NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

APPLICATION OF SURVIVAL ANALYSIS IN THE INSURANCE INDUSTRY

Predicting Motor Policy Tenure and Unveiling Factors Impacting Retention

Skander Chaabini

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

APPLICATION OF SURVIVAL ANALYSIS IN THE INSURANCE INDUSTRY

Predicting Motor Policy Tenure and Unveiling Factors Impacting Retention

by

Skander Chaabini

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Bruno Jardim, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2024

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to all the professors at Nova IMS who have imparted their knowledge and wisdom during my academic journey. Their dedication and passion for teaching have been a source of inspiration. A special thank you to my academic supervisor, Bruno Jardim, for his invaluable guidance, support, and encouragement throughout the course of this thesis. His expertise and insights have been instrumental in shaping this work.

I am immensely grateful to the data science team at Generali Tranquilidade, with whom I had the privilege to work during my internship. Their collaboration and support have significantly enriched my experience and understanding of real-world data science applications. A special thank you goes to my professional mentor, Miguel Lince, for his mentorship and advice. His professional guidance has been crucial in my development and understanding of the practical aspects of data science.

I would also like to thank my friends for their unwavering support and encouragement, which have kept me motivated throughout this journey.

Lastly, I am profoundly grateful to my family for their endless love, patience, and support. Their belief in me has been my greatest strength and motivation. I would also like to acknowledge myself for the hard work, dedication, and perseverance that have gone into completing this thesis.

Thank you all for your contributions and support, without which this thesis would not have been possible.

ABSTRACT

This thesis explores the application of survival analysis in predicting motor policy tenure and identifying factors impacting retention within the insurance industry. Effective customer retention is crucial for insurers to maintain financial stability and growth in a competitive market. Survival analysis models are developed to estimate the probability of a policy to be active over time, allowing for proactive identification of at-risk policies and targeted retention strategies. Key factors influencing policy duration are analyzed using machine learning techniques such as Penalized Cox Proportional Hazards, Random Survival Forest, and Gradient Boosting. Among these methods, Penalized Cox Proportional Hazards is favored for its superior discriminatory power and interpretability. By enhancing understanding of policyholder behavior and improving retention strategies, this research contributes to optimizing business practices and sustaining growth for insurers. The insights gained can inform strategic decisions in customer relationship management, pricing strategies, and service offerings, thereby maximizing customer lifetime value and profitability.

KEYWORDS

Survival Analysis; Motor Insurance; Policy Tenure Prediction; Penalized Cox Proportional Hazards; Customer Retention

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CPH: Cox Proportional Hazards Model

RSF: Random Survival Forest

GBS: Gradient Boosting

C-index: Concordance Index

t-AUC: Time-dependent Area Under the ROC Curve

IBS: Integrated Time-dependent Brier Score

1. INTRODUCTION

The insurance industry has developed significantly within the last century, although the essence of insurance has always remained a straightforward transaction. It is an agreement where the customer, called policyholder, pays regular premiums to the insurance company in exchange for financial protection and compensation in case of a covered unexpected event.

It is crucial for an insurance company to consider its clients' satisfaction and loyalty and prevent them from leaving for competitor companies, particularly in the competitive motor insurance market. The insurance business model relies heavily on retaining policyholders to ensure sustainability and growth. When a customer decides to leave, it is termed as "churn", and it is a significant issue since the company loses money and needs to spend more to attract a new customer.

Hence, a major challenge for insurance is predicting for how long a policy will remain active before being cancelled and understanding which policyholders are at risk of churning and what are the factors that influence the churn event.

This research studies the application of survival analysis, a statistical method suited for analyzing time-to-event data. In the context of motor insurance, survival analysis allows to model policy tenure duration until the occurrence of churn. Through this approach, the study aims to achieve two key objectives:

- Predicting policy tenure: The construction of a survival model enabling the estimation of the probability of a policy to be active at any given point in time based on the defined study period. This will empower insurers to proactively identify policies at risk of cancellation and implement targeted retention strategies.

- Exploring factors impacting retention: The analysis of the survival model revealing the factors influencing policy duration. These factors include demographics, vehicle information, claims history, and pricing strategies. Understanding these drivers allows the company to develop data-driven retention strategies that address specific customer needs and behaviors.

This research contributes to the insurance industry by demonstrating the effectiveness of survival analysis in predicting policy tenure and identifying key drivers of customer loyalty. By employing powerful statistical tools and comprehensive data analysis, this study has the potential to significantly improve customer retention and ensure sustained growth for insurance companies.

2. LITERATURE REVIEW

2.1. INSURANCE INDUSTRY AND RELATED DATA SCIENCE APPLICATIONS

The insurance industry holds significant importance within the European Union's financial landscape. It acts as a major institutional investor and offers a broad spectrum of risk mitigation solutions to a vast customer base. Insurance Europe (2020) reports an average of €248 was spent per person on motor insurance in Europe and that life insurance had the highest average spend per capita at approximately 1,163 euros. In 2023, Insurance Europe stated that the average per capita expenditure exceeding €2,000 across 35 European countries, highlighting the continued reliance on insurance for financial protection and risk mitigation.

Risk mitigation is the core nature of the insurance industry and therefore comes the important role of data science. Traditionally, the scarcity of data posed a significant challenge in measuring operational risks. Nevertheless, in recent years, insurance companies have increasingly gathered statistical information to address this limitation. The increasing role of digitalization and data science in the insurance industry is transforming various aspects, as highlighted by Albrecher et al. (2019).

The influx of data within the insurance industry, fueled by advancements in data collection and storage technologies, has opened doors for the transformative power of data science. These sophisticated techniques are revolutionizing various aspects of insurance:

- Risk Assessment and Pricing: Advanced analytics leverage historical data, telematics (vehicle sensor data), and even weather patterns to create more accurate risk models, leading to personalized pricing that reflects individual customer profiles (Anh & Duc, 2024).
- Fraud Detection: Machine learning algorithms can identify fraudulent claims based on patterns in claims history, behavior analysis, and external data sources (Agarwal, 2023).
- Customer Churn Prediction: Data science helps insurers predict customer churn by analyzing factors like demographics, policy history, and online behavior. This allows for targeted interventions to retain valuable customers (Jones & Sah, 2023).
- Product Development: Data-driven insights enable insurers to develop innovative products tailored to specific customer needs and risk profiles, promoting market competitiveness (Pugnetti & Seitz, 2021).

These are just a few examples of how data science is reshaping the insurance industry. As data collection and analytical capabilities continue to evolve, more transformative applications are expected in the years to come.

2.2. FUNDAMENTALS OF SURVIVAL ANALYSIS

Survival analysis, also known as time-to-event analysis, is a statistical method used to analyze the time-until an event of interest occurs (Cleves et al., 2004). The term "survival analysis" originates from its early applications in clinical research, where predicting the time to death was a primary objective (Klein & Moeschberger, 2003).

Survival analysis shares similarities with regression analysis, as both aim to establish relationships between variables. However, survival analysis has a unique twist: it deals with data that may be censored (Tutz & Schmid, 2016). Censored data arises when the event of interest is not observed for all individuals during the study period.

By considering the time dimension of events, survival analysis provides valuable insights that go beyond simple event prediction. It unveils the factors influencing event occurrence and allows for the estimation of probabilities and risk assessments (Vandenbroucke et al., 2007). This analytical approach empowers researchers in various fields to develop predictive models and make data-driven decisions for improved outcomes.

2.2.1. Survival Data and Censorship

As explained by Fu & Wang (2014), to construct survival analysis data, it is crucial to distinguish between calendar time and duration time within the study.

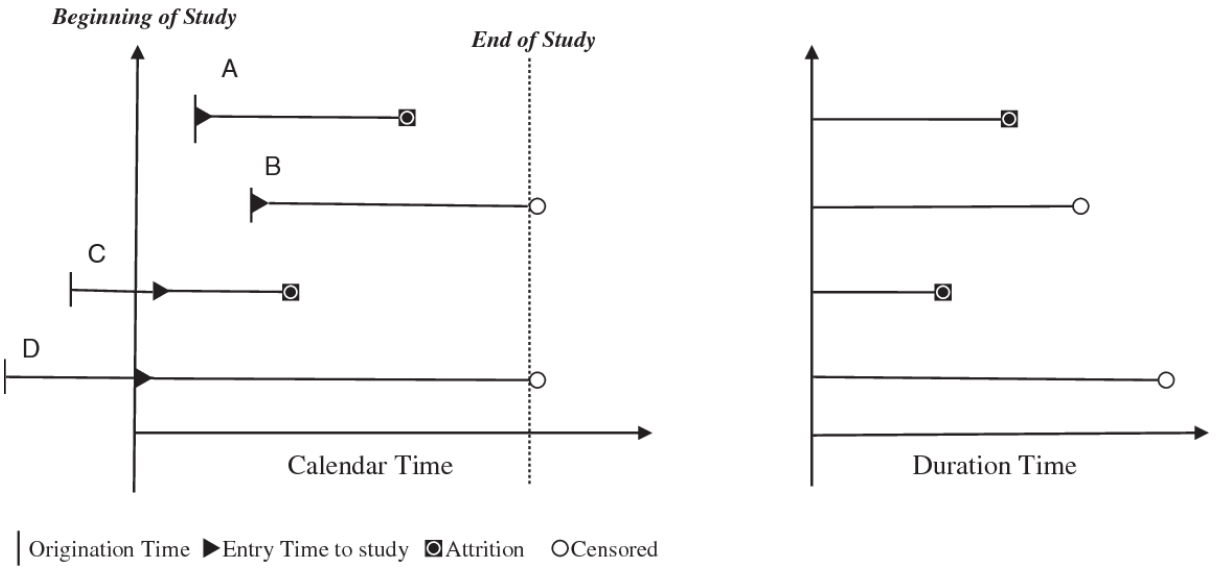

Figure 1: Calendar time vs. duration time in survival analysis (Fu & Wang, 2014)

Figure 1 illustrates the concept of duration time. For a better understanding, the explanation of figure 1 is going to be adapted to the insurance context and individuals (A, B, C, and D) are going to be referred to as policies.

In the calendar time graph, the vertical line "|" denotes the original inception date of policies, and each black triangle signifies the entry time into the study. In the diagram, four

policies (A, B, C, and D) with different inception times are depicted. Policies B and D remained active until the study's conclusion, represented by simple circles to indicate censoring. Conversely, policies A and C, marked with black squares, experienced attrition during the study period. Therefore, in calendar time, both entry and exit times of policies are staggered and can occur at any point throughout the study duration. Duration time refers to the length of time each policy was part of the study, starting at zero and ending either at attrition or at the study's conclusion, referred to as censored.

Data censoring can happen in two cases; the first case is when the study ends before the individual experiences the event, this case is called right-censoring. The second case is when data is not complete, an external event happened that interrupted the study or the occurrence of the target event for a specific individual. In figure 1, policies B and D are right-censored because the cancellation did not occur before the end of the observation period.

Since survival times are susceptible to right-censoring, additional information is needed to capture the complete survival experience. Therefore, survival data requires two essential variables for each individual:

- Survival Time: This represents the observed duration of the state of interest (e.g., active policy) until either the event of interest occurs (e.g., policy cancellation) or the end of the observation period (censoring).
- Status: This variable indicates whether the observed survival time represents the actual duration until the event of interest (uncensored) or the duration until the end of observation (censored).

By considering both survival time and status, it becomes possible to effectively analyze censored data and understand the likelihood of events occurring over time.

2.2.2. Prediction functions in Survival Analysis

In survival analysis, unlike predicting a single future event, the focus often shifts to predicting functions that describe the likelihood of an event occurring over time (Hosmer et al., 2008). Two key functions are used for this purpose: the survival function and the hazard function.

- The Survival Function S(t): It estimates the probability of an individual surviving beyond a specific time point t. In mathematical notation, it's represented as $S(t)=P(T>t)$.

  In the insurance context, the survival function translates to the probability that a policy will remain active beyond a specific time point t. In simpler terms, it tells how likely it is for a policy to survive cancellation for a certain period (Hosmer et al., 2008).

- The Hazard Function h(t): It represents the instantaneous risk of an event occurring at a specific time point t, given that the individual has survived up to that point. In contrast to the survival function, which describes the absence of an event, the hazard function provides information about the occurrence of an event (Therneau & Grambsch, 2000).
  In insurance policies case, it can be thought of as a measure of the "cancellation rate" at a specific time, but only considering policies that haven't been cancelled yet.

  The Cumulative Hazard Function H(t) is a function built upon the hazard function and represents the total accumulated risk of experiencing the event of interest up to a specific time point t. It takes into account all the hazard rates experienced up to that point, including policies that might have already been cancelled before t (Van Houwelingen, 2006). Mathematically, it's calculated by integrating the hazard function over the time interval from zero to t: $H(t) = \int_0^t h(u)\, du$.

These prediction functions form the foundation for various survival analysis models that can be used to estimate event probabilities. Therefore, understanding the survival and hazard functions is crucial to interpret the results of these models and draw meaningful insights from survival data.

2.3. APPLICATIONS OF SURVIVAL ANALYSIS

Survival analysis, a statistical method with origins in medical research, has transcended its initial domain and found valuable applications in diverse fields. Its core strength lies in its ability to analyze the likelihood and timing of events over time. This unique capability makes it a powerful tool for understanding phenomena where the duration until an event occurs is a critical aspect. Here are some examples of its application in various sectors:

- **Medicine**: In the healthcare realm, survival analysis plays a crucial role. As highlighted by Collett (2014) , researchers and clinicians leverage it to predict patient survival times after diagnosis, a crucial factor in treatment planning and prognosis. For instance, survival analysis can be used to estimate the likelihood of a cancer patient surviving five years post-diagnosis. Additionally, it helps analyze recurrence rates of diseases, allowing for early detection and intervention strategies. Furthermore, studying the time to response to a treatment can provide valuable insights into treatment efficacy and guide medication adjustments. (Hosmer et al., 2008)

- **Engineering**: Survival analysis is a valuable asset for engineers, according to Yang et al. (2022). It helps estimate the time to failure of mechanical components, which is crucial for predicting equipment lifespan and ensuring preventative maintenance schedules are in place. By analyzing product lifespans, manufacturers can optimize product design and improve durability. Additionally, survival analysis can be used to

predict warranty claims, allowing companies to allocate resources effectively and set appropriate warranty reserves.

- **Social Sciences**: Survival analysis finds applications in various social science disciplines. As noted by Box-Steffensmeier & Jones (2004), it assists researchers in investigating the duration of unemployment spells, a key metric for understanding labor market dynamics. In marketing, analyzing customer churn – the rate at which customers stop using a service – is crucial for retention strategies. Survival analysis helps identify factors contributing to churn and predict which customers are at risk of leaving. Demographers can also leverage this technique to study the time to marriage in populations, providing insights into social trends and marriage patterns.

- **Insurance**: The insurance sector has found survival analysis particularly valuable (Scriney et al., 2020). Unlike traditional methods that might solely focus on whether an event occurs, survival analysis provides valuable insights into the timing of these events**.** By analyzing the time until policy cancellations, customer churn, or claim occurrences, insurers can gain a deeper understanding of customer behavior and risk profiles. This analytical approach empowers insurance companies to develop predictive models, make data-driven decisions, and optimize strategies for reducing risks and improving outcomes (Kleinbaum & Klein, 2012). For instance, survival analysis can help insurers identify policyholders at risk of cancelling their coverage and develop targeted retention programs.

In conclusion, survival analysis has become an indispensable tool for researchers and practitioners across diverse fields. Nevertheless, its application within the insurance industry holds particular significance. By analyzing the timing of events, insurers gain a deeper understanding of customer behavior and risk profiles. This empowers them to make data-driven decisions, develop targeted strategies, and ultimately achieve improved outcomes. As the insurance industry continues to embrace data science, survival analysis is poised to play an increasingly crucial role in its future success.

3. METHODOLOGY

3.1. DATA EXTRACTION

This section explains how data is gathered from the insurance company's data warehouse to make a dataset suitable for studying survival. The gathered data includes details about active car insurance policies, policy cancellations, customer information, and claims. This dataset forms the basis for analyzing survival patterns in the insurance field.

3.1.1. Data Sources

The data used for this analysis is sourced from multiple tables within the insurance company's data warehouse:

- Table of Current Active Policies: contains information about active automobile insurance policies, including policy numbers, policy start dates, policy end dates, premium payments, coverage details, policy reductions, and other relevant policy-related information.
- Table of Cancellations: stores records of policy cancellations, including the cancellation date, reason for cancellation, and any associated fees or penalties.
- Table of Customer Personal Data: holds demographic and personal information of insurance customers, including age, gender, location, contact details, marketing consents, and other relevant customer-related information.
- Table of Vehicle data: includes information about the insured vehicle, such as brand, model, age, weight, energy consumption, engine power, and other relevant data.
- Table of Claims: contains details of insurance claims filed by policyholders, including claim dates, claim amounts, types of claims (such as accidents, theft, or damage), and the status of each claim (e.g., pending, approved, denied).

3.1.2. Extraction Process

The extraction process for this survival analysis focused on motor policies active on December 31st, 2018. This specific starting point allows to study the survival behavior, for how long the policies tend to survive, for a time range of at least five years. The latest date to observe the survival was set to May 31st, 2024, as the most recent date that could be included, resulting in maximum survival period of 5,4 years. Using SQL queries, data was fetched from the company’s data warehouse. These queries combined information from the previously mentioned tables.

The tables were merged using left joins and inner joins to get the needed data from each table. Left joins were employed to include all records from the primary table of active policies and matching records from the secondary tables of cancellations, customer data, and claims. While, inner joins were used to combine data that existed in all tables, ensuring

that only relevant information was included in the dataset. This method was applied for temporary tables that were created within the process to facilitate the merging of data from different sources. Temporary tables served as intermediary steps in the extraction process, allowing for the integration of information from various tables while maintaining data integrity and consistency. (Ben-Gan, 2016)

During the SQL query process, filters were applied within the joins, such as cancellation motives, type of transaction, and type of client, to ensure data consistency and accuracy. These additional filters helped refine the dataset by including only the most relevant and reliable information for analysis. By incorporating these filters directly into the join conditions, the extraction process ensured that the resulting dataset was coherent and aligned with the research objectives.

Additionally, during the extraction process, two primary target variables were created: "status" and "survival in years". The "status" variable was defined as "true" for policies that were canceled during the study period and "false" for policies that survived until the end of the study period. This binary classification allowed for the differentiation between policies that experienced the event of cancellation and those that did not. The "survival in years" variable was computed to determine the tenure period of each policy in years. This metric measures the time from the beginning of the study on December 31st 2018 until its cancellation, if applicable, or until the end of the study set to May 31st 2024 for policies that remained active. These target variables are crucial components for the model training and predicting the survival probabilities over time for each policy.

In summary, the extraction process brought together all the necessary data for survival analysis from various sources using appropriate joins and filters. This dataset contains policies that were active on December 31st 2018 and their relevant information at the starting date of the study, additionally to the survival status and period that are captured until the end of the observation period.

3.2. DATA EXPLORATION

This section delves into exploring the dataset obtained from the extraction process to gain insights and understand its characteristics. The dataset comprises a total of 300,000 policies and 104 features. These features span various categories including policy details, policyholder demographics, insured vehicle information, other owned policies, claim history, and ultimately, target survival data. This exploration helps identify data characteristics and prepare it for further analysis, ensuring a solid foundation for building insurance risk assessment models.

3.2.1. Exploration of target variables: Status and Survival in years

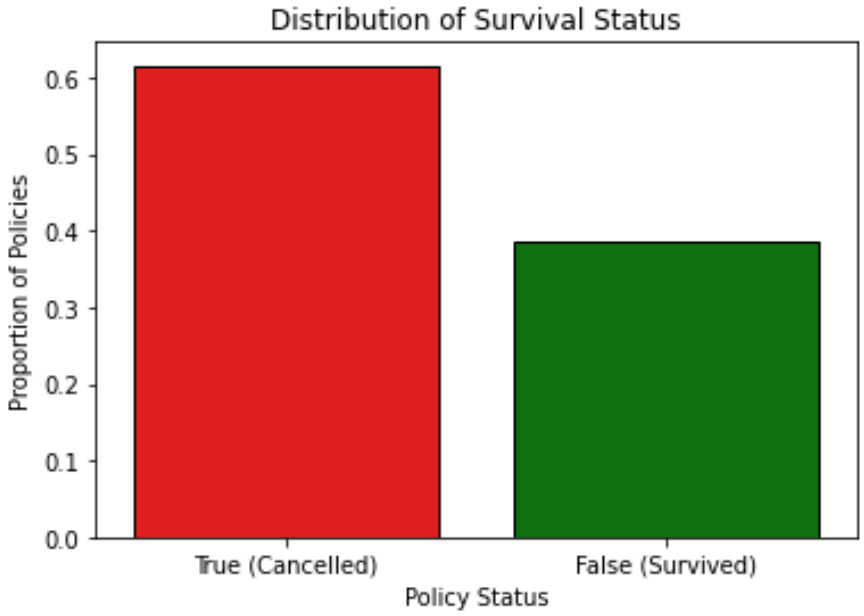

Figure 2: Distribution of survival status

Figure 2 displays the distribution of survival status. 61.5% of the policies were flagged with a status of true, indicating that cancellation occurred during the observation period. These policies are considered uncensored, representing instances where the event of interest was observed within the specified time frame. Conversely, the remaining policies, representing 38.5% of total data, survived and did not experience cancellation until the end of the study period and are thus considered censored.

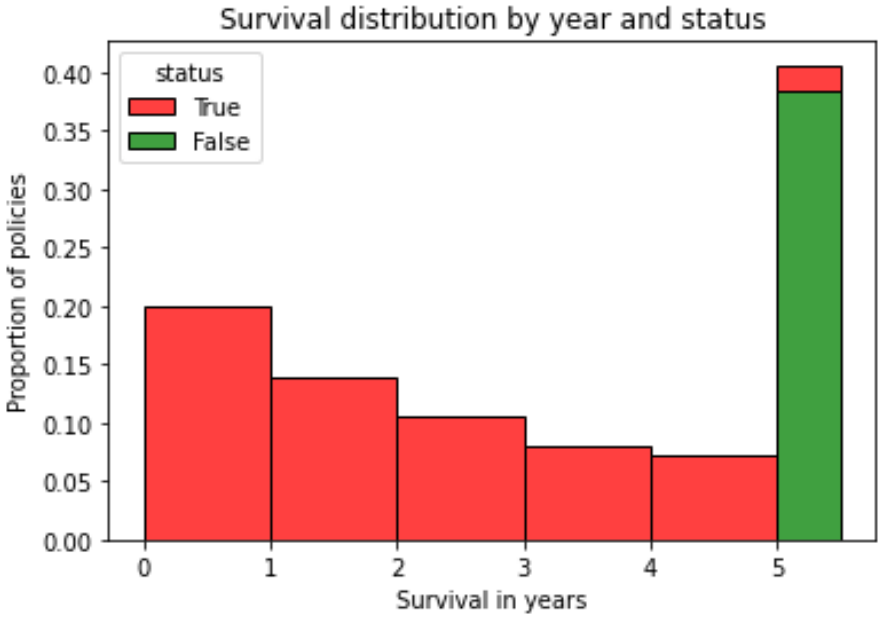

Figure 3: Survival distribution by year and status

According to figure 3, the number of cancellations decreases each year showing 20% of the policies cancelled during the first year of the study, followed by 14% in the second year, 10% in the third year, 8% in the fourth year and 7% in the fifth year. The survived policies are all present in the last year as these policies did not experience a cancellation and they all have a maximum survival period equal to 5.4 years. Additionally, the last bin in the graph displays a small proportion of policies that were cancelled during the period from January 1st 2024 to May 31st 2024.

3.2.2. Effect of explanatory variables on target

- **Policy Age**

Policy age is the difference between the start date of the policy and the beginning date of the study.

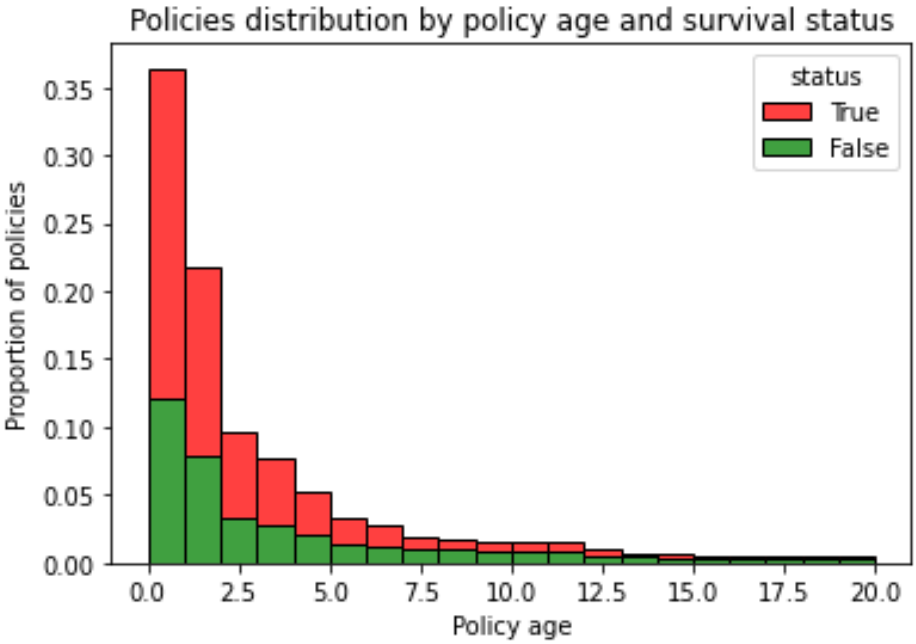

Figure 4: Policies distribution by policy age and survival status

Figure 4 shows that new policies are more present in the dataset than old policies, with policies with 2 years old or less representing 60% of the data. Policies older than 5 years represent 18% of the data. Another aspect that can be detected from figure 4 is the proportion of non-survival among the new policies, cancelled policies represent two thirds of the policies aged 4 years or less. The proportion of cancellations decreases with the increase of policy age. Therefore, older policies have a higher chance to survive.

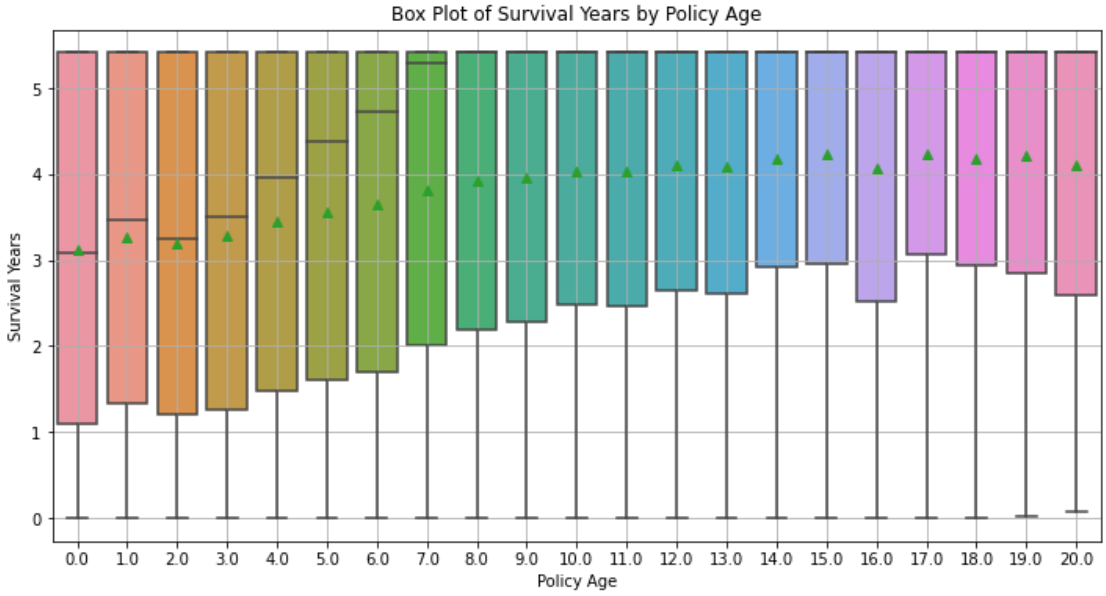

Figure 5: Box plots of Survival years by Policy age

Figure 5 displays box plots of the survival years by policy age. The median line, representing the middle 50% of the data, starts at 3 years of survival for policies aged 0 and has an increasing trend until it reaches the maximum survival value of 5.4 years on policies with 8 years old, indicating that policies aged 8 years and older have more than 50% probability to survive for longer than 5 years. Additionally, the lower quartile, indicating the first 25% of data, increases simultaneously with the age of policy.

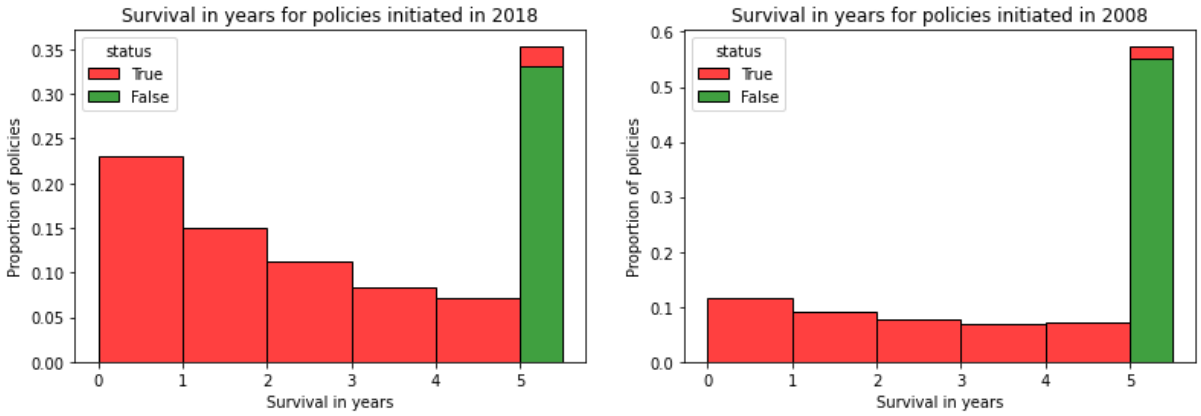

Figure 6: Survival for 0 years old policies versus 10 years old policies

To further analyze the effect of policy age on survival, figure 6 was created to compare the survival for policies initiated in 2018 with 0 years old versus policies initiated in 2008 with 10 years old. First, the survival percentage is 32% for 0 years old policies compared to 55% survival for 10 years old policies. Second, the proportion of cancellations decreases significantly for new policies, going from 23% to 15% to 11% in the first three years after their initiation. On the other hand, policies with 10 years old have a stable cancellation rate, around 10% in all years of the study.

- **Annual value of the policy**

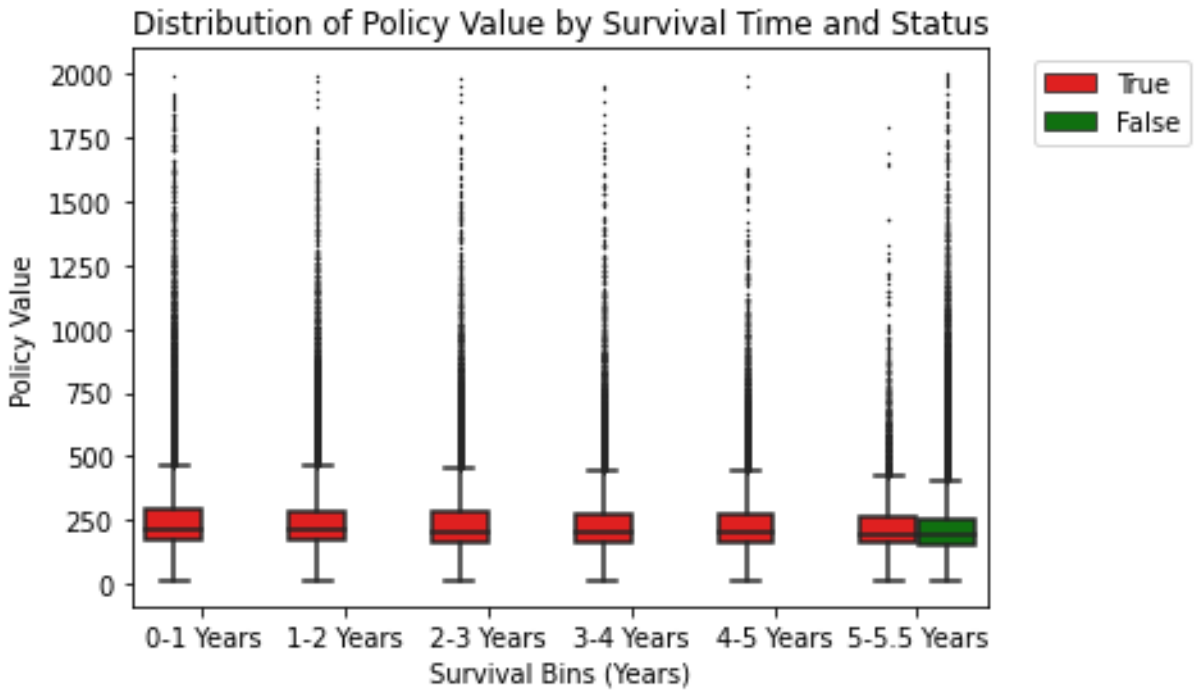

Figure 7: Box plots of Annual Policy Value by Survival Time and Status

Figure 7 shows box plots of annual policy value by survival time and status. The box plots do not show any significant difference in the annual value for the cancelled policies depending on the year. The value of policies that survived is slightly lower, but the difference is negligeable. It can be concluded that the annual value of a policy does not have a significant effect on its survival.

- **Insurance Type and Product Categories**

In motor policies, there are two types of insurance: MTPL and MOD. MTPL refers to Motor Third-Party Liability, it is mandatory for all vehicles, and it covers damages caused to third parties by the insured vehicle. MOD refers to Motor Own Damage, it is optional and an additional coverage that supplements MTPL and protects the policyholder's own vehicle. The studied dataset is composed of 76.6% MTPL policies and 23.4% MOD policies.

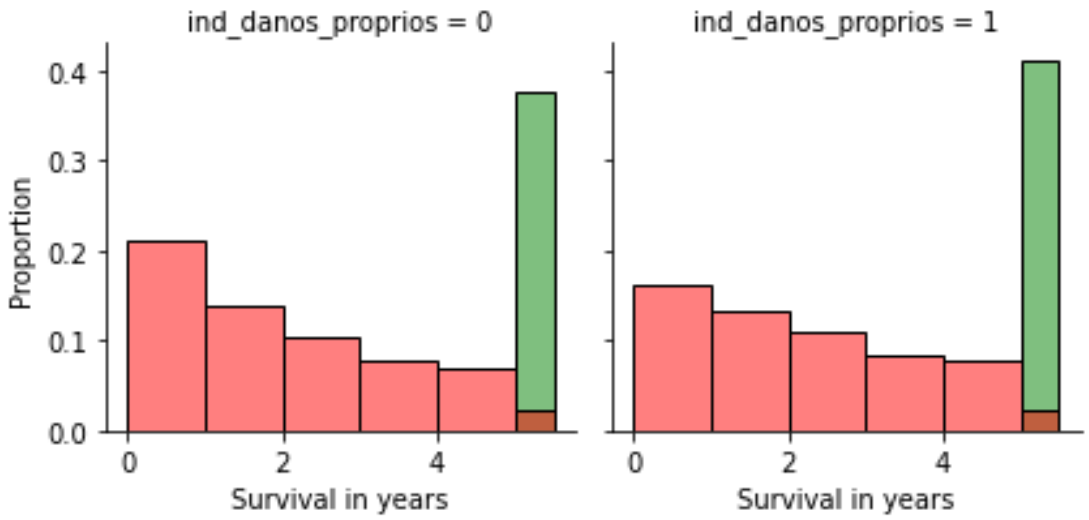

Figure 8: Facet Grid of survival years by insurance type

Figure 8 shows that MTPL policies experience more cancellations on the first year of the study with 21% cancellation rate compared to 16% for MOD policies. On the remaining study year, the percentages are similar. On the survival rate, there is a slight advance for the MOD policies, which can be explained by the difference in the first year's cancellations.

- **Client age**

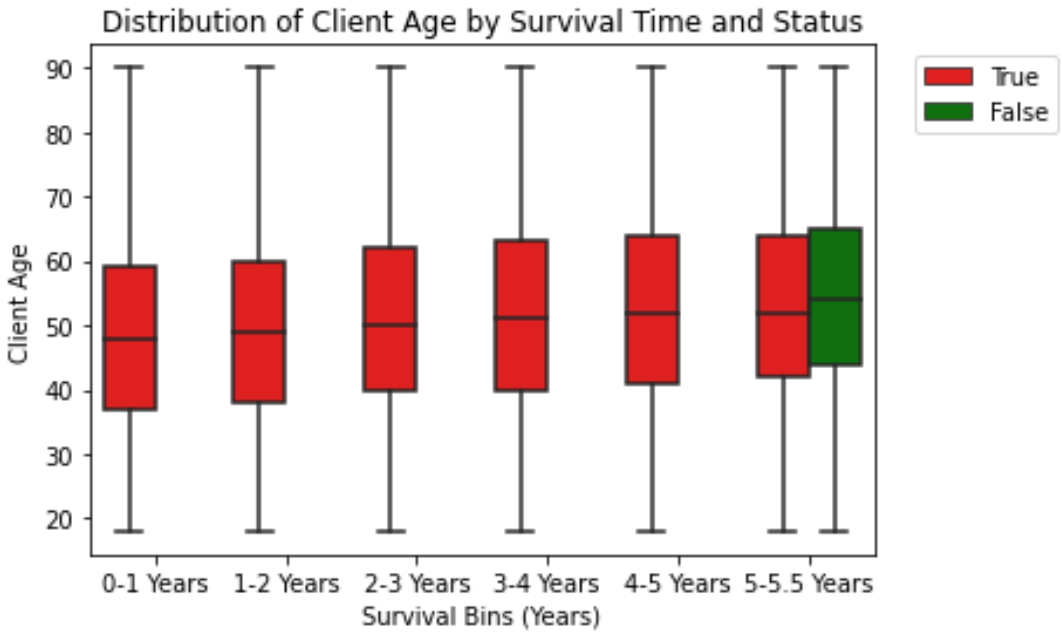

Figure 9: Box plots of Client age by survival years and status

Figure 9 displays box plots of client age by survival years and status. The client age for the non-survived policies increases slightly in each year of the study period. Policies that survived have a client age marginally higher than the ones that did not survive.

- **Client gender**

The dataset of motor policies is composed of 70% male clients and 30% female clients.

Figure 10: Facet Grid of survival time by client gender

Figure 10 shows the survival time and status based on the gender of the client. There are no significant differences between the survival within male and female clients.

3.3. DATA PREPROCESSING

This section outlines the preprocessing steps undertaken to prepare the dataset for further analysis. Data preprocessing involves cleaning, transforming, and organizing the data to ensure its quality and suitability for modeling.

3.3.1. Train Test Split

Before any preprocessing steps were applied, the dataset was divided into training and testing sets. This initial split is crucial to prevent data leakage, which occurs when information from outside the training dataset is used to create the model, leading to overly optimistic performance estimates and poor generalization to new data (Brownlee, 2016). Conducting the train-test split before preprocessing ensures that the model is evaluated on unseen data and thus provides a more accurate measure of its performance in real-world scenarios.

An 80/20 train-test split was used with a fixed random state, with 80% of the data allocated for training and 20% reserved for testing. This approach aligns with standard machine learning practices, ensuring enough data for both training the model and evaluating its performance.

3.3.2. Missing Values Handling

The handling of missing values involved various strategies tailored to different scenarios within the dataset. For the majority of missing values, a pragmatic approach was taken, filling them with the median value. This method ensures robustness against extreme values while maintaining the central tendency of the data. (Jazayeri et al., 2020)

Additionally, in specific cases where missing values were associated with categorical variables, such as the age range of clients, a more nuanced approach was adopted. For example, the 'client seniority' referring to the years a client was registered with the company, was filled using group medians based on the age range of clients, as younger clients have lower years within the company. For instance, if a client's age fell within the range of 18 to 25 years, the missing value was replaced with the median value corresponding to this age group.

Furthermore, when dealing with categorical data related to insurance policies, such as coverage type and vehicle characteristics, missing values were imputed based on predefined mappings specific to the type of policy or insurance coverage.

3.3.3. Outliers Detection and Treatment

Outliers are data points that deviate significantly from the majority of the data, can negatively impact model performance by skewing results. Outliers can arise from various sources like data entry errors, measurement inaccuracies, or inherent data skewness (Hawkins, 1980). Left untreated, they can distort the model's understanding of the underlying relationships between features and the target variable, leading to biased predictions.

To address outliers within the dataset, a systematic approach was implemented to identify and remove extreme values. Visual analysis using boxplots facilitated the detection of these anomalies. Upper and lower thresholds for identifying outliers were manually set based on logical criteria rather than statistical methods, ensuring a practical approach to data cleaning. These thresholds were set conservatively to avoid losing significant amounts of data, removing only the most extreme and anomalous values.

Figure 11: Box plots of some variables to detect outliers

Figure 11 illustrates the distribution of client age, vehicle construction age, and policy annual value variables. The box plots highlight the outliers, making it easier to detect the extreme values and anomalies. The first box plot shows that the age of clients can reach 175 and can be below 18. These values are mostly incorrect as there are no minor clients and no humans have reached such high ages. The second box plot displays negative vehicle construction ages, which is impossible. The third boxplot displays the policies' annual values, with a maximum value of 20,000. Although this value is true, it is much higher than the other values and is considered an extreme value.

Outliers were manually detected to identify anomalies and abnormalities in the data. The few data points identified as outliers were removed, resulting in minimal data loss. Specifically, for the client age variable, outliers were defined with a minimum threshold of 18 years and a maximum threshold of 100 years. For the vehicle construction age, the minimum threshold was set to 0 years and the maximum to 80 years. Similarly, for the policy annual value, the thresholds were set to remove values above 3,000.

A total of 1,133 data points were removed, representing 0.47% of the training set. This subjective approach in setting thresholds is based on logical criteria to effectively identify and remove outliers. This removal is crucial to preserving the integrity and quality of the dataset for more accurate modeling and analysis.

3.3.4. Features Encoding

Most machine learning models require numerical features for computations. Feature encoding transforms categorical variables, which represent discrete labels or classifications, into numerical representations suitable for modeling. This process improves the model's ability to learn relationships between features and the target variable (James et al., 2023). Common encoding techniques include label encoding, one-hot encoding, and binary encoding.

Before performing feature encoding, it is necessary to group categorical features. This involves consolidating categories with low frequencies or similar meanings into broader,

more general categories. Grouping helps reduce the number of unique values in a categorical feature, addressing the issue of high cardinality and enhancing model performance by reducing noise and improving computational efficiency. For instance, if a feature contains various minor categories, these can be grouped into a single "Other" category. This step ensures that the model does not overfit to infrequent categories and can generalize better on unseen data.

Moreover, quasi-constant features, which exhibit little variability and have the same value for a large majority of the observations, were identified and dropped to streamline the dataset and avoid introducing noise into the model. A threshold of 95% was used, meaning that if a feature had the same value in 95% or more of the observations, it was removed from the dataset. This decision was made to enhance model performance by eliminating features that do not contribute meaningful information. During this process, 31 quasi-constant features were dropped, significantly reducing the dimensionality of the dataset. This threshold ensures that only the most static features are removed, retaining those with slight but potentially important variability.

Thereafter, in the feature encoding process, a function was developed to transform categorical variables into numerical representations suitable for modeling. This custom function performs various encoding techniques on categorical features, addressing dimensionality concerns and facilitating efficient model training. Here's a breakdown of its functionalities:

- Dropping Features with High Cardinality: The function identifies and drops features with a high number of categories exceeding a predefined threshold (max_levels). Features with too many categories can introduce noise and negatively impact model performance.
- Binary Feature Conversion: For categorical features with only two distinct levels, the function directly converts them into a single binary indicator column. This simplifies the representation and avoids creating unnecessary dummy variables.
- One-Hot Encoding with Limits: For categorical features with more than two levels but within a specified limit (max_dummies), the function employs one-hot encoding. This approach creates a separate dummy variable for each unique category, effectively expanding the feature space. However, the function can limit the maximum number of dummy variables generated to prevent excessive dimensionality growth.

Each indicator variable took a value of 1 if the corresponding category matched the specified condition and 0 otherwise. This binary encoding facilitated the incorporation of categorical variables into predictive models, enhancing their interpretability and predictive performance.

It is important to note that the preprocessing pipeline also involved dropping certain features alongside applying hot encoding for specific categorical variables. This resulted in a final set of 109 features ready for model training.

3.4. FEATURES SELECTION

Feature selection is a crucial step in the data modeling process, aimed at identifying the most relevant features that contribute significantly to predictive performance while discarding irrelevant or redundant ones. Given that the dataset contains 109 features, it is essential to reduce the dimensionality and focus on the most informative features to enhance model interpretability, reduce overfitting, and improve computational efficiency.

3.4.1. Boruta Selection

The Boruta selection technique is a wrapper method for feature selection used in machine learning. It is employed to identify the most relevant features for predictive modeling. This technique utilizes a random forest classifier, where features are evaluated based on their importance in predicting the target variable. (Kursa & Rudnicki, 2010)

The Boruta feature selection function is applied to the training dataset, which returns three sets of features: accepted features, irresolution features, and rejected features. Accepted features are deemed significant for predictive modeling, irresolution features are less important but can be included in the model training, while rejected features are considered non-significant.

This technique reduced the number of features from 109 to 22 relevant features, resulting in optimizing model performance and enhancing the robustness of the predictive models.

3.4.2. The Elastic Net Penalized Cox Model for Features Selection

In survival analysis, penalized Cox proportional hazards models offer a powerful technique for variable selection, particularly when dealing with high-dimensional data. The standard Cox proportional hazards model estimates the relationship between a set of covariates and the risk of an event (e.g., policy cancellation) over time. It doesn't directly perform feature selection, but provides hazard ratios for each feature, indicating their association with survival. With a large number of features, the Cox model can become unstable and prone to overfitting. Penalized Cox models introduce a penalty term into the model fitting process. This penalty term discourages the coefficients of some features from becoming too large, essentially shrinking them towards zero. There are two main types of penalties used for feature selection in this context:

- L1 Regularization (LASSO): This penalty encourages sparsity. With a strong L1 penalty, features with weak associations become more likely to have coefficients shrunk to zero, effectively removing them from the model. This leads to a more interpretable model with a focus on the most important features.

- L2 Regularization (Ridge): This penalty encourages stability by reducing coefficients but keeping all features in the model. The L2 penalty generally doesn't drive coefficients to zero, but it shrinks them towards zero. This reduces the impact of irrelevant features and improves the overall stability of the model.

The Elastic Net penalized Cox proportional hazards model is an approach that combines L1 (LASSO) and L2 (Ridge) regularization penalties. The Elastic Net introduces a hyperparameter alpha that controls the balance between L1 and L2 penalties in the objective function. Tuning this parameter is crucial for achieving optimal model performance and feature selection.

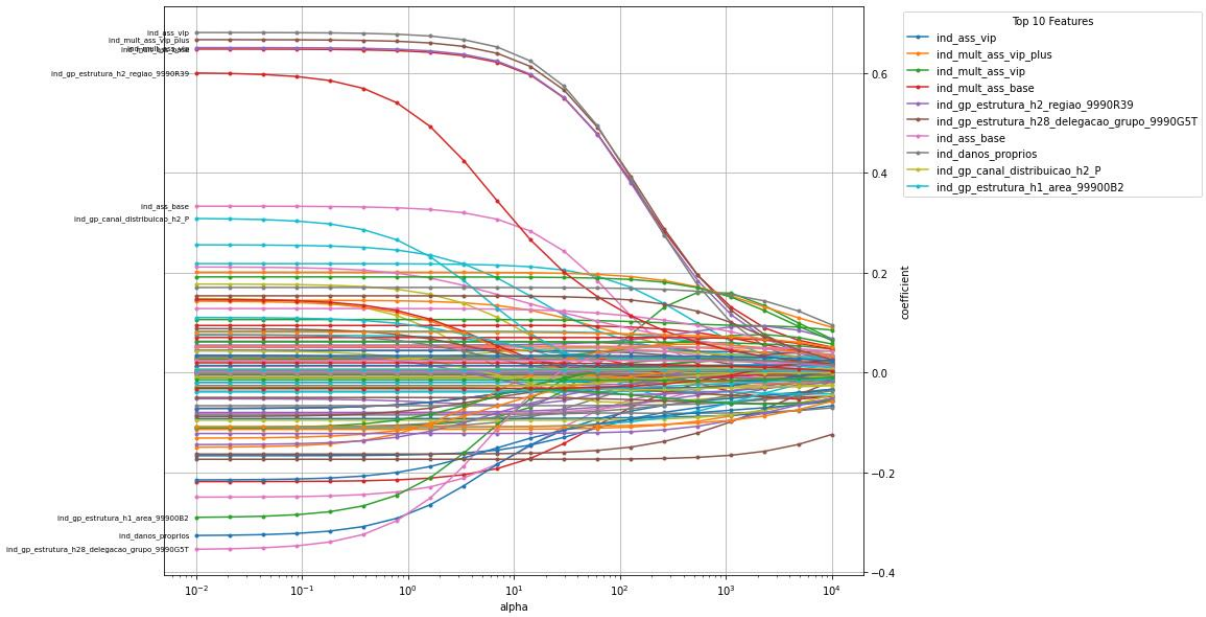

Figure 12: Alpha tuning for the Elastic Net Cox

Figure 12 illustrates the impact of the hyperparameter alpha on the coefficients of features in the Elastic Net Cox model. A higher alpha supports a discriminative behavior similar to L1. It can drive more coefficients to zero, leading to a smaller set of features but also potentially missing some relevant ones. Oppositely, a lower alpha leans towards L2 behavior, keeping all features but shrinking their coefficients. It might not achieve as much sparsity as a higher alpha, but it can improve model stability.

The Elastic Net Cox model was applied for feature selection with an alpha of 0.02 and an L1 ratio of 0.9 after experimenting with multiple combinations. This configuration prioritizes sparsity due to the high dimensionality of the preprocessed data while still maintaining a sufficient number of features to fit the model effectively, resulting in a model with 35 features having non-zero coefficients. These features are the most influential factors for predicting survival outcomes in the data, while features with coefficients driven to zero by the penalty can be considered less relevant for this specific analysis.

3.5. MODEL TRAINING

Survival analysis models are well-suited for predicting the time until an event occurs, such as policy cancellation in insurance applications. This chapter explores three survival analysis models offered by scikit-survival: Penalized Cox Proportional Hazards Model, Random Survival Forests, and Gradient Boosting Survival Models.

3.5.1. Penalized Cox Proportional Hazards Model

The Cox proportional hazards (PH) model, first proposed by Sir David R. Cox (1972), is a widely used semi-parametric approach for survival analysis. It estimates the relationship between a set of predictor variables and the hazard function, which represents the instantaneous probability of an event occurring given that the subject has survived up to that time. Unlike other models, Cox PH doesn't directly model the survival function itself.

The penalized Cox PH model extends the traditional approach by incorporating a penalty term during model fitting. This penalty term helps to control model complexity and prevent overfitting. The scikit-survival library offers the “coxph” class for implementing penalized Cox PH models. This model is a good choice when the proportional hazards assumption is met and interpretability of the results is important, as the coefficients can be interpreted similarly to those in a logistic regression model.

A key hyperparameter in this model is alpha, which controls the strength of the penalty term. Grid Search CV technique was employed to efficiently evaluate the model's performance across a range of alpha values. This was crucial to find the optimal balance between model complexity and capturing important relationships within the data.

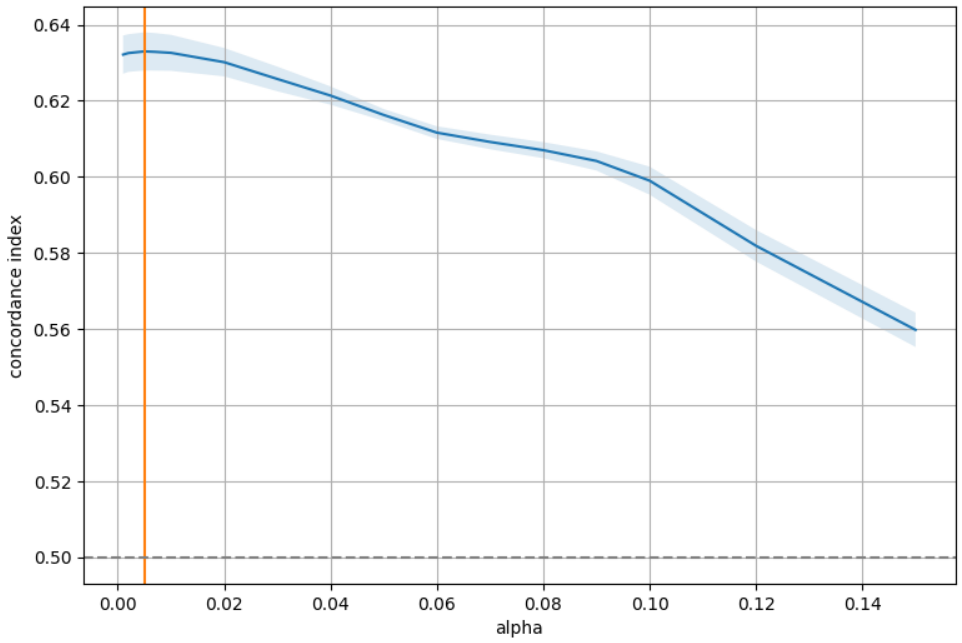

Figure 13: Concordance index for the alpha values in the grid search

According to Figure 13, a low value of 0.005 for alpha was chosen for the Cox model since this value provides the highest concordance index. In this part of model training, a low alpha value was opted for, which does not discriminate against features as this was already done in the feature selection. This choice ensures that the model captures important relationships and slightly penalizing features to avoid overfitting.

3.5.2. Random Survival Forest

Random survival forests, introduced by Hemant Ishwaran et al. (2008), are ensemble methods that combine the predictions of multiple decision trees to improve the overall model performance and robustness. Each decision tree in the forest is built on a random subset of features and a random subset of the training data. This helps to reduce the variance of the model and prevent overfitting.

The scikit-survival library provides the RandomSurvivalForest class for building random survival forests. This model is a good option when dealing with high-dimensional data or when the relationship between features and survival time is complex and non-linear.

To optimize the Random Survival Forest model, hyperparameter tuning was conducted using Random Search CV technique. This involved evaluating the model's performance across a range of values for both the number of trees (n_estimators) and the maximum depth of trees (max_depth). Random Search was chosen for its efficiency in exploring a defined parameter space. By randomly sampling a fixed number of parameter combinations, it provides a good balance between thoroughness and computational feasibility.

3.5.3. Gradient Boosting

Gradient boosting is a powerful machine learning technique that can be adapted for survival analysis. The application of Gradient Boosting in Survival analysis was introduced by Chen et al. (2013). Similar to random survival forests, gradient boosting models combine the predictions of multiple weak learners in a sequential manner. At each stage, a new learner is added to the ensemble that focuses on improving the overall prediction error for the previously misclassified samples.

To optimize the Gradient Boosting model, hyperparameter tuning was conducted using Random Search. This involved evaluating the model's performance across a range of values for several key parameters, including the learning rate, number of boosting rounds, and maximum depth of trees.

3.6. MODEL ASSESSMENT

Evaluating the performance of survival analysis models requires metrics that assess their ability to rank subjects based on predicted survival times and predict the likelihood of an event occurring within a specific timeframe. These metrics differ from traditional accuracy

metrics used in classification tasks, as they account for the time-dependent nature of survival data.

- **Harrell's Concordance Index (C-index)**

The C-index is a widely used concordance measure that evaluates a model's ability to correctly rank subjects based on their predicted survival times. It calculates the proportion of subject pairs (i, j) where the subject with the higher predicted risk of experiencing the event (i) indeed experiences the event before the subject with the lower predicted risk (j).

A C-index value of 1 indicates perfect concordance, meaning the model always correctly ranks subjects. Conversely, a value of 0.5 implies no better performance than random guessing. However, it is important to note that Harrell's C-index can be biased upwards, particularly when dealing with a significant amount of censored data. This can lead to an overestimation of the model's true performance.

- **Uno's C-index**

Uno's C-index offers a more robust alternative to Harrell's C-index, especially when a substantial portion of the data is censored. This estimator takes censored data into account by incorporating inverse probability of censoring weights, leading to a more accurate reflection of the model's ability to rank subjects based on their predicted survival times.

- **Time-dependent Area Under the ROC Curve (tAUC)**

Unlike the traditional ROC curve used in classification problems, the tAUC metric is specifically designed for survival analysis. It evaluates the model's ability to discriminate between subjects at risk of experiencing an event within a specific time horizon (t) as opposed to a single point in time. This is particularly valuable when the exact timing of the event is uncertain, but predicting its occurrence within a timeframe is crucial. The tAUC is calculated by integrating the sensitivity and 1-specificity values over a range of possible thresholds for the predicted risk score at each time point t. A higher tAUC value (closer to 1) indicates superior performance in predicting event occurrence within a specific timeframe.

- **Integrated Time-dependent Brier Score (IBS)**

The IBS builds upon the concept of mean squared error (MSE) used in regression tasks. It measures the average squared difference between the predicted probability of an event occurring at each time point (t) and the actual observed probability. A lower IBS indicates better agreement between the model's predicted probabilities and the actual observed data across the entire follow-up period. The time-dependent aspect ensures we consider the entire timeframe, not just a single point. This integrated IBS provides a comprehensive assessment of both model discrimination (separating high-risk from low-risk subjects) and calibration (accuracy of predicted probabilities).

3.7. SCORING NEW DATA

The final step in the model deployment process is scoring new data to make predictions on unseen instances, including updated values for existing policies and new policies. Periodically, typically every three months, the updated data is passed through the trained model, which generates predictions for each policy's probability of survival for the upcoming years.

The scoring process entails applying the trained model to the updated dataset, where each policy's features are input into the model to produce corresponding survival probability estimates. These estimates reflect the likelihood of a policyholder's policy enduring for the specified duration, considering any changes or updates to their circumstances captured in the updated data.

Once the predictions are obtained, clients with policies identified as having low survival probabilities are targeted in marketing campaigns. These campaigns aim to engage with clients proactively, offering tailored solutions or incentives to minimize their churning risk.

In summary, this chapter outlined the methodological process for conducting survival model analysis in the insurance sector. Covering data extraction to model assessment, each stage was explained thoroughly to support the objective of using machine learning to retain clients. By implementing these methods, insurers can utilize data-driven insights to inform decision-making and effectively manage risks, thus improving client retention strategies within the insurance industry.

4. RESULTS AND DISCUSSION

This chapter analyzes the performance of three machine learning survival models developed to predict the survival function among insurance policies: Penalized Cox Proportional Hazards Model, Random Survival Forest, and Gradient Boosting. The focus here is to assess their effectiveness in predicting policy survival probabilities over a five-year horizon. For this purpose, advanced evaluation metrics are applied such as the Concordance Index, the Time-dependent Area Under the ROC Curve, and the Integrated Time-dependent Brier Score. By analyzing these metrics and their corresponding visualizations, the chapter aims to identify the model that demonstrates the most effective survival prediction capabilities.

4.1. MODELS PERFORMANCE

Following the development of three survival prediction models in the previous chapter, this section evaluates their performance in predicting policy survival probabilities over a five-year horizon. A combination of metrics is used to assess their effectiveness.

Table 1: Comparison of models' performance using C-index and IBS metrics

Model	C-index	IBS
Penalized Cox Proportional Hazards	0.631	0.185
Random Survival Forest	0.623	0.182
Gradient Boosting	0.621	0.187
Random	0.500	0.250

Table 1 summarizes the performance of three survival prediction models (Penalized Cox Proportional Hazards, Random Survival Forest, and Gradient Boosting) along with a random guessing baseline, evaluated using the Concordance Index (C-index) and Integrated Brier Score (IBS). Random guessing was included to establish a baseline for comparison and understand how much better the models perform compared to simply picking survival probabilities at random.

All three models achieve a C-index above 0.6, indicating a moderate ability to discriminate between high-risk and low-risk policies. There's a slight difference between the models, with the Penalized Cox Proportional Hazards model having the highest C-index (0.631) and Gradient Boosting the lowest (0.621). However, the difference is very small, suggesting all models might perform similarly in terms of ranking policies.

In terms of Integrated Brier Score, all models show lower IBS compared to the random guessing baseline (0.250). This indicates they outperform random guessing in terms of calibration, meaning their predicted survival probabilities are closer to the actual observed outcomes. Among the models, Random Survival Forest has the lowest IBS (0.182), followed by Penalized Cox Proportional Hazards (0.185) and Gradient Boosting (0.187). This suggests Random Survival Forest might have a slight edge in terms of calibration.

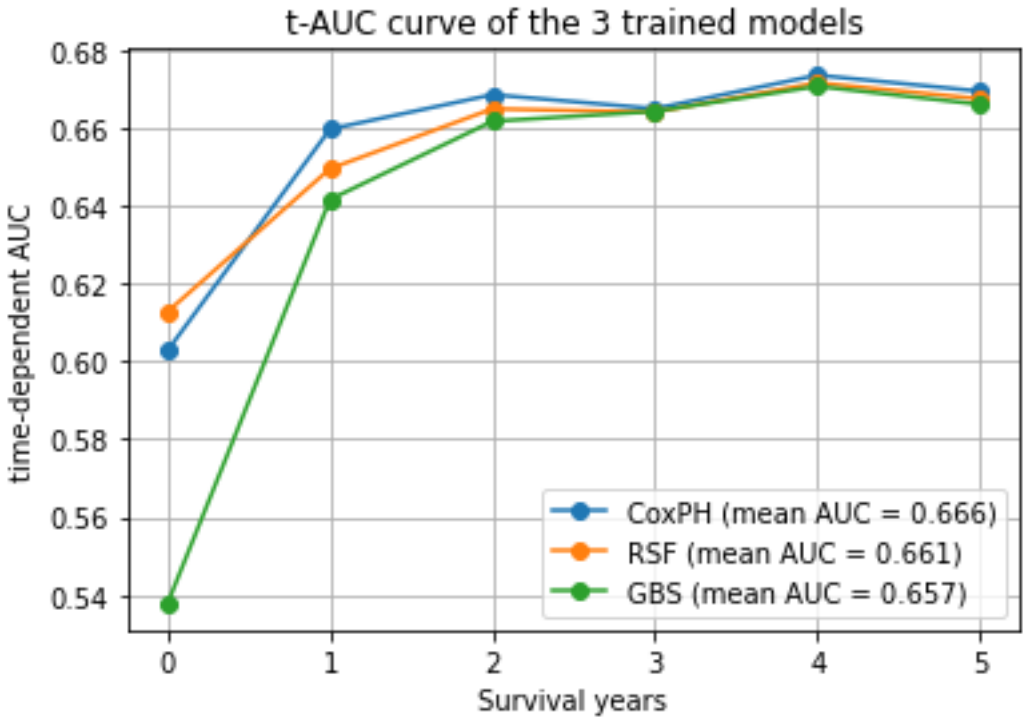

Figure 14: Comparison of the t-AUC curves of the trained models

Figure 14 depicts the time-dependent Area Under the ROC Curve (t-AUC) for the three survival models. The three models exhibit challenges in predicting survival probabilities for year 0, notably, Gradient Boosting showing the weakest performance. The predicting performance improves significantly for the three models in the next years of study and converges over time, suggesting similar discrimination capabilities in later years. This indicates an increasing ability to discriminate between high-risk and low-risk policies as more time and survival data become available.

While the mean t-AUC values are relatively close (CPH: 0.666, RSF: 0.661, GBS: 0.657), a closer look reveals some key differences. CPH maintains a relatively stable discrimination ability throughout the five years, making it a strong choice for consistent risk assessment across the entire timeframe is crucial. Comparably, RSF offers a good balance between performance and stability. While its average t-AUC is close to CPH, it exhibits less variation, suggesting a more consistent performance across years compared to CPH. While GBS displays the lowest t-AUC average with the most significant variation.

4.2. Model Selection and Trade-offs

After evaluating the performance of the three models using the C-index, IBS, and t-AUC, this section discusses the key trade-offs involved in selecting the most suitable model for the insurance policy survival prediction task. Due to its consistently lower performance across all metrics, the Gradient Boosting model was excluded from further consideration.

Evaluating both CPH and RSF models revealed very close performance. Despite both achieving moderate discrimination ability (C-index) and outperforming random guessing in prediction accuracy (IBS), key differences emerged. While RSF displayed a slight advantage in calibration accuracy (IBS) and exhibited less variation in discrimination power throughout the five years (t-AUC), CPH consistently maintained a high level of discrimination (t-AUC) and displayed a slightly higher C-index.

Beyond the close performance in discrimination (C-index) and calibration (IBS), other factors influenced the model selection. Computationally, CPH is generally known for its efficiency compared to RSF. This translates to both faster training times and less resource consumption during prediction, making it advantageous in scenarios where computational resources are limited. Additionally, CPH offers some interpretability through its estimated coefficients. This allows for understanding how specific features influence the predicted risk of policy lapse, which can bring valuable insights for decision-makers. Considering these factors, the Penalized Cox Proportional Hazards model emerged as the most suitable choice for the insurance policy survival prediction task.

4.3. Predicting Policy Tenure

The Penalized Cox Proportional Hazards model has been employed to predict policy tenure. The CPH model's strength lies in its ability to generate a survival probability for each policy, offering valuable insights into their expected longevity. While visualizing individual survival curves for each policy would be ideal for a closer look, the high volume of data makes this impractical.

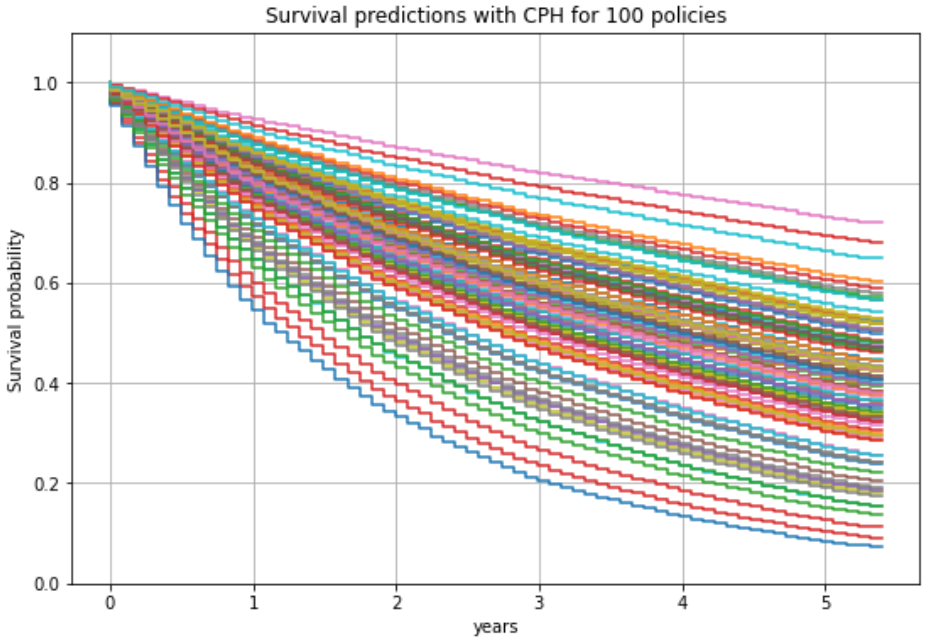

Figure 15: Survival predictions of CPH for a sample of 100 policies

Figure 15 depicts a sample of predicted survival probability curves. As displayed, there is significant variability among the curves, highlighting the model's ability to personalize risk assessment for each policy. Some policies show a steeper decline, indicating a higher risk of lapsing in the earlier years. Others display a more gradual decrease, suggesting a lower risk and potentially longer policy life. Furthermore, the predicted survival probabilities exhibit a general decline over time, reflecting the natural course of insurance policies where some lapse or are canceled as years progress.

The model empowers the insurance company to leverage the predicted survival probabilities and prioritize policies for retention efforts. High-risk policies with a steeper decline in predicted survival can be targeted with interventions to mitigate lapse risk. Early identification of at-risk policies is another area where the model can be particularly valuable. By analyzing the predicted survival probabilities, the model can potentially flag policies with a high chance of lapsing soon after issuance. This allows for proactive measures to be taken before the policyholder decides to cancel.

It's important to acknowledge that the model's predictions are based on historical data and may not capture all factors influencing policy survival. Unforeseen events or external variables not included in the model could impact renewal decisions.

4.4. Unveiling Factors Impacting Retention

Understanding the factors that influence policy retention is crucial for insurance companies to develop effective strategies for minimizing churn and maximizing customer lifetime value. The analysis identified several features that influence, to different extents, the survival of a policy.

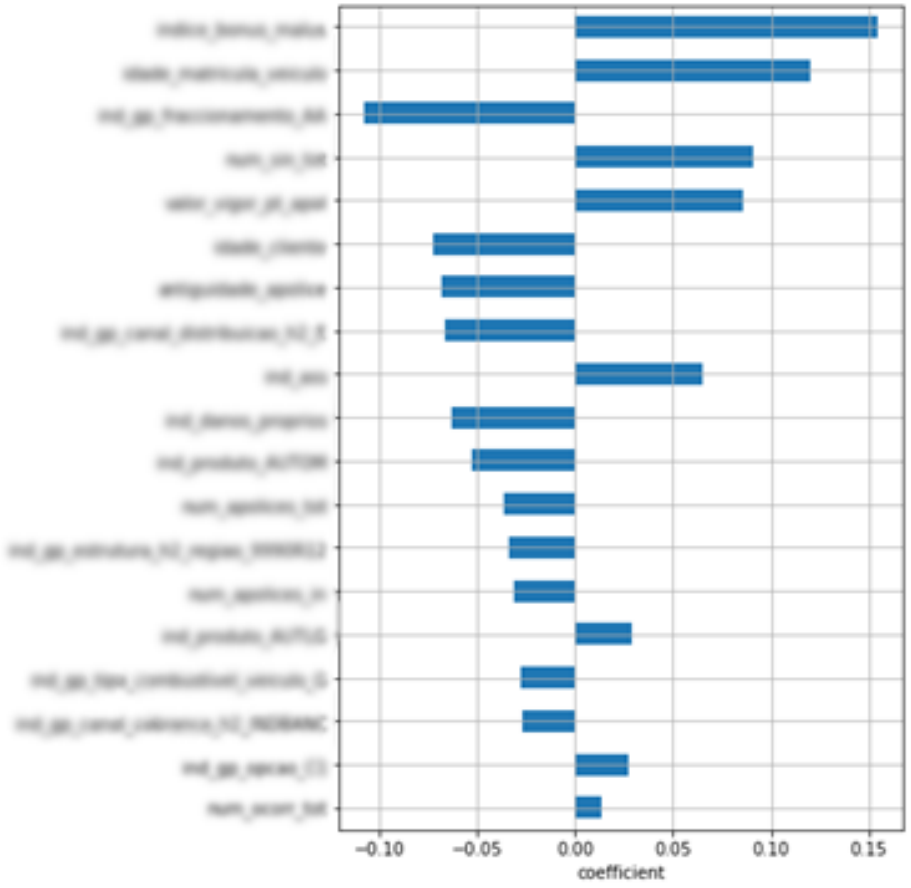

Figure 16: Coefficients of the main features influencing the survival function

(Feature names have been blurred to maintain confidentiality)

Figure 16 illustrates the coefficients of the main features influencing the survival prediction. The most significant factor identified was related to the policyholder's risk rating, which reflects the increase in premiums based on the client's history, mainly previous claims. Clients with a higher risk rating have a higher probability of canceling their policies, likely due to the rise in paid premiums.

The second most significant factor is the age of the vehicle's registration. Older registrations tend to correlate with lower renewal rates, possibly because older vehicles are more likely to be sold or decommissioned, making the policy unnecessary. Additionally, the third most important factor are clients categorized under a specific company code showing higher survival chances than those in other categories.

The total number of claims and the annual value of the policy are the next on the list. As these variables increase, the risk of cancellation also increases. Oppositely, a higher age of the client and a higher age of the policy influence positively the retention. Older clients and older policies tend to survive longer than the younger ones. It is possible to say that they became loyal to the company.

In conclusion, this analysis has revealed valuable insights into the key factors influencing policyholder retention. By understanding these factors, insurance companies can develop targeted strategies to address customer concerns and improve renewal rates. For instance, offering competitive premiums for safe drivers, with low risk rating, remains crucial. However, retaining policyholders with higher risk ratings requires different strategies. This could involve implementing risk-based pricing models while also offering additional benefits like accumulating discounts points for driving safely. This will reduce their risk, which is beneficial for the insurance and for them to benefit from lower prices and be safe. Similarly, while older clients and policies demonstrate higher retention, strategies are necessary to retain younger policyholders with potentially less established loyalty. This might involve offering introductory discounts, targeted marketing campaigns emphasizing the value proposition for younger demographics, or providing flexible payment options.

5. CONCLUSIONS AND FUTURE WORKS

This thesis explored the application of machine learning for predicting policy survival probabilities in the motor insurance industry. Policy cancellation poses a significant challenge for insurance companies, impacting their financial stability and growth.

The research employed survival analysis, a statistical technique ideal for analyzing time-to-event data such as policy tenure. This approach enabled the modeling of policy duration until cancellation occurs. The investigation pursued two key objectives:

- **Predicting Policy Tenure**: The aim was to construct a survival model capable of estimating the probability of a policyholder renewing their policy at any given time point. This empowers insurers to proactively identify policies at risk of cancellation and implement targeted retention strategies.

- **Exploring Retention Factors**: By analyzing the survival model, the research sought to identify factors influencing policy duration. These factors encompass demographics, vehicle information, claims history, and pricing strategies. Understanding these drivers empowers marketing departments to develop data-driven retention strategies that address specific customer needs and behaviors.

The research compared the performance of three machine learning models adequate for survival analysis: Penalized Cox Proportional Hazards, Random Survival Forest, and Gradient Boosting. The three models demonstrated moderate ability to discriminate between high-risk and low-risk policies. Ultimately, Penalized Cox Proportional Hazards emerged as the preferred choice due to its:

- Strong and consistent discrimination power throughout the study period compared to Gradient Boosting.
- Faster training and prediction times compared to Random Survival Forest.
- Ability to offer interpretability through estimated coefficients, providing insights into how features influence policy cancellation risk.

This research lays the groundwork for further exploration in the area of policy survival prediction and retention analysis. Here are some potential future research directions:

- **Model Improvement**: Investigate techniques for further enhancing the predictive accuracy of survival models. This could involve exploring advanced feature engineering methods, hyperparameter tuning strategies, or incorporating additional data sources.
- **Deep Learning Approaches**: Explore the application of deep learning architectures like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for survival analysis. These models could potentially capture complex temporal

patterns in policy data that might not be easily captured by traditional survival models. (Wiegrebe et al., 2024)

- **Study Causal Impact of Retention Strategies**: Future research could explore causal inference techniques to understand the true causal effect of specific retention strategies (discounts, loyalty programs) on policyholder renewal rates. This would go beyond identifying influential factors to establish a clear link between interventions and their impact on customer behavior, enabling data-driven decisions for maximizing retention efforts.

By exploring these possible future works, the field can continually improve the accuracy and interpretability of policy survival prediction models. This, in turn, empowers insurance companies to develop more effective retention strategies, ultimately enhancing customer lifetime value.

BIBLIOGRAPHICAL REFERENCES

Agarwal, S. (2023). An intelligent machine learning approach for fraud detection in medical claim insurance: A comprehensive study. *Scholars Journal of Engineering and Technology, 11*(9), 191-200. https://doi.org/10.36347/sjet.2023.v11i09.003

Albrecher, H., Bommier, A., Filipović, D., et al. (2019). Insurance: Models, digitalization, and data science. *European Actuarial Journal, 9*(2), 349-360. https://doi.org/10.1007/s13385-019-00209-x

Anh, N. V., & Duc, T. M. (2024). Big data-driven predictive modeling for pricing, claims processing, and fraud reduction in the insurance industry globally. *International Journal of Responsible Artificial Intelligence, 14*(2), 12–23. https://neuralslate.com/index.php/Journal-of-Responsible-AI/article/view/82

Ben-Gan, I. (2016). *T-SQL fundamentals*. Microsoft Press.

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling: A guide for social scientists* (Online version). Cambridge University Press. https://doi.org/10.1017/CBO9780511790874

Brownlee, J. (2016). Overfitting and underfitting with machine learning algorithms. *Machine Learning Mastery, 21*, 575.

Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/2013/873595

Cleves, M. A., Gould, W. W., & Gutierrez, R. G. (2004). *An introduction to survival analysis using Stata* (revised ed.). Stata Press. https://www.stata-press.com/books/preview/saus3-preview.pdf

Collett, D. (2014). *Modelling survival data in medical research* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18041

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Fu, L., & Wang, H. (2014). Estimating insurance attrition using survival analysis. *Variance, 8*(1), 55-72. https://www.casact.org/sites/default/files/2021-08/Estimating-Insurance-Attrition-Fu-Wang.pdf

Hawkins, D. M. (1980). Multivariate outlier detection. In *Identification of outliers. Monographs on applied probability and statistics* (pp. 104-114). Springer, Dordrecht. https://doi.org/10.1007/978-94-015-3994-4_8

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data*. John Wiley & Sons. https://doi.org/10.1002/9780470258019

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in Python*. Springer Cham. https://doi.org/10.1007/978-3-031-38747-0

Jazayeri, A., Liang, O. S., & Yang, C. C. (2020). Imputation of missing data in electronic health records based on patients' similarities. *Journal of Healthcare Informatics Research, 4*(3), 295–307. https://doi.org/10.1007/s41666-020-00073-5

Jones, K. I., & Sah, S. (2023). The implementation of machine learning in the insurance industry with big data analytics. *International Journal of Data Informatics and Intelligent Computing, 2*(2), 21-38. https://doi.org/10.59461/ijdiic.v2i2.47

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer. https://doi.org/10.1007/b97377

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4419-6646-9.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Pugnetti, C., & Seitz, M. (2021). Data-driven services in insurance: Potential evolution and impact in the Swiss market. *Journal of Risk and Financial Management, 14*(5), 227. https://doi.org/10.3390/jrfm14050227

Scriney, M., Nie, D., Roantree, M. (2020). Predicting Customer Churn for Insurance Data. In: Song, M., Song, IY., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) *Big Data Analytics and Knowledge Discovery. DaWaK 2020*. Lecture Notes in Computer Science(), vol 12393. Springer, Cham. https://doi.org/10.1007/978-3-030-59065-9_21

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. *Statistics for Biology and Health*. Springer. https://doi.org/10.1007/978-1-4757-3294-8

Tutz, G., & Schmid, M. (2016). *Modeling discrete time-to-event data*. *Springer Series in Statistics*. Springer. https://doi.org/10.1007/978-3-319-28158-2

Van Houwelingen, H. C. (2006). Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics, 34*(1), 70–85. https://doi.org/10.1111/j.1467-9469.2006.00529.x

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., & STROBE Initiative (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology (Cambridge, Mass.), 18*(6), 805–835. https://doi.org/10.1097/EDE.0b013e3181577511

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review, 57*(1), 65. https://doi.org/10.1007/s10462-023-10681-3

Yang, Z., Kanniainen, J., Krogerus, T., & al., et. (2022). Prognostic modeling of predictive maintenance with survival analysis for mobile work equipment. *Scientific Reports, 12*, 8529. https://doi.org/10.1038/s41598-022-12572-z

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa