# Predicting customer churn for an insurance company by utilizing behavioural features

Dreyer, Simon A.

*Award date:*
2023

# Predicting customer churn for an insurance company by utilizing behavioural features

Master Thesis

*Author:*
Simon Dreyer
0969192

*Supervisors:*
**Eindhoven University of Technology**
Dr. Ir. Laurens Bliek
Dr. Baris Ozkan
**Company supervisor**
Dr. Berlinda Hermsen
Antal Nusselder, MSc

*A thesis submitted in partial fulfilment of the requirements for the degree of*
*Master of Science*
*in Operations Management and Logistics*

Information Systems
Industrial Engineering & Innovation Sciences

Nijmegen, 5 February 2023

# Abstract

Customer churn prediction is a field that utilizes machine learning techniques to predict the probability of a customer leaving the company or classification of the customers as leaving or not. The main objective of this thesis is to study churn prediction and more specifically analyze what the influence of behavioural features is on the predictive performance of predictive churn models, applying the knowledge to a Dutch insurance company. This study begins with exploring the relevant concepts, reviewing the literature conducted in the field of machine learning, churn prediction and other relevant topics. The literature review is followed by an empirical study utilizing data provided by the insurance company, comparing different models. Before comparing the different models, various machine learning algorithms are analyzed to determine the most suitable algorithm, followed by an analysis of the effect of class imbalance which is present in the provided dataset. In the algorithm comparison it was found that the boosting algorithms performed the best in this case, with LightGBM being slightly better than the other boosting algorithms. The models that include behavioural features are compared against a model that has no behavioural features included and a model based on the current practice of the company. Following this comparison, it was seen that the differences amongst models were small, however focusing on the top-decile showed slightly better results for the model involving behavioural features aggregated on a yearly basis and the model that utilized normalized behavioural features. The evaluation, which was done using metrics like AUC, $F_1$ score, precision, recall, and top-decile lift, showed that the inclusion of customer behaviour could be potentially beneficial for a predictive churn model in the insurance industry context.

**Based on the information in this thesis, this thesis can only be used for educational purposes. No other purposes are allowed.**

# Management Summary

This research is part of finalizing the Master program Operations Management and Logistics at the Eindhoven University of Technology. This project researches how the inclusion of features based on customer behaviour can influence the prediction model for customer churn, in the context of a Dutch insurance company. Based on the research objectives and the context, the main research question was defined as:

*How does the inclusion of customer behaviour influence the predictive modelling of customer churn within the insurance industry?*

The prediction of customers that are likely to leave the company, the prediction of customer churn, is considered to be an important topic for many businesses, especially for companies operating in a highly competitive and saturated market like the Dutch insurance industry. The costs of attracting a new customer is estimated to be at least three times more expensive than retaining an existing customer, emphasizing the need of having a well-functioning customer retention strategy. Predictive churn models play a major role in this strategy, making customer churn models an important theme for business to research and improve.

Predicting the probability of a customer leaving the company is complex, where various aspects play a significant role which is not easily modelled. The inclusion of customer behaviour is seen as a possible extension to existing churn models, while current churn models often only incorporate socio-demographical information and information regarding the product. Multiple studies have shown the benefits of including customer behaviour as features in predictive churn models, often entailing the use of the RFM framework. These features measures information regarding the behaviour of a customer, measuring the recency of interaction with the customer, frequency of interaction with the customer, and the monetary value of a customer. Other extensions have also been proposed with mixed results, making behavioural features an interesting topic to research in the context of customer churn models.

Customer churn models are the application of machine learning techniques to classify customers as potential churners, assigning a probability of churning to the customers. The modelling of customer churn starts with a sampling phase, a phase where the data is selected for the model fitting. The sampling phase is an important aspect of churn modelling while class imbalance is inherent to churn prediction. A company often has significantly more non-churning customers than churning customers, complicating the application of models. Sampling can be done through a wide range of strategies, of which some have been analyzed in this research. The sampling phase is followed by the model fitting phase, where machine learning algorithms are utilized to build a predictive model. Lastly, the models are evaluated using various evaluation criteria to determine the optimal model.

This study entails all three phases, analyzing aspects of each phase, with the main focus being on the inclusion of customer behaviour in current predictive churn models. Before analyzing

which model with certain features performs best, an analysis is performed to review the most frequently used techniques. From this review a wide range of techniques were explored, selecting the Logistic Regression, Decision Tree as baseline for the other techniques. Based on literature reporting promising results of ensemble learning algorithms, four ensemble techniques were included. Three gradient boosting algorithms and Random Forest were included, comparing all techniques on $F_1$ score, AUC, precision, recall, and the top-decile lift. The boosting algorithms showed in comparison a relatively good performance, obtaining the highest score in $F_1$, AUC and top-decile lift while reporting relatively high performance in the other metrics. Looking specifically at the top-decile lift, a better performance of the boosting algorithms is also observed. The LightGBM algorithm performed better than the other boosting algorithms, reporting minor improvements.

The LightGBM model is subsequently used for the analysis of the models with and without customer behaviour incorporated. Three behavioural models were defined, the FI-monthly, RFI-yearly, and RFI-normalized models. FI-monthly measures the behaviour using frequency and intensity features on a monthly basis, where RFI yearly aggegrates the recency, frequency and intensity features on a 12-month basis. Lastly, the RFI-normalized normalizes the monthly values of recency, frequency, and intensity using the deviation. The models are compared against a model that has no behavioural features included and a model that utilizes two behavioural features, a model that would be implemented by the company. Following the comparison, comparable results were seen amongst the various models. The values of the metrics showed minor differences across the models, while analyzing the top-decile provided more insights and showed a minor improvement for both the RFI-yearly and RFI-normalized model when focusing on the top 10% of the policies that have the highest chance of being cancelled.

Further investigating the variable importance and the partial dependence mainly showed that the models utilizing customer behaviour became a bit more interpretable, having partial dependence plots that were easily interpreted. However, variable importance showed the relatively low importance of the behavioural features, explaining the minor differences seen in the values of the metrics. Motivated by the performance of the models, an analysis was performed the analyze whether the predictive performance of the models could be enhanced through a sampling strategy. This analysis showed that the predictive performance for the LightGBM models remained fairly constant across various class distributions, while the sampling strategies caused a higher degree of variation in terms of predictive performance for the Random Forest and Decision Tree models.

Answering the main research question as a conclusion, it is concluded that the inclusion of customer behaviour in predictive churn model marginally increases the predictive power of the models, which is already useful for the company due to the low predictive power of the models. A second influence of including behavioural features is seen in the top-decile, causing the number of actual cancelled policies to be slightly higher than the model without any customer behaviour included. Lastly, the inclusion of customer behaviour in predictive churn models increases the interpretability of the model. The effect of behavioural features is relatively easy interpreted, enhancing the interpretation of the model.

### Recommendations for the company

Following the results and the observation that the RFI-yearly performed slightly better in terms of the top-decile lift, a recommendation would be that the company can investigate the impact of including recency in their model. Currently, the foreseen model includes customer behaviour by measuring the number of phone calls in the last twelve months and the number of online and application interactions. Based on the feature importance of the recency variables in the RFI-yearly model and the slightly better performance in the top-decile lift, the model could possibly be improved by including recency as a feature. Secondly, the model did not include any policy that was active for less than a month. This group has been excluded based on not having data for a month and the graphical analysis which potentially indicates a difference in interactions for these policies. Further researching this difference, checking whether this is statistically significance can provide valuable insights for the company, potentially enhancing their retention strategy for this group of customers.

**Limitations of the research**

- Each policy has twelve data points when the policy was active for the whole year of 2021, which is less than twelve for churned policies or new policies. This introduces bias and increases the imbalance, which potentially affects the predictive performance of the model.
- The monthly observation period of predicting whether the policy is being cancelled next month is highly specific, making the prediction task even more complex.
- No feature selection has been performed while the research was focused on including the behavioural features. However, feature selection can potentially improve the models, increasing the predictive performance.
- Costs and benefits have not been accounted for in the metrics, which is a main driver in the weighing of the company.

**Future Research**

- Investigating a more qualitative approach of incorporating customer behaviour, utilizing emotions or the visiting of specific web pages.
- Investigating the performance of semi-supervised for churn prediction within an insurance context, being less dependent on labelled entries.

# Acknowledgements

This thesis is the result of five months of hard work in order to fulfill my duties at the TU/e, finalizing my master degree in Operations Management and Logistics. This thesis has been conducted from September 2022 to February 2023 at the Dutch insurance company OHRA and would not be the same without the help of some people. I would therefore like to take this opportunity to thank some people who helped or motivated me during the master and during my master thesis.

I would like to thank my supervisors from the university, my first supervisor Laurens Bliek who was always available when I had questions and ensured that I stayed on track with the weekly meetings. I would also like to thank my second supervisor, Baris Ozkan, who provided some valuable insights regarding the conceptual aspects, causing me to view my research through other perspectives.

Secondly, I would like to thank the supervisors from the company in which this research was conducted. My first supervisor Berlinda Hermsen was always available for questions and helped me tremendously with ensuring that all the required data was provided while also having time all the various meetings that we had. Also many thanks to Antal Nusselder, who took over when Berlinda was not available and provided some valuable insights regarding the technical approaches that have been taken in this research.

Lastly, I would like to thank my family and friends, who always supported me during the process and made sure that I could focus on my thesis whenever it was needed. Also a special thanks to my girlfriend, Fleur, who was always there for me and helped me through difficult periods.

This research project concludes my six years at the university, a time that I thoroughly enjoyed and formed me as the person that I am right now. It has been an important period that I will always be remembering.

Simon Dreyer

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Service providers within the insurance industry experience an ever-increasing competition, in a market that is already seen as highly competitive due to the yearly renewal of contracts and competitors that offer almost identical products. Retaining the existing customers is therefore becoming increasingly important for any organization that desires to remain competitive. A customer moving from one company to a similar company is defined as the *customer churn*, often abandoning a service provider or organization either by inactivity or by formal termination of a contract. Tools in the field of Customer Relationship Management (CRM) attempt to mitigate the probability of a customer churning to increase the competitiveness of the organization, enhancing the profitability and value of a customer. These tools involve the identification of the customers that are most likely to leave the company in the near future, hence the use of predictive churn models. *Churn prediction* is concerned with predicting what the probability of a customer churning is and is considered to be the main subject of this thesis.

The research topic of predictive churn models has been researched in a wide range of different industries like the telecommunication industry (Alboukaey, Joukhadar, & Ghneim, 2020), the financial industry (Günther, Tvete, Aas, Sandnes, & Ørnulf Borgan, 2014; Keramati, Ghaneei, & Mirmohammadi, 2016), and the gaming industry (Lee, Kim, & Lee, 2017). Predictive churn models often involve logistic regression, decision trees or neural networks to predict the churn probability (Eria & Marikannan, 2018; Mahajan, Misra, & Mahajan, 2015), utilizing various features based on information like socio-demographic information, billing data, and usage data. The incorporation of behavioural information has been studied in multiple studies, however when looking at these behavioural features a common approach is observed within the feature handling process. Regularly, the behavioural features are aggregated in such manner that the information over a specific time span is reduced to one data point for each user, often done through the use of a mean function. Although this approach simplifies the task of incorporating behavioural traits in a predictive churn model, valuable information in the form of latent temporal information is not incorporated. For example, a customer that barely had any interaction with the company but increasingly became more active, can potentially be seen as a customer that is not as satisfied anymore and subsequently having a higher probability of churning. This information is lost with the aggregation often seen within the incorporation of behavioural features like the frequency or recency of interaction.

Including behavioural features can for instance also be done through the use of normalization, where the features are normalized among the year average of the variable. This requires the use of more dynamic variables in the form of rolling means and averages, while the interaction of customers is time dependent. By implementing this strategy, it is hypothesized that more information is included that could potentially be beneficial for the predictive performance of the customer churn models.

This research attempts to investigate how the inclusion of behavioural features influences the predictive performance of customer churn models, comparing the strategy among other approaches where sequential data is for example aggregated in one single data point. Therefore, five different models are compared using various techniques. All models will be compared using six specific algorithms, analyzing the performance per algorithms but also researching how the context of customer churn in the form of class imbalance influences the models. The main objective of this research is to explore the research field of customer churn prediction, finding and analyzing potential improvements to predictive churn models in the form of behavioural patterns, and comparing the various techniques utilized.

Understanding the area of subject and the context is an important part of conducting the research, where literature can aid in understanding the present state of a specific research area and identifying existing gaps in current knowledge regarding the research subject (Arshed & Dansen, 2015). While focusing on the implementation of predictive churn models that capture information regarding behavioural patterns and changes in behavioural patterns, this research project provides insight in customer churn and predicting this aspect within an insurance context. This chapter discusses the research area, the research objectives and the design of the study. The study starts with a description of the research area and the context in Section 1.1. Following the context, the problem outline and relevance are discussed in Section 1.2, after which the research strategy and outline of the project are defined in Section 1.3.

## 1.1 Research Area and Context

In order to understand the research area, a complete overview of the research area, problem outline and the relevance of the problem are needed. Firstly, the research area will be described to obtain a general understanding of the research context, after which the company context is provided and the corresponding problem outline.

### 1.1.1 Research Area

Customer retention is concerned with retaining the customers for a specific company, where customer retention can be defined as a measure of the customer's intention to remain with a service provider like an insurance company (Edward & Sahadev, 2011). Customer retention is described by Alkitbi, Alshurideh, Al Kurdi, and Salloum (2020) as a continuous process, consisting of both finding new customers and retaining those customers. Customer retention enables businesses to become more cost-effective (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Torkzadeh, Chang, & Hansen, 2006), but also enables the company to improve the value of the customer in the long term (Hsin Chang, 2007). Although customer retention can provide high value to the company, it is a complex combination of various issues like the pricing, service, personal preferences, convenience and many other (Smith, Willis, & Brooks, 2000). Due to these interactions between various issues, the subject of customer retention is an interesting topic both from an industry perspective as well as a research perspective.

Customer retention is part of Customer Relationship Management (CRM), which is described through various definitions. Kumar and Reinartz (2018) define CRM as a strategic process with the goal of optimizing the combined value of customer for a company, Chalmeta (2006) describe CRM as a customer-focused business strategy with its origin in the relational marketing area, while others deem CRM as a tool to manage customer relationships (Meena & Sahu, 2021). Despite having multiple interpretations the aspect of the relationship between the customer and the business is central, which can be attributed to the fact that the origin of CRM is within relational marketing. Relational marketing is a research area focusing on the strategy to attract, maintain, and enhance customer relationships which has been described by Ndubisi (2005) as lacking empirical validity. CRM has emerged in a response to this lack of empirical validity while maintaining the importance of customer relationships.

CRM has become a research field with increased attention as a marketing concept among both academia and businesses (Wahlberg, Strandberg, Sundberg, & Sandberg, 2009). Customer Relationship Management has gained attention from businesses while companies nowadays aim for having long-term relationships in an attempt to maximize the value of the customers (Risselada, Verhoef, & Bijmolt, 2010), indicating the importance of managing the relationship between business and customer. Central to this process or strategy is the involvement of the concept of customer value, the value that a customer provides to the company also known as Customer Lifetime Value (CLV). Swift (2001) defines four main pillars within CRM using the perspective of the firm, respectively customer acquisition, customer retention, customer loyalty, and customer profitability. Gupta et al. (2006) combines customer loyalty and customer profitability in customer expansion, but furthermore also builds upon customer acquisition and customer retention. As mentioned before, the aspect of customer retention will be the aspect of CRM that will be mainly researched throughout this research project.

There are many different business approaches for retaining customers ranging from focusing on a positive customer-employee relationship (Hawkins & Hoon, 2019), using integrated marketing communications (Thaichon & Quach, 2015), and using models to early identify customers which are more likely to leave the business (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015). Nowadays technology have an important role within the field of customer retention through enabling faster and more ways of communication with customers, more marketing channels, but also through enabling the use of (predictive) models to determine whether customers are likely to leave the company. Predictive modelling is a research field of great importance and used in many different areas of research. Besides being an important research area, predictive modelling is commonly used by industries, governments and science for all different kind of purposes (Mackenzie, 2015). Predictive modelling builds upon the analysis of data sets to discover patterns within the data, also known as data mining, and using those patterns to predict the likelihood of future events (Crockett & Eliason, 2016). Predictive modelling can enhance the customer retention efforts made by companies through providing a model which predicts the likelihood that a customer is leaving the company.

Models that predict the likelihood of customers leaving the company are generally known as customer churn prediction models. By predicting customer churn, a company is able to segment customers based on their probability of leaving the company, but also enables to calculate the potential value of the customer based on their lifetime. Following the increased attention for Customer Relationship Management and customer retention, these models also gradually received more attention. Churn models generally are based on machine learning algorithms like Logistic Regression, Decision Trees learning or Linear Regression, but other techniques like Neural Networks or Support Vector Machines are also used (Vafeiadis et al., 2015). The research in the area of customer churn does not only focus on the various techniques available, but also on the context of implementation. Research has been mainly conducted within the telecommunication industry (Alboukaey et al., 2020; Wei & Chiu, 2002), but has also been conducted within the insurance industry (Günther et al., 2014), the banking industry (Keramati et al., 2016), and other industries like the gaming industry (Lee et al., 2017). The context in which the churn models are deployed is highly relevant for the predictive power of the churn models, while important aspects like customer behaviour and the relationship between the customer and the firm differ per industry. Customer churn and the predicting of this customer churn is therefore a versatile research area which can have significant implications for the industry and companies involved. Based on this versatility and the combination of utilizing data analyses to implement more effective marketing measures, the research area has been defined to be predictive churn models.

### 1.1.2 Company context

The research entails a case study where the focus will be on one specific company, a Dutch insurance company known as OHRA. OHRA is an insurance company founded in 1925 and has two main divisions, a health insurance division and a non-life insurance division. The research is conducted within the non-life insurance division, the division that offers a wide-range of non-life insurances ranging from car insurance to home insurance. The division of OHRA concerned with non-life insurance is part of the Nationale-Nederlanden group (NN group), an umbrella organization with several brand names. The other division, the life insurance division, is part of the CZ group, and will not be taken into consideration while the businesses are considered to be two fairly distinct entities. NN group is the second largest insurer in the Dutch non-life insurance industry with a market share of 20.4 % (KPMG, 2021). However, OHRA is only a minor part of the Nationale-Nederlanden group, meaning that OHRA can be considered to be a relatively small insurance company in the Dutch insurance industry.

The non-life insurance division has a varied portfolio with insurances like pet insurance, liability insurance, home insurance, and car insurance. The research focuses on one specific insurance within the non-life insurance division, the car insurance. This insurance is one of the larger products within the portfolio of OHRA, therefore being of high importance for the business. The main goal of OHRA across both divisions is to offer insurance without unnecessary administrative tasks, offering easy and accessible insurance for every customer.

The Dutch insurance industry in general can be described as a highly competitive market where operating costs are an important aspect of the industry. This is observed within the consolidation of the market, where various companies have merged together causing a decrease in the number of registered insurers (KPMG, 2021). Being a heavily consolidated market, it is observed that the three largest non-life insurers have a total market share of 61.9%, as seen in Table 1.1, whereas the other 38.1% is divided by 61 other insurers. The market is considered to be a saturated market, being a market that does not significantly grow but does experience increasing costs for the insurers. Damages that have to be covered by the insurers in terms of monetary value are increasing due to more complex technologies used within products and increasing costs for personal injuries. This causes the profitability of insurance companies to be under pressure, while increasing the premium for customer is not favourable in a highly saturated and competitive market.

### 1.1.3 Problem outline and Relevance

As mentioned in subsection 1.1.2, the market in which an insurance company like OHRA operates is considered to be highly competitive and saturated. Therefore, operating costs and the reduction of operating costs is an important aspect of operating in the insurance industry. One of the main

Table 1.1: Description of the largest Dutch insurers in 2020 (KPMG, 2021)

| Insurer | Gross premium (millions) | Market Share (%) |
|---|---|---|
| Achmea | 3627 | 23.6 |
| Nationale-Nederlanden | 3128 | 20.4 |
| ASR | 2749 | 17.9 |
| **Total** | **9504** | **61.9** |

areas of interest for reducing the operating costs is the area of customer retention, while research suggests that gaining a new customer is up to twelve times more expensive than retaining an existing customer (Torkzadeh et al., 2006). This difference in costs emphasizes the need for customer churn management within a highly cost competitive industry like the insurance industry. In addition to costing less, Mozer et al. (2000) have concluded that marketing campaigns for retaining customers are also significantly more efficient than marketing campaigns aimed for attracting new customers. Decreasing the number of customers leaving is therefore a good method of reducing the operating costs and subsequently increasing the companies value, being an interesting method for various industries.

Decreasing the customer churn is often approached by businesses through the use of predictive churn models. OHRA currently has multiple churn models in place for distinct product, where OHRA utilizes these models for two distinct objectives, respectively gaining insights and calculation of the customer value. Using the churn models OHRA can deduce which customers are most likely to leave the company and subsequently act upon this information or use this information for other models. Furthermore, OHRA also uses the information for the calculation of the customer value. This customer value is defined by OHRA as the expected margin obtained from the customer calculated over the expected maturity of a customer. Based on the fact that the industry in which OHRA operates is a highly competitive market where the operating costs should be minimized, OHRA is constantly attempting to improve their predictive churn models. A strategy identified by OHRA that could benefit from improvements are *trigger models*, a model that identifies the probability of a customer churning in the coming month. This model is mainly used for marketing purposes, where a customer likely to churn will receive a marketing expression with the objective of retaining the customer.

Following the urge to improve their predictive churn models, OHRA is searching for techniques or methods to increase the predictive performance of their customer churn models. This search of finding predictive models with a higher predictive performance is also observed within the scientific field, where research has focused for many years on utilizing new techniques to increase the predictive performance of predictive churn models (Alboukaey et al., 2020; Coussement & Van den Poel, 2008; De Caigny, Coussement, De Bock, & Lessmann, 2020; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). OHRA identified possibilities within the area of customer behaviour, while customer behaviour and corresponding changes in behaviour can possibly be a good indicator of the probability of a customer leaving (Chen, Fan, & Sun, 2012; Y. Zhang, Bradlow, & Small, 2015). Researching how customer behaviour and the changes in customer behaviour can influence the performance of predictive churn models within an insurance industry context is therefore considered to be an important research area to remain competitive by OHRA.

In conclusion, it is identified that the competitive and saturated environment in which insurance companies like OHRA operate, asks for measures to further reduce the operating costs. Although it has been concluded that the utilization of predictive churn models can help in this reduction, a higher predictive performance could increase the competitiveness of the company and therefore being of high importance. This led to the following problem statement:

**Problem statement**
The competitive environment of non-life insurance companies demands for increasing competitiveness through the use of methods like predictive churn models to increase customer retention. Identifying methods to enhance predictive churn models is required to remain competitive.

## 1.2 Research Objective and Questions

Based on the problem statement, the main research objective and the research questions can be formulated. These will be used to direct the research and ensure that the required results are met.

**Research Objective**
Understand how existing churn models and methods enable the prediction of churn, explore how the predictive performance of a predictive churn model can be improved using customer behaviour, and investigate the effect of class imbalance on the predictive performance.

Following the main research objective, research questions are derived from this objective resulting in one main research question and various sub-research questions. The results of answering these research questions will enable the insurance company to improve their predictive churn models, but also contributes to the scientific research area of customer churn and the predictive modelling of customer churn.

**Main research question**
How does the inclusion of customer behaviour influence the predictive modelling of customer churn within the insurance industry?

**Sub-research question 1**
Which techniques can be used to predict the probability of a customer churning?

**Sub-research question 2**
How can customer behaviour be used within the context of predictive churn modelling?

**Sub-research question 3**
How do different predictive churn modelling techniques compare?

**Sub-research question 4**
How do predictive churn models that utilize customer behaviour compare to models that do not utilize customer behaviour?

**Sub-research question 5**
What is the influence of class distribution and the handling of this class distribution on predictive churn models?

## 1.3 Research Design

This study is approached through the use of the CRISP-DM framework, a framework that is often used within projects related to data mining, data science, and data analytics (Wirth & Hipp, 2000). This framework for carrying out data projects does not depend on the industry sector or the technologies used, causing the framework to be highly suitable for this project. According to the CRISP-DM framework there are six distinct phases in a data mining project, respectively *Business Understanding*, *Data Understanding*, *Data Preparation*. *Modeling*, *Evaluation*, and *Deployment*.

The objectives and what the company requires as needed for the phase of Business Understanding have already been defined in the previous sections. Besides understanding the business, a formal understanding of customer churn, predictive churn modelling, and the techniques involved in predictive churn modelling is required. Chapter 2 describes the background of this research, covering these various topics and subsequently providing a formal understanding of the research context. Chapter 3 extends this by exploring research that covered similar topics, analyzing whether the results of these studies can be utilized within this specific study. By providing the background of the research and describing the related work, information regarding the first sub-research question is given.

Following the definition of the context and the related work that has been performed, the two phases of understanding the data used for the project and the preparation of data are described in Chapter 4. This chapter provides a detailed description of the dataset that will be utilized, while also providing an exploration of the data. After describing the data, an overview of how the data has been prepared is given. Building upon the related work, it is also analyzed how customer behaviour can be extracted from this specific dataset.

Chapter 5 defines the methodology used for the modelling, providing an overview of the methods utilized to obtain the results and corresponding answers to research question. This chapter covers the modelling phase of the CRISP-DM framework, providing an extensive description on how customer behaviour is modelled within this research. The chapter describing the methodology is followed by Chapter 6, the chapter that describes the results of the previously defined modelling approaches. It provides the foundation for the answering of the last three sub-research questions, which will be extended on in the discussion as depicted in Chapter 7. The discussion will entail the answering of all the various sub-research questions, the answering of the main research question, and other remarks regarding the research.

After the evaluation phase, a conclusion is provided in Chapter 8. This chapter concludes the research project and summarizes the work performed, providing an overview of the work that has been done and most important results.A complete overview of the various chapters, the expected results per chapter and the corresponding phase is provided in Figure 1.1.

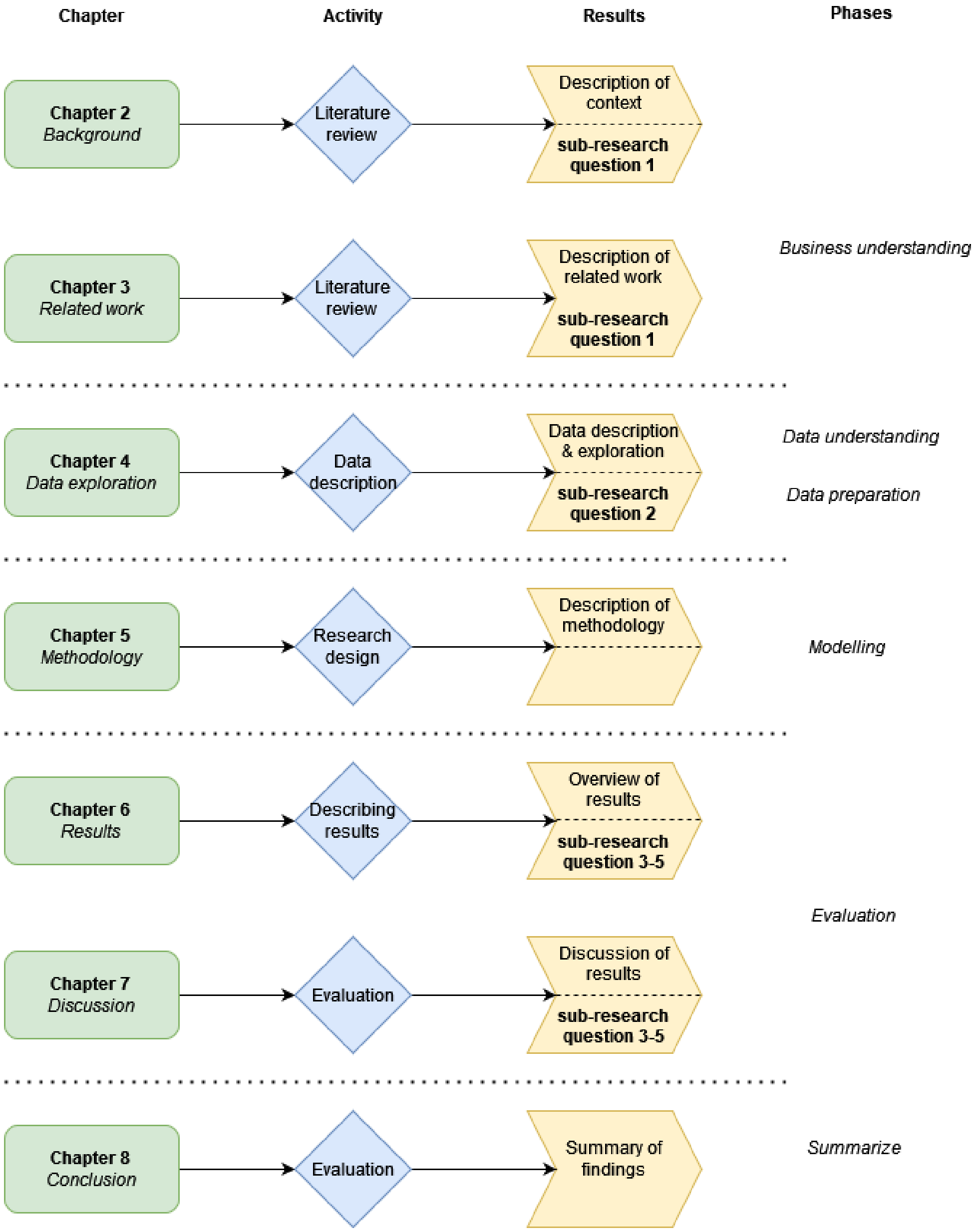

Figure 1.1: Overview of research structure

# Chapter 2

# Background

In order to obtain a better understanding of this research and the contributions of this research, certain aspects and techniques require an introduction. Various aspects are application-specific as in mainly used within the research field of customer churn and predictive churn modelling, while others are more broadly used. Providing a description of the aspects is the main objective of this chapter.

## 2.1 Customer behaviour

Understanding how a consumer or customer behaves is crucial for all businesses, government agencies, and other organizations. Investigating and analyzing the behaviour of the consumer helps in directing the organization, but also in specific business processes like marketing. Consumer behaviour and patterns observed within this behaviour has been subject to many studies in various domains, where this section attempts to obtain a better understanding of consumer behaviour and the measurement of this behaviour within predictive churn models.

### 2.1.1 Definition of Consumer Behaviour

Consumer behaviour is defined by Hoyer, MacInnis, and Pieters (2012) as "Consumer behavior reflects the totality of consumers' decisions with respect to the acquisition, consumption, and disposition of goods, services, activities, experiences, people, and ideas by (human) decision-making units (over time)"(p. 3). Hence, consumer behaviour can be explained as all the decisions made by a decision-maker regarding the consumption of an organization's offering. Kardes, Cronley, and Cline (2014) extends this by defining consumer behaviour as "all consumer activities associated with the purchase, use, and disposal of goods and services, including the consumer's emotional, mental and behavioral responses that precede, determine, or follow these activities" (p. 8). These activities carried out by the consumer affects the consumer, organization and the relationship between the consumer and organization. In addition to this effect, an organization can also influence the behaviour and activities. Understanding the activities and implications of these activities can be of high value for an organization.

Both Hoyer et al. (2012) and Kardes et al. (2014) distinguish consumer behaviour to have three types of activities, respectively *acquisition/purchase*, *use*, and *disposal* activities. This distinction is made while the consumers' response to certain incentives depends on the type of activity, causing the response to a specific incentive differ per activity type. Therefore, categorizing activities is important for grasping the full context of consumers' behaviour. The first category, *purchase activities*, relatea to activities through which a customer acquires a certain good or service. Purchase or acquisition activities entail for example the evaluation of information about the product before buying, choosing where to buy the product and also the method of purchase. After buying a service or product, a customer will start to use the offering that has been bought. Subsequently, the *use activities* will start, which entail all the activities that describe where, when and how the consumption takes place. Particular examples for said activities are the start time of consumption, reasoning for consumption, and how the product is consumed. Lastly, *disposal activities* are involved which are the activities that describe how the customer dispose of the product or service offered by the organization. Disposal activities entail for example the discarding of a product, recycling activities or the resale activities that occur with the product.

In addition to having specific activities within customer behaviour, a central aspect of customer behaviour is also the consumer's mental, behavioural, and emotional responses to the goods or services offered by the organization and corresponding marketing (Kardes et al., 2014). Responses of the customer depends on the activities described earlier, while responses by themselves also elaborate on consumer behaviour. Consumer responses are determined to have one of three specific categories, being an *emotional*, *mental* or *behavioural* response. An *emotional* response is a response that reflects the emotions and feelings of a consumer, for example the emotions that a consumers feels after using the offering of an organization. Although being an important response, measuring the emotions of a customer without explicitly asking can be quite difficult. The second category of response are the *mental* responses, which include the thought processes, opinions, attitudes, and beliefs of the customer. Just like emotional responses, these responses can be relatively hard to measure for an organization. Lastly, *behavioural* responses are defined to be decision and actions during the various stages of activity. In contrast to the other two responses, actions that are done in the influence sphere of an organization can be relatively easy measured. Examples of such actions are reading information on the organization's website before purchasing a certain offering of the organization, testing specific components of a product, or asking questions to the organization. In summary, consumer behaviour can be defined as all activities carried out by the consumer and the responses that are related to these activities. Consumer behaviour discloses information about the attitude and stance of a consumer while also describing the personality of a customer, causing the customer behaviour to be of high value for an organization or business.

### 2.1.2 Importance of Consumer Behaviour

Studying consumer behaviour and factors influencing the consumer behaviour are of great importance for an organization due to a wide range of reasons. Studying consumer behaviour and utilizing the results can significantly improve the performance of a business through the creation of better products, more effective promotion of their offerings, and introducing marketing strategies that enhance the sustainable competitive advantage (Kardes et al., 2014). The insights of studying consumer behaviour are mainly used within elements of marketing, which is the process of providing value to customers (Hoyer et al., 2012). According to Schiffman and Wisenblit (2019) the core principle of marketing is to satisfy consumer needs in an effective way by only providing products that the consumers are most likely to buy. Studying the consumer behaviour enables an organization to identify the consumer needs, causing the organization to focus on the products that the consumer really want and subsequently enhancing the organization's performance. Secondly, understanding the behaviour of the consumers enable an organization to better segment the market and enhance the targeting for marketing strategies. Having an understanding of the customer behaviour enables marketeers to target the appropriate customer groups for their marketing tactic, increasing the efficiency of their marketing strategies.

Besides having a general importance for an organization or business, understanding the consumer behaviour also benefits the development of better customer churn prediction models.

The inclusion of behavioural aspects within business analytics is an emerging research field also known as the computational social science field, where research considers the everyday activities of customers to be of high importance for describing the conscious behaviours and decisions of customers (Kaya et al., 2018). The everyday activities of a customer entails a significant amount of information on the behaviour of the customer, whereas socio-demographic characteristics describes the customer but does not elaborate on the behaviour of a customer. The behaviour of a customer is heavily dependent on the context of the product or service that the customer utilizes. A customer can hold multiple products of which some products are more likely to be churned than others. Subsequently, customer behaviour differs per contract, causing the everyday activities of customers to be more important than customer demographics which do not capture the differences between various contracts (Wei & Chiu, 2002).

Using customer behaviour as input for a predictive customer churn model is also important while customer behaviour information is easily available and reliable (Wei & Chiu, 2002; Y. Zhang, Liang, Li, Zheng, & Berry, 2011). Chen et al. (2012) state that behavioural data is often separately stored within a transactional database, where each activity is treated as a data instance. Furthermore, logging activities of a customer is less prone to missing data than customer demographics. Logging which activity a consumer performs in the organization's environment is based on the activity itself, an activity entailing interaction with any of the platforms of the organization. This in contrast to the data with customer demographics which is usually provided by the customer, leading to more missing data or incorrect data. Hence, behavioural data is seen as a solid source of information relevant for the prediction of customer churn.

Due to the information providing information and being more reliable and accessible, recent studies show that including consumer behaviour features besides including customer demographics in customer churn prediction increases the predictive performance (Alboukaey et al., 2020; Ali & Arıtürk, 2014; Hung, Yen, & Wang, 2006). Furthermore, the research of Wei and Chiu (2002) has shown that predicting customer churn by solely using customer behavioural features is possible with satisfactory results. Adding customer behavioural information in a predictive customer churn model also enables the model to dynamically predict the probability of a customer churning (Chen et al., 2012). Data containing customer demographics like age, place of residence, and gender are fairly static, being data that does not change frequently and therefore often have the same implication for the probability of the customer churning over time. This is a different matter for data regarding customer activities, which differs over time and subsequently imposes different implications per time frame. By adding information regarding the transaction and customer-organization interaction data which represents the behavioural activities of customers, an organization can extend their customer churn predictive model with a reliable data source. Implementing customer behavioural activities and patterns within a customer churn predictive model is therefore becoming increasingly popular within the field of churn prediction.

## 2.2 Customer churn and its context

### 2.2.1 Customer Churn

One of the main goals of businesses and enterprises, profit or non-profit, is to provide value to certain customers or users. Growth of an organization is often linked to these customers while adding more customers or users will enable the growth of a company. This linkage also works the other way around, while customers that are leaving will lead to a decline within the company. Consequently, organizations are interested in not only obtaining new customers but also retaining existing customers or users. Investigating methods to minimize the amount of customers leaving, also known as churn, is therefore an important topic both for industries and scientific research.

Originally, churn was used in the context of the farming industry where it both existed as a plural or verb. Churn was used to describe a container in which cream is shaken or stirred to make butter, but is also used to describe the violently stirring of a substance (Merriam-Webster, n.d.). Following the use in a farming context, the concept of churn is nowadays also used in a general business context where it is defined by Gold (2020) as the moment when a customer or user decides to quit using a service or cancels their subscription. Besides being used to describe the formal termination of a service or a subscription, customer churn is also defined as the movement of a customer from one company to another company (Berson & Thearling, 1999). When looking at different domains in the business context however, differences in the definition of churn are observed. Lazarov and Capota (2007) define three types of customer churn: *active or deliberate* means that the customer has decided to quit and switch to a different provider, *rotational or incidental* is the quitting of the customer without the aim of switching to a different provider, and lastly *passive or non-voluntary* where the company itself discontinues the contract. Besides dividing based on the motivation of customer churn, churn can also be divided based on how the customer cancels their agreement. *Total* is defined as the official cancellation of the agreement, *hidden* is defined to be a situation where the contract is not officially cancelled but the customer is not actively using the service since a longer period of time. Lastly, *partial* churn is a situation without official discontinuation but where the customer uses only parts of the service and utilizes the service of a competitor. Depending on the industry or company, a specific definition of customer churn is used.

Differences in the specific definition of customer churn is seen throughout the literature, while services that are subscription based like insurance companies (Günther et al., 2014) and the telecom service industry (Hung et al., 2006) define churn through a total cancellation of agreement. Industries like the social gaming industry denote customer churn as a customer not returning to the game within a predetermined amount of days (Milošević, Živić, & Andjelković, 2017). This introduces uncertainty in terms of the period of inactivity while the period of inactivity is up to debate, but also introduces uncertainty through the definition of inactivity. Handling of these uncertainties depends on the context and requirements of the company.

### 2.2.2 Customer Retention

Companies are highly invested in customer churn and focus their marketing efforts often on retaining their customers while the attraction of a new customer is significantly less expensive that acquiring a new customer. The costs of acquiring a new customer are estimated to be 3-12 times higher than the costs of retaining an existing customer (Jamalian & Foukerdi, 2018; Torkzadeh et al., 2006). Customers that are longer at the business are more valuable for the company and losing customers cause a rise in opportunity costs (Verbeke, Martens, Mues, & Baesens, 2011). A second motivation for businesses in retaining their customers is that the effectiveness of marketing regarding retaining customers is significantly more effective than marketing towards acquiring new customers, therefore enhancing the overall cost effectiveness of a company (Mozer et al., 2000). Consequently, the process of retaining customers and increasing customer retention rates is seen as a key aspect of business operations. This process is defined by Smith et al. (2000) as a complex combination of various issues, like pricing, service, personal preferences, and many other. The complexity of customer retention is reinforced by the fact that personal preferences involve customer behaviour, which is often not easily explained.

Customer retention utilizes the loyalty of a customer to the company, where the loyalty of a customer is seen as a prerequisite for retaining a customer (Larsson & Broström, 2019). The loyalty of a customer determines the chance of a customer to continue making transactions at the company, where migration to another company is becoming more likely when loyalty among customers is low. Customer retention focuses on the aspects that a company can influence like the segmentation of customers based on their loyalty profile, in which these aspects all attempt to evaluate and potentially increase the loyalty among customers. The research of Coyles and Gokey (2005) has shown that customer loyalty is influenced by three underlying forces, respectively the attitude of a customer, the needs that a customer experiences, and the satisfaction of a customer towards the company. These underlying forces significantly influence the loyalty and subsequently can be used in segmentation of customers. Aspinall, Nancarrow, and Stone (2001) argue that customer retention has a multifaceted nature, where customer retention is driven by behavioural and attitudinal reasoning or driven by a more practical reasoning in which the customer simply is not able to migrate to another company. Distinguishing these different types of reasoning is important for the understanding of why a customer could possibly churn, indicating the importance of this research field for customer churn.

Business often attempt to enhance the retaining of customers through explicitly defined strategies. These strategies can focus on a whole range of factors which all influence the customer retention. Hawkins and Hoon (2019) analyzed the strategy of maintaining a positive customer-employee relationship, concluding that focusing on this positive relationship could increase the customer's trust and increasing the repurchase rate in a small business context. Thaichon and Quach (2015) analyzed how the use of an integrated marketing communication strategy could increase the customer retention rate, while adequately managing the customer's expectations through integrated marketing influences the customers' satisfaction and trust. By influencing these factors a business can ultimately enhance the customer retention. Besides attempting to influence the customer retention rate through the managing of expectations and building relationships, a business can also approach customer retention through specific targeting of customers that are likely to leave. This specific targeting starts by detecting which customers are most likely to leave the company and focusing on these customers with marketing efforts or price discrimination (Ascarza, 2018).

### 2.2.3 Customer Relationship Management

The concept of customer retention, corresponding underlying factors, and the business approaches involved with customer retention is considered to be one of the most important topics (Aspinall et al., 2001), along with being one of the most critical challenges in Customer Relationship Management (CRM) (Miguéis, Van den Poel, Camanho, & e Cunha, 2012). Customer Relationship Management is a research field which has been profoundly researched in the last 20 years and has been extended on by not only researchers, but also by companies and other stakeholders. CRM is defined by Kumar and Reinartz (2018) as a concept in which the analysis and use of marketing databases in combination with the leverage of communication technologies is adopted to determine corporate policies and methods utilized for the increasing of customer value. In general, CRM can be seen as a strategic process with the goal of optimizing the customer value for a company. Guerola-Navarro, Gil-Gomez, Oltra-Badenes, and Sendra-García (2021) define CRM as a fundamental tool for businesses which aid the consistent management of customer information by identifying the most valuable clients, attracting them as clients with high trust, retain those clients by introducing loyalty policies and subsequently developing a lasting partnership between the customer and the business. Therefore, four main dimensions of CRM can be identified, respectively customer identification, customer attraction, customer retention, and customer development (Ngai, 2005). Customer identification entails the actions coordinated and directed by the company in order to analyze the potential customers of the company which is often performed through customer analysis and segmentation of customers. Following the identification phase, the customer attraction dimension focuses on the steps that a business can undertake to attract the interest of customers. Thirdly, the customer retention dimension has already been described as one of the most important topics of CRM. It entails any strategy aimed at building more customer loyalty and establishing valuable relationships with customers. Lastly, the customer development dimension mainly entails expanding commercial relationships with existing customers. By utilizing the different dimensions of CRM, a business can systematically approach the relationship management of customers in order to enhance their customer value.

Due to the rapid growth of digital systems, involved information technologies and subsequently the generation of customer data, current CRM systems and approaches are highly digital focused. It is seen that customer churn and the prediction of customer churn is becoming an important feature within CRM systems for a large share of industries like the telecommunication industry (Lu, Lin, Lu, & Zhang, 2012) and the insurance industry (Günther et al., 2014). Customer churn analyses are used by these industries within the dimension of customer retention, where customer churn can be utilized in the process of retaining customers by enabling more specific targeting of customer that are most likely to churn or using the analyses to focus their marketing efforts. Within the context of customer retention and customer churn, Gold (2020) identify two main uses of customer churn analyses that are potentially beneficial for a company, customer segmentation which can be based on the results of the churn analysis and implementing measures to reduce the customer churn. By improving the existing product, introducing marketing efforts like engagement campaigns and customer interactions, and adjusting the pricing, a business can attempt to reduce the customer churn.

Through the implementation of measures to reduce the customer churn, businesses aim to optimize the value of customer. Customer Relationship Management involves the notion of Customer Lifetime Value (CLV), referring to the present value of the net benefit to the firm from the customer over time (Borle, Singh, & Jain, 2008). Generally, the net benefit can be seen as the revenues gained through the customer minus the costs that are necessary for maintaining the relationship between the business and the customer. CLV is seen as an important metric to evaluate the marketing decisions that are taken with CRM, in which customer churn significantly influence the metric through an indication of the lifetime of a customer. Therefore, customer churn is seen as highly important by businesses not only for their customer retention efforts but also for their CRM strategies and analyses.

The dynamics of customer churn depends on the nature of the industry that the business is operating in, where the lifetime of a customer is highly dependent on the product that a business offers. A customer that is buying a car acts differently and has a different lifetime than a customer looking for a new insurance that is often yearly renewed. Although differences are observed between these customers, managing their customers and customer churn is of high importance for both industries. Devriendt, Berrevoets, and Verbeke (2021) and Verbeke et al. (2012) argue that marketing efforts should be focused on customers that are most likely to churn, increasing the cost-efficiency and a better allocation of the marketing resources. Furthermore, targeting customers with a relatively high CLV enables businesses to increase their share of long-term customers, which generate higher profits for the company. Based on this importance, businesses often utilize customer churn prediction models. These models attempt to predict which customers have the highest propensity to attrite, where customer churn is often depicted as a binary variable dividing the customers in non-churners and churners. Various techniques for predictive churn modelling exist (Vafeiadis et al., 2015), all aiming to identify early signals of potential churn and recognizing the customers with an increased likelihood to voluntarily leave the company. It is argued that for the customer retention dimension of CRM, the main use of customer churn is within the predictive churn modelling while aiming and targeting for customers that already churned is not cost-effective and often overdue. Accurately predicting customer churn is seen as a key aspect of the relationship between customer churn and customer retention, where maximizing the predictive power can provide significant benefits for a business.

In conclusion, churn is nowadays mainly used in a business context as customer churn, the moment when a customer decides to quit using a service or cancel the subscription. Although customer churn is dependent on the context of the company and differences are observed between different industries, similarity is seen within the companies' stance towards customer churn. Managing customer churn is seen to be crucial, delivering value to the company and increasing the cost effectiveness. Therefore, customer churn is identified to be an important factor within the customer retention strategies and subsequently the area of Customer Relationship Management. Predicting customer churn can enable more cost-efficient marketing strategies and enable a more accurate calculation of the Customer Lifetime Value, a metric often used to assess their marketing efforts. Generally, customer churn and the prediction of customer churn is seen as beneficial for any business, which is also observed within the increased attention of both the scientific research and various industries.

## 2.3 Churn Modelling

The challenging problems of customer churn and the need of the industry to temper customer switching behaviour, especially for the subscription-based service industry, led to a growing interest in accurately predicting and modelling the customers that are most likely to leave the company (Coussement & Van den Poel, 2008). Customer churn modelling in a business context not only entails the systematic prediction of the churn likelihood, but businesses also like to explore the customer's motivation to churn. Exploring the causes of customer churn behaviour enables a company to substantiate their profiling of churning customers, as well as enabling more efficient marketing campaigns (Geiler, Affeldt, & Nadif, 2022). Therefore, the research area of customer churn modelling is a research area where a trade-off is observed between accuracy and interpretability. Any business prefers a highly accurate customer churn prediction model, but also requires an interpretable model in order to explore the motivations behind customer churn. However, models that can be seen as relatively accurate in customer churn prediction like the models of Alboukaey et al. (2020) and Ali and Arıtürk (2014), are less interpretable than predictive models based on methods like ensemble techniques and logistic regression. This trade-off causes the selection of features for the predictive models and techniques used to predict customer churn to be crucial.

Customer churn prediction is an application within the field of business analytics, where techniques and approaches are used to analyze data. These data analyses are used for the evidence-based decision making, aiming for an increase in efficiency, efficacy, and subsequently profitability (Devriendt et al., 2021). This is also seen within the customer churn prediction (CCP), where customer retention efforts are supported with a CCP model. Customer churn prediction models are known as classification models, which estimate the probability of a customer churning during a specific period of time. Churn prediction can be modeled as a binairy classification task, where it attempts to model the conditional probability of a customer churning. Formally, this can be depicted as $P(Y = y_i|x_i)$ also known as the *class posterior*. The prediction of customer churn follows a general pipeline as depicted in Figure 2.1, an adjusted version of the pipeline defined by Geiler et al. (2022). This pipeline consists of three distinct phases, respectively sampling, model fitting, and evaluation. As can be seen in the figure, the three phases utilize different techniques, as well as various techniques that are used within the different phases. In order to understand what customer churn prediction modelling entails, a full understanding of the different phases and the various techniques involved is needed.

### 2.3.1 Data and corresponding features

Customer churn prediction is done on churn datasets, datasets generated by a company containing data of customers that both have churned and remained within the company. This dataset often consists of various types of data and can be used to describe the customer to a great extent. Most churn datasets describe the customer through demographic data, subscription/product data, and data regarding customer-company interactions (Hung et al., 2006). Demographic data is used to describe the customer through demographics like gender and age, involving fundamental information of customers. The dataset extends this fundamental information by including data regarding the subscription or product that a customer has with the company, providing information on the cost and content of the product and payment information. Lastly, the third category of data mainly entails information about the contact points with the customer. This can be in the form of usage of the mobile application and website of the company, usage of the actual product, and also service contact information.

Churn datasets are often large sets with a large share of high-dimensional data and using this data for the pipeline is done through feature selection and feature extraction. Feature selection and extraction enables the reduction of data dimensionality, improving the performance, computational efficiency and decreasing the memory storage requirements (Li et al., 2017). Feature selection is the direct selection of a subset of relevant features, whereas feature extraction is the projection of original high-dimensional features to a low-dimensional feature space. Both feature selection and feature extraction are crucial processes for churn modelling and can entail manual selection of features, selection based on expert knowledge or research, or feature selection and extraction based on extraction. Within churn modelling, the features mainly consists of the categories also seen in the general dataset like demographic features, features decribing the interactions between customer and company, and product related features.

### 2.3.2 Sampling Phase

The first phase of the customer churn prediction pipeline is the sampling phase, the phase that entails the selection of a subset from within the original churn dataset. As mentioned before, this dataset contains data regarding customers of both costumers that have left the business and customers that still remain within the company. Sampling is done to make the data more accessible for processing and to enable further analyses. Inherent to using a churn dataset is that the dataset is unbalanced while customer churn is an infrequent event. Although happening quite often, customer churn is in most cases only observed within a relatively small number of customers. This class

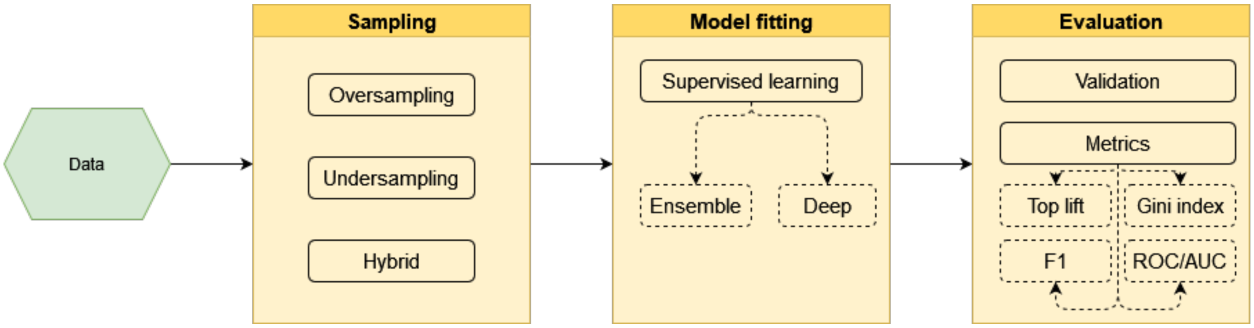

Figure 2.1: General pipeline of customer churn prediction

imbalance interferes with the objective of accurately predicting customer churn for all customers, causing the main objective of the sampling phase to be the obtaining of a similar dataset that is better balanced in terms of churning and non-churning customers (Geiler et al., 2022). In order to achieve this objective *undersampling*, *oversampling*, or a combination of both approaches (*hybrid*) can be used to achieve a better balanced dataset (Burez & Van den Poel, 2009). Achieving a better balanced dataset can be beneficial for the prediction accuracy of the models, but research has shown that this will not always be the case. Exploring the various approaches and the potential advantages of these approaches can aid in potentially improving the prediction models.

The first approach, oversampling, is a technique that normally consists of duplicating instances in the minority class or synthesizing new instances from the available data (Geiler et al., 2022). The most straightforward approach to oversampling is the approach of random oversampling, where randomly data points are selected to be replicated. This simplistic approach is often substituted by more complicated techniques like the *Synthetic Minority Oversampling Technique* (SMOTE) and the *Adaptative Synthetic Sampling* (ADASYN) approach. Both techniques focus on generating data instances for the minority class, where SMOTE uses a $k$-nearest neighbours approach. The feature vector of a minority class observation is taken after which the difference is calculated between the original feature vector and the feature vector of the neighbour. This difference is then multiplied by a random number between 0 and 1, after which the multiplied difference is added to the original feature vector, causing a new observation to exist (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). ADASYN generates new data instances based on their distributions, using a density distribution as criterion (H. He, Bai, Garcia, & Li, 2008). According to Burez and Van den Poel (2009) this generation of new data instances can significantly increase training time, but can also lead to a situation in which the prediction model is overfitted due to the generation of exact copies of existing data instances. Due to these reasons, undersampling is often preferred over oversampling.

Undersampling is a second approach that can be utilized for the subject of class imbalance, and is in its simplest form the elimination of majority-class examples in order to increase the balance within the sample, with the downside of potentially discarding useful data instances. The research of Zhu, Baesens, and van den Broucke (2017) also identified more advanced techniques based on undersampling like the *one-sided selection* (OSS) technique and the *cluster-based undersampling algorithm* (CLUS). The OSS technique selectively removes the data instances from the majority class which are redundant, while the CLUS approach organizes the training data into homogeneous groups with similar characteristics and subsequently downsizes the number of samples in each cluster. Besides undersampling and oversampling, a class imbalance problem can also be solved through a hybrid strategy utilizing a hybrid sampling technique. There are various techniques, ranging from a combination of oversampling with data cleaning to applying computational intelligence to identify useful instances. These techniques are used to counter some of the disadvantages of using one specific sampling strategy, but do not fully eliminate the disadvantages of the methods used in the combination for example (Geiler et al., 2022). Approaching the sampling through a data-level solution enables the potential creation of more accurate customer churn prediction models. The disadvantages of the described techniques can however have a significant impact on this accuracy. In order to enable the full potential of tackling the class imbalance, the method of handling this class imbalance should be carefully chosen.

### 2.3.3 Model Fitting Phase

The second phase in the general pipeline of customer churn prediction is the model fitting phase, the phase in which data mining techniques are utilized for the customer churn prediction task.

According to Geiler et al. (2022) the most widespread techniques can be divided in supervised learning techniques and semi-supervised learning techniques. When looking at the research of KhakAbi, Gholamian, and Namvar (2010) and Eria and Marikannan (2018), the five most frequently used data mining techniques as seen in Table 2.1 for customer churn prediction are *Neural Networks*, *Decision tree*, *Logistic Regression*, *Support Vector Machine*, and *Naive Bayes*. Therefore, these techniques will be explored during this section to gain a complete understanding of the data mining techniques useful for approaching the customer churn prediction task. As seen in the general pipeline, ensemble methods can also be used for the model fitting phase. Therefore, ensemble methods are also explored.

Table 2.1: Overview of data mining techniques per literature review

| **Data Mining Technique** | **Literature Review by KhakAbi et al. (2010)** | **Literature Review by Eria and Marikannan (2018)** |
|---|---|---|
| *Neural Networks* | 15 | 7 |
| *Decision Tree* | 13 | 5 |
| *Logistic Regression* | 13 | 4 |
| *Support Vector Machine* | 7 | 7 |
| *Naive Bayes* | 3 | 5 |

### Neural Networks

Neural Networks (NNs) are a form of supervised learning, also known as deep learning, and represent a broader family of data mining methods that are inspired by biological neural networks like the human brain. Neural Networks convert the data into an algorithm that mimics the brain neuron system, often leading to significantly higher results (Geiler et al., 2022) in terms of accuracy and efficiency, but also leading to more complexity and less interpretability (Eria & Marikannan, 2018). This complexity can be seen in the use of hidden layers as depicted in Figure 2.2. The main difference between deep learning algorithms and the traditional machine learning algorithms is the aspect of feature engineering, where deep learning algorithms automatically performs feature engineering. This is not the case for traditional machine learning algorithms, where domain knowledge is required for the feature engineering aspect. It is however not always preferred within the context of churn prediction, while using deep learning algorithms like NNs reduces the interpretability which is often important for the companies implementing the customer churn prediction models.

### Decision Tree

A second data mining technique that is often seen within the context of customer churn prediction modelling is *Decision Tree*, a method relying on a symbolic learning technique. Extracted information from a dataset is organized in a hierarchical structure, composed of nodes and branches (Nie, Rowe, Zhang, Tian, & Shi, 2011). Within a binary decision tree, every non-terminal node represents a decision made by the approach. Based on the decision made at the node, the approach continues with the left or right branch where a node is at the end to depict the outcome. The class labels are represented by the tree leaf nodes, whereas the conjunctions leading to these nodes are represented as tree branches (Eria & Marikannan, 2018). Tree-based methods is a highly interpretable technique which supports logical human thinking, where the organisation in the form of a tree facilitates interpretability of the model. This high interpretability causes the Decision Tree approach to be a frequently used data mining technique within customer churn prediction.

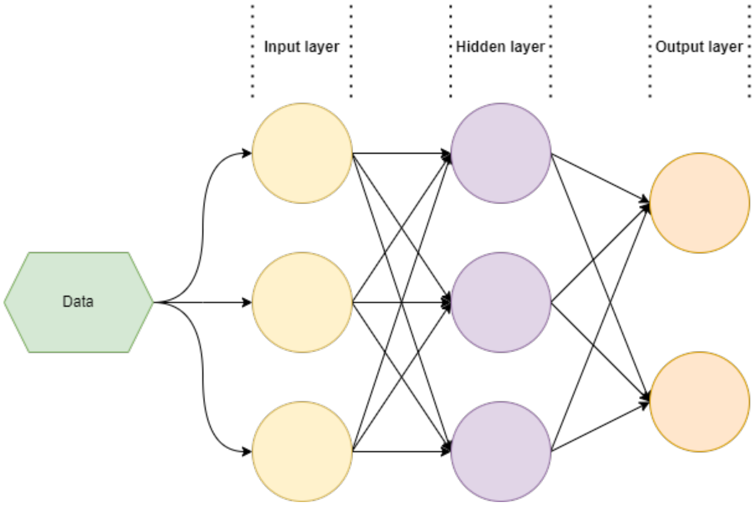

Figure 2.2: Structure of a Neural Network with one hidden layer

### Logistic Regression

*Logistic Regression* is one of the simplest techniques in terms of supervised learning and is still frequently used while the model is considered to be conceptually simple and shows robust results in general studies (Burez & Van den Poel, 2009). It is a form of regression analysis where the dependent variable is binary and utilizes a specific form as seen in Equation 2.1 and the logit transformation seen in Equation 2.2 (Nie et al., 2011). Logistic regression is a highly popular in the customer churn prediction context due to its predictive performance in combination with its good comprehensibility (De Caigny et al., 2020).

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2.1}$$

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{2.2}$$

### Support Vector Machine

*Support Vector Machine* (SVM) is a supervised learning model used for both the classification and regression task. SVM attempts to construct a hyperplane that optimally separates two linearly separable classes, where the classes in a customer churn prediction context are non-churning and churning customers. SVM aims to minimize the distance between hyperplanes and the classes to separate the classes as much as possible. According to the research of Eria and Marikannan (2018) SVM is frequently used in the customer churn prediction context, but the technique does require additional parameter tuning and performs significantly less when the dataset is too skewed.

### Naive Bayes

*Naive Bayes* or a Bayes classifier applies Bayes' theorem, which in the context of customer churn prediction attempts to describe the probability that a customer will churn based on conditions that might be related to that event. It can be described as an independent feature model (Vafeiadis et al., 2015), a classifier that is assuming that the presence of customer churn is unrelated to the presence of any other feature. Therefore, given a set of independent variables $(X_1, ..., X_m)$ and a dependent categorical variable (churn) (Y), the probability of a customer churning is calculated using Equation 2.3. In the field of churn prediction, Naive Bayes classifiers have achieved relatively good results, but has the downside of being highly computational intensive.

$$P\left(\frac{Y}{(X_1, ..., X_m)}\right) = \frac{P\left({}^{(X_1, ..., X_m)}/_{Y}\right) P(Y)}{P(X_1, ..., X_m)} \tag{2.3}$$

**Ensemble methods**
Ensemble methods are algorithms that create a combination or set of classifiers, utilized to classify data points by using a type of weighted voting (Dietterich, 2000). There are two distinct types of ensemble methods, respectively *bagging* and *boosting* (Opitz & Maclin, 1999). Bagging takes random values specifically from the original dataset, even taking a specific value multiple times, and create multiple classifiers. This aspect is extended by boosting methods, which assigns weight to the various data points that are related to the error rates, causing the wrongly predicted values to be more present in the next weak learner.

A bagging method that is commonly seen is the *Random Forest* approach, a method that utilizes multiple classification trees to classify a new object from a specific input vector. Random Forest is an ensemble method also known as a meta-algorithm, combining several models into one predictive model to decrease the variance of the model (Geiler et al., 2022). This method blends elements of random subspaces and bagging, where at each node a subset of the selected features is randomly chosen and subsequently selecting the best split available across those randomly chosen features (Burez & Van den Poel, 2009). Random Forest generate the classifications or predictions by taking an average of the predictions from individual trees. The method is less affected by missing data than most other machine learning algorithms and is able to handle large datasets relatively well, causing the technique to be both an efficient and interpretable technique in the context of customer churn prediction.

As mentioned before, boosting algorithms are built sequentially by building succeeding learners based on the error rate of the previous classifier. The models use a gradient-descent based formulation, where the gradient is used to minimize a loss function (Natekin & Knoll, 2013). The gradient can be utilized to find the direction in which the model parameters should be changed to optimally reduce the error in the next round. There are various boosting algorithms, with two of the most popular variants being XGBoost and LightGBM. Boosting algorithms often report high performance, having a high generalization accuracy or fast training speed (Bentéjac, Csörgő, & Martínez-Muñoz, 2021).

Overall, there is no real consensus on which data mining method to use in the research field of customer churn prediction. There is the observed trade-off between interpretability and predictive performance, through which logistic regression was originally seen as one of the most feasible options based on its comprehensible and robust results (De Caigny et al., 2020). Other research have concluded that other algorithms are more accurate like the research of Mozer et al. (2000) and Vafeiadis et al. (2015). The observed differences in the field of customer churn prediction can be partly attributed to the nature of the studies performed within the research field. Many studies evaluate a limited amount of data mining techniques while performing the analysis on a single dataset, causing the results to vary among studies (Verbeke et al., 2012). Since there is no general consensus on which method to use within the prediction of customer churn, a study regarding customer churn prediction should focus on a range of techniques to ensure that the most accurate prediction result is achieved.

### 2.3.4 Evaluation Phase

After preprocessing and implementing the data mining technique, it is required to evaluate the predictive performance of the model(s), also known as the evaluation phase. This phase consists of model validation, estimation of model effectiveness for the prediction of unseen data instances, and the model evaluation which uses specific evaluation metric to assess the predictive models of the designed models. The model validation utilizes a subset of unseen data to evaluate the predictive performance and the strategy chosen for the model validation depends on the sample size (Verbeke et al., 2012). When dividing the full dataset in a subset used for training and a subset used for testing, also known as the *holdout* strategy, a portion of the data is lost which can be undesirable when having a relatively small dataset. Therefore, other validation strategies like the K-fold cross validation strategy are often used to counter this aspect. The strategy of K-fold cross validation is a strategy where the dataset is split in K subsets of similar size, after which the model is fitted on $K-1$ folds. The error of prediction is than calculated per $k^{th}$ unseen subset, after which the whole strategy is repeated $K$ times, where K is often 5 or 10 (Geiler et al., 2022). By combining all $K$ estimates, a more extensive cross-validation strategy is performed which counters some elements of overfitting and bias.

Another strategy that is often seen within the customer churn prediction is the *stratified K-fold cross-validation* strategy. This strategy addresses the imbalance within a dataset which is often seen in the domain of customer churn prediction. Stratified validation utilizes the target variable, in this case customer churn, to ensure that each fold within the K-fold cross-validation strategy contains the same distribution of class labels (Geiler et al., 2022). Noteworthy is that stratified cross-validation strategy can introduce sampling bias for certain machine learning techniques (Oommen, Baise, & Vogel, 2011). Stratified K-fold cross-validation therefore can aid in the countering of class imbalance against the cost of introducing sampling bias, causing the choice of cross-validation strategy to be of high impact for the predictive models.

Besides focusing on the validation strategy the prediction model also has to be evaluated, which is done through the use of evaluation metrics. The first metric, *top-decile lift*, is an evaluation metric with roots in the marketing domain, hence often used within the context of customer churn prediction (Geiler et al., 2022). Top-decile lift is a metric which considers the data instances in order of predicted probability, therefore in the context of churn modelling considering the customers in order of churn probability. Generally, the top-decile lift focuses on a specific probability like 10%, and calculates this subset of 10% by choosing the highest probability of churning. After calculating the 10% riskiest customers, the portion of churning customer within this set is compared against the whole portion of churning customers in the dataset. Consequently, this metric evaluates whether customers that are predicted to be potentially churners are actually at risk. This is especially important for practitioners in the marketing domain where budgets and resources are constrained, emphasizing the need for a focus strategy (Coussement & Van den Poel, 2008). A second metric closely related to the top-decile lift is the *Gini coefficient*, which also takes the customers less likely to churn into account. It indicates the effectiveness of the model in differentiating between customers that are likely to churn and the customers less likely to churn. The Gini coefficient is often used to compare models and evaluate the predictive power of a model.

The third metric, the $F_1$ *score*, summarizes two metrics from the confusion matrix, respectively *Precision* and *Recall*. This score is a more accurate metric while calculating accuracy for a predictive churn model within an unbalanced dataset like a churn dataset is not appropriate (Geiler et al., 2022). The $F_1$ has its origin in the confusion matrix as depicted in Table 2.2. In order to calculate the $F_1$ score, precision is calculated through the formula $Precision = \frac{TP}{TP+FP}$ and recall is calculated through $Recall = \frac{TP}{TP+FN}$. The $F_1$ score is subsequently calculated through $F_1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$, being more applicable to a churn context than using simply *Accuracy.*

The last metric is one of the most widely used evaluation metric in both a churn prediction context as well as other machine learning domains. This metric, the *ROC & AUC*, utilizes a two-

Table 2.2: Confusion matrix in the context of customer churn prediction

| | **Actual churn** | |
|---|---|---|
| **Churn prediction** | **0** | **1** |
| **0** | True Negative (TN) | False Positive (FP) |
| **1** | False Negative (FN) | True Positive (TP) |

dimensional graph known as the ROC curve that displays the True Positive Rate ($TPR = \frac{TP}{TP+FN}$) as a function of the False Positive Rate ($FPR = \frac{FP}{FP+TN}$). Following the ROC curve, the area under the curve is depicted through the AUC metric which provides an aggregated performance measure, interpreted as the probability that the model correctly classifies an instance as positive compared against the probability that an instance is correctly classified as negative.

# Chapter 3

# Related work

Predictive churn modelling is a research field with a considerable amount of studies, studying various aspects of the research field. Studies found within the field of predictive churn modelling often entails the comparison of multiple predictive churn models. Within the large share of studies that have been conducted in the the research field of customer churn prediction, the main topic that is seen is the evaluation of different algorithms along a industry-specific dataset, evaluating algorithms in the telecom, mobile gaming, or financial industry. Understanding how the models have been defined using the available information and how the various techniques relate to each other in several research domains is considered to be highly relevant for implementing and analyzing predictive churn models.

Related to studies within the domain of predictive churn modelling is the strategy of handling class imbalance, an aspect that is frequently observed within datasets used for predicting customer churn. Utilizing imbalanced datasets require specific strategies to enhance the predictive performance of the suggested models, enabling the model to learn as much as possible from the minority class. Multiple studies have focused on handling the class imbalance within the field of customer churn prediction, utilizing various strategies or introducing novel methods. Analyzing the effects of these studies and the implications can potentially be beneficial for the predictive performance of the models found within this research. Other studies have also focused on the inclusion of customer behaviour within predictive churn models, focusing on both static and dynamic methods of including behaviour and behavioural patterns. This study attempts to analyze the effects of including behavioural aspect, therefore requiring an understanding of the effects and results found in studies that explored the inclusion of customer behaviour.

This chapter attempts to analyze and summarize a wide range of studies found within the research field of customer churn prediction that have a similar research objective as this research. By introducing studies on the main topic of model comparison, the aspect of class imbalance and the inclusion of behavioural features within predictive churn models, a solid theoretical understanding of the research context is provided. Therefore, this chapter focuses on the following aspects:

- Analyzing the studies that define customer churn models and compare various predictive churn models among different algorithms, as well as identifying the corresponding effects and results found within these studies.

- Understanding the effect of class imbalance on the predictive performance of customer churn models and exploring strategies that can be utilized to address the class imbalance.

- Exploring the effects and results of including customer behaviour within predictive churn models.

## 3.1 Comparison of predictive churn models

Identifying customers that are likely to leave a company or organization is an important aspect within the field of customer retention and customer relationship management. Customer retention is significantly related to the economical value of a customer for a company, while the costs of acquiring a new customer surpasses the costs of retaining a customer and increasing the lifetime of a customer enhances the profitability of that specific customer for the company (Ali & Arıtürk, 2014). Subsequently, customer churn prediction is an established research topic within scientific studies. Customer churn prediction has been studied in a wide range of domains with the main research domains being industries with a large data flow like the telecommunications industry (Alboukaey et al., 2020; Huang, Kechadi, & Buckley, 2012; Hung et al., 2006; Ullah et al., 2019; Verbeke et al., 2012) or the gaming industry (Castro & Tsuzuki, 2015; Milošević et al., 2017; Perišić & Pahor, 2021). In addition to data-heavy industries being included as the research domain, other industries like the banking industry (Ali & Arıtürk, 2014) and the insurance industry (Günther et al., 2014) are also researched, albeit less extensively.

The importance of the research domain is observed within various aspects of customer churn prediction like the aspect of defining customer churn. As mentioned earlier, the definition of churn varies per industry, often involving one of the types of churn described by Lazarov and Capota (2007). Studies mainly focus on the question of how the customer has churned when defining the customer churn for their predictive models, frequently entailing the type of *total* churn or *hidden* churn. Research conducted in the context of a subscription-based industry commonly define churn as the official cancellation of a contract (Coussement & Van den Poel, 2008; Günther et al., 2014), while studies conducted in an industry that does not utilize subscriptions have commonly defined churn through a certain inactivity period (Alboukaey et al., 2020) or period of low activity (Hadiji et al., 2014).

A second aspect that varies among studies in the definition of customer churn is the level that defines whether a customer has churned or not. When looking specifically at the domain of interest, the subscription-based industries and corresponding insurance industry, differences are observed. Morik and Köpcke (2004) defined churn as the policy termination before the official end date, showing similarities to the research of Scriney, Nie, and Roantree (2020) which defined customer churn as an indicator whether the policy has been renewed. This is different from research like the study of Günther et al. (2014) and Risselada et al. (2010), which defined churn as the customer having an insurance policy leaving. The varying level of churn also has implications for the time period which is considered when defining churn, where defining churn based on the policy termination is dependent on the duration of the policy. This in contrary to other strategies which utilize an observation period.

Studies found in the research area of customer churn prediction often compare methods and utilize a wide range of methods for this purpose, as seen in the comparison of literature reviews depicted in Table 2.1. Traditionally, methods like Logistic Regression and Decision Tree have been commonly utilized, based on the high interpretability and relatively good performance (Neslin, Gupta, Kamakura, Lu, & Mason, 2006; Risselada et al., 2010). This follows from the trade-off that is observed within predictive churn modelling, an exchange between comprehensibility and predictive performance (De Caigny et al., 2020). Although a growing attention is observed within the literature for new prediction methods, Logistic Regression and Decision Tree are still frequently used. Logistic Regression is used by recent studies mainly in combination with other methods for comparison like in the research of Vafeiadis et al. (2015) and Mena, De Caigny, Coussement, De Bock, and Lessmann (2019). These studies have shown better predictive performance for the other models at the cost of less comprehensibility. Looking specifically at applications of Logistic Regression in the context of churn within the insurance industry, Bolancé, Guillen, and Padilla-Barreto (2016) and Spiteri and Azzopardi (2018) compared Logistic Regression against other techniques while Günther et al. (2014) solely evaluated customer churn prediction through a Logistic Regression model with a generalised additive model (GAM). Both comparisons show similar results as research in other industries, concluding that Logistic Regression is performing less compared to other models.

Decision Tree is considered by García, Nebot, and Vellido (2017) as one of the most popular types of predictive models within business applications. This popularity is also seen within the context of churn prediction, with multiple studies including customer churn in the domain of the financial industry (Keramati et al., 2016; Nie et al., 2011), telecommunications industry (Huang et al., 2012; Hung et al., 2006), and other industries like the gaming industry (Hadiji et al., 2014) and the online gambling industry (Coussement & De Bock, 2013). Just like the studies including Logistic Regression, most studies compare Decision Tree among other models. The results among these studies vary greatly, with different conclusions per study. Nie et al. (2011) concluded that Logistic Regression performs better than Decision Tree, where other studies show relatively high predictive performance (Hadiji et al., 2014; Keramati et al., 2016). Huang et al. (2012) concluded that the performance of Decision Tree is mainly beneficial when interested in the true churn rate. The observed variation in predictive performance and conclusions regarding the models indicate that Decision Tree significantly depends on the context of the study, highlighting the potential benefit of including a model within the research.

With regards to churn prediction in the insurance industry, Decision Tree models have been mainly compared against other models. Scriney et al. (2020) compare Decision Tree models against Naive Bayes models, SVM models, and Artificial Neural Network (ANN) models. They concluded that although models involving ANN scored better in terms of accuracy, recall and $F_1$ score, Decision Tree models follow closely. Bolancé et al. (2016) concluded that there is no distinctively optimal model among Logistic Regression, Decision Tree, SVM, and ANN model, supporting the observation that the predictive performance of Decision Tree models is comparable to other models. The results of Spiteri and Azzopardi (2018) showed that the Decision Tree model is closely following the Random Forest model in terms of accuracy, making it the second best model among Random Forest, Naive Bayes, SVM, and Logistic Regression models. In general, studies have shown the potential of Decision Tree models not only in the overall domain of churn prediction but also in the domain of churn prediction within an insurance industry. while maintaining a certain degree of comprehensibility.

Besides the previously mentioned studies of churn predicting in an insurance context, the study of Ahn, Hwang, Kim, Choi, and Kang (2020) analyzed a wide range of studies in the context of multiple studies, identifying in total four other studies that were conducted in the industry domain. These studies applied varying algorithms based on their distinct objectives, obtaining differing results. Morik and Köpcke (2004) compared a Decision Tree model, SVM model, Naive Bayes model, and an Apriori model with the objective of including more time information in their predictive churn model for a life insurer. The study sampled the dataset in such manner that in total 10,000 observations were used for the evaluation. In comparison, the best predictive performance was observed when utilizing the Decision Tree model, showing the highest performance on average in terms of the $F_1$ score, Precision and Recall. Hur and Lim (2005) conducted a study to analyze the feasibility of implementing a SVM model within the context of an online car insurer, comparing a SVM model against an ANN model and Logistic Regression model. Trained on a dataset of 13,200 observations, the SVM model showed the highest performance in terms of accuracy, having a significant better performance than the Logistic Regression and ANN model.

Risselada et al. (2010) analyzed the performance of predictive churn models in a number of periods later than the observation period, examining both Logistic Regression Models and Decision

Tree models. The comparison was performed on two industry-specific datasets, where one dataset is obtained from a health insurer with at most 1789 observations. Within the context of the health insurance dataset, it is seen that a Decision Tree model combined with a bagging procedure shows the best predictive performance in terms of top-decile lift and Gini coefficient. The authors of the study observed that the findings are stronger for the non-insurance dataset, providing the explanation that a larger dataset would be more beneficial for the applied techniques. Lastly, the study of R. Zhang, Li, Tan, and Mo (2017) analyzed the performance of a combination model, combining both deep learning in the form of a Neural Network and a Generalized Linear Model (GLM). The combination model is trained on a dataset of 15,167,324 instances gathered from a life insurance company. In comparison with eleven other models, the combination model showed the best performance in terms of accuracy, $F_1$ score, and AUC metric.

Analyzing various literature reviews and surveys that have been performed in the field of customer churn prediction, it is observed that other techniques are also frequently utilized besides the most commonly found methods of Logistic Regression and Decision Tree (Eria & Marikannan, 2018; KhakAbi et al., 2010). Neural Networks have been widely present in studies regarding predictive churn modelling using various types of Neural Networks (Buckinx & Van den Poel, 2005; Huang et al., 2012; Mena et al., 2019). Models utilizing Neural Networks often discover patterns in an incomprehensible manner, limiting the scope of application but providing a high degree of predictive performance (García et al., 2017). Buckinx and Van den Poel (2005) utilize an ARD Neural Network based on the property of built-in bayesian hyperparameter tuning, however comparing the model against Logistic Regression and Random Forest showed that there was no significant difference between models. Huang et al. (2012) applied a Multilayer Perceptron Neural Network in the context of churn prediction, however the model showed similar results as other models like Logistic Regression, Decision Tree, and SVM models. Based on the computational cost of training the Neural Network model, it was concluded that a MLP Neural Network model was not feasible for a large data set and other models are more favourable.

More recent research within industries with a large flow of information has focused on the application of other types of Neural Networks (Alboukaey et al., 2020; Mena et al., 2019). The research of Alboukaey et al. (2020) attempted to predict churn through the use of behavioural features, utilizing both a Convolutional Neural Networks and LSTM model while comparing the models with other models like a Recency, Frequency, and Monetary (RFM) based model. They concluded that the LSTM model showed the best predictive performance, closely followed by the RFM-based model. Mena et al. (2019) predicted churn through a Recurrent Neural Network (RNN) while incorporating customer behaviour using Recency, Frequency and Monetary variables, comparing the model against a Logistic Regression model. Within the study it was concluded that the RNN model performed better in terms of the top-decile lift and EPMC, however the RNN model required significantly more hyperparameter tuning and was less interpretable.

Models based on Support Vector Machines (SVM) are researched by various studies in the research field of predictive churn modelling like the studies of Coussement and Van den Poel (2008) and Chen et al. (2012). Coussement and Van den Poel (2008) researched the predictive performance of a predictive churn model based on SVM in an industry that sell subscriptions. In a comparison with models based on Random Forest and Logistic Regression on both the AUC and top-decile lift, it was concluded that with the use of a balanced dataset SVM could potentially outperform other models, while using SVM with an unbalanced dataset would provide worse results. The study of Chen et al. (2012) utilized a multiple kernel SVM, showing better performance than Logistic Regression in terms of AUC and lift, while being surpassed by a model using the Random Forest algorithm. However, utilizing SVMs is relatively computationally expensive, making the application less feasible when handling large datasets (Tsang, Kwok, Cheung, & Cristianini, 2005).

Considering the wide range of studies performed within the research area of churn prediction and more specifically churn prediction in the context of an insurance industry, an absence of ensemble methods used for predictive churn modelling is observed. Various studies have concluded that the application of ensemble methods can improve the predictive performance of the predictive churn models. Coussement and De Bock (2013) describes models that use an ensemble method as more robust, showing higher predictive performance in terms of the top-decile lift and lift index than both the general additive model and decision tree model. These results coincide with the study of Lemmens and Croux (2006) which concludes that models based on ensemble methods perform substantially better than a single Logistic Regression model, emphasizing the potential of the ensemble method models.

Vafeiadis et al. (2015) applied a boosting algorithm to both a SVM and Decision Tree model, exploring the impact of boosting whilst comparing with other models. From the comparison it was deduced that both the accuracy and the $F_1$ score were improved using the boosting algorithm, causing the SVM model in conjunction with the boosting algorithm to be the best performing model in terms of accuracy and $F_1$ score. The study concluded that the use of boosting techniques can optimize the performance of current models, therefore highlighting the performance of ensemble methods. The research of Burez and Van den Poel (2009) applied two ensemble methods in the field of predictive churn modelling, while exploring applications for handling the class imbalance problem. They concluded that the use of an ensemble method in corporation with class weights, the weighted Random Forest, outperformed other methods like Logistic Regression and Gradient Boosting. Based on the AUC and lift metrics it was deduced that the use of an ensemble method can be optimal in the context of class imbalance.

Based on the various results and conclusion of the various studies, it was expected that ensemble methods would be commonly present in the context of the insurance industry. However, only a small number of studies using ensemble methods are observed. Some studies regarding churn prediction in the insurance industry applied the ensemble method of Random Forest (Risselada et al., 2010; Spiteri & Azzopardi, 2018). Both studies have shown that the application of Random Forest can be beneficial for insurers, having the highest predictive performance among various models. Y. He, Xiong, and Tsai (2020) also utilized ensemble methods in their comparison, using both Random Forest, Extra Tree Classifier, and a Gradient Boosting model in their comparison. The study concludes that the ensemble methods are optimal in terms of AUC, highlighting the performance of both the Extra Tree Classifier and Gradient Boosting. However, the dataset used within the study is relatively small with only 25,275 observations. Other studies observed in the insurance context have mainly focused on other methods, therefore causing the application of ensemble methods for predictive churn modelling in the context of insurance companies to be significantly less researched.

## 3.2 Inclusion of Customer Behaviour

Inclusion of customer behaviour in predictive churn models is a topic that is gaining more interest, causing various approaches to utilize customer behaviour in predictive churn models to surface. As previously mentioned in Chapter 2, there are different types of customer behaviour activities and these different types can also be observed within the implementation of customer behaviour features (Buckinx & Van den Poel, 2005). The two types that are mainly observed within various studies are features based on the activities in the *acquisition/purchase* and *usage* phase. *Disposal* activities mainly entail activities of a customer that is disposing a product, whereas a predictive model attempts to predict whether a customer will dispose or not. Within the research field of predictive customer churn models, mainly customer behavioural features regarding the acquisition and use phase are considered.

An important property of behavioural data of a customer is the temporal nature of the data, where all data instances often contain a specific timestamp, duration or time of occurrence (Chen et al., 2012). This temporal nature is often disregarded in predictive churn models through transformation of the temporal behavioural variables into static variables (Buckinx & Van den Poel, 2005; Eichinger, Nauck, & Klawonn, 2006; Y. Zhang et al., 2011). By extracting the behavioural features during the whole observation period, the prediction of customer churn through behavioural features is approached as a simplified static problem, achieving higher predictive accuracy than models without behavioural features (Berger & Kompan, 2019; Buckinx & Van den Poel, 2005; Vafeiadis et al., 2015). The transformation of behavioural features into static features is generally done through various aggregation techniques, where aggregation simplifies the feature while still containing information regarding the behavioural activities. However, aggregating data instances can lead to a loss of information regarding the development of the data over time. Implementing static features subsequently imposes a trade-off between simplifying behaviour and accuracy of representation.

### 3.2.1 RFM Features

The implementation of static behavioural features within the context of predictive churn modelling mainly involves the notion of Recency, Frequency, and Monetary (RFM) variables. The RFM model originates from the research field of direct marketing, where these variables are well known as predictors for depicting customer's loyalty within the group of behavioural variables (Buckinx & Van den Poel, 2005). Many different operationalizations of the RFM model can be found within the predictive churn literature due to its simplicity and relatively good predictive performance (Mitrović, Baesens, Lemahieu, & De Weerdt, 2021). Fundamentally, a RFM model attempts to summarize customer's behaviour through *Recency* (i.e. the time since last purchase or renewal), *Frequency* (i.e. number of purchases or renewals), and *Monetary* (i.e. total monetary amount purchased) variables (Coussement & Van den Poel, 2009). This fundamental model can be extended by aggregation or by including other dimensions like time and direction, while research by Berger and Kompan (2019) has shown the feasibility of including other variables like intensity in the RFM model. Exploring different operationalizations and its implications can therefore aid in understanding the inclusion of behavioural features within predictive customer churn models.

The study of Mitrović et al. (2021) distinguishes two distinct types of customer behavioural features based on the RFM model, respectively summary RFM features and detailed RFM features. Summary RFM features are usually only different in terms of the type (i.e. recency, frequency, monetary) and the dimension of measurement (i.e. calls/seconds/sessions). Detailed RFM features can consider different dimensions (i.e. aggregations, time, direction) and often utilizes different transformations like averages and ratios. Research often combines the two types of features, using the summary RFM features as a foundation and further enhancing the predictive model through the use of detailed RFM features. Summary RFM features are frequently seen within the research of churn prediction, where *Recency* is often depicted as number of days since last usage, session or transaction until the end of the analysis period (Alboukaey et al., 2020; Buckinx, Moons, Van den Poel, & Wets, 2004; Castro & Tsuzuki, 2015; Miguéis et al., 2012; Perišić & Pahor, 2021). Differences in the definition of usage are observed, however the dimension of time is similar among various studies. According to Miguéis et al. (2012) the recency metric is considered to be the most powerful predictor of future behaviour, where a lower recency indicates a lower switch probability for customers of a company that offers fast moving customer goods (Buckinx & Van den Poel, 2005). The study of Perišić and Pahor (2021) that was conducted in the online gaming industry concluded that recency is beneficial for the prediction of customers that are likely to have switching behaviour, underwriting the imporance of including recency.

The second type of feature in the RFM model is the *Frequency* feature, which attempts to measure the strength of the customer relationship with the business (Miguéis et al., 2012). The summary feature of frequency measures the total number of activity within an observation period. Within the various studies differences are observed regarding the unit of activity based on the context of the study. Buckinx et al. (2004) define frequency as the number of visits within an observation period, Alboukaey et al. (2020) as the total number of purchases, while both Perišić and Pahor (2021) and Castro and Tsuzuki (2015) define frequency as the number of sessions within an observation period. Although differing in terms of defining the activity, all features within the studies attempt to measure the strength of the interaction between the customer and the organization. Differences are also observed within the third feature of the RFM model, the *Monetary* feature. This metric attempts to capture the monetary amount spent at the business, being mostly valuable in combination with the recency and frequency features (Miguéis et al., 2012). Multiple studies utilize the monetary feature, all defining the feature by the amount of money or currency spent in a specific observation period (Alboukaey et al., 2020; Buckinx et al., 2004; Perišić & Pahor, 2021).

The results of studies utilizing summary RFM variables are promising and have potential for the application within the insurance industry. Castro and Tsuzuki (2015) concluded that the model utilizing summary RFM features in the context of online gaming performs relatively well while maintaining a high degree of interpretability. Alboukaey et al. (2020) have shown that the use of a RFM-based model in the context of the telecommunication industry is the second best model after a LSTM model, which has significantly more complexity than the RFM-based model and as a consequence less interpretable. The study of Mena et al. (2019) indicated that the incorporation of RFM variables improved the predictive performance of both the LSTM model and Logistic Regression model, using a dataset of a company operating in the financial industry.

### 3.2.2 Extended RFM Features

Following the summary RFM features, various adjustments have been proposed in order to extend even further on the behaviour of a customer. These adjustments a wide range of variations, ranging from aggregation levels to averages (Mitrović et al., 2021). Adjustments to the RFM models are made while extending the behavioural variables can be useful in the capturing of more relevant data, therefore improving the predictive accuracy of models utilizing said features. One of the main categories of detailed RFM features is the transformation through taking the average or calculating a ratio. Transformations are for example seen in the research of Perišić and Pahor (2021), Hung et al. (2006), Wei and Chiu (2002), and Vafeiadis et al. (2015), where average or sum functions are used to capture more information. Besides using these functions, studies also frequently use a ratio function to compare the data instance to an average (Berger & Kompan, 2019) or to the length of the relationship (Buckinx et al., 2004).

A second type of adjustment to the RFM model that is frequently seen are the adjustments in dimensionality, obtaining different dimensions within the various RFM features. The study of Castro and Tsuzuki (2015) distinguishes different time slices for the RFM, whereas Perišić and Pahor (2021) utilizes features based on the short-term and long-term. Another dimension is proposed by Vafeiadis et al. (2015), distinguishing the features based on whether the activities were performed during the evening or not. Change in dimensionality is also seen within the direction of activities (incoming/outgoing) and a variety of different communication channels (Alboukaey et al., 2020; Motahari et al., 2014; Vafeiadis et al., 2015). In general, various transformations and dimensionality adjustments are observed within predictive churn research. A summary of various adjustments is seen in Table 3.1, providing an overview of various detailed RFM features as seen in the literature.

Table 3.1: Overview of several adjusted RFM features

| Variable | Description | RFM | Source |
|---|---|---|---|
| **Transformations** | | | |
| Session gap change | Difference between previous session gap time and average session gap | Recency | Berger and Kompan (2019) |
| rLorFrequency | Total number of visits relative to length of the relationship | Frequency | Buckinx et al. (2004). |
| rLorMonetary | Total spendings relative to length of relationship | Monetary | Buckinx et al. (2004) |
| LTV | Monetary value of a customer over its lifetime | Monetary | Perišić and Pahor (2021) |
| **Dimension transformations** | | | |
| Call-amount/SMS-amount | Number of contacts per communication channel | Frequency | Alboukaey et al. (2020) |
| Ses_Nr/LifeSes_Nr | Number of sessions in short term and long term | Frequency | Perišić and Pahor (2021) |
| Rec_C1/Rec_C2 | Number of days from last currency C1/C2 spent | Recency | Perišić and Pahor (2021) |
| Incoming/Outgoing | Number of incoming and outgoing calls | Frequency | Motahari et al. (2014) |
| $\frac{total\_night\_charge}{total\_day\_charge}$ | Total amount charged for day and night calls | Monetary | Vafeiadis et al. (2015) |

Following the use of summary and detailed RFM features, extensions to the RFM model in the context of predictive churn have also been proposed by multiple studies within different contexts. Various studies utilize additional features that extend the original RFM model for various reasons, ranging from adjusting the model to suit the research context (Berger & Kompan, 2019) or to tackle identified gaps to increase the predictive performance (Y. Zhang, Bradlow, & Small, 2013). These extensions can be useful for actual research and its context, therefore a descriptive analysis of the various extensions is performed.

The extension to the RFM model that is often seen within the literature is the use of an *intensity* feature as the monetary metric. Within various industries like the insurance industry and online gaming industry, no monetary value is directly involved while the customer simply has a subscription or does not have to pay for the usage. Therefore, measuring the monetary metric does not provide any valuable insights and can be replaced with a new intensity feature (Castro & Tsuzuki, 2015). This feature attempts to measure the intensity of the past activities and can be described in various ways. Castro and Tsuzuki (2015) describe the intensity as the sum of all session times recorded during the observation period, whereas Berger and Kompan (2019) describe the intensity or activity as the number of actions per session. This study even further extends on this intensity by comparing the number of actions in the last session against the average number of actions per session. Measuring the intensity of actions or sessions is mainly used as a viable alternative for the monetary feature when there is no monetary value observed, it can however also be jointly used with monetary features (Perišić & Pahor, 2021).

Another extension to the RFM model that entails the inclusion of a new variable is the extension made by Y. Zhang et al. (2015), who studied the inclusion of the *clumpiness* variable. Clumpiness is defined as the degree of clumpy data, meaning how close the various contact points are in terms of time compared to the overall observation period. Clumpiness can indicate an effect on the churn probability, causing this feature to be a possible improvement in terms of predictive accuracy. Besides including new distinct features based on the timing and frequency of customer behaviour, extensions can also be observed within the content of customers' actions. (Coussement & Van den Poel, 2009) studied the inclusion of adding extra information regarding the customer-business interaction, where emotions where obtained from the interaction using specific words. By abstracting these emotions, the predictive churn model was improved leading to increased predictive performance.

Utilizing models that entail detailed RFM features can be beneficial when compared to other models. Adding detailed RFM features showed better performance in terms of precision and recall in the study of Berger and Kompan (2019), whom researched churn prediction in the e-commerce industry. Buckinx et al. (2004) concluded that features that can be considered to be detailed RFM features are the most influential for classifying customers in the retail industry context. Extended RFM models like the models of Y. Zhang et al. (2015) and Coussement and Van den Poel (2009) also displayed positive results for the predictive performance. For example in the study of Coussement and Van den Poel (2009), it was concluded that the inclusion of extended RFM variables derived from the emotions of customers of a Belgian newspaper company improved their churn model in terms of AUC.

Looking specifically at research conducted in the context of the insurance industry, no significant studies are observed that have included customer behaviour. The study of Scriney et al. (2020) only included features regarding the policy, whereas Risselada et al. (2010) solely focused on including relationship and socioeconomic variables in the case of the insurance model. Hur and Lim (2005) focused mainly on the characteristics of the car that has been insured, which in the case of Bolancé et al. (2016) and Günther et al. (2014) has been expanded with policy and policyholder features. The inclusion of behavioural features within the research domain of the insurance industry is therefore not commonly seen, indicating a possible improvement for the models that can be used by insurers.

Overall, customer behaviour can be important for the predictive performance of churn models as seen in various studies. Data regarding customer behaviour in terms of interaction information is highly available, providing reliable features for the prediction of customer churn. Multiple studies have researched the inclusion of RFM variables and extensions to this model, however in the field of churn prediction within an insurer the model is relatively under researched. Including some approach for the handling of behavioural features can be potentially beneficial, emphasizing the need to research this specific topic in the context of an insurance company.

## 3.3 Class Imbalance

Datasets found within the research area of churn prediction commonly have a characteristic that can be problematic for the training of predictive churn model, which is the significant uneven distribution of churning and non-churning customers. The class imbalance significantly influences the training in such manner that it introduces bias, causing the model to have difficulties with predicting the minority class of churning customers. Burez and Van den Poel (2009) considered three problems as most relevant in the context of churn modelling, respectively the use of improper evaluation metrics, the relative rarity of the minority class and the greater effect of noise on the minority cases in terms of prediction. Each problem has multiple solutions which should be explored to potentially enhance the predictive models.

Utilizing a binary classifier for the purpose of predicting potential churning customer enables the implementation of evaluation metrics, evaluating the labeling of churning and non-churning customers. The use of commonly used evaluation metrics like accuracy and error rate are not appropriate for predictive churn modelling while these metrics display misleading results based on their strong dependence on the class distribution (Menardi & Torelli, 2014). Subsequently, different evaluation metrics are required to compare and assess predictive churn models. One of the most common metrics in the case of churn prediction with an imbalanced dataset is the AUC metric which follows from the receiver operating characteristics (ROC) curve. This metric enables the comparison of various models that consider imbalanced data and are therefore commonly seen as one of the main metrics in studies regarding churn prediction (Alboukaey et al., 2020; Ali & Arıtürk, 2014; Coussement & Van den Poel, 2009; Mena et al., 2019; R. Zhang et al., 2017). The AUC metric is considered by Burez and Van den Poel (2009) as a better overall evaluation metric compared to accuracy and proposes lift as a second metric that can be beneficial for the evaluation of predictive churn models. Although being more sensitive to class imbalance than the AUC metric, the lift metric is highly interpretable when utilizing in a marketing context like churn prediction. Especially the top-decile lift is commonly observed, being a metric that quantifies the gain when the model is utilized for marketing purposes (Ali & Arıtürk, 2014; Devriendt et al., 2021; Lemmens & Croux, 2006). A third metric that is sometimes considered in studies within the domain is the $F_1$ score (Alboukaey et al., 2020; Keramati et al., 2016; Mishra & Reddy, 2017). By combining both the precision and recall metric, a harmonic mean is found that can be used to assess a model's performance. In terms of graphical representation, the ROC curve is often used while it aids in the comparison of different trade-offs per distinct classifier (Menardi & Torelli, 2014). Furthermore, the lift chart is often seen in literature regarding churn prediction to visualize the results from a marketing perspective (Burez & Van den Poel, 2009).

Besides focusing on the evaluation metrics, literature often addresses the class imbalance problem by sampling the data in such manner that the distribution of churning customer is improved for the training of the models. As mentioned in Chapter 2, the strategies of undersampling, oversampling, and hybrid sampling can be used in an attempt to improve the predictive performance of the churn model. One of the most commonly used strategies, undersampling, is often done when the corresponding dataset is large and have a significant number of records with a churn label. For example, Scriney et al. (2020) and Perišić and Pahor (2020) utilized a random undersampling strategy to obtain an equal distribution between churning and non-churning customers, losing possible observations that can be beneficial for the predictive performance. Both studies only utilized undersampling, therefore no conclusions are provided regarding the benefits of undersampling in the context of churn prediction. Berger and Kompan (2019) tries to counter the downside of losing potentially important observations by first dividing the customers into several clusters, ensuring that as much data patterns as possible are included. This approach performed better in terms of the AUC, precision, and recall metrics than a random undersampling approach at the cost of increasing model complexity. It is noteworthy that the random undersampling approach performed better than a model trained on the whole imbalanced dataset in terms of precision and recall, however the random undersampled model was outperformed by the whole dataset in terms of the AUC.

The comparative study of Amin et al. (2016) attempted to find the most optimal oversampling strategy in the context of churn prediction, comparing six distinct oversampling strategies. Within the study, all techniques show a relatively high degree of precision and recall. The compared models all entail some sort of oversampling strategy, which prevents conclusions from being drawn in relation to other strategies like undersampling or utilizing the whole dataset. Verbeke et al. (2012) compared the oversampling strategy using multiple datasets and techniques, extensively researching the approach of oversampling. It was concluded that oversampling is not significantly improving the predictive performance of the model, however an observation was made that the impact of oversampling was highly context-dependent. Therefore, it was advised to empirically test whether an oversampling technique is beneficial for the predictive performance. This coincides with the research of Zhu, Baesens, Backiel, and Vanden Broucke (2018) that after extensively comparing various sampling strategies concluded that the impact of using a certain sampling strategy depends on the context regarding the evaluation metrics used and the used algorithms.

In conclusion, the aspect of class imbalance have some implications for predictive churn models. The evaluation metrics on which various models are compared should be carefully chosen in order to reduce the effect of class imbalance and corresponding bias on the evaluation. It was shown that some evaluation metrics are less susceptible to the class imbalance and therefore should be preferred. Another method of handling the class imbalance was identified to be the sampling strategy, a strategy in which the distribution of churning customers and non-churning customers is improved. Various strategies have been compared in multiple studies, often emphasizing the importance of the model's context. The effect of sampling strategies is dependent on aspects like the evaluation metrics, the algorithms used for the models, and the original dataset. Subsequently, an empirical comparison is needed to fully capture the effect of the class imbalance on predictive churn models.

# Chapter 4

# Data Understanding

Obtaining a dataset that is suitable for the process of predicting the probability that a customer is leaving is considered to be a comprehensive task. Businesses obtain a lot of information regarding the customer, interactions between the customer and the company, and other important pieces of information. This chapter will describe the dataset that is being used for the predictive churn modelling, provide an overview of the data frame, and define the preprocessing steps that have been undertaken to enable the analyses.

## 4.1 Dataset Creation

In order to conduct this research, permission was granted to utilize the data sources of a Dutch insurance company. The dataset has been obtained from the data warehouse of the company and entails information regarding demographics of the customer, subscription data, and data regarding the interactions of a customer with the company, a combination of information that is often seen within predictive churn modelling (Hung et al., 2006). Every day information is gathered of customers and their actions, interactions with the company, and usage of their applications. The amount of raw data obtained by a business is ever-increasing, where all this information is stored within the databases of the company. Because a significant amount of data has already been gathered for a long time, the databases of a business like OHRA are extensive and detailed, causing the incorporation of all gathered data to be out of scope for this research project. It is therefore necessary to determine a subset of information that will be utilized within this project, also known as the process of *sampling*.

The process of creating a dataset starts off with determining the unit of analysis. Within this research, the unit of analysis has been defined to be a specific type of policy, the car insurance policy. A customer purchases a car insurance policy to insure a specific car, and this policy can be cancelled every month. Analyzing whether the policy will be cancelled is the main objective of this study and avoids the issue of defining customer churn, while a cancelled policy is easily quantified as customer churn based on the contractual terms. OHRA utilizes two different IDs, a policy ID and relation ID. The relation ID is mainly used to link multiple policies to the specific person, whereas the policy ID is used to distinguish the policies. This research focuses on a specific type of insurance, the car insurance, and therefore the policy ID should be utilized. The policy ID will be used through the research project to distinguish the policies, enabling the prediction of each unit of analysis separately. The relation ID is however used for linking interactions of customers to policies.

Being the center of the analysis, the policy holder has a lot of information stored within the dataset and the databases of OHRA. In terms of information, two specific general types of information can be distinguished, respectively information that does not change along a temporal dimension and data that does change along a temporal dimension, also known as time-invariant and time-variant data. It is considered that time-invariant data can be described as static, remaining constant over time. Following the two general types of information, three specific types of information are also observed within the data provided by OHRA. The first type of information describes the policy holder through socio-demographic variables, while the second type of information describes information of the product and other services provided to the customer through the use of product-related variables. Lastly, the interactions between the customer and the company are described using customer interaction variables. These three types are often seen within studies regarding predictive churn modelling (Hung et al., 2006). Looking at studies that cover predictive churn modelling within an insurance context, it is indeed seen that socio-demographic variables, customer interaction variables, and product-related variables are often used for the predictive modelling of customer churn (Günther et al., 2014; Risselada et al., 2010).

In addition to the unit of analysis, the target variable is a variable which depicts whether a customer has churned or not. Within the data this is depicted through the use of a binary variable, where a *1* indicates that the customer has left the company in the following month and a *0* indicates that the customer has not left the company. This variable however also contains a 1 when a customer has died, causing the inclusion of non-voluntary churn within the variable. Chapter 2 distinguished various types of customer churn, entailing information on how a customer can leave the company. While Lazarov and Capota (2007) define multiple definitions of churn, the type of voluntary churn has been defined to be the type of interest. The unit of analysis, a policy, enables a straightforward method of measuring customer churn, based on the contractual terms. Customers that were excluded by the company are not included in the provided dataset based on the sensitive nature of these customers.

The data in the data warehouse consists out of time-variant and time-invariant data, where

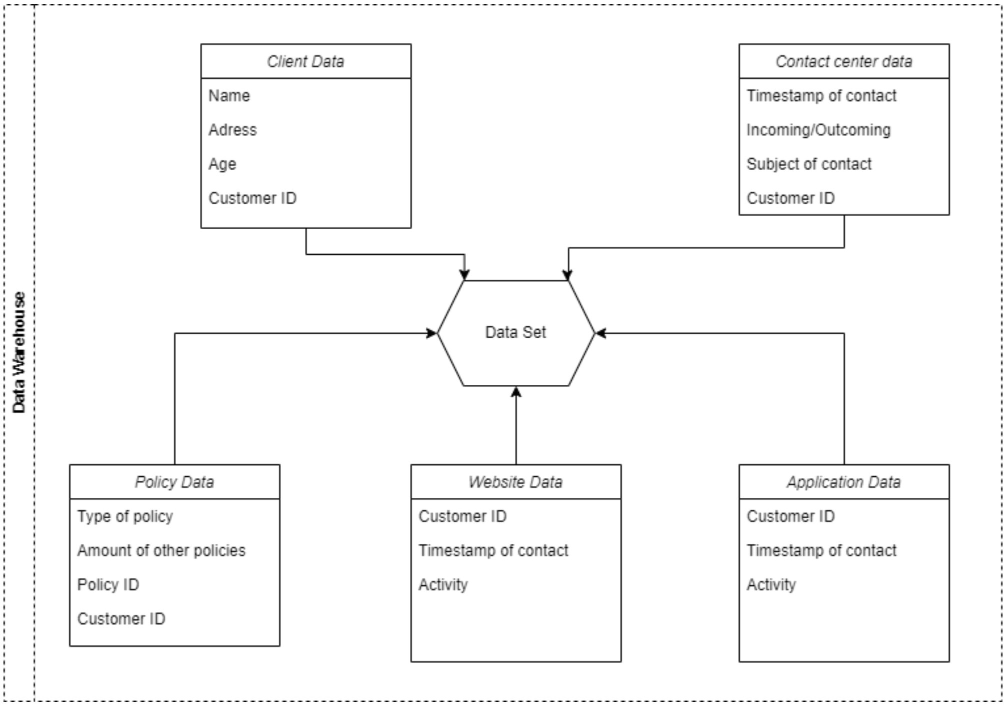

Figure 4.1: Overview of the dataset

the objective is to aggregate both types of data into one dataset. OHRA desires to predict the probability of a policy being cancelled on a monthly level, in order to being able to revisit the model that triggers when a policy is likely to be cancelled every month. Therefore, the time window will be a month, causing the dataset to have 12 data points per policy if the customer has not cancelled the policy. When a policy holder has voluntarily left the company, the policy subsequently has less data points than 12. Multiple data points per policy indicates that the data is longitudinal data, enabling the prediction of customer churn to utilize lagging variables or time-variant data more extensively. The data is organized in such manner that every data point has a reference data which is the first day of the month. Every variable in the dataset is measured on this exact reference data with two exceptions. Data regarding interaction with the customer is aggregated on a monthly level, causing the related variables to be different from the other variables. Secondly, the variable regarding the customer churn is as mentioned before binary and indicates whether the customer will churn. This project attempts to research how the inclusion of the time-variant information can affect the predictive performance of the predictive customer churn models, focusing on the interaction data.

In general, the data provided by the company has the structure as depicted in Figure 4.1. The demographics of the policy holder are obtained from the client data, while the information regarding the policy has been obtained from a different data source. Lastly, all the contact data has been obtained from a specific database that corresponds to the method of contact. Following the time windows previously mentioned, the client and policy data is simply measured on the specific reference date specific to that data point. The data point is further extended on by including information per type of contact, including all information regarding interactions with the company in the month before the reference data and aggregating on a monthly basis. After these steps all the data is joined based on customer ID or policy ID, obtaining a dataset that entails a significant amount of information per policy per month. The year 2021 is taken as measurement year, whereas the data regarding interactions with the company had a window of 12 months before the reference date. The information regarding interactions which can be found in the contact center, application, and website data, therefore have two years worth of data. The data of 2020 is used to build a historical image of the customer behaviour, while the data of 2021 is utilized to evaluate current behaviour of the customer. The completely processed dataset has 1,965,932 data points with 192,293 distinct policies and 25,207 churned policies. Therefore, the churn rate in all the data points is 1.3%, whereas the churn rate in terms of policies is larger than 10%.

An analysis of examined research as depicted in Table 4.1 shows that there is no clear indication of a relationship between the sample size, the churn percentages, and features used within the scientific field of churn prediction. The scientific research in the analysis utilizes a sample size of 1,474 to 1,500,00 with churn percentages ranging from 2.00% to 30.57%. Secondly, a wide range of variables used within the predictive models is observed. Within the various studies a substantial variety is observed within the three aspects of sample size, churn percentage, and number of variables used. Using a dataset containing 192,293 distinct policy holders can be concluded to be suitable for performing analyses on predictive churn modelling. Before utilizing the processed dataset of 192,293 distinct policies, an understanding of how the provided dataset is processed to obtain this number of distinct policies and general overview of the dataset is required.

## 4.2 Data overview and preprocessing

The data as collected by OHRA was found in the data warehouse of the company, where the data was retrieved from various sources, which often are referred to as the CRM systems of a company. Within these CRM systems three distinct types of data can be found that will be utilized for this

Table 4.1: Analysis of sample size, churn percentage, and number of variables per scientific paper

| Paper | Sample size | Churn percentage | Variables |
|---|---|---|---|
| (Mozer et al., 2000) | 2,876 | 6.2% | 134 |
| (Coussement & Van den Poel, 2008) | 90,000 | 30.57% | 32 |
| (Risselada et al., 2010) | 1,474 | *Unknown* | 6 |
| (Verbeke et al., 2012) | 338,874 | 14.10% | 22 |
| (Günther et al., 2014) | 127,961 | *Unknown* | 17 |
| (Alboukaey et al., 2020) | 1,500,00 | 12% | 10 |
| (De Caigny et al., 2020) | 607,125 | 2.00% | 17 |
| (Devriendt et al., 2021) | 200,903 | 20.51% | 162 |

research, the customer level data, the contract level data, and the interaction data. The data spans a time horizon of two years in the past, 2020 and 2021, causing the data gathering process to be entirely completed by the company. The phase of data gathering is subsequently not of interest for this study, enabling the immediate start of the preprocessing phase of the data. While the data has been gathered from the data warehouse, the data was already structured to a certain extent. By providing an overview of the data and describing the preprocessing steps, a better understanding of the dataset is obtained.

### 4.2.1 Data overview

Having three distinct types of data, the overall dataset is considered to be a quite extensive dataset. Before the start of the data cleaning phase, the dataset contains 91 different variables. Of these 91 variables only one variable is not considered to be one of the three distinct types of data, respectively the reference date. Regarding the customer-level data, 16 variables are considered to be customer-level data, whereas 18 variables are considered to be interaction data. Lastly, the remaining 59 variables are describing some information regarding the policy of the customer, also entailing the target variable whether the policy was cancelled ($churn = 1$) or not ($churn = 0$). A summary of the variables found within the dataset is provided in Table 4.2. As can be seen in this summary, the policy-level variables are the main source of information, which coincides with the unit of analysis being the policy holder. Through the use of 16 variables in the dataset the policy is described, whereas 17 variables are used to describe the car that is insured by the policy.

The information is extended with the two other types of variables, where the customer-level variables are the second largest. Customer-level variables are mainly describing the sociodemographics of the customer that has the policy and the other policies of OHRA that the customer has. Lastly, the interaction-level variables describe the different types of interaction that a customer has with the business. The types of interaction that are observed within the dataset are contact by phone, email, Whatsapp, and the use of both the website and the mobile application. These interactions are described by a timestamp, a variable that describes what the activity entailed, and variables that describe the content of activity. Although some session IDs are registered, not every interaction has a corresponding session ID variable. Within the dataset, both the use of the mobile application and the contact by phone do not have any corresponding session ID variable. Lastly, the reference date is provided in every entry. This reference date can be used for aggregation, aggregating multiple entries based on the specific reference data.

Although the dataset has already been structured by the company to a certain extent and can be considered of relatively high quality, it is still required to process the dataset in order to utilize the data within statistical models for the sake of predicting customer churn. Within

Table 4.2: Summary of information found in dataset

| **Customer-level variables** | **Number of variables** |
|---|---|
| Customer ID | 1 |
| Socio-demographic (e.g. age, place of living) | 9 |
| Other policies of customer | 14 |
| Customer' preferences | 3 |
| Lifetime of customer | 2 |
| *Total* | *29* |
| **Policy-level variables** | **Number of variables** |
| Policy ID | 1 |
| Churn related | 2 |
| General information (e.g. payment, duration) | 7 |
| Car policy | 16 |
| Car characteristics | 17 |
| *Total* | *43* |
| **Interaction-level variables** | **Number of variables** |
| Session ID | 3 |
| Timestamp | 7 |
| Content of activity | 8 |
| *Total* | *18* |
| **Other variables** | |
| Reference date | 1 |

the unprocessed dataset duplicate entries and other data discrepancies are observed, causing the need for further data cleansing and filtering. The unprocessed dataset contains 12,165,279 rows or entries, in which more than 200,000 distinct policies are found. These policies are purchased by less customers, indicating that there are various customers that have multiple car insurance policies. This unprocessed dataset will eventually lose some entries due to the sampling, which is elaborated later on.

### 4.2.2 Defining the data frame

Based on the unit of analysis being a policy and the requirements of the company to predict for each month what the probability of the cancellation of a policy was, the data had to be setup in a specific way. For the dataset, twelve reference dates are needed per policy except for the policies that are new or policies that have been cancelled. By using twelve different data points per policy, the data is considered to be of a longitudinal nature. This nature has implications for the data, where time-variant data is dependent on these various reference dates. Most customer-level variables and policy-level variables are time-invariant data, therefore being static and not requiring any further processing to have the specific setup. The interaction level variables whatsoever depend on the time, due interactions between the company and the consumer occurring throughout the year and not happening in a static manner. In order to match the setup of having one data point per month, the activities are aggregated per interaction channel. All the activities that happened during a specific month are grouped by using a simple aggregation per interaction channel, summing all interactions per month per channel. The resulting variable refers to the total amount of interaction specific to a communication channel, describing the frequency of contact.

Besides aggregating the activities, the intensity of the activities and the recency of the activities are also aggregated in such manner that there is one data point per month. The intensity of activities are aggregated by summing the amount of URLs visited or clicks that happens per session, both for interaction via the website and interaction with the mobile application. The resulting variable describes how many actions happened during the sessions that occurred in that month. Lastly, the timestamp of the activity are consolidated by taking the most recent timestamp, resulting in a specific variable per interaction channel. The resulting data frame is shown in Table 4.3, in which the longitudinal nature of the data can be seen. Based on the target variable whether the policy was cancelled ($churn = 1$) or not ($churn = 0$), the number of data points per policy is reduced when the customer has cancelled the policy. Defining the data frame using this approach directly affects the modelling, while the amount of non-churning policies per month are significantly more than the churning policies per month. This implication should be taken into consideration when evaluating the models and further details are presented in later chapters.

Table 4.3: Visualization of utilized data framework

| **Data frame** | | | | | | |
|---|---|---|---|---|---|---|
| **Policy ID** | **Customer ID** | **Reference date** | **Variable X** | **Variable Y** | … | **Churn** |
| 1 | a | $i$ | $X_i$ | $Y_i$ | … | 0 |
| | | $i+1$ | $X_{i+1}$ | $Y_{i+1}$ | … | 0 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | $i+11$ | $X_{i+11}$ | $Y_{i+11}$ | … | 0 |
| 2 | a | $i$ | $X_i$ | $Y_i$ | … | 0 |
| | | $i+1$ | $X_{i+1}$ | $Y_{i+1}$ | | 1 |
| 3 | b | $i$ | $X_i$ | $Y_i$ | … | 0 |
| | | $i+1$ | $X_{i+1}$ | $Y_{i+1}$ | … | 0 |
| | | $i+2$ | $X_{i+2}$ | $Y_{i+2}$ | … | 0 |
| | | $i+3$ | $X_{i+3}$ | $Y_{1+3}$ | … | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Policy ID<br>$N$ | Customer ID<br>$M$ | $i$ | $X_i$ | $Y_i$ | … | 0 |
| | | $i+1$ | $X_{i+1}$ | $Y_{i+1}$ | … | 0 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | 0 |
| | | $i+11$ | $X_{i+11}$ | $Y_{i+11}$ | … | 0 |

### 4.2.3 Processing steps

After creating the data frame, more processing steps are undertaken to enable the process of predictive modelling using statistical models. The data includes various categorical variables that require processing before being used within the algorithms used to predict the customer churn. Handling categorical variables with distinct categories like type of contract can be done in various ways, creating numerical variables from categorical variables. For this project the choice has been made to one-hot encode the various categorical variables, based on the reasoning that creating dummy variables is considered to be simplistic and helping in the interpretation of the various variables. This however increases the amount of variables drastically, which should be taken into consideration when evaluating the model.

Secondly, the various numerical variables found in the dataset have to be normalized to avoid any bias towards features. All the numerical variables are normalized through the use of standard rescaling which utilizes the equation as depicted in Equation (4.1). This scaling procedure creates the scaled data $z$ by subtracting the mean of the training sample $u$ and dividing by the standard deviation of the training sample, therefore scaling to unit variance and obtaining a standardized distribution.

$$z = \frac{(x - u)}{s} \tag{4.1}$$

Datasets like the dataset used in this research inevitably have missing data and this dataset is no exception. Missing data can be for example found within categorical variables like the province where the customer lives and the type of car that is insured with the policy. As can be seen from the data, the relative share of missing data is small, respectively 0.48% and 0.55%, but the missing data still requires to be addressed. For categorical variables the missing data is handled as a new category, causing the missing data to become a new variable based on the one-hot encoding. This approach was deemed feasible while the percentage of missing data is considered to be low, but adding the information about the missing variables could add some information to the predictive models. With regards to the numerical data, even lower percentages of missing values were observed. These values have been imputed with the median, being a relatively robust measure while still maintaining the other information of the row.

## 4.3 Sampling

When the data has been processed, some sampling decisions are made to prevent additional bias from being included. In order to obtain the most realistic dataset, only a small number of sampling decisions are made. First of all, the deceased customers are removed from the dataset while this type of churn is different and deemed unnecessary to include. Secondly, the decision has been made to include only policies that have a policy duration for longer than a month. It is observed that policies with a policy duration of less than 31 days have a different usage pattern than older policies, potentially being a source of bias in the behavioural features. As can be seen in Figure 4.2, the usage differs substantially. Besides being different, the exclusion is also based on the fact that the policies do not have a complete month of data. Furthermore, no other sampling decisions were made, causing the complete dataset to have 1,965,392 data points with 192,293 distinct policies.

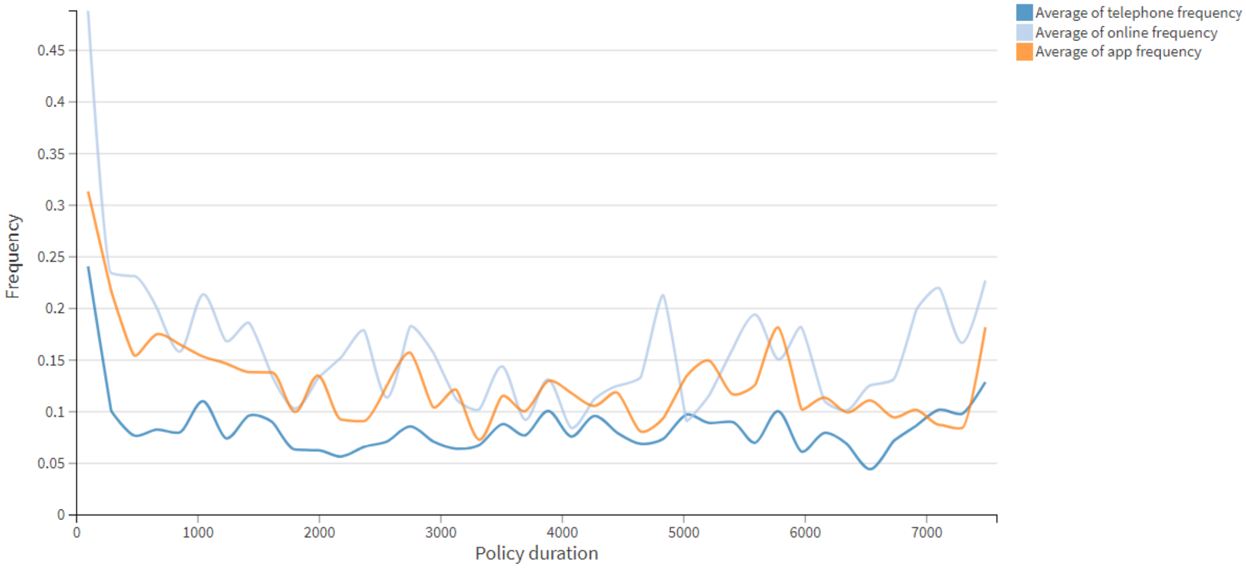

Figure 4.2: Overview of frequency variables across policy duration

# Chapter 5

# Methodology

Conducting experiments and research that are reproducible and as unbiased as possible is one of the main challenges of conducting research in general. The complexity of predictive churn modelling and the process of predicting whether a customer makes a certain decision demands for standardized experiments. This chapter attempts to provide an overview on how the experiments were conducted, which decision were made with regards to the research design, and how certain aspects like the sampling strategy and validation were performed.

## 5.1 Algorithms

Selecting algorithms is a delicate process in which various considerations have to be made with regards to predictive performance, applicability to the experiments, and the preferences of the organization at which the experiments took place. An extensive overview of various machine learning algorithms used within the research field of predictive churn modelling and other comparable research areas is provided in Chapter 2 and Chapter 3. Each algorithm has certain implications for the analysis, therefore a variety of algorithms have been chosen. The chosen algorithms are depicted in Table 5.1.

Table 5.1: Algorithms used within the analysis

*Algorithms*

| **Algorithm** | **Type of algorithm** |
|---|---|
| Logistic Regression | Regression |
| Decision Tree | Decision Tree |
| Random Forest | Ensemble |
| Gradient Boosted Trees | Ensemble |
| LightGBM | Ensemble |
| XGBoost | Ensemble |

As can be seen in Table 5.1, three types of algorithms used within this research can be distinguished. The first type, *Logistic Regression*, is commonly used in churn prediction as can be concluded from the reviews of KhakAbi et al. (2010) and Eria and Marikannan (2018). This is supported by Günther et al. (2014) who describe Logistic Regression as relatively simple and robust, showing overall a good performance. Based on the frequent use within customer churn prediction and the robustness of the model, Logistic Regression is included within the analysis. The second algorithm used within the analyses, *Decision Tree*, is also a frequently used algorithm within the research area of predictive churn modelling (Neslin et al., 2006). Decision Tree is widely known for its comprehensibility, although models involving Decision Tree can easily be overfitted to the data. However, both Decision Tree and Logistic Regression are considered algorithms that show lower predictive accuracy while being significantly more interpretable (Risselada et al., 2010). Therefore, the Decision Tree algorithm is included within this analysis.

Following the two more interpretable algorithms, four ensemble algorithms are also included within the analysis. As described in Chapter 3 and Chapter 2, ensemble methods essentially combine multiple classifiers to obtain a more accurate classifier. According to Opitz and Maclin (1999) two of the most popular methods of ensemble learning are *bagging* and *boosting*, therefore both methods are included within the research. The method of bagging is represented by the use of *Random Forest*, which combines multiple decision trees. Although lacking some interpretability, Random Forest has shown promising results within analyses like the analysis of Burez and Van den Poel (2009) and the analysis of Ullah et al. (2019). Subsequently, Random Forest is considered to be a good bagging method to include within the analysis.

Lastly, three *boosting* algorithms as part of the *ensemble* methods are included. Research within the field of customer churn prediction has shown the potential of boosting algorithms (Lemmens & Croux, 2006; Vafeiadis et al., 2015). Based on the potential and the fact that the company where the research is conducted also frequently uses boosting algorithms within their churn models, multiple algorithms have been included. XGBoost is one of the most widely known boosting algorithms and is known for its good performance and scalability. The second method, LightGBM, is an algorithm which performs relatively well with large datasets while having a significantly higher efficiency. Lastly, the algorithm of Gradient Boosted Trees is also utilized. Although this algorithm is not as memory efficient as the other two algorithms, performance of this model is still considered to be feasible for this analysis. Ensemble methods can significantly increase the predictive performance compared to other methods when dealing with an imbalanced dataset, therefore justifying the extensiveness of analyzing the various ensemble methods (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011).

## 5.2 Sampling Strategy

Predicting customer churn is almost always accompanied with an imbalanced dataset, where companies that have a relatively normal service level find significantly more non-churning customers in their data than churning customers. Customer churn is considered to be a rare object that is not often occurring, causing the imbalance found within the dataset (Amin et al., 2016; Burez & Van den Poel, 2009). Having said class imbalance, this imbalance within a dataset negative influences the performance of a machine learning model while the classification will be biased towards the majority class (Abd Elrahman & Abraham, 2013). When a model is trained on a highly imbalanced dataset, the model tend to predict all instances as part of the majority class, resulting in high accuracy but an unacceptable degree of precision (Zhu et al., 2017).

Reducing the effect of class imbalance can be done through various ways as seen in Chapter 3, where one of the most frequently used measures to increase the performance is the use of a sampling technique. Based on the importance of class imbalance within the research field of churn prediction and data mining in general, the subject of reducing the effects of class imbalance through sampling strategies is a widely researched topic. Commonly, three types of sampling techniques are found for the handling of significant imbalanced dataset. The first type is *undersampling*, where instances from the majority class are removed by using a selection technique that has been predefined. The second type, *oversampling*, defines synthetic instances that closely related to actual instances found in the minority class. Lastly, *hybrid sampling*, combines both over-sampling and undersampling in one strategy. This strategy commonly entails the creation of new synthetic instances through an oversampling strategy, after which the majority class is reduced.

The results of applying sampling techniques on the predictive performance of predictive churn models varies per study. Burez and Van den Poel (2009) concluded that undersampling can be beneficial for the prediction of customer churn, however the ratio to which should be undersampled is dependent on the method of undersampling and the application. Zhu et al. (2018) argue that it is not feasible to determine the most optimal strategy within a general context, emphasizing the importance of the context. Based on this dependence on context and methods, three different sampling strategies will be evaluated among the two best performing algorithms per model. An overview of the utilized sampling strategies are depicted in Table 5.2. Random undersampling is performed to obtain a churn percentage of approximately 10%, therefore including every churning policy and randomly selecting one instance per policy ID of the policies that were not cancelled. For the over-sampling strategy, the choice was made to include SMOTE-NC. SMOTE-NC is considered to be one of the most important oversampling techniques and seen as a benchmark of other oversampling strategies Amin et al. (2016). In terms of the hybrid sampling strategy, the technique of SMOTE-NC was used in combination with random undersampling.

Table 5.2: Utilized sampling strategies

*Sampling strategies*

| **Sampling strategy** | **Method** |
|---|---|
| Undersampling | Random Undersampling |
| Hybrid sampling | SMOTE-NC and Random Undersampling |
| Oversampling | SMOTE-NC |
| Whole dataset | No sampling utilized |

## 5.3 Cross-validation

Within the field of supervised machine learning, it is common to split the dataset using an approach in a *train*, *validation* and *test* set. This splitting is done to receive an unbiased analysis of the performance of the models trained using the predefined algorithms, as seen in Section 5.1. The *training* set is utilized to train the model using the algorithm, where the algorithm attempts to find (hidden) patterns within the training data. The output of this training process is a machine learning model, that will according be evaluated using unseen data in the *test* set, and subsequently preventing bias in the model evaluation. When the model is evaluated along data that is also trained on in the *training* set, the model evaluation will be unrealistically optimistic. As a consequence, the model's evaluation can not be generalized for data that the model has never seen. This process of overfitting should be avoided to have more reproducible and transparent results.

Machine learning algorithms have certain parameters that have be chosen before the optimization of the model, the so-called *hyperparameters*. Optimization of parameters demands for a method of evaluating the various parameters, measuring how good a specific set of parameters compares to other sets of available options. This process of tuning and evaluating the hyperparameters requires a third set of unseen data, while using data from the test set is not possible. When data from the test set is used for the evaluation of the hyperparameters, the data does not remain unseen and therefore increases the bias in the models. Subsequently, the process of hyperparameter optimization requires a third subset of the data, known as the *validation* set.

The division of the data set into three subsets decreases the number of data points per set, causing a trade-off between available data points for training and the introduction of bias. One of the most frequently used resampling method is the method of *cross-validation*, which aids in the assessing of the generalization ability and prevents overfitting (Berrar, 2019). Cross-validation has various methods that samples differently, however the choice has been made for this project to utilize the method of *k-fold cross-validation*. Within this method, the available training set is partitioned into $k$ subsets of approximately the same size. These subsets are referred to as "folds", hence the name *k-fold cross-validation*. The division into approximately equal subsets is performed randomly without any replacement. Following the division, the model is trained using $k-1$ subsets which will be the training set, after which the model is applied to the remaining subset to validate the performance of the hyperparameters. This process is repeated until each $k$ subset has been used as a validation set, where the average of $k$ performance measurements is known as the cross-validated performance.

The method of cross-validation is further extended for this research project by involving the use of *stratified* random sampling when defining the various folds. This involves a sampling method for the various fold that ensures that the proportion of the target variable is the same for the different folds as it is for the complete learning set (Berrar, 2019). By keeping the proportion constant, the ratio of churning policies and non-churning policies is the same among the various folds. Constant proportions are required to prevent the introduction of bias and is especially important for the predictive modelling of customer churn. Classification of customer churn is known for the imbalance that occurs among churning and non-churning, causing it to be extra liable for this type of bias.

For this research project, a strategy of *stratified 5-fold cross-validation* is applied. The complete dataset is split using a 80/20 splitting strategy which is considered feasible based on the size of the dataset, resulting in a train set of 1,572,314 data points and a test set of 393,078 data points. The train set is after the first partitioning divided in 5 subsets, which are used for the hyperparameter tuning‘and result in a final model. Resulting models are afterwards used for the prediction of customer churn on the test set. The whole validation strategy of this research is depicted in Figure 5.1.

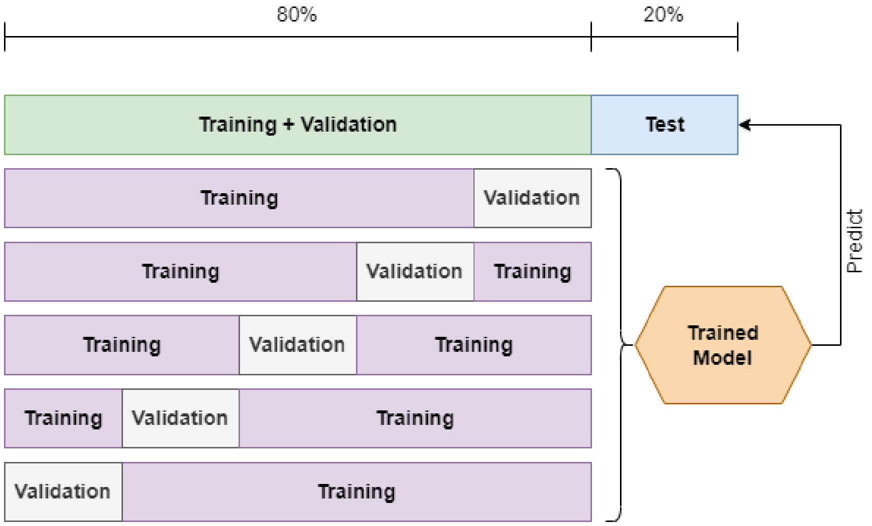

Figure 5.1: Overview of cross-validation strategy

## 5.4 Hyperparameter Optimization

Inherent to the use of machine learning algorithms is the process of optimizing hyperparameters, also known as hyperparameter optimization (HPO). The tuning of hyperparameters can significantly improve the performance of a model but the optimization of hyperparameters in a practical application also contains a trade-off between objectives (Feurer & Hutter, 2019). Within HPO a trade-off can be observed between the objective of obtaining the highest performance of a model and the objective of keeping the usage of computational resources at a feasible level. Every machine learning algorithm has a different set of parameters, where the hyperparameters of an algorithm are the parameters that have been defined before the phase of learning. The optimization of these parameters can be done through a wide range of methods, of which grid search is seen as the most basic method. When using grid search, a specific set of values per hyperparameter is provided by the user, after which the specific set is evaluated based on the performance of the training algorithm. The previously mentioned trade-off of objectives is clearly distinguished within the use of this search strategy, while grid search suffers from high-dimensional search spaces when aiming for the highest model performance possible (Liashchynskyi & Liashchynskyi, 2019).

An alternative to the grid search is the random search strategy, a strategy for finding the optimal hyperparameters which entails the use of random configurations of hyperparameters until a specific budget for the search is exhausted (Feurer & Hutter, 2019). The parameter values are randomly sampled from a predefined distribution for a set number of rounds, having a processing budget that is not dependent on the parameters and possible values of these parameters. This strategy is able to outperform the grid search strategy, however this is mainly the case when only a small set of hyperparameters is important for the performance of the training algorithm. Furthermore, random search entails easier parallelization and flexible resource allocation. Although this search is able to outperform the grid search strategy, it often takes far longer than guided search methods to find one of the most optimal hyperparameter configurations.

This research project however will utilize an optimization framework that recently gained traction due to its promising results in various research areas, an optimization framework known as Bayesian optimization.(Feurer & Hutter, 2019; Shahriari, Swersky, Wang, Adams, & De Freitas, 2015). The central idea of Bayesian optimization is to create a model that can be updated and queried to evaluate optimization decisions. Bayesian optimization is an iterative algorithm with two main aspects, a probabilistic surrogate model and an acquisition function. The probabilistic surrogate model is fitted to all observations of the target function that has been iteratively made, while the acquisition function is used to decide which point will be evaluated next. Based on the promising results in recent research areas as mentioned by (Feurer & Hutter, 2019), Bayesian optimization is seen as a feasible search strategy for the experiments. Besides using a Bayesian optimization search strategy for finding the optimal hyperparameters, hyperparameter search plots that plot the AUC score against the various values of hyperparameters and partial dependence plots are analysed to ensure an optimal set of hyperparameters. These plots are mainly used to check for any outliers and increase the interpretability of the used parameters.

Based on the algorithms mentioned in Section 5.1, every algorithm has a different set of parameters. In order to optimize the parameters, a predefined search space has to be defined. This search space is defined based on previous research, extensions to the search space provided by the software used, and decisions based on the previously mentioned trade-off between performance and resource consumption. However, an extra and final evaluation is implemented through the use of hyperparameter search plots and partial dependence plots. When the plots indicate a possible extension to the search space, the hyperparameter search space is extended even further.

### 5.4.1 Logistic Regression

The first algorithm that is utilized, Logistic Regression, has two main parameters that can be optimized, respectively the inverse of regularization strength ($C$) and the *method of regularization*, which both are depicted in Table 5.3 The method of regularization is used to address overfitting and generally utilizes the L1 regularization technique or the L2 regularization technique. The L1 technique is known for being more robust than the L2 technique, while the L2 regularization technique utilizes a different loss function that is more sensitive to outliers. Both techniques will be evaluated in the hyperparameter search to evaluate the differences and the influence of the outliers on the model.The second parameter, the inverse of regularization strength, a range was chosen on a log-uniform distribution between 0.01 and 100. With this parameter a smaller value means stronger regularization and evaluating various values in a wide search space can potentially be beneficial for the model.

Table 5.3: Defined hyperparameters of Logistic Regression

*Logistic Regression*

| **Parameter** | **Search Space** |
|---|---|
| Method of Regularization | L1, L2 |
| Inverse of regularization strength | 0.01-100 (log-uniform distribution) |

### 5.4.2 Decision Tree and Random Forest

The second algorithm, Decision Tree, has more hyperparameters to tune as seen in Table 5.4. For this algorithm a choice is required on the maximum depth of the tree, which criterion to use to measure the quality of the tree split, how many samples per leaf minimal is and which strategy needs to be used to select the split at each node. The criterion that will be used to measure the quality of a split in a decision can be either Gini or Entropy, whereas the split strategy can be either based on the best split or a random split. Following the research of Kiguchi, Saeed, and Medi (2022), all options are considered to be reasonable hyperparameters and therefore included in the search space. Maximum depth of the tree is defined to have a search space that follows an uniform distribution of [1-32], where 1 is the minimum value and 32 is seen as a feasible option in terms of resource consumption. Lastly, the minimum number of samples per leaf is defined to have a search space that has follows an uniform distribution of [1-50], a search space to be considered sufficient by Mantovani et al. (2018).

Table 5.4: Defined hyperparameters of Decision Tree

*Decision Tree*

| **Parameter** | **Search Space** |
|---|---|
| Criterion | Gini, Entropy |
| Split strategy | Best, Random |
| Maximum depth of tree | [1-32] (uniform distribution) |
| Minimal samples per leaf | [1-50] (uniform distribution) |

Random Forest is an algorithm that combines many decision trees, therefore having similar hyperparameters as the decision tree algorithm. Random Forest has two hyperparameters that are also found within decision, the maximum depth of tree and the minimal samples per leaf. The search space for the maximum depth of tree is reduced to an uniform distribution of [6-20] to reduce the resource consumption, where a higher maximum depth generally increases the predictive performance but substantially increases the computation time. In terms of the minimal samples per leaf, a search space of [1-50] has been defined that follows an uniform distribution, and has the same value as the decision tree algorithm. Lastly, the number of trees has been defined to have a search space of [10-500] which also follows an uniform distribution. All corresponding hyperparameters and search space are depicted in Table 5.5.

Table 5.5: Defined hyperparameters of Random Forest

*Random Forest*

| **Parameter** | **Search Space** |
|---|---|
| Number of Trees | [10-500] (uniform distribution) |
| Maximum depth of tree | [6-20] (uniform distribution) |
| Minimal samples per leaf | [1-50] (uniform distribution) |

### 5.4.3 Gradient Boosting Algorithms

XGBoost, LightGBM, and Gradient Boosted Trees are all algorithms that utilize gradient boosting. XGBoost is an algorithm which includes more than 25 hyperparameters and therefore has a complicated hyperparameter optimization process (Feng, Wang, Yin, Li, & Hu, 2020). Covering the whole process of optimizing every hyperparameter is considered to be out of scope for this research project that mainly focuses on implementing more customer behaviour in the predictive churn models, therefore this section does not cover every hyperparameter. With regards to XGBoost, several hyperparameters are seen as important and are listed in Table 5.6. For the booster hyperparameter both Gradient Boosted Trees and DART will be evaluated, while the maximum number of trees is 500, the maximum as seen within Random Forest. In terms of maximum depth, a search space of [3-20] has been selected for evaluation to reduce the resource consumption. The learning rate has a search space of [0.1-0.5], where the rate is used to adjust the weight of each step in order to improve the robustness. Lastly, the columns subsample ratio for trees and subsample ratio for splits are both set to (0.5-1] and the gamma hyperparameter search space is defined to be [0-1]. Characterised by a relatively fast training speed and low memory usage, the LightGBM algorithm is capable of handling large dataset quite efficiently. As a consequence, the search space for the different hyperparameters can be extended in comparison to the XGBoost algorithm. LightGBM also has a relatively large number of hyperparameters, which is deemed out of scope based on the same reasons as XGBoost. The most important hyperparameters are provided in Table 5.7. For the boosting type, both types will be evaluated. The maximum number of trees is kept constant with the other ensemble learning algorithms, therefore having a search space of [1-500]. The maximum depth of the three has been unconstrained, whereas the learning rate has been defined to be between [0.1-0.5]. Lastly, the number of leaves is defined to be between 20 and 500 and the subsample ratio is kept between 0.5 and 1.0.

Table 5.6: Defined hyperparameters of XGBoost

*XGBoost*

| **Parameter** | **Search Space (distribution)** |
|---|---|
| Booster | Gradient Boosted Trees, DART |
| Maximum number of trees | [1-500] (uniform distribution) |
| Maximum depth of tree | [3-30] (uniform distribution) |
| Learning rate | [0.1-0.5] (uniform distribution) |
| Gamma | [0-1] (uniform distribution) |
| Subsample ratio for trees | [0.5-1] (uniform distribution) |
| Subsample ratio for splits | [0.5-1] (uniform distribution) |

Lastly, the gradient boosted trees algorithm is evaluated. This algorithm has four specific

Table 5.7: Defined hyperparameters of LightGBM

*LightGBM*

| **Parameter** | **Search Space (distribution)** |
|---|---|
| Boosting Type | Gradient Boosting Decision Tree, Gradient One-Side sampling |
| Maximum number of trees | [1-500] (uniform distribution) |
| Maximum depth of tree | Unconstrained |
| Learning rate | [0.1-0.5] (uniform distribution) |
| Number of leaves | [20-500] (uniform distribution) |
| Subsample ratio for trees | [0.5-1] (uniform distribution) |

hyperparameters as seen in Table 5.8. Due to this algorithm being computational expensive and taking a significant amount of time, the search space of this algorithm has been mainly based on the trade-off between performance and computational resource usage. The number of boosting stages has been defined to be between 80 and 300, whereas the search space for the learning rate is [0.05-0.5]. Lastly, both *Deviance* and *Exponential* are used and the search space for the maximum depth of a tree is defined as [3-10].

Table 5.8: Defined hyperparameters of Gradient Boosted Trees

*Gradient Boosted Trees*

| **Parameter** | **Search Space (distribution)** |
|---|---|
| Loss | Deviance, Exponential |
| Learning rate | [0.05-0.5] (uniform distribution) |
| Number of boosting stages | [80-300] (uniform distribution) |
| Maximum depth of tree | [3-10] (uniform distribution) |

## 5.5 Training of models

Analyzing whether the inclusion of behavioural features increases the predictive performance of predictive churn models requires multiple models to fully analyze the impact. Following these requirements, five models have been defined that will be analyzed in Chapter 6. The first model, the *general* model, is a model without any behavioural features included, based on a model approximation of a predictive churn model that the insurance company utilizes. This model mainly functions as a baseline model which can be employed as benchmark for the other models that do include behavioural features in some manner. Besides being a baseline, the other models are also built using the features used within the general model and extended through the inclusion of behavioural features. Secondly, the *model approximation* is a model that contains all the features that an existing model deployed by the insurance company has. Although the model is already deployed by the insurance company, the method of predicting is different in terms of observation window, making this analysis interesting for the insurance company. The *model approximation* contain the same features as seen in the general model, however two extra features are included that measure customer behaviour through the interactions that the customer had with the company. These features are depicted in Table 5.9.

Table 5.9: Behavioural features of model approximation

*Model approximation*

| **Feature** | **Description of feature** |
|---|---|
| aantal_inb_tel | Number of inbound calls in the last 12 months |
| aantal_overig_contact | Total number of application uses, website visits in the last 12 months |

Three new models are proposed that potentially can increase the predictive performance by extensively including customer behaviour. The inclusion of customer behaviour is done through the use of the RFM variables as mentioned in Chapters 2 and 3. *FI-monthly* is a model that extends the measuring of customer behaviour by including new features regarding the frequency and intensity of the interactions. This model distinguishes the interactions per communication channel (e.g. telephone, mobile application, website). FI-monthly utilizes the features as depicted in Table 5.10, where the frequency is measured per month and the intensity is measured as the amount of actions per online session or mobile application use. Following the RFM variables, the recency variable has not been included for the monthly measurement.

Table 5.10: Behavioural features of FI-monthly

*FI-monthly*

| **Feature** | **Description of feature** |
|---|---|
| frequency_telephone | Number of inbound calls in the last month |
| frequency_website | Number of website visits in the last month |
| frequency_app | Number of application uses in the last month |
| intensity_website | Number of pages visited in the last month |
| intensity_app | Number of clicks in the mobile application in the last month |

*RFI-yearly* is the fourth model and utilizes also RFM, however this model extends the observation period compared to the monthly model and aggregates over a whole year. This has mainly been argued while interaction data within an insurance company is sparse and increasing the observation period potentially increases the predictive power of the variables. The behavioural variables included within the RFI-yearly model are depicted in Table 5.11. Here, recency variables are also added which depict the days since last interaction per communication channel. Noteworthy is that while the observation period is defined to be 12 months, the maximum number of days is 365. For instance, when a customer has no interaction via the mobile application, the recency is 365 days while the frequency and intensity feature has a value of 0.

Lastly, a hybrid model is defined as *RFI-normalized*. This model normalizes the monthly values against the year average for frequency and intensity, while comparing the recency against the average days per session. This is depicted in Table 5.12. By comparing the monthly values against the average values, weight is provided to the more recent interactions while normalizing the values against a whole year, possibly indicating changes in the behaviour of the customer. When the lifetime of the policy has been a year or higher, the frequency, intensity, and recency variables are normalized against the mean of the last twelve months. If the policy is less than twelve months old, the lifetime of the policy is first transformed to represent the lifetime as the total share of a year, after which this value is utilized for calculating a year average.

Table 5.11: Behavioural features of RFI-yearly

*RFI-yearly*

| **Feature** | **Description of feature** |
|---|---|
| recency_telephone | Days since last contact with telephone |
| recency_website | Days since last online session |
| recency_app | Days since last application use |
| frequency_telephone_yearly | Number of inbound calls in the last 12 months |
| frequency_website_yearly | Number of website visits in the last 12 months |
| frequency_app_yearly | Number of application uses in the last 12 months |
| intensity_website_yearly | Number of pages visited in the last 12 months |
| intensity_app_yearly | Number of clicks in the mobile application in the last 12 months |

Table 5.12: Behavioural features of RFI-normalized

*RFI-normalized*

| **Feature** | **Description of feature** |
|---|---|
| recency_telephone_norm | Days since last contact with telephone compared against average days per session |
| recency_website_norm | Days since last online session compared against average days per session |
| recency_app_norm | Days since last application use compared against average days per session |
| frequency_telephone_norm | Monthly inbound calls compared against average monthly calls |
| frequency_website_norm | Monthly website visits compared against average monthly visits |
| frequency_app_norm | Monthly app use compared against average monthly frequency |
| intensity_website_norm | Monthly website intensity compared against average monthly intensity |
| intensity_app_norm | Monthly app intensity compared against average monthly intensity |

## 5.6 Evaluation of Models

Evaluating the performance of the various models involve multiple evaluation metrics, of which some already have been given in Chapters 2 and 3. One of the main sources of metric is the *confusion matrix*, which is a table representing the performance of the classifier. Within this 2x2 table, each cell contains the count of how many instances were classified as the category. It helps visualizing and will be used to visualize the performance of the models in a more visual manner. Based on the confusion matrix, various numerical metrics are obtained as described in Chapter 2. The *recall*, *precision*, and $F_1$ *score* aid in the assessing of the predictive performance, whereas other frequently used numerical metrics like accuracy and error rate are less applicable to predictive churn modelling due to the class imbalance (Abd Elrahman & Abraham, 2013; Chawla et al., 2002). These metrics are also displayed in decision chart, showing the various metrics in the context of the threshold. These decision charts aid in the understanding of how the various metrics develop.

The models are also evaluated using the Receiver Operating Characteristic (ROC) curve, which is a standard technique used to assess the predictive performance (Chawla et al., 2002). The ROC curve is a graphical representation incorporating the true positive rate and false positive rate at various thresholds, indicating the predictive performance when utilizing different thresholds. A numerical metric that is obtained from the ROC curve is the area under the curve (AUC), which is considered to be a summary metric for the ROC. By utilizing the AUC, various models can easily be compared and is therefore included as an evaluation metric.

A third evaluation measure is the top-decile lift (TDL) and the corresponding lift charts. The lift charts provide insights in how many of the churned policies are reached when a subsample of the whole dataset is targeted. Lift charts are especially useful for predictive churn models, while the whole population of policies can not be targeted for marketing purposes. In a realistic setting, a subsample is targeted, therefore emphasizing the importance of the lift chart. Based on this lift chart, a top-decile lift can be calculated. The top-decile lift focuses on 10% of the data, more specifically the top 10% of policies who have the highest probability of churning, representing a proper group for targeting purposes (Lemmens & Croux, 2006). This lift is the proportion of churning policies in the top 10% divided by the proportion of churning policies in the complete dataset. Both the lift charts and the top-decile lift will be used for evaluation based on the marketing purpose of this model.

The evaluation of the various features is done through the use of partial dependence plots (PDP) and the feature importance. Both are methods to describe the relationship between input features and model outcome, causing these methods to be important for classifying the performance of the introduced features (Molnar et al., 2021). Partial dependence plots visualize the average effect of a feature on the prediction, where feature importance describes how much each feature is improving the predictive performance of the model. When including both methods, features can be analyzed thoroughly which is necessary in the context of the research objective.

## 5.7 Utilized software

The conducted empirical analyses are conducted through the use of Dataiku, a data science and machine learning platform (Dataiku, 2023). This platform combines multiple packages in a visual tool to reduce the amount of programming. All the data preprocessing is conducted in this program, however for the analyses several Python libraries are used. The open source scikit-learn library is used for most of the machine learning analysis (Pedregosa et al., 2011), however not all analyses could be conducted with this library. Therefore, the use of the *LightGBM* and *XGBoost* library is required to include the algorithms of LightGBM and XGBoost (LightGBM, 2023; XGboost, 2023). Lastly, the library of *imblearn* was utilized to perform the class imbalance analyses (imblearn, 2023).

# Chapter 6

# Results

This chapter provides an overview of the results of the various experiments conducted for the research. Based on the research objectives, by conducting the experiments an attempt is made to find which supervised machine learning algorithm is the most optimal, how different models using behavioural features compare, and how the class imbalance frequently found within the area of customer churn influences the predictive performance of the models. The following experiments have been conducted:

- Comparing various supervised machine learning techniques based on a model approximation utilized by the Dutch insurer.
- Measuring the effect of including more extensive measurements of customer behaviour within predictive churn models as found within the company.
- Measuring the effect of over-sampling, under-sampling or a hybrid sampling strategy on the performance of the model.

The results of these experiments will be described in this chapter in a concise manner, while the main analysis of the results will be performed in Chapter 7. Due to the various experiments that have been conducted, presenting all the elaborate results would be too extensive for a single chapter. Some of the data from the results therefore has been depicted in Appendix A, where the additional results can be analyzed when deemed necessary. The methodology and experimental setup has already been addressed in Chapter 5, however some experimental setup decisions have to be emphasized:

- Several evaluation metrics of the model, the $F_1$ score, precision, and recall metric, are based on a cut-off threshold that is optimized to achieve the most optimal $F_1$ score. The course of the various evaluation metrics along various cut-off thresholds are depicted in the decision charts. The top-decile lift (TDL) does not depend on any threshold and is therefore not depicted in the decision charts.
- The sampling strategy when analyzing which algorithm is the most optimal is simply utilizing the whole dataset. Based on the results of the analysis which sampling strategy provides the best predictive performance, the various models are compared using this sampling strategy.
- If not stated otherwise, the prediction window is a month, meaning that every month the probability of a policy being cancelled in the following month is recalculated.

## 6.1 Comparison of Algorithms

Objective of the first experiment is to compare the various algorithms using the dataset provided by the insurance company. Analysis of the models is conducted using the method and the various models as described in Chapter 5, extensively analyzing various algorithms per model. The functioning of specific algorithms can be dependent on the context of the model, causing a potential difference in the performance among algorithms when comparing models with and without customer behaviour. The impact of the various algorithms on the different models will be analyzed in this section, providing an overview of the performance of each algorithm within the context of three different models. Although the metrics used to compare the various algorithms already entails information regarding the overall performance of the models, Section 6.2 will focus on extensively analyzing the effect of incorporating more behavioural features within the predictive model. This section focuses on comparing the performance of the various algorithms within different model contexts. All the models are trained on the whole available dataset, implying that the training has be performed on an imbalanced dataset.

### 6.1.1 General Model

The first analyzed model is the model that operates as benchmark within the model comparison, the *general model*. This model contains no behavioural features and contains the features as found within the predictive model applied by the insurance company. As can be seen in Table 6.1, the ensemble algorithms involving boosting (*Gradient Boosted Trees, LightGBM, XGBoost*) have higher values compared to other models when looking at the AUC metric with values close to 0.750. Regarding the $F_1$ score, it is observed in the table that the boosting methods show the highest values together with the Decision Tree algorithm. The relatively high $F_1$ score of Decision Tree can largely be attributed to the high precision of the Decision Tree model, which has the highest value for precision among all algorithms. Lastly, the recall metric is the highest when using the Random Forest algorithm, however at the cost of having a low precision.

When assessing the overall performance of the algorithms within the general model by considering the five metrics altogether, it is observed that the ensemble methods utilizing boosting perform slightly better than other models, while being closely followed by the Decision Tree model in terms of predictive performance. The boosting models are the three highest scoring models in terms of AUC, $F_1$ score, and top-decile lift, while only being surpassed by one model in terms of precision and recall. Among the boosting models, LightGBM scores on average the highest by displaying the highest value in terms of AUC, precision, and top-decile lift.

Table 6.1: Results of General Model

| **Algorithm** | **AUC** | **$F_1$ score** | **Precision** | **Recall** | **TDL** | **Threshold** |
|---|---|---|---|---|---|---|
| Decision Tree | 0.732 | 0.149 | **0.218** | 0.113 | 3.695 | 0.900 |
| Gradient Boosted Trees | 0.749 | 0.153 | 0.159 | 0.146 | 3.797 | 0.075 |
| LightGBM | **0.753** | 0.152 | 0.185 | 0.128 | **3.854** | 0.850 |
| Logistic Regression | 0.725 | 0.095 | 0.073 | 0.137 | 3.394 | 0.775 |
| Random Forest | 0.735 | 0.097 | 0.072 | **0.150** | 3.404 | 0.650 |
| XGBoost | 0.751 | **0.155** | 0.178 | 0.137 | 3.836 | 0.850 |

Another noteworthy aspect within the analysis of the algorithms, is the threshold found related to the use of the Gradient Boosted Trees algorithm. The significance of this threshold can be observed within the decision chart as depicted in Figure 6.1 or Figure A.1, where the orange line corresponding to the Gradient Boosted Trees algorithm immediately increases for both the precision

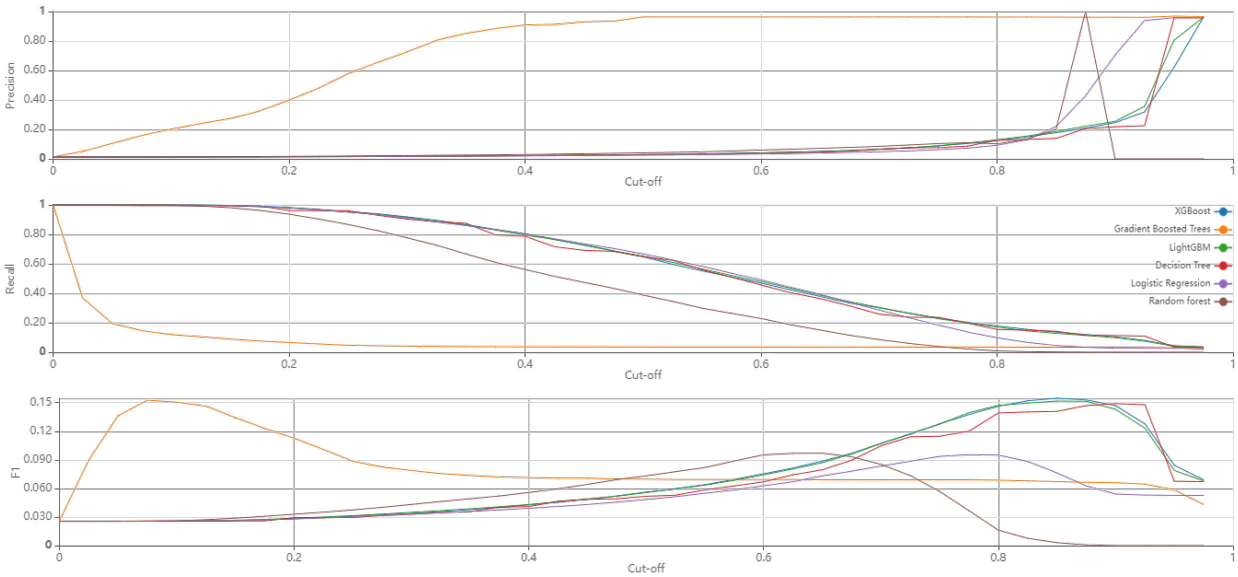

Figure 6.1: Decision chart general model

and $F_1$ metric with the $F_1$ showing a substantial steepness. Furthermore, it can be observed that the other models relatively follow the same trajectory where the precision steeply increases when reaching a cut-off point of 0.8, whereas the recall is gradually declining among various cut-off percentages starting at a cut-off threshold of 0.2. A second distinct trajectory is observed for the model utilizing the Random Forest algorithm. The model has a zero value for precision, recall and $F_1$ when reaching the 0.9 threshold as can be observed in the decision chart. This indicates that the random forest predicts no observation having a higher probability of churning than 0.875, causing the various metrics to become zero at a cut-off of 0.9.

### 6.1.2 Model Approximation

Following the analysis of the general model without any behavioural features, the second model utilizes two behavioural features that are currently found within the model of the Dutch insurance company as depicted in Table 5.9. When analyzing the performance among AUC, $F_1$, precision/recall, and the top-decile lift, the boosting algorithms perform relatively well as can be seen in Table 6.2. The boosting algorithms account for the highest scores for the AUC, $F_1$ score, and top-decile lift. The LightGBM model is only surpassed in terms of recall by Random Forest, scoring the highest value across all other metrics. Within the context of the model approximation, the relatively simple algorithm of logistic regression performs worse than the more complex algorithms considered. Logistic regression scores lowest on all metrics except for precision and recall, emphasizing the potential benefit of utilizing more complex algorithms. However, when looking at another relatively simple algorithm like Decision Tree, the $F_1$ score and precision are among the highest of all algorithms. Although algorithms like XGBoost and LightGBM still score higher for these metrics, the relatively small difference is noteworthy while the interpretability of decision tree is significantly better than for ensemble learning algorithms. The ensemble learning algorithm of Random Forest has the highest value for the recall metric, causing the algorithm to have the highest proportion of correct predictions among cancelled policies.

When looking at the best performing algorithms in terms of all metrics, the boosting algorithms, the differences are relatively small in terms of $F_1$ score and AUC, indicating a similar performance on the testing set. The precision metric of Gradient Boosted Trees is however lower

Table 6.2: Results of Model Approximation

| Algorithm | AUC | $F_1$ score | Precision | Recall | TDL | Threshold |
|---|---|---|---|---|---|---|
| Decision Tree | 0.732 | 0.149 | 0.214 | 0.114 | 3.684 | 0.900 |
| Gradient Boosted Trees | 0.751 | 0.153 | 0.161 | 0.146 | 3.818 | 0.075 |
| LightGBM | **0.755** | **0.154** | **0.217** | 0.119 | **3.850** | 0.875 |
| Logistic Regression | 0.728 | 0.095 | 0.070 | 0.146 | 3.422 | 0.775 |
| Random Forest | 0.737 | 0.099 | 0.068 | **0.182** | 3.402 | 0.625 |
| XGBoost | 0.751 | 0.151 | 0.179 | 0.131 | 3.807 | 0.850 |

compared to XGBoost and LightGBM, meaning that the Gradient Boosted Trees algorithm has a bit more difficulties with correctly predicting the policies that have been cancelled. Overall, the boosting algorithms again score highest on almost all considered metrics.

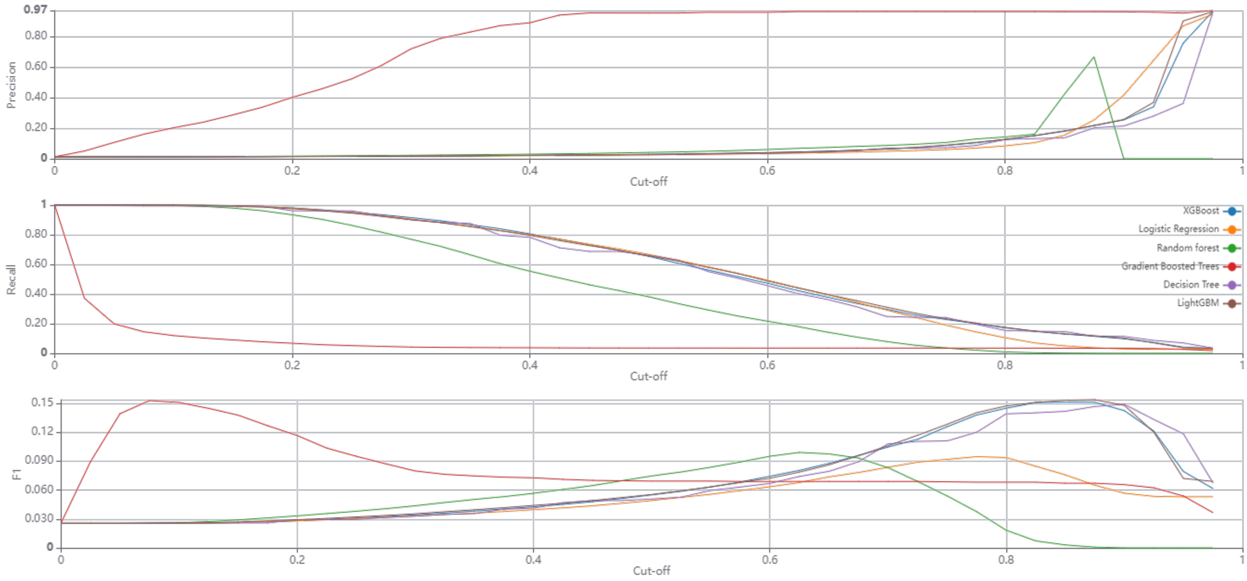

Figure 6.2: Decision chart Model Approximation

The thresholds of the various algorithms within the context of the model approximation show similar results as the general model, having a relatively high threshold for all algorithms except for Gradient Boosted Trees and Random Forest. As can be observed in Figure 6.2 or Figure A.2, the Gradient Boosted Trees algorithm has the optimal $F_1$ score when having a low threshold due to the steepness of the recall variable. Although the optimal threshold is found to be low, the values for precision, recall and $F_1$ remain fairly constant over the various thresholds. Looking at the model involving the Random Forest algorithm, a threshold of 0.625 is observed. This indicates that the model assigns lower probabilities than the other algorithms, which is also seen in the decision chart. The model has zero values for all metrics when reaching the threshold of 0.9, predicting that no observation has a higher churn probability than 0.875. For the other algorithms, it is observed that the focal point is more towards the higher thresholds. The trajectories for both the $F_1$ metric and the recall metric are fairly gradual, whereas the precision metric steeply increases with a probability threshold of 0.8 or higher. The precision chart for Logistic Regression is observed to be less steep than the other models, explaining the lower $F_1$ score. Furthermore, the tresholds at approximately 0.8 indicates that the models are not as strict as the Gradient Boosted Trees or Random Forest, which already have the thresholds at respectively 0.075 and 0.625.

### 6.1.3 FI-monthly

The third model, utilizing both the intensity and the frequency of interactions, is considered to be an extension to the model approximation. This model entails more information per interaction channel and also captures information regarding the content of the interaction, resulting in the metrics as depicted in Table 6.3. Within the third model, the Decision Tree algorithm scores the highest on the precision metric while having the lowest recall of all models. This leads to a $F_1$ metric of 0.150, which is among the highest across the models. However, the interpretability as said before is high for decision tree, causing this algorithm to be a feasible option for a slightly less optimal performance than the boosting algorithms. The AUC however is also the lowest, causing the boosting algorithms to score higher on average compared to the decision tree.

The boosting algorithms perform the highest in terms of AUC, $F_1$ and top-decile lift, while showing relatively high precision/recall values compared to other algorithms. Although being surpasses by Random Forest in terms of recall and Decision Tree in terms of precision, the boosting algorithms on average perform the highest. This is in line with the other two models, showing the same results. Logistic Regression has a low $F_1$ score mainly due its low precision value, causing this benchmark algorithm to be less optimal when analyzing on $F_1$ and precision. In terms of the Random Forest algorithm, the same can be said however the recall is the highest of all algorithm, showing a relatively good performance in the number of correctly predicted cancelled policies.

Table 6.3: Results of FI-monthly

| **Algorithm** | **AUC** | **$F_1$ score** | **Precision** | **Recall** | **TDL** | **Threshold** |
|---|---|---|---|---|---|---|
| Decision Tree | 0.731 | 0.150 | **0.219** | 0.114 | 3.857 | 0.875 |
| Gradient Boosted Trees | 0.757 | 0.151 | 0.190 | 0.125 | 3.915 | 0.100 |
| LightGBM | **0.763** | **0.156** | 0.188 | 0.133 | **3.946** | 0.850 |
| Logistic Regression | 0.733 | 0.097 | 0.081 | 0.121 | 3.485 | 0.800 |
| Random Forest | 0.744 | 0.100 | 0.068 | **0.192** | 3.506 | 0.625 |
| XGBoost | 0.759 | 0.152 | 0.215 | 0.118 | 3.852 | 0.875 |

Looking at the decision chart for the FI-monthly model as seen in Figure 6.3 or Figure A.3, the trajectory of the Gradient Boosted Trees is relatively steep compared to the other algorithms with both the recall and $F_1$ metric. However, the precision is more gradually increasing among various thresholds, where the other algorithms are seen as steep. Within the Random Forest algorithm trajectory of the $F_1$ score, it can be observed that the $F_1$-score at a threshold of 0.9 becomes 0 and remains 0 for the higher thresholds. This has also been the case for the first two models, while other algorithms do not show this effect. The models utilizing Gradient Boosted Trees and Random Forest do predict churn probabilities in a different manner than the other models, assigning lower probabilities.

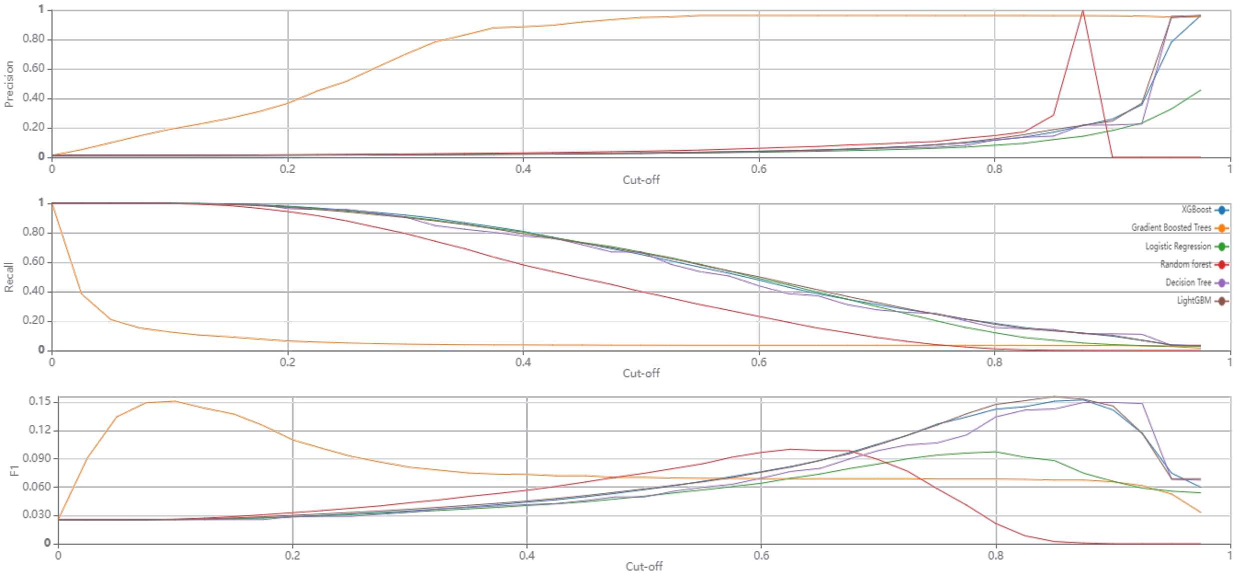

Figure 6.3: Decision chart FI-monthly model

## 6.2 Analysis of Models

Following the analysis of the algorithms, the various models are compared based on the various metric to determine the effect of including the behavioural features to the models. The results of the LightGBM algorithm are depicted in Table 6.4. This algorithm scored the highest in terms of AUC and top-decile lift and was generally among the top algorithms when comparing based on $F_1$ score and precision/recall. Within the results of the various models, similar AUC values are found for the FI-monthly, RFI-normalized, and RFI-yearly model, with RFI-yearly reporting the highest AUC of 0.766. The other two models, the general model and the model approximation, have respectively an AUC value of 0.753 and 0.755. Subsequently, the inclusion of yearly behavioural variables improves the AUC value of the model with 1.46% compared to a model without any behaviour.

In terms of the $F_1$ score, minor differences are observed with the FI-monthly model having the highest value of 0.156. The lowest value for the $F_1$ score is seen within the general model with a value of 0.152, causing the maximum difference to be 0.004. Therefore, all models score fairly similar in terms of $F_1$, indicating that the harmonic mean of the precision and recall is comparable among the various models. The precision metric is observed to be more diverse, reporting the highest value in the model approximation with a value of 0.217. The model as used by the Dutch insurer therefore has the highest share of actual cancelled policies in their churn predictions, with RFI-yearly being the second highest with a precision of 0.204. The percentage of correctly predicted cancelled policies amongst the total number of predicted cancelled policies is 1.3% higher for the model approximation, being slightly more precise. The models of FI-monthly and general model have a similar precision value of respectively 0.188 and 0.185, being slightly less that the RFI-normalized precision of 0.195. The recall, the metric that indicates how many cancelled policies have been predicted as cancelled, is observed to be the highest in the FI-monthly model with a value of 0.133. Three models, the general model, RFI-normalized, and RFI-yearly report a recall value in the range of $[0.126 - 0.128]$, displaying the relative similarity between the models in terms of recall. The model approximation is observed to have a value of 0.119, displaying the lowest number of correctly predicted actual cancelled policies.

Table 6.4: Results of models with LightGBM algorithm

| Model | AUC | $F_1$ score | Precision | Recall | TDL | Threshold |
|---|---|---|---|---|---|---|
| FI-monthly | 0.763 | **0.156** | 0.188 | **0.133** | 3.946 | 0.850 |
| General Model | 0.753 | 0.152 | 0.185 | 0.128 | 3.854 | 0.850 |
| Model Approximation | 0.755 | 0.154 | **0.217** | 0.119 | 3.850 | 0.875 |
| RFI-normalized | 0.764 | 0.154 | 0.195 | 0.127 | **3.950** | 0.850 |
| RFI-yearly | **0.766** | 0.155 | 0.204 | 0.126 | **3.950** | 0.850 |

The top-decile lift is better in the models of FI-monthly, RFI-yearly, and RFI-normalized than in the general model and model approximation, performing better when aiming for 10% of the data that have the highest probability of churning. The FI-monthly, RFI-yearly, and RFI-normalized models report a top-decile lift of respectively 3.946 and 3.950, indicating that these models find approximately 39.5% of the actual cancelled policies in the dataset. That is higher the models with a top-decile lift of 3.850 and 3.854 which find approximately 38.5% of the actual cancelled policies in the dataset. This is visualized in the graphical representation known as the lift chart, where on the horizontal axis the percentage of the targeted population is shown and on the vertical axis the percentage of found positive records. Looking at the lift chart as depicted in Figure 6.4 or Figure A.4, a steep lift chart in the beginning is observed while becoming more gradual when increasing the population targeted. Targeting 10% of the population with the highest probability of churning, the top-decile lift, provides the previously mentioned TDL values. The main advantage of the model is observed in the first decile based on the steepness of the curve, however added value is still observed for the other deciles compared to a random model. The various models are close in terms of the performance when looking at the lift chart, however differences can be observed when analyzing the individual lift charts and corresponding per-bin lift charts.

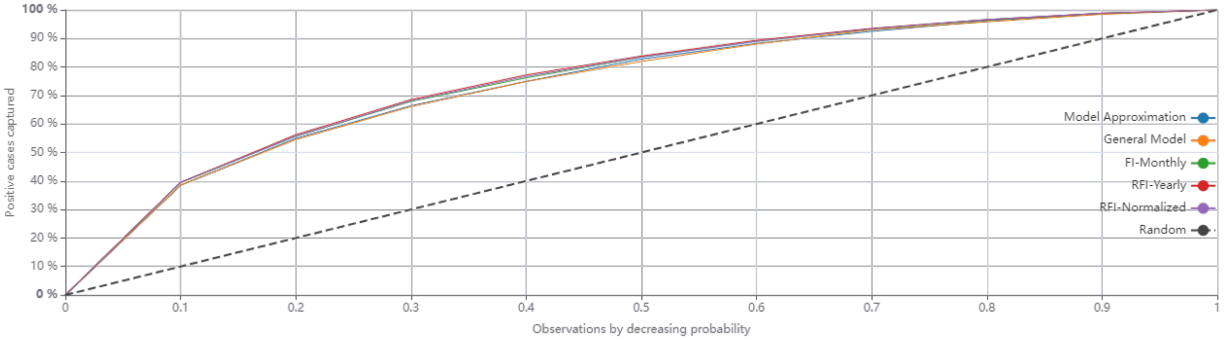

Figure 6.4: Lift chart LightGBM

Comparing the results of the various models is also done through the analysis of the ROC curve, where a steeper curve is a potential indication for a better predictive performance. When looking at the various models in the ROC curve as depicted in Figure 6.5 or Figure A.5, the models overall follow the similar trajectories with minor differences. The differences between the various models are more delicate than easily seen within the ROC curve, addressing the need for more detailed analysis of the various models which is conducted in the following subsections.

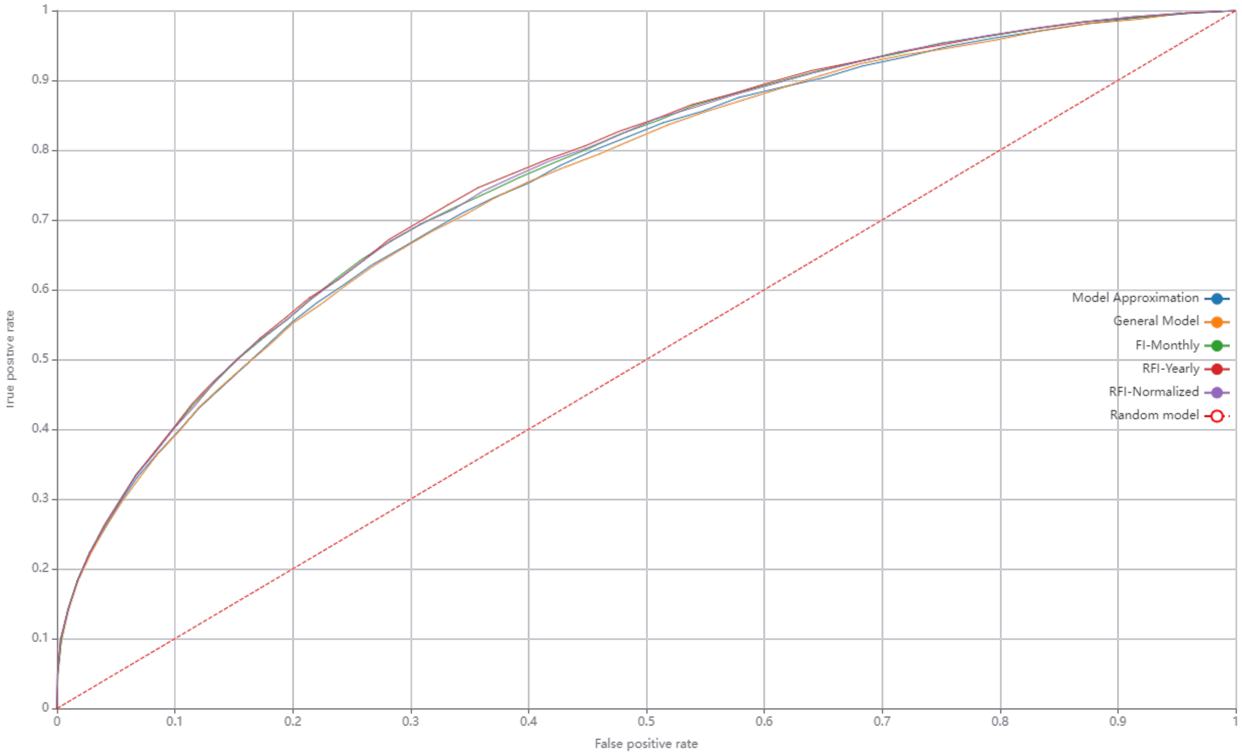

Figure 6.5: ROC curve of all models

### 6.2.1 General model

Analyzing the models requires the use of a general model as a benchmark which does not incorporate any behavioural feature. Analyzing the confusion matrix as showed in Table 6.5, the amount of non-churners is highly distinct in the confusion matrix which can be attributed to the class imbalance found in the dataset. Besides indicating the class imbalance, the confusion matrix also displays the relatively low value of precision and recall. The challenge of detecting whether a policy is going to get cancelled in the coming month is difficult due to the amount of non-cancelled policies in the dataset, introducing bias in the models to mainly classify as non-cancelled policies. This motivated the experiment described in Section 6.3. In the confusion matrix it is observed that the model predicts 653 cancelled policies correctly, while predicting 4438 policies as non-churning while the policies have actually been cancelled. Subsequently, the recall is calculated to be 0.128. When looking at the 3521 policies predicted as churn, 653 of these policies were actually cancelled while the other 2868 had not been cancelled in the following month. The precision is calculated to be 0.185, causing the precision to have a higher value than the recall.

Table 6.5: Confusion matrix of general model

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 653 | 4438 | 5091 |
| | Non-churn | 2868 | 385119 | 387987 |
| | Total | 3521 | 389557 | 393078 |

Looking at the importance of the various features considered by the model, the results are shown in Table 6.6. Within this model, the main features of importance are the discount provided for the policy, duration of the policy, information regarding the customer and the specific policy. The most important feature is considered to be the discount, which is policy specific and also specific for the company, causing the price to be higher or lower compared to other companies. Analyzing the top 10 most important features, eight features are related to the policy or the context of the policy for which the churning probability is estimated, whereas the other two variables are related to the customer.

Table 6.6: Feature importance of general model

| **Feature** | **Feature description** | **Feature importance (%)** |
|---|---|---|
| Policy feature 1 | Provided discount | 25% |
| Policy feature 2 | Duration of policy | 14% |
| Policy feature 3 | Content of policy | 8% |
| Customer feature 1 | Customer lifetime | 7% |
| Policy feature 4 | Age of insured property | 7% |
| Policy feature 5 | Age | 6% |
| Policy feature 6 | Property of car | 5% |
| Policy feature 7 | Property of car | 5% |
| Customer feature 2 | Customer policies | 2% |
| Policy feature 8 | Property of car | 2% |

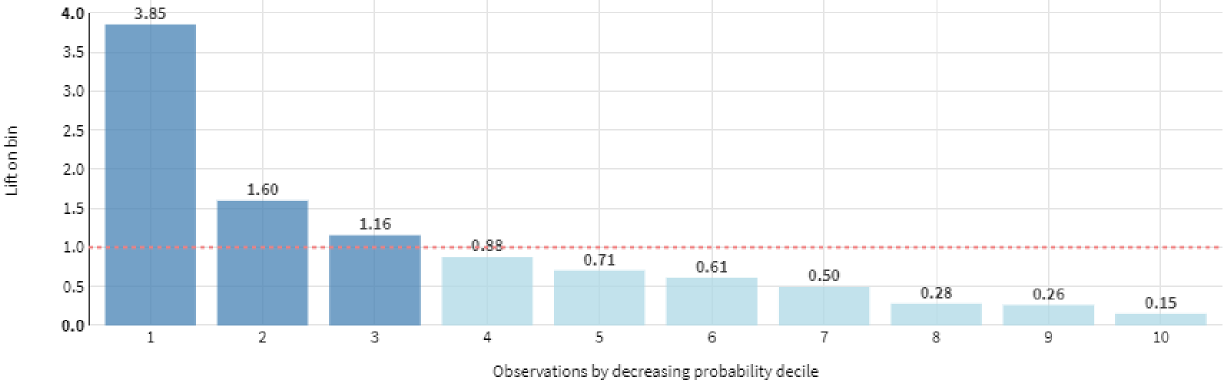

Figure 6.6: Lift chart General Model

Generally, the general model showed significant better results in the lift chart as seen in Figure 6.4 compared to a random classifier. Extending the analysis by sorting the observations per decile, the lift for each bin can be calculated as shown in Figure 6.6. The lift chart visualizes the benefits of targeting a subset of the population, which is of high importance for the marketing purpose of the applied churn models. Within the lift chart the red line depicts a random model, emphasizing the results for the first three deciles. Focusing on the first decile, containing 10% of the observations which have the highest probability of churning, targeting with the general model would yield 3.85 times as many positive results as utilizing random sampling. When targeting the second decile, the effect already decreases to 1.60 times compared to random sampling, causing the targeting of the top 20% to be 2.74 times better than a random model. The lift drops even further to 1.16 when targeting the third decile, capturing 66.1% of the cancelled polices compared to 30%

when using a random model. Further targeting would lead to results lower than random sampling, therefore the main benefit of the model is observed in the first three deciles.

### 6.2.2 Model Approximation

Predicting customer churn through the use of the model approximation lead to the following predictions as shown in Table 6.8. This model predicts in total 606 actual cancelled policies as churn, predicting in absolute terms less churned policies correctly than the general model. The recall is therefore 0.119, being slightly lower than the general model. The precision of the general model is however 16.9% higher with a value of 0.217, decreasing the degree of false prediction of cancelled policies as non-cancelled.Furthermore, the difficulties seen within the general model are also identified for the model approximation, showcasing a distinct bias towards predicting policies as not cancelled..

Table 6.7: Confusion matrix of Model Approximation

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 606 | 4485 | 5091 |
| | Non-churn | 2189 | 385789 | 387987 |
| | Total | 2795 | 390283 | 393078 |

Table 6.8: Confusion matrix of Model Approximation

Table 6.9: Feature importance of Model Approximation

| **Feature** | **Feature description** | **Feature importance (%)** |
|---|---|---|
| Policy feature 1 | Provided discount | 28% |
| Policy feature 2 | Duration of policy | 12% |
| Policy feature 3 | Age of insured property | 12% |
| Policy feature 4 | Content of policy | 10% |
| Customer feature 1 | Customer lifetime | 8% |
| Policy feature 5 | Property of car | 6% |
| Policy feature 6 | Age | 6% |
| Policy feature 7 | Property of car | 2% |
| aantal_inb_tel | Number of inbound calls | 2% |
| aantal_overig_contact | Number of online/app interactions | 1% |

Regarding the various features and its corresponding features in Table 6.9, similar results to the general model are seen within the model approximation, with one major difference with regards to the behavioural features. The top 10 most important features consists out of seven policy features, two behavioural features and one customer feature, reducing the amount of customer-related features in favour of the behavioural features. Looking specifically at the behavioural features, it is observed that the number of inbound calls in the last twelve months is more important than the number of online and app interactions, having respectively a feature importance of 2% and 1%. Computing both partial dependence plots on the whole training set, similar results are found when analyzing the trajectories of the partial dependence in Figure 6.7 and Figure 6.8. Both trajectories follow a steep curve, where having none inbound calls or other interactions has a minor reducing effect on the probability of churn. This effect changes when the number of interactions increases, while the probability of a customer churning steeply increases when the number of interactions increases. Having multiple interactions therefore increases the chance of a customer churning according to the model approximation. The distribution of the customers having said interactions is however small, reducing the overall effect of the features.

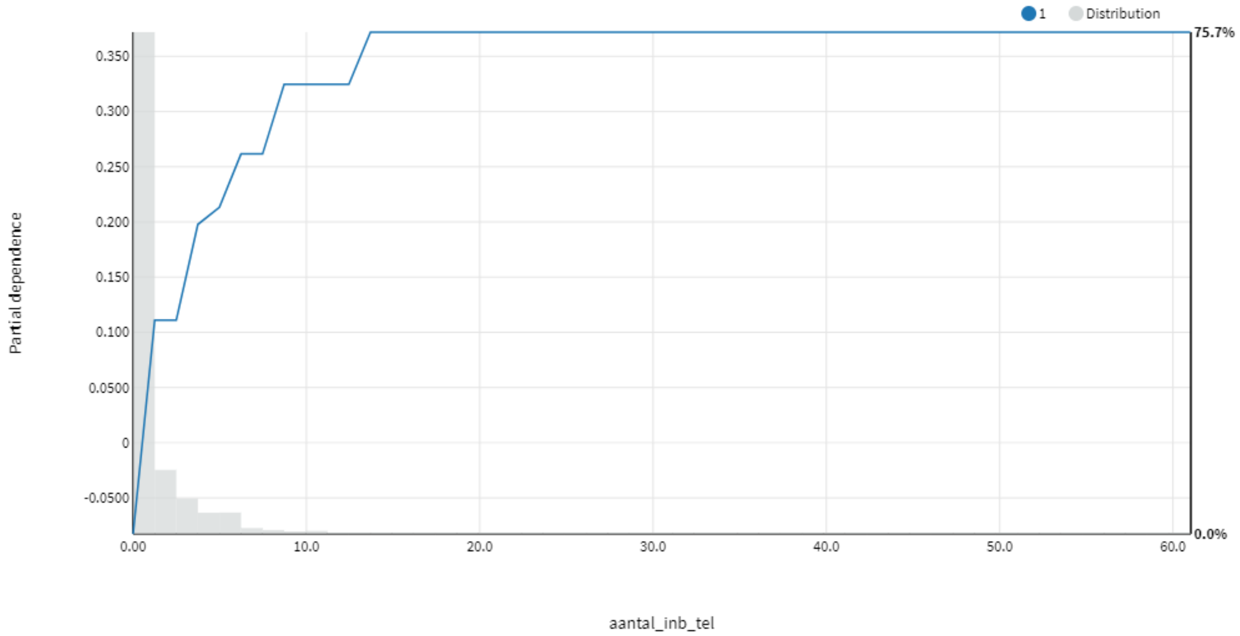

Figure 6.7: Partial dependence plot of aantal_inb_tel

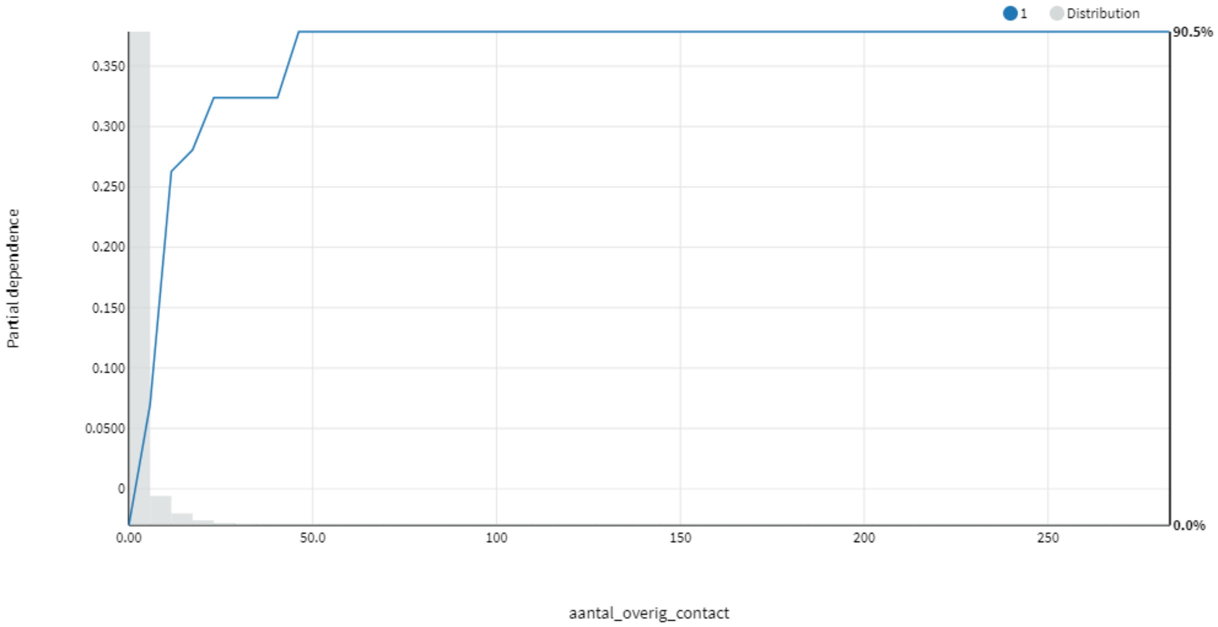

Figure 6.8: Partial dependence plot of aantal_overig_contact

Analyzing the model approximation through the lift chart in Figure 6.9, a similar lift value of 3.85 is observed for the first decile when compared to the general model. Prediction of customer churn through the use of the model approximation is therefore equal to the model without any behavioural features for the 10% of the policies that have the highest chance of being cancelled by the customer. Predicting the churn within the second decile shows a slightly better performance with a lift of 1.64 compared to a value of 1.60 for the general model. The lift within the third decile is 0.02 lower than the general model, having a value of 1.14. This model performs slightly better than the general model when focusing on the first two deciles, with 54.9% of the cancelled policies captured against 54.5%. In general, the lift is better for the first three deciles compared to a random model, after which the predictive performance is levelled out.

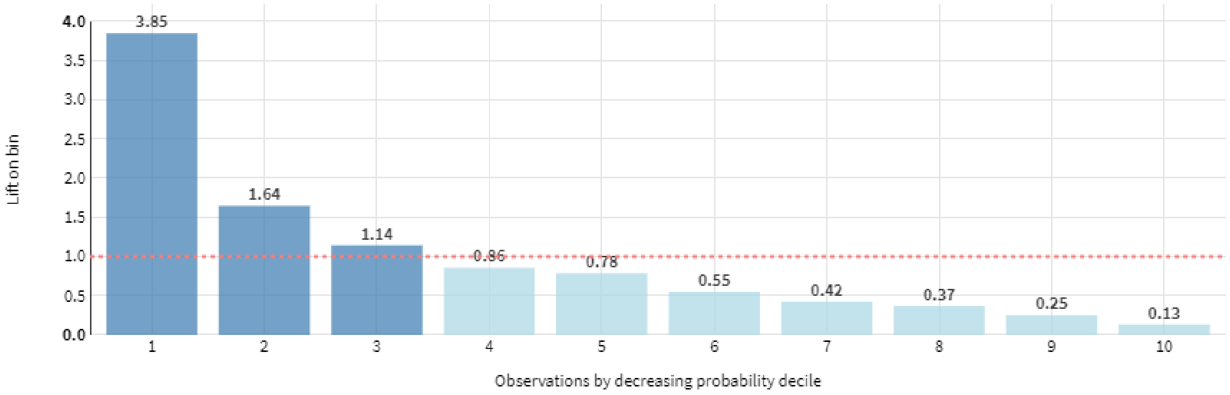

Figure 6.9: Lift chart of model approximation

### 6.2.3 FI-monthly

FI-monthly is the model which extended the behavioural features by separating each interaction channel in a specific feature and also including the intensity of the interactions, while measuring the features over a month. Prediction utilizing the FI-monthly model provided the results of Table 6.10. The model predicts 678 cancelled policies correctly, a difference of 25 with the general model and 71 policies with the model approximation. Furthermore, the FI-monthly model predicts 4413 churn incorrectly as non-churning, which is less than the general model and the model approximation with respectively 4438 and 4485 wrongly classified real cancelled policies. The proportion of correct prediction among the cancelled policies group is 13.3%, causing the recall to be the highest of all models. The precision is calculated to be 18.8%, which is higher than the general model but lower than the model approximation. The combined score found using the $F_1$ score is 0.156, the highest of all models. Just like the other models, difficulties are observed while the model is significantly biased towards predicting non-cancelled policies.

Table 6.10: Confusion matrix of FI-Monthly

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 678 | 4413 | 5091 |
| | Non-churn | 2923 | 385064 | 387987 |
| | Total | 3601 | 389477 | 393078 |

Table 6.11: Feature importance of FI-monthly

| Feature | Feature description | Feature importance (%) |
|---|---|---|
| Policy feature 1 | Provided discount | 23% |
| Policy feature 2 | Duration of policy | 11% |
| Policy feature 3 | Age of insured property | 10% |
| Policy feature 4 | Content of policy | 8% |
| Customer feature 1 | Customer lifetime | 8% |
| Policy feature 5 | Property of car | 5% |
| Policy feature 6 | Age | 4% |
| online_acties | Number of online websites visited in the month | 3% |
| telefoon_gebruik | Number of inbound calls in the month | 3% |
| Policy feature 7 | Property of car | 2% |

Feature importance of the FI-monthly model is depicted in Table 6.11, having the same top 3 as the general model and the model approximation. The total number of behavioural features within this model is two, while having seven policy-related variables and one customer-related variable. Of the two behavioural features, the intensity of the website sites visited is considered more important with a feature importance of 4% whereas the number of inbound calls is considered to have a feature importance of 3%. Although significant differences are observed between the top 3 features and the behavioural features, inclusion in the top 10 most important features shows the contribution to the predictive performance of the model. Other behavioural features are not seen within the top 10 most important features of the model, an indication of the minor contribution to the predictive performance of this model.

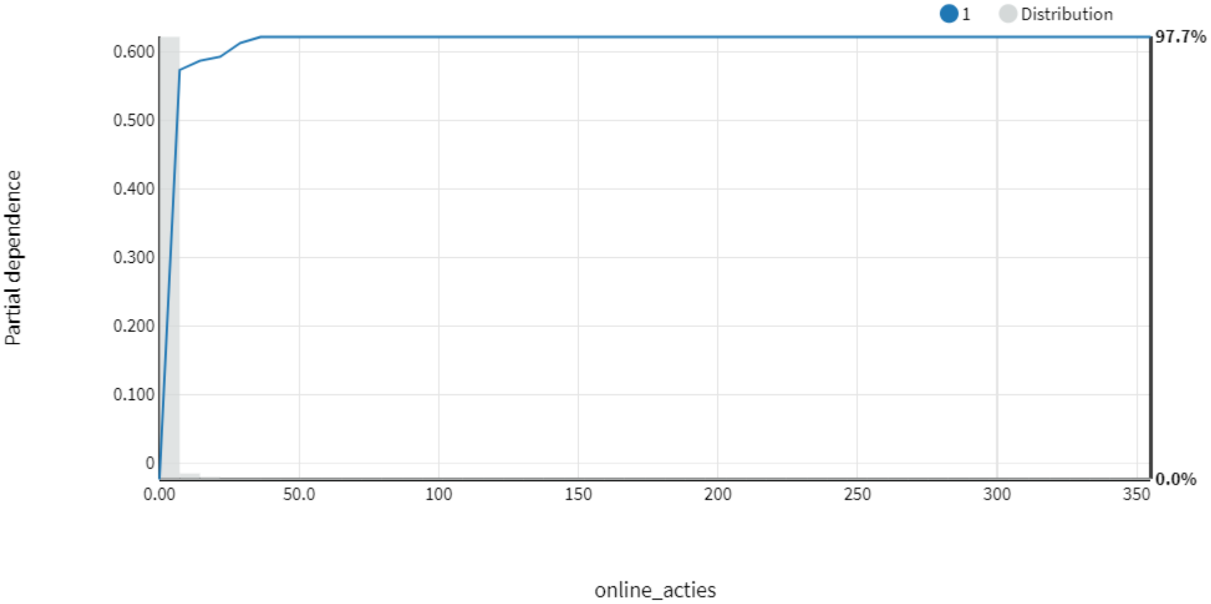

Figure 6.10: Partial dependence plot of online_acties

Looking at the partial dependence plots of the two most important behavioural features, the monthly number of site visits aggregated over the various sessions has a steep increase in the effect on the churn probability as seen in Figure 6.10. Having more than 1 website visit steeply increases the churn probability, however the effect levels out and remains fairly constant after 10 website visits. The steep increase between zero and two website visits indicates that difference between no website visits and at least one website visit is the most important difference in frequency. For the other important behavioural feature regarding the frequency of calls per month, the partial dependence plot is given in Figure 6.11. Here, two steep increases are observed with the highest increase seen between zero calls and one call, indicating the importance of having at least one phone call. However, a second increase is seen between three and four which increases the partial dependence with 0.150, far smaller than the first increase of 0.650. Furthermore, the curve remains fairly constant among various numbers of phone calls.

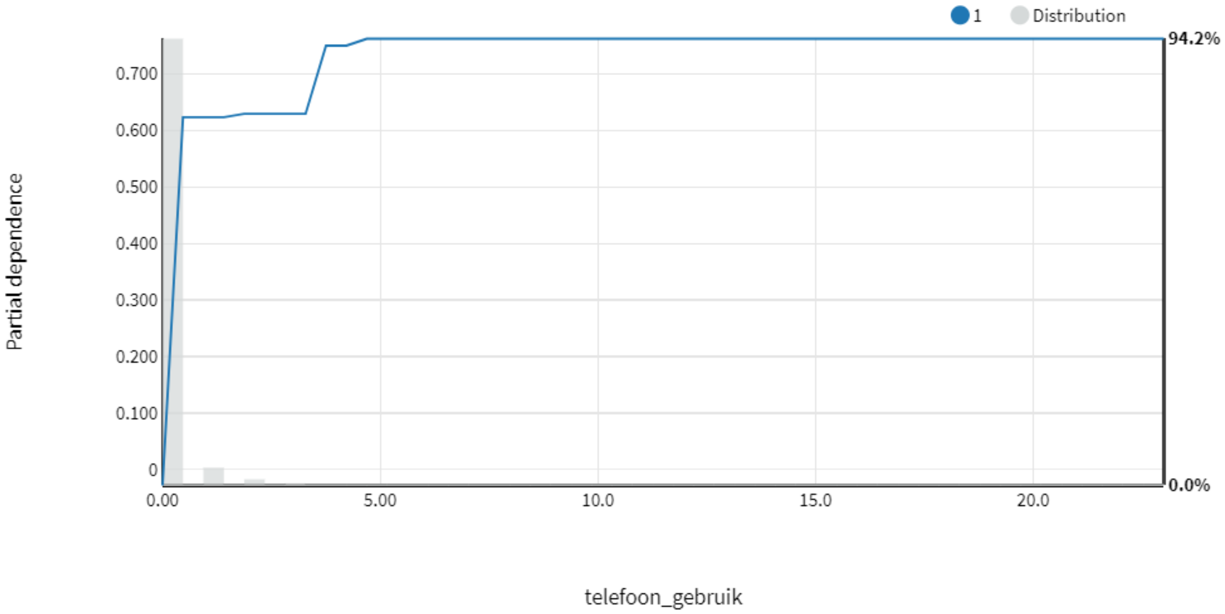

Figure 6.11: Partial dependence plot of telefoon_gebruik

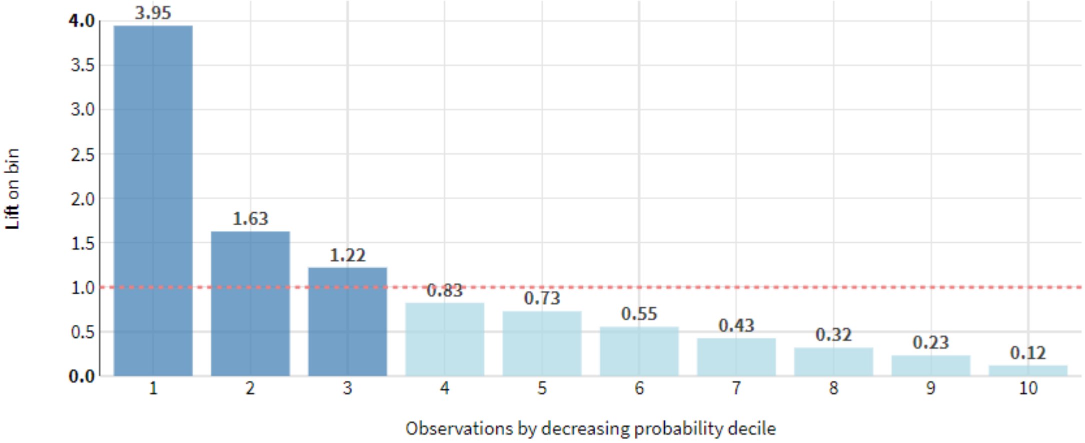

Figure 6.12: Lift chart of FI-monthly

Looking at the lift chart in Figure 6.12, the first three deciles are performing better compared to the general model and model approximation, with values of respectively 3.95, 1.63, and 1.22. Furthermore, it performs better among the three deciles than the model approximation, showing an increase in the predictive performance. Targeting 10% of the policies that have the highest churning probabilities using the model would provide 3.95 times as many positive results as a random sampling, while focusing on all three deciles would provide a population of actual churners targeted that is 2.22 times bigger than a random model.

### 6.2.4 RFI-yearly

Predictions of one of the most extensive models with regards to behavioural features, the RFI-yearly model, are shown in Table 6.12. The RFI-yearly model predicts 639 cancelled policies correctly as churned while predicting 4452 policies as non-churned, causing the recall to be 0.126. This is less than the FI-monthly and general model, but higher than the model approximation. The RFI-yearly model surpasses most models in terms of precision with 639 correctly predicted policies out of 3137 policies predicted as churn, causing the precision to be the second highest after the model approximation with a value of 0.204.

Table 6.12: Confusion matrix of RFI-yearly

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 639 | 4452 | 5091 |
| | Non-churn | 2498 | 385489 | 387987 |
| | Total | 3137 | 389941 | 393078 |

Table 6.13: Feature importance of RFI-yearly

| **Feature** | **Feature description** | **Feature importance (%)** |
|---|---|---|
| Policy feature 1 | Provided discount | 19% |
| Policy feature 2 | Age of insured property | 9% |
| Policy feature 3 | Duration of policy | 9% |
| Customer feature 1 | Customer lifetime | 8% |
| online_recent_dagen | Number of days since last online interaction | 7% |
| Policy feature 4 | Content of policy | 7% |
| telefoon_recent_dagen | Number of days since last phone call | 5% |
| Policy feature 5 | Age | 5% |
| Policy feature 6 | Property of car | 4% |
| app_recent_dagen | Number of days since last app interaction | 2% |

Analyzing the feature importance of the features used within the model shows that three behavioural features, six policy-related features, and one customer-related feature are the most important features of the model. The most important variables are nearly identical for the other models, showcasing the importance for the percentage of discount, age of the car and the current duration of the policy. When looking at the differences, three behavioural features are present. The three behavioural features are all related to the recency variable, defined as the number of days since the last interaction online. *Online_recent_dagen* is the most important behavioural variable with a variable importance of 7%, followed by the variable *telefoon_recent_dagen*, which is the number of days since the last interaction by phone with a variable importance of 5%. Lastly, the number of days since last app interaction is reported to have a variable importance value of 2%, being the lowest of the recency variables.

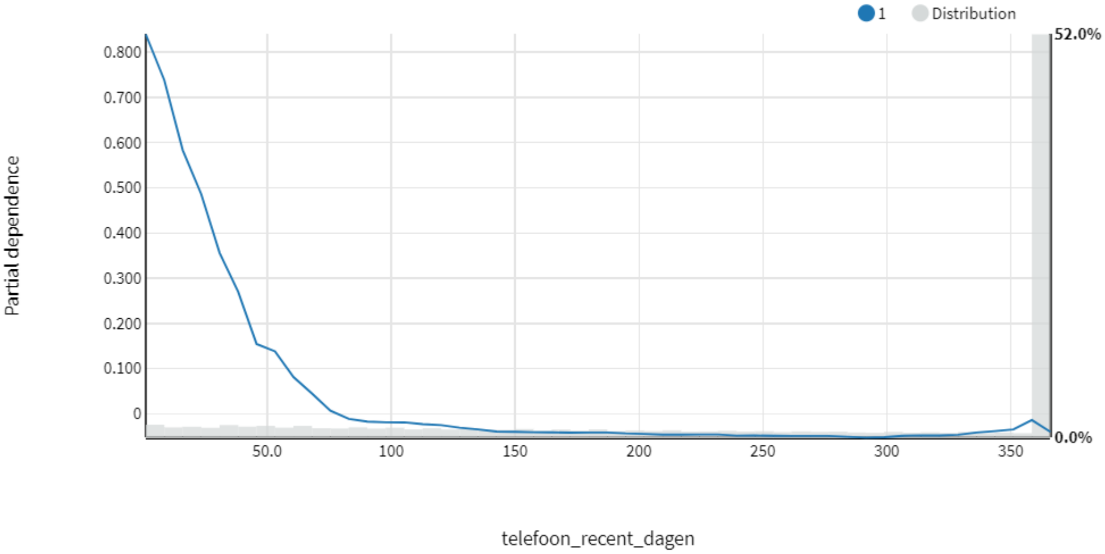

Figure 6.13: Partial dependence plot of telefoon_recent_dagen

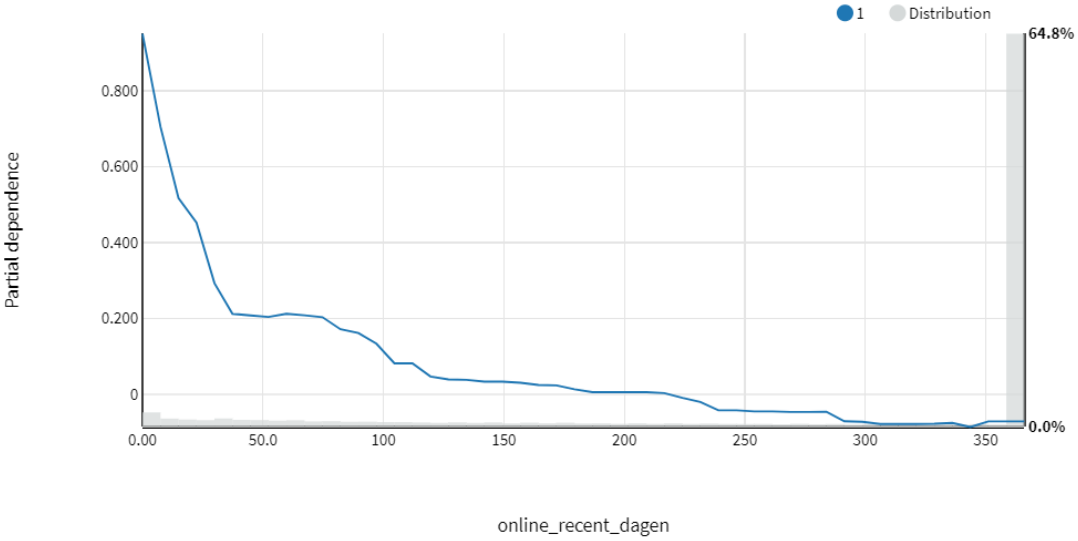

Figure 6.14: Partial dependence plot of online_recent_dagen

Further analyzing the various behavioural features, the partial dependence plots are considered. The partial dependence plot of *telefoon_recent_dagen* is shown in Figure 6.13, showing a relatively steep downwards sloping line. Having more inactive days decreases the chance of churning, where a recent interaction provides an indication that the probability of churning is potentially higher. The partial dependence plots also shows that the distribution is skewed, indicating that most customers have a recency that is 365 days or higher. Looking at the partial dependence plot of the most important variable in terms of recency, *online_recent_dagen*, a similar curve can be

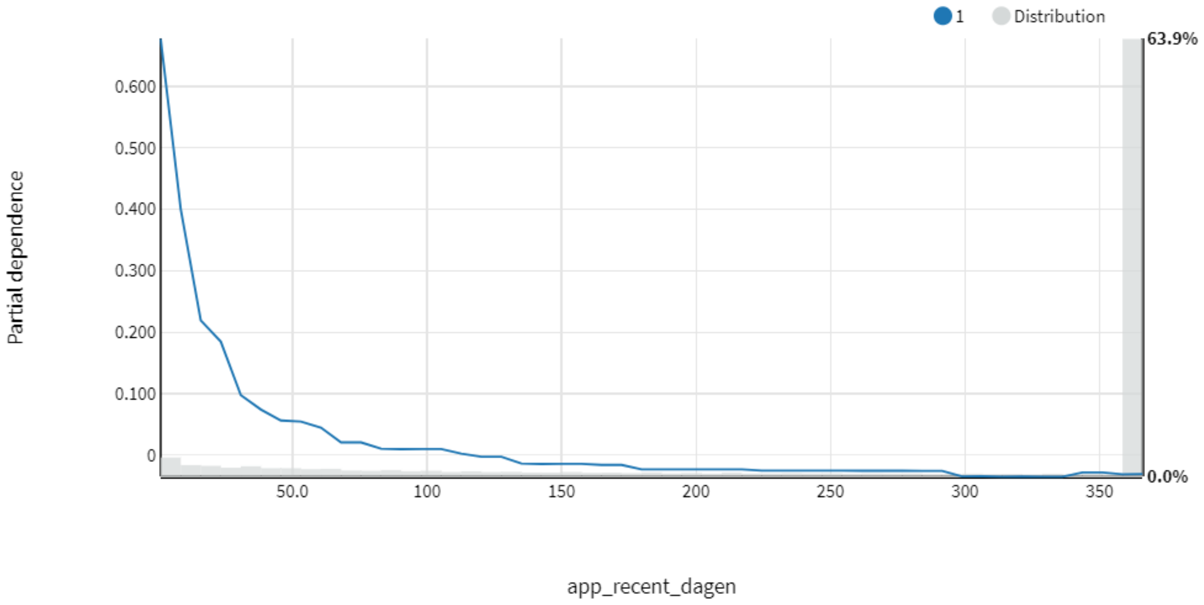

Figure 6.15: Partial dependence plot of app_recent_dagen

observed in Figure 6.14. Within the recency of the online interactions, a downwards sloping curve is related to the partial dependence, where the effect remains fairly constant between 45 and 75 days, after which the curve slopes downwards again albeit less steeply. Subsequently, the first 30 days have the largest effect on the probability, which is similar to the telephone recency feature. When considering the whole timespan of a year, the effect levels out after 100 days, having a partial dependence of almost 0, which turns into a negative effect on the churn probability when becoming more dthan 225 days. The last behavioural feature regarding the recency of the mobile application interactions, the *app_recent_dagen* feature, has a similar trajectory as the other two recency features as seen in Figure 6.15. The curve slopes downwards relatively steeply for the first 30 days, after which the effect becomes less significant.

RFI-yearly is a model that performs equally to the FI-monthly model when looked at the the first decile in the lift chart, achieving a lift of 3.95 in the first decile as observed in Figure 6.16. The lift of 3.95 indicates that targeting the top 10% of customers that are most likely to churn provides approximately 4 times more positive results as random sampling, which is better than the top-decile lifts of the model approximation and general model. The second decile however performs better than all the other models with a lift of 1.67, outperforming the second best with 0.03.

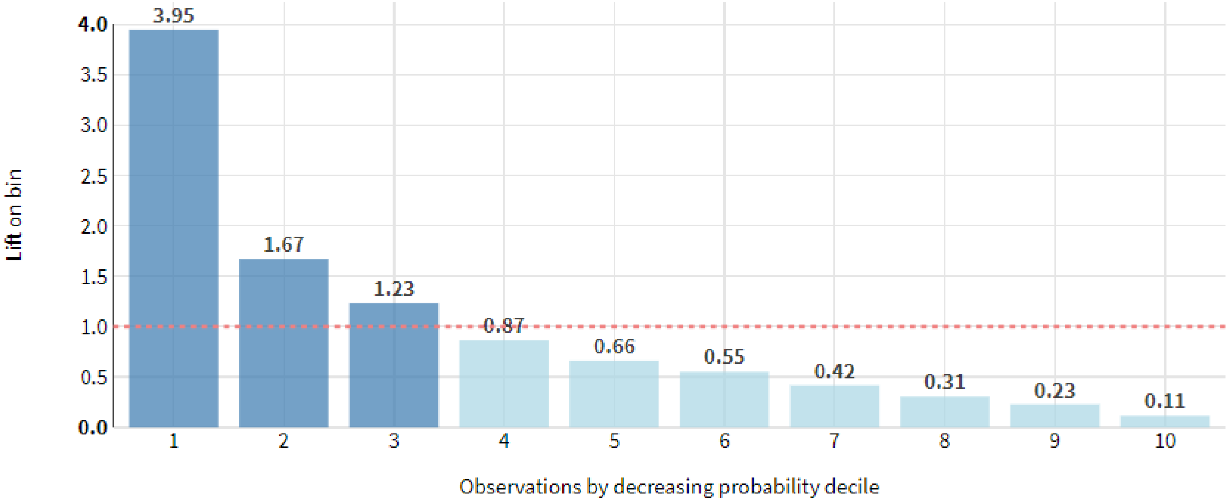

Figure 6.16: Lift chart of RFI-yearly

### 6.2.5 RFI-normalized

The last model attempts to capture the change relative to the interactions observed in a whole year by normalizing the monthly features against the year average. Looking at the confusion matrix in Table 6.14, the model predicts 648 cancelled policies that have actually been cancelled while predicting 2675 policies falsely as churned. This causes the precision to be 0.195 which is the third highest after the model approximaton and the RFI-yearly model. The recall, calculated to be 0.127 with 4452 false positives and 385489 true negatives, is lower than the recall of FI-monthly and the General Model.

Table 6.14: Confusion matrix of RFI-normalized

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 648 | 4443 | 5091 |
| | Non-churn | 2675 | 385312 | 387987 |
| | Total | 3323 | 389755 | 393078 |

The predictions of the model are mainly influenced by the features provided in Table 6.15, displaying the most importance for the same features found in the other models with the highest importance assigned to the percentage of discount on the policy. The top 10 of the most important features of this model consists out of seven policy-related features, two normalized behavioural features and one customer-related feature. The two normalized behavioural features, *online_gebruik_norm* and *telefoon_gebruik_norm* have an assigned importance of respectively 4% and 3%. Furthermore, it is noteworthy that the customer lifetime feature is not present in the top 10 of most important features, while being of significance for the other models.

Table 6.15: Feature importance of RFI-normalized

| Feature | Feature description | Feature importance (%) |
|---|---|---|
| Policy feature 1 | Provided discount | 25% |
| Policy feature 2 | Duration of policy | 11% |
| Policy feature 3 | Age of insured property | 10% |
| Policy feature 4 | Content of policy | 9% |
| Customer feature 1 | Customer lifetime | 7% |
| Policy feature 5 | Property of car | 5% |
| Policy feature 6 | Age | 5% |
| online_gebruik_norm | Monthly online interactions normalized | 4% |
| telefoon_gebruik_norm | Monthly telephone interactions normalized | 3% |
| Policy feature 7 | Property of car | 2% |

Analyzing the partial dependence plots of the behavioural features, the partial dependence plot of *telefoon_gebruik_norm* shows a distinct curve in Figure 6.17 with the point of emphasis laying around 0. Here, the effect is almost 0, indicating that a frequency relatively similar to the average over the year has no real effect on the churn probability. However, when the frequency is less than the year average, a small increase in the churn probability is indicated. The steepest increase whatsoever is found within more phone use than the year average, with the most gain found between 0-2. The other frequency variable, the *online_gebruik_norm*, show a similar trajectory only having a negative effect on the churn probability when the online frequency of interaction is lower compared to the year average. The steep increase of having a higher frequency is also seen within this variable, showcasing the steepest increase between 0-2. Around 0, the effect of the feature on the churn probability is considered to be 0, indicating that normal usage does not influence the churn probability as seen in Figure 6.18.

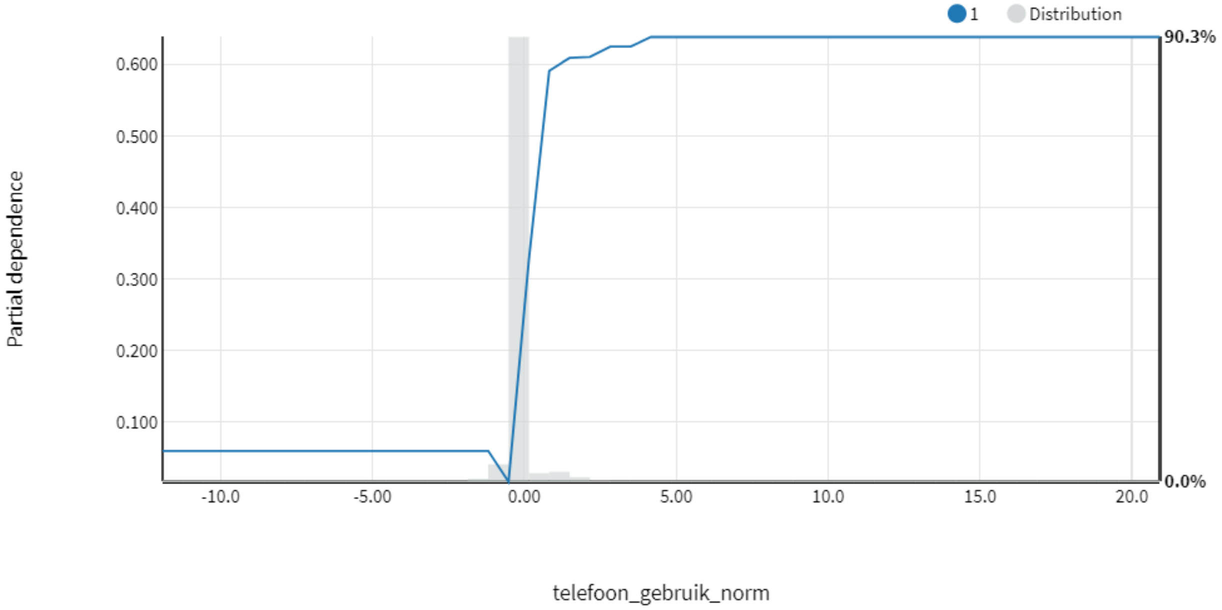

Figure 6.17: Partial dependence plot of telefoon_gebruik_norm

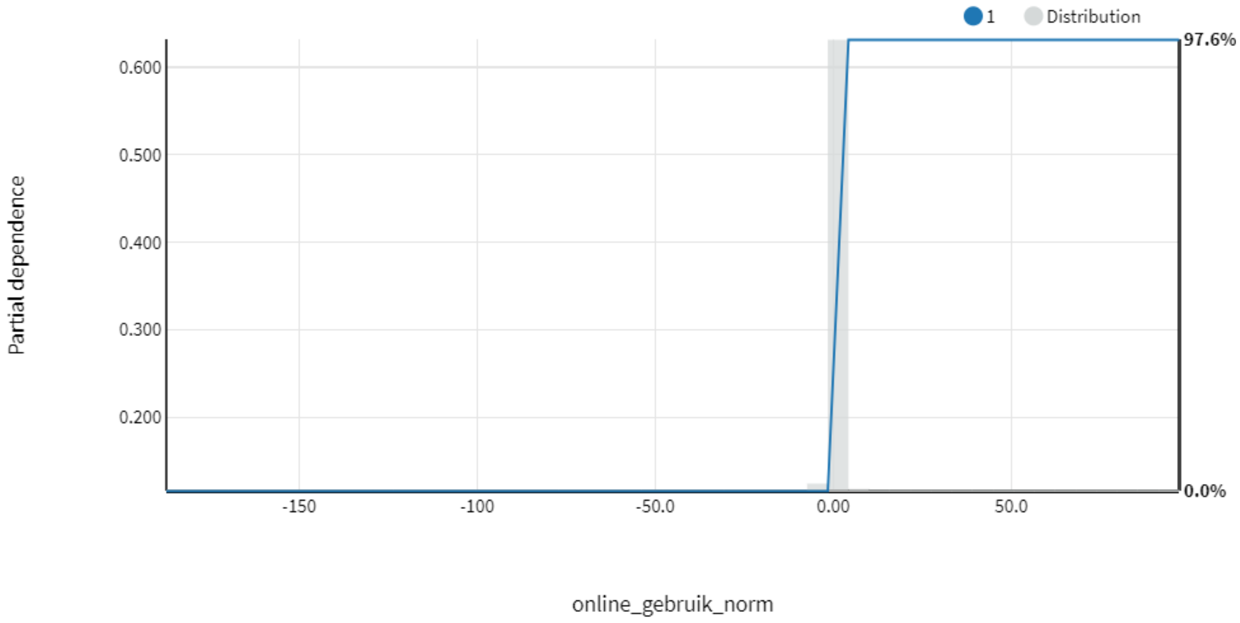

Figure 6.18: Partial dependence plot of online_gebruik_norm

Lastly, the predictive performance of the model in the first decile is the together with the RFI-yearly model the highest with a reported lift of 3.95, being higher than the other three models. Within the second decile, the lift is reported to be 1.62 which is lower than most models, but reports the highest lift of all models in the third decile with a lift of 1.24. This model therefore performs just like the other models better in the first three deciles of the sample, after which the model does not perform any better than a random model.

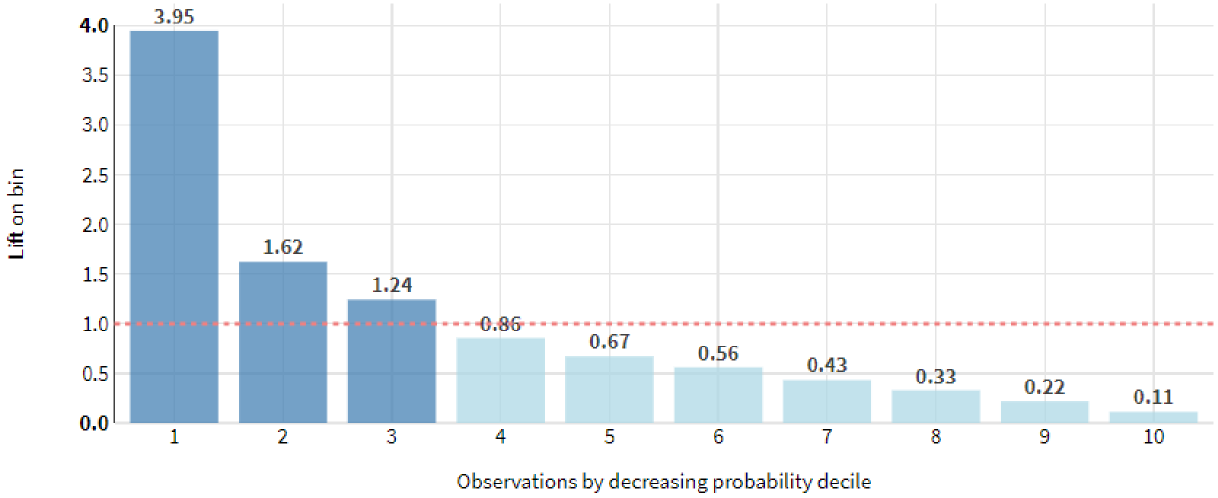

Figure 6.19: Lift chart of RFI-normalized

### 6.2.6 Top-decile lift analysis

Models like predictive churn models are often applied on a sub-sample while focusing on the whole customer base is often inefficient and costly in terms of company resources. A metric that enables to justifiably focus on a sub-sample is the top-decile lift, which already have been provided in the model comparison. However, focusing on this specific top-decile for the other metrics is also possible to assess the performance in a somehow realistic environment. Therefore, the observations of the top-decile are labelled as churn, while the rest is labelled as non-churn. Using the result of the top-decile, the $F_1$ score, precision and recall are given.

The confusion matrices per model are given in Table A.1-Table A.5. From these tables it is observed that the proportion of the actual churned policies found is relatively high, with the models finding between 1950 and 2050 actual churned policies. This is significantly higher than the number of actual churned policies found by the models that utilize a threshold, displaying the advantage when focusing on finding a relatively large number of churning policies. By aiming for the top-decile, the number of false positives amongst the different models also distinctively increases compared to utilizing a threshold, causing the average number of false positives to be approximately 37,300. Aiming for a larger proportion of actual churning policies therefore comes at the cost of an increased number of policies falsely predicted as churn.

Based on the various confusion matrices, the other metrics are given in Table 6.16. As seen in this comparison, the models are reasonably equivalent in the predictive performance. The RFI-normalized and RFI-yearly model both have the same number of churned policies in the top-decile, causing the results to be exactly the same and emphasizes the minor difference between the models. When considering these metrics, both RFI models are the highest performing when considering the $F_1$ score and the precision even though the differences are small. Furthermore, the recall is higher than the General model and Model Approximation. Due to the significant higher value of recall, utilizing a top-decile analysis is mainly useful when having the requirement of including many policies that have been cancelled by a customer, while the precision and $F_1$ are less compared to the models utilizing a threshold.

Table 6.16: Results different models in the top-decile

| **Model** | **$F_1$ score** | **Precision** | **Recall** |
|---|---|---|---|
| FI-monthly | 0.0905 | 0.0511 | 0.395 |
| General Model | 0.0883 | 0.0499 | 0.385 |
| Model Approximation | 0.0888 | 0.0501 | 0.387 |
| RFI-normalized | **0.0906** | **0.0512** | **0.395** |
| RFI-yearly | **0.0906** | **0.0512** | **0.395** |

## 6.3 Class imbalance analysis

Based on the results of the Section 6.2, it was motivated that analyzing the effect of having different class distribution could be beneficial for the predictive performance of the model. Therefore, the choice was made to analyze several class distributions. For this evaluation the predictive performance is assessed on the various metrics, altering the class distribution of the training set while maintaining the original distribution in the training set. The evaluation is done with three different models that utilize Decision Tree, Random Forest, and LightGBM.

### 6.3.1 Oversampling

The first strategy that has been applied is the strategy of oversampling using the SMOTE-NC technique. For oversampling, four distributions have been assessed with respectively the original distribution of 1%, 2%, 5%, and 10%. It was hypothesized that more than ten synthetic examples against one real observation would introduce too much bias, therefore focusing mainly on these distributions. The exact results of applying this technique in terms of metrics are depicted in Table 6.17, with the graphical representation of the $F_1$ score and the top-decile lift depicted in respectively Figure 6.20 and Figure 6.21. Within the LightGBM model, the $F_1$ score and top-decile lift remain fairly constant when the share of churned policies is artificially increased using oversampling, observing a minor decrease in the metric for the last two class distributions as seen in the graphical representations. The model utilizing the Decision Tree algorithm also shows a decrease in $F_1$ score, with a relatively large decrease when having a distribution of 95% non-churning policies. The Decision Tree model has a certain variation in the metric as also seen in the top-decile lift, observing that the Decision Tree technique is sensitive to various class distributions. Lastly, the Random Forest model reports an increasing $F_1$ score when increasing the share of churning policies in the training set. This is however not the case for the top-decile lift, which shows a rather small improvement for the 98% and 95% share of non-churn before decreasing to 2.35 for the 90% distribution. Overall, the LightGBM ensemble technique is more robust, having less variation when having different class distribution compared to both Decision Tree and Random Forest techniques.

Table 6.17: Results oversampling

| **Class Distribution (churn:non-churn)** | **Algorithm** | **AUC** | **$F_1$ score** | **Precision** | **Recall** | **TDL** |
|---|---|---|---|---|---|---|
| 1:99 | Decision Tree | 0.726 | 0.139 | 0.153 | 0.128 | 3.00 |
| | LightGBM | 0.763 | 0.148 | 0.160 | 0.138 | 3.97 |
| | Random Forest | 0.771 | 0.118 | 0.115 | 0.120 | 3.81 |
| 2:98 | Decision Tree | 0.724 | 0.137 | 0.195 | 0.106 | 3.53 |
| | LightGBM | 0.763 | 0.151 | 0.239 | 0.110 | 3.96 |
| | Random Forest | 0.771 | 0.131 | 0.164 | 0.109 | 3.60 |
| 5:95 | Decision Tree | 0.640 | 0.0945 | 0.106 | 0.085 | 1.72 |
| | LightGBM | 0.761 | 0.149 | 0.216 | 0.114 | 3.87 |
| | Random Forest | 0.753 | 0.129 | 0.107 | 0.164 | 3.60 |
| 10:90 | Decision Tree | 0.729 | 0.136 | 0.129 | 0.144 | 3.42 |
| | LightGBM | 0.757 | 0.146 | 0.142 | 0.149 | 3.85 |
| | Random Forest | 0.809 | 0.136 | 0.0888 | 0.289 | 2.35 |

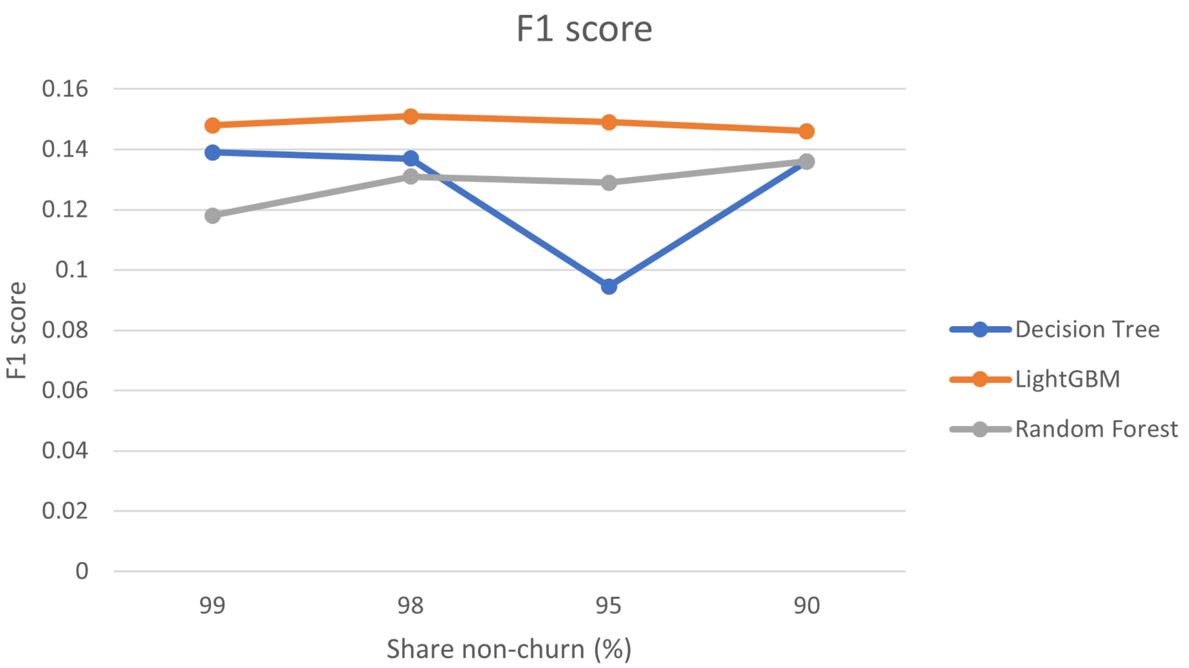

Figure 6.20: F1 scores for different oversampled class distributions

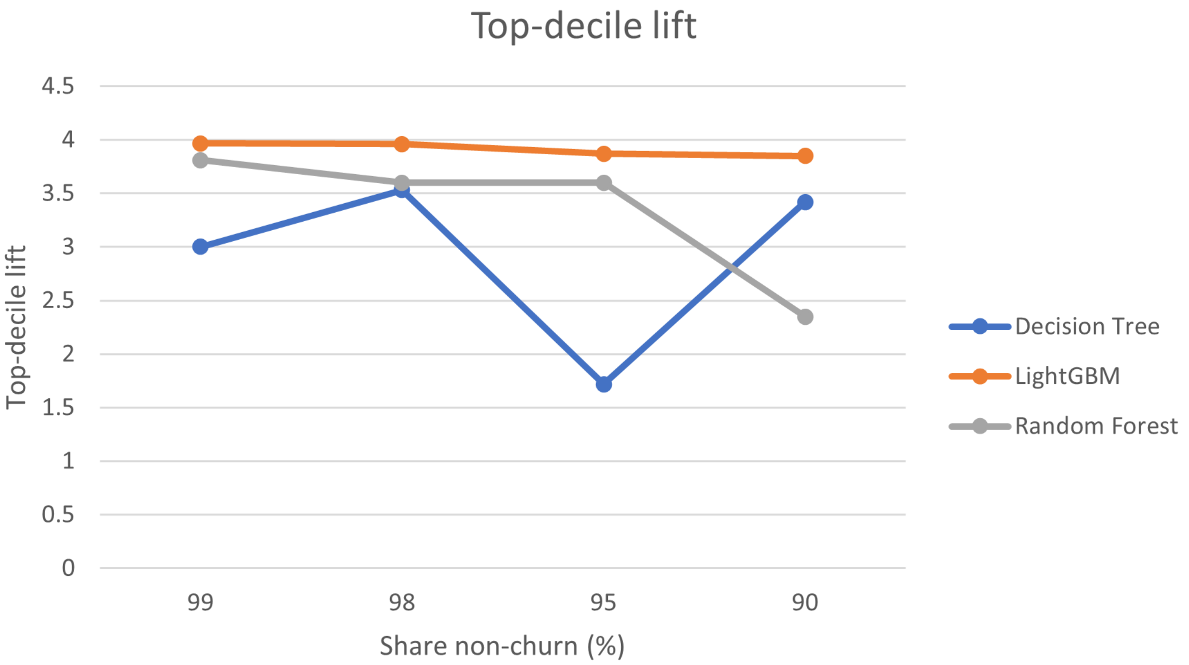

Figure 6.21: Top-decile lift for different oversampled class distributions

### 6.3.2 Undersampling

The second strategy of undersampling has been extensively researched, applying the technique of random undersampling to obtain more optimal class distributions. For the undersampling, a range of class distributions is analyzed ranging from 1% churned policies (the original dataet) to 50:50, while this does not introduce any synthetic examples. Based on the results provided in Table 6.18 and the graphical representation of the $F_1$ score in Figure 6.22 and the top-decile lift in Figure 6.23, it is observed that the LightGBM model is fairly constant with only a minor increase seen in the $F_1$ score when increasing the share of churned policies. Both the Decision Tree model and the Random Forest model are more affected by different class distributions. The Decision Tree model is constant when considering the $F_1$ score with only a small increase when having 80% non-churn, however displays significant variation in terms of the top-decile lift. For the top-decile lift, the Decision Tree model is observed to perform better when the share of non-churn is decreased in the dataset. Regarding the Random Forest algorithm, the $F_1$ score shows a significant degree of variation while remaining fairly constant in terms of top-decile lift when adjusting the class distribution. The Random Forest model shows the highest $F_1$ when having a share of non-churn of 95%, after which a gradual decline is observed.

Table 6.18: Results undersampling

| Class Distribution (churn:non-churn) | Algorithm | AUC | $F_1$ score | Precision | Recall | TDL |
|---|---|---|---|---|---|---|
| 1:98 | Decision Tree | 0.726 | 0.139 | 0.153 | 0.128 | 3.00 |
| | LightGBM | 0.763 | 0.148 | 0.160 | 0.138 | 3.97 |
| | Random Forest | 0.771 | 0.118 | 0.115 | 0.120 | 3.81 |
| 2:98 | Decision Tree | 0.727 | 0.138 | 0.155 | 0.125 | 3.15 |
| | LightGBM | 0.763 | 0.152 | 0.192 | 0.126 | 3.97 |
| | Random Forest | 0.787 | 0.089 | 0.052 | 0.304 | 3.79 |
| 5:95 | Decision Tree | 0.683 | 0.139 | 0.184 | 0.111 | 2.23 |
| | LightGBM | 0.763 | 0.155 | 0.250 | 0.112 | 3.95 |
| | Random Forest | 0.758 | 0.149 | 0.152 | 0.147 | 3.82 |
| 10:90 | Decision Tree | 0.729 | 0.136 | 0.129 | 0.144 | 3.42 |
| | LightGBM | 0.764 | 0.158 | 0.184 | 0.139 | 3.97 |
| | Random Forest | 0.756 | 0.143 | 0.156 | 0.132 | 3.79 |
| 20:80 | Decision Tree | 0.710 | 0.150 | 0.205 | 0.119 | 2.81 |
| | LightGBM | 0.764 | 0.160 | 0.262 | 0.115 | 3.97 |
| | Random Forest | 0.755 | 0.135 | 0.124 | 0.148 | 3.73 |
| 30:70 | Decision Tree | 0.728 | 0.131 | 0.148 | 0.117 | 3.54 |
| | LightGBM | 0.764 | 0.153 | 0.177 | 0.135 | 3.97 |
| | Random Forest | 0.751 | 0.117 | 0.0992 | 0.143 | 3.52 |
| 40:60 | Decision Tree | 0.716 | 0.128 | 0.140 | 0.118 | 3.43 |
| | LightGBM | 0.765 | 0.158 | 0.198 | 0.131 | 3.97 |
| | Random Forest | 0.752 | 0.116 | 0.0901 | 0.162 | 3.72 |
| 50:50 | Decision Tree | 0.736 | 0.137 | 0.164 | 0.118 | 3.57 |
| | LightGBM | 0.764 | 0.156 | 0.184 | 0.135 | 4.00 |
| | Random Forest | 0.748 | 0.106 | 0.0767 | 0.171 | 3.57 |

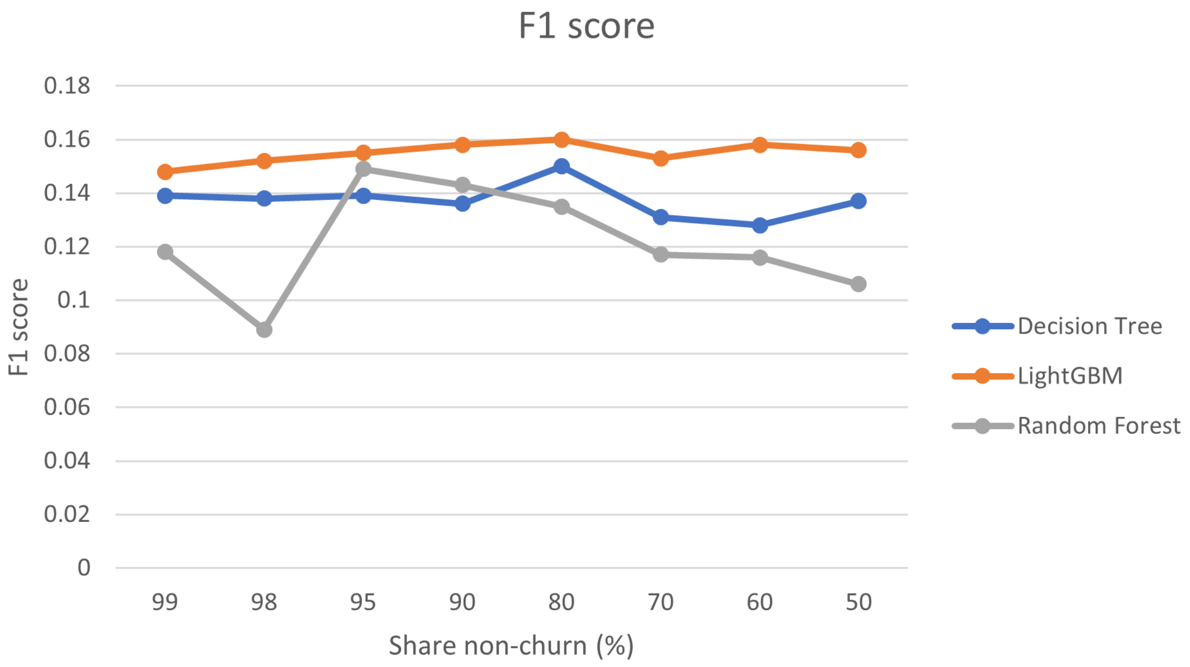

Figure 6.22: F1 scores for different undersampled class distributions

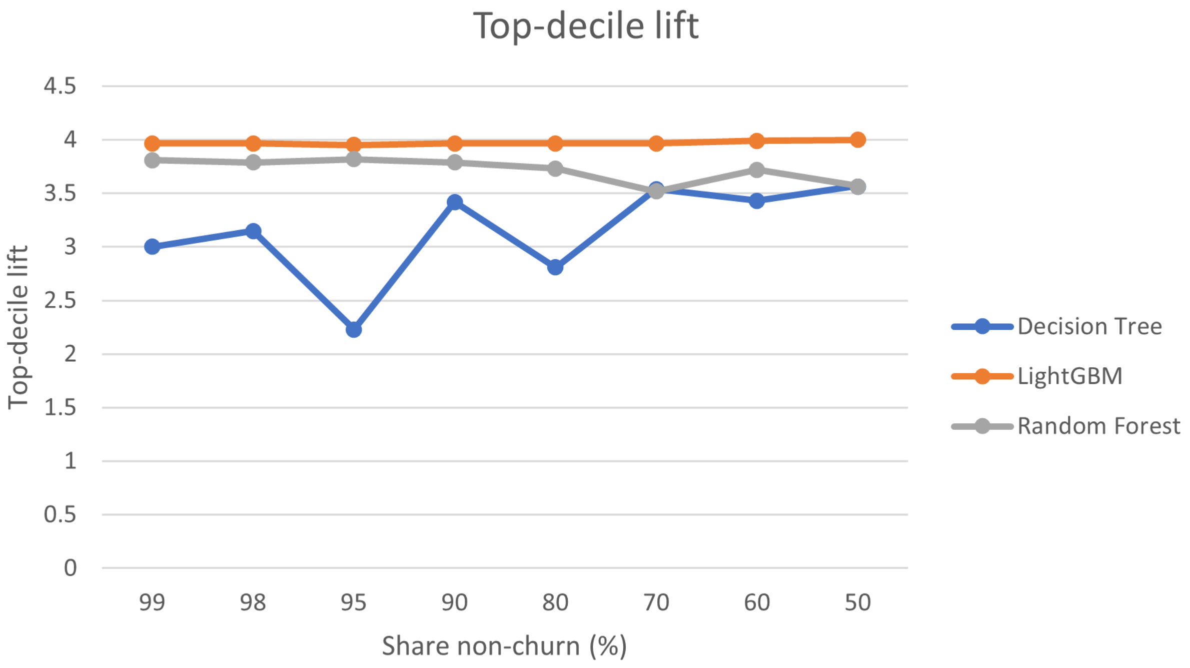

Figure 6.23: Top-decile lift for different undersampled class distributions

### 6.3.3 Hybrid sampling

The last sampling strategy that has been applied is the hybrid sampling strategy, where both oversampling and undersampling are combined. The hybrid sampling strategy uses a churn rate of 2% for the oversampling strategy, which is a 100% increase compared to the original dataset. Furthermore, three class distributions for undersampling are used to analyze the varying results. Considering the various metric as see in Table 6.19 and the graphical representations in Figure 6.24 and Figure 6.25. The LightGBM model is just like the other two strategies fairly constant, displaying a minor increase in terms of the $F_1$ while showing no distinctive increase or decrease in terms of the top-decile lift. The The performance of the Decision Tree model regarding the $F_1$ score decreases when differing the class distribution from the original dataset. Furthermore, the model shows a high degree of variation when adjusting the class distribution in terms of top-decile lift, with the highest top-decile lift observed when the churn is oversampled to 2% and followed by an undersampling of 50%. The Random Forest model shows minor variation in the $F_1$ scores, with the highest performance obtained when oversampling to 2% churn and afterwards undersampling the churn ratio in the data to 10%. The top-decile lift is not improved in any different class distributions for the Random Forest model, only showing a high degree of variation.

Table 6.19: Results hybrid sampling

| **Class Distribution (oversampled—undersampled)** | **Algorithm** | **AUC** | **$F_1$ score** | **Precision** | **Recall** | **TDL** |
|---|---|---|---|---|---|---|
| Original dataset | Decision Tree | 0.726 | 0.139 | 0.153 | 0.128 | 3.00 |
| | LightGBM | 0.763 | 0.148 | 0.160 | 0.138 | 3.97 |
| | Random Forest | 0.771 | 0.118 | 0.115 | 0.120 | 3.81 |
| 0.02—0.05 | Decision Tree | 0.668 | 0.109 | 0.184 | 0.0774 | 3.17 |
| | LightGBM | 0.763 | 0.155 | 0.208 | 0.123 | 3.94 |
| | Random Forest | 0.785 | 0.119 | 0.0723 | 0.339 | 1.89 |
| 0.02—0.1 | Decision Tree | 0.668 | 0.118 | 0.187 | 0.086 | 1.43 |
| | LightGBM | 0.763 | 0.151 | 0.195 | 0.124 | 3.94 |
| | Random Forest | 0.747 | 0.135 | 0.123 | 0.150 | 3.70 |
| 0.02—0.5 | Decision Tree | 0.667 | 0.105 | 0.166 | 0.0770 | 3.42 |
| | LightGBM | 0.757 | 0.148 | 0.152 | 0.145 | 3.90 |
| | Random Forest | 0.716 | 0.0850 | 0.0610 | 0.141 | 3.23 |

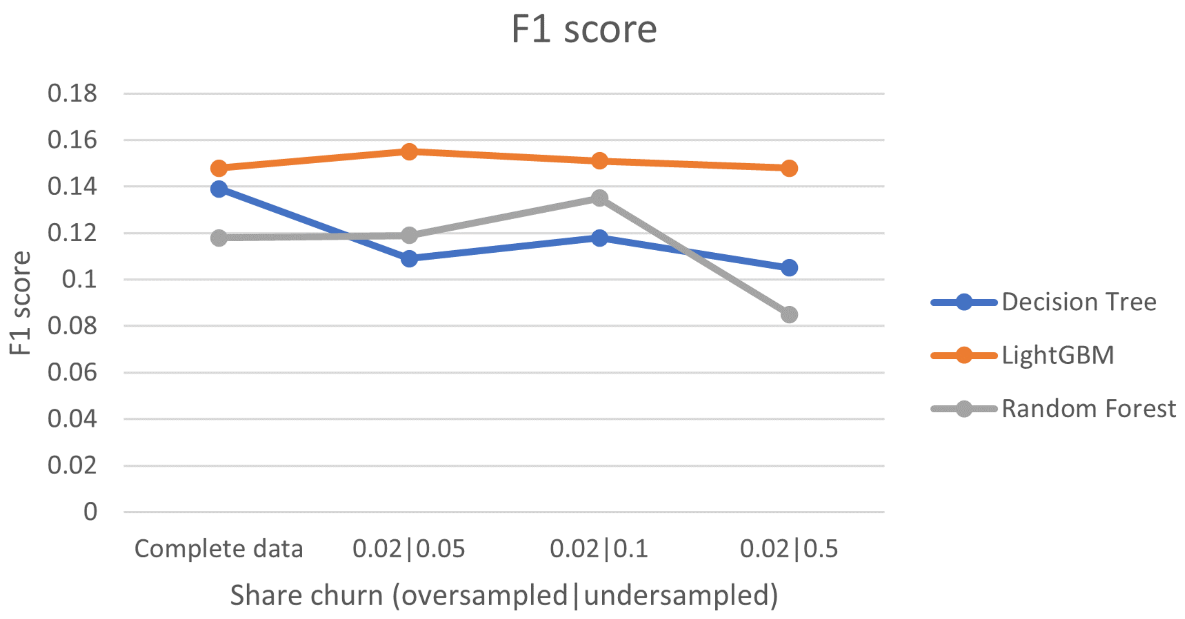

Figure 6.24: F1 scores for different hybrid sampled class distributions

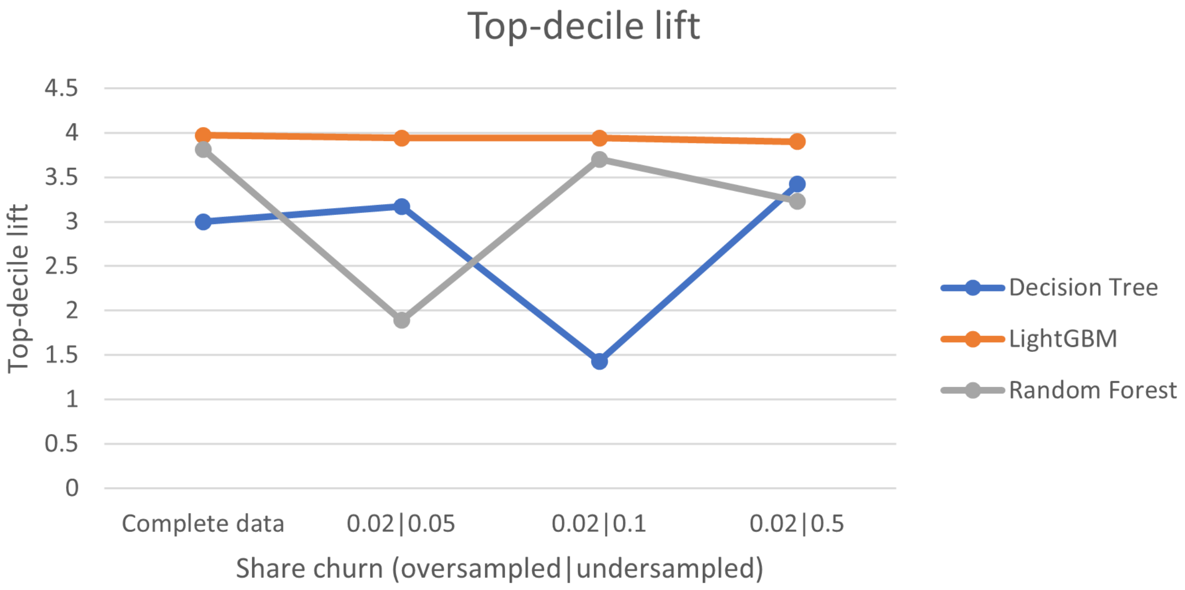

Figure 6.25: Top-decile lift for different hybrid sampled class distributions

# Chapter 7

# Discussion

One of the most important challenges of any business is the challenge of retaining their customers, being especially true for businesses in a competitive market like the Dutch insurance industry. Reasons to leave a company are numerous and often based on personal preferences, illustrating the complexity of retaining customers and preventing customers for leaving. Businesses are constantly searching for opportunities to improve their customer retention, reducing the costs of lost customers and improving their competitiveness. An opportunity for customer retention is the use of a predictive churn model, a model that attempts to predict the probability of a customer leaving the company. This research focused on the predictive churn model of a major Dutch insurance company, attempting to improve the predictive performance of the model by including more extensive customer behaviour and changes in this behaviour. Furthermore, various sampling strategies were explored to research the effect of the sampling strategy on the performance. This chapter answers the main research question and sub-research questions, provide recommendations for the insurance company, discuss the limitations of the research, and describe possibilities for future research.

## 7.1 Review of Research Questions

To enable the answering of the main research question, five sub-research questions were defined in Section 1.2. First, the sub-research questions will be answered before this section is concluded with an answer to the main research question.

**Sub-research question 1**
*Which techniques can be used to predict the probability of a customer churning?*

The review of literature conducted in the area of predictive churn modelling in Chapter 2 and Chapter 3 showed a wide range of supervised learning techniques that can be applied for the prediction of customer churn. Based on Table 2.1 and the extended analysis of related literature, it was observed that Logistic Regression and Decision Tree were often used in the context of churn prediction, having a high level of interpretability while reporting promising results in various studies. Furthermore, Neural Networks are frequently used for prediction of churn as a type of deep learning, observing multiple types of Neural Networks like Recurrent Neural Networks and Convolutional Neural Networks. Following the literature reviews, it was also observed that Support Vector Machines and Naive Bayes could be applied in the context of churn prediction.

Based on the extensive analysis of related work in Chapter 3, applications of ensemble methods were also seen as a viable option for the prediction of churn. Although having less interpretability than methods like Logistic Regression and Decision Tree, ensemble methods like

Random Forest and Boosting algorithms have been utilized in various studies. When considering the context of churn prediction within the insurance industry, the commonly seen methods of Logistic Regression, Decision Tree, Random Forest, Artificial Neural Networks, Support Vector Machine have also been frequently applied in this specific context. However, the use of ensemble methods in the churn prediction for an insurance company is less researched with only the Random Forest technique being commonly applied.

**Sub-research question 2**
*How can customer behaviour be used within the context of predictive churn modelling?*

Customer behaviour is used within the context of predictive churn modelling through various ways. One of the most common methods of including customer behaviour is the use of the RFM framework, a framework that consists out of a *Recency*, *Frequency*, and *Monetary* variable and originates from the direct marketing and has been commonly applied in the context of churn prediction in an attempt to include more behavioural data for the churn models. Within this framework, a distinction is made between summary RFM features that measure the various types through a basic dimension of measurement and the detailed RFM features that transform RFM variables in an attempt to capture more information. Transformations are seen by averaging or normalizing the information, varying the dimension of the feature, distinguishing the time or other variations. Multiple studies have also proposed extensions to the RFM framework by including new features that capture other behaviour like the *intensity* of the actions performed by the customer or more qualitative approaches by including the emotions found in the customer interactions. Overall, customer behaviour is mainly included through features based on the data gathered from the interaction of the customers with the company, including the information through a quantitative or qualitative approach.

**Sub-research question 3**
*How do different predictive churn modelling techniques compare?*

The performance of the techniques is compared using three distinct models on the AUC, $F_1$ score, precision, recall, and the top-decile lift. By analyzing the various techniques for different models, a general comparison can be made regarding the performance of the various techniques. LightGBM scores the highest for all three models on the top-decile lift and AUC, while scoring amongst the highest in terms of the other metrics. This indicates a certain robustness which is also seen in the other two boosting algorithms of Gradient Boosted Trees and XGBoost. These algorithms appear amongst the highest scoring in terms of the $F_1$ score, AUC, and top-decile lift while scoring reasonably to high on the other metric when compared to the other algorithms. This robustness is not seen with Random Forest and Logistic Regression, where the models based on these algorithms report a low precision compared to the other algorithms.

Noteworthy is the performance of the Decision Tree algorithm, scoring relatively high on precision across the various models and corresponding $F_1$ score. Decision Tree is known for its high interpretability, using a flowchart-like structure to decide the probability of churn. This is potentially useful for the company utilizing the predictive churn model, making the reasons for churning easily interpretable. Predictive churn models often focus on a subsample, aiming for the highest lift. The top-decile lift is used to gain more insights of the performance, and within this top-decile lift it is seen that the boosting algorithms perform higher than the other algorithms when focusing on 10% of the customer base. When this 10% is obtained from the models, the boosting algorithms score consistently higher on the proportion of actual churned policies in the 10% than the other algorithms.

In conclusion, the boosting algorithms perform reasonably well compared to other algorithms, with the LightGBM being slightly better than the other boosting algorithms. When the model is mainly utilized to obtain a consistent model in terms of metrics, the boosting algorithms and LightGBM in particular is a solid choice. When focusing on the precision of the model, the decision tree is in comparison the highest performing while having a high degree of interpretability and simplicity.

**Sub-research question 4**
*How do predictive churn models that utilize customer behaviour compare to models that do not utilize customer behaviour?*

Various models have been analyzed in the model comparison on various metrics, variable importance, and lift per bin. This comparison utilized the LightGBM algorithm to only study differences in the models based on the behavioural traits. Based on the results of the various models regarding the metrics, the models are comparable in terms of the harmonic mean between the precision and recall. Subsequently, when a certain model scores high on precision, the recall will be lower compared to the other models. This is for example seen in the model as foreseen by the company, the Model Approximation model, which scores high in terms of precision but lowest on recall. The area under the curve is comparable, with slightly higher values for the FI-monthly, RFI-normalized and RFI-yearly model.

Based on the top-decile, the corresponding top-decile lift and the analysis specific to this top-decile, it is observed that the RFI-yearly and RFI-normalized capture the most policies that have been cancelled when focusing on the top 10% of the policies. Using the aggegrated recency, frequency, and monetary variables or the normalized variants, targeting 10% of the customers lead to the capturing of 39.5% of the actual cancelled policies. This is higher compared to the General Model and Model Approximation with 38.5%, quantitatively leading to the inclusion of 40 extra cancelled policies on a test set of 39308. A minor improvement is therefore seen in more extensive behavioural models when targeting a subsample of the testing set.

The analysis of the various partial dependence plots and feature importances showed that the inclusion of behavioural features provided some weight to the models. For the model approximation, the number of inbound calls and number of online and app interactions had an assigned variable importance of respectively 2% and 1%. For the FI-monthly model, the online intensity and the frequency of phone usage aggregated on a monthly basis both showed a variable importance of 3%. The recency variable is the main feature of importance for the RFI-yearly model, having all three types of communication included. The online recency is the most important with a feature importance of 7% within the model, where the phone recency and app recency have a value of respectively 5% and 2%. RFI-normalized has different features that are important with regards to the behavioural features with normalized online frequency and phone frequency account for 4% and 3%. In comparison, the important features differ per model, with RFI-yearly reporting the highest feature importance of the behavioural features. Although the various types of interaction, mobile application usage, online visits, and telephone interaction are all included within the various models, features regarding the mobile application interactions are only included once as a standalone feature, with the recency feature seen in the RFI-yearly model. Within the partial dependence plots, the same effect of frequency is seen across all models that include behavioural features. More frequency causes the probability to churn to become higher, where the normalized features show the main effect when having more frequency than on average. The plots of the recency features show an increase in the churn probability when the interaction has been more recent.

Models that utilize customer behaviour perform comparable in terms of metrics while showing a minor increase in the performance when targeting a subsample. Based on the importance of the various behavioural features, online and telephone interactions are considered to be more important than application interactions. The partial dependence plots provided insights in the effect of the behavioural features, with the probability of churning increasing when the frequency or intensity is increased. Lastly, more recent contact is an indicator of a higher chance of churning, emphasizing the importance of these interactions.

**Sub-research question 4**
*What is the influence of class distribution and the handling of this class distribution on predictive churn models?*

A wide range of class distributions have been analyzed, obtained through oversampling, undersampling, and hybrid sampling strategies. The effect of increasing the amount of churned policies in the training set is dependent on the algorithm used, where the LightGBM algorithm is hardly influenced by an increase in the number of churned policies found in the dataset. For all three strategies it is observed that the top-decile lift remains around 3.97, which is only slightly increased when having an undersampled 50:50 distribution. The $F_1$ score also has similar scores across varying class distributions, only showing a small increase when undersampling. For the other two algorithms, a relatively high degree of variation is seen amongst various class distributions. It is observed that the algorithms have significant differences when increasing the share of churned policies, making it difficult to find a general trend for the class distributions.

Looking specifically at the oversampling strategy, it is seen that the Random Forest model has a slight increase in $F_1$ score while performing worse in terms of top-decile lift when increasing the share of churned policies. For the decision tree a minor decrease in $F_1$ score and increase in top-decile lift is observed, with a significant difference when having 95% non-churned policies in the testing set. For the undersampling, Random Forest models seem to be fairly constant in the top-decile lift in various class distributions, having a higher degree of variation in terms of $F_1$ score. The opposite is true for a Decision Tree model, showing a high degree of variation in terms of $F_1$ score while having minor variation in the top-decile lift. Lastly, the hybrid sampling decreases the $F_1$ score for the Decision Tree model while increasing and decreasing this score for the Random Forest model. Hybrid sampling introduces a significant degree of variation for the top-decile lift for both models, increasing and decreasing the performance.

In general terms, the effect of class distribution is context-dependent where LightGBM is hardly affected by different class distributions and sampling strategies. Other models show a higher degree of variation, where undersampling can be beneficial for the top-decile lift of decision tree models. Hybrid sampling is observed to slightly affect the performance in terms of $F_1$ score, while introducing a high degree of variation for Random Forest and Decision Tree models.

**Main research question**
*How does the inclusion of customer behaviour influence the predictive modelling of customer churn within the insurance industry?*

The inclusion of customer behaviour influences the predictive churn modelling for a Dutch insurer by adding features that have some value for the predictive power of the models. Including behavioural features marginally increases the predictive performance in terms of metrics, increasing the area under the ROC curve while obtaining a better $F_1$ score and lift. The predictive power of the models is relatively low, causing any improvement to be important.

The main area of improvement has been in the targeting of the top-decile. Behavioural features influence the top-decile lift, causing the actual churned policies present in the top-decile to be higher for the models with more extensive behavioural features. Although the differences are small, aggregating behavioural on a yearly basis helps in the reaching of more policies that have actually been cancelled. Because the difference in costs between retaining customers and gaining new customers is high, reaching more churned policies is highly valuable for the company and targeting even a small number of policies that are potentially being cancelled extra will be beneficial for the company. Including customer behaviour also influences the performance while a new source is included, causing the model to be less dependent on policy variables. The model without any behavioural features include eight policy-related features, while most models including behavioural features report less policy-related features in their most important features. Including other sources besides information regarding the policy is important, making the dependency on these features smaller. The decision of cancelling a policy is based on various aspects and including some of these aspects can be beneficial.

Lastly, the inclusion of behavioural features increases the interpretability of the models. The effect of behavioural features is easily understood as seen in the partial dependence plots and the description of these plots, in contrast to features like the fuel economy of the car. Enhancing the interpretability of the models can be used for obtaining marketing insights or understanding why a customer has left the company, all valuable information for the company.

## 7.2 Company recommendations

Every company benefits from an improved predictive churn model, especially companies operating in a highly competitive and saturated market like the Dutch insurance company in which the study was conducted. Based on the variable importance and the relative importance within the model, a recommendation would be to include more customer behaviour in their predictive churn model, with recency being the main candidate based on the performance of the RFI-yearly model and frequency already being included within the model as foreseen by the insurance company. Although it only slightly enhances the model, focusing on the customers with the highest churn probability with the behavioural features included can be beneficial.

The predictive churn models found in this study do not incorporate customers who have been customer for less than a month, therefore excluding a significant proportion of the customer base. These customers have been excluded based on the graphical analysis provided in Chapter 4, which found that the frequency of interactions differ from long-term customers. Understanding this difference and the reasons behind this will enable the company to enhance their retention strategies for this group. These customers can be considered to still be in the onboarding phase, where a high frequency of interaction may indicate that something is going wrong and should be addressed by the company.

## 7.3 Limitations

Customer churn modelling is a complicated subject while customer churn depends on a large number of aspects, ranging from policy-related aspects like the price of a policy to customer-related aspects like the personal preference. Predicting customer churn is therefore not easily done, however some limitations are also observed in this study that can influence the performance of the models. The model entails data which has been aggregated per month, causing each customer to have at maximum twelve entries. Although this provide a lot of information, it also introduces bias towards the policies that have not been cancelled. The policies that have been cancelled have less than twelve entries, causing the imbalance to become higher and therefore introducing more bias towards the majority class. Although this was necessary to predict each month per customer, this method most likely had influence on the study.

Following the use of a monthly observation period, the model attempted to predict the customer churn for the coming month, making the model highly specific by aiming for a prediction the next month. This causes the targeting to be specific, focusing on customers that are already thinking about cancelling their policy, which can be seen as an advantage. However, the complexity of the prediction further increases whilst the prediction of churn on a yearly basis is already a difficult subject. By making the period of churn so specific, the predictive performance is most likely affected.

A third limitation of this study is the aim of including customer behaviour in the models, therefore mainly focusing on what the differences are between models that include customer behaviour and the model without any behavioural features. Subsequently, any bias found in the model without customer behaviour is also transferred to the models with customer behaviour. The models have a feature selection strategy included based on domain knowledge, however it may be beneficial for the predictive performance to include a more systematically method of feature selection like focusing on a correlation analysis or Principal Component Analysis (PCA). A fourth limitation that is also related to the inclusion of customer behaviour are the data sources used to find the customer interactions and corresponding behaviour. Because the prediction is done on a policy-level, bias is introduced while the behaviour relates to a customer. A customer can have different policies, being a small proportion of the whole dataset, however it potentially still affects the predictive power of the behavioural features.

Lastly, a limitation that is observed in this study is that the comparison of models is done through various metrics, while one of the main drivers of costs and benefits have not been accounted for. Including a metric that represented the total benefit for the company, e.g. the monetary benefits, would maybe help in differentiating the models while the differences in the metrics are small. Including a benefit metric would also enhance the usability of the comparison for the company, where the purpose of the model is mostly marketing. Targeting people requires certain costs of the company and by providing a cost metric regarding the various models, an assessment can be made whether the differences are worth it.

## 7.4 Future Work

An approach that has not been included in this research but which can potentially be beneficial for the predictive churn models is the inclusion of qualitative behavioural features. This research had a quantitative approach, focusing on how many interactions, the intensity, and how many days had passed since the last contact. A different approach would be focusing on the content of the interactions, for example which specific web pages are visited. Certain web pages can potentially indicate a customer reviewing their policy, like web pages that have all the contract terms or cancellation conditions. Another approach could be using the contents of the phone interactions, utilizing text mining to derive information about the customer's attitude.

A second topic that could be included by research is the application of semi-supervised learning, utilizing a combination of labelled and unlabelled data to predict churn. This approach would enable the use of more recent data, potentially reduce the effect of the imbalance found in the dataset, and capturing the underlying data distribution that can help in the generalizability of the model.

# Chapter 8

# Conclusion

The main objective of this research was to study the effect of including customer behaviour in predictive churn models, a model that predicts the probability of a customer leaving the company. For the analysis multiple classifiers have been used, using traditional algorithms like Decision Tree and Logistic Regression, ensemble algorithms like Random Forest, XGBoost, LightGBM, and Gradient Boosted Trees. This study first explored these algorithms, concluding that the boosting algorithms were robust and that these algorithms were scoring the highest of all algorithms in terms of the $F_1$ score, AUC and the top-decile lift. However, the performance of the Decision Tree algorithm was relatively comparable, showing a minor decrease in predictive performance while being significantly more interpretable than the ensemble algorithms.

From the results of the algorithm comparison it was motivated that a better performance could potentially be achieved through a higher percentage of actual cancelled policies in the dataset. However, the various sampling strategies of undersampling, oversampling, and hybrid sampling mainly introduced a degree of variation among various class distributions for Random Forest and Decision Tree, whereas LightGBM remained fairly constant for the varying churn distributions.

The objective of this study was achieved by analyzing various models based on the LightGBM algorithm, the algorithm that performed the best amongst the boosting algorithms and scored relatively well for all metrics. Besides the predictive performance, the LightGBM algorithm is also less affected by the class distribution. By analyzing the various models, it was observed that the extensions proposed in terms of behavioural features caused a slightly higher predictive performance, being almost comparable in terms of performance. When the analysis focused on the top-decile, the 10% of the observations with the highest probability of churning, a more distinctive result was observed, indicating that the RFI-yearly and RFI-normalized enabled a better targeting of the actual churned policies. The increase was modest, however the targeting of actual cancelled policies is difficult and achieving a slightly higher proportion of cancelled policies can already be beneficial for the company using the predictive churn model. It was observed in the model comparison that the number of policy-related variables often decreases when looking at the ten most important features, causing a new source of information to be present. The behavioural features are relatively easy to understand, emphasizing another benefit of using features based on the customer behaviour. Following these aspects, including customer behaviour is therefore potentially seen as beneficial for a Dutch insurance company despite the minor differences in metrics.

# References

Abd Elrahman, S. M., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*(2013), 332–340. Retrieved from https://www.semanticscholar.org/paper/A-Review-of-Class-Imbalance-Problem-Elrahman-Abraham/bb2e442b2acb4530aa28d24e45578f84447d0425

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. *IEEE Access*, *8*, 220816–220839. doi: 10.1109/access.2020.3042657

Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic behavior based churn prediction in mobile telecom. *Expert Systems with Applications*, *162*, 113779. doi: 10.1016/j.eswa.2020.113779

Ali, Ö. G., & Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, *41*(17), 7889–7903. Retrieved from https://www.infona.pl/resource/bwmeta1.element.elsevier-37264b2d-e6f5-33f2-b5de-7701273c5c91

Alkitbi, S. S., Alshurideh, M., Al Kurdi, B., & Salloum, S. A. (2020). Factors affect customer retention: A systematic review. In *International conference on advanced intelligent systems and informatics* (pp. 656–667). doi: 10.1007/978-3-030-58669-0_59

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940–7957. Retrieved from https://napier-repository.worktribe.com/output/1792667/comparing-oversampling-techniques-to-handle-the-class-imbalance-problem-a-customer-churn-prediction-case-study

Arshed, N., & Dansen, M. (2015). *The literature review* (2nd ed.; K. O'Gorman & R. MacIntosh, Eds.). Goodfellow Publishers. doi: 10.23912/978-1-910158-51-7-2790

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80–98. doi: 10.1509/jmr.16.0163

Aspinall, E., Nancarrow, C., & Stone, M. (2001). The meaning and measurement of customer retention. *Journal of Targeting, Measurement and Analysis for Marketing*, *10*(1), 79–87. doi: 10.1057/palgrave.jt.5740035

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. doi: 10.1007/s10462-020-09896-5

Berger, P., & Kompan, M. (2019). User modeling for churn prediction in e-commerce. *IEEE Intelligent Systems*, *34*(2), 44–52. doi: 10.1109/mis.2019.2895788

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, 542–545. doi: 10.1016/b978-0-12-809633-8.20349-x

Berson, A., & Thearling, K. (1999). *Building data mining applications for crm.* Retrieved from https://dl.acm.org/doi/abs/10.5555/580792

Bolancé, C., Guillen, M., & Padilla-Barreto, A. E. (2016). Predicting probability of customer churn in insurance. In *International conference on modeling and simulation in engineering, economics and management* (pp. 82–91). doi: 10.1007/978-3-319-40506-3_9

Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer lifetime value measurement. *Management science*, *54*(1), 100–112. doi: 10.1287/mnsc.1070.0746

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, *26*(4), 509–518. Retrieved from https://biblio.ugent.be/publication/212066

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European journal of operational research*, *164*(1), 252–268. doi: 10.1016/j.ejor.2003.12.010

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626–4636. doi: 10.1016/j.eswa.2008.05.027

Castro, E. G., & Tsuzuki, M. S. (2015). Churn prediction in online games using players' login records: A frequency analysis approach. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(3), 255–265. doi: 10.1109/tciaig.2015.2401979

Chalmeta, R. (2006). Methodology for customer relationship management. *Journal of systems and software*, *79*(7), 1015–1024. doi: 10.1016/j.jss.2005.10.018

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357. doi: 10.1613/jair.953

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of operational research*, *223*(2), 461–472. doi: 10.1016/j.ejor.2012.06.040

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*(9), 1629–1636. doi: 10.1016/j.jbusres.2012.12.008

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, *34*(1), 313–327. doi: 10.1016/j.eswa.2006.09.038

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, *36*(3), 6127–6134. doi: 10.1016/j.eswa.2008.07.021

Coyles, S., & Gokey, T. C. (2005). Customer retention is not enough. *Journal of Consumer Marketing*, *22*(2), 101–105. doi: 10.1108/07363760510700041

Crockett, D., & Eliason, B. (2016). What is data mining in healthcare. *Insights: Health Catalyst*. Retrieved from https://www.healthcatalyst.com/wp-content/uploads/2014/06/What-is-data-mining-in-healthcare.pdf

Dataiku. (2023). *About dataiku.* Retrieved from https://www.dataiku.com/company/

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*, *36*(4), 1563–1578.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, *548*, 497–515. doi: 10.1016/j.ins.2019.12.075

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems: First international workshop, mcs 2000 cagliari, italy, june 21–23, 2000 proceedings 1* (pp. 1–15). doi: 10.1007/3-540-45014-9_1

Edward, M., & Sahadev, S. (2011). Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage. *Asia Pacific Journal of Marketing and Logistics*. doi: 10.1108/13555851111143240

Eichinger, F., Nauck, D. D., & Klawonn, F. (2006). Sequence mining for customer behaviour predictions in telecommunications. In *Proceedings of the workshop on practical data mining at ecml/pkdd* (pp. 3–10). Retrieved from `https://publikationen.bibliothek.kit.edu/1000005948`

Eria, K., & Marikannan, B. P. (2018). Systematic review of customer churn prediction in the telecom sector. *Journal of Applied Technology and Innovation*, *2*(1).

Feng, Y., Wang, D., Yin, Y., Li, Z., & Hu, Z. (2020). An xgboost-based casualty prediction method for terrorist attacks. *Complex & Intelligent Systems*, *6*(3), 721–740. doi: 10.1007/s40747-020-00173-0

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning* (pp. 3–33). Springer, Cham. Retrieved from `https://link.springer.com/chapter/10.1007/978-3-030-05318-5_1?error=cookies_not_supported&code=0417f1e5-7258-4630-b5cc-86414147d856`

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484. doi: 10.1109/tsmcc.2011.2161285

García, D. L., Nebot, À., & Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. *Knowledge and Information Systems*, *51*(3), 719–774. doi: 10.1007/s10115-016-0995-z

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, 1–26. doi: 10.1007/s41060-022-00312-5

Gold, C. (2020). *Fighting churn with data*. Manning Publications.

Guerola-Navarro, V., Gil-Gomez, H., Oltra-Badenes, R., & Sendra-García, J. (2021). Customer relationship management and its impact on innovation: A literature review. *Journal of Business Research*, *129*, 83–87. doi: 10.1016/j.jbusres.2021.02.050

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., … Sriram, S. (2006). Modeling customer lifetime value. *Journal of service research*, *9*(2), 139–155. doi: 10.1177/1094670506293810

Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., & Ørnulf Borgan. (2014, 2). Modelling and predicting customer churn from an insurance company. *http://dx.doi.org/10.1080/03461238.2011.636502*, 58-71. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/03461238.2011.636502` doi: 10.1080/03461238.2011.636502

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., & Bauckhage, C. (2014). Predicting player churn in the wild. In *2014 ieee conference on computational intelligence and games* (pp. 1–8). doi: 10.1109/cig.2014.6932876

Hawkins, D., & Hoon, S. (2019). The impact of customer retention strategies and the survival of small service-based businesses. *SSRN Electronic Journal*, *5*(564), 1–19. doi: 10.2139/ssrn.3445173

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328).

He, Y., Xiong, Y., & Tsai, Y. (2020). Machine learning based approaches to predict customer churn for an insurance company. In *2020 systems and information engineering design symposium (sieds)* (pp. 1–6).

Hoyer, W. D., MacInnis, D. J., & Pieters, R. (2012). *Consumer behavior.* Cengage Learning.

Hsin Chang, H. (2007). Critical factors and benefits in the implementation of customer relationship management. *Total quality management*, *18*(5), 483–508. doi: 10.1080/14783360701239941

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414–1425. doi: 10.1016/j.eswa.2011.08.024

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515–524. doi: 10.1016/j.eswa.2005.09.080

Hur, Y., & Lim, S. (2005). Customer churning prediction using support vector machines in online auto insurance service. In *International symposium on neural networks* (pp. 928–933).

imblearn. (2023). *Imbalanced-learn documentation.* Retrieved from https://imbalanced-learn.org/stable/

Jamalian, E., & Foukerdi, R. (2018). A hybrid data mining method for customer churn prediction. *Engineering, Technology & Applied Science Research*, *8*(3), 2991–2997. doi: 10.48084/etasr.2108

Kardes, F., Cronley, M., & Cline, T. (2014). *Consumer behavior.* Cengage Learning. doi: 10.1140/epjds/s13688-018-0165-5

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science*, *7*(1), 41.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1), 1–13. doi: 10.1186/s40854-016-0029-6

KhakAbi, S., Gholamian, M. R., & Namvar, M. (2010). Data mining applications in customer churn management. In *2010 international conference on intelligent systems, modelling and simulation* (pp. 220–225). Retrieved from https://www.academia.edu/7708720/Data_Mining_Applications_in_Customer_Churn_Management

Kiguchi, M., Saeed, W., & Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, *118*, 108491. doi: 10.1016/j.asoc.2022.108491

KPMG. (2021, 12). *Analyse van de nederlandse verzekeringsmarkt 2020.* Retrieved from https://kpmg.com/nl/nl/home/insights/2021/12/analyse-van-de-nederlandse-verzekeringsmarkt-20210.html

Kumar, V., & Reinartz, W. (2018). *Customer relationship management.* Retrieved from https://link.springer.com/content/pdf/10.1007/978-3-662-55381-7.pdf

Larsson, A., & Broström, E. (2019). Ensuring customer retention: insurers' perception of customer loyalty. *Marketing Intelligence & Planning*, *38*(2), 151–166. doi: 10.1108/mip-02-2019-0106

Lazarov, V., & Capota, M. (2007). Churn prediction. *Bus. Anal. Course. TUM Comput. Sci*, *33*, 34.

Lee, E.-B., Kim, J., & Lee, S.-G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems*. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/IMDS-12-2015-0509/full/html

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276–286. doi: 10.1509/jmkr.43.2.276

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45. doi: 10.1145/3136625

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*. Retrieved from https://arxiv.org/abs/1912.06059

LightGBM. (2023). *Lightgbm documentation.* Retrieved from https://lightgbm.readthedocs.io/en/latest/index.html

Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659–1665. doi: 10.1109/tii.2012.2224355

Mackenzie, A. (2015). The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, *18*(4-5), 429–445. doi: 10.1177/1367549415577384

Mahajan, V., Misra, R., & Mahajan, R. (2015). Review of data mining techniques for churn prediction in telecom. *Journal of Information and Organizational Sciences*, *39*(2), 183–197.

Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. d. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*. Retrieved from https://arxiv.org/abs/1812.02207

Meena, P., & Sahu, P. (2021). Customer relationship management research from 2000 to 2020: An academic literature review and classification. *Vision*, *25*(2), 136–158. doi: 10.1177/0972262920984550

Mena, C. G., De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2019). Churn prediction with sequential data and deep neural networks. a comparative analysis. *arXiv preprint arXiv:1909.11114*. Retrieved from https://arxiv.org/abs/1909.11114

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, *28*, 92–122. Retrieved from https://link.springer.com/article/10.1007/s10618-012-0295-5?error=cookies_not_supported&code=a09a984f-fd6e-4127-b0f8-9106340c57e8

Merriam-Webster. (n.d.). *Definition of churn.* Retrieved from https://www.merriam-webster.com/dictionary/churn

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & e Cunha, J. F. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert systems with applications*, *39*(12), 11250–11256. Retrieved from https://biblio.ugent.be/publication/3121034

Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, *83*, 326–332.

Mishra, A., & Reddy, U. S. (2017). A novel approach for churn prediction using deep learning. In *2017 ieee international conference on computational intelligence and computing research (iccic)* (pp. 1–4). doi: 10.1109/iccic.2017.8524551

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2021). tcc2vec: Rfm-informed representation learning on call graphs for churn prediction. *Information Sciences*, *557*, 270–285. doi: 10.1016/j.ins.2019.02.044

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021). Relating the partial dependence plot and permutation feature importance to the data generating process. *arXiv preprint arXiv:2109.01433*. Retrieved from https://arxiv.org/abs/2109.01433

Morik, K., & Köpcke, H. (2004). Analysing customer churn in insurance data–a case study. In *European conference on principles of data mining and knowledge discovery* (pp. 325–336).

Motahari, S., Jung, T., Zang, H., Janakiraman, K., Li, X.-Y., & Hoo, K. S. (2014). Predicting the influencers on wireless subscriber churn. In *2014 ieee wireless communications and networking conference (wcnc)* (pp. 3402–3407).

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000, 5). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, *11*, 690-696. doi: 10.1109/72.846740

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21. doi: 10.3389/fnbot.2013.00021

Ndubisi, N. O. (2005). Customer loyalty and antecedents: a relational marketing approach. In *Allied academies international conference. academy of marketing studies. proceedings* (Vol. 10, p. 49).

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, *43*(2), 204–211. doi: 10.1509/jmkr.43.2.204

Ngai, E. W. (2005). Customer relationship management research (1992-2002): An academic literature review and classification. *Marketing intelligence & planning*, *23*(6), 582–605. doi: 10.1108/02634500510624147

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273–15285. Retrieved from https://www.academia.edu/19420670/Credit_card_churn_forecasting_by_logistic_regression_and_decision_tree

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, *43*(1), 99–120. doi: 10.1007/s11004-010-9311-8

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198. Retrieved from https://arxiv.org/abs/1106.0257

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Perišić, A., & Pahor, M. (2020). Extended rfm logit model for churn prediction in the mobile gaming market. *Croatian Operational Research Review*, 249–261.

Perišić, A., & Pahor, M. (2021). Rfm-lir feature framework for churn prediction in the mobile games market. *IEEE Transactions on Games*, *14*(2), 126–137. doi: 10.17535/crorr.2020.0020

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2010, 8). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*, 198-208. doi: 10.1016/J.INTMAR.2010.04.002

Schiffman, L. G., & Wisenblit, J. (2019). *Consumer behavior*. Pearson.

Scriney, M., Nie, D., & Roantree, M. (2020). Predicting customer churn for insurance data. In *International conference on big data analytics and knowledge discovery* (pp. 256–265). doi: 10.1007/978-3-030-59065-9_21

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. doi: 10.1109/jproc.2015.2494218

Smith, K. A., Willis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the operational research society*, *51*(5), 532–541. doi: 10.1057/palgrave.jors.260094

Spiteri, M., & Azzopardi, G. (2018). Customer churn prediction for a motor insurance company. In *2018 thirteenth international conference on digital information management (icdim)* (pp. 173–178). doi: 10.1109/icdim.2018.8847066

Swift, R. S. (2001). *Accelerating customer relationships: Using crm and relationship technologies*. Prentice Hall Professional.

Thaichon, P., & Quach, T. N. (2015). From marketing communications to brand management: Factors influencing relationship quality and customer retention. *Journal of Relationship Mar-*

*keting*, *14*(3), 197–219. doi: 10.1080/15332667.2015.1069523

Torkzadeh, G., Chang, J. C. J., & Hansen, G. W. (2006, 11). Identifying issues in customer relationship management at merck-medco. *Decision Support Systems*, *42*, 1116-1130. doi: 10.1016/J.DSS.2005.10.003

Tsang, I. W., Kwok, J. T., Cheung, P.-M., & Cristianini, N. (2005). Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, *6*(4).

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, *7*, 60134–60149.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. doi: 10.1016/j.simpat.2015.03.003

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, *218*(1), 211–229. doi: 10.1016/j.ejor.2011.09.031

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, *38*(3), 2354–2364. doi: 10.1016/j.eswa.2010.08.023

Wahlberg, O., Strandberg, C., Sundberg, H., & Sandberg, K. W. (2009). Trends, topics and under-researched areas in crm research: a literature review. *international Journal of Public information systems*, *3*, 191–208.

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, *23*(2), 103–112. doi: 10.1016/s0957-4174(02)00030-1

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29–39).

XGboost. (2023). *Xgboost documentation.* Retrieved from https://xgboost.readthedocs.io/en/stable/index.html#

Zhang, R., Li, W., Tan, W., & Mo, T. (2017). Deep and shallow model for insurance churn prediction service. In *2017 ieee international conference on services computing (scc)* (pp. 346–353). doi: 10.1109/scc.2017.51

Zhang, Y., Bradlow, E. T., & Small, D. S. (2013). New measures of clumpiness for incidence data. *Journal of Applied Statistics*, *40*(11), 2533–2548. doi: 10.1080/02664763.2013.818627

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From rfm to rfmc. *Marketing Science*, *34*(2), 195–208. Retrieved from https://www.jstor.org/stable/24544955

Zhang, Y., Liang, R., Li, Y., Zheng, Y., & Berry, M. (2011). Behavior-based telecommunication churn prediction with neural network approach. In *2011 international symposium on computer science and society* (pp. 307–310). Retrieved from https://www.semanticscholar.org/paper/Behavior-Based-Telecommunication-Churn-Prediction-Zhang-Liang/f1e1437ad6cada93dc8627f9c9679ffee02d921c

Zhu, B., Baesens, B., Backiel, A., & Vanden Broucke, S. K. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, *69*(1), 49–65. doi: 10.1057/s41274-016-0176-1

Zhu, B., Baesens, B., & van den Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information sciences*, *408*, 84–99. Retrieved from https://www.semanticscholar.org/

paper/An-empirical-comparison-of-techniques-for-the-class-Zhu-Baesens/cffe585aefee3a82fd1e41d00cd22a44eea02824

# Appendix A

# Extended results

## A.1 Top-decile analysis

Table A.1: Confusion matrix of General Model regarding the top-decile

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 1971 | 3120 | 5091 |
| | Non-churn | 37337 | 350650 | 387987 |
| | Total | 39308 | 353770 | 393078 |

Table A.2: Confusion matrix of Model Approximation regarding the top-decile

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 1960 | 3131 | 5091 |
| | Non-churn | 37348 | 350639 | 387987 |
| | Total | 39308 | 353770 | 393078 |

Table A.3: Confusion matrix of FI-monthly regarding the top-decile

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 2009 | 3082 | 5091 |
| | Non-churn | 37299 | 350688 | 387987 |
| | Total | 39308 | 353770 | 393078 |

Table A.4: Confusion matrix of RFI-yearly regarding the top-decile

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 2011 | 3080 | 5091 |
| | Non-churn | 37297 | 350690 | 387987 |
| | Total | 39308 | 353770 | 393078 |

Table A.5: Confusion matrix of RFI-normalized regarding the top-decile

| | | **Predicted value** | | |
|---|---|---|---|---|
| | | Churn | Non-churn | Total |
| **Actual value** | Churn | 2011 | 3080 | 5091 |
| | Non-churn | 37297 | 34622 | 37297 |
| | Total | 39308 | 353770 | 393078 |

## A.2 Other Analyses

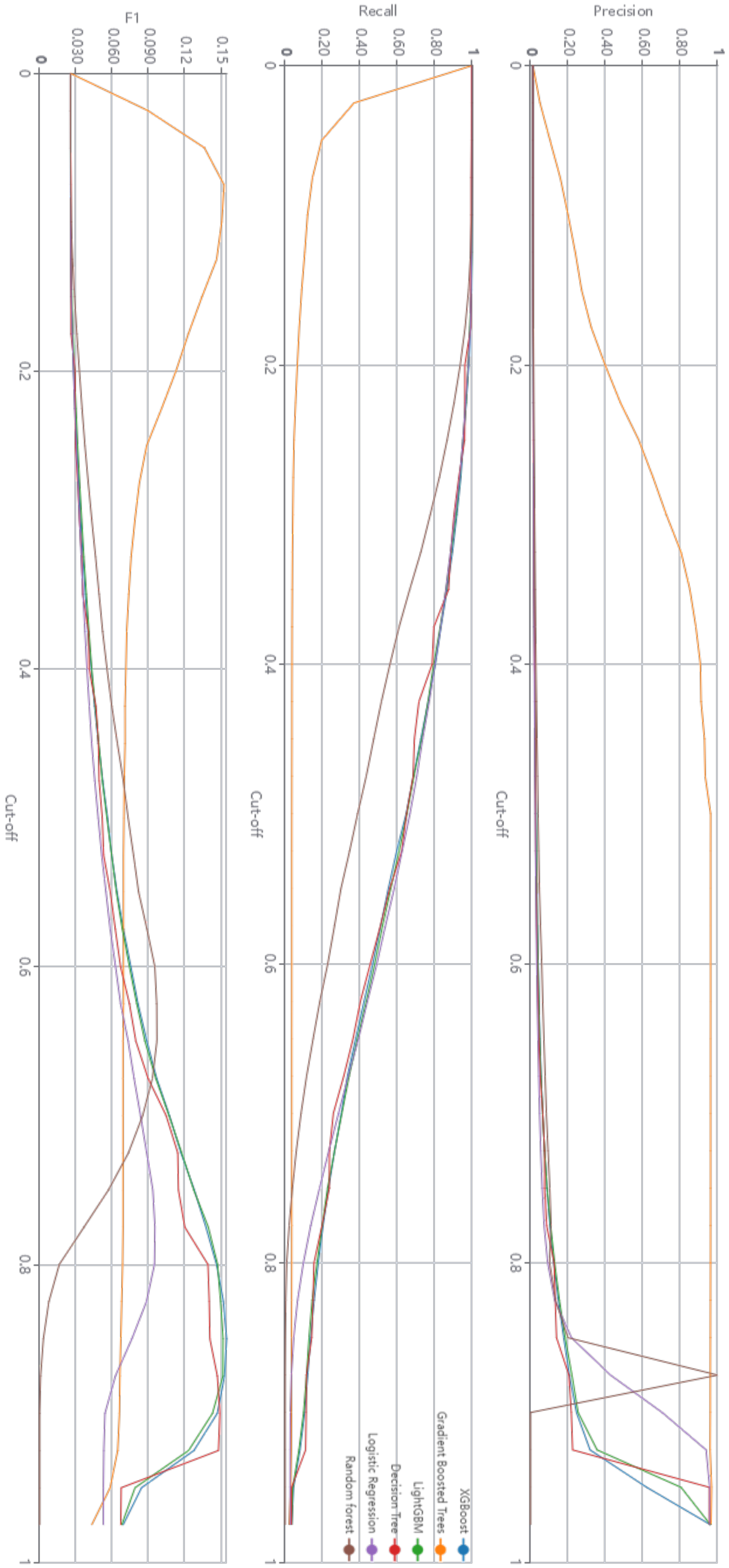

Figure A.1: Decision chart general model

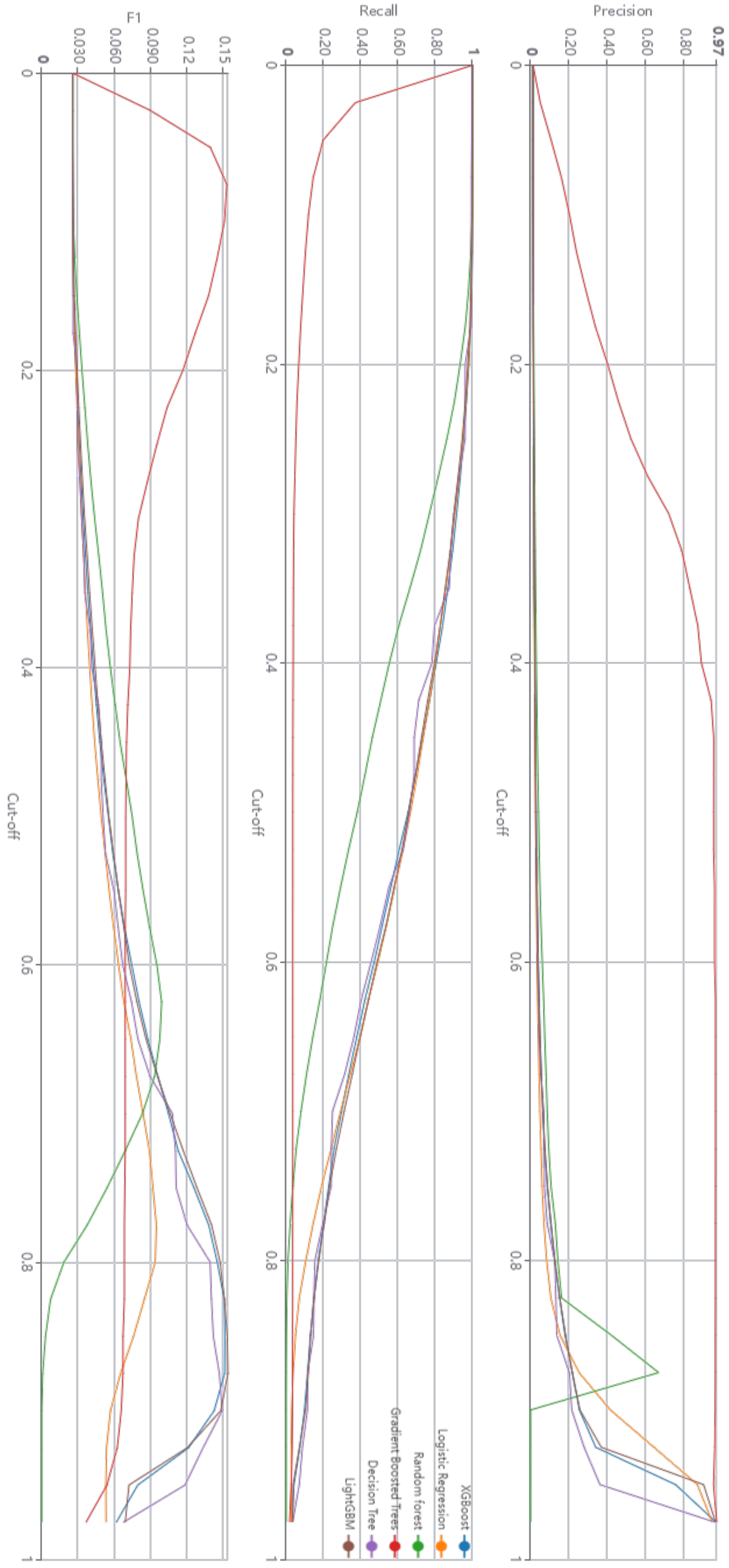

Figure A.2: Decision chart Model Approximation

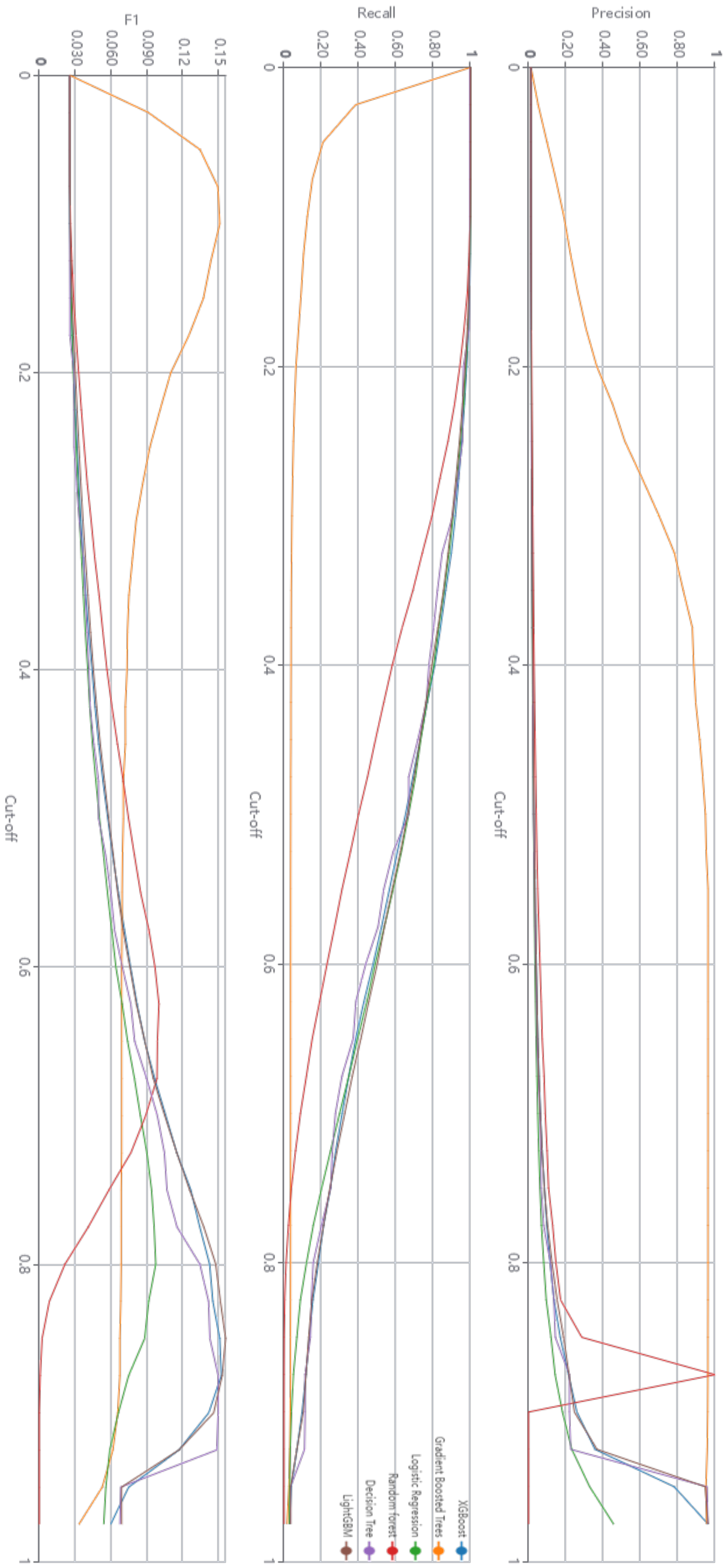

Figure A.3: Decision chart FI-monthly model

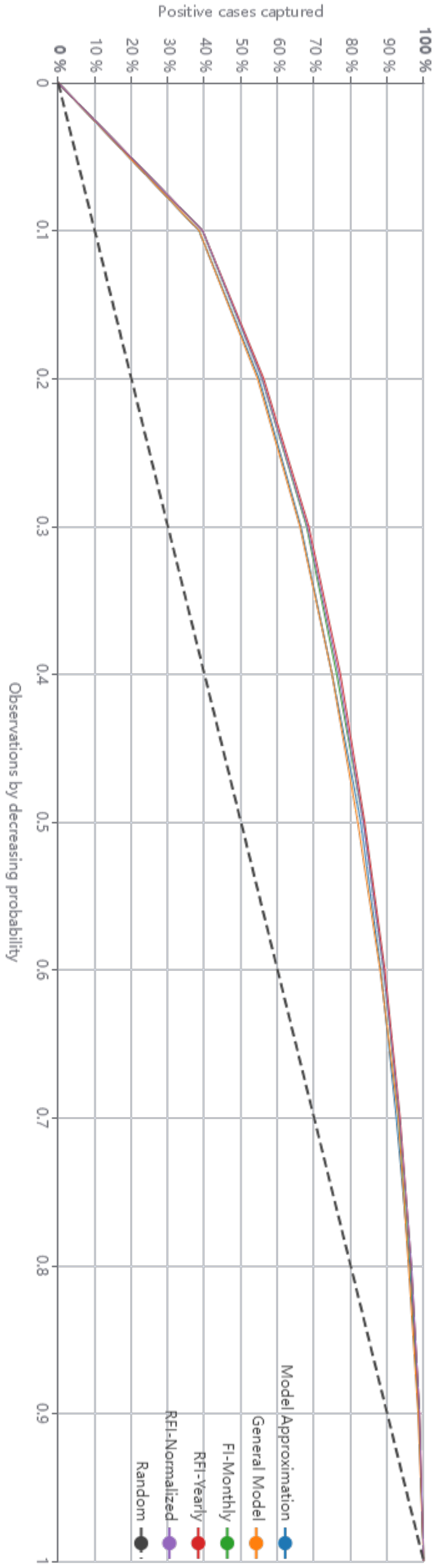

Figure A.4: Lift chart of various models using the LightGBM algorithm

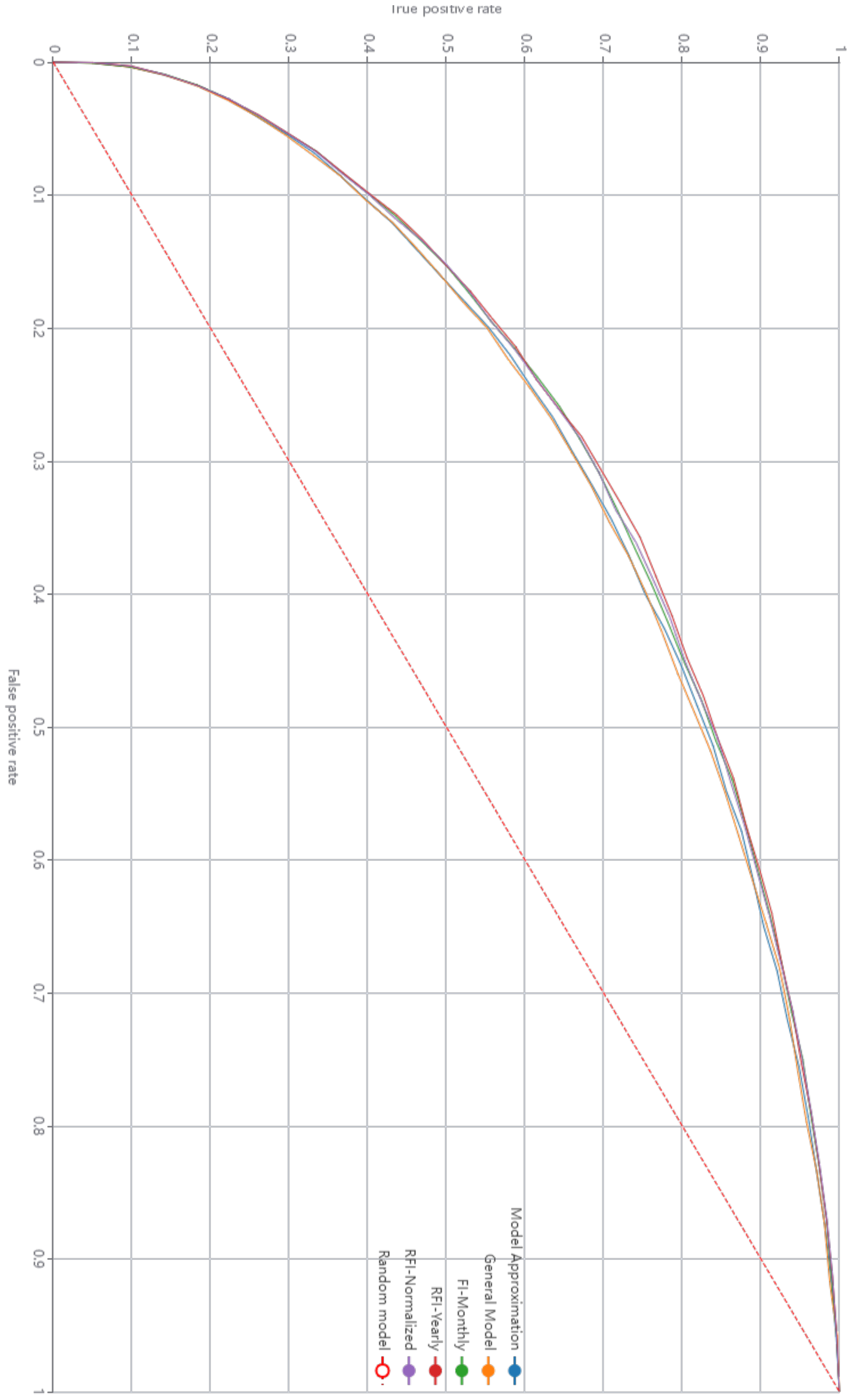

Figure A.5: ROC curve of all models