



**Politecnico  
di Torino**

POLITECNICO DI TORINO

Corso di Laurea in Digital Skills for Sustainable Societal Transitions

Tesi di Laurea

**Data-driven approach to predict  
policy cancellation and improve  
customer retention in the  
insurance industry**

**Relatore**

Gianvito Urgese

**Candidato**

Nargiza Mirzabekova

December 2024



# Abstract

The world is rapidly changing and restructuring with the development of technology. In order not to lose their competitiveness, companies are also adapting to these changes, digitalizing and optimizing processes. This also applies to the insurance sector. If yesterday insurance was associated with papers, now this sector is adopting the latest practices provided by the world of technology and AI. The success of an insurance business directly depends on the satisfaction of customers who purchase this insurance. Customers may have different needs, ranging from price sensitivity to the fact that they need to receive effective communication on time on questions that have arisen regarding coverage. Today, when it is easy to compare service alternatives, customer retention is becoming more difficult. To improve customer retention, it is necessary to study patterns in their behaviour. Digitalization simplifies this task by providing access to data that can be analysed. One of the valuable sources of insights is insurance cancellation data. This data contains open and hidden reasons why customers decided to leave. Exploratory data analysis can answer simple questions like how many people decided to stop their relationship with an insurance company and when. However, to dig deeper, the up-to-date technologies in the world of AI can be used.

This thesis addressed the problem of customer retention in the insurance industry, by analysing the cancellation data provided by the real insurtech company with around 7300 entries. The objective was not to create a retention plan but to provide a data-driven understanding of which customers were at risk of churn and needed to be targeted by marketing strategies. This means that based on the provided data, it was important to reveal which patterns lead to cancellation.

To achieve that, the factors affecting the cancellation were hypothesized, and then the data points which could represent it were found. To predict which users might cancel their coverage, machine learning and deep learning techniques were used. These methods included such algorithms as Decision Trees, Logistic Regression, Ensemble models such as Random Forest and Gradient Boosting, deep learning models such as Artificial Neural Networks (ANN) and state-of-the-art Large Language Model (LLM). All these models were used to test if they could identify which users might keep their policies active and which might cancel. Such a predictive tool can be useful for insurance companies to better understand the cancellation

data and take preventive measures. Such models can stand as a risk assessment tool to detect potential customer loss. Earlier detection and further retention actions can keep the customers and save the business revenue for a company.

To implement this approach, first, the data was collected and cleaned. The next step was selecting features to be used. The dataset had many columns but not all of them were needed. All personal information about customers was removed, so there was no way to identify individuals. The columns which can leak post-cancellation information were also removed. The following step was to create workflows in the KNIME analytics platform, which allows selecting well-known machine learning techniques, testing them and deploying a model. Also, the model can be optimized and evaluated.

Four conventional models which are Decision Tree, Logistic Regression, Random Forest and Gradient Boosting were developed and evaluated using KNIME. Ensemble models achieved the highest accuracy, with 87.47% for Gradient Boosting and 86.01% for Random Forest classifier. The accuracy of the Decision Tree was lower at 84.59%, but still better than Logistic Regression, which achieved 80.99% accuracy. The lower performance of the Logistic Regression highlighted that there were nonlinear relationships in the dataset which it could not capture as well as more advanced techniques as Gradient Boosting could.

Next, to try a more flexible option to test deep learning in the classification task, the ANN was developed using the machine learning library TensorFlow and Python. In contrast to the no-code platform, KNIME, this approach allowed taking direct control in building and optimizing the model. The accuracy achieved by ANN was 83% which could be further improved by hyperparameter-tuning.

The last method tested was LLM. For that NotebookLM which is based on Google Gemini was used. Since the use of OpenAI or Gemini API was costly for the dataset, the concept was evaluated using an open-source tool. This tool allows creating a knowledge base which means by uploading the data, the LLM can reply within the context of that data and make predictions based on historical patterns. Since NotebookLM lacks API integration, testing was done manually through interaction with the user interface by sending the prompt and getting the answer. The LLM achieved 72% which was the lowest among all models. However, other models were tested with 2194 entries, while LLM was tested only with 100 entries because of manual interaction instead of automated one. The accuracy can be more reliable when this approach is developed using paid LLM API and automating the processes. Even though the accuracy was not high, the major advantage of LLM was its ability to predict not only cancellation status but also cancellation reason. For each status prediction, it provided clear reasoning behind that, explaining its choice. This made the model the most interpretable one.

All models were tested using metrics such as accuracy, precision, recall, and Cohen's kappa. The challenge which affected the results was that the dataset was imbalanced. Only 35% of the data referred to cancellations, while the rest of the

policies were active. This issue can be addressed in the future by resampling methods. Such a solution can be used when a larger amount of data will be available, so sampling does not lead to overfitting because of reducing the size of the dataset.

Overall, machine learning and deep learning techniques demonstrated a strong predictive potential which can be used in the insurance sector for managing cancellations. The predicted results can be helpful in decision-making for marketing strategies to retain customers. The methodology provided can be used not only for policy cancellations but also can be applied to subscription services in different industries such as telecommunications, entertainment and other SaaS services.

# Contents

|  |    |
|--|----|
| <b>List of Figures</b>   | 8  |
| <b>List of Tables</b>  | 9  |
| <b>1 Introduction</b>  | 11 |
| <b>2 Background</b>  | 13 |
| 2.1 Churn vs. Cancellation in Insurance . . . . .                    | 13 |
| 2.2 Policy Cancellation in Business Insurance . . . . .              | 14 |
| 2.3 Predictive Modelling for Policy Cancellation . . . . .           | 15 |
| 2.3.1 Machine Learning in Predicting Customer Churn . . . . .        | 15 |
| 2.3.2 Logistic Regression and Decision Trees . . . . .               | 16 |
| 2.3.3 Ensemble Models . . . . .                                      | 17 |
| 2.3.4 Deep Learning Techniques . . . . .                             | 18 |
| <b>3 Materials and methods</b>                                       | 21 |
| 3.1 Data Collection . . . . .  | 21 |
| 3.2 Data Preprocessing . . . . .                                     | 22 |
| 3.3 Feature Selection . . . . .                                      | 23 |
| 3.4 Modelling . . . . .  | 25 |
| <b>4 Results and discussion</b>                                      | 33 |
| 4.1 Exploratory Data Analysis (EDA) . . . . .                        | 33 |
| 4.2 Model Performance Comparison . . . . .                           | 36 |
| 4.2.1 Decision Tree and Logistic Regression Models Results . . . . . | 36 |
| 4.2.2 Ensemble Models Results . . . . .                              | 39 |
| 4.2.3 Artificial Neural Networks Model Results . . . . .             | 40 |
| 4.2.4 Large Language Model Results . . . . .                         | 41 |
| 4.3 Models Comparison . . . . .                                      | 44 |
| 4.4 Challenges and Perspectives . . . . .                            | 46 |
| <b>5 Conclusion</b>  | 47 |



# List of Figures

|      |  |    |
|------|--|----|
| 3.1  | Feature Correlation Matrix. . . . .                    | 25 |
| 3.2  | Decision Tree Classifier workflow in KNIME. . . . .    | 26 |
| 3.3  | Decision Trees View. . . . .                           | 28 |
| 4.1  | Annual vs. Monthly Policies. . . . .                   | 33 |
| 4.2  | Histogram of Days Before Cancellation. . . . .         | 34 |
| 4.3  | Cancelled and Active Policies By Product Type. . . . . | 35 |
| 4.4  | Cancellation Rate. . . . .                             | 35 |
| 4.5  | Decision Tree Prediction Results. . . . .              | 37 |
| 4.6  | Logistic Regression Prediction Results. . . . .        | 38 |
| 4.7  | Random Forest Prediction Results. . . . .              | 39 |
| 4.8  | Gradient Boosting Prediction Results. . . . .          | 40 |
| 4.9  | Artificial Neural Networks Prediction Results. . . . . | 41 |
| 4.10 | LLM Prediction Results. . . . .                        | 42 |
| 4.11 | Prompt template sent to NotebookLM. . . . .            | 43 |
| 4.12 | LLM reply to the user prompt. . . . .                  | 44 |
| 4.13 | ROC Curve Comparison Across Models. . . . .            | 45 |
| 4.14 | Accuracy Comparison Across All Models. . . . .         | 45 |



# List of Tables

|     |                                  |    |
|-----|----------------------------------|----|
| 3.1 | Product type encoding . . . . .  | 23 |
| 3.2 | Business type encoding . . . . . | 23 |
| 4.1 | Cancellation Reasons . . . . .   | 36 |



# Chapter 1

## Introduction

Most people associate insurance with paperwork, bureaucracy, complex legal language and slow processes. However, in the world of rapidly developing technologies, the insurance industry has received a significant boost. Insurtech companies have been especially successful in this, improving the speed and quality of processes through digitalisation. Now, you can buy insurance in a couple of clicks while sitting at home. One of the benefits of such digital provision of services is that valuable customer data is collected every day, which can allow a company to understand better its target audience, and their needs and make products more customer-centric.

It is difficult to get a client in the insurance sector, but even more difficult to retain them. Customer loss affects business revenue, so insurance companies are interested in developing preventive measures against customer churn. In particular, cancellation data can be valuable for supporting decision-making in retention strategies. However, raw data alone cannot tell you anything unless it is analysed using various methods. Depending on the complexity, the data can be analysed using statistics, correlation visualization, machine learning algorithms or more advanced deep learning techniques. These methods can help to find hidden patterns that are not noticeable at first glance.

This thesis aims to thoroughly analyse real-world insurance data from the insurtech company to understand cancellation behaviour. The objective is to compare the effectiveness of different approaches in customer churn prediction based on evaluation metrics. The prediction results can be further used by the company to target marketing strategies towards customers at risk.

To achieve this objective, the methodology is divided into several steps. Firstly, the data is preprocessed to make it clean and ready for further steps. The next is the exploratory data analysis (EDA) is done to get simple correlations in the dataset. This stage provides insights into the distribution of features in data. Following EDA, the predictive models are developed and tested using an analytics platform and libraries. The methods include simple machine learning techniques

such as Decision Tree and Logistic Regression, ensemble models such as Random Forest and Gradient Boosting, and more advanced deep learning methods such as Artificial Neural Networks (ANNs) and Large Language Models (LLMs).

Finally, the results are evaluated based on performance metrics and compared for all models. The findings can be further used for developing proactive retention strategies.

# Chapter 2

## Background

Keeping customers can be challenging in the insurance industry. This requires understanding their behaviour, using data analysis and applying advanced tools like machine learning. Machine learning can catch small signals in client behaviour. It may help insurers take action to retain clients. Customers often reevaluate their needs because of economic changes or their business operations. With predictive techniques, insurers can adapt to new dynamics, respond to competition, and improve customer relationships.

### 2.1 Churn vs. Cancellation in Insurance

Churn and cancellation are related, but they are distinct terms. Churn broadly refers to a customer ending their relationship with a company. Günther et al. [1] define churn in insurance as

When a customer cancels all his/her policies, either to switch insurance provider or because the need of insurance is no longer present, the customer has churned. [1, p. 58]

This statement includes cancellation as one form of churn, but churn can also have other reasons for leaving.

In contrast to some subscription services where customers can easily return after cancellation, insurance cancellations mostly lead to a loss of coverage, and it is difficult to re-engage customers. de la Llave et al. [2] note that churn in insurance often represents a final decision and retention after cancellation is challenging.

In insurance, lapse is also a type of churn which occurs when policies are not renewed. This in turn impacts an insurer's financial stability. Loisel et al. [3] highlight that lapses and cancellations both lead to customer churn, but they may require different retention approaches. Understanding these nuances can help insurers develop targeted strategies for addressing different forms of churn.

## 2.2 Policy Cancellation in Business Insurance

Insurance plays a fundamental role in risk management within the modern economy. Business insurance, especially in the context of liability and indemnity policies, helps companies operate with financial security and operational confidence [4]. Liability and indemnity policies cover a variety of financial risks and potential legal liabilities. This allows companies to protect themselves against losses from litigation, property damage and other unexpected events that may impact their operations [5]. As a result, business insurance policies are often considered a prerequisite for stable economic activity, since they offer companies the safety net they need to grow in competitive and uncertain markets [6].

Despite clear importance, the insurance industry faces challenges in customer retention. This is mostly true for non-life insurance sectors, where the rate of switching or cancellation is high. This customer retention challenge is associated with customer sensitivity to price, coverage options and the availability of competitive options from alternative providers [7, 8]. Small improvements in retention rates can result in considerable revenue gains for insurers [9]. This highlights the critical importance of effective customer relationship management [1].

Customer churn is a common issue in the insurance sector. Getting new customers rather than keeping the existing ones is more difficult and costly [10, 2]. Depending on the industry, the acquisition cost of a new client can be from 5 to 25 times more expensive than the retention cost [11]. Research shows that small and medium-sized businesses tend to cancel policies if they feel their insurance needs are no longer adequately met or if they face financial struggles that limit their ability to afford coverage [12].

Several factors affect policy cancellations in business insurance. Business closure is one of the most common reasons for cancellations because companies in liquidation or restructuring may no longer require liability or indemnity coverage [13]. Premium prices also play a significant role in cancellations because companies are more likely to cancel policies when they find premiums excessive. It is especially common in sectors with low-profit margins. This price sensitivity also correlated with competitive market conditions, since businesses are likely to look for more cost-effective insurance options from other providers [2]. Switching providers to secure better terms is another common reason for policy cancellations. This decision may happen because of gaps in coverage or restrictive terms that do not fully address the specific business risks. When policy limitations become obvious, for example, due to recent claims, industry risks, or regulatory compliance requirements, businesses are more likely to change a provider if a new one offers more flexible or comprehensive solutions [14]. Economic and regulatory factors also may have a significant influence on policy retention. Economic downturns, caused by global crises, force companies to cut expenses, including insurance costs. In such situations, some companies may see insurance as a non-essential expense that can

be minimised or removed at all to save financial resources [13].

A thorough understanding of these factors is essential for insurers who want to develop targeted retention strategies. The study [8] highlights the benefits of customer-centric approaches that emphasize direct engagement and tailored solutions. By focusing on policy customization, proactive claims management, and flexible premium options, insurers can better align their offerings with the needs of business customers and thereby improve retention in a competitive market.

## 2.3 Predictive Modelling for Policy Cancellation

Predictive modelling is an important tool for insurance companies to effectively manage customer churn. Since business profitability is highly correlated with customer retention, insurers are widely using machine learning techniques to analyse customer behaviour, predict churn risk, and optimise policy retention strategies. With advanced algorithms, companies can better understand cancellation behaviour in high-churn sectors such as insurance [15]. Predictive modelling in insurance involves different machine learning approaches. These can be traditional statistical models or more complex ensemble and deep learning models. Each offers unique advantages.

### 2.3.1 Machine Learning in Predicting Customer Churn

In the insurance industry, predictive models analyse historical and behavioural data to predict a policy cancellation probability. This approach allows insurers to develop data-driven retention strategies. In its turn, this can lead to customer satisfaction improvement and a reduction of revenue losses from cancellations. Research shows that accurate churn prediction can be achieved by combining traditional and advanced ML methods [15]. The approach choice depends on the types of data and forecasting objectives. The application of machine learning on churn prediction can start with creating simple models such as logistic regression and decision trees. These models provide a foundation for understanding basic churn patterns. It is possible to see how specific features, such as premium levels, claim frequency, and payment history, correlate with the likelihood of churn [16]. With the advancement of predictive models, ensemble techniques like Random Forest and Gradient Boosting have become central to handling more complex datasets and improving prediction accuracy. Ensemble models combine the predictions of multiple simpler models. This allows to capture a broader array of patterns that can be missed by single models [17].

Neural networks, a common tool of deep learning, have also become increasingly popular for churn prediction. Neural networks can be useful if companies have high-dimensional and nonlinear data [18]. For example, when they want to analyse

complex customer behaviours over time. In contrast to traditional models, deep learning techniques automatically learn patterns in data. This makes them ideal for modelling churn in industries with large and diverse customer datasets [18].

An effective churn prediction system often uses a mix of these methods. It is important to balance the interpretability of traditional models with the enhanced accuracy of ensemble and deep learning approaches. For instance, a study by Liu et al. [19] demonstrated how backpropagation neural networks could identify high-risk customers in life insurance renewals. Moreover, combining ensemble methods followed by a neural network layer demonstrated improved accuracy without losing interpretability. These techniques allow them to focus resources on high-risk clients and can lead to an overall reduction in cancellation rates [19, 8].

### 2.3.2 Logistic Regression and Decision Trees

Two basic machine learning methods which can be used to predict customer churn are decision tree and logistic regression. They are considered highly interpretable models because these models provide straightforward insights into the factors influencing customer behaviour. These techniques can be valuable for understanding why customers might cancel their policies. Logistic regression is suitable for binary classification tasks, such as finding if it is more likely that a customer will leave or stay. It works by calculating the probability of a specific outcome based on different predictor variables. For example, a customer's claims history, regularity of payment, and premium changes can be taken into account [20]. Logistic regression is simple and transparent, so it is useful for companies that need interpretable results for compliance reasons. Decision trees, on the other hand, are more flexible because they involve recursively dividing data into branches based on feature values. As a result, insurers can get classifications which help to categorise customers based on churn risk. For instance, a decision tree may classify customers who made multiple claims in a year and got their profiles as high-risk, which led to the increased amount of money they paid. Such classification can assist insurers in prioritising accounts for retention strategies [21]. The tree structure is also easy to interpret because each split clearly outlines how different variables impact the decision path. Decision trees can capture nonlinear relationships, which can be common in customer behaviour patterns. So, they can be effective in segmenting data with categorical variables like demographic information [1].

Research demonstrates the practical applications of these models in the insurance industry. For example, Calderon and Garcia-Bedoya [16] used logistic regression and decision trees to develop a retention model for policy cancellations. They found that customers who had recently filed claims or faced premium increases were more likely to cancel. This result highlights that these techniques have a predictive power.

Even though these models are considered effective, they have some limitations.



Logistic regression assumes that there is a linear relationship between predictors and the outcome. Such an assumption can restrict its accuracy if there is complex, nonlinear customer behaviour. Decision trees are flexible, but they tend to overfitting in case they are not properly tuned. This means they may reduce their performance on unseen data [15]. To address these limitations, insurers often combine these two approaches with ensemble techniques. This way, they can improve accuracy by using the strengths of multiple models. In general, decision trees and logistic regression make a significant contribution to predictive modelling in insurance. They are simple, interpretable, and effective in identifying key churn indicators.

### 2.3.3 Ensemble Models

The result of a combination of multiple machine learning methods is ensemble models. They stand out with their robustness and accuracy in comparison to what can be achieved using individual models. The two well-known ensemble techniques are Random Forest and Gradient Boosting. Each of them provides unique strengths and can work with complex data and minimise the limitations of simpler models.

Random Forest is an ensemble of decision trees. Each tree is trained on a random subset of data and features. The final prediction is an average of all the trees or a majority vote in classification tasks. This way of prediction reduces the risk of overfitting [17]. In the insurance industry, Random Forest is considered effective in customer segmentation based on churn risk. It can identify predictors, such as payment patterns, claim history, and demographic details. He et al. [15] demonstrated that Random Forest can be used in analysing insurance renewals. In this paper, the model successfully identified customers who were likely to cancel their policy because they were not satisfied with coverage or rising premiums.

Gradient Boosting is another popular ensemble approach. It builds sequential decision trees, where each new tree corrects the errors made by the previous ones. This iterative process allows Gradient Boosting to focus on difficult cases and achieve high prediction accuracy. AlShourbaji et al. [22] introduced an enhanced Gradient Boosting model called CP-EGBM, which achieved up to 97.79% accuracy on certain datasets for predicting customer churn in telecommunications. This result illustrates its efficiency in handling imbalanced datasets, which can be common in insurance applications.

Ensemble methods have some limitations. Both Random Forest and Gradient Boosting are computationally intensive. Especially, it can be noticed with large datasets, since they require the training of multiple trees. Also, Gradient Boosting is sensitive to parameter tuning. If parameters are poorly chosen, it can lead to overfitting or underfitting. It will negatively affect model performance on new data [16]. Additionally, ensemble models tend to be less interpretable than single models. It can be a drawback in insurance, where decision transparency can be important

for compliance. To solve this, many insurance companies use feature importance scores to interpret Gradient Boosting models. This way they can balance accuracy with interpretability.

Despite these challenges, ensemble models are still powerful in customer behaviour analysis. In practice, insurers can combine Random Forest and Gradient Boosting with simpler models and achieve both accuracy and interpretability.

### 2.3.4 Deep Learning Techniques

Deep learning can uncover intricate patterns within customer behaviour that may influence policy cancellation. In contrast to traditional models, deep learning techniques, such as neural networks, do not require much manual feature engineering. Instead, they can automatically learn relevant features from raw data. This ability makes them useful for unstructured data. Data points can be customer interactions, textual feedback, and behavioural patterns over time [19].

Neural networks are a foundational deep learning technique, and they are widely applied in churn prediction. A key advantage of neural networks is their ability to model nonlinear relationships. For example, changes in behaviour can be correlated with churn, but not linearly. Instead, a mix of multiple small changes in behaviour can collectively indicate a higher churn risk. Neural networks can capture these subtle relationships more effectively than traditional methods [23]. It makes NNs suitable for complex datasets. Khattak et al. [24] showed the effectiveness of an advanced deep learning model using a combination of Bidirectional Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) for predicting customer churn in telecommunications. Their model achieved 81% accuracy on test datasets. This approach can win in comparison to traditional models because of its sophistication in capturing both the sequence and context of customer behaviour.

Moving to the limitations of deep learning models, they are computationally intensive. This means they often require large datasets and substantial processing power. What is more, deep learning models are generally considered *black boxes* because their internal decision processes are less interpretable in comparison to simpler models [25]. To improve interpretability, insurers use additional explainability methods by applying mathematical theory to estimate the contribution of attributes.

Large language models (LLMs), such as OpenAI's GPT and Google's BERT, are state-of-the-art innovations in deep learning techniques for natural language processing. They show strong performance in tasks where it is necessary to process data in unstructured or non-numeric form, like text analysis. Unlike other predictive models in machine learning that need to be trained at the start of modelling, LLMs are already pre-trained on huge datasets. This approach allows these models to be tuned to very specific tasks in an effective way. Thus, this makes them powerful tools for extracting information and predicting results. At the same time, understanding and generating human-natural texts create a lot of convenience when analysing churn and customer retention in real-life situations. This creates many unique opportunities when compared with other models. For example, in cases where it is necessary to analyse unstructured data such as call transcripts, email content, and feedback forms [26]. These sources often have very valuable information to catch patterns of customer dissatisfaction. Such information is often difficult to "mathematize", which is not applicable in cases such as logistic regression or decision trees.

LLMs use deep learning architectures, specifically transformers, to model the relationships between words and text elements. In doing so, they build connections. This enables sentiment analysis of texts, which can then be used to classify text into churn categories. For example, complaints about unresolved claims and applications, as well as subscription price increases, can be labelled as high-risk predictors of churn [27].

Despite the listed benefits of LLMs specifically in the policy cancellations domain, they, like other models, have a number of drawbacks. Due to the complex processes, they require special equipment or cloud access to this equipment. Most often, this access in commercially known models requires funds and subscriptions to services. The above-mentioned drawback in the Deep Learning section with the transparency of the process is especially relevant here since the insurance sector is heavily regulated. And this leads to mistrust of the results of models in these types of businesses [28]. Although LLMs are a relatively new phenomenon in the insurance industry, their adoption is growing [29]. Their ability to detect subtle patterns makes them an important component in the next generation of predictive models.



# Chapter 3

## Materials and methods

This study involved a systematic approach to develop and evaluate a predictive model for customer churn. The methodology involved data preprocessing, model selection, hyperparameter tuning, holdout validation, and final evaluation. KNIME analytics platform was used to create workflows per simple and ensemble techniques. Additionally, an ANN binary classifier was implemented using a software library, Tensorflow, to explore deep learning techniques for enhanced predictive accuracy. Finally, the LLM model was applied using a note-taking online tool NotebookLM, based on Google Gemini. This approach is designed to be robust and adaptable which can be applied beyond the current context to analyse subscription cancellation across various industries, such as telecommunications, media streaming, and SaaS (Software as a Service) products. Each stage of the methodology is detailed below.

### 3.1 Data Collection

To understand the trends behind the cancellation, various data points were merged and involved in the analysis. Data was collected from three sources: insurance provider dashboard, payment platform Stripe, and customer service platform Intercom. The insurance provider dashboard allows exporting the policy's status, entity type, payables and fees, starting and cancellation dates, and product name. Payment history including payment failures and successes was extracted from the payment platform Stripe. All customer interaction through messages and emails was available on the customer service platform Intercom. In Intercom, it was not possible to automatically download the user notes, so the data was labelled manually for the users whose policies were cancelled manually. The rest of the policies were cancelled automatically due to non-payment, and that data was available through Stripe.

Manual data labelling was done according to the following steps:

1. Organising the list of users whose policies were cancelled manually based on the insurance provider and payment platforms
2. Finding a user using an email filter on Intercom
3. Opening the conversation with the user
4. Checking the notes from the customer success team
5. Checking the answer of the user to the question “Can you tell us the reason you’d like to cancel your policy?” which is standard protocol in the cancellation process for each customer
6. Categorise the answer into one of the ten categories: “No longer self-employed, working in the relevant field, or pursuing independent work”, “Business no longer operating”, “Only required insurance for a short or specific period”, “Found a new insurance provider”, “Covered under association or employer-provided insurance”, “Wants to repurchase with different terms (e.g., lower limits)”, “Wants to pause for an unknown time (e.g. renovation, moving to a new place)”, “Maternity leave/Paternity leave”, “Travelling overseas”, and “No details”
7. Leave a free-writing comment in the Comments column

Historical manual cancellation data included interaction with 1593 customers. The policies of 1116 customers were automatically cancelled due to non-payment, so a total of 2709 cancellations were labelled. The breakdown is shown in Table 3.1.

## 3.2 Data Preprocessing

To get high quality data and prepare the dataset for training the model, preprocessing was performed. First, data was cleaned in Excel, removing all personal data, so now there is no way to identify the customer. Totally irrelevant columns were removed. Duplicated rows were cleaned. As a result, there were 7435 rows in the dataset. The further preprocessing phase included the following steps:

Missing values were handled by removing these rows. It was rare to have missing data, mostly it was encountered in the business revenue, business age and business size columns. After removing rows with missing data, the dataset contained **7313** rows.

Categorical variables were pre-converted to numeric format during data processing in Excel. Categorical variables such as Product type and Business type were replaced by the following numbers in the Tables 3.1 and 3.2:

|      |   |
|------|---|
| 0    | Healthcare Professionals Civil Liability (PI & PL)  |
| 1    | Public & Products Liability   |
| 2    | Cyber & Privacy Liability   |
| 3    | Management Liability  |
| 4    | Sole Traders & Partnerships Management Liability  |
| 5    | Professional Indemnity  |
| 6    | Association Liability   |
| 51   | Professional Indemnity, Public & Products Liability   |
| 702  | Healthcare Professionals Civil Liability (PI & PL), Cyber & Privacy Liability                       |
| 12   | Public & Products Liability, Cyber & Privacy Liability  |
| 703  | Healthcare Professionals Civil Liability (PI & PL), Management Liability                            |
| 7032 | Healthcare Professionals Civil Liability (PI & PL), Management Liability, Cyber & Privacy Liability |
| 32   | Management Liability, Cyber & Privacy Liability   |
| 126  | Public & Products Liability, Cyber & Privacy Liability, Association Liability                       |
| 10   | Public & Products Liability, Healthcare Professionals Civil Liability (PI & PL)                     |

Table 3.1: Product type encoding

|   |                 |
|---|-----------------|
| 0 | Association     |
| 1 | Co-operative    |
| 2 | Employee        |
| 3 | Partnership     |
| 4 | Private company |
| 5 | Sole Trader     |

Table 3.2: Business type encoding

This data preprocessing framework is flexible and can be used for different subscription-based datasets.

### 3.3 Feature Selection

The dataset contains different columns. Some can contribute to the prediction, while the rest do not add any predictive value. To filter out what features can be useful, first, the correlation matrix was built to catch any linear correlation in the

dataset with policy cancellation status. Setting 1 as Cancelled and 0 as Active, findings from the correlation matrix Fig. 3.1 are:

- Positive correlation (0.3284) between Cancellation status and binaryPolicyType (1 - Monthly, 0 - Annual), which means Monthly policies tend to be cancelled.
- Positive correlation (0.1806) between Cancellation status and outstandingLoanAmount, which means the more amount of money to be paid, the more likely to cancel. For example, if customers paid 2 monthly installments and have 10 installments left for one-year cover, then they are more likely to cancel than those who already paid 10 installments and have only 2 installments left.
- Positive correlation (0.1032) between Cancellation status and monthlyFees, which means higher monthly fees, more likely to cancel.
- Positive correlation (0.0546) between Cancellation status and numBusinessType. In Table 3.2, the business type is encoded using numbers, according to it, cancellation increases going to the bottom of the table from 0 (less likely) to 5 (more likely).
- Positive correlation (0.0445) between Cancellation status and paymentFailures, which means more payment failures, more likely to cancel.
- Negative correlation (-0.1895) between Cancellation status and amountReceived. It is the opposite of outstandingLoanAmount, which means, the more money they already paid, the less likely will cancel.
- Negative correlation (-0.1335) between Calculation status and logBusinessRevenue, which means the higher revenue of a business, a policy is likely to be cancelled.
- Negative correlation (-0.12) between Cancellation status and underwriterFee, which means the higher the underwriter fee they paid, they are less likely to cancel.
- Negative correlation (-0.0493) between Cancellation status and businessAge, which means older a business, it is less likely to cancel.
- Negative correlation (-0.0289) between Cancellation status and quotePrice, which means the higher the price of the policy, the less likely they cancel. High-earning businesses can have a higher quote price, and these companies may be less likely to cancel their policies. However, this correlation value is very small, close to neutral.



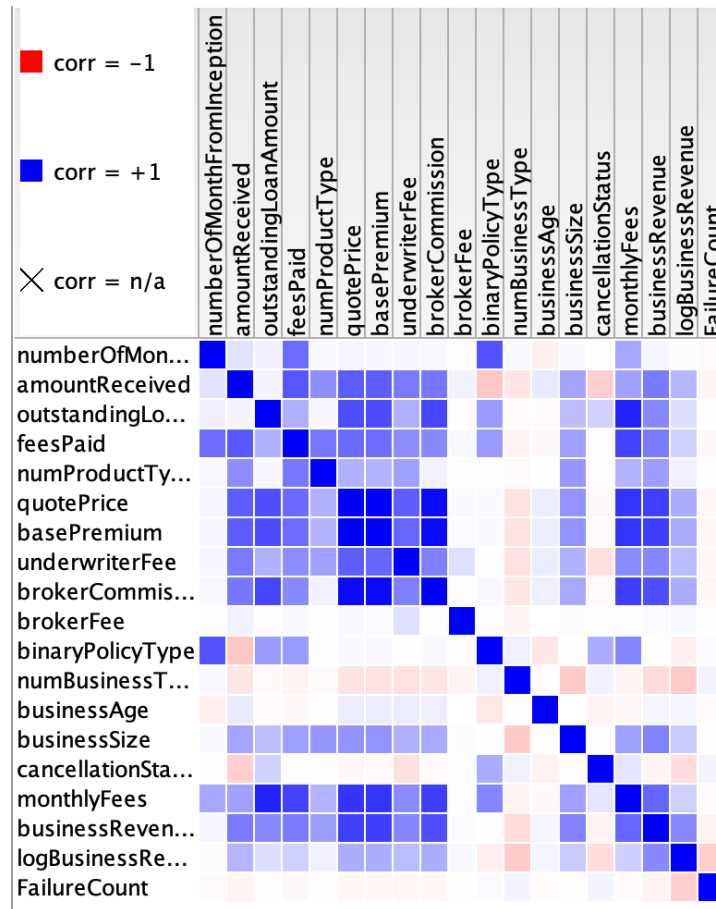


Figure 3.1: Feature Correlation Matrix.

Based on this correlation matrix, the following features could be considered: *binaryPolicyType*, *outstandingLoanAmount*, *monthlyFees*, *numBusinessType*, *paymentFailures*, *amountReceived*, *logBusinessRevenue*, *underwriterFee*, *businessAge*, *quotePrice*. Even though linear correlation values can be slight, they may have a nonlinear correlation, and features in combination with each other can have significant predictive power. The final list of features was obtained by using the Column Filter node and ROC curve checking the effect of the feature on the predicting accuracy.

### 3.4 Modelling

After cleaning and encoding data, the next step was to create workflows in KNIME. KNIME, the Konstanz Information Miner, is an open-source analytics platform. It also allows reporting data, visualization and integration with other platforms. KNIME is easy to use and effective because there is no need to code and implement

the algorithms. With KNIME, it is possible to access popular machine learning techniques, test, validate and deploy a model, automate tasks and get insights from data. The KNIME workflow's building blocks are nodes. The nodes represent individual tasks. Each node consists of an input and output port and also settings. For example, the Column Filter node can be set to filter out unnecessary columns from the data to be used. The example of the workflow created for the Decision Tree Classifier is shown in Fig. 3.2. Each step of the workflow will be further explained in this section.

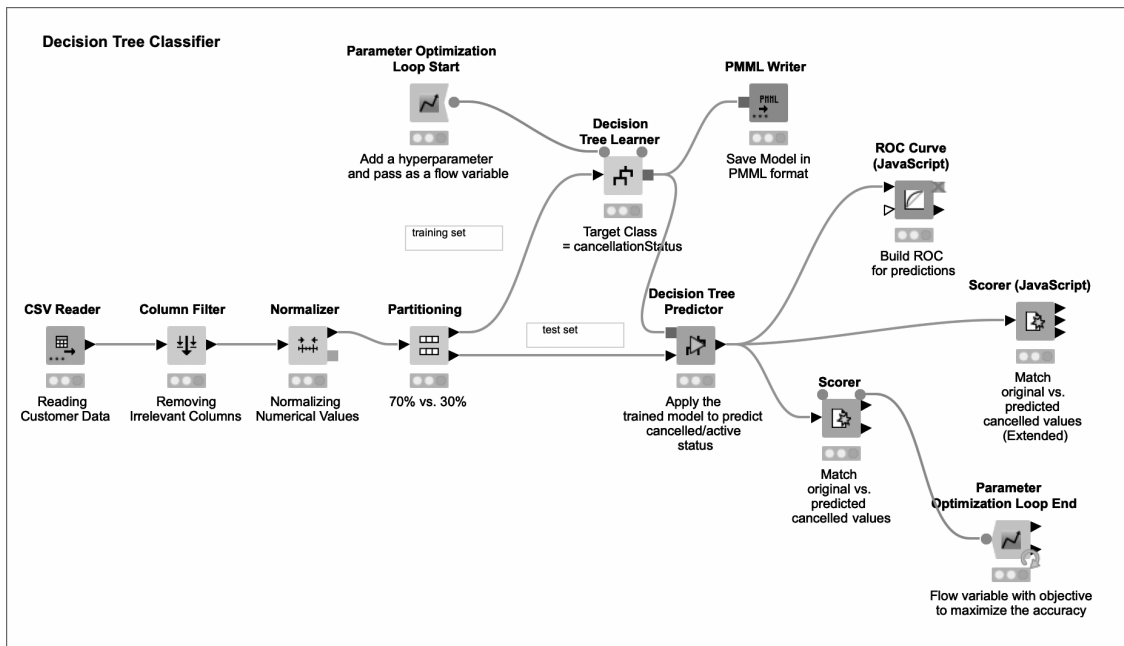


Figure 3.2: Decision Tree Classifier workflow in KNIME.

In this thesis, the four sets of workflows were created to build models. They included Decision Trees, Random Forest, Logistic Regression and Gradient Boosting classifiers.

## Data Preprocessing in KNIME

Each workflow started with data preprocessing. The CSV Reader node was used to open the data. Since the data was already cleaned in Excel and encoded, there was no need to manage missing values and use encoding. Using the Column Filter node, some columns were removed from the analysis because they were not relevant to cancellation or they were post-cancellation information which could leak future information. Since some machine learning algorithms (e.g. logistic regression) are sensitive to feature scales, the Normalizer node was used to standardize the numeric features. This ensured a consistent scale having continuous values between 0 and 1

for all predictors. The Min-max normalization method was used on all numerical features.

### Data Partitioning

In Fig. 3.2, the next step after data preprocessing is partitioning. Using the Partitioning node, the data was divided into a 70% training set and a 30% testing set. Since the data was imbalanced with only 35% of cancellations, the stratified sampling method was used with fixed random seed for all models. The fixed random seed allowed to split the data in the same way for all 4 models, so they can be fairly compared with each other. The stratified sampling is well-known for applying to imbalanced data ensuring that all instances of cancellation status are represented in both training and testing sets. The Partitioning node has 2 output ports, representing two splits of data.

### Decision Tree Classification

After splitting the data, further steps were choosing the learner and predictor. For the decision tree classification, there were the Decision Tree Learner and Decision Tree Predictor nodes. The Decision Tree Learner node allows setting the quality measure which can be Gini index or Gain ratio. In this analysis, the Gini index was chosen because it resulted in higher accuracy. The Gini index is a measure of impurity which means it helps to determine how mixed or pure the classes (Active or Cancelled) are in each group created by the decision tree. When the tree splits the data, it tries to make the resulting groups as "pure" as possible, so it tries to reach that most of the data points in each group belong to one class, either Active or Cancelled. A lower Gini index indicates a better, more "pure" split. For example, in Fig. 3.3, at the root of the tree, most customers are classified as Active, with 3320 out of 5119 having this status. The tree starts by classifying the majority of the data as Active, but then it looks for ways to further split the data into more specific groups to improve accuracy. The first significant split happens based on the *outstandingLoanAmount* feature, with a threshold of 0.0097, which is the normalized value of an outstanding loan amount. Customers with an outstanding loan amount less than or equal to 0.0097 are grouped together, and in this group, 2445 out of 3123 customers are Active. In the second group, where the *outstandingLoanAmount* is greater than 0.0097, 1121 out of 1996 customers are still Active. The tree continues to split these groups further based on other features, such as the underwriter fee and number of months from inception in order to refine the prediction. Since the Gini index focuses on minimizing the impurity within each group, it leads to splits that make the prediction of Active or Cancelled more accurate and clear. This is why the Gini index provided better results in this case compared to the Gain ratio which normalizes splits and is useful for features

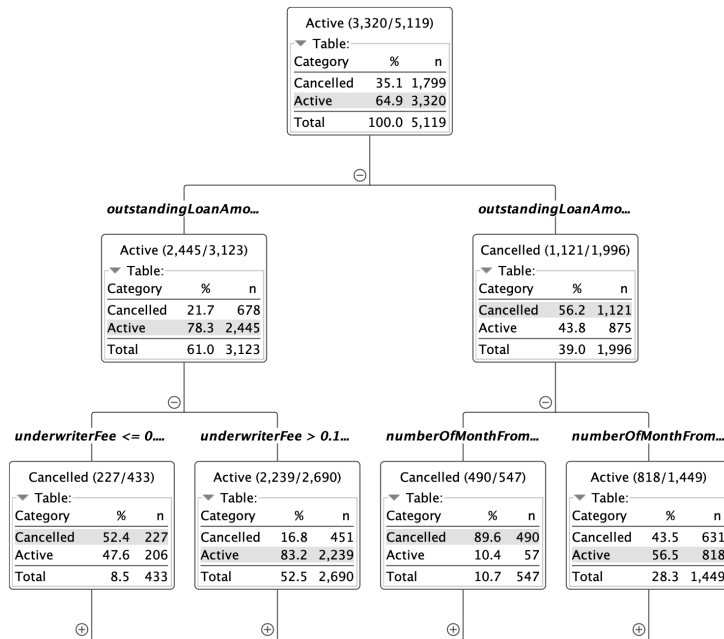


Figure 3.3: Decision Trees View.

with many unique values, not evenly distributed. Also, the MDL pruning method was applied. By default, there is no pruning. Setting the pruning helps to reduce the size of the trees by removing branches which add a little value to prediction. This way tree remains simpler, preventing overfitting, so the tree is more likely to perform well on unseen data. Next, the Decision Tree Predictor node was applied to unseen testing data to predict the cancellation status. A standard holdout validation was performed by splitting the data into training and testing sets.

### Logistic Regression, Random Forest, and Gradient Boosting

Similar steps were done for other algorithms but using Logistic Regression Learner, Logistic Regression Predictor, Random Forest Learner, Random Forest Predictor, Gradient Boosting Learner, Gradient Boosting Predictor nodes correspondingly. For Random Forest, the algorithm creates an ensemble of decision trees to predict cancellation status. Each tree is trained on a random subset of the data, and at each split, it considers only a random set of features. The final prediction is made by voting among all the trees. This randomness helps to avoid overfitting and noise in the data. Gradient Boosting also uses decision trees but builds them one after another. Each new tree focuses on correcting the errors made by the earlier trees. The final prediction is a weighted sum of all tree predictions. What about Logistic Regression, it estimates the probability of a target class by applying a logistic function to a linear combination of data features.

### Artificial Neural Networks

After creating and testing these machine learning models, the next step was to test the deep learning models on the dataset. For that, Jupyter Notebook was used as the environment to write, run the code, and visualize results. It was easy to import the libraries needed such as Tensorflow for building a model, Numpy for handling numerical operations, Pandas for normalization, Matplotlib and Seaborn for visualization of results. With Tensorflow, the Artificial Neural Networks model was created. The data preprocessing steps for the ANN were similar to those used above. First, numerical features were normalized. Using stratified sampling, the dataset was divided into 70% for training and 30% for testing. The ANN architecture included three layers. The first layer was with 64 neurons, the second was with 32 neurons, and an output layer with one neuron for predicting cancellation. Neurons are like small decision-making units which process information and pass it through the network. The hidden layers (64 and 32 neurons) help the model learn complex patterns in the data.

For the hidden layers, the activation function called ReLU (Rectified Linear Unit) was applied, so this function allows a model to learn nonlinear patterns. To optimize the model's performance and prevent overfitting, two callbacks, which are early stopping and learning rate reduction, were implemented. Early stopping callback helps by stopping training if the model's performance does not improve after several tries. The learning rate reduction callback adjusts the learning rate during training. This callback lowers the learning rate when the validation performance does not improve. As a result, these callback helps in achieving better performance. Further optimization of the model was achieved using the Adam optimizer for adjusting weights and binary cross-entropy loss to measure prediction accuracy for binary outcomes.

### Large Language Models

The last method used is Large Language Models, a subset of deep learning. LLMs can process complex relationships and generate human-like interpretations of data. LLMs are trained on large datasets and are advanced in identifying patterns, reasoning, and making predictions based on provided inputs. To test the application of LLMs in the context of cancellation analysis, a note-taking online tool NotebookLM which is based on Google Gemini was used in this thesis. So, what stands out for Google's free access tool NotebookLM is that it allows the creation of a knowledge base. The knowledge base works as a database which can be accessed by asking questions, and AI replies based on the data uploaded. This way, the model "thinks" within the context of the training data. Such a powerful feature makes NotebookLM well-suited for testing LLM. While conventional machine learning models often output probabilities without context, NotebookLM provides textual reasoning for prediction relying on data patterns. This feature not only wins in

the interpretability of predictions but also enhances trust in the outcomes because each prediction goes with clear explanations and identified potential cancellation reasons. This tool was chosen as a cost-effective alternative to paid APIs such as OpenAI’s GPT or Gemini. While these APIs are scalable and robust, their usage requires substantial costs when processing a large dataset. To automate the creation of prompts for NotebookLM, a Python script was developed. The script generated prompts based on a predefined template and saved them for further use. Each prompt asked NotebookLM to analyse specific policy details, predict the cancellation status, and, when relevant, determine the most likely cancellation reason. The prompt template is shown in Fig. 4.11. Predictions were obtained manually using the NotebookLM user interface because the platform currently lacks API integration. Since the manual sending of a prompt and getting the results is time-consuming, only 100 entries were checked. The results were systematically verified against actual policy statuses and recorded in an Excel sheet with the prompt, predicted status, actual status, predicted cancellation reason, actual cancellation reason, and NotebookLM explanation. Finally, a confusion matrix was created using Matplotlib and Seaborn libraries to evaluate the model’s performance.

### Hyperparameter Tuning and Evaluation

To tune hyperparameters and experiment with different values for number of records per node, the values were configured as a flow variable. For that, the Parameter Optimization Loop nodes were used where the start value, stop value, and step size for the hyperparameters were set. After identifying the best value for the hyperparameter, the optimization loop was disabled, and the selected value was manually set for the model to avoid overfitting to the validation set during hyperparameter tuning. For Decision Tree, the best value for a minimum number of records per node was found and set. Gradient Boosting and Random Forest models were tuned using hyperparameters as the maximum tree depth and minimum node size.

And finally, the Scorer(Javascript) node was applied to evaluate a model. This node created a confusion matrix with True Cancelled, False Cancelled, True Active, and False Active policies, and provided the overall accuracy of the model. The evaluation metrics for models were accuracy, precision, recall and Cohen’s kappa. Accuracy is calculated by dividing the correct predictions by the total number of predictions, using the formula 3.1:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

where

- $TP$  = True Positive, which is True Cancelled
- $TN$  = True Negative, which is True Active

- $FP$  = False Positive, which is False Active
- $FN$  = False Negative, which is False Active

Precision is calculated as the proportion of correctly predicted positive cases out of all cases predicted as positive, following the formula 3.2:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.2)$$

The next metric is recall, which measures the proportion of actual positive cases that are correctly found by the model, calculated following the formula 3.3:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.3)$$

And lastly, Cohen’s kappa compares the agreement between the predictions of the model and the true labels with the agreement that would be expected by random chance. It is calculated by equations 3.4 and 3.5:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (3.4)$$

Where  $P_e$  is calculated as:

$$P_e = \frac{(TP + FP)(TP + FN) + (TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2} \quad (3.5)$$

These evaluation metrics were retrieved for each model after testing on unseen data. Their performance was compared to identify the best approach for predicting cancellation status.





# Chapter 4

## Results and discussion

### 4.1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis is used to understand basic patterns in customer cancellations. The findings from EDA will support the development of machine learning models and interpretation of the results. In Fig. 4.1, 80.2% of the customers, who cancelled their policies, were opting for monthly payments. It is convenient for customers who can not afford to pay the annual insurance upfront, so they pay in monthly installments. With a monthly payment plan, customers have more frequent opportunities to cancel, which results in a higher customer churn rate.

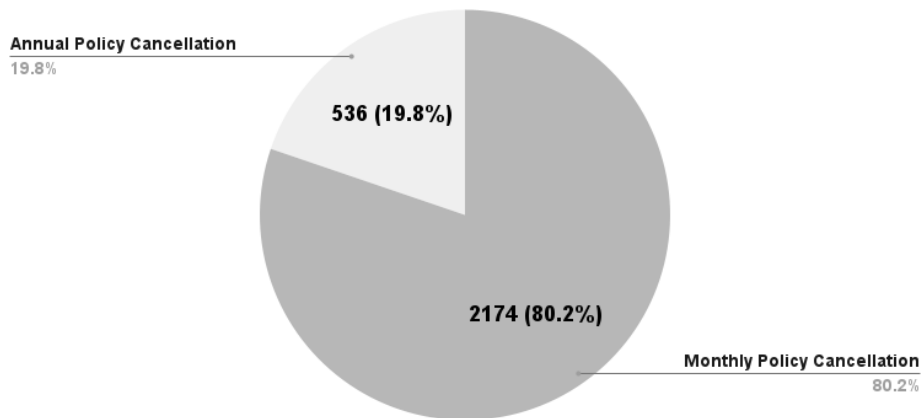


Figure 4.1: Annual vs. Monthly Policies.

From Fig. 4.2, a clear trend can be seen following the red line. Every 30 days, the cancellation number rockets. Most cancellation happens in the first seven days after purchase, then it rises after 30 days when there is a date to pay the next installment. Upcoming charges may prompt clients to reconsider their insurance,

but also these numbers are related to payment failures after auto-renewal attempts. Payment failures may happen because of insufficient funds, card issues such as lost, expired or stolen cards, transaction restrictions, or due to technical issues.

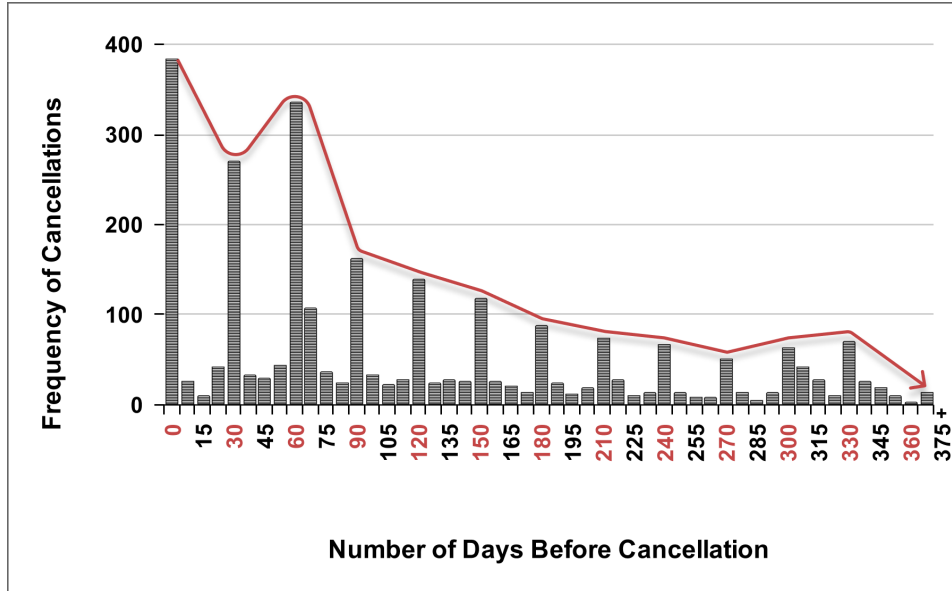


Figure 4.2: Histogram of Days Before Cancellation.

Fig. 4.3 shows the cancellation status of customers based on the product they bought. The most popular insurance product is Healthcare Professionals Civil Liability (PI & PL). The target clients for this product category are disability support workers, nurses, fitness instructors, yoga teachers and other health professionals. Some of these professionals, such as support workers, may only work in these roles temporarily while they seek other career opportunities. So, they tend to cancel their policies more often.

The next Fig. 4.4 demonstrates that even though Healthcare professionals are the vast majority of customers, the cancellation rate is higher for clients who buy Public and Product Liability insurance. The top reasons for the cancellation of this product are payment failures, switching the field, and the business is no longer operating. Also, in Fig. 4.3, there are only two people who bought the Sole Traders and Partnerships Management Liability product, and both of them cancelled. This explains why in Fig. 4.4 the cancellation rate of this product is 100%.

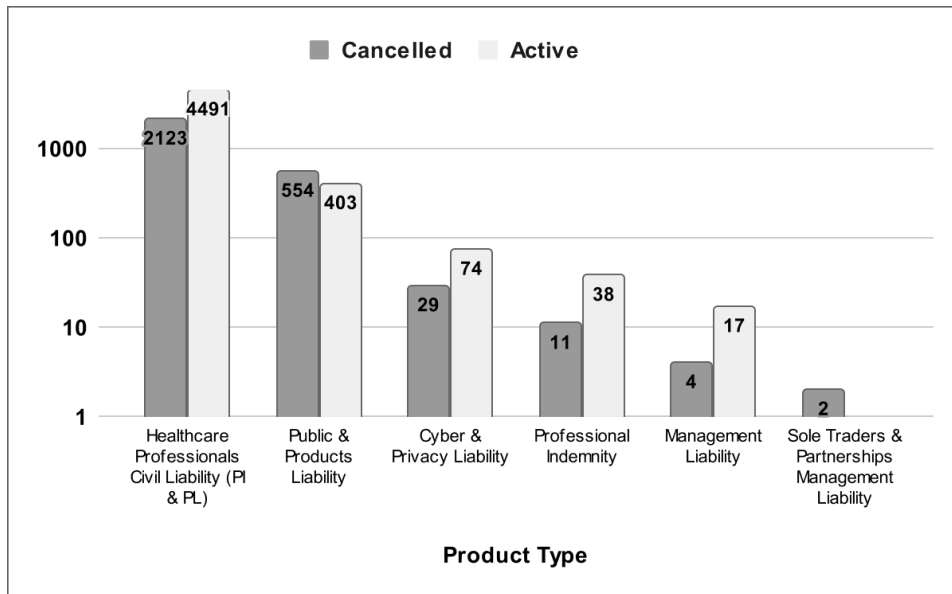


Figure 4.3: Cancelled and Active Policies By Product Type.

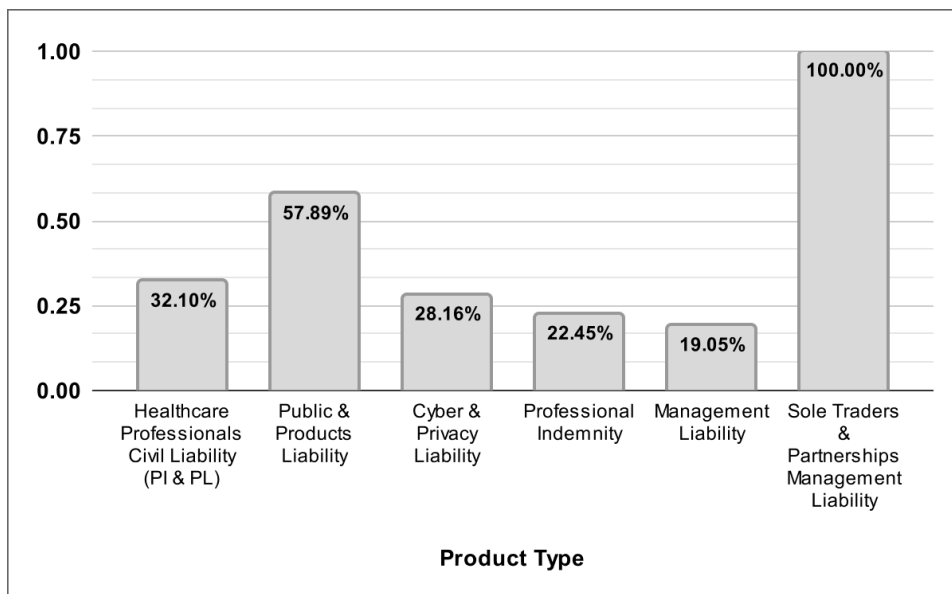


Figure 4.4: Cancellation Rate.

In Table 4.1, a breakdown of all cancellation reasons is provided. This table shows that even though almost half of the cancellations were after payment failures, the other half of cancellations were requested manually. The majority of the requested cancellations refer to the change of employment, which is related to 510 customers.

|    |  |             |
|----|--|-------------|
| 1  | No longer self-employed, working in the relevant field, or pursuing independent work | 510         |
| 2  | Business no longer operating   | 217         |
| 3  | Only required insurance for a short or specific period                               | 24          |
| 4  | Found a new insurance provider   | 133         |
| 5  | Covered under association or employer-provided insurance                             | 62          |
| 6  | Wants to repurchase with different terms (e.g., lower limits)                        | 97          |
| 7  | Wants to pause for unknown time (e.g. renovation, moving to a new place)             | 33          |
| 8  | Maternity leave/Paternity leave  | 17          |
| 9  | Travelling overseas  | 31          |
| 10 | No details   | 469         |
| 11 | Payment failure  | 1116        |
| #  | <b>Total</b>   | <b>2709</b> |

Table 4.1: Cancellation Reasons

The next major one is the No details label, which means that the customers did not want to explain why they wanted to cancel. As a result, Payment failure and No details in combination show that around 58% of users churned, i.e. stopped relations with the company, without providing the exact reason for the cancellation of their policies. Some considerable part of the requested cancellations happened because of external reasons in the life of customers. For example, maternity/paternity leave, travelling overseas, pausing for an unknown time which can happen for family or health reasons, renovations, or moving to another place. Reasons like change of employment, or closure of business can be highly correlated with payment failures, making a payment failure a symptom for the base reason of financial problems of the business. The reason for cancellation which states that the customer wants to repurchase with other limits is different from other cancellations since it does not show the intention of the customer to end the relationship with the company. Taking into account all the correlated features, machine learning algorithms can help here to further explore it.

## 4.2 Model Performance Comparison

### 4.2.1 Decision Tree and Logistic Regression Models Results

The Decision Tree Classifier achieved an overall accuracy of 84.59%, showing a strong alignment between predictions and actual labels. However, 15.41% of the predictions were incorrect, and it highlights areas where the performance could

be improved. The recall for the Active class in Fig. 4.5 was particularly strong, being 94.87%. This metric shows that the model effectively identified most of the active policies. This recall value means that active policies are rarely misclassified as cancelled. On the other hand, the recall for the Cancelled class was lower at 65.63%. It indicates that the model struggled to capture all actual cancellations. It can lead to missing some customers who are at risk of policy cancellation. In terms of precision, the model performed well for both classes. For the Cancelled class, the precision was 87.39%. This value suggests that most of the policies predicted as cancelled were indeed correctly classified. This adds value to the model because fewer incorrect cancellation predictions mean fewer unnecessary retention interventions. The Active class had a slightly lower precision of 83.59%. It means some active policies were incorrectly predicted as cancelled. However, the customers whose policies were incorrectly classified as cancelled might be those who need attention in the future. The retention strategies can target such clients because they are showing some patterns that can eventually lead to cancellation. For example, payment failures can be an indicator that a user may cancel after some time, having financial struggles or going through a business closure. The Cohen's kappa value of 0.642 indicates moderate to substantial agreement between the model predictions and the actual outcomes. It reflects that the model performs significantly better than random guessing. However, there is still room for improvement. The lower recall for cancellations highlights the need to address the class imbalance in the dataset. In the long run, when more data is available, resampling techniques can be used to reduce class imbalance. Even though stratified sampling was used to address this issue, the number of active policies is much higher than cancelled ones.

### Decision Tree Scorer View

Confusion Matrix

|                    | Active (Predicted) | Cancelled (Predicted) |        |
|--------------------|--------------------|-----------------------|--------|
| Active (Actual)    | 1350               | 73                    | 94.87% |
| Cancelled (Actual) | 265                | 506                   | 65.63% |
|                    | 83.59%             | 87.39%                |        |

Overall Statistics

| Overall Accuracy | Overall Error | Cohen's kappa ( $\kappa$ ) | Correctly Classified | Incorrectly Classified |
|------------------|---------------|----------------------------|----------------------|------------------------|
| 84.59%           | 15.41%        | 0.642                      | 1856                 | 338                    |

Figure 4.5: Decision Tree Prediction Results.

The Logistic Regression model reached an overall accuracy of 80.99%, which is lower compared to the Decision Tree classifier. The recall for the Active class was 92.76% which can be seen in Fig. 4.6. It indicates that the model was highly effective at correctly identifying active policies. However, the recall for the Cancelled class was considerably lower at 59.27%. This value suggests that the model missed a significant proportion of actual cancellations. This disproportion shows

that there is a bias in the model towards predicting the majority class, Active. This bias results in a higher recall for Active and negatively affects the model's ability to identify the minority class, Cancelled. The precision for the Active class was 80.78% and 81.61% for the Cancelled class. These precision values indicate that the model could fairly accurately identify cancellations among its predictions, yet less consistent than the Decision Tree classifier. The Cohen's kappa value for this model was 0.555. This kappa value further confirms that the model struggles to balance predictions across the two classes effectively. This happened most likely due to the imbalanced nature of the dataset. Also, the Logistic Regression's reliance on linear decision boundaries may have contributed to its difficulty in capturing complex nonlinear relationships.

### Logistic Regression Scorer View

Confusion Matrix

|                    | Active (Predicted) | Cancelled (Predicted) |        |
|--------------------|--------------------|-----------------------|--------|
| Active (Actual)    | 1320               | 103                   | 92.76% |
| Cancelled (Actual) | 314                | 457                   | 59.27% |
|                    | 80.78%             | 81.61%                |        |

Overall Statistics

| Overall Accuracy | Overall Error | Cohen's kappa ( $\kappa$ ) | Correctly Classified | Incorrectly Classified |
|------------------|---------------|----------------------------|----------------------|------------------------|
| 80.99%           | 19.01%        | 0.555                      | 1777                 | 417                    |

Figure 4.6: Logistic Regression Prediction Results.

### 4.2.2 Ensemble Models Results

The Random Forest classifier demonstrated robust and confident performance with an overall accuracy of 86.01%. It effectively classified the predictions with the actual labels in most cases. Of the total predictions, 1887 were correct and only 307 were misclassified. This performance was reflected in the Cohen’s kappa value of 0.681 in Fig. 4.7. Random Forest showed significant agreement between the model’s predictions and the actual results. However, further improvements could be done. This is especially evident in solving the problem of misclassifying the Cancelled class. The recall for the Active class was 93.75%, which, like Decision Tree and Logistic regression, highlights that the model can accurately identify the majority of active policies. Such a strong recall minimizes the risk of misclassifying active policies as cancelled, and it ensures that the majority of active customers remain in their original category. On the other hand, the recall for the Cancelled class was lower at 71.73%, indicating that the model struggled more in identifying all actual cancellations. However, this result is considerably better than for Decision Tree and Logistic Regression. In terms of precision, the Random Forest model maintained a balanced performance. The accuracy for the Cancelled class was 86.14%, indicating that most policies predicted as cancelled were indeed correctly classified. Similarly, the Active class achieved a precision of 85.95%, ensuring fewer misclassifications of active policies.

#### Random Forest Scorer View

Confusion Matrix

|                    | Active (Predicted) | Cancelled (Predicted) |        |
|--------------------|--------------------|-----------------------|--------|
| Active (Actual)    | 1334               | 89                    | 93.75% |
| Cancelled (Actual) | 218                | 553                   | 71.73% |
|                    | 85.95%             | 86.14%                |        |

Overall Statistics

| Overall Accuracy | Overall Error | Cohen's kappa ( $\kappa$ ) | Correctly Classified | Incorrectly Classified |
|------------------|---------------|----------------------------|----------------------|------------------------|
| 86.01%           | 13.99%        | 0.681                      | 1887                 | 307                    |

Figure 4.7: Random Forest Prediction Results.

The Gradient Boosting classifier outperformed Random Forest with an overall accuracy of 87.47%, correctly classifying 1919 policies and misclassifying only 275. The Cohen’s kappa value of 0.714 in Fig. 4.8 indicates a stronger agreement between the predictions and the actual results compared to Random Forest. This improvement demonstrates the ability of Gradient Boosting to handle complex patterns in the data. The recall for the Active class was 94.73%, showing that the model was highly effective in identifying active policies as well as other models. For the Cancelled class, the recall score improved to 74.06%, which outperforms Random Forest and indicates a better ability to capture at-risk customers. Thus,

this model is less likely to miss customers prone to churn.

### Gradient Boosting Scorer View

Confusion Matrix

|                    | Active (Predicted) | Cancelled (Predicted) |        |
|--------------------|--------------------|-----------------------|--------|
| Active (Actual)    | 1348               | 75                    | 94.73% |
| Cancelled (Actual) | 200                | 571                   | 74.06% |
|                    | 87.08%             | 88.39%                |        |

Overall Statistics

| Overall Accuracy | Overall Error | Cohen's kappa ( $\kappa$ ) | Correctly Classified | Incorrectly Classified |
|------------------|---------------|----------------------------|----------------------|------------------------|
| 87.47%           | 12.53%        | 0.714                      | 1919                 | 275                    |

Figure 4.8: Gradient Boosting Prediction Results.

Both ensemble models demonstrated significant benefits in predicting customer churn compared to simpler models such as decision trees and logistic regression. These results demonstrate the potential of ensemble methods to achieve a balance between accuracy and robustness. The Gradient Boosting demonstrated the highest accuracy, and it can be explained by its ability to capture more complex patterns because it builds trees sequentially, focusing on the errors made by previous trees.

### 4.2.3 Artificial Neural Networks Model Results

In Fig. 4.9, the confusion matrix for the Artificial Neural Networks binary classifier is provided. The ANN achieved an overall accuracy of 83%, indicating a strong match between its predictions and actual results. However, this accuracy is slightly lower compared to the ensemble models discussed earlier. The model correctly classified 1,821 cases but misclassified 373 cases. The recall was 93.29% for the Active class, which demonstrates the model's strength in accurately recognizing active policies. On the other hand, the recall for the Cancelled class was 63.65% which shows that the model had difficulty identifying all cancelled policies effectively. This relatively low recall indicates that a significant portion of actual cancellations went undetected. The calculated Cohen's kappa was 0.6029 which showed that the predictions were reasonably consistent with the actual data. In terms of precision, the ANN model performed moderately well. The precision of the Active class was 82.83% which indicates that most policies predicted as active were correctly classified. Similarly, the precision of the Cancelled class was 83.48%, indicating that most policies predicted as cancelled were indeed cancelled. However, some active policies were incorrectly classified as cancelled. This misclassification can be a warning that customers show early signs of dissatisfaction or payment issues. While the ANN model showed promising results, similar to other models, its performance for the Cancelled class was noticeably affected by



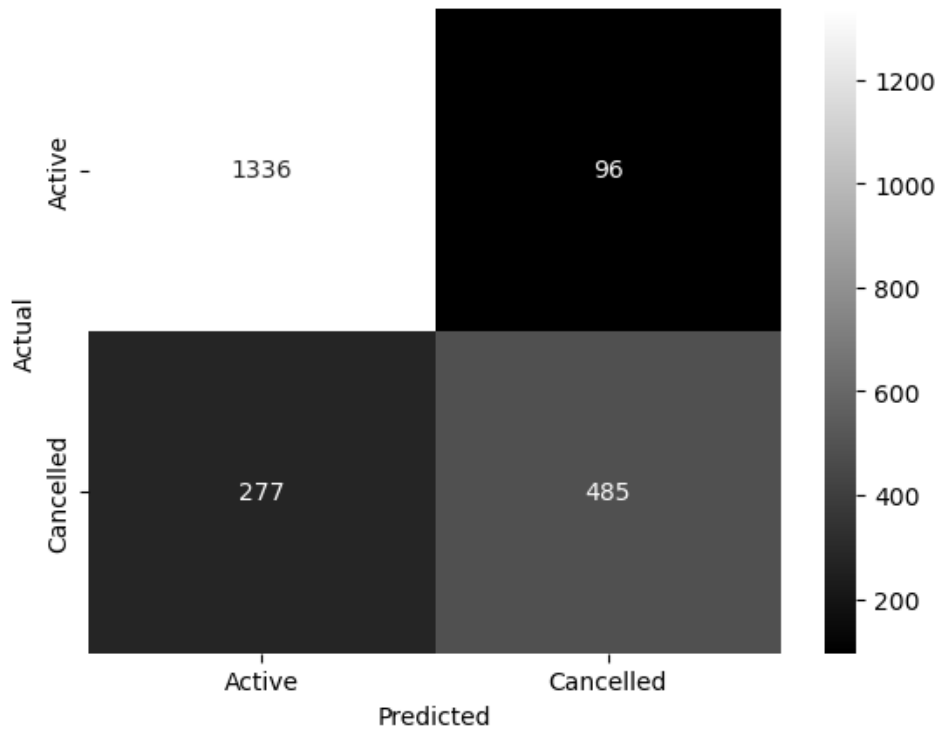


Figure 4.9: Artificial Neural Networks Prediction Results.

the dataset being imbalanced. The model could benefit from additional improvements. For example, oversampling the minority class, undersampling the majority class, or using advanced algorithms to more effectively handle the imbalance. Also, further fine-tuning of hyperparameters can improve the results.

#### 4.2.4 Large Language Model Results

The LLM approach achieved an overall accuracy of 72%, which is lower than the other models tested. However, because of manual testing of LLM, only 100 rows of data were analysed, of which 47 were active and 53 were cancelled. Despite this, the recall for the class cancelled was 73.58%, which can be seen in Fig. 4.10 making it the second-highest of all models in this metric after Gradient Boosting. This indicates that LLMs are particularly effective at dealing with cancelled policies, which is valuable because it creates a base for a retention strategy. For the Active class, the ensemble recall is 70.21%, and it suggests a moderate ability to correctly classify active policies. The accuracy of the Cancelled class, supported by the high recall, shows that most policies marked as cancelled were indeed cancellations. This is critical to minimising false positives since unnecessary retention measures can be costly and ineffective. While recall for active policies is lower than other models, it

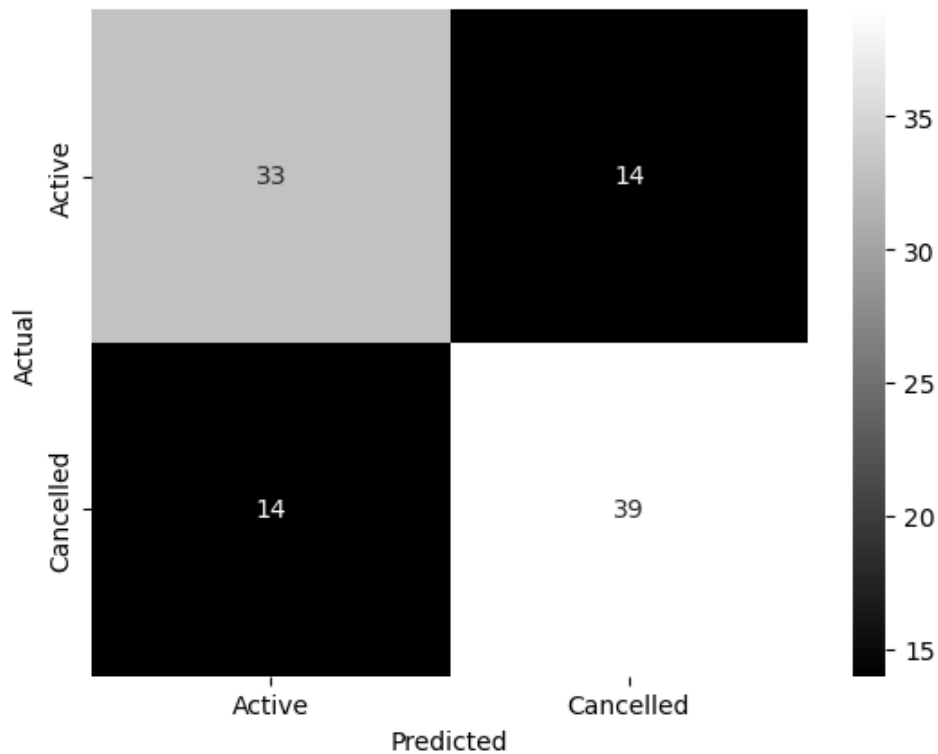


Figure 4.10: LLM Prediction Results.

is worth noting that LLM was better at detecting cancellations, while other models struggle with it. This can be explained by the flexibility of prompt engineering, where considerations such as insurance terms and conditions, and the company's cancellation practices can be provided via prompt, making the model smart in understanding the insurance context for the specific company. The prompt template used for interacting with the LLM is shown in Fig. 4.11. It included considerations regarding payment patterns, information about the company's cancellation practice for annual policies, and consideration about revenue outliers,

These results can not be highly reliable because they were tested on small data. Further automatization of testing by accessing the paid LLM API can provide more accurate results. However, this approach revealed that in general, LLM provides the highest interpretability since it provides clear reasoning for each predicted outcome. The main advantage of LLM compared to other models, it was possible to use cancellation reason labels to provide higher context and get the prediction of not only the cancellation status but also potential cancellation reason. This makes the LLM model more powerful than other conventional machine learning models. In Fig. 4.12, the reply from the NotebookLM is demonstrated. The NotebookLM provides the rationale for the predicted instance by analysing the historical data which was

"You are analyzing unseen policyholder data based on historical patterns of cancellations and active policies. Using your knowledge of past cancellation reasons and their corresponding statuses, predict whether the following policy is likely to be ""Cancelled"" or remain ""Active."" Additionally, provide a rationale for your prediction based on the provided details.

Key Considerations:

1. Payment Patterns: No payment failures are a strong indicator of non-cancellation. Check for alignment among:
  - Amount Received
  - Monthly Fees
  - Quote Price
  - Outstanding Loan Amount  
Inconsistent payments or missing amounts should be prioritized when assessing the likelihood of cancellation.
2. Annual Policies: For annual policies with outstanding loan amounts always equal to zero, cancellations are still possible. Take into account the possibility of cancellation within the first 30 days (Number of Months from Inception is 0) for a full refund.
3. Revenue as a Low Priority: While lower revenue often correlates with business closure or transitions out of self-employment, treat revenue trends as a secondary factor. Focus on payment patterns first.
4. Outliers: Recognize that there may be outliers in the data, such as unusually low or high business revenue, which may not align with general trends.
5. If predicting ""Cancelled,"" specify the most likely reason for cancellation based on observed trends (e.g., missed payments, business closure, or other factors).

Entry 5:

numberOfMonthFromInception: 9 amountReceived: 407.97 outstandingLoanAmount: 118.24  
feesPaid: 53.21 productType: Healthcare Professionals Civil Liability (PI & PL) quotePrice: 473  
basePremium: 280 policyType: Monthly businessType: Sole Trader businessAge: 1 businessSize: 1  
monthlyFees: 70.95 businessRevenue: 1000 FailureCount: 8

"

Figure 4.11: Prompt template sent to NotebookLM.

uploaded as the training set. During the testing, there were False Cancelled. Even though these predictions decreased the accuracy, they can be reasonable since the model explained why they might cancel. These instances fall under customers at risk and can be targeted with retention strategies.

**Policy Status Prediction for Entry 5:**

**Prediction:** Cancelled

**Rationale:**

- **High Failure Count:** Entry 5 has a FailureCount of 8, a significant number that strongly suggests a history of missed or unsuccessful payments. [Source 1] This is a primary indicator for policy cancellation. [Source 1]
- **Potential Payment Issues:** Although the amountReceived (407.97) plus the outstandingLoanAmount (118.24) roughly aligns with the quotePrice (473), the high FailureCount raises serious doubts about the policyholder's ability to maintain consistent payments. [Source 1]
- **Monthly Policy Considerations:** As a monthly policy, regular payments are crucial. The high FailureCount suggests chronic payment issues, making cancellation highly likely. [Source 1]
- **Possible Cancellation Reasons:** Considering the significant FailureCount, the most probable reason for cancellation is **repeated missed payments**. [Source 1] This aligns with the Key Considerations provided, which emphasize payment patterns as the primary factor. [Source 1]

**Conclusion:** The high FailureCount (8) in Entry 5, despite a relatively consistent total payment amount in relation to the quotePrice, points towards **cancellation** due to a pattern of missed payments. [Source 1] This prediction aligns with the emphasis placed on payment patterns as the primary predictor for policy cancellations. [Source 1]

📄 👍 💬 🚩 Save to note

Figure 4.12: LLM reply to the user prompt.

### 4.3 Models Comparison

The Decision Tree, Logistic Regression, Random Forest and Gradient Boosting classifiers were all created and tested in the KNIME analytics platform using the same conditions, splitting the data with constant random seed. So, the comparison can be made between these four models, given they were deployed in the same way. The Gradient Boosting demonstrated the overall best performance among the rest of the models. The precision metric for Gradient Boosting was higher than for the other three models. However, the Decision Tree showed a slightly better result in recall for the Active class. In Fig. 4.13, the ROC curve for four models is illustrated. This curve plots the true positive rate against the false positive rate. The graph also includes AUC, which stands for the area under a curve and shows the model's ability to distinguish between the two classes. The ensemble techniques achieved the highest results in differentiating the cancelled and active classes, with AUC equal to 0.91 and 0.89.

Comparing all models created and tested, the accuracies were summarised in Fig. 4.14. The ANN model had the same splitting ratio for training and testing sets as the previously mentioned four models, with the same number of entries. So, it makes its conditions close to four models created in KNIME. In contrast, LLM lacked the testing, limited to only 100 entries. LLM's actual accuracy can be compared to the other models if it gets further automation of testing and tuning.

### 4.3 – Models Comparison

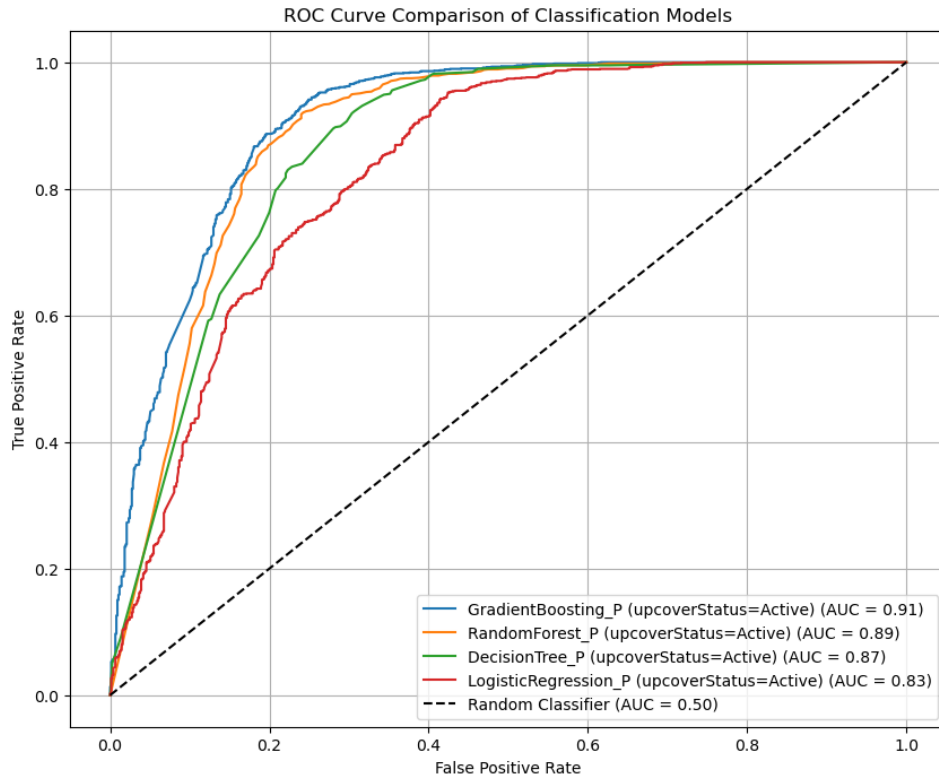


Figure 4.13: ROC Curve Comparison Across Models.

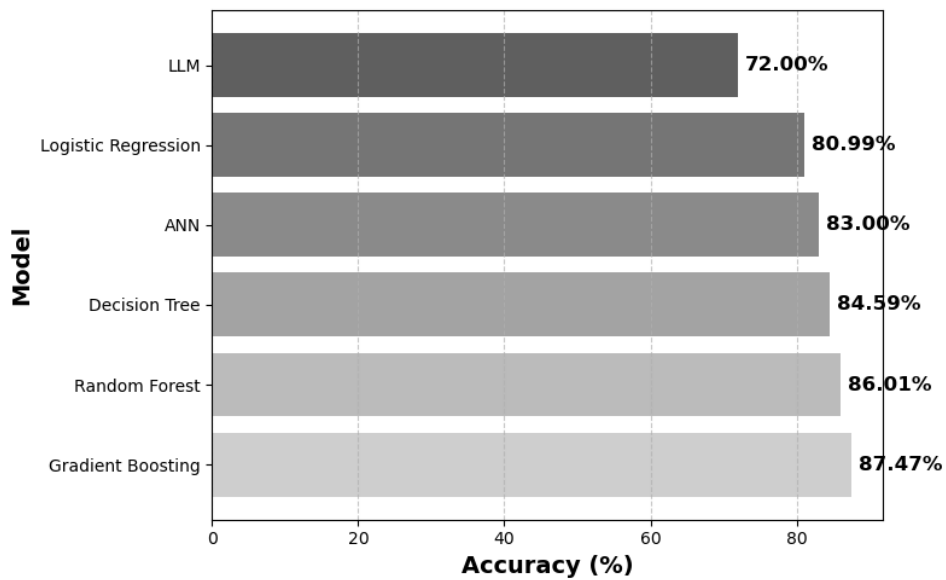


Figure 4.14: Accuracy Comparison Across All Models.

## 4.4 Challenges and Perspectives

During the exploratory data analysis in section 4.1, based on Table 4.1, it was found that a considerable part of the cancellations were done because of external reasons. This may create noise in the dataset when customers do not show any sign of dissatisfaction but end up cancelling their policies. For example, they can cancel because of family issues or health concerns, but the created models do not take into account any features related to that, so this part of customer information is unknown. As a result, it creates limitations in prediction, and it explains why recall for the Cancelled class is low for all models. In contrast, False Cancelled policies can indicate that those customers have signs of cancellation in future. The other challenge to target is the imbalanced dataset. So, better results can be achieved when a larger amount of data is collected, so it can allow having equal sampling or other alternatives. For the current dataset, equal sampling was not applied because it would reduce the size of the dataset, which can lead to overfitting. LLM also have the potential to provide higher accuracy with more extensive data provided. Creating a knowledge base with paid LLM API like Gemini can open new perspectives for tuning and testing. For all models, further feature engineering and optimization can help to reach a better performance.

# Chapter 5

## Conclusion

This thesis demonstrates the power of machine learning and deep learning techniques in predicting customer churn and understanding the root causes of insurance policy cancellations. Through a systematic approach including data preprocessing, feature selection, model development, and results evaluation, this study highlights the practical applications and limitations of different prediction models. Ensemble methods such as Random Forest and Gradient Boosting proved to be the most accurate methods, achieving up to 87.47% accuracy while effectively balancing predicting performance across classes. Artificial Neural Networks showed competitive results, but required optimization to better handle imbalanced datasets. Large Language Model, despite their lower accuracy in this case, proved its value in interpreting complex cancellation patterns and generating actionable insights.

The study's systematic approach and results evaluation of different models highlight applicability beyond the insurance domain. Industries such as telecom, media streaming, and SaaS can use similar approaches to address subscription cancellations, improve customer retention, and refine targeted marketing efforts. However, challenges such as manual data labelling, unbalanced datasets, and limitations of LLM use remain areas for improvement. Further work should focus on addressing these limitations with a more comprehensive approach to interpreting the findings across business processes. Expanding behavioural and financial datasets can further improve the predictive models.

Overall, this study contributes to the growing body of knowledge on predictive analytics and churn detection. The findings may already offer insurers and similar industries areas to focus on, actionable tools to address churn, adapt to dynamic customer needs and maintain a competitive advantage.





# Bibliography

- [1] Clara-Cecilie Günther et al. «Modelling and predicting customer churn from an insurance company». In: *Scandinavian Actuarial Journal* 2014.1 (2014), pp. 58–71.
- [2] Miguel Ángel De la Llave, Fernando A López, and Ana Angulo. «The impact of geographical factors on churn prediction: an application to an insurance company in Madrid’s urban area». In: *Scandinavian Actuarial Journal* 2019.3 (2019), pp. 188–203.
- [3] Stéphane Loisel, P. Piette, and C. Tsai. «APPLYING ECONOMIC MEASURES TO LAPSE RISK MANAGEMENT WITH MACHINE LEARNING APPROACHES». In: *ASTIN Bulletin* 51 (2019), pp. 839–871. DOI: 10.1017/asb.2021.10.
- [4] Patrick M Liedtke. «What’s insurance to a modern economy?» In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 32 (2007), pp. 211–221.
- [5] Liyan Han et al. «Insurance development and economic growth». In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 35 (2010), pp. 183–199.
- [6] Snežana Živković and Aleksandra Ilić Petković. «Insurance as a Prerequisite for Risk Management in Working and Living Environment». In: *Facta Universitatis, Series: Working and Living Environmental Protection* (2020), pp. 139–150.
- [7] Catalina Bolancé, Montserrat Guillen, and Alemar E Padilla-Barreto. «Predicting probability of customer churn in insurance». In: *Modeling and Simulation in Engineering, Economics and Management: International Conference, MS 2016, Teruel, Spain, July 4-5, 2016, Proceedings*. Springer. 2016, pp. 82–91.
- [8] Manuel Leiria, Nelson Matos, and Efigénio Rebelo. «Non-life insurance cancellation: a systematic quantitative literature review». In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 46.4 (2021), pp. 593–613.

- [9] Montserrat Guillén et al. «Time-varying effects in the analysis of customer loyalty: A case study in insurance». In: *Expert systems with Applications* 39.3 (2012), pp. 3551–3558.
- [10] S. Johnson. *Customer Retention is the Most Cost-Effective Path to Growth*. Retrieved from <https://iireporter.com/customer-retention-is-the-most-cost-effective-path-to-growth>. Insurance Innovation Reporter. 2023.
- [11] A. Gallo. *The Value of Keeping the Right Customers*. <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>. Harvard Business Review. Aug. 2022.
- [12] PwC. *Building Trust with SMEs in the Insurance Industry*. <https://www.pwc.co.uk/industries/insurance/insights/building-trust-with-sme.html>. PwC UK. Aug. 2022.
- [13] Pius Babuna et al. «The impact of Covid-19 on the insurance industry». In: *International journal of environmental research and public health* 17.16 (2020), p. 5766.
- [14] Amelie Gamble, E Asgeir Juliusson, and Tommy Gärling. «Consumer attitudes towards switching supplier in three deregulated markets». In: *The Journal of Socio-Economics* 38.5 (2009), pp. 814–819.
- [15] Yunxuan He, Ying Xiong, and Yiting Tsai. «Machine learning based approaches to predict customer churn for an insurance company». In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE. 2020, pp. 1–6.
- [16] Andres Yecid Rodriguez Calderon and Olmer Garcia-Bedoya. «Design of a predictive model for customer retention in the cancellation of an insurance policy». In: *2020 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE. 2020, pp. 1–6.
- [17] Fatima Enehezei Usman-Hamza et al. «Intelligent decision forest models for customer churn prediction». In: *Applied Sciences* 12.16 (2022), p. 8270.
- [18] Arno De Caigny et al. «Incorporating textual information in customer churn prediction models based on a convolutional neural network». In: *International Journal of Forecasting* 36.4 (2020), pp. 1563–1578.
- [19] Dayong Liu et al. «Research on Renewal Prediction of Life Insurance Policy Based on Back Propagation (BP) Neural Network». In: *2022 International Conference on Big Data, Information and Computer Network (BDICN)*. IEEE. 2022, pp. 664–668.
- [20] Nikhil Mathur et al. «Analyzing Consumer Behavior Predictions: A Review of Machine Learning Techniques». In: *2022 International Conference on Advances in Computing, Communication and Materials (ICACCM)*. IEEE. 2022, pp. 1–5.

- [21] Shiyunyang Zhao. «Customer Churn Prediction Based on the Decision Tree and Random Forest Model». In: *BCP Business & Management* 44 (Apr. 2023), pp. 339–344. DOI: 10.54691/bcpbm.v44i.4840. URL: <https://bcpublication.org/index.php/BM/article/view/4840>.
- [22] I. AlShourbaji, N. Helian, Y. Sun, et al. «An efficient churn prediction model using gradient boosting machine and metaheuristic optimization». In: *Scientific Reports* 13 (2023), p. 14441. DOI: 10.1038/s41598-023-41093-6.
- [23] Sen Hu et al. «A spatial machine learning model for analysing customers' lapse behaviour in life insurance». In: *Annals of Actuarial Science* 15.2 (2021), pp. 367–393.
- [24] A. Khattak, Z. Mehak, H. Ahmad, et al. «Customer churn prediction using composite deep learning technique». In: *Scientific Reports* 13 (2023), p. 17294. DOI: 10.1038/s41598-023-44396-w.
- [25] James E Dobson. «On reading and interpreting black box deep neural networks». In: *International Journal of Digital Humanities* 5.2 (2023), pp. 431–449.
- [26] Vishvesh Soni. «Large Language Models for Enhancing Customer Lifecycle Management». In: *Journal of Empirical Social Science Studies* 7.1 (Feb. 2023), pp. 67–89. URL: <https://publications.dlpress.org/index.php/jesss/article/view/58>.
- [27] Yun Li et al. «A deep multimodal autoencoder-decoder framework for customer churn prediction incorporating chat-GPT». In: *Multimedia Tools and Applications* (2023), pp. 1–27.
- [28] Caesar Balona. «ActuaryGPT: Applications of large language models to insurance and actuarial work». In: *Available at SSRN 4543652* (2023).
- [29] Daniël van Dam and Raymond van Es. *The potential of large language models in the insurance sector*. Accessed: 2024-12-01. Feb. 2024. URL: [https://www.milliman.com/en/insight/potential-of-large-language-models-insurance-sector?utm\\_source=chatgpt.com](https://www.milliman.com/en/insight/potential-of-large-language-models-insurance-sector?utm_source=chatgpt.com).