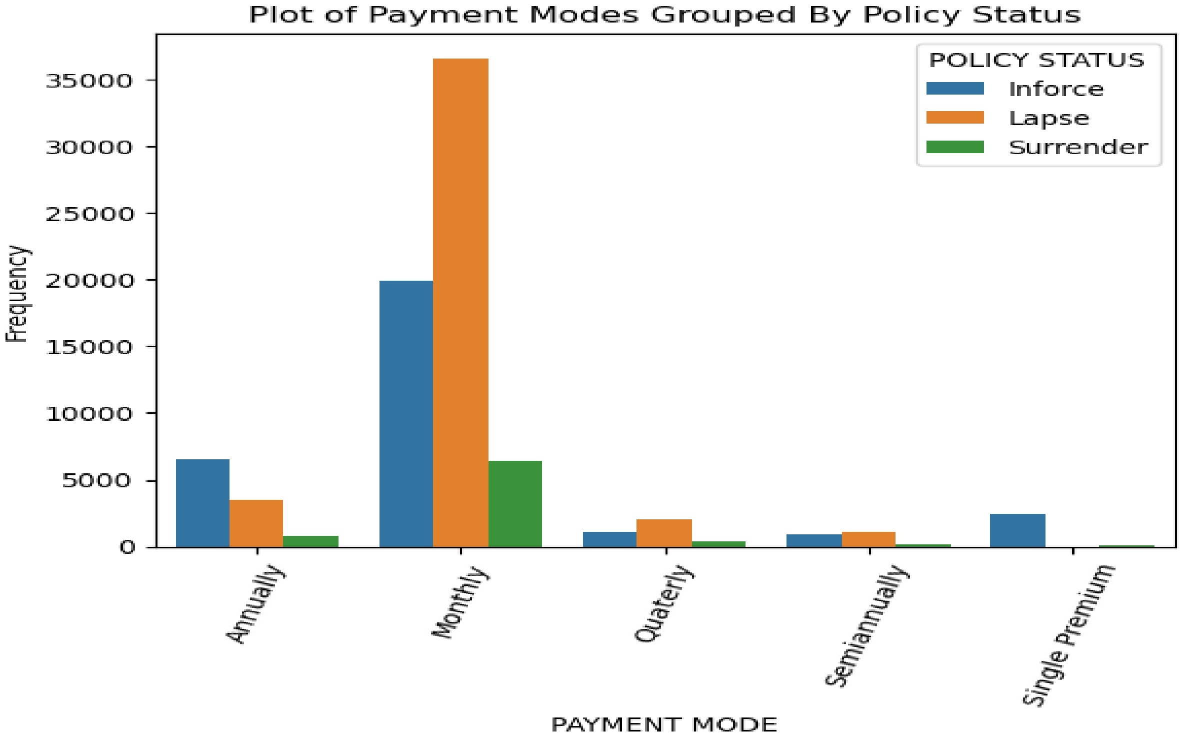

Figure 4.4: Policy status categorized by payment mode.

Figure 4.4 shows the policy status of policyholders categorized by payment mode per each policyholder. It can be observed from the diagram that Inforce, Lapse, and Surrender are higher for policyholders paying premiums monthly as compared to other payment mode choices. Hence payment mode can be considered as a predictive factor in policyholders surrendering their life insurance policy.

The rate of a policyholder remaining in force (active), lapsing, or surrendering as contained in policy status to other continuous variables is represented.

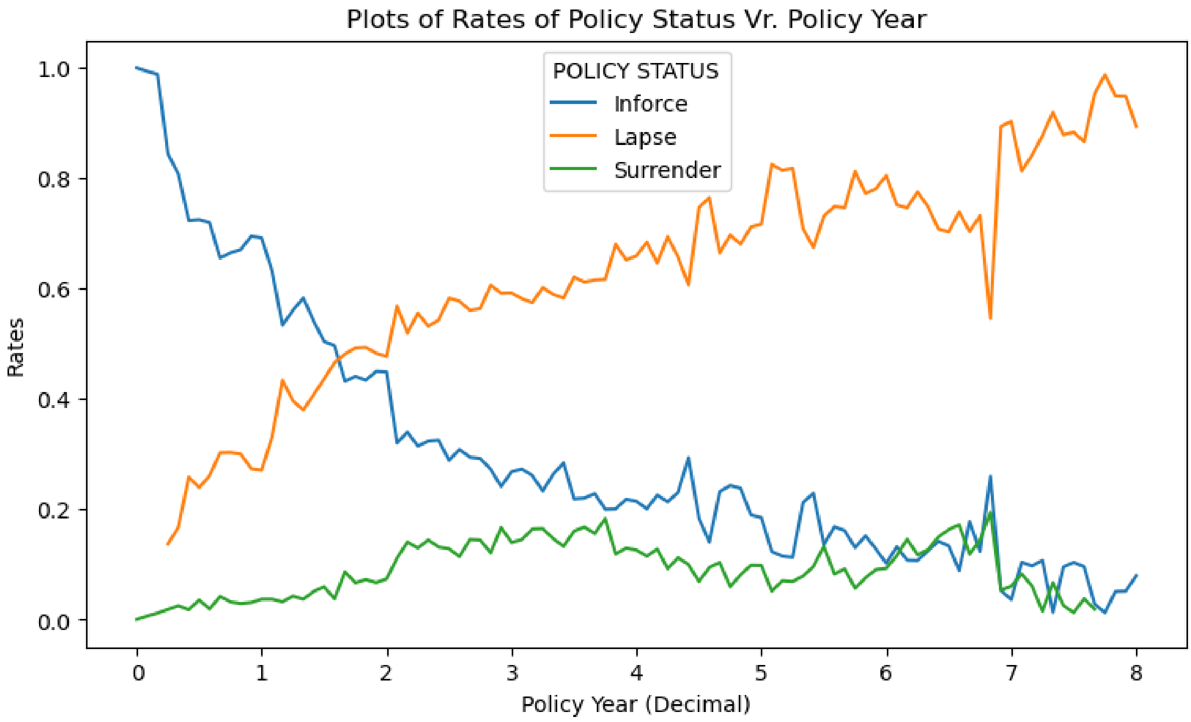

Figure 4.5: Rates of policy status and policy year (decimal) and policy year

From Figure 4.5 which depicts the rates of policy status (inforce, lapse, and surrender), it can be observed that, as the number of years of coverage of a policyholder increases, the rate of the policyholder being active decreases. While there is a positive correlation between lapse and policy year, there is an increasing rate of surrender within the early years of inception of the policy which tends to decrease from the fourth year of the policy. This indicates policy year (Decimal) and policy year can outlay the dynamism within the policy status data and therefore are a good predictive variable.

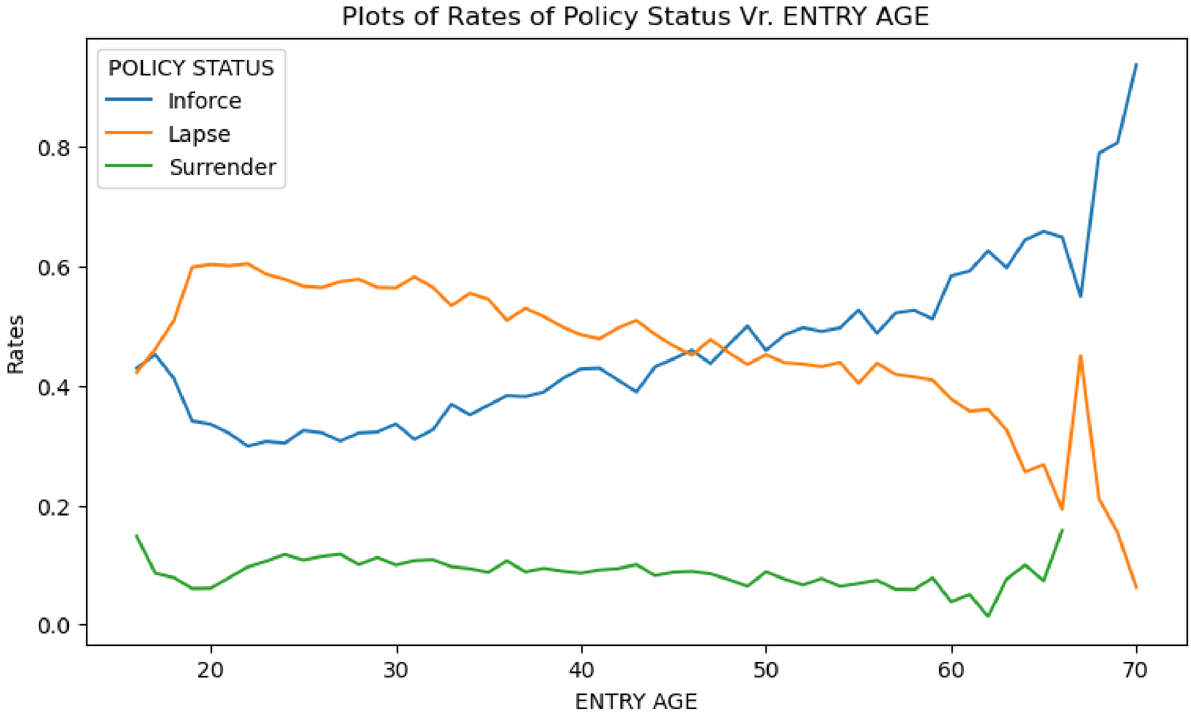

Figure 4.6: Rates of policy status against the entry age of the policyholder

Similarly, Figure 4.6 depicts the rates of policy status (inforce lapse, and surrender) as the entry age of a policyholder increases. Though the rates are fairly constant as the entry age increases, policyholders who joined the scheme at age 30 and above have an increasingly higher rate of being inforce(active) with decreasing lower rate of surrendering and lapse. This figure shows the differences within different entry ages and therefore a good candidate for prediction.

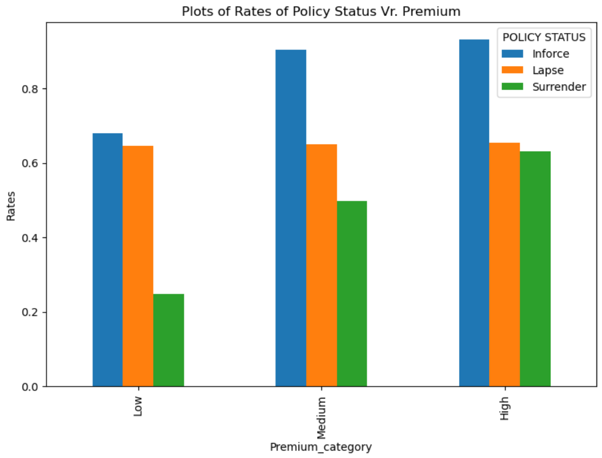

Figure 4.7: Plot of rates of policy status and Premium

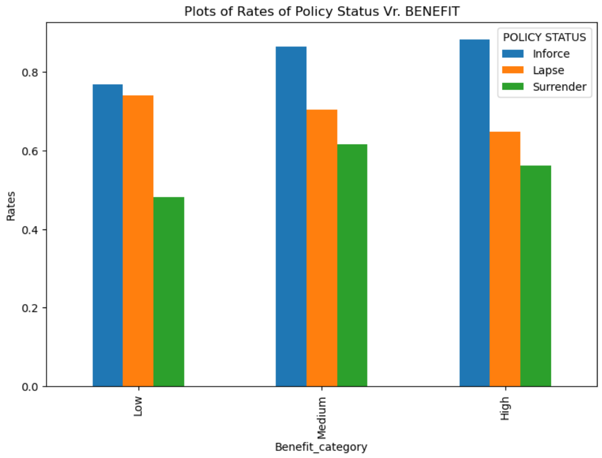

Figure 4.8: Plot of Rates of policy status and benefit

Figure 4.7 illustrates the rates of policy status and Premium. It can be observed that an increase in premium results in an increase in surrender and a gradual increase in lapse. The number of active policyholders also increases as the premium increases. Generally, the rate of policy status is positively correlated to premium. Hence Premiums can be considered a predictive factor in policyholders surrendering their life insurance policy.

Figure 4.8 also shows that an increase in benefits received by a policyholder results in a gradual decrease in lapse. Rates of surrender are higher for policyholders receiving medium benefits as compared to those receiving lower and higher benefits. The rate of active policyholders increases as benefit increases. This means retention of policyholders can be achieved by increasing the benefits of policyholders. Hence, benefits is a good predictive factor since it can reveal the complex structure within different policy status. Other variables depicted an unclear and complex structure that cannot be readily interpreted by observation and are allowed for the explainability of features by the models through data mining.

4.4 Model Building

In this section, the Scikit Learn module plays a crucial role. It was utilized to split the data into a training dataset (70%) and a testing dataset (30%) using the *train_test_split* function. The training set contains 57,348 observations, while the test set contains 24,579 observations. Additionally, the Scikit Learn module was employed to train the model using the fit function. The models trained include a Decision tree, Random forest, and Support vector machine. In the case of the Decision tree, a $5 - foldcross - evaluation$ is used to choose the observations for training and testing using the $GridSearchCV$ function. The Support vector machine was built using the $RadialBasis$ function (RBF) kernel. The model building aims to contrast the accuracy from models of all categories of the target variable and that of the accuracy when considering lapse and surrender.

4.5 Feature Importance

The Decision Tree algorithm and Random forest feature importance functions serve as an important data mining technique in obtaining the most contributing dataset variables to prediction in the machine learning algorithm. These important features are computed based on a mean decrease in impurity at each node of the tree. This section considers the exploration of important features from datasets that contributed to model performance.

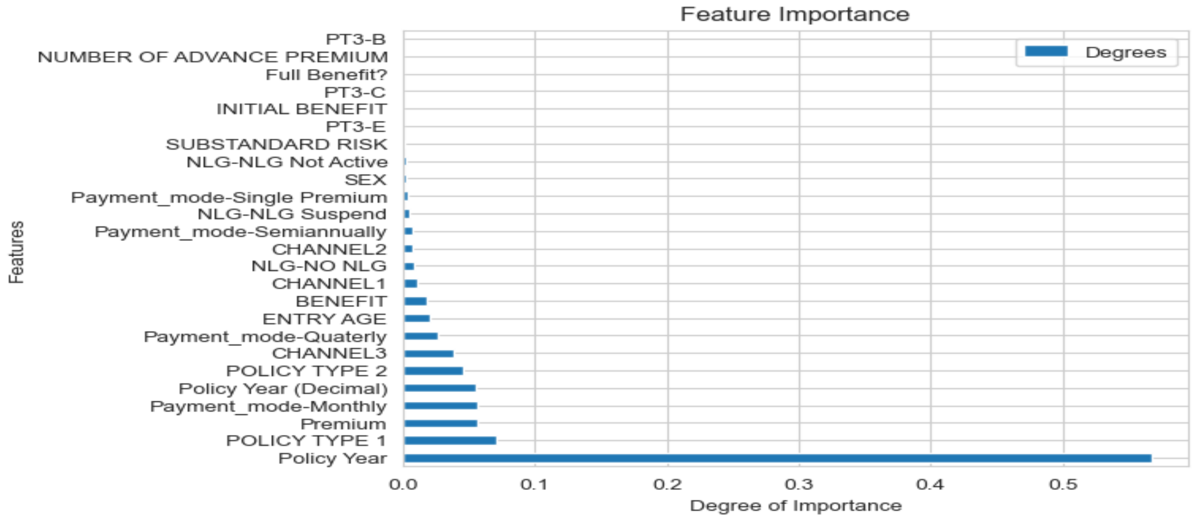

Figure 4.9: Feature importance for Decision Tree

The figure illustrates that policy year is the most important feature with a degree of 0.567, which account for the number of years a policyholder has been covered by the insurance. This is directly followed by policy type 1. Other features including Payment mode, Policy types 2, Premium, Channel 1, Entry Age, Benefits, etc were also deemed to be important. Sex is the least important feature which is almost negligible.

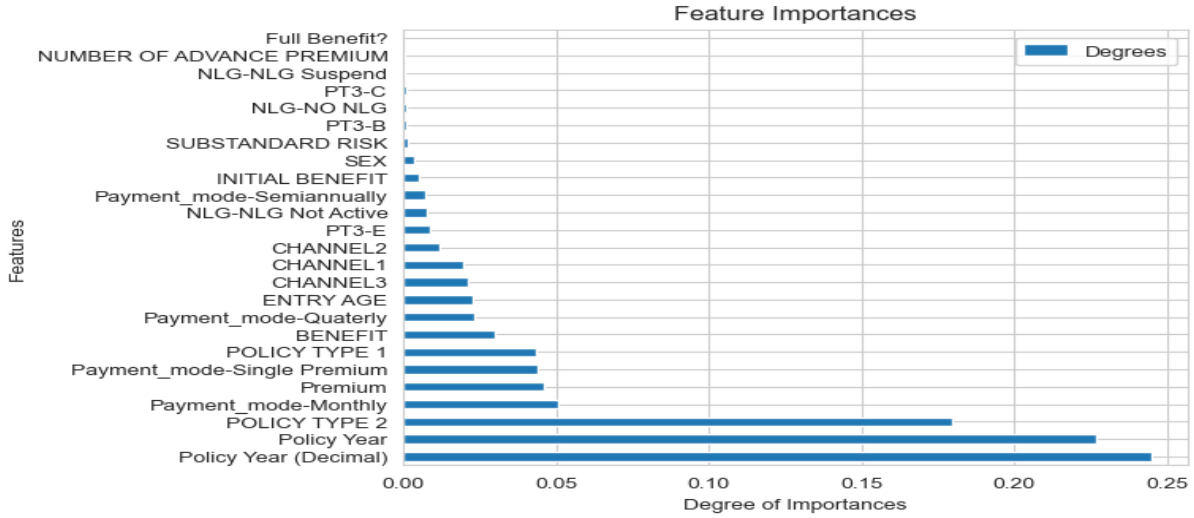

Figure 4.10: Feature importance for Random Forest

According to the Random Forest algorithm, the policy year (decimal) which accounts for the actual number of years and months a policyholder has been covered by the insurance is the most important feature with a degree of 0.245. This is directly followed by the policy year. Other features include Payment mode, Policy types 2, Premium, Channel 1, Entry Age, Benefits, etc. The Substandard risk is the least important feature with a degree of 0.0003 which is almost negligible.

4.6 Model Evaluation

The figures below illustrate the count for correctly and incorrectly classified observations.

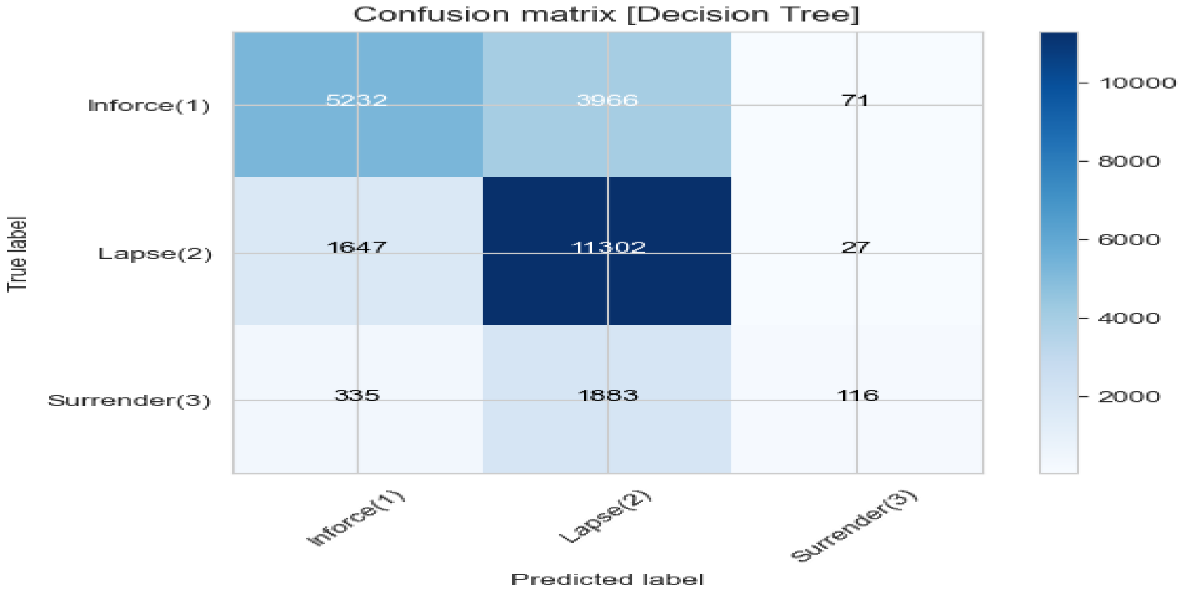

Figure 4.11: Confusion Matrix for Decision Tree

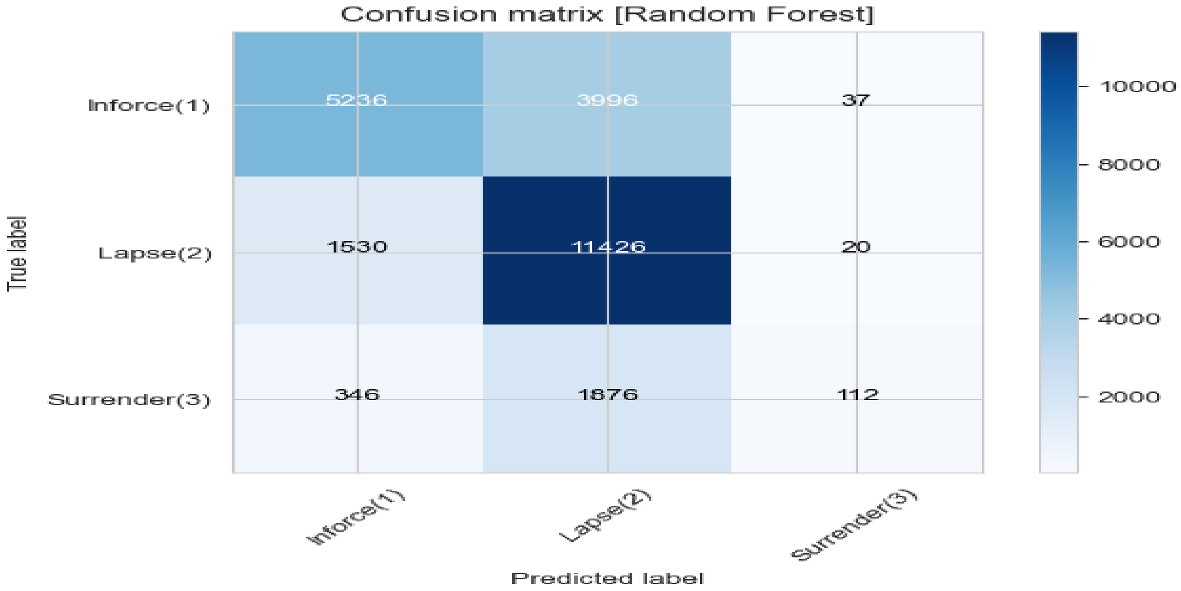

Figure 4.12: Confusion Matrix for Random Forest

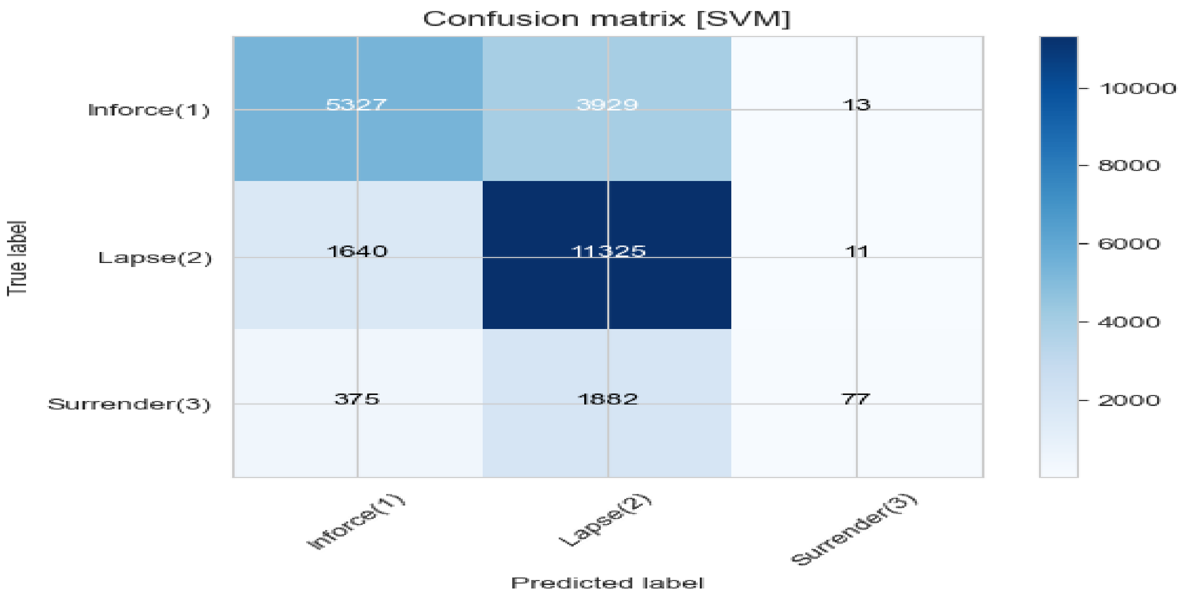

Figure 4.13: Confusion Matrix for Support Vector Machine

Generally, all the models achieved 68% accuracy in predicting the policy status. Random Forest correctly predicted 68.25% of the policy status(i.e. Inforce, Lapse, and Surrender) and misclassified 36.43% of observations in the policy status. Similarly, Support Vector Machines correctly predicted 68.06% of all categories of the policy status but misclassified 36.67% of cases. On the other hand, Decision Tree correctly predicted 67.74% of the observations of policy status and misclassified 37.21% of all the cases.

4.7 Algorithms Performance Evaluation.

Table 4.3: Model performance evaluation Result

MODELS	ACCURACY	MSE	PRECISION	RECALL	F1 SCORE
DT	0.6774	0.3721	0.6729	0.6774	0.6442
RF	0.6825	0.3643	0.6893	0.6825	0.6481
SVM	0.6806	0.3667	0.6949	0.6806	0.6450

Table 4.4 presents the performance evaluation metrics for each of the algorithms. The Support Vector Machine achieves the highest Precision at 69.49%, indicating its superior ability to predict true positives compared to the other algorithms. This implies that the

Support Vector Machine consistently captures the most accurate observations, reducing the false-negative rate. The Random Forest algorithm follows closely with the second highest precision at 68.93%. The Precision of the Decision Tree algorithm is the lowest among the algorithms at 67.29%.

Random forest has the highest recall of 68.25%, Support vector machine has the second highest recall of 68.06% and Decision Tree has the least recall of 67.79%.

Considering the F1 score metric, Random Forest is the best-performing model amongst the three models in predicting policy status(policy surrender) with an F1 percentage of 64.81%. The algorithm with the least F1 measure is the Decision Tree Classifier with an F1 score of 64.42%.

Random Forest recorded the highest accuracy of 68.25% followed by Support Vector Machines with an accuracy of 68.06%. Random Forest is the best-performing model in terms of accuracy, Mean square error measure, recall and f1 score. The Support Vector Machine outperformed the Decision tree in terms of all the evaluation metrics considered above.

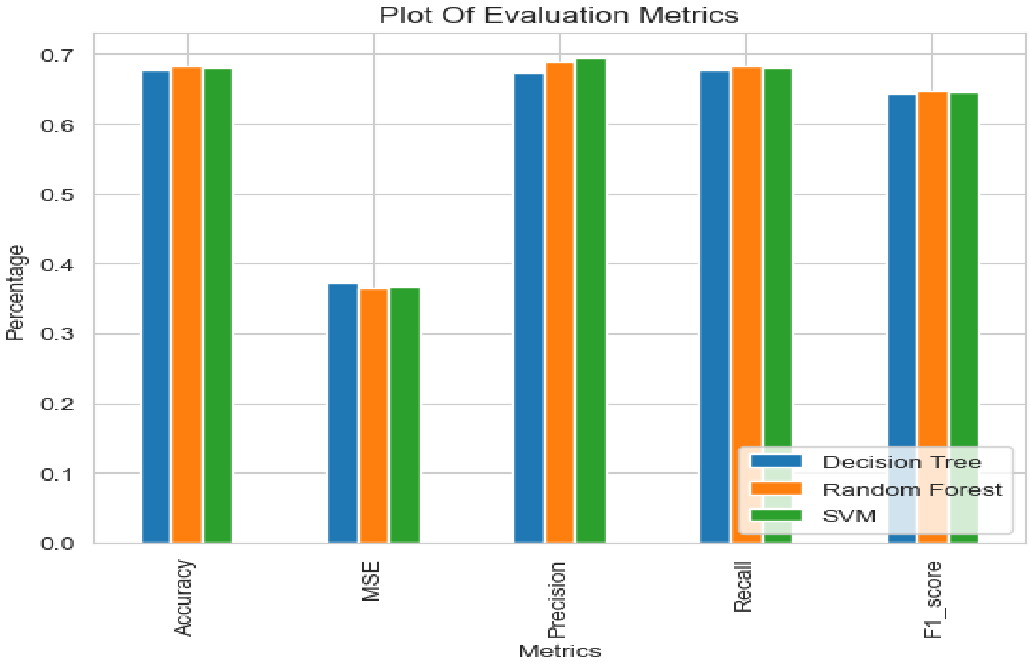

Figure 4.14: Graphical representation of Evaluation Metrics

From the graph, Random Forest has the highest value for Recall, accuracy, f1 score, and the smallest Mean square error. It indicates that it is the best model for predicting the policy surrender of a policyholder. The second-best model in terms of accuracy and precision is the support Vector Machine, which performed better than the Decision Tree concerning mean square error, recall, and F1 score.

4.8 Chapter Summary

This chapter presents the research findings, and the analysis results are illustrated using tables and figures in Chapter 4. The chapter covers descriptive features and exploratory data analysis, including bivariate analysis to assess the impact of independent variables on the target variable. The study utilized the ROC metric to evaluate the performance of supervised machine learning algorithms. Metrics like precision, accuracy, F1 measure, and Mean were employed to evaluate the algorithms. Based on these evaluation metrics, the Random Forest algorithm as an ensemble of several decision trees emerged as the most effective model in terms of accuracy and overall performance. The Support Vector machine with the highest precision outperformed the decision tree algorithm. These findings give similar results to that of (Kgare, 2021), where they employed SVM, Gradient Boost, Random Forest and CART algorithms to build predictive models on lapse. In conclusion, Random Forest and Gradient Boost outperformed single classifier models in terms of statistical accuracy with an accuracy of 86% and 92% respectively.

Chapter 5

Summary, Conclusion, and Recommendations

5.1 Introduction

This chapter talks about the summary of the findings, a conclusion, and recommendations. The objectives stated in chapter one which served as a direction for the study influenced the study's findings and conclusions, which are presented in this section. The recommendation in this section was also based on the study's findings. In this chapter, practical and straightforward recommendations are given and their effectiveness is also assessed.

5.2 Summary of Findings

The research illustrated that standardization is crucial for a dataset of features with different scales since higher variance affects the performance of a model. Exploratory data analysis is essential to understand the structure and dynamism within the data. Data splitting was also necessary because models were trained to test data it was not accustomed to. The results of the research are that the Support vector machine is the best performing model where precision is required. The Random forest model as an ensemble model is the best of all the three models based on accuracy, recall, f1 score and Mean square error measurement.The least performing model is the Decision tree.

5.3 Conclusion

This section comprises the concluding remarks of the research, which are drawn from the finding guided by the research objectives.

First and Foremost, the research concludes that the Random Forest algorithm is the best-performing model in producing accurate predictions in policy status (surrender) classification problems. Though the Support Vector Machine is the second best performing model, it is more precise, and using it to predict the rate of policyholder surrendering their life insurance policy where precision is required will be beneficial. Random Forest is an Ensemble model and therefore has an improved performance than the stand-alone models such as Decision Tree.

In addition, The most important features in predicting the policy surrender status of a policyholder are policy year (decimal), Policy year, Premium, Payment mode, channel 1-3, Benefit, Entry age, Sex, Policy type 1-3, and Non-lapse guaranteed.

Moreover, surrender rate is directly correlated with policy year and policy year (decimal) whiles policyholders who enter the scheme at age 30 and above have decreasing rate of surrender. Male policyholders have a higher number of surrenders as compared to their female counterparts. It is imperative to conclude that higher premiums are associated with higher rates of surrender whiles benefits tend to be associated with a more complex structure in terms of surrender. Therefore, features of at-risk policyholders are higher premiums, an entry age of 30 and below, and longer policy years.

Finally, Though machine learning models can accurately predict the policy surrender and lapse status of policyholders, not all algorithms are equal in terms of performance; therefore, machine learning experts or enthusiasts must choose their models based on some measure of accuracy. According to the study, the Random Forest is the best algorithm for predicting policy surrender status in life insurance. Also, results from exploratory data analysis should be incorporated to enhance the explainability of models during prediction.

5.4 Recommendations

This section presents the study's recommendations based on the research findings, analysis, interpretation, and discussion. Under the study-specific objectives, the following recommendations are made based on the study findings: Other Underwriting risk assessment models involving machine learning techniques can be developed to ensure policy is issued to low-risk policyholders, to ensure higher retention of policyholders. Random Forest algorithm was found to outperform other algorithms in terms of accuracy, recall, and F1 Score. These were the main evaluation metrics used to evaluate model performance. On that note, the following recommendations are made to assist other researchers in their future research. The recommendations for individuals, insurance organizations, underwriters, other insurance agencies, and researchers are as follows:

- Life insurance companies should consider; policy type, length of coverage of the insured, entry age, premiums, sex, and expected benefit of the insured before issuing insurance coverage to prospective proposers.

- Insurance companies and underwriters should apply machine learning techniques during their underwriting works to determine accurately the probability of a prospective insured surrendering from the insurance coverage.

- Machine learning experts should combine comprehensive data analysis and ML models to better understand the complex structure and dynamics within insurance data to improve quality risk assessment.

- Researchers, machine learning experts, and AI enthusiasts should consider comparing several machine learning techniques and other advanced methods in determining the most performing models before making their final decision.

References

Milhaud, X., Loisel, S., & Maume-Deschamps, V. (2010). Surrender triggers in life insurance: classification and risk predictions. Laboratoire de Sciences Actuarielle et Financiere (Working Paper).

Tiwari, A., Hadden, J., & Turner, C. (2010). A new neural network-based customer profiling methodology for churn prediction. In Computational Science and Its Applications–ICCSA 2010: International Conference, Fukuoka, Japan, March 23-26, 2010, Proceedings, Part IV 10 (pp. 358-369). Springer Berlin Heidelberg.

Eling, M. and Kochanski, M. (2013), "Research on the lapse in life insurance: what has been done and what needs to be done?", Journal of Risk Finance, Vol. 14 No. 4, pp. 392-413. https://doi.org/10.1108/JRF-12-2012-0088

Russell, D. T., Fier, S. G., Carson, J. M., & Dumm, R. E. (2013). An empirical analysis of life insurance policy surrender activity. Journal of Insurance Issues, 35-57.

Eling, M., & Kiesenbauer, D. (2014). What policy features determine life insurance lapse? An analysis of the German market. Journal of Risk and Insurance, 81(2), 241-269.

Barsotti, F., Milhaud, X., & Salhi, Y. (2016). Lapse risk in life insurance: Correlation and contagion effects among policyholders' behaviors. Insurance: Mathematics and Economics, 71, 317-331. Babaoglu, C., Ahmad, U., Durrani, A., & Bener, A. (2017, October). Predictive modeling of lapse risk: An international financial services case study. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 16-21). IEEE.

Aleandri, M. (2017). Modeling dynamic policyholder behaviour through machine learning techniques. Submitted to Scuola de Scienze Statistiche. Zhang, R., Li, W., Tan, W., & Mo, T. (2017). Deep and shallow model for insurance churn prediction service. In 2017 IEEE International Conference on Services Computing (SCC) (pp. 346-353). IEEE.

Russo, V., Giacometti, R., & Fabozzi, F. J. (2017). Intensity-based framework for surrender modeling in life insurance. Insurance: Mathematics and Economics, 72, 189-196.

Binder, S., & Mußhoff, J. (2017). Global Insurance Industry Insights. An in-depth perspective McKinsey Global Insurance Pools.

Stehno, C., Guszcza, J., Batty, M., Tripathi, A., Kroll, A., Wu, C. P., ... & Katcher, M. (2018). Predictive Modeling for Life Insurance. Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4(2), 145-154.

Milhaud, X., & Dutang, C. (2018). Lapse tables for lapse risk management in insurance: a competing risk approach. European Actuarial Journal, 8, 97-126.

Santharam, A., & Krishnan, S. B. (2018). Survey on customer churn prediction techniques. International Research Journal of Engineering and Technology, 5(11), 3.

Miguel Ángel de la Llave, Fernando A. López & Ana Angulo (2019) The impact of geographical factors on churn prediction: an application to an insurance company in Madrid's urban area, Scandinavian Actuarial Journal, 2019:3, 188-203, DOI: 10.1080/03461238.2018.1531781

Xong, L. J., & Kang, H. M. (2019). A comparison of classification models for life insurance lapse risk. International Journal of Recent Technology and Engineering, 7(5), 245-250.

De la Llave, M. Á., López, F. A., & Angulo, A. (2019). The impact of geographical factors on churn prediction: an application to an insurance company in Madrid's urban area. Scandinavian Actuarial Journal, 2019(3), 188-203. DOI: 10.1080/03461238.2018.1531781

He, Y., Xiong, Y., & Tsai, Y. (2020). Machine learning based approaches to predict customer churn for an insurance company. In 2020 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-6). IEEE.

Anuganti , (2020). Confusion matrix, Analytics Vidhya: What is a confusion matrix? Everything you Should Know about — by Anuganti Suresh — Analytics Vidhya —

Medium and Published On 18/05/2021 and Last Modified On 01/07/2021, accessed on 28/07/2023 at 10:20 pm

Sahlol, A. T., Abd Elaziz, M., Al-Qaness, M. A., & Kim, S. (2020). Handwritten Arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set. IEEE Access, 8, 23011-23021.

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

Akyildirim, E., Cepni, O., Corbet, S., & Uddin, G. S. (2021). Forecasting mid-price movement of Bitcoin futures using machine learning. Annals of Operations Research, 1-32.

Hu, S., O'Hagan, A., Sweeney, J., & Ghahramani, M. (2021). A spatial machine learning model for analyzing customers' lapse behavior in life insurance. Annals of Actuarial Science, 15(2), 367-393.

Kgare, M. (2021) Predicting lapse rate in life insurance using machine learning algorithms (Doctoral dissertation).

Whitepaper AI in the enterprise, (2021): Unleashing opportunity through data: what is machine learning? — IBM: https://www.ibm.com/topics/machine-learning, accessed on 20/06/2023 at 10:40 am

Reck, L., Schupp, J., & Reuß, A. (2022). Identifying the determinants of lapse rates in life insurance: an automated Lasso approach. European Actuarial Journal, 1-29.

Kiermayer, M. (2022). Modeling surrender risk in life insurance: theoretical and experimental insight. Scandinavian Actuarial Journal, 2022(7), 627-658.

Huang, W. T., Luo, S. N., & Chen, L. P. (2022, November). An Application of Machine Learning to Policyholder Behavior Prediction System Design. In 2022 10th International Conference on Orange Technology (ICOT) (pp. 1-4). IEEE.

Azzone, M., Barucci, E., Moncayo, G. G., & Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. Expert Systems with Applications, 191, 116261.

Alcaide, D. D. C. (2023). Predicting Lapse Rate in Life Insurance: An Exploration of Machine Learning Techniques (Doctoral dissertation). Machine learning (2023, July 18) In Wikipedia:

https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1167534378

Somasundaram, A., & Reddy, U. S. (2016, September). Data imbalance: effects and solutions for classification of large and highly imbalanced data. In international conference on research in engineering, computers and technology (ICRECT 2016) (pp. 1-16).

APPENDIX

Table 5.1: Data description after standardization

FEATURES	MEAN	ST. DEVIATION	MINIMUM	MEDIAN	MAXIMUM
CHANNEL1	0	1	-1.52	0.882	1.842
CHANNEL2	0	1	-2.606	-0.823	0.96
CHANNEL3	0	1	-0.71	-0.64	5.043
ENTRY AGE	0	1	-1.843	-0.18	3.439
SEX	0	1	-1.224	0.817	0.817
POLICY TYPE 1	0	1	-1.375	-0.169	3.208
POLICY TYPE 2	0	1	-1.27	0.203	2.435
BENEFIT	0	1	-0.427	-0.181	61.163
NUMBER OF ADVANCE PREMIUM	0	1	-0.013	-0.013	171.056
INITIAL BENEFIT	0	1	-0.088	-0.088	60.655
POLICY YEAR (DECIMAL)	0	1	-1.648	0.049	2.878
POLICY YEAR	0	1	-1.386	0.309	3.134
PREMIUM	0	1	-0.227	-0.151	74.98

Visit this link https://github.com/Ibrahim-Hak/Final_Year_Project for the code sample and data.