



Leveraging Data Analytics for Lapse Reduction in Life Insurance

**A dissertation submitted for the Degree of Master of
Business Analytics**

**W.M.U.C. KAPILABANDARA
University of Colombo School of Computing
2023**


Leveraging Data Analytics for Lapse Reduction in Life Insurance

**W.M.U.C. KAPILABANDARA
2023**

Declaration

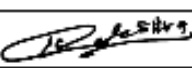
Name of the student: W.M.U.C.Kapilabandara
Registration number:2019/BA/012
Name of the Degree Programme: Master of Business Analytics
Project/Thesis title: Leveraging Data Analytics for Lapse Reduction in Life Insurance

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	Date (DD/MM/YYYY)
	16/10/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	Supervisor 2	Supervisor 3
Name	Dr. L.N.C. De Silva		
Signature			
Date	16/10/2024		

I would like to dedicate this thesis
to my parents for their eternal love and encouragement,
to my beloved husband Udara for believing in me and supporting me in everything.

Acknowledgment

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Lasanthi De Silva from the University of Colombo School of Computing, for her invaluable advice, support, and constant supervision throughout the research.

And also, I would like to acknowledge my department supervisor from work, my company CEO, and the group IT CEO for granting me access to the information. This research could not have been accomplished without their trust and kindness.

Finally, I thank my loving husband, Udara, my parents, and my sisters for the encouragement given over the entire period to make this research successful.

Abstract

Insurance companies operate on the concept of pooling losses among their insureds. An insurer invests the premiums to earn enough money not only to pay for losses but also to operate the company and gain a profit. Thus, the insurance company must reasonably predict the payments that will be made for loss and charge affordable premiums to insure a risk. The term "lapse" refers to the termination of an insurance policy by the policyholder for any reason other than the death of the policyholder. When an insurance policy lapses, it will decrease the performance of the product, the initial year's expense of the policy may not be covered and it will create a loss of public image.

Since retaining existing customers is much cheaper and more profitable than getting a new customer, it is crucial to identify policies which are likely to lapse. Even though the insurance industry uses a number of mathematical, statistical, and financial concepts to understand the behaviour of policyholders and quantify future liabilities and risks, those approaches have major drawbacks.

This study focuses on predicting individual policyholder lapse rate and identifying scenarios which reduce lapses in Sri Lankan Insurance industry. To conduct this, firstly data of policyholder need to be collected and analysed. This information is gathered from an Insurance Company. Policies that commence from 2013 to 2022 are included. 32 parameters are considered for the analysis and variable importance is calculated. Then Random Survival Forest (RSF) and Cox net Survival Analysis are used to predict the lapse rate. Those techniques let the model to construct survival functions with different shapes for each insured.

Model performance is high in random survival forest compared to cox net survival analysis as it captures linear, non-linear relationships and interactions between many factors. Hence variable importance is calculated using random survival forest. Then scenarios are performed to identify policy characteristics which give low lapse rate, in other words high survival rate.

By the findings, it was successfully concluded that, through machine learning teachings, insurance companies will not only be able to predict the lapse rate of individual policyholders but also be able to identify policy characteristics that give a high survival rate.

Table of Contents

Acknowledgment.....	ii
Abstract.....	iii
List of Figures.....	v
List of Tables.....	vii
Chapter 1 Introduction.....	1
1.1 Introduction to the problem	1
1.2. Motivation	2
1.3 Aims and Objectives.....	3
1.4 Scope of the study.....	3
1.5 Dataset	4
1.6 Structure of the Thesis.....	4
Chapter 2 Literature Review & Background.....	5
2.1 Background.....	7
2.1.1 Survival Analysis.....	8
2.1.2 Statistical Models for Survival Analysis	9
2.1.3 Machine Learning Models for Survival Analysis.....	11
2.1.4. Model Evaluation	12
Chapter 3 Methodology	13
3.1 Proposed Approach & Methodology	13
3.2 Research Solution Design.....	14
3.2.1 Data.....	14
3.2.2 Algorithms	14
3.2.3 Model Evaluation	14
3.2.4 Identification of most important features for lapse prediction.	15
3.2.5 Performing scenarios	15
Chapter 4 Results and Evaluation.....	15
4.1 Data Pre-processing	15
4.1.1 Load Dataset	17
4.1.2 Exploratory Data Analysis (EDA).....	18
4.1.3 Data Visualization	22
4.2 Predicting Policy lapsation	30
4.2.1 Random survival forest.....	30
4.2.2 Cox net Survival Analysis	31
4.2.3 Model Evaluation	31
4.2.3.1 C-index	31
4.2.3.2 Time-dependent area under the ROC curve	32
4.3 Important Features.....	32
4.4 Performing scenarios to identify policy characteristics.....	33
4.5 Discussion.....	37
4.5.1 Challenges	37
4.5.2 Assumptions	38
Chapter 5 Conclusion	38

List of Figures

Figure 2.1:Cardiovascular Disease Clinical Study -Scikit-survival	8
Figure 2.2:Survival Curve Graph Example	9
Figure 3.1:Proposed Process.....	14
Figure 4.1:dataset sample	17
Figure 4.2:data types.....	18
Figure 4.3:Summary Statistics 1	18
Figure 4.4:Summary Statistics 2.....	19
Figure 4.5:Summary Statistics 3.....	19
Figure 4.6:Null values	19
Figure 4.7:Null values as a percentage	20
Figure 4.8:Zone wise percentage of policy count.....	20
Figure 4.9:percentage of policy count based on policy status	21
Figure 4.10:percentage of policy count based on premium frequency.....	21
Figure 4.11:percentage of policy count based on premium band.....	21
Figure 4.12:duplicate values.....	21
Figure 4.13:percentage of policy count on policy status after conversion	22
Figure 4.14:proportion of observations of the policy status	22
Figure 4.15:proportion of observation of the categorical variables on policy status in sample 1	23
<i>Figure 4.16:proportion of observation of the categorical variables on policy status in sample 2</i>	<i>24</i>
Figure 4.17:proportion of observation of the categorical variables on policy status in sample 3	25
Figure 4.18:Distribution of policy duration by term, age & lapse in the sample	26
Figure 4.19:correlation matrix	27
Figure 4.20:correlation matrix after removing correlated variables.....	28
Figure 4.21:survival curve of the sample	28
Figure 4.22:survival curve by channel code of the sample	29
Figure 4.23:survival curve by premium frequency of the sample	29
Figure 4.24:survival curve by product type of the sample	29
Figure 4.25:data transformation	30
Figure 4.26:data splitting	30
Figure 4.27:train data using random survival forest.....	30
Figure 4.28:model performance	30
Figure 4.29:model prediction -random survival forest.....	30
Figure 4.30:predicted survival curve	31
Figure 4.31:Fit the model using Cox net survival analysis	31
Figure 4.32:Model performance-Cox net survival analysis	31
Figure 4.33:C-index of selected algorithms	31
Figure 4.34:calculating time-dependent AUC.....	32
Figure 4.35:time-dependent AUC	32
Figure 4.36:Finding Important features.....	32
Figure 4.37:survival rates	34
Figure 4.38:Effect of policy terms on survival rate.....	34
Figure 4.39:Effect of premium frequencies on survival rate.....	35
Figure 4.40:Effect of sum assured on survival rate	35
Figure 4.41:Effect of premium frequencies on survival rate 2.....	36
Figure 4.42:Effect of sum assured on lapse rate 2.....	36

Figure 4.43:effect of product type/riders on survival rate	37
Figure 4.44:Effect of age on survival rate	37

List of Tables

Table 4.1: Variables of Data Set.....	16
Table 4.2: Variables of Data Set.....	17
Table 4.3: Important features.....	33

Chapter 1 Introduction

1.1 Introduction to the problem

Insurance refers to a contractual agreement between an individual or an entity, known as the insured, and an insurance company, referred to as the insurer. It is a financial arrangement that protects against potential financial losses or damages resulting from specified risks or events. In this arrangement, the insured pays a regular premium to the insurer in exchange for the assurance that if a covered loss occurs, the insurer will compensate the insured according to the terms and conditions outlined in the insurance policy. Insurance aims to lighten the financial impact of unexpected events and provide peacefulness to the insured party (Ch and Ramesh, 2011).

Insurance companies operate on the concept of pooling losses among their insureds. The risks of loss that insurers are willing to cover are called insurable risks. An insurer pools money from a group of people (insureds) to reimburse the insured who suffers a loss covered by an insurance policy. An insurer invests the premiums to earn enough money not only to pay for losses but also to operate the company and gain a profit. Thus, the insurance company must reasonably predict the payments that will be made for loss and charge affordable premiums to insure a risk (Ch and Ramesh, 2011).

There are two main classes of insurance, namely life, and non-life. In life insurance, the insurer agrees to pay a sum upon the insured's death or other events, such as illnesses, for the payments (premiums) made by the insured at regular intervals or in lump sums. On the other hand, non-life insurance deals with services other than those for the lives of human beings. Here, the focus area is life insurance (Ch and Ramesh, 2011).

The term "lapse" refers to the termination of an insurance policy by the policyholder for any reason other than the death of the policyholder. The coverage will be lost because the policyholder failed to pay the premium. The term "surrender" is used when the surrender value is paid when the policyholder terminates the policy (Xong and Kang, 2019).

This happens as the result of allowing policyholders to choose among many options. For example, a policyholder may receive a surrender value if he continues the policy only for three years and then terminates it or choose to discontinue premium payments without any payment if he terminates the contract before three years (Xong and Kang, 2019).

When an insurance policy lapses, it affects the company in below ways.

- The performance of the product is highly relying on the persistence of the business. Persistency is the measure of how long policies remain in force. The persistency rate is the number of policies in force at the end of a given year as a ratio to the number of policies at the beginning of the year. This arrives after deducting cancelled, lapsed, or ceded insurance policies. Therefore, the higher the persistency rate, the higher the product performance. The product policies that have been in force for a long time are more profitable for the insurance company than policies that lapse quickly.
- Also, the initial year's expense of the policy is very high compared to the premium paid in the same period for the life insurance company due to the payment of high commissions to the agent, stamp duty, fixed costs, administration costs, etc. Therefore, insurance companies tend to profit from a policy after the completion of three (3) years. If a policy lapses in the first three years of the policy term, it is a significant loss for the company (Ch and Ramesh, 2011). Furthermore, the increase in the rate of lapsation creates a loss of public image. It will result in the lapse of even more policies. And also, it will create disinterest in the public towards choosing the policies of the company. Thus, it will harm new businesses as well.

Customer satisfaction is vital for life insurance companies since there are now so many more players in the competition than in the past, and customers can easily change insurers. Also, retaining existing customers is much cheaper and more profitable than getting a new customer, as some expenses have already been covered by previously paid premiums (Verhoef and Donkers, 2001b). From the perspective of an insurer, when policies lapse, the business will deteriorate due to a lack of accuracy in their estimates, as lapse risk accounts for capital requirements in life insurance. Therefore, insurers should thoroughly understand lapsation to define the reasonable capital standard (Ch and Ramesh, 2011). Considering all the above facts, a suitable mechanism should be identified for the lapse reduction of life insurance policies.

1.2. Motivation

The insurance industry possesses vast amounts of data and is in the process of modernizing its core systems. However, it has yet to fully harness the potential of this data. Today, the insurance industry has opportunities to leverage data in new ways. By using data analytic techniques, insurance companies can gain knowledge about data and improve business decisions.

For the reduction of policy lapsation, it is important to learn about policyholders and their behaviour. The insurance industry uses a number of mathematical, statistical, and financial concepts to understand the behaviour of policyholders and quantify future liabilities and risks. Those approaches have major drawbacks, like not accounting for the value that different policyholders place on certain features (number and type of fund choices available within a life insurance policy or annuity contract, liquidity versus guarantees), and not accounting for how strongly social, and emotional factors influence policyholders' financial decisions (job insecurity and the need for liquidity)(Lombardi and Paich, n.d.).

Data analytic techniques like predictive modelling give a better understanding of policyholder behaviour as they consider the interactions between many factors that influence a policyholder's decisions (Devale, 2012). Also, from data analytics techniques, hidden information can be revealed, and that information will help insurance companies retain existing customers and acquire new customers. As considering above benefits of data analytic techniques, usefulness of using those for reducing lapses in company by targeting right policies is shown in this study.

1.3 Aims and Objectives

Main aim of this study is to predict individual policyholder lapse rate and identify policy characteristics which reduce lapses in Sri Lankan Insurance industry. To achieve the above-mentioned target, below objectives have to be met.

- Collect and analyse data related to policyholder characteristic from the insurance company.
- Enhance customer acquisition and retention strategies through exploring a variety of data analytics models Experiment with different algorithms to build a predictive model.
- identify the most influential attributes for accurately predicting lapsation.
- Evaluate the potential performance of the deployed model.

1.4 Scope of the study

This project will provide reasons for policy lapsation and identify policies that are about to lapse, which will help management take the necessary actions to reduce policy lapsation.

Policies that are about to lapse are identified through survival models such as a random survival forest and cox net survival analysis. The performance of the model will be assessed via concordance index (C-index) and time-dependent area under the receiver operating characteristic curve (ROC).

In this project, data from only one insurance company is considered, and only attributes made available by the company will be used for modelling. Unlabelled data will be omitted from the analysis. Policy statuses (In-force/Lapse etc.) are considered as at a particular date.

1.5 Dataset

Insurance company dataset with the company consent, is considered for the study. Policies that commence from 2013 to 2022 are included. Around 30 parameters are considered for the analysis and variable importance is calculated.

1.6 Structure of the Thesis

The structure of the thesis is organized into several chapters. Chapter 2, titled "Literature Review & Background" provides discussion on previous studies, essential contextual information, theories, and concepts necessary for comprehending the research context. Chapter 3 titled "Design & Methodology" provides in detail view of methods, models, and techniques, along with the design. Chapter 4 titled "Results and Evaluation" presents the obtained results and evaluations of the results. And finally, chapter 5 titled "Conclusion" presents the conclusion arrived based on the study.

Chapter 2 Literature Review & Background

Life insurance has always been a competitive industry. Applying data analytics to the field is an added advantage for competition.

Devale and Kulkarni discussed how data mining can help firms get business advantages to support decision-making and the importance of data mining to get involved in the market. And also, it pointed out how association rules usable to retain existing customers by finding the necessary combination, and then new policies can be sold to existing customers to retain them. Furthermore, they discussed how the K Nearest Neighbours algorithm can be used to find a customer segment (Devale, 2012).

Rao, in the paper, discussed how customers can be acquired and retained by using data analytics techniques and examined data analytics models for that purpose. Rao also discussed methods for reducing policy lapsation. In the paper, it says that the most profitable customers need to be identified to meet the target, and they point out a model for predicting customer life-time value. Furthermore, it pointed out the common barriers to employing predictive analysis. The barriers that are mentioned in the article are start-up costs, processing expenses, interoperability, cultural constraints, and lack of expertise (Rao, n.d.).

Verhoef and Donkers pointed out that keeping existing customers is more profitable than attracting new ones. Also, they pointed out that by using customer information contained in a database, insurance companies can identify customers who are valuable to the company. According to the article, there are four major segments: low potential and low current value; high potential and low current value; low potential and high current value; high potential and high current value. In this paper, they have mainly focused on the modelling of customer potential value. They have discussed and compared different statistical models for that purpose (Verhoef and Donkers, 2001).

Checcacci, in his paper, pointed out how variables like age, gender, unemployment, loss of spouse, retirement, and divorce affect the lapse decision, and he mainly discussed how an agent-based model can be used by the insurance industry for modelling lapse risk. He pointed out that an agent-based model is the most suitable model to explain the interaction between policyholder choices, the insurer's attempts to retain customers, and the state of the economy in general. And also, general linear models, which are mentioned in various research papers, have been discussed. Those models are the Negative Binomial Model, Logit Regression, Poisson Models,

and Logistic Regression for Voluntary Lapse Decision. There, it made comparisons between those research papers (Matematico-Statistiche, n.d.).

The American Risk and Insurance Association pointed out the factors affecting early termination and the impacts of that on insurance. The reasons that have been pointed out in the article are: a riskier occupation; collecting premiums through an insurer's agent; and providing periodic living benefits. Furthermore, it demonstrates how variables such as age, gender, policy term, channel, occupation, commission ratio, premium payment mode, and product type influence early termination (Insurance Association, n.d.).

Chandra and Ramesh discussed the situations that create high lapse rates. They pointed out that when considering premium-paying modes, policies that are in monthly and quarterly modes have higher lapse rates. And also, when considering the premium-paying term, policies with terms between 0 and 10 have the highest lapse rates. Furthermore, doing sales through brokers and corporate agents causes the policies to become more lapsed (Ch and Ramesh, 2011).

Estany, Marín, and Zanón, in their article, mentioned the importance of considering the different policies owned by the same customer together. They also analysed the profit generated by each customer based on three dimensions. There are three types of profit: historical profit, which accumulates over time; prospective profit, which is generated if the customer does not cancel the policy; and potential profit, which is generated when the customer purchases a new policy that differs from the current one. They have also proposed a method for determining expected losses due to policy cancellation. It is done by estimating the profit generated by each policyholder and then predicting their probability of cancellation (Guillén et al., 2011).

Xong & Kang, in their article, compared four classification algorithms, such as logistic regression, k-nearest neighbour, neural network, and support vector machines, for constructing life insurance lapse risk assessments. They found that both SVM and NN produced high prediction accuracy (Xong and Kang, 2019).

Oshini discussed in her paper factors that affect policy early termination and methods to identify a subset of customers who have a high probability of early termination. She has used classification methods such as decision trees and neural networks for her analysis (Goonetilleke and Caldera, 2013).

Karyn Azevedo mentioned that to evaluate time-to-event data, it is required to have a specific set of mathematical tools capable of dealing with the behaviour of survival data. Therefore,

survival analysis will be a good approach for such data. Karyn had used survival analysis on kidney transplant data. And also, Karyn stated that for a study, it is important to select the right set of features to get good results (de Azevedo, n.d.).

Hemant Ishwaran introduced an ensemble tree method for the purpose of analyzing right-censored survival data, which is random survival forests (RSF). Ishwaran had extended Breiman's random forests method to derive random survival forests. In that paper, independent bootstrap samples are used to grow each tree with a randomly selected subset of variables at each node. Then, using a survival criterion that involved survival time and censoring status, a node is splitted. And also, the Nelson-Aalen estimator is used to estimate Cumulative Hazard Function (CHF) for terminal node. The ensemble is derived by averaging these CHFs. Ishwaran had stated that RSF is convenient for applying to real data, which has complex interrelationships between variables. (Ishwaran et al., 2008)

Jorge Andrade applied machine learning and traditional survival analysis techniques to insurance data for the purpose of modelling the lapse rate. In that paper, traditional Cox Proportional Hazards (CPH), Random Survival Forests (RSF) and Conditional Inference Forests (CIF) machine learning models are compared. It demonstrates the better performance of machine learning algorithms for predicting survival function using the C-index and Brier Score compared to traditional survival analysis. (Andrade et al., 2021)

In the above papers, they discussed targeting new customers and preventing early termination. Factors affecting lapsation can vary from company to company and region to region. The above papers have discussed the factors in general. Some variables, like payment method, which can be directly affected by lapsation, are not considered for most of the papers. And also, it shows the importance of using survival analysis techniques in the insurance industry.

The research aims to predict policy lapsation using survival analysis techniques and identify factors that will reduce policy lapsation for Sri Lankan insurance companies. And also, to give the insurance company knowledge on what type of policies they focus on when commencing new business.

2.1 Background

Below section described the theories which needed for the study

2.1.1 Survival Analysis

Survival analysis is used to evaluate time to event. In our study, the event means the lapsation of the policyholder. It is very unexpected as for the study period, it may occur or not occur. As we can't exactly say the time which occur the event of interest, data is incomplete.

This happens because studies are limited in time with a start date and an end date. This will result in a limited time frame for the event to be observed. And also, the participant will die or drop out of the study before the event occurs, which will cause incomplete data (de Azevedo, n.d.).

The picture 3.1 illustrates clinical study that investigates cardiovascular disease. It has been carried out over a 1-year period (Scikit-survival user guide)

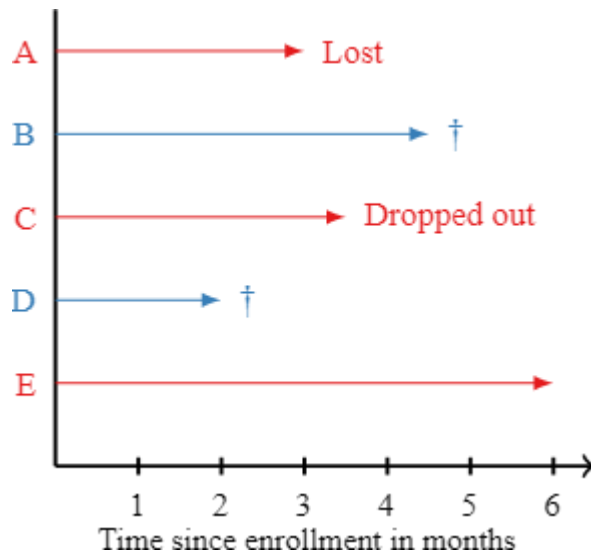


Figure 2.1: Cardiovascular Disease Clinical Study -Scikit-survival

(Source: Scikit-survival user guide)

The patients B and D both had events of interest observed during the study time frame. It was regardless of the enrolment time. These are the complete data, as events of interest occurred during the investigation period. These data are known as uncensored data. The patient A has lost, and the patient C dropped out during the study before the event of interest happened. The patient E was observed during the entire study, but the study period ended before the event occurred. So, for patient E, it is unknown whether he experienced or did not experience the event. Therefore, patients A, C, and E are event-free up to their last follow-up. That means the data is incomplete or censored.

Even though patents A, C, and E have not experienced the event of interest, that information is useful to estimate the likelihood of the event. Survival analysis techniques use not only uncensored data but also censored data (de Azevedo, n.d.).

Prediction of Survival Analysis consists of predicting survival function.

Survival probability $S(t)$ returns the probability of survival time being greater than certain time(t), given a random variable corresponding to a patent's survival time T . This gives the probability of surviving beyond t where $t \geq 0$.

$$S(t) = P(T > t) \quad (3.1)$$

Then the survival function, also known as the cumulative survival rate can be obtained as 3.2

$$S(t) = 1 - P(T > t) \quad (3.2)$$

Here $S(t)$ is the cumulative survival rate. It is a non-increasing function with a probability of surviving that goes to 0 as time goes to infinity (de Azevedo, n.d.).

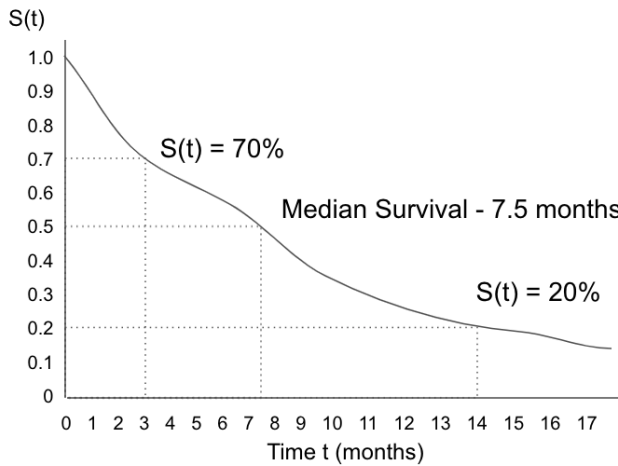


Figure 2.2: Survival Curve Graph Example

(Source: (de Azevedo, n.d.))

$S(t)$ can be estimated by 3.3. If the data is complete, which means everyone dies before the study ends, we can observe the exact survival time of all subjects.

$$S(t) = \frac{\text{number of patients surviving beyond } t}{\text{total number of patients}} \quad (3.3)$$

This cannot be used for censored data as the numerator is not always defined (Scikit-survival user guide).

2.1.2 Statistical Models for Survival Analysis

Below described the most used techniques for survival analysis, which are Kaplan–Meier method and Cox proportional hazards regression.

Kaplan-Meier estimator

Kaplan-Meier's approach has uses for performing survival analysis that work for censored data.

The following assumptions are made under this approach:

- Censoring does not impact the probability of developing the event of interest, i.e., censored and non-censored individuals have the same survival probability.
- Subjects with different enrolment times (i.e., later or earlier enrolment) in the study have the same survival probabilities;
- The time of the event happens at the specified time.

To define the estimator using the Kaplan-Meier survival function, we only need two variables, which are time to the event and censored subjects. However, in many studies, it is also interesting to evaluate the influence of other variables on survival time (de Azevedo, n.d.).

Cox Proportional Hazards (CPH)

One of the most widely used functions of survival analysis is Cox Proportional Hazards (CPH). The assumption of proportional hazards made by the survival function $S(t | X_i)$ is adjusted by applying the classic Cox model technique. The adjustment is made considering the model assumptions and using the linear estimation model for the log of the hazard rate based on an exponential distribution (Andrade et al., 2021).

For this model, all the covariates are added to the final model without considering their relevance. Therefore, this model can be used for a smaller set of features, whereas for a large set of features, this will result in a non-singular matrix due to the higher risk of correlation between features (de Azevedo, n.d.).

Penalized Cox Models

To overcome the problem of getting a non-singular matrix for CPH, penalized cox models have been used. There are three types of penalized Cox models: Ridge, LASSO, and Elastic Net (de Azevedo, n.d.).

In Ridge's Penalized Cox Model, it uses a penalty variable that has the ability to shrink certain coefficients to zero. Only less relevant variables will get the penalty from this approach, but it does not exclude the redundant features. Since this approach has a high computational cost, the LASSO approach is introduced, which proposes to continuously select the most predictive subset of features. This method also has drawbacks, which are the inability to select a number

of features greater than the sample size and the random selection of a feature when the model experiences a set of features closely related (de Azevedo, n.d.).

2.1.3 Machine Learning Models for Survival Analysis

Random Survival Forests (RSF)

Random forests are one of the most important machine learning techniques for classification and regression. An advantage of this model is that it's totally non-parametric. This also captures linear and non-linear relationships between the explained variable and the predictor variables. Another important feature is that it finds interactions between covariates because the learning comes from the ensemble decision trees. Outliers in data do not affect it, nor do they suffer from convergence problems.

Survival trees are built by splitting each parent node into two daughter nodes starting at the root, which comprises the full dataset. A split is performed according to a survival criterion that maximizes the difference between daughter nodes; such a split is repeated on each subsequent node in a binary manner. This process is repeated to build n trees, and then ensemble techniques are used to obtain the final estimators, in this case the average of all trees.

The algorithm has double randomness. First, a random sample is obtained by replacing the original data in each new tree. Second, the parent node is split into two daughters using a randomly selected covariate x_j . Due to the law of large numbers, this double randomness leads to the convergence of the prediction error (PE). It describes the number of trees where the PE converges with a higher accuracy (Andrade et al., 2021).

The algorithm is described below.

The first step is to draw bootstrapping samples of average size 63% of the original dataset. The unselected data, which is 37% of the original data, is called out-of-bag (OOB) data (de Azevedo, n.d.).

In the second step, a survival tree is grown for each sample, and at each tree node, specific number of covariates are randomly selected. Here, that node is split based on the covariate and its value, which gives the maximum survival difference between the daughter nodes of it (Andrade et al., 2021).

Finally, the tree is expanded to its maximum size, provided that each leaf should have defined event cases (Andrade et al., 2021).

Once the Random Forest is built, the cumulative hazard function (CHF) is calculated for each terminal node and averaged over the bootstrap samples.

2.1.4. Model Evaluation

To evaluate the quality of the model, prediction error will be calculated using the Concordance Index (C-index) (Andrade et al., 2021).

Concordance index (C-index)

This measures how well the predictors rank the two randomly selected policyholders with respect to survival. This is equivalent to the area under the curve (AUC) (Andrade et al., 2021).

The C-index is defined as the ratio of concordant pairs to comparable pairs. Here, concordant pair means correctly ordered pair (Scikit-survival user guide).

The higher the C-index score, the higher the model performance is (Andrade et al., 2021).

Time-dependent area under the receiver operating characteristic curve (ROC)

This is an extended version of the receiver operating characteristic curve (ROC curve) to cater to censored survival times. The ROC curve is a commonly used performance measure for binary classification tasks. ROC curves are used to find how well estimated risk scores can separate a false positive rate from a true positive rate (Scikit-survival user guide).

From ROC, for a specific time t , how well a predictive model can differentiate subjects who will experience an event by time t from those who will not by time t can be estimated. But this needs to be estimated for a given list of time points. A time-dependent area under ROC is implemented as an estimator for such a case (Scikit-survival user guide).

Chapter 3 Methodology

3.1 Proposed Approach & Methodology

This section explains the proposed approach and methodology, which involve data preparation, analysing data, lapse rate prediction for individual policy holder using statistical methods/machine learning techniques and identifying scenarios which improve survival rate for individual policy holder.

This research mainly aims to reduce lapses in the life insurance domain. Insurance data for a specific period is collected and analysed for research purposes.

The data set consists of policies that commenced from May 28, 2012, to December 31, 2021. Policy status has been obtained as of December 31, 2018 and as of December 31, 2021. Policy states on two specific dates have been collected to check the impact of the economic condition in Sri Lanka on policy status. This will impact the most important features for lapse reduction as policy holder's needs may have changed and they may not need the policy.

The original dataset for the study contains 32 variables and 116,543 records. A sample is constructed to build the predictive model, as the large volume affected the performance of the laptop using the stratified sampling method. Then analyse the new dataset to see whether it has the same characteristics as the original dataset and to gain insights about the data.

The built predictive model using the above data set will be a basic model as customer sensitive information such as salary range and agent information is not available in the dataset.

The proposed methodology consists of two phases: building a predictive model to identify the most important variables for policy lapsation and identifying scenarios that improve policy survival by using the built predictive model.

To build the predictive model, the study will utilize random survival forest, and Cox net survival analysis. Finally, based on the selected two approaches, the model will be evaluated using the C-index and time-dependent area under the ROC curve. Results will be performed on the selected model. Then the most important features will be identified, considering policy status at two specific dates.

After the model building phase, several scenarios are performed to identify the factors that improve policy survival. This will be beneficial for the company to issue policies with a longer survival period.

Selected Tools

- Jupyter Notebook (Anaconda 3)
- Libraries: Pandas, Numpy, Matplotlib, Scikit-learn, Scikit-survival

Below graph shows the above-described process.

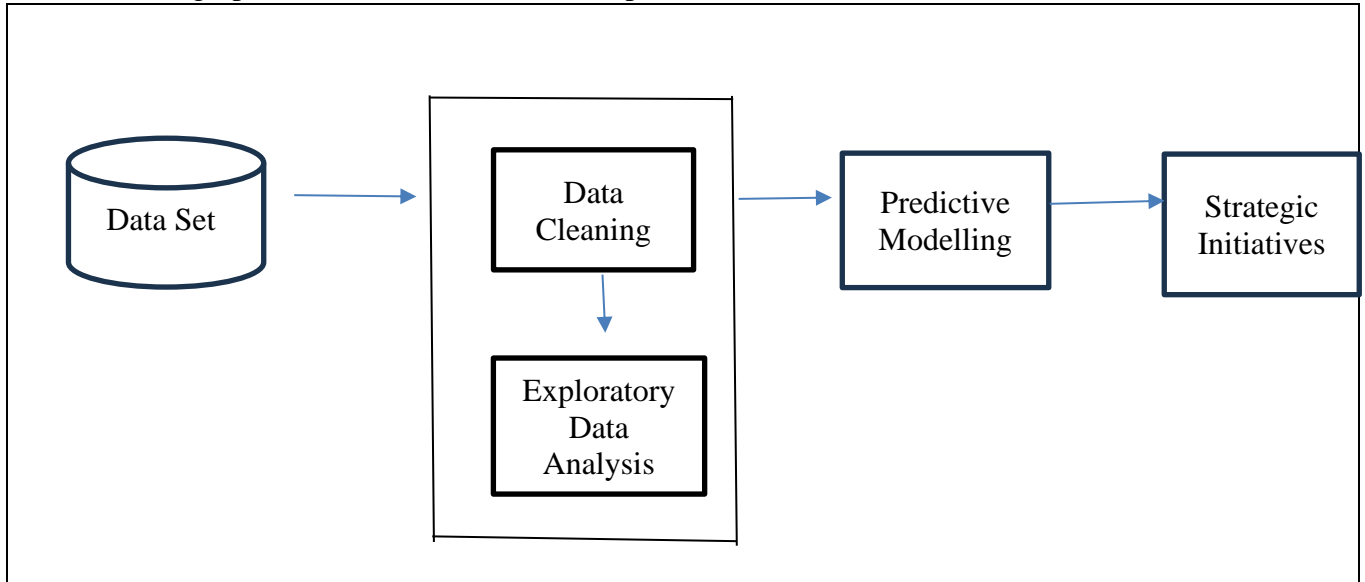


Figure 3.1:Proposed Process

3.2 Research Solution Design

Below summarized the solution components.

3.2.1 Data

The insurance policy data was collected from an insurance company with the company's consent. Exploratory data analysis techniques are used to clean up and get an overall idea of the portfolio.

3.2.2 Algorithms

Two approaches have been used to build predictive models. One is a machine learning algorithm and the other one is traditional survival analysis techniques. These are Random Survival Model and Cox net Survival Analysis respectively.

3.2.3 Model Evaluation

To evaluate the model, two techniques have been used. Those are,

- C-Index
- Time dependent Area under the curve

3.2.4 Identification of most important features for lapse prediction.

After selecting the model based on the evaluation criteria, the most important features will be identified based on mean, std value as there will be many features to consider for the research and not all won't give much contribution to the lapse prediction.

3.2.5 Performing scenarios

After identifying the most important variables from the selected model, several scenarios are performed to identify characteristics of a policy which gives a higher survival rate than the existing survival rate.

Chapter 4 Results and Evaluation

4.1 Data Pre-processing

Data Set

A Life insurance company data was used for this study. It contains the characteristics of the policyholders covering a period of 9 years from 2012 to 2021.

Below table summarized the variables which used for the study. Total datapoints/policies available for the study is 116, 543 and total number of variables available for the study is 32.

Some variables which can be used to identify certain policyholder are removed from the study to due to the sensitive nature of those variables.

Variable	Description
ID	Policy number (integer numbers are put here, removed the actual number due to sensitivity of the data)
V_PLAN_CODE	The code of the product which a policyholder has bought
V_PYMT_DESC	How often policyholder should pay the premium (Premium frequency). There are 4 types. MONTHLY, HALF YEARLY, QUARTERLY & YEARLY.
N_TERM	Term of the policy
COMMENCEMENT_YEAR	Year of the policy commencement
V_REL_CODE	Relationship to the main policy. Here only self is considered, otherwise records get duplicated.
AGE	Age of the policyholder
V_SEX	Gender of the policyholder
N_CUST_REF_NO	Unique number of the policyholder. If policyholder bought more than 1 policy, this number would be unique
PREMIUM_BAND	Premium of the policy
N_SUM_COVERED	Sum assured of the policy (Life Cover)
V_LAST_REC_INST	Last payment method
N_CHANNEL_CODE	From which channel policyholder bought the policy. There are 4 channels. BA(Bancassurance),BR(Broker),LO(LOLC) & RE(Agent)
ZONE	Zone of the policyholder
V_OCCUP_CLASS	Risk level of the Occupation
PROD_TYPE	Product category
RIDER_INDICATOR	Whether policy holder bought riders or not
NO_OF_MEMBERS_WITH_RIDERS	Number of members insured with Riders
NO_OF_RIDERS	Number of riders bought
NO_OF_MEMBERS	Number of members insured with the policy
CHB	whether policy has bought child rider
CI	whether policy has bought critical illness rider
CI_CH	whether policy has bought critical illness rider for child
DEATH	whether policy has bought death rider
DISABILITY	whether policy has bought disability rider
HEALTH	whether policy has bought Health rider
OTHER	whether policy has bought a rider other than mentioned above
WOC_CI	whether policy has bought WOC cover for critical illness
WOC_DTH	whether policy has bought WOC cover for death
WOC_TPD	whether policy has bought WOC cover for disability
V_STATUS_DESC	Status of the policy
POLICY_DURATION_DAYS	Policy Duration

Table 4.1:Variables of Data Set

As illustrated in table 5.2, the “V_STATUS_DESC” column and “POLICY_DURATION_DAYS” (response variables) indicate the status of the policy (active or inactive) and policy duration (from the commencement date to the event date) respectively. V_STATUS_DESC includes five statuses namely LAPSE, DEATH L/A, MATURITY, IN-FORCE, SURRENDERED.

V_STATUS_DESC	Description
LAPSE	The policyholder has not paid the premium, so the policy is inactive. The company is liable for the policy if the policyholder has already paid premiums for more than 3 years.
DEATH L/A	The policyholder has died and the company is no longer liable for the policy.
MATURITY	Policy is matured and company is no longer liable for the policy.
IN-FORCE	The policy is active and the policyholder has paid the premium. The company is liable for the policy.
SURRENDERED	The policy is surrendered and the company is no longer liable for the policy.

Table 4.2: Variables of Data Set

4.1.1 Load Dataset

Firstly, python libraries and dataset are imported for the analysis. Then checked the no of records as in Figure 5.1

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.model_selection import StratifiedShuffleSplit
from sklearn.ensemble import RandomForest
from sklearn.linear_model import LogisticRegression, LogisticRegressionCV, LogisticRegressionF
from sklearn import set_config
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.pipeline import make_pipeline
from sklearn.inspection import permutation_importance
import warnings
from sklearn.exceptions import FitFailedWarning

#Import Dataset

df = pd.read_csv('DATASET_2021.csv')

df.head()
```

ID	V_PLAN_CODE	V_PYMT_DESC	N_TERM	COMMENCEMENT_YEAR	V_REL_CODE	AGE	V_SEX	N_CUST_REF_NO	PREMIUM_BAND	...	CI_CH	DEATH	
0	1	PRODUCT6	MONTHLY	15	2012	SELF	46	M	86672	5000	...	0	2
1	2	PRODUCT6	MONTHLY	15	2012	SELF	38	M	86673	1000	...	0	2
2	3	PRODUCT6	MONTHLY	15	2012	SELF	34	F	86675	1000	...	0	2
3	4	PRODUCT5	MONTHLY	20	2013	SELF	41	F	115955	1000	...	0	2
4	5	PRODUCT6	MONTHLY	20	2012	SELF	50	M	86666	1000	...	0	2

5 rows x 32 columns

```
df.shape
```

(116543, 32)

Figure 4.1:dataset sample

4.1.2 Exploratory Data Analysis (EDA)

Prior to prediction of individual policy lapse rates, conducted exploratory data analysis to get insight about the data set.

data types:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116543 entries, 0 to 116542
Data columns (total 32 columns):
#   Column                               Non-Null Count  Dtype
---  -
0   ID                                   116543 non-null  int64
1   V_PLAN_CODE                         116543 non-null  object
2   V_PYMT_DESC                         116543 non-null  object
3   N_TERM                             116543 non-null  int64
4   COMMENCEMENT_YEAR                  116543 non-null  int64
5   V_REL_CODE                         116543 non-null  object
6   AGE                                116543 non-null  int64
7   V_SEX                              116543 non-null  object
8   N_CUST_REF_NO                      116543 non-null  int64
9   PREMIUM_BAND                      116543 non-null  int64
10  N_SUM_COVERED_BAND                 116543 non-null  int64
11  V_LAST_REC_INST                    54699 non-null   object
12  N_CHANNEL_CODE                     116543 non-null  object
13  ZONE                               116456 non-null  object
14  V_OCCUP_CLASS                      116543 non-null  object
15  PROD_TYPE                          116543 non-null  object
16  RIDER_INDICATOR                    116543 non-null  object
17  NO_OF_MEMBERS_WITH_RIDERS          116543 non-null  int64
18  NO_OF_RIDERS                       116543 non-null  int64
19  NO_OF_MEMBERS                     116543 non-null  int64
20  CHB                                116543 non-null  int64
21  CI                                  116543 non-null  int64
22  CI_CH                              116543 non-null  int64
23  DEATH                              116543 non-null  int64
24  DISABILITY                         116543 non-null  int64
25  HEALTH                             116543 non-null  int64
26  OTHER                              116543 non-null  int64
27  WOC_CI                             116543 non-null  int64
28  WOC_DTH                           116543 non-null  int64
29  WOC_TPD                           116543 non-null  int64
30  V_STATUS_DESC                      116543 non-null  object
31  POLICY_DURATION_DAYS               116543 non-null  int64
dtypes: int64(21), object(11)
memory usage: 28.5+ MB
```

Figure 4.2:data types

As shown above, the data set contains 116,543 observations and 32 columns. There are 11 categorical variables and 21 numerical variables. And also, integer and object values presented in the dataset.

```
df.describe()
```

	ID	N_TERM	COMMENCEMENT_YEAR	AGE	N_CUST_REF_NO	PREMIUM_BAND	N_SUM_COVERED_BAND
count	116543.000000	116543.000000	116543.000000	116543.000000	1.165430e+05	1.165430e+05	1.165430e+05
mean	58272.000000	16.508748	2017.805153	39.646028	2.465532e+06	1.118913e+04	4.505866e+05
std	33643.210548	5.475623	2.337270	10.457120	1.510032e+06	6.019004e+04	3.927293e+06
min	1.000000	1.000000	2012.000000	18.000000	7.200000e+01	0.000000e+00	1.000000e+02
25%	29136.500000	15.000000	2016.000000	31.000000	9.242135e+05	1.000000e+03	1.000000e+05
50%	58272.000000	15.000000	2018.000000	39.000000	2.582003e+06	1.000000e+03	1.000000e+05
75%	87407.500000	20.000000	2020.000000	47.000000	3.950870e+06	5.000000e+03	5.000000e+05
max	116543.000000	40.000000	2021.000000	73.000000	4.583897e+06	1.000000e+07	1.000000e+09

8 rows x 21 columns

Figure 4.3:Summary Statistics 1

```
df.describe()
```

NO_OF_MEMBERS_WITH_RIDERS	NO_OF_RIDERS	NO_OF_MEMBERS	...	CI	CI_CH	DEATH	DISABILITY	HEALTH
116543.000000	116543.000000	116543.000000	...	116543.000000	116543.000000	116543.000000	116543.000000	116543.000000
1.367109	6.185065	1.428537	...	0.564290	0.079438	1.696807	1.494899	1.172829
0.945657	3.918414	0.885417	...	0.648289	0.351641	1.091244	1.176219	1.120592
0.000000	0.000000	1.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	3.000000	1.000000	...	0.000000	0.000000	1.000000	0.000000	0.000000
1.000000	6.000000	1.000000	...	0.000000	0.000000	2.000000	2.000000	1.000000
1.000000	8.000000	1.000000	...	1.000000	0.000000	2.000000	2.000000	1.000000
6.000000	24.000000	6.000000	...	2.000000	4.000000	6.000000	4.000000	10.000000

Figure 4.4:Summary Statistics 2

OTHER	WOC_CI	WOC_DTH	WOC_TPD	POLICY_DURATION_DAYS
116543.000000	116543.000000	116543.000000	116543.000000	116543.000000
0.083703	0.002840	0.308375	0.781883	490.856353
0.337215	0.053218	0.461824	0.412969	598.984178
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	91.000000
0.000000	0.000000	0.000000	1.000000	242.000000
0.000000	0.000000	1.000000	1.000000	639.000000
2.000000	1.000000	1.000000	1.000000	3206.000000

Figure 4.5:Summary Statistics 3

In considered data, average term of the policy is 16.5 years, average age is 39.6 years, average no of riders is 6 and average policy duration is 490.85 days as in figures 5.3,5.4 & 5.5.

```
df.isnull().sum()
```

```
ID 0
V_PLAN_CODE 0
V_PYMT_DESC 0
N_TERM 0
COMMENCEMENT_YEAR 0
V_REL_CODE 0
AGE 0
V_SEX 0
N_CUST_REF_NO 0
PREMIUM 0
N_SUM_COVERED 0
V_LAST_REC_INST 61844
N_CHANNEL_CODE 0
ZONE 87
V_OCCUP_CLASS 0
PROD_TYPE 0
RIDER_INDICATOR 0
NO_OF_MEMBERS_WITH_RIDERS 0
NO_OF_RIDERS 0
NO_OF_MEMBERS 0
CHB 0
CI 0
CI_CH 0
DEATH 0
DISABILITY 0
HEALTH 0
OTHER 0
WOC_CI 0
WOC_DTH 0
WOC_TPD 0
V_STATUS_DESC 0
POLICY_DURATION_DAYS 0
dtype: int64
```

Figure 4.6:Null values

```
df.isna().mean().round(4) * 100
```

```
ID 0.00
V_PLAN_CODE 0.00
V_PYMT_DESC 0.00
N_TERM 0.00
COMMENCEMENT_YEAR 0.00
V_REL_CODE 0.00
AGE 0.00
V_SEX 0.00
N_CUST_REF_NO 0.00
PREMIUM 0.00
N_SUM_COVERED 0.00
V_LAST_REC_INST 53.07
N_CHANNEL_CODE 0.00
ZONE 0.07
V_OCCUP_CLASS 0.00
PROD_TYPE 0.00
RIDER_INDICATOR 0.00
NO_OF_MEMBERS_WITH_RIDERS 0.00
NO_OF_RIDERS 0.00
NO_OF_MEMBERS 0.00
CHB 0.00
CI 0.00
CI_CH 0.00
DEATH 0.00
DISABILITY 0.00
HEALTH 0.00
OTHER 0.00
WOC_CI 0.00
WOC_DTH 0.00
WOC_TPD 0.00
V_STATUS_DESC 0.00
POLICY_DURATION_DAYS 0.00
dtype: float64
```

Figure 4.7: Null values as a percentage

In figure 5.6 & 5.7 shows that “V_LAST_REC_INST” and ZONE have null values. Missing percentage of “V_LAST_REC_INST” is 53.07%. That means half of the records are missing. So, it has been removed from the analysis. Since the missing percentage of the “zone” is only 0.07%, replace the missing value with the most frequent zone, which is “UVA & EASTERN” as per the figure 5.8

```
df.ZONE .value_counts()/len(df)*100
```

```
UVA & EASTERN 14.886351
SOUTHERN 14.828003
NORTH CENTRAL 13.728838
NORTHERN & EASTERN 12.617660
HEAD OFFICE 10.760835
WESTERN & NORTH WESTERN 10.354976
NORTH WESTERN 8.760715
METRO 8.265619
SABARAGAMUWA 3.570356
CENTRAL & UVA EASTERN 0.782544
SOUTHERN & SABARAGAMUWA 0.627236
NORTHERN 0.417872
EASTERN 0.324344
Name: ZONE, dtype: float64
```

Figure 4.8: Zone wise percentage of policy count

```
df.V_STATUS_DESC.value_counts()/len(df)*100
```

```
V_STATUS_DESC
LAPSE          73.208172
IN-FORCE       26.186043
MATURITY        0.434175
DEATH L/A       0.150159
SURRENDERED     0.021451
Name: count, dtype: float64
```

Figure 4.9:percentage of policy count based on policy status

In the dataset 73% of the policies is lapsed and only 26% of the policies are in-force as in figure 5.9

```
df.V_PYMT_DESC.value_counts()/len(df)*100
```

```
V_PYMT_DESC
MONTHLY        81.164892
YEARLY         9.267824
QUARTERLY       6.096462
HALF YEARLY     3.470822
Name: count, dtype: float64
```

Figure 4.10:percentage of policy count based on premium frequency

81% of the policies pay premium monthly.

```
df.PREMIUM_BAND .value_counts()/len(df)*100
```

```
PREMIUM_BAND
1000          60.785290
5000          19.715470
10000         12.301039
100000         3.527453
50000         2.871043
500000         0.416155
100           0.251409
1000000        0.119269
0              0.011155
5000000        0.000858
10000000       0.000858
Name: count, dtype: float64
```

Figure 4.11:percentage of policy count based on premium band

There are policies which have 0 for the premium variable. Those have been by the most frequent premium.

```
df.duplicated().sum()
```

```
0
```

Figure 4.12:duplicate values

There are no duplicate policies in the data set.

4.1.3 Data Visualization

For the study, lapse rate needs to be predicted for individual policy holder. Therefore, interested events will lapse and non-lapse. For that, “V_STATUS_DESC” variable is converted the as follows:

```
df["V_STATUS_DESC"].value_counts()/len(df)*100  
  
V_STATUS_DESC  
LAPSE      73.208172  
NO LAPSE   26.791828  
Name: count, dtype: float64
```

Figure 4.13:percentage of policy count on policy status after conversion

Under the response variable, there are 2 variables: “V_STATUS_DESC” & “POLICY_DURATION_DAYS”. The influence of each independent variable on the “V_STATUS_DESC” is analysed first.

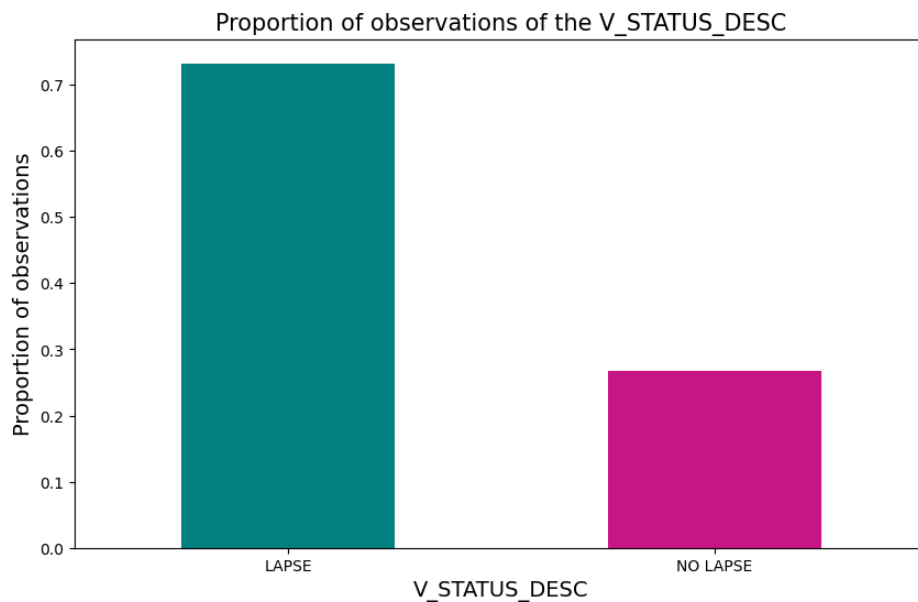


Figure 4.14:proportion of observations of the policy status

The above bar plot shows the percentage of observations that correspond to each class of the “V_STATUS_DESC”: lapse and no lapse. It shows that 73.21% of policies have lapsed.

Since the policy count is large, choose a sample size of 8000 for the study. Used the stratified sampling method to choose a sample based on “V_STATUS_DESC” and analysed the sample before prediction.

A normalized stacked bar plot is used here to compare how the response variable varies across all groups of independent variables.

Information



Figure 4.15: proportion of observation of the categorical variables on policy status in sample 1

Information

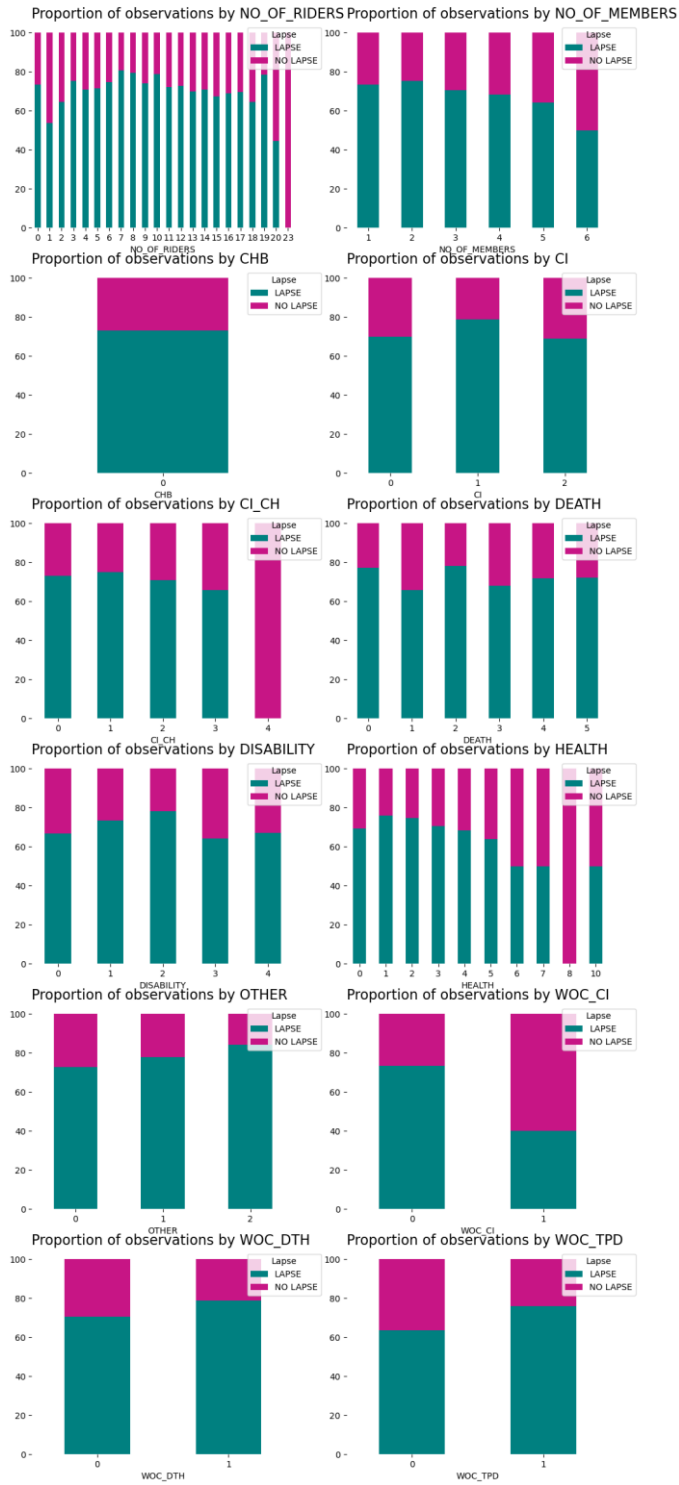


Figure 4.16: proportion of observation of the categorical variables on policy status in sample 2

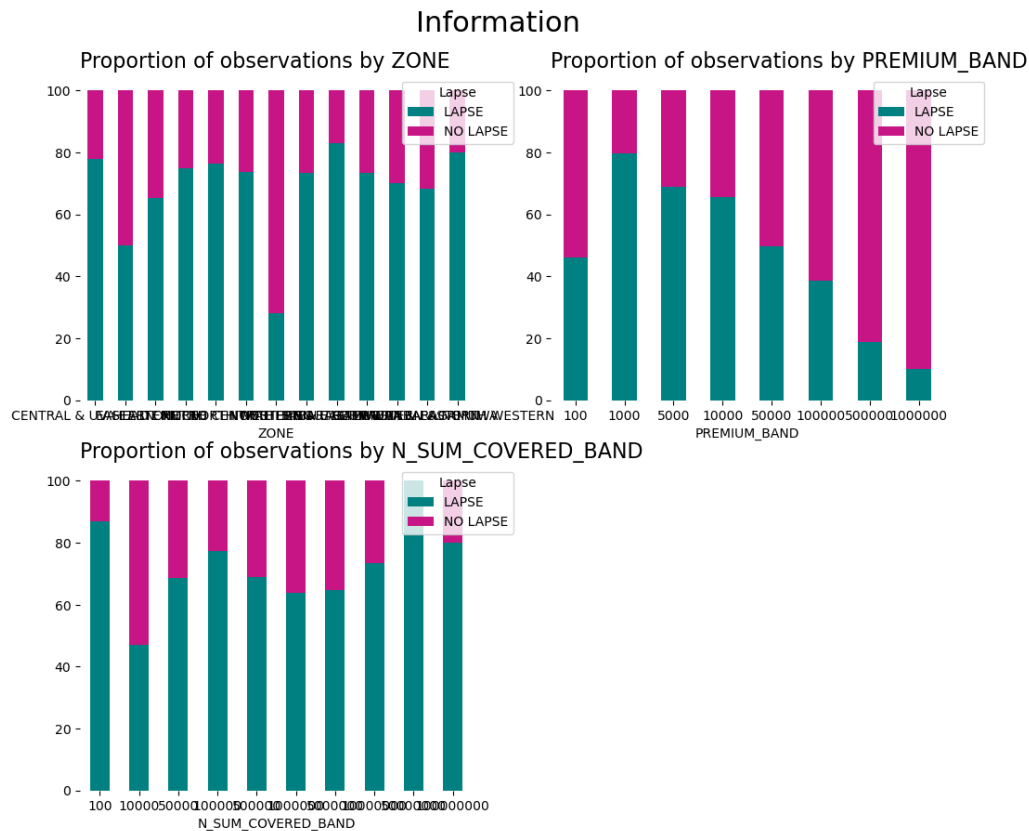


Figure 4.17: proportion of observation of the categorical variables on policy status in sample 3

The following information can be extracted by analysing the above variables:

- A similar percentage of lapsation is shown both when a customer is a man or a woman.
- Yearly and half-yearly policies have less lapsation than monthly and quarterly policies.
- Policies that commenced in 2020 and 2021 are more likely to lapse than in other years.
- From channels, “BR” has fewer lapses than other channels following “BA” and “RE”.
- There is no significant difference in lapsation between with rider and no rider.

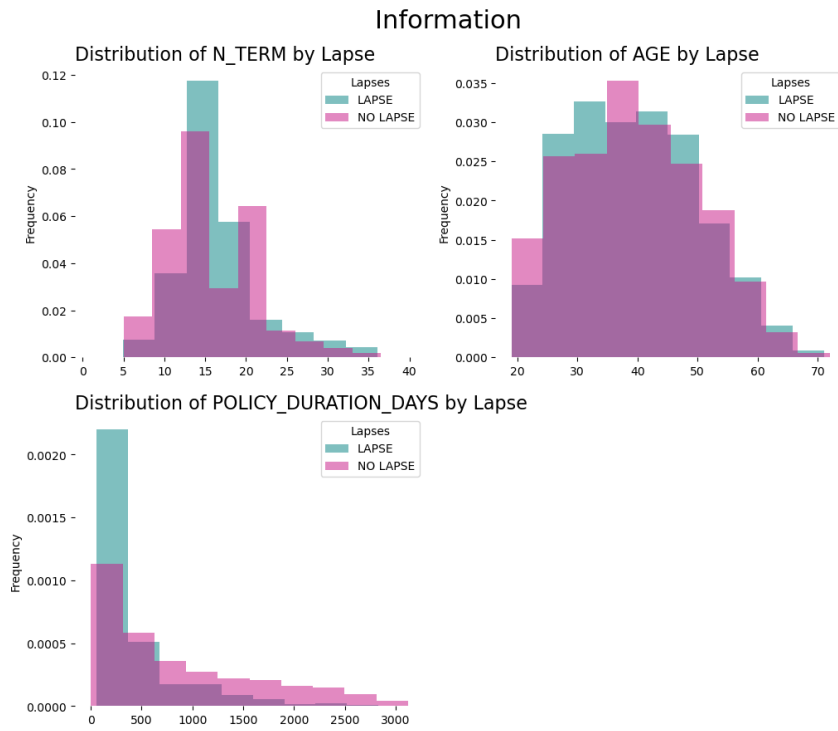


Figure 4.18: Distribution of policy duration by term, age & lapse in the sample

The following information can be extracted by analysing the above variables:

- Term 15–20 years have more lapses than other terms.
- Lapse rates tend to be higher when the policy duration is 0-500 days.

Checked the correlation matrix for all features and found below information

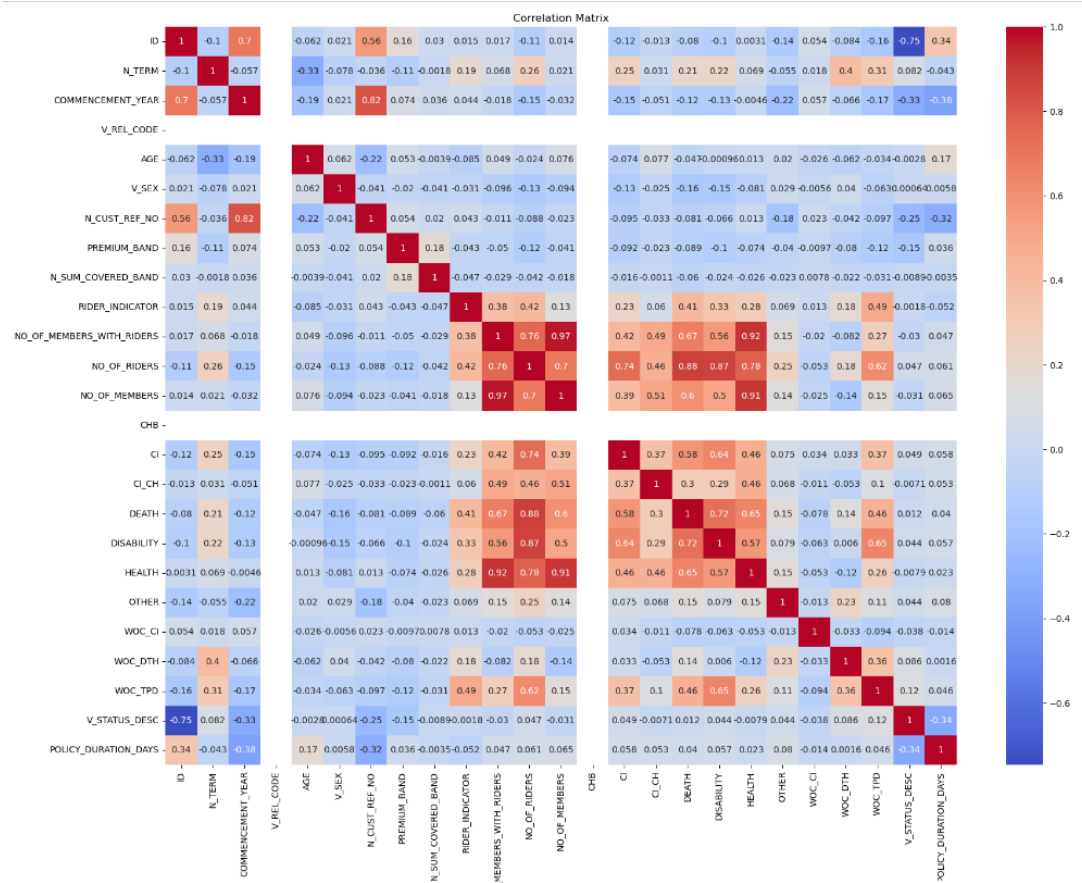


Figure 4.19:correlation matrix

NO_OF_MEMBERS_WITH_RIDERS and NO_OF_MEMBERS are highly correlated. Therefore, NO_OF_MEMBERS_WITH_RIDERS has been removed from the data.

Removed N_CUST_REF_NO, ID, and V_PLAN_CODE, as well as those won't be affected by the response variable.

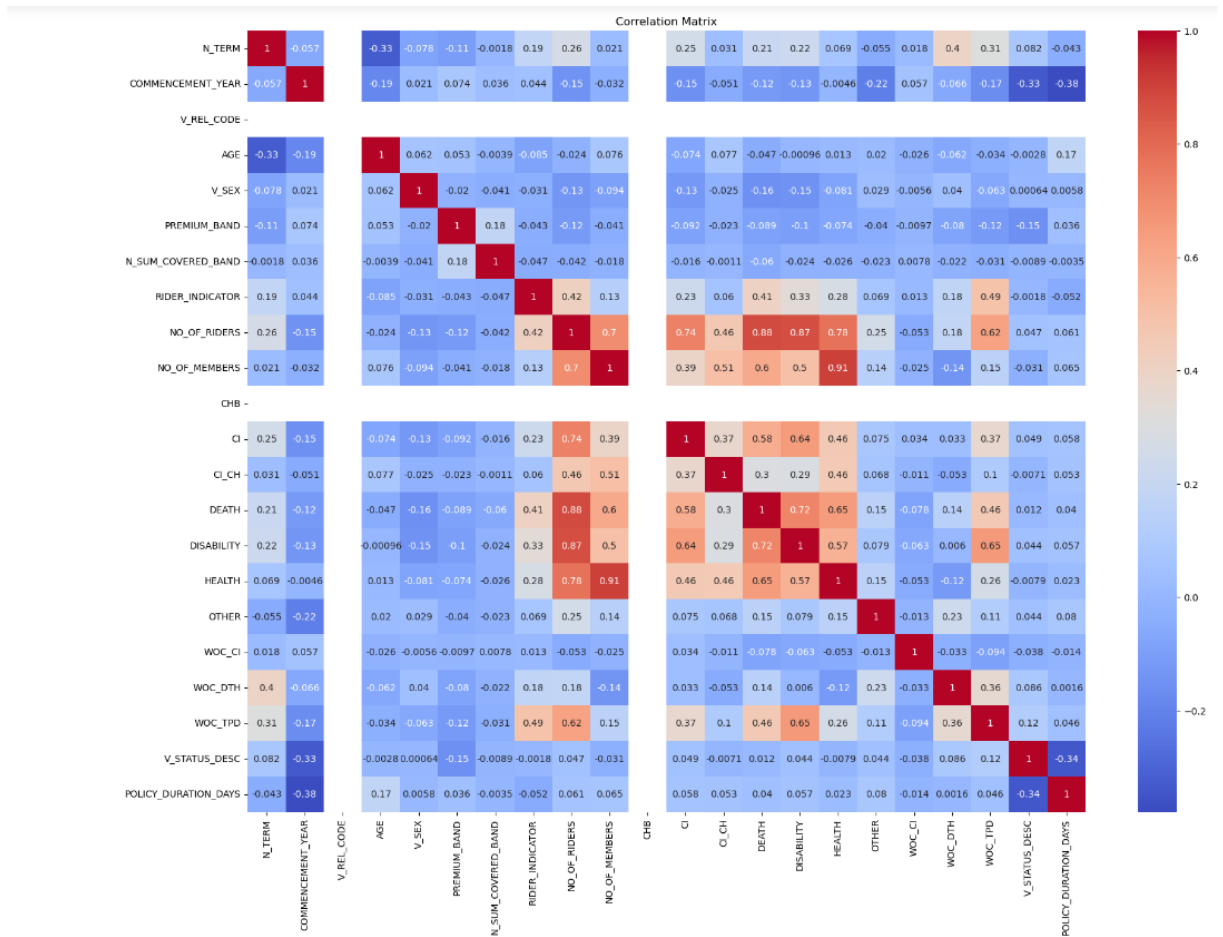


Figure 4.20:correlation matrix after removing correlated variables

Next, evaluated “POLICY_DURATION_DAYS” using survival analysis.

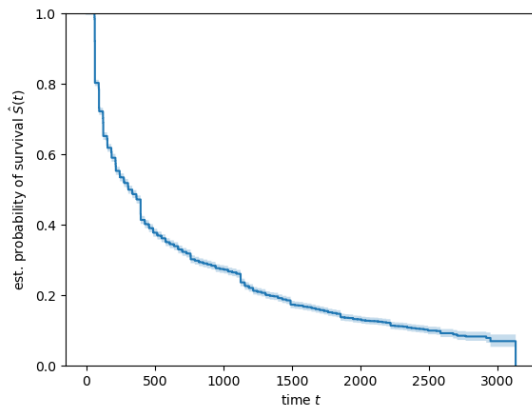


Figure 4.21:survival curve of the sample

From the plot, we can see that most policies lapsed in the first 1000 days, as indicated by the steep slope of the estimated survival function in the first 1000 days.

Then analysed survival function with other variables.

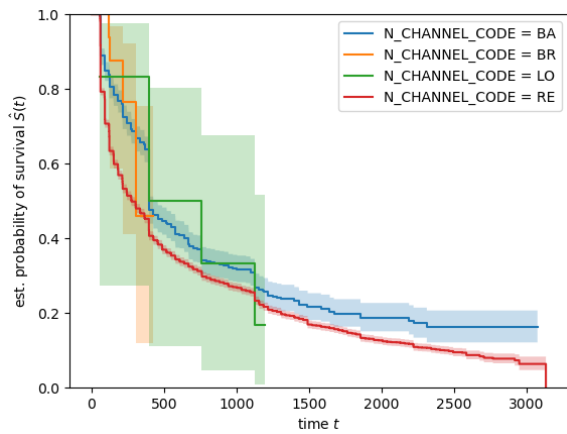


Figure 4.22:survival curve by channel code of the sample

Policies that came through the RE and BA channels have lapsed in the first 1000 days.

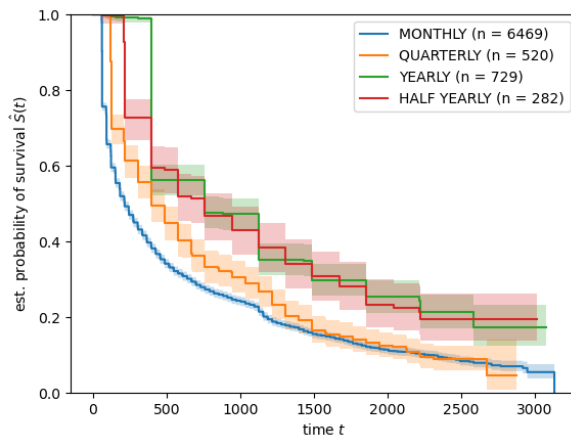


Figure 4.23:survival curve by premium frequency of the sample

Policies that have “YEARLY” and” HALF YEARLY” modes seem to have a better survival rate compared to policy holders with other modes.

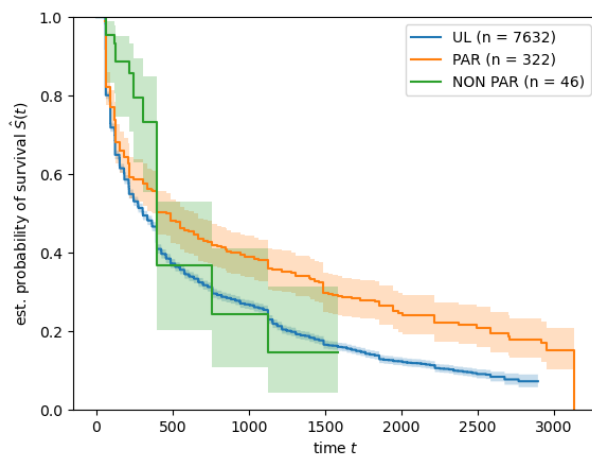


Figure 4.24:survival curve by product type of the sample

Policies that have a “PAR” product type seem to have a better survival rate compared to policy holders with other product types.

4.2 Predicting Policy lapsation

4.2.1 Random survival forest

First, we need to convert categorical data to numeric data.

```
data_X_sample2 = pd.get_dummies(data=data_X_sample1, columns=["V_PYMT_DESC", "N_CHANNEL_CODE", "ZONE", "V_OCCUP_CLASS", "PROD_TYPE"])
```

Figure 4.25: data transformation

Then split data set into training and test data.

```
X_train, X_test, y_train, y_test = train_test_split(data_X_sample2, data_y_sample1, test_size=0.2, stratify=data_y_sample1["V_STA"])
```

Figure 4.26: data splitting

Trained the data using random survival.

```
rsf = RandomSurvivalForest(n_estimators=100, min_samples_split=10, min_samples_leaf=15, n_jobs=-1, random_state=20)
rsf.fit(X_train, y_train)

RandomSurvivalForest(min_samples_leaf=15, min_samples_split=10, n_jobs=-1,
                      random_state=20)
```

Figure 4.27: train data using random survival forest

Finding model performance:

```
: rsf.score(X_test, y_test)
: 0.6246438520486562
```

Figure 4.28: model performance

Predicting survival function for selected data:

```
: X_test_sorted = X_test.sort_values(by=["AGE", "PREMIUM_BAND"])
X_test_sel = pd.concat((X_test.head(3), X_test.tail(3)))

: pd.Series(rsf.predict(X_test_sel))

: 0    410.041671
  1    262.268202
  2    337.403865
  3    390.483993
  4    444.006990
  5    458.821843
dtype: float64

: import matplotlib.pyplot as plt

: surv = rsf.predict_survival_function(X_test_sel, return_array=True)

for i, s in enumerate(surv):
    plt.step(rsf.unique_times_, s, where="post", label=str(i))
plt.ylabel("Survival probability")
plt.xlabel("Time in days")
plt.legend()
plt.grid(True)
```

Figure 4.29: model prediction -random survival forest

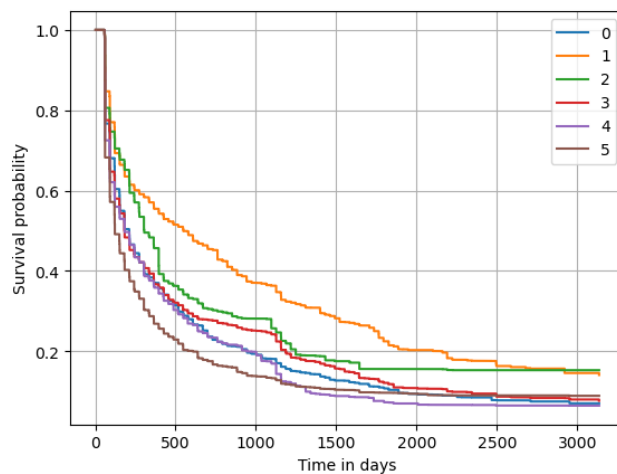


Figure 4.30: predicted survival curve

4.2.2 Cox net Survival Analysis

Fitting the model:

```
import warnings
from sklearn.exceptions import FitFailedWarning
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler

coxnet_pipe = make_pipeline(StandardScaler(), CoxnetSurvivalAnalysis(l1_ratio=0.9, fit_baseline_model=True))
warnings.simplefilter("ignore", UserWarning)
warnings.simplefilter("ignore", FitFailedWarning)
coxnet_pipe.fit(X_train, y_train)

Pipeline(steps=[('standardscaler', StandardScaler()),
                 ('coxnetsurvivalanalysis',
                  CoxnetSurvivalAnalysis(fit_baseline_model=True,
                                          l1_ratio=0.9))])
```

Figure 4.31: Fit the model using Cox net survival analysis

Model performance:

```
coxnet_pipe.score(X_test, y_test)

0.5783550277086575
```

Figure 4.32: Model performance-Cox net survival analysis

4.2.3 Model Evaluation

Model performance is evaluated the by C-index and time-dependent area under the ROC.

4.2.3.1 C-index

```
score_cindex = pd.Series(
    [
        rsf.score(X_test, y_test),
        coxnet_pipe.score(X_test, y_test),
        0.5,
    ],
    index=["RSF", "CNH", "Random"],
    name="c-index",
)

score_cindex.round(3)

RSF      0.625
CNH      0.578
Random   0.500
Name: c-index, dtype: float64
```

Figure 4.33: C-index of selected algorithms

Higher the C-index score, higher the model performance is (Andrade et al., 2021). From above it can be seen that random survival forest perform better than cox net survival analysis.

4.2.3.2 Time-dependent area under the ROC curve

```
rsf_chf_funcs = rsf.predict_cumulative_hazard_function(X_test, return_array=False)
rsf_risk_scores = np.row_stack([chf(va_times) for chf in rsf_chf_funcs])

rsf_auc, rsf_mean_auc = cumulative_dynamic_auc(y_train, y_test, rsf_risk_scores, va_times)

C:\ProgramData\anaconda3\Lib\site-packages\sksurv\metrics.py:482: RuntimeWarning: invalid value encountered in divide
  true_pos = cumsum_tp / cumsum_tp[-1]

plt.plot(va_times, coxnet_auc, "o-", label=f"CoxNH (mean AUC = {coxnet_mean_auc:.3f})")
plt.plot(va_times, rsf_auc, "o-", label=f"RSF (mean AUC = {rsf_mean_auc:.3f})")
plt.xlabel("days from enrollment")
plt.ylabel("time-dependent AUC")
plt.legend(loc="lower center")
plt.grid(True)
```

Figure 4.34:calculating time-dependent AUC

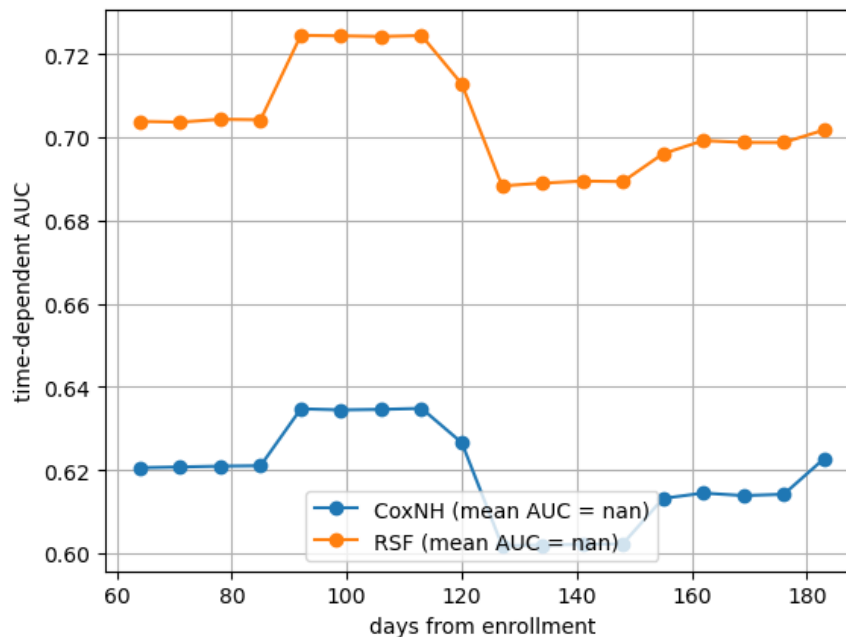


Figure 4.35:time-dependent AUC

It can be seen that random survival forest perform better than cox net survival analysis

4.3 Important Features

From the above evaluation, random survival forest has been used for further analysis.

Finding important features using random survival forest:

```
#Feature Importance
from sklearn.inspection import permutation_importance

result = permutation_importance(rsf, X_test, y_test, n_repeats=15, random_state=20)

pd.DataFrame(
    {
        k: result[k]
        for k in (
            "importances_mean",
            "importances_std",
        )
    },
    index=X_test.columns,
).sort_values(by="importances_mean", ascending=False)
```

Figure 4.36:Finding Important features

Policy Status as at 31/12/2018	mean	std
V_PYMT_DESC_MONTHLY	0.029093	0.003773
V_PYMT_DESC_YEARLY	0.011844	0.00278
ZONE_SOUTHERN	0.007441	0.001464
PREMIUM_BAND	0.007357	0.001859
ZONE_NORTH CENTRAL	0.007339	0.001598
ZONE_NORTHERN & EASTERN	0.005403	0.002683
AGE	0.002369	0.004814
ZONE_WESTERN & NORTH WESTERN	0.002281	0.000915
V_PYMT_DESC_HALF YEARLY	0.001947	0.000692
ZONE_UVA & EASTERN	0.001834	0.001065
CI	0.001714	0.000703
N_SUM_COVERED_BAND	0.001441	0.000994
N_TERM	0.001036	0.001374
V_SEX	0.001014	0.000743
DISABILITY	0.000833	0.000521
ZONE_SABARAGAMUWA	0.000246	0.000172
V_OCCUP_CLASS_MEDIUM	0.000207	0.000188
WOC_TPD	0.000183	0.000227

Policy Status as at 31/12/2021	mean	std
AGE	0.0146001	0.00338
PREMIUM_BAND	0.01249414	0.003055
V_PYMT_DESC_MONTHLY	0.01040074	0.002791
V_PYMT_DESC_YEARLY	0.0063567	0.002651
ZONE_SOUTHERN	0.004834227	0.001729
ZONE_NORTH CENTRAL	0.003976795	0.001307
N_TERM	0.001742688	0.001156
ZONE_WESTERN & NORTH WESTERN	0.001411175	0.000624
ZONE_UVA & EASTERN	0.001165317	0.00141
ZONE_CENTRAL & UVA EASTERN	0.00104708	0.000435
COMMENCEMENT_YEAR	0.000952537	0.001214
NO_OF_MEMBERS	0.000949331	0.001009
V_PYMT_DESC_HALF YEARLY	0.000789887	0.000908
DISABILITY	0.000623899	0.000737
ZONE_NORTH WESTERN	0.000608203	0.000447
CI	0.000528893	0.00066
RIDER_INDICATOR	0.000410657	0.000123
N_SUM_COVERED_BAND	0.000346846	0.001024
ZONE_METRO	0.000209938	0.000617
PROD_TYPE_UL	0.000201809	0.000646
V_SEX	0.000174051	0.001319
ZONE_SABARAGAMUWA	0.000147581	0.000432

Table 4.3: Important features

Other features are insignificant for predicting policy lapses.

It seems that those are varied based on the investigation period. 2018 is before COVID, and 2021 data is after COVID.

Therefore, for current usage, the most important variables, as of December 31, 2021, have been used.

4.4 Performing scenarios to identify policy characteristics

Then new data is applied to the built model, and the below information is obtained.

- For premiums up to 6,000 and sum assured up to 50,000, for policy term 5, low-age monthly policies will tend to have a higher lapse rate than premiums up to 100,000, sum

assured up to 1000000, and for policy term 10/15, high-age yearly policies.

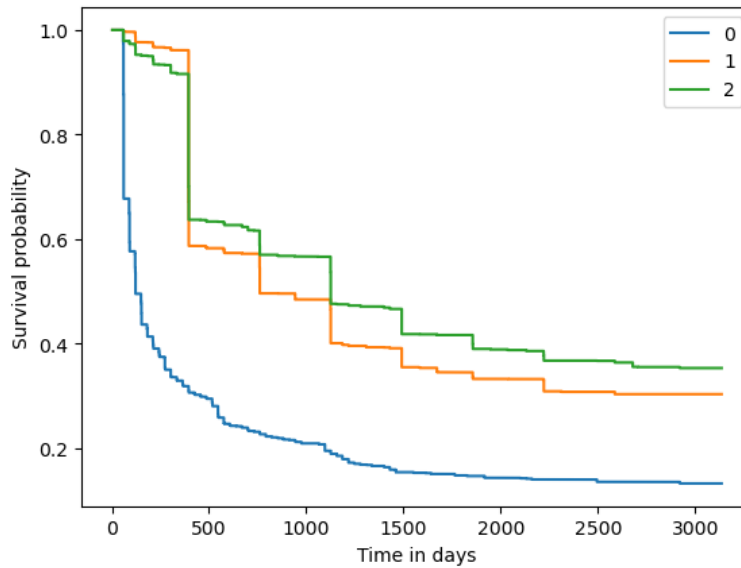


Figure 4.37:survival rates

0: AGE: 23, PREMIUM_BAND =6,000, N_SUM_COVERED_BAND:500,000, N_TERM=5, V_PYMT_DESC_MONTHLY=1, V_PYMT_DESC_YEARLY=0
1: AGE: 43, PREMIUM_BAND =100,000, N_SUM_COVERED_BAND:1,000,000, N_TERM=15, V_PYMT_DESC_MONTHLY=0, V_PYMT_DESC_YEARLY=1
2: AGE: 40, PREMIUM_BAND =500,000, N_SUM_COVERED_BAND:1,000,000, N_TERM=10, V_PYMT_DESC_MONTHLY=0, V_PYMT_DESC_YEARLY=1

The probability of policy lapsation can be reduced by applying the below scenarios:

For premiums up to 6,000, sum assured up to 50,000, for policy term 5, age 23 monthly policies

- If the term is increased to 10 it will get a higher survival rate.

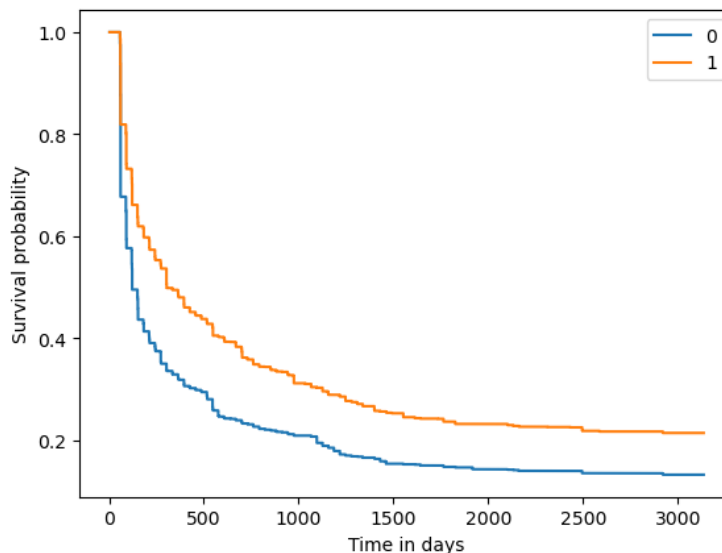


Figure 4.38:Effect of policy terms on survival rate

0: term 5 year, **1:** term 10 year

- If premium mode is converted to yearly, it will give would higher survival rate.

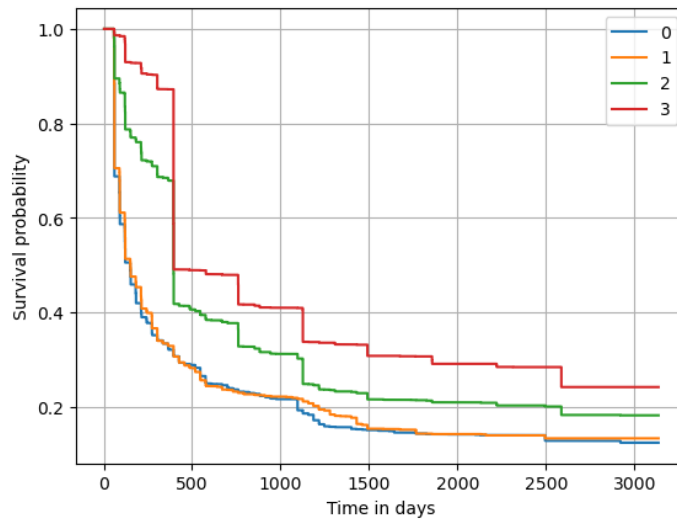


Figure 4.39: Effect of premium frequencies on survival rate
0,1- premium mode monthly **2,3** -premium mode yearly

If it has a high sum assured (1,000,000 instead of 100,000) survival probability will improve for the same premium.

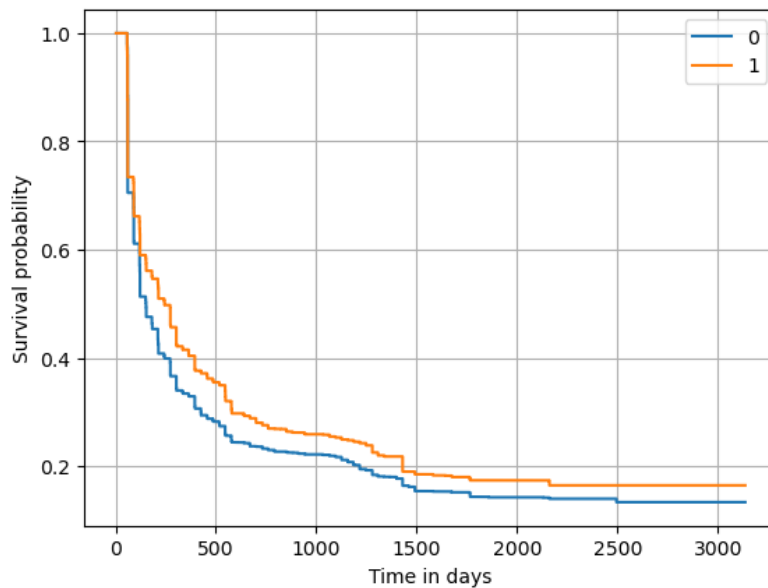


Figure 4.40: Effect of sum assured on survival rate
0-sum assured 100,000, **1**- sum assured 1000,000

For premium $\leq 10,000$ and sum assured 100,000, policy term 10

- For premium mode, yearly policies tend to lapse less than monthly policies.

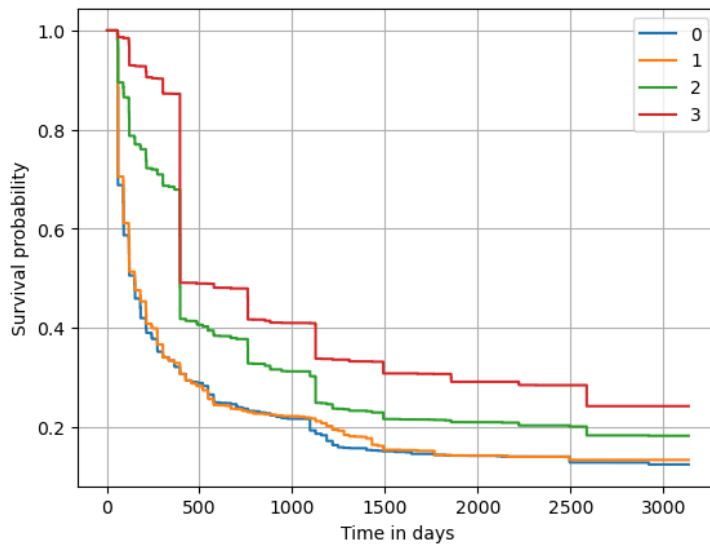


Figure 4.41:Effect of premium frequencies on survival rate 2

0,1: Premium mode - Monthly

2,3: Premium mode – Yearly

- If the sum assured is higher, the probability of policy lapse is low.

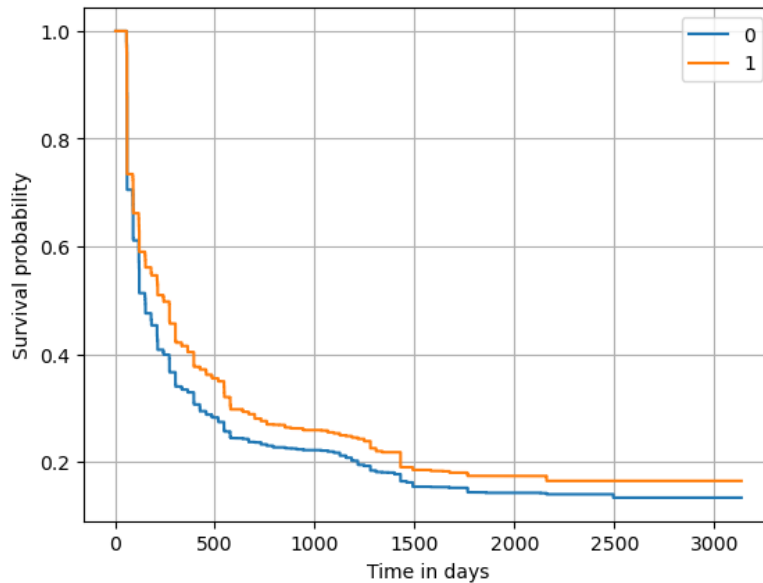


Figure 4.42:Effect of sum assured on lapse rate 2

0: Sum Assured 100,000

1: Sum Assured 1,000,000

If Product type =NP (Non-Par) and if there is CI cover, the probability of policy lapse is low.

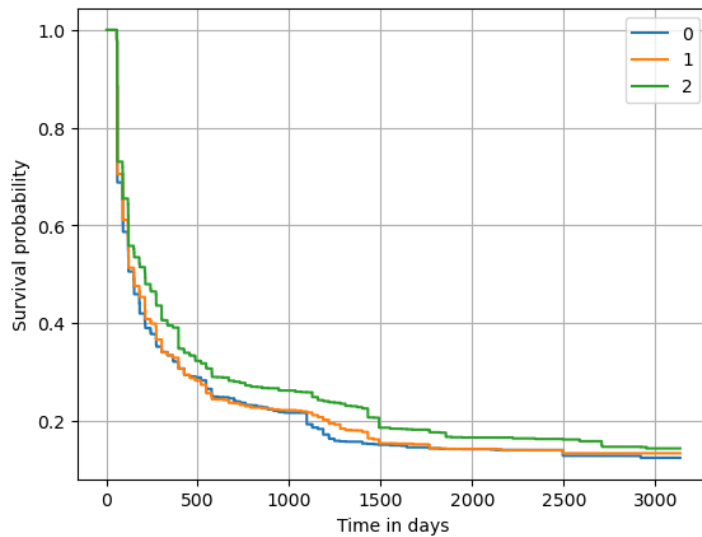


Figure 4.43:effect of product type/riders on survival rate

0: PREMIUM_BAND =1,000, **1,2:** PREMIUM_BAND =10,000

0,1: PROD_TYPE_UL=1, CI=0, DISABILITY=1

2: PROD_TYPE_UL=0, CI=1, DISABILITY=0

- Policies with higher age tend to lapse less than polies with lower age.

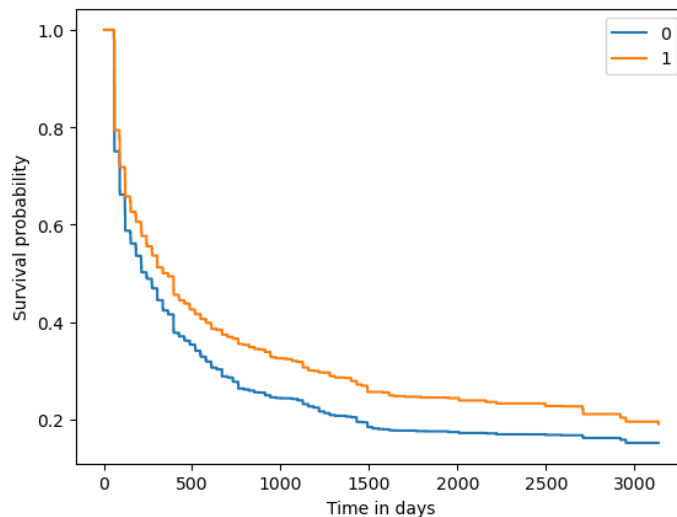


Figure 4.44:Effect of age on survival rate

0: AGE =23, PREMIUM_BAND =5000, N_SUM_COVERED_BAND:100,000, N_TERM=10

1: AGE =43, PREMIUM_BAND =5000, N_SUM_COVERED_BAND:100,000, N_TERM=10

4.5 Discussion

4.5.1 Challenges

The below mentioned challenges were faced when doing the research.

Large Data Volume

High data volume would improve the model performance as a large amount of data is contributed to the model fitting. However, computation would be difficult for such data. As a

result, a dataset with 116,543 records had to be reduced to 8,000 records. This would have an impact on model building and predicting survival probability.

Data Integrity

Most important variables like payment method have more null values. Therefore, it had to be removed from the studies. This may have a significant contribution to lapse prediction.

4.5.2 Assumptions

- One policyholder has bought one policy
- For the research, one policy status at a time is considered. Multiple policy status are not considered for the study.
- Premium reflects the income of the policyholder.
- Effects of the economic conditions on the policyholder are reflected in policy status.

Chapter 5 Conclusion

In conclusion, by using machine learning techniques policy lapse rate for each policyholder can be predicted accurately as it considered nonlinear relationships among variables. To evaluate this machine learning technique such as random survival forest as well as traditional survival techniques such as cox net survival analysis are used.

Due to computational difficulties stratified sample is chosen from the dataset. After fitting the model using the above-mentioned algorithms, models are evaluated using C-index and time dependent area under the ROC curve. It can be seen that model performance of the random survival forest is higher than cox net survival analysis.

Then the most important variables which contributed to policyholder lapse prediction are selected using permutation importance from random survival forest. This was done on datasets of two specific period to see the impact on economic conditions for policyholders.

Finally, several scenarios are performed to identify policyholder characteristics. It showed that policies with higher premium, higher sum assured and yearly premium payment method have higher survival rate. And also, policies with product type =NP (Non-Par) and if there is CI cover, those have higher survival rate. Hence above scenarios improve the survival probability of policyholders, those will be beneficial for the company when commencing new business.

References

- Aleandri, M., n.d. Modeling Dynamic Policyholder Behavior through Machine Learning Techniques.
- Andrade, Andrade, J.L., Valencia, J.L., 2021. Modeling lapse rates using machine learning: a comparison between survival forests and cox proportional hazards techniques. *Anales del Instituto de Actuarios Españoles* 161–183. https://doi.org/10.26360/2021_7
- Andrade, J.L., Valencia, J.L., 2022. A Fuzzy Random Survival Forest for Predicting Lapses in Insurance Portfolios Containing Imprecise Data. *Mathematics* 11, 198. <https://doi.org/10.3390/math11010198>
- Bauer, D., Gao, J., Moenig, T., Ulm, E.R., Zhu, N., n.d. Policyholder Exercise Behavior in Life Insurance: The State of Affairs.
- Ch, S.C., Ramesh, J., 2011. LAPSING OF POLICIES IN LIFE INSURANCE SECTOR – NEED FOR COMPETITIVE STRATEGIES 1.
- Chan, M., Li, K., Lombardi, Louis, Lombardi, Lucian, Purushotham, M., Rao, A., 2013. A survey and literature review.
- de Azevedo, K.S., n.d. Survival Model Analysis applied to Kidney Transplant Data.
- Devale, A.B., 2012. Applications of Data Mining Techniques in Life Insurance. *IJDKP* 2, 31–40. <https://doi.org/10.5121/ijdkp.2012.2404>
- Eling, M., Kochanski, M., 2013. Research on lapse in life insurance: what has been done and what needs to be done? *The Journal of Risk Finance* 14, 392–413. <https://doi.org/10.1108/JRF-12-2012-0088>
- Geschiere, M., n.d. Predicting the Lapse Rates of AllSecur.
- Goonetilleke, T.L.O., Caldera, H.A., 2013. Mining Life Insurance Data for Customer Attrition Analysis. *JIII* 1, 52–58. <https://doi.org/10.12720/jiii.1.1.52-58>
- Groll, A., Wasserfuhr, C., Zeldin, L., 2022. Churn modeling of life insurance policies via statistical and machine learning methods -- Analysis of important features.
- Grover, G., Ravi, V., Saini, R., Varshney, M.K., n.d. Predictive Modelling of Lapsation of Life Insurance Policies in India.
- Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. *Ann. Appl. Stat.* 2. <https://doi.org/10.1214/08-AOAS169>
- Kaushik, K., Bhardwaj, A., Dwivedi, A.D., Singh, R., 2022. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *IJERPH* 19, 7898. <https://doi.org/10.3390/ijerph19137898>
- Liebenberg, A.P., Carson, J.M., Dumm, R.E., 2012. A Dynamic Analysis of the Demand for Life Insurance. *J of Risk & Insurance* 79, 619–644. <https://doi.org/10.1111/j.1539-6975.2011.01454.x>
- Loisel, S., Piette, P., Tsai, C.-H.J., 2021. APPLYING ECONOMIC MEASURES TO LAPSE RISK MANAGEMENT WITH MACHINE LEARNING APPROACHES. *ASTIN Bull.* 51, 839–871. <https://doi.org/10.1017/asb.2021.10>
- Lombardi, L., Paich, M., n.d. Behavioral Simulations: Using agent-based modeling to understand policyholder behaviors.
- Matematico-Statistiche, E., n.d. UNIVERSITÀ DEGLI STUDI DI TORINO SCUOLA DI MANAGEMENT ED ECONOMIA.
- Mills, A., n.d. Complexity Science: an introduction (and invitation) for actuaries. *Complexity science*.
- Shumrak, H.M., Darley, V., 1999. Applying Complex Adaptive Systems to Actuarial Problems.
- Vasilev, I., Petrovskiy, M., Mashechkin, I., 2022. Survival Analysis Algorithms based on Decision Trees with Weighted Log-rank Criteria:, in: *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. Presented

- at the 11th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS - Science and Technology Publications, Online Streaming, --- Select a Country ---, pp. 132–140. <https://doi.org/10.5220/0010987100003122>
- Verhoef, P.C., Donkers, B., 2001. Predicting customer potential value an application in the insurance industry. *Decision Support Systems* 32, 189–199. [https://doi.org/10.1016/S0167-9236\(01\)00110-5](https://doi.org/10.1016/S0167-9236(01)00110-5)
- Xong, L.J., Kang, H.M., 2019. A Comparison of Classification Models for Life Insurance Lapse Risk 7

