



# Increasing Retention in Insurtechs Through Churn Prediction

Oskar Christiansen & John Rapp Farnes

DIVISION OF INNOVATION ENGINEERING | DEPARTMENT OF DESIGN SCIENCES  
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY  
2021

MASTER THESIS



Hedvig®

Picture © Hedvig AB

# Increasing Retention in Insurtechs Through Churn Prediction

Oskar Christiansen & John Rapp Farnes



**LUND**  
UNIVERSITY

# Increasing Retention in Insurtechs Through Churn Prediction

Copyright © 2021 Oskar Christiansen & John Rapp Farnes

*Published by*

Department of Design Sciences

Faculty of Engineering LTH, Lund University

P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Innovation Engineering (INTM01)

Division: Innovation Engineering

Supervisor: Lars Bengtsson

Examiner: Jessica Lagerstedt Wadin

# Abstract

Over the last decades, the Swedish insurance industry has seen decreased entry barriers due to deregulation and emerging new technologies, which have the potential to disturb the stagnated and consolidated competitive landscape of the industry. Initiated by newcomers like American insurance startup Lemonade, and later Swedish Hedvig among others, there is an increased push toward digitalization, transparency, and automation in the industry. This thesis examines how Insurtechs can increase retention by identifying customers at-risk of churning, as well as what actions they can take in order to make customers more likely to stay, with the digital insurance company Hedvig as a case study. Various machine learning methods for predicting churn are examined in a literature review, and a model is developed and proposed for Hedvig. Seven levers for increasing retention, 1) Understanding Churn, 2) Customer Intake, 3) Product Improvement, 4) Lock-in, 5) Targeting at-risk Churners, 6) Save Desk, and 7) Organizational Setup, are identified and presented with documented best practices from expert interviews. The conclusion is that churn could not be predicted accurately as the proposed model, a Gradient Boosted Tree model, achieved an ROC value of 62%, which is considered low, and an unsatisfactory precision and recall curve. In the discussion section, we propose that the reason behind this is that there is not enough signal in the data, that the two classes are very homogeneous. In order to improve the predictive accuracy, more usage data from the customers, that have a stronger correlation with the outcome variable, churn, should be collected. Besides predicting churn, the thesis discusses some alternative ways to increase retention, based on discussions with industry professionals, and presents some company specific recommendations in the discussion chapter.

**Keywords:** Non-life insurance, Property and casualty insurance, Customer retention, Churn prediction, Predictive analytics, Classification, Machine learning

# Sammanfattning

Under de senaste decennierna har den svenska försäkringsbranschen sett minskade inträdesbarriärer på grund av avreglering och framväxande ny teknik med potential att förändra branschens stagnerade och konsoliderade konkurrenssituation. Initierat av aktörer som bland andra det nystartade amerikanska försäkringsbolaget Lemonade och senare det svenska Hedvig, finns det ett ökat tryck mot digitalisering, transparens och automatisering i branschen. Det här examensarbetet undersöker hur Insurtechs kan minska kundbortfall genom att identifiera kunder som riskerar att säga upp sina avtal, samt vilka åtgärder de kan vidta för att göra kunder mer benägna att stanna, där det digitala försäkringsbolaget Hedvig utgör en fallstudie. I rapporten undersöks först olika maskininlärningsmetoder för att förutsäga kundbortfall genom en litteraturstudie, och en maskininlärningsmodell utvecklas och presenteras sedan specifikt för Hedvig. Sju spakar att tillgå för att öka bibehållandet av kunder, 1) Förstå kundbortfallet, 2) Kundintag, 3) Produktförbättring, 4) Inlåsnings, 5) Riktade åtgärder mot riskkunder, 6) Save Desk och 7) Organisation, identifieras och presenteras med dokumenterade bästa praxis från expertintervjuer. Sammanfattningsvis kunde kundbortfallet inte förutsägas med tillräckligt hög precision och den föreslagna modellen, en Gradient Boosted Tree-modell, uppnådde 62% på precisionsmåttet ROC, vilket anses vara låg för syftet. I diskussionsavsnittet föreslår vi att orsaken bakom detta, att det helt enkelt inte finns tillräckligt med signal i informationen då kunder som säger upp och de som inte gör det är väldigt lika från företagets perspektiv. För att förbättra precisionen skulle en Insurtech dra nytta av att samla in mer användningsdata från kunderna som har en starkare korrelation med kundbortfall. Förutom att förutsäga kundbortfall diskuterar examensarbetet några alternativa sätt att minska kundbortfall, baserat på diskussioner med branscheexperter, och presenterar några företagsspecifika rekommendationer i diskussionskapitlet.

**Nyckelord:** Skadeförsäkring, Kundlojalitet, Kundbortfall, Prediktiv analys, Klassificering, Maskininläring

# Preface

This thesis examines how Insurtechs can increase retention by identifying customers that are at-risk of churning, and what actions these companies can take to make customers more likely to stay. The thesis is in its nature rather multifaceted, as it on one hand includes data analysis and the development of a statistical models, and on the other hand the understanding of the business aspects and the implications of deploying a model of this kind.

As both authors, Oskar Christiansen and John Rapp Farnes, are on the verge of completing our Master of Science in Industrial Engineering and Management, both with specialization in Financial Engineering and Risk Management, we found the topic of high interest for a multitude of reasons. Firstly, the program itself is rather broad, integrating mathematics and business, a perspective we valued when we first entered the program five years ago. Additionally, our specialization in finance and risk management has enlarged our interest in mathematical modeling and risk and therewith the insurance industry. Ultimately, both authors have during these years had the opportunity to work for various startups, igniting a curiosity for business model innovation and the deployment of new technology in traditional industries. When faced with the opportunity to write our Master's thesis for the Insurtech Hedvig on the topic of customer retention and churn modeling, we felt that it was just the right fit for us

This thesis has been undertaken in collaboration with the Insurtech Hedvig AB and the Department of Design Sciences at the Faculty of Engineering at Lund University. We want to thank Kajsa Alenmyr, Director of Operations at Hedvig, and Akash Patel, Head of Analytics, for their collaboration throughout this thesis. Moreover, we want to extend our gratitude to Lars Bengtsson, for his interest in overseeing our thesis. Finally, we want to thank the long list of interviewees for voluntarily discussing this topic with us and thereby increasing the quality of this thesis.

Lund, May 2021

Oskar Christiansen & John Rapp Farnes

# Table of Contents

<b>1</b>	<b>Introduction .....</b>	<b>9</b>
1.1	Background .....	9
1.2	Purpose .....	14
1.3	Contributions .....	16
1.4	Thesis Structure .....	17
<b>2</b>	<b>Theoretical Framework .....</b>	<b>18</b>
2.1	Retention Management.....	18
2.2	Literature on Churn Prediction.....	21
2.3	Theory on Churn Prediction .....	25
<b>3</b>	<b>Methodology.....</b>	<b>47</b>
3.1	Identify .....	48
3.1.1	Prediction Goal.....	48
3.1.2	Considerations .....	48
3.1.3	Performance Evaluation .....	51
3.1.4	Dataset.....	52
3.1.5	Model Selection.....	55
3.1.6	Churn Drivers .....	55
3.2	Actions.....	56
3.2.1	Research Approach.....	56
3.2.2	Interview Content and Format.....	56
3.2.3	Interview Considerations.....	57
3.2.4	Industry and Company Selection.....	58
3.2.5	Qualitative Analysis .....	61

<b>4</b>	<b>Data Analysis .....</b>	<b>62</b>
4.1	Constructing the Dataset.....	62
4.2	Modeling .....	73
<b>5</b>	<b>Results.....</b>	<b>78</b>
5.1	Identify .....	78
5.1.1	Model Performance .....	78
5.1.2	Churn Drivers .....	80
5.2	Actions.....	81
5.2.1	Impact of Retention Work.....	81
5.2.2	Levers to Increase Retentions.....	83
5.2.3	Best Practices to Increase Retention.....	85
<b>6</b>	<b>Discussion .....</b>	<b>87</b>
6.1	Summary and Conclusion.....	87
6.2	Model Performance .....	89
6.3	Churn Drivers .....	93
6.4	Retention Management.....	95
6.5	Recommendations to Hedvig .....	98
6.6	Transferability .....	104
6.7	Further Research.....	104
6.8	Ethics .....	105
	<b>References .....</b>	<b>107</b>
	<b>Appendix A Participating Respondents in Interview Study.....</b>	<b>113</b>
	<b>Appendix B Interview Guide.....</b>	<b>115</b>
	<b>Appendix C Hyperparameter Optimization .....</b>	<b>117</b>
	<b>Appendix D Churn Drivers .....</b>	<b>127</b>
	<b>Appendix E Best Practices to Increase Retention.....</b>	<b>135</b>



# 1 Introduction

*This chapter introduces the thesis by providing the relevant background of the topics and presenting the research questions and goals of the thesis.*

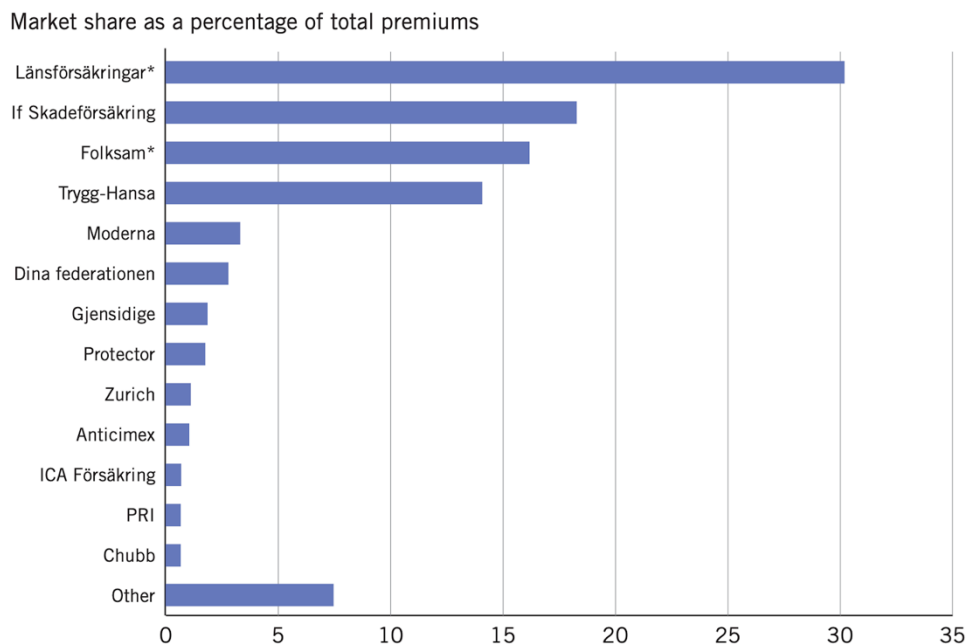
## 1.1 Background

### 1.1.1 The Insurance Market

Insurance as a product and industry has existed for a long time. The earliest trace of insurance dates back to 300 BC in Babylon in the slave trade. Another early record of insurance was in Venice, which provided risk sharing for sea transportation. There are mentions of a Swedish insurance company from the 1200s, insuring against fire damage, so called ‘brandstod’ (Svensk Försäkring, 2021). Non-life insurance today is generally based on a subscription model, where customers pay a fee per period to insure themselves against various risks, rare events that would entail greater costs for them. An insurance company reduces its risk as more customers buy its insurance due to the diversification effect, and the business model is relatively trivial: the insurance premium is set to exceed the average current cost per customer and a risk premium to take on the customers' respective risk.

#### 1.1.1.1 *The Swedish Insurance Market*

In Sweden, the insurance industry has over the past 250 years shifted from a laissez faire approach to stringent regulatory interventions stipulated by the Financial Supervisory Authority. The Swedish insurance market has over time become highly consolidated, partially due to the historically strongly formalized regulatory framework, including strict licensing requirements for new entrants (Lindmark, Andersson, & Adams, 2006). In 2019, out of the total non-life insurance premiums of almost SEK 88 billion, the four largest insurers - Länsförsäkringar, If Skadeförsäkring, Folksam and Trygg-Hansa - accounted for almost 80 percent of the market, as shown in figure 1.1 (Insurance Sweden, 2019).



**Figure 1.1 Swedish market shares for non-life insurance (Insurance Sweden, 2019).**

Although the Swedish insurance industry is still highly regulated, it has since the 1990s seen notable relief. With this deregulation, increased competition has followed, mainly from insurance subsidiaries of Swedish banks and captive insurances of multinational corporations. As an example, ICA Försäkringar is the fastest growing insurance company in the consumer market with a compound annual growth of about 30%, roughly eight times the overall market growth (Dagens Industri, 2020; Ica Försäkring, 2021). Additionally, the industry has been affected by new technology, promising operational efficiency. Together, these changes have somewhat loosened up the barriers for new entrants and increased information cost for the incumbents, e.g., monitoring new competition, customer acquisition and retention activities, and risk assessment (Lindmark et al., 2006).

Apart from lowering the barriers to enter for newcomers, recent regulatory changes have decreased the switching cost in the industry. The Swedish Insurance Contracts Act of 2005 imposed a limit for the binding period of an insurance contract to be no longer than a year, enabling customers to change their insurance provider more frequently than before (SFS, 2005).

#### 1.1.1.2 *Insurtechs*

Insurance has been notably static throughout history and slow to leverage new technologies compared to other industries. The main reasons for the lack of innovation have been state intervention, and the complex nature of the business, coupled with the resulting dominance of the incumbents. This consolidation has

ensured stable profits for a few large players, which therefore have lacked sufficient incentives to reinvent themselves (Yan et al., 2018). However, the industry has recently seen an increased degree of innovation, partly driven by new entrants, and it has been argued that insurance is experiencing a similar ‘wave of disruption’ as that of the financial industry has since the 2008 crisis (Puertas et al., 2017). Initiated by newcomers like American insurance startup Lemonade, and later Swedish Hedvig, there is an increased push toward digitalizing insurance and to increase transparency of operations and in communication, aiming to result in higher customer value and potentially improve both customer acquisition and retention (Dexe, Franke, Nöu, & Rad, 2020). Throughout this paper, we specifically refer to this type of digital non-life insurance providers when we use the term ‘Insurtechs’, even though the term has been used in other contexts with reference to a wider set of companies.

While insurance today is still a highly conservative industry, advancements in e.g., digitalization, information technology and machine learning show promising potential, in terms of efficiency and automation as well as customer experience (Dexe et al., 2020). The environment for new innovative companies to succeed in the insurance industry is optimistic, as these advancements could be highly valuable for companies leveraging them successfully, both in terms of cost leadership and service differentiation. Further, providing improved customer value is long overdue, as insurance is still among the industries with lowest customer satisfaction and loyalty (Dickinson, 2015).

In the light of new innovative insurance startups, the incumbents are forced to reinvent themselves to keep their market positions from the Insurtechs leveraging the digital transformation and advancements in areas such as mobile phones, telematics, Internet of Things, blockchain, cloud computing, artificial intelligence, and predictive modeling (Cappiello, 2020). The potential for these Insurtech startups has been further enhanced by the current low interest rates and the increased amounts of alternative capital in the market, marked by the sudden rapid surge in funding for Insurtech companies (Braun & Schreiber, 2017).

### **1.1.2 The Company**

Hedvig AB is a digital insurance company founded in 2017 with the mission to improve the customer experience in consumer non-life insurance. What the founders identified was that the traditional insurance companies' business models lack incentives to make it easier for customers to receive their payments in the event of claims. Hedvig wants to change this through a new business model where a fixed share of the premiums, 20%, goes to Hedvig and the rest, 80%, goes to a pool to pay out claim reimbursements. This means that Hedvig has an incentive to make the customer experience as good as possible, which they achieve with the help of new technology. With Hedvig, insurance can be purchased within a few minutes through

their app, and claims are handled quickly and efficiently. Additionally, Hedvig goes beyond their competition in terms of flexibility for the customers, allowing customers to cancel their contract at any time, i.e., a binding time of one month rather than the usual annual binding period that we see in the industry. With this flexibility comes an increased significance of offering a high level of customer service and in other ways work on retaining customers by ensuring that they are satisfied and wish to keep their insurance contract, unassisted by any binding clause. Hedvig currently operates solely within home insurance in Sweden, Norway, and Denmark, but have plans to extend this offering both in terms of geographical and contractual coverage. Hedvig offers their service both to rentals, housing cooperatives, and house owners, but to this date students have been one of the main customer segments.

### **1.1.3 Churn and Retention**

In subscription companies, retention is the term used to describe the process of keeping paying customers, and level of retention refers to the share of customers that stays within a measured period, e.g., one month. Churn is the opposite of retention and refers to customers who are leaving, level of churn meaning the number of customers that leave within a measured period (Kumar & Reinartz, 2006). In this thesis, we refer to customers who terminate their contract as a churner, and a customer who is considering, or on their way to churn, as an at-risk churner.

As insurance is based on a subscription business model and is in essence reliant on recurring revenue from premiums, it is important for insurance companies that a large share of their customers keep their contract for as long as possible. Connecting new customers to the platform comes with a customer acquisition cost in the form of marketing and the operational costs of onboarding new users. This makes it important for insurance companies to work actively to retain their existing customers, as they generate recurring revenue without giving rise to customer acquisition costs. The annual average customer retention rate worldwide in the insurance industry is around 83% (Statista, 2018). Studies have shown that the cost of acquiring a new customer is about ten times higher than to retain an existing customer (Daly, 2002), and that, depending on the industry, an increase in customer retention from 85% to 90% produces an increase in net present value profits from 35% to 95%, shedding light on the critical role of retaining customers (Reichheld & Sasser, 1990). Ascarza et al. (2018) highlight this in a different notion, by stating that the average retention elasticity is around 4.9, i.e. that a 1% increase in customer retention gives rise to almost a 5% increase in customer equity. According to Ascarza et al. (2018), 85% of customers have the opinion that companies could do more to retain them and 49% of top executives are unsatisfied with their ability to support their retention goals framework.

#### **1.1.4 Identifying and Targeting at-risk Churners**

One method of increasing retention is to target customers who are at-risk of churning with directed actions. Focusing on the customers who are actually at-risk, rather than targeting all customers, have many advantages including not wasting the marketing budget on incentives to customers who were not considering leaving in the first place. Companies work with this as part of their Customer Relationship Management (CRM) activities, where customer retention is a cornerstone and is arguably the most important component of a customer's lifetime value (Ascarza et al., 2018). In order to target these customers effectively, one first needs to identify which customers are at the doorstep of leaving the company. The ways of identifying these customers are varying, with different levels of sophistication: from using simple heuristics of customer segments who are likely to churn based on experience or company research, to more recent approaches leveraging predictive methods based on the advancements in machine learning.

Even the most advanced predictive machine learning models do not add any value on their own if no actions are taken to affect the at-risk customers. If retention is to be increased, customer retention programs must be developed, and the identified customers targeted in a suitable way. Campaigns or actions that decrease the likelihood of such an identified customer to churn are varied and include e.g. calling, sending notifications, or offering discounts. Finally, churn prediction models have two main purposes: to be able to proactively contact customers at-risk of churning, and to help in the planning of marketing campaigns by understanding the drivers behind churn (Grize, Fischer, & Lützel Schwab, 2020).

#### **1.1.5 Previous Retention Work at the Company**

Hedvig has so far worked with retention almost exclusively in the form of developing an attractive product. Recently, the approach has become somewhat more prioritized, and data driven by examining which types of customers tend to be more likely to churn. As an example, a project at Hedvig recently has been to contact customers with failed payments to better understand the underlying reasons. However, Hedvig has not further in detail examined the drivers of churn, which customers are most likely to leave the platform, the interaction between various drivers, or how they should best allocate their resources to minimize the number of terminated agreements. This has led Hedvig to express a need to investigate customer retention more in depth, and to examine how Hedvig could leverage more advanced methods of identification, i.e., machine learning, in their customer retention efforts. In the long run, they seek a way to be able to identify at-risk customers in a scalable way that can be integrated into their various business processes.

## 1.2 Purpose

The purpose of this thesis is to increase the knowledge of customer retention in Insurtechs, with the digital insurance company Hedvig as a case study.

### 1.2.1 Research Questions

#### 1.2.1.1 *Research question*

The research question of this thesis is: How can an Insurtech increase retention by identifying customers at-risk of churning, and what actions can be taken to make customers more likely to stay?

This research question is further broken down into sub-questions concerning identification of at-risk churners and retention actions that an Insurtech can take in order to make customers more likely to stay.

*Identifying at-risk customers:*

- How accurately can at-risk customers be identified using machine learning at Insurtechs and specifically at Hedvig?
- Which machine learning model and data are the most suitable to predict churn at Insurtechs and specifically for Hedvig?
- How can drivers of churn be extracted at Insurtechs, and what are the strongest drivers for Hedvig?

*Actions to make customers more likely to stay:*

- What are the levers for increasing retention at Insurtechs?
- Within each lever, which actions have proved to be the most effective, and which are best suited to be employed at Insurtechs, specifically Hedvig?

### 1.2.2 Goals

The goal of this thesis is to examine various machine learning methods for predicting churn in Insurtechs. As the information on this topic for Insurtechs to date is very limited, traditional insurance companies, and subscription companies in adjacent industries, are used as a proxy. Further, this thesis aims to compare the performance of different models and methods using the data from Hedvig as a case study, presenting a proof-of-concept model. The model should perform better than a comparable benchmark. Additionally, the aim is to provide an account of the best practices of what actions are best suited to increase retention for Insurtechs, obtained from expert interviews. This account should consist of a set of levers that are used

to increase retention, how these can be used for this purpose, and what impact they can be expected to have on retention. This account should be focused on actions related to predictive modeling.

### **1.2.3 Problem Description**

More specifically, the thesis sets out to make use of Hedvig's customer data in order to create and evaluate statistical models. These models should, as accurately as possible, be able to classify each customer into whether they will terminate their insurance contract in a given period, i.e. the categories "churn" or "not churn", based on the available data on the customer and its behavior. Together with this classification, the prediction should output the probability of belonging to each class for each customer. This model should be accompanied by an interpretation and explanation of how the algorithm understands churn, from which churn drivers can be extracted. The account of industry best practices should be organized within the framework of a set of levers that will be identified.

### **1.2.4 Delimitations**

We will delimit the thesis to customers in Sweden, partly as Hedvig only has limited operations outside, and as the conclusions are assumed to apply to a large extent also in other geographies where Hedvig is or plans to be active in. Hedvig conducts operations in home insurance, both to customers in housing cooperatives and tenancies, and the analysis is therefore focused on this product rather than other non-life insurance categories. We will also limit the churn prediction to voluntary churn, i.e. customer-initiated insurance contract termination. Hence, we exclude involuntary churn, i.e. termination initiated by the insurance company, caused by e.g. failed payments or insurance fraud. Further, we will limit the discussion to business-to-customer companies, offering their services directly to consumers.

There are studies indicating higher profit potential for a company by instead predicting remaining customer lifetime at the company or to model the profit potential, uplift, or expected effect on a customer of the retention action (Lemmens & Gupta, 2013; Ascarza, 2018; Devriendt, Berrevoets, & Verbeke, 2021). In this thesis we focus on the task of predicting the likelihood of churn, as these alternative approaches require more assumptions and data from Hedvig, which is not feasible at this stage.

The most common and most promising models for churn prediction will be evaluated. There exist other models and approaches (e.g. survival models and other time series approaches) that could perform equally well or perhaps better than the ones tested. These methods however require data that is not reliably available at this time and will therefore not be evaluated. Rather, they are possible extensions for

Hedvig to employ in their continued churn modeling work beyond the scope of this thesis.

Finally, it is not within the scope of the thesis to carry out a full statistical analysis in the sense that we do not aim to find statistically significant model parameters, parameter intervals, or cause-and-effect relationships. Rather, a practical iterative approach will be employed to find the model with the best predictive performance for the provided data.

### 1.3 Contributions

This thesis aims to contribute to academia in several ways. Firstly, it aims to present an updated review on how churn is modeled and predicted in the insurance industry and which models and the approaches that have proved to be the most successful at accomplishing this. In addition, rather than only developing and describing a prediction model, this thesis includes a model interpretation section, which is seldom found in the churn prediction literature in the insurance context. Further, the literature review is accompanied by interviews with decision makers who work with customer retention, which will create a broader account on which actions are best employed by Insurtechs in order to increase retention, and what the best practices are in the broader subscription business landscape. Thirdly, this thesis puts a greater emphasis on digitally focused entrants in the traditional industry of insurance, and the interviews hence aim to discuss the actions with highest potential from startup or scale-up companies that are the most relevant for an Insurtech company like Hedvig.



## 1.4 Thesis Structure

**Chapter 2 - Theoretical Framework** introduces the reader to the relevant theoretical frameworks that the thesis builds upon in terms of machine learning.

**Chapter 3 - Methodology** then describes the methodology of identifying at-risk customers, as well as interviewing decision makers in the industry.

**Chapter 4 - Data Analysis** describes the data analysis process of processing the data, transformation of variables, and comparing machine learning models.

**Chapter 5 - Results** presents the results of the predictive modeling and the interview study.

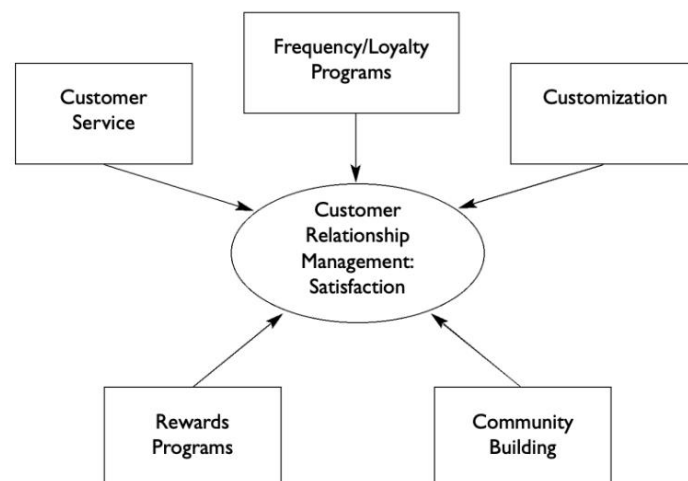
**Chapter 6 - Discussion** discusses the results of the thesis, our recommendations to Hedvig, together with the transferability of the thesis, and suggests areas of further research.

## 2 Theoretical Framework

*This chapter presents the relevant theoretical framework that the thesis builds upon in terms of previous literature on retention management and machine learning.*

### 2.1 Retention Management

Most of the literature on customer retention has been focused on modeling and predicting churn. However, there has been little emphasis on when to target these customers, how they should best be approached, and generally on the design of retention campaigns and the integration of these programs into the overall marketing strategy (Ascarza et al., 2018). This being said, there has been some work in this area. In 2001, Winer presented a framework for customer retention programs shown in figure 2.1, encompassing customer service, frequency/loyalty programs, customization, rewards programs, and community building.



**Figure 2.1 Framework for Customer Relationship Management (Winer, 2001).**

Winer (2001) identifies these five areas as critical in customer retention programs and highlights aspects such as reactive and proactive customer service, increased switching costs through loyalty programs, the creation of products and services for individual customers (referred to as *versioning*), and the impact of building communities that create a more personal bond to the company and to the *family* of other customers.

Later, in 2018, Ascarza et al. underlines that the vast majority of customer retention research has been on churn prediction and less on other important aspects of retention management. The authors present an integrated framework for retention management, shown in figure 2.2, that leverages opportunities such as new data sources and machine learning approaches, provided by recent technological advancements.

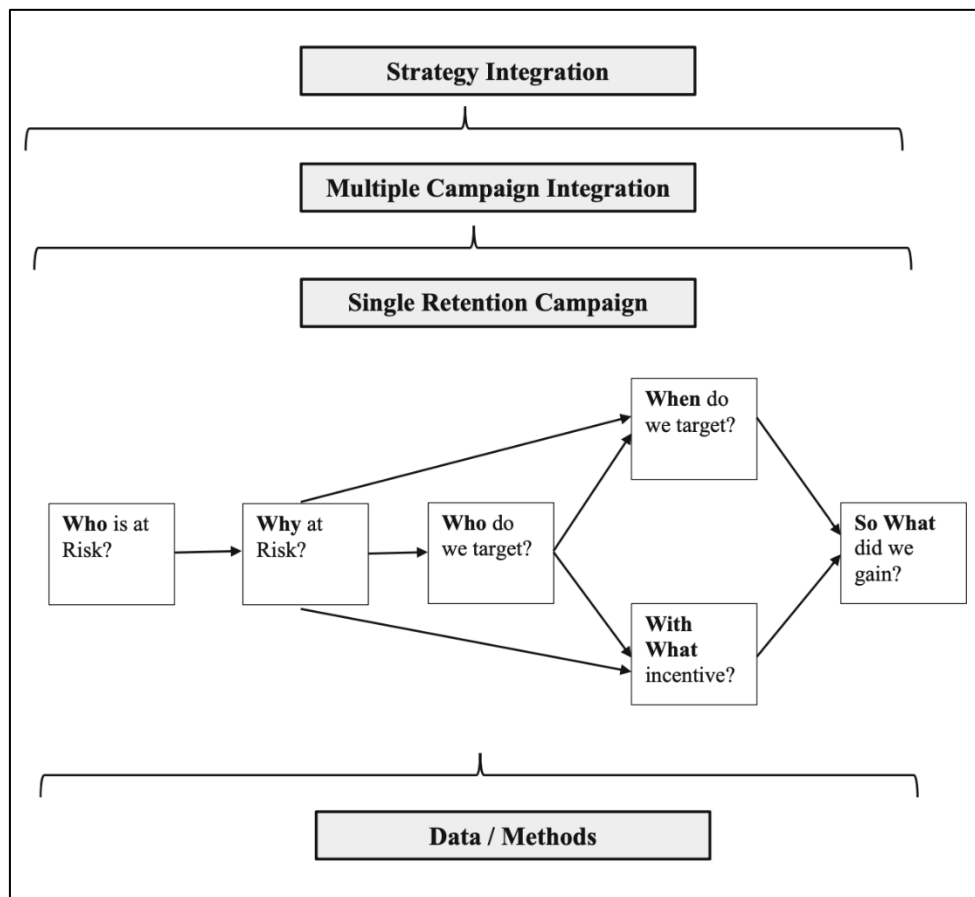
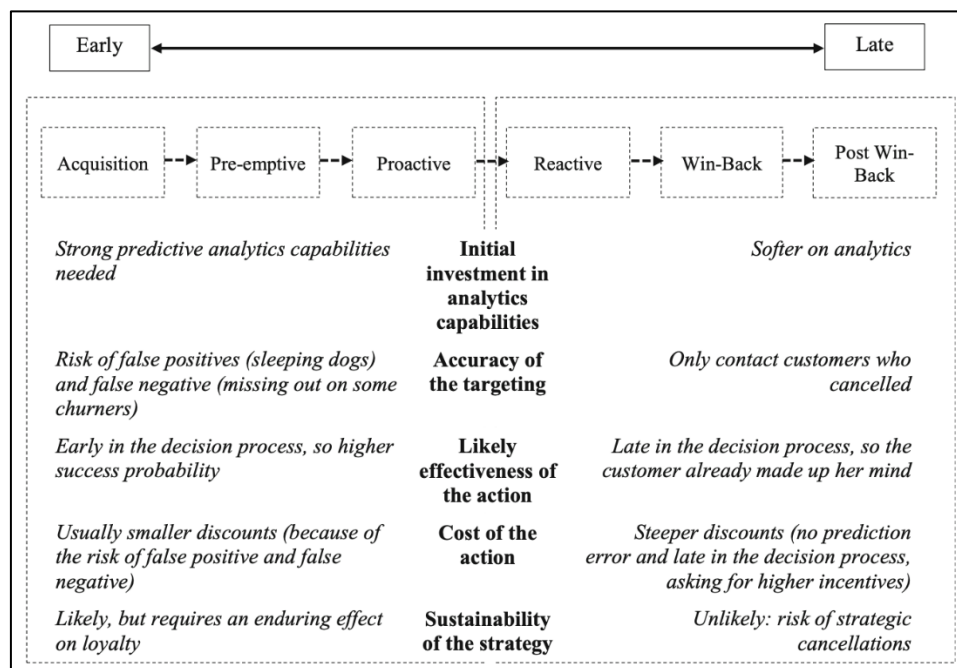


Figure 2.2 Framework for managing customer retention (Ascarza et al., 2018).

Ascarza et al. (2018) presents the framework for managing customer retention with the aim of bridging this gap in the literature on retention. Previous literature on

churn prediction have mainly been focused on identifying who is at-risk, and some efforts have been made, even though limited, on “Explainable AI” as discussed in section 2.2.4 *Drivers of Churn in Insurance*, dealing with why the identified customers are at-risk. The framework presented by Ascarza et al. (2018), however, also incorporated the equally important questions regarding who should be targeted, with what initiatives, and when this should be carried out. The topic of who should be targeted is discussed in section 2.3.4 *Machine Learning Models for Churn Prediction*, has begun to gain interest in the field, and had been discussed by e.g. Lemmens & Gupta (2013), Devriendt et al., (2021), and by Ascarza (2018), one of the authors behind this framework. These studies suggest that the customers that are at the highest risk of churning are not necessarily those that should be targeted, wherefore uplift modeling has been proposed. Further, the question regarding what initiatives should be used and when these should be employed extends previous literature and the authors suggest that it is key to distinguish between reactive and proactive campaigns, as illustrated in figure 2.3.



**Figure 2.3 Framework for timing of retention campaigns (Ascarza et al., 2018).**

Reactive campaigns refer to initiatives carried out after a customer has initiated the contract termination with the aim to ‘save’ these customers and is typically achieved by financial incentives. Proactive campaigns on the other hand, aim to target the underlying issue that will lead to churn before it happens. A firm typically uses a combination of discrete proactive campaigns and reactive ones, and it is critical that these are coordinated and integrated with the firm’s overarching marketing strategy (Ascarza et al., 2018).

## 2.2 Literature on Churn Prediction

### 2.2.1 History of Churn Prediction

Recent advancements in machine learning have laid the foundation for a multitude of new predictive methods for identifying at-risk churners. The area has been studied extensively, especially in the telecommunication industry, due to its high competitiveness and its historical liberalization in many parts of the world (Wei & Chiu, 2002; Hung, Yen, & Wang, 2006). An early approach in the industry was to use empirical Bayes methods (Greis & Gilstein, 1991). After that, a multitude of different modeling approaches have been attempted and compared, including bagging and boosting (Lemmens & Croux, 2006) and Monte-Carlo simulations for optimal parameter combinations (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015). More recently, the topic of customer churn modeling has been extended to other industries such as e-commerce (Yu, Guo, Guo, & Huang, 2011), gaming (Castro & Tsuzuki, 2015; Milošević, Živić, & Andjelković, 2017), financial services (Larivière & Van den Poel, 2004), and insurance (Morik & Köpcke, 2004; Risselada, Verhoef, & Bijmolt, 2010; Günther et al., 2014; Boucher & Couture-Piché, 2015; Bolancé, Guillen & Padilla-Barreto, 2016; Zhang, Li, Tan, & Mo, 2017; De la Llave, López, & Angulo, 2019; Henao Madrigal, 2020). Zooming in on Swedish academia and the major Swedish engineering faculties in particular, churn prediction has been a popular topic for Master's thesis projects during the past couple of years, primarily within telecommunications, fintech, gaming, and streaming services.

### 2.2.2 Machine Learning Models for Churn Prediction

There are many different approaches for modeling classification problems, each with its advantages and drawbacks. As different models outperform others in different settings, it is not possible to state a single model to be superior to all others in modeling churn irrespectively of industry and company settings. As such, it is more interesting to study the popularity of various models in order to get an understanding of the developments in the area. In a machine learning competition in 2005, where 33 academics and practitioners with experience in customer churn modeling, the 44 submitted entries explored on average 3.3 different estimation techniques. The most popular models were logistic regression (used by 45%), decision trees (23%), and neural networks (11%) (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Based on this contest, the authors argue that the most commonly used models in the submissions, logit and tree approaches, are reasonable models to start off with for most churn modelling problems, as they are widely used and tend to

perform relatively well (Neslin et al., 2006), at the same time as they are fairly simple and have a high degree of interoperability (Günther et al., 2014).

Historically, statistical techniques such as logistic regression (Burez & Van den Poel, 2007; Brockett et al., 2008; Risselada et al., 2010; Günther et al., 2014; Bolancé et al., 2016) were mainly used to predict churn, largely due to the simplicity of the models and their interpretability in terms of odds ratios (Günther et al., 2014). Standard regression models, however, assume linear relationships between the explanatory variables and the log odds. This is a very strong assumption that is rarely fulfilled in practice. In order to mitigate these shortcomings, alternative models such as generalized additive models (GAM) have been attempted with data from a Norwegian insurance company (Günther et al., 2014). Another study took this further and made use of Multivariate Adaptive Regression Spline (MARS), a method of automating the selection of nonlinear terms in regression models (Friedman, 1991; James, Witten, Hastie, & Tibshirani, 2013 p. 145-151; De la Llave et al., 2019).

Later, algorithmic techniques such as decision trees (Morik & Köpcke, 2004; Hung et al., 2006; Risselada et al., 2010; Yu et al., 2011; Bolancé et al., 2016), neural networks (Hung et al., 2006; Yu et al., 2011), ensemble methods, e.g. random forests (Burez & Van den Poel, 2007), and support vector machines (Morik & Köpcke, 2004; Yu et al., 2011; Zhao, Li, Li, Liu, & Ren, 2005; Bolancé et al., 2016) have been deployed for this purpose. More recently, alternative approaches such as deep learning methods and bagging and boosting techniques (Lemmens & Croux, 2006; Risselada et al., 2010) have been developed and in many cases outperformed established classification techniques (Vafeiadis et al., 2015; Grize et al., 2020). In a churn modeling competition in 2018, a model using extreme gradient boosting (XGBoost) achieved first-place out of 575 contestants. The model combined the two frequently used gradient boosted tree algorithms XGBoost and LightGBM, leveraging recent advancements in machine learning with more complex but often more accurate techniques (Gregory, 2018).

Other approaches have been attempted such as Naive Bayes (Morik & Köpcke, 2004), hazard or survival models (Bolton, 1998; Burez & Van den Poel, 2007; Brockett et al., 2008; Henaó Madrigal, 2020), association rule learning (Morik & Köpcke, 2004), spatial probit models (De la Llave et al., 2019), and a hybrid approach by combining a neural network with decision trees (Hu, Yang, Chen, & Zhu, 2020). Finally, recent studies have argued for predicting each customer's remaining lifetime at the company. Another suggestion is to use prescriptive analytics, and more specifically uplift model, i.e. expected effect on a customer of a given retention action (Lemmens & Gupta, 2013; Ascarza, 2018; Devriendt et al., 2021). The argument behind this is that modeling uplift better aligns with the actual business objective as they not only predict whether a customer will churn, but also if they will be retained when targeted. However, no comparative studies have been performed of uplift models versus predictive models, and hence there is no empirical

evidence of uplift models outperforming predictive models in this context (Devriendt et al., 2021).

### **2.2.3 Churn Prediction in the Insurance Industry**

Customer churn has as mentioned been a critical aspect also in the insurance industry. Customer loyalty was investigated in the insurance industry in the 1980s, examining relationship marketing and insurance consumers' satisfaction with the contact person, the core service, and the institution (Crosby & Stephens, 1987). Following this, the effect of relational constructs (e.g., satisfaction, trust, and affective and calculative commitment) on customer referrals and purchasing behavior was studied by Verhoef, Franses, and Hoekstra (2002), using a Poisson regression model.

An early modeling approach introduced a two-stage segmentation process in order to identify and target health insurance groups effectively, yielding a 7% increase in customer retention (Cooley, 2002). Another paper compared the methods decision tree, support vector machines, naive bayes, and association rule learning, using a data transformation method from information retrieval to create time-related features, which outperformed models not including them (Salton & Buckley, 1988; Morik & Köpcke, 2004). Different methods, logit models and classification trees, both with and without applying a bagging procedure, were compared, showing highest performance for classification trees with bagging procedure, but overall low staying power, as predictive performance deteriorates considerably with time (Risselada et al., 2010). Additionally, traditional regression approaches were extended to account for nonlinearities using GAM in the preprocessing of data from a major Norwegian insurance company (Günther et al., 2014). After this, an approach was tested using queuing theory to model the rate of insurance policy cancellation for a car insurer (Boucher & Couture-Piché, 2015). Other approaches, such as traditional logistic regression, decision trees, and support vector machines, were compared in predicting churn for a home insurance company, resulting in the support vector machines model outperforming the others (Bolancé et al., 2016).

Recently, efforts have been made on insurance data to combine the memorization power of shallow models such as logistic regression with the generalization capability of deep models, resulting in a proposed deep and shallow model with improved performance compared to many of the previously proposed approaches (Zhang et al. 2017). Another approach has been to make use of the spatial information of customers, showing that customers in close proximity to their insurance company's office had a lower risk of churning, while customers in the surroundings of competing firms' offices had a higher risk of churning. The same study found spatial autocorrelation in churn probability, indicating a higher churn risk if other customers close by also churned (De la Llave et al., 2019). Another elaborate approach was carried out where their customers' information was linked

to anonymous online quotes made on their website, concluding that customers that had requested a quote online were easier to predict whether they were at-risk of churning or not (Mau, Pletikosa, & Wagner, 2017).

Finally, it is worth mentioning that churn prediction in insurance is in essence an imbalanced problem, where the number of non-churners in the data outnumber the number of churners (Fernández et al., 2018, p. 35). The annual average customer retention rate worldwide in the insurance industry is around 83%, which corresponds to a monthly churn rate of around as little as 1-2% (Statista, 2018). In order to achieve high performing models in these settings, considerations must be made in the modeling stage, as the minority class otherwise tends to be overlooked (Fernández et al., 2018, p. 19).

#### **2.2.4 Drivers of Churn in Insurance**

The selection of variables included in the churn models in previous papers are usually first and foremost chosen by availability, and most of the variables are rather self-explanatory such as demographic variables (age, gender, partner), policy attributes (policy types, number of policies, premium, discount, duration of customer relationship), and other information e.g. if the customer owns several homes, if the customer has rejoined the company, or if the discount has expired (Günther et al., 2014).

When constructing features for the classification task, time series variables such as the timeline history of claims must be transformed into a tabular format. For this task, there have been several approaches that have been proposed. Salton and Buckley (1988) proposed a data transformation method from information retrieval called frequency-inverse document frequency. The winning entry to a churn modeling competition in 2018, transformed such information into relative features measuring e.g. the time since the last event, and absolute features measuring e.g. the day of the year that a particular event happened (Gregory, 2018).

When considering the structure of the data, recent research has studied the effect of granularity of data and confirmed that it has a significant impact on both the performance of the predictions but also the model selection and presents an approach to generate and impute missing data (Scriney, Nie, & Roantree, 2020). The literature covering churn prediction in the insurance industry generally focuses on developing the most accurate model rather than interpreting the model and understanding the churn drivers. A recent study related to insurance, conducted by Gramegna and Giudici (2019), interprets their model using methods within the growing field of “Explainable AI”. This project does not model churn however, and no papers taking advantage of this approach were found in our review for churn prediction in the insurance industry.



### 2.2.5 Summary

In summary, our review of related works shows that churn has been successfully predicted in the past, and that a variety of approaches have been tried for the task. For identification, various models have been attempted in predicting customer churn in the insurance industry. Depending on the particular study, the industry or the company, different modeling techniques have proven to be best suited to explain the churn behavior, and therefore there does not exist a general solution for this problem that always results in the best predictive performance.

Within churn classification, some models have however been used extensively in the literature and proven to give high predictive performance. These include logistic regression (LR), tree-based methods, i.e. decision trees (DT), random forests (RF) and gradient boosted trees (GBT), support vector machines (SVM), and neural networks, including the most basic form called feed-forward neural networks (FFNN). Among these, nonlinear kernel SVM, bagging and boosted tree-based methods as well as FFNN tend to outperform logistic regression and are expected to perform well for Insurtech churn prediction. However, because of the simplicity and interpretability of LR, it is included in this project to serve as a benchmark for the other more advanced models.

## 2.3 Theory on Churn Prediction

### 2.3.1 Supervised Learning and Classification

In this thesis, the problem of churn prediction is viewed as a classification problem, where customers are to be classified into two groups: “churn” or “no churn”. Additionally, customers may be not only strictly put in a category, but rather assigned a probability of classes. These classes can for example be encoded by the labels 0 and 1. Hence, to encode probabilities, the output of the model is a number between 0 and 1 encoding the probability of the customer belonging to the 1, or “churn” class.

#### 2.3.1.1 Training

At a high level, this kind of classification problem belongs to the field of supervised learning, where an unknown function  $f(X) = Y + \epsilon$  is to be found or approximated, where  $Y$  is the target classification variable, and  $X$  is the data, and  $\epsilon$  the irreducible error in the data. For this problem, a wide range of models are available. These models can generally be described as a set such functions  $f$  where each class includes a set of parameters. In supervised machine learning, these parameters are determined in order to best approximate  $\hat{f}(X) \approx Y$  according to some measure of

fit, i.e. a loss function. This process of determining model parameter is known as “training”, and is performed by using a training set of examples  $(X_{train}; Y_{train})$ , which an algorithm uses to minimize the loss function by adjusting the parameters, and hence resulting in the best approximation  $f(X_{train}) = Y_{train}$ . For most algorithms, this minimization is done by a combination of numerical optimization and heuristics encoded in the algorithms (James et al., 2013 p. 15-58).

### 2.3.1.2 Testing and Validation

As the goal of machine learning is to approximate  $f(X)$ , careful consideration must be made in the process in order to generate a function that generalizes to data unseen by the model, beyond the training set. When a model performs well on the training set but does not generalize beyond it, it is referred to as overfitting. Overfitting is part of the bias-variance trade-off and can be mitigated with less flexible models or early stopping. In order to evaluate the true performance of the model on unseen data, the model is evaluated on a separate dataset called test set. This mitigates the problem of optimism of the training error rate (James et al., 2013 p. 15-58).

Another frequently used measure to evaluate the performance of a model more accurately, is to use validation. This is used for tuning and model selection. With validation, the model is trained on the training set and decisions regarding model selection and parameters are made based on the performance on a validation set. During this process, the test set is hidden from the model and researcher and only used in the final step of assessing the performance on unseen data. With a simple three-set validation approach, the data is randomly split into the train, validation, and test part with some chosen fraction, e.g., 50%, 25%, 25% (James et al., 2013 p. 175-189). This three-set validation approach is shown below in figure 2.4.



**Figure 2.4 Three-set validation and test approach (Hastie, Tibshirani, & Friedman, 2009, p. 222).**

Based on this idea, there are several other methods that try to address the bias-variance trade-off of validation, including leave-one-out cross-validation and  $k$ -fold cross-validation. In  $k$ -fold cross-validation, the training set is split into  $k$  folds, and sequentially the model is trained on all but one of the folds and the performance is evaluated on the left-out fold. This process results in  $k$  measures of performance, which can then be compared across models e.g., by taking the average performance on the folds (James et al., 2013, p. 176-186).

### 2.3.1.3 Hyperparameters

In addition to the model parameters, which are learned in the training process, machine learning models tend to have hyperparameters, which must be determined outside of the training process. Usually, these hyperparameters are determined by

optimization, where a set of models are trained with different sets of parameters, and the best performing set of parameters is chosen. This set of parameters can be chosen by a variety of methods.

The most straightforward method is called grid search, where a set of possible values of each parameter is chosen, and the algorithm then evaluates the performance of each combination of these parameters in order to determine the best set. As the number of combinations of parameters quickly can grow with this method, random sampling approaches can be employed (James et al., 2013, p. 227-228). This random sampling can either be employed by naive random draws of parameters from a specified parameter space, or more “smart” algorithms can be deployed, using sampling theory from e.g., Monte Carlo methods and Bayesian statistics (Vafeiadis et al., 2015; Ma, Tan, & Shu, 2015). The performance of different sampling algorithms has been studied, and it has been shown both empirically and theoretically, that random search is able to find better performing models given the same computation time, hence this method will be used in this project (Bergstra & Bengio, 2012).

### **2.3.2 Performance Evaluation**

In order to measure the accuracy of a prediction model, and compare the performance of different models, accuracy must first be defined. In a classification context, one model might for example classify all points as 1, leading to perfect classification of the 1 class, but poor classification on the 0 class. Whether this is a good or poor prediction depends on the metric used. For this purpose, a series of classification metrics have been developed. The best metric to employ depends on the problem at hand, as different types of errors have different costs in different applications.

#### *2.3.2.1 Confusion Matrix*

The building block of most classification metrics is the confusion matrix. This matrix has four cells, covering all combinations of predicted and actual values. The confusion matrix is illustrated in figure 2.5 below. In this figure, TP stands for True Positive, FP stands for False Positive (also called Type I error), FN stands for False Negative (also called Type II error), and TN stands for True Negative (Fawcett, 2006).

		<u>True class</u>	
		<b>p</b>	<b>n</b>
<u>Hypothesized class</u>	<b>Y</b>	True Positives	False Positives
	<b>N</b>	False Negatives	True Negatives

**Figure 2.5 The confusion matrix (Fawcett, 2006).**

### 2.3.2.2 Accuracy

Accuracy measures the percentage of correctly classified items, and is calculated as following:

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

Accuracy is a commonly used metric in classification problems. However, it does not handle imbalance problems such as churn prediction well (Fawcett, 2006). In imbalanced problems, the positive “churn” class is usually very small in relation to the negative “no churn” class. As such, a high accuracy can be achieved by simply classifying all observations as “no churn”. Therefore, imbalanced problems call for other metrics of performance evaluation.

### 2.3.2.3 Precision and Recall

Two measures, which can be better employed for imbalanced problems, are precision and recall (Fawcett, 2006).

Precision is the percentage of positive predictions that are truly positive (Fawcett, 2006). In the churn context, precision can be interpreted as the share of observations that the model predicts as churning that will actually churn. The equation for precision is:

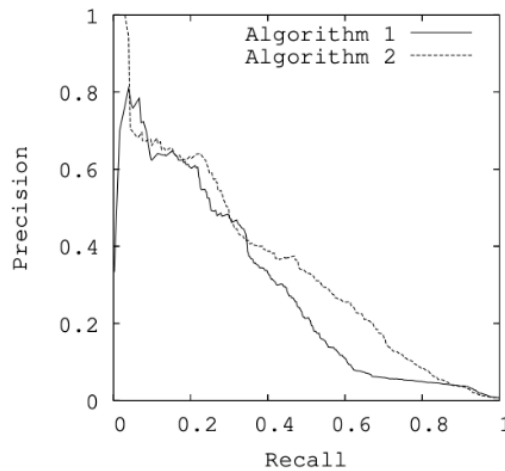
$$\text{Precision} = \frac{TP}{TP + FP}$$

A problem with only measuring precision is that very conservative models would be favored. For example, a model that would only classify one customer as churning, but be very confident and right about its prediction, would get 100% precision. This classifier would however perform badly on the majority of customers. As such, precision is often measured together with recall, the percentage of the positive observations that were correctly predicted, also called sensitivity (Fawcett, 2006).

In the churn context, recall can be interpreted as the share of churners “found” by the model:

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

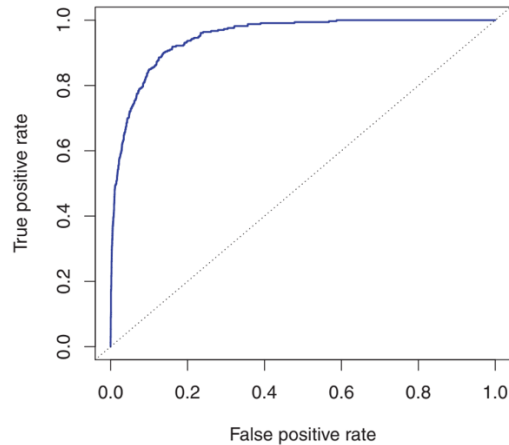
Precision and recall tend to have an inverse relationship, where a model that is more conservative in labeling churn tends to have high precision and low recall, while a model that labels most observations as churn tends to have high recall and low precision. For models that output not only a label, but an estimated probability of churning, the precision and recall can be measured for different probability thresholds, resulting in a curve of possible precision, recall combinations possible with the model. This is called a Precision-Recall curve, and an example is provided below in figure 2.6 (Fawcett, 2006).



**Figure 2.6 Example of Precision-Recall curves comparing two models (Davis & Goadrich, 2006, p. 234).**

#### 2.3.2.4 ROC Curve and AUC

In addition to precision and recall, a frequently used tool to study the performance for binary classifiers is the Receiver Operating Curve (ROC). Similarly to the precision-recall curve, this curve is constructed by varying the probability threshold, whereby a series of points are generated. In the ROC curve, recall (also known as the true positive rate or TPR) is plotted against the false positive rate  $FPR = \frac{FP}{FP+TN}$ . An example of an ROC curve is shown below in figure 2.7.



**Figure 2.7 Example of an ROC curve (James et al., 2013, p. 148).**

In this curve, the perfect classifier would only have a point in the upper left corner, with a 100% TPR and 0% FPR. The dotted line in figure 2.7, which goes from (0,0) straight to (1,1) represents random guessing.

In order to aggregate the ROC to one number describing the performance of a classifier, the Area Under the Curve (AUC) is often taken, often calculated as the interpolated numerical integral of the curve. With this measure, a value above 0.5 means performance over random guessing, while a 1 represents a perfect classification (Fawcett, 2006). In this project the area under the ROC curve is referred to as the AU-ROC.

### 2.3.2.5 Comparing PR and ROC Curves

Both PR and ROC curves are commonly used in the churn prediction setting. Some argue that the PR curve and the area underneath is more suitable for cases where imbalance is large (Davis & Goadrich, 2006). In this project however, we have opted to use the AU-ROC as the main performance metric, and the PR curve only to differentiate between models with the same or similar AU-ROC. The reason why we choose to not compute or present the AU-PR is because this could reveal the underlying churn rate of Hedvig, which the company has requested to anonymize. Consider for example a benchmark model that randomly guesses which customers will churn or not. This model will have the overall churn rate as precision for all levels of recall, and the AU-PR will hence be equal to the underlying churn rate.

### 2.3.3 Class Balancing

Previous studies have shown that classification problems with severe imbalance between classes, one class being more frequent than the other in the dataset, may

lead to challenges in developing models (Fernández et al., 2018, p. 35). This is in particular true for churn modeling in the insurance industry because of the low churn rate. The intuition for why models fail to learn properly on datasets with severe imbalance, is that models learn that most examples belong to the majority class rather than focusing on the differences that tell the classes apart. In general there are three ways to deal with this issue and restore balance to classes (Kuhn & Johnson., 2016, p. 419-444):

1. Achieve balance in data collection by measuring an equal number of observations.
2. Re-balancing data by either undersampling the majority class or oversampling the minority class, or a hybrid approach during training.
3. Model specific approaches that weight the classes differently or sample from them with different frequencies to correct for the imbalance during training. These methods are described in Fernández et al. (2018, p. 123-196).

As the churn rate observed in the data set is the true churn rate at the company, the first option is not applicable to this case. Rather, this section will focus on the second point, as they are general model agnostic balancing methods, while the third method is model specific. For the second option, a series of methods have been proposed and compared. In this paper, we will include the simpler models random over and under sampling, as well as Synthetic Minority Oversampling Technique (SMOTE), as these are methods that are fairly simple to implement for the scope of this paper and have shown good results in previous studies (Kuhn & Johnson, 2016, p. 419-444; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In addition, previous research has shown that there is no clear winner among sampling methods, but rather that the best method depends on the dataset and model at hand (Kuhn & Johnson, 2013, p. 429; Fernández et al., 2018, p. 79-122). For an extension of the project, more class balancing approaches could be tried, such as One Class SVM (OCSVM) or hybrid  $k$ -NN sampling, which have shown indications of increased performance (Fernández et al., 2018, p. 79-122). Additionally, there have been many proposed extensions to SMOTE, which are not presented or used in this project (Fernández et al., 2018, p. 101-113).

In order for the performance evaluation to be as unbiased as possible, the sampling methods are not used when predicting, rather the original data is used in this context. Balancing is only performed during the training stage, in order to increase the model understanding of the difference between classes (Kuhn & Johnson, 2016, p. 429).

#### 2.3.3.1 *Random Under Sampling*

A naive method of balancing is to randomly under sample from the majority class until the classes are equal in occurrence. Advantages of this approach is that all samples are real customers, and no duplicates in the data. The disadvantage of the

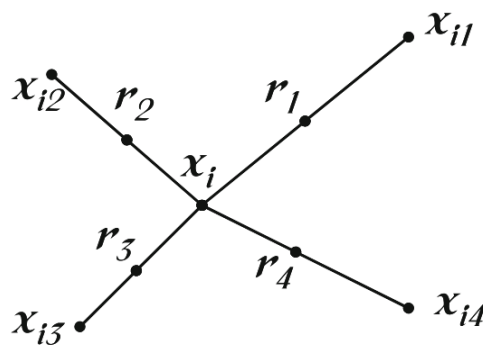
approach on the other hand is that observations are removed, and hence only a subset of the data is used to train the model, which can be especially problematic for small datasets often found in startups such as Hedvig with a relatively small customer base.

### 2.3.3.2 *Random Over Sampling*

In contrast to random under sampling, random over sampling oversamples the minority class with replacement, effectively duplicating observations in the minority class. Advantages of the approach is that all samples are used, however the resulting dataset will contain duplicate observations, which can lead to overfitting on the minority observations at hand (Fernández et al., 2018, p. 80).

### 2.3.3.3 *SMOTE*

In order to overcome the problem of overfitting to the minority observations, SMOTE has been proposed. Rather than using duplicates of the minority class samples, SMOTE generates synthetic samples from the existing observations, reducing the problem of overfitting. The SMOTE algorithm iteratively selects a sample randomly from the minority class to base a synthetic sample on. From this selected sample, the  $K$  nearest neighbors (where  $K$  is a parameter of the algorithm, 5 being the default value originally proposed) is identified. Next,  $K$  synthetic samples are created by interpolating between the selected sample each of the neighbors, by adding a random number  $\in (0,1)$  multiplied by the difference between the two points (Fernández et al., 2018, p. 98-101). This process is visualized in figure 2.8, where  $X_i$  is the selected point and the points  $r_k$  are the generated samples.



**Figure 2.8 Visualization of SMOTE (James et al., 2013, p. 148).**

Advantages with SMOTE compared to random oversampling, is that the risk of overfitting is lower since the synthetic samples are not direct duplicates of observations. Nevertheless, a disadvantage is that the generation of synthetic samples is fairly naive, done with simple linear interpolation which for example



may lead to unrealistic observations, and in some cases interfering with the classification boundary between the classes (Fernández et al., 2018, p. 98-101).

### 2.3.4 Machine Learning Classification Models for Churn

As discussed in section 2.2 *Literature on Churn Prediction*, a multitude of different modeling approaches have been attempted in order to identify at-risk churners, and there is no general model that has proven to always perform the best. That being said, there are some models that have been used more frequently throughout the literature in a churn classification context and that have yielded high predictive performance, which we will focus on in this thesis:

- Logistic Regression (LR)
- Tree-based methods: Decision Trees (DT), Random Forests (RF) and Gradient Boosted Trees (GBT)
- Support Vector Machine (SVM)
- Feed-Forward Neural Networks (FFNN)

The first and foremost goal of predictive modeling is to achieve a high predictive performance. As more advanced machine learning approaches have been developed, more accurate models have been created. The interpretability of the models is often of secondary interest, but inherently there is a trade-off between high predictive accuracy and understanding what lead to a certain prediction (Kuhn & Johnson, 2013, p. 4). As mentioned in 2.2 *Literature on Churn Prediction*, both SVM, bagging and boosted tree-based methods, as well as FFNN tend to outperform logistic regression. However, LR is included in this project to serve as a benchmark for the other more advanced models and because of its simplicity and interpretability.

#### 2.3.4.1 Logistic Regression

Logistic regression, often called *logit*, is a fairly simple but highly interpretable classification modeling approach that was first presented in the 1940s. Later, in the 1970s, Nelder and Wedderburn presented generalized linear models (GLM) as a generalization of previous linear regression models, of which logistic regression is a subset (James et al., 2013, p. 6). The logit model, in its most simple form, uses the logistic function to model a binary variable such as customer churn, encoded by 0 or 1, as a function of a number of predictors. This is achieved by modeling the probability that a sample belongs to either class, whereafter one can adjust the probability threshold between the two classes. The antecedent multiple linear regression model is insufficient in achieving probabilities strictly between 0 and 1. In order to overcome this issue, the S-shaped logistic function is used, defined as:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

The model is fit by estimating the parameters using maximum likelihood, and the entity  $\frac{p(X)}{1-p(X)}$  is called the *odds*. If we take the logarithm of this expression, we get what is called the *log-odds*, or *logit*, and this can in a multivariable setting be expressed:

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

We can clearly see that the resulting model for the log-odds is linear in  $X$ , and the model is interpretable in terms of changes in the odds or log-odds. There are multiple extensions of the logit model, e.g., allowing for multi-category dependent variables or ordered variables (Kuhn & Johnson, 2013, 2013 p. 130-138).

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + f_1(X_1) + \dots + f_p(X_p)$$

#### 2.3.4.1.1 Regularization

In order to increase performance of the logit approach, regularization is often applied. Regularization “penalizes” models with large weight vectors, which makes the resulting model less complex and less prone to overfitting. Regularization is applied in the objective function that is minimized in order to determine the weight vector  $\beta$ . In general, regularization applies a penalty  $p(\beta)$  to the weights, in addition to the residual sum of squares (RSS) that define logistic regression, leading to the optimization problem:

$$\min J(\beta) = RSS + p(\beta)$$

The most common forms of regularization for logistic regression are a  $L^1$  or  $L^2$  penalty on the weights, or a combination of them. A  $L^1$  penalty corresponds to  $p(\beta) = C \sum_{j=1}^p |\beta_j|$ , where  $C$  is a hyperparameter representing how heavily the weights are penalized and is referred to as a LASSO (Least Absolute Shrinkage and Selection Operator). Because of the “sharp shape” of the  $L^1$  ISO curves, using a LASSO tend to lead to sparse models where many weights are set to zero, which leads to a simple and interpretable model. An  $L^2$  penalty corresponds to

$p(\beta) = C \sum_{j=1}^p \beta_j^2$ , and is referred to as ridge regression, which penalizes complex models but does not tend to be as aggressive as the LASSO. Finally, elastic nets are a proposed method that combines  $L^1$  or  $L^2$  regularization with a factor  $\alpha \in (0,1)$  that determines the weight of the  $L^1$  and  $L^2$  term. In general, none of the approaches have been proven to dominate the other, rather this depends on the problem and data set at hand (James et al., 2013, p. 214-228).

### 2.3.4.2 Tree Based Methods

Classification and regression trees were first introduced in the 1980s by Breiman, Friedman, Olshen and Stone (James et al., 2013, p. 6). These are non-parametric models, meaning they do not make any assumptions about the functional form of the underlying function  $f$  (James et al., 2013 p. 23). Additionally, tree-based models are highly interpretable and flexible in the sense that they can handle various types of data, and there is no need for preprocessing of e.g. skewed or sparse data (Kuhn & Johnson, 2013, p. 174).

Tree-based models have proven highly accurate and proven to many times outperform more complex and flexible models such as neural networks for tabular data. This is especially true for ensemble methods such as gradient boosted trees. Additionally, tree-based models are often more interpretable than linear models, as linear models often suffer from strong assumptions, e.g. regarding distributions and functional forms of the relationships, and are hence sensitive to model-mismatch (Lundberg et al., 2020).

#### 2.3.4.2.1 Decision Trees

The logic behind decision trees is based on nested if-then statements, where some or all the predictors are used to partition the observations into two or more groups, based on some criteria, illustrated in figure 2.9 (Kuhn & Johnson, 2013, p. 369).

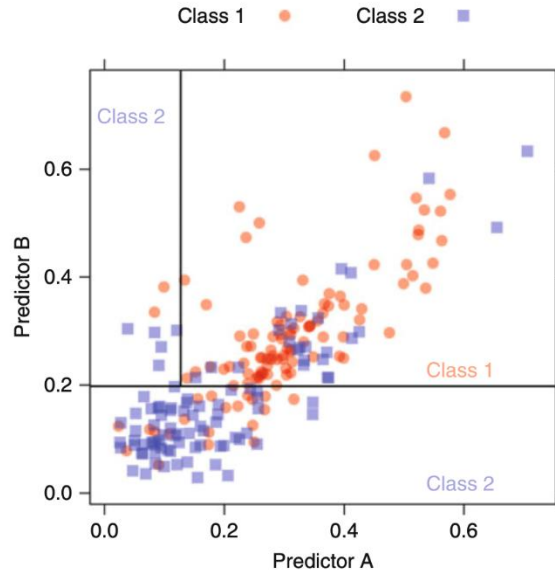


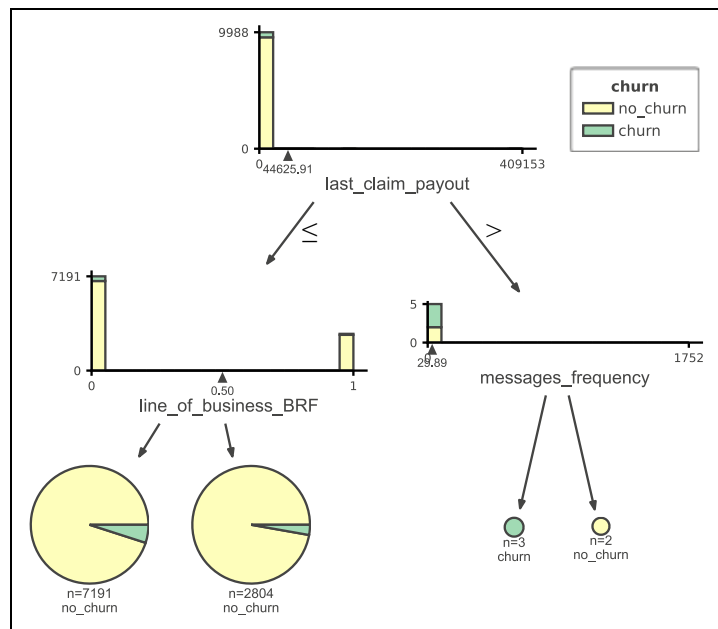
Figure 2.9 Class prediction using a tree-based approach (Kuhn & Johnson, 2013, p. 370).

The groups on either side of the split should be more homogeneous than before, also referred to as the nodes being purer, i.e. containing a larger proportion of one of the classes. The split can be determined based on the Gini index or the cross entropy,

used as a measure of the purity of the nodes. In this two-class classification problem of “churn” or “no churn”, the Gini index can be expressed as:

$$p_1(1 - p_1) + p_2(1 - p_2) \quad \text{or} \quad 2p_1p_2$$

where  $p_1$  and  $p_2$  are the two class probabilities. Partitioning algorithms then evaluate various splitting options and partitions the data where the purity criterion is minimized. Thereafter, this process is continued for each node, repeatedly until a predetermined stopping criterion is fulfilled. An example is illustrated in figure 2.10. This criterion could for example be a maximum tree depth or a minimum number of observations on either side of a partitioning. In order to avoid overfitting, a penalty can be introduced for an increasing number of final nodes (Kuhn & Johnson, 2013, p. 370-372).



**Figure 2.10** Example of a decision tree showing the splits of the tree, as well as churn rates for each group in the bottom row.

#### 2.3.4.2.2 Bagging and Random Forest

Bagging, short for bootstrap aggregation, is an ensemble technique that was first proposed in the 1990s. The approach is based on generating  $m$  bootstrap samples of the original data, and then training a model on this sample. This is then repeated for the  $m$  samples, and each model votes on the predicted class, resulting in the bagged model’s predicted probability vector after averaging all  $m$  predictions. Bagging reduces variance and increases model prediction stability (Kuhn & Johnson, 2013, p. 192-194).

Bagging decision trees is commonly referred to as random forest and tends to include around 1000 trees trained on bootstrap samples of the original training dataset. The term random forest comes from the fact that there are multiple trees and that there is a random component included in the process. The random component is included in order to de-correlate the trees, for the trees not to have the same structure or be too correlated to one another, improving the predictive accuracy. In order to tune the algorithm, one needs to adjust the number of trees and the number of randomly selected predictors to choose from at each partitioning (Kuhn & Johnson, 2013, p. 198-200).

#### 2.3.4.2.3 Boosting and Gradient Boosted Trees

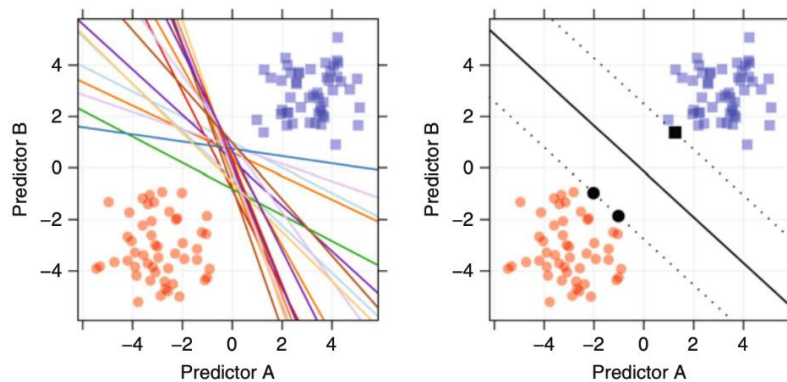
Boosting was first proposed for classification problems in the mid 1980s. The idea behind boosting is to combine many weak classifiers into an ensemble classifier with a superior performance. Many different boosting algorithms have been proposed, among which AdaBoost, short for adaptive boosting, proposed in 1999, and the succeeding gradient boosting machines are most notable. In the AdaBoost algorithm, each observation is given an initial weight. This weight is then updated at each iteration, each new tree being trained, and incorrectly classified observations receive a larger weight in the next iteration and correctly classified observations a smaller weight. Hence, the weight for observations that are repeatedly incorrectly classified will be given an increasingly large weight, until the observations hopefully are classified correctly. Hence, the idea is that the sequence of weak classifiers each build upon the shortcomings of the previous models, and together are able to make a more accurate prediction.

Shortly after AdaBoost was proposed, the idea behind boosting was carried further to include more advanced statistical concepts, finally resulting in gradient boosting machines. The intuition is to find an additive model at each iteration that minimizes the loss function and incorporate this model into the previous sequence. Then, the gradient is calculated, and the iterations continue until a predetermined stopping criterion. In order to reduce the risk of overfitting, a stochastic component is introduced. At each iteration, a new additive model is trained on a random subset, a so-called bagging fraction, of the observations. An issue is that this strategy ensures an optimal model at each iteration but does not guarantee a global optimal solution. To account for this, a learning rate or shrinkage is introduced. This is a tuning parameter that slows down the learning process by taking smaller steps at each iteration and hence shrinks the contribution of each tree by a given rate. The parameter is set to be between 0 and 1. Small values, a slower learning process, often results in a more accurate prediction, however, will also increase the training time of the model and increase the required memory. It is also possible to boost other models than tree-based, however, trees have proven highly suitable for this type of algorithm. (Kuhn & Johnson, 2013, p. 203-208; 389-392). As discussed in section 2.2 *Literature on Churn Prediction*, the boosting algorithm XGBoost

(eXtreme Gradient Boosting) is a popular gradient boosting algorithm that has proven very useful and has recently gained popularity (Gregory, 2018).

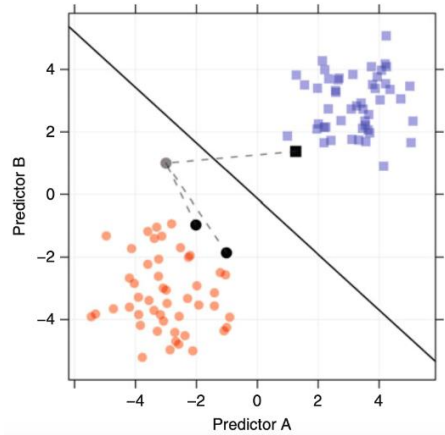
### 2.3.4.3 Support Vector Machines

Support vector machines are a class of statistical models introduced by the Russian statistician Vladimir Vapnik in the mid-1960s and are to date considered one of the most flexible and effective models for classification problems. The main idea behind the approach in its most basic form is illustrated in figure 2.11, where two variables are used to predict two classes. In this example, the classes are completely separable by a straight line. This line can be drawn in an infinite number of ways by adjusting the intercept and slope. In order to decide which line to use, one makes use of the dotted lines in the figure, which indicates where the closest data point for each class is located in the data set. The distance between the dotted line and the solid classification boundary is called the margin, and the dashed lines are located at the maximum distance from the classification boundary, with the boundary being located with equal margin on both sides. The classification boundary that maximizes these margins is referred to as the maximum margin classifier (Kuhn & Johnson, 2013, p. 343-345).



**Figure 2.11** Left: Data with well-separable classes and possible lines to separate them. Right: The linear maximum margin classifier as a solid line and the support vectors as solid black points (Kuhn & Johnson, 2013, p. 344).

Once the class boundary has been determined, new observations will be classified based on what side of the boundary line they are located, as illustrated in figure 2.12.



**Figure 2.12 Classification of a new observation using a support vector machine (Kuhn & Johnson, 2013, p. 346).**

The approach was extended in the 1990s to account for nonlinearities by introducing a kernel function, making the SVM models extremely flexible. For example, the boundary can be set by using a nonlinear kernel such as:

$$\text{polynomial} = (\text{scale}(\mathbf{x}'\mathbf{u}) + 1)^{\text{degree}}$$

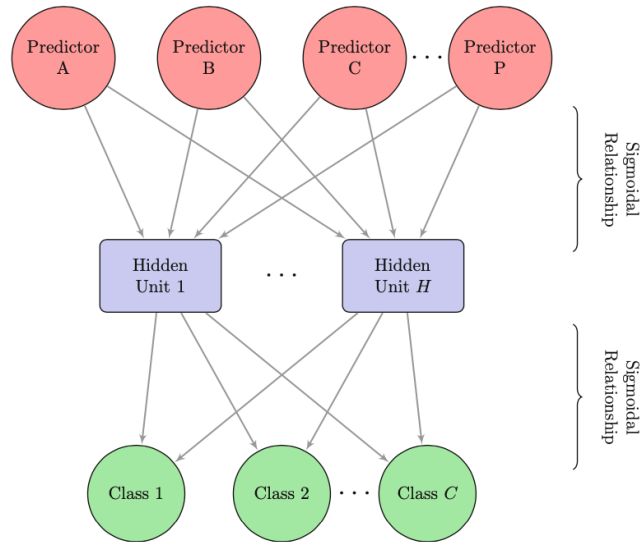
$$\text{radial basis function} = \exp(-\sigma|\mathbf{x} - \mathbf{u}|^2)$$

$$\text{hyperbolic tangent} = \tanh(\text{scale}(\mathbf{x}'\mathbf{u}) + 1)$$

Further, we need to consider the case when the classes are not completely separable. In this case, a cost is introduced for misclassified observations or when an observation is located on the boundary itself. This cost parameter adjusts the complexity of the class boundary, and a too large cost value will most likely result in overfitting. A final consideration is that the predictors need to be standardized before the SVM model is fitted to the data, so that large-valued attributes do not have excess influence on the model (Kuhn & Johnson, 2013, p. 346-351).

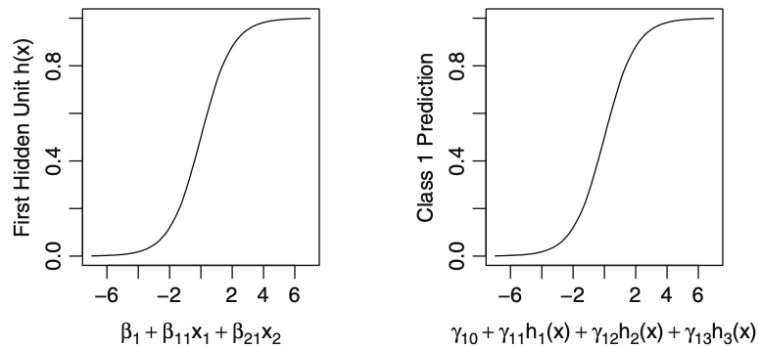
#### 2.3.4.4 Feed-Forward Neural Networks

Neural networks are a group of nonlinear models with nodes and links, mimicking the structure of the brain's neurons. Figure 2.13 shows a feed-forward neural network, which in its most basic form consists of a set of predictors, the class outcome predictions, and an intermediate layer, a set of hidden variables or units that are unobserved.



**Figure 2.13 Single-layer feed-forward neural network (Kuhn & Johnson, 2013, p. 334).**

The hidden units are estimated by a linear combination of the original predictors, usually transformed by the logistic function, also called the sigmoidal function, as illustrated in figure 2.14.



**Figure 2.14 Modeling approach with the Sigmoidal function for the hidden units and class predictions in a single-layer FFNN (Kuhn & Johnson, 2013, p. 334).**

The output of each node is computed as:

$$h_x(x) = g\left(\beta_{0k} + \sum_{i=1}^P x_j \beta_{jk}\right), \text{ where } g(u) = \frac{1}{1 + e^{-u}}.$$

Here,  $h_k(x)$  is the  $k$ :th hidden node, the  $\beta$  coefficients are the original predictors, and  $g(u)$  is the nonlinear transformation, usually the logistic functions. The same equation is also used in the final layer to compute the class prediction output of the model (Kuhn & Johnson, 2013, p. 141-144, 333-334).



The Sigmoidal transformation ensures that the class predictions lie between zero and one. However, these predictions are not to be interpreted as probabilities, as the sum of the class predictions is not one. Therefore, the predictions are normalized using the Softmax transformation, yielding a probability distribution over the class prediction output that is proportional to the exponentials of the initial class predictions. This is achieved by computing the following function:

$$f_{il}^*(x) = \frac{e^{f_{il}(x)}}{\sum_l e^{f_{il}(x)}}$$

Here,  $f_{il}(x)$  is the initial prediction of the  $l$ :th class and the  $i$ :th sample, and  $f_{il}^*(x)$  is the transformed, probability-like output of the  $l$ :th class. Finally, the sample is classified as the class with the highest probability (Kuhn & Johnson, 2013, p. 248, 333-334).

The modeling structure described above is called a single-layer, feed-forward neural network and can then be extended to involve multiple hidden layers, loops that go in both directions between the hidden layers and including manual alterations such as removing specific node connections as part of the final optimization process (Kuhn & Johnson, 2013, p. 141-144, 333-334). When introducing too many layers, and hence a large number of estimated parameters, the network tends to overfit the data, leading to poor generalizability to new observations. There are several approaches to overcome this. Firstly, the optimization algorithm can be stopped before a minimum is reached, and instead be stopped when the estimated test error starts to increase. The logic behind this is that, when minimizing the training error, the test error will start to increase when the model starts to overfit to the training data set. This procedure is often called early stopping. Another approach is to introduce a penalization for large coefficients, referred to as a weight decay. By doing this, a large coefficient is only allowed if the parameter has a significant impact on the model. The weight decay value can be adjusted, and a large penalization term leads to a smoother model and a lower risk of overfitting. Additionally, several networks can be fitted using different starting values in the optimization algorithm, and the average of the class probabilities can be used for the classification. These models could look very different but perform equally well, as they can find different local minima. This approach often yields improved results compared to the regular approach (Kuhn & Johnson, 2013, p. 143-144, 335-336). Yet another approach is bagging, by instead taking the average of the predicted class probability from different networks that have been trained on random bootstrap samples of the original training data (Hastie et al., 2009, p. 401).

As described, there is a trade-off in the selection of the number of hidden features between model flexibility and risk of overfitting. However, it is generally better to have too many hidden units than not having enough, as this allows the model to be flexible enough to include the necessary nonlinearities, while excess hidden units can be ruled out with proper regularization through weight decay. The appropriate

number of hidden units varies substantially depending on the data set and task at-hand, but generally tends to be in the range of 1-100 (Hastie et al., 2009, p. 400).

### 2.3.5 Model Interpretation

In order to understand what the model has learned, and thereby what drivers of churn it has identified, interpretation methods can be used. This process is useful also for inspecting the model to see if it seems to have learned a true causal pattern rather than overfitting to the data. Fitting a model and inspecting it, is also an effective way of understanding the relationship between the data and churn. Rather than trying to find drivers by data exploration methods, the machine learning models work by trying to find patterns that predict churn well and those can then be inspected (Molnar, 2020).

Machine learning interpretability, or explainable artificial intelligence as it is sometimes called, is a relatively new field which is quickly developing (Molnar, Casalicchio, & Bischl, 2020). This development is especially driven by e.g., medical applications where explanations are vital. In general, this is a trade-off between the flexibility and interpretability of a model. As such, machine learning models that are more advanced and potentially more accurate, tend to be less interpretable. Intuitively, this is because a more flexible model requires more parameters, and more complex interactions, which are harder to visualize and explain. A representation of this trade-off with a selection of statistical models is shown in figure 2.15. As historically the more intrinsically interpretable models such as logistic regression have been used, the focus in the machine learning community has not been around developing interpretation methods. More recently however, as more complex models such as SVM, ensemble tree methods and neural networks have gained popularity, new tools are being developed (Adadi & Berrada, 2018).

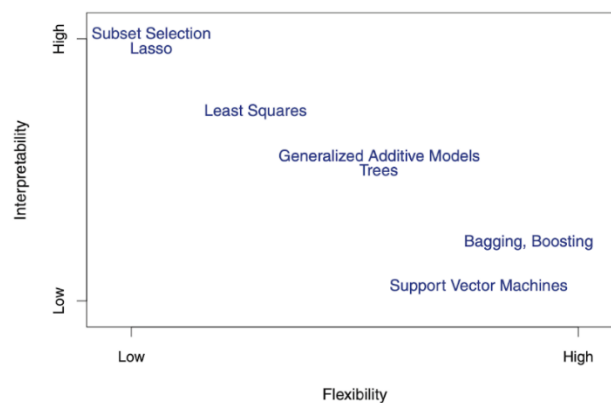


Figure 2.15 Representation of the trade-off between flexibility and interpretability (James et al., 2013 p. 25)

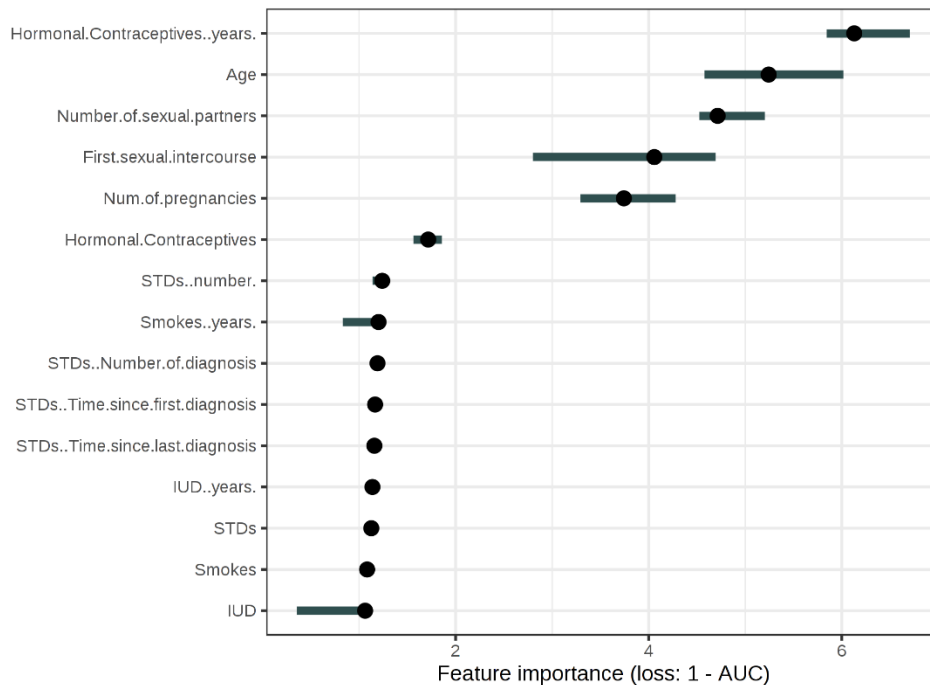
Generally, interpretation methods can be categorized into 1) intrinsically or post hoc 2) model specific or model agnostic and 3) local or global. Intrinsic methods are models that are considered intrinsically interpretable because of their simple structure, such as logistic regression or shallow decision trees, whereas post hoc are methods applied to analyze the model after training. Model specific methods can only be applied to a particular model whereas model agnostic ones can be applied to any model. Finally, local explanations are tools that explain predictions of certain observations or a neighborhood of observations, whereas global explanations apply to the entire space (Molnar, 2020).

As interpretable machine learning is still a developing field, where new methods are proposed, no single tool has been identified as the best for the task. Rather, different tools have different strengths and weaknesses, and can be used together. Following is a selection of interpretation tools that have been proposed for tabular data, which will be described further and used in this project (Adadi & Berrada, 2018; Burkart & Huber, 2021):

- Permutation Feature Importance (PFI)
- Individual Conditional Expectation (ICE) and Partial Dependence Plot (PDP)
- Shapley Additive Explanation (SHAP)

#### 2.3.5.1 *PFI*

Feature importance is a measure that attempts to estimate the effect on the accuracy of a model when a predictor is used or omitted. A feature importance plot assigns a value to each of the predictors and displays them in a plot, see figure 2.16. There are many ways these values can be estimated depending on the definition, and assumptions of the model. For certain models there exist model specific values for feature importance, such as the parameter weights in logistic regression, or mean Gini gain for random forests. Permutation feature importance (PFI) is a global model agnostic importance measure, which calculates the performance of the model rather than the effect on the output prediction  $\hat{Y} = \hat{f}(X)$ . To calculate PFI, the performance of the model based on some metric, e.g. AUC-ROC is calculated for the full data set. Next, for each feature, its values are randomly permuted across the observations, and the performance is calculated on this data set and the loss in performance is recorded. This process simulates removing, or rather placing nonsense values in place of the actual values for the feature and seeing how much worse the model performs (Molnar, 2020).



**Figure 2.16 Example of a Feature Importance plot with error bars (Molnar, 2020).**

One disadvantage with PFI is that the permutation procedure samples values from all other observations. This can be problematic with correlated variables, as observations that are unrealistic are used in the calculation. For example, if the predictor is the square meters of a home, and the data set includes the premium of the insurance contract as a feature, combinations such as very small homes with very high premiums may be included in the calculation of PFI. This causes the calculation to be biased towards unrealistic observations. This effect can be minimized by decreasing the number of correlated features in the data set (Molnar, 2020). With correlated features, the importance is also split among the correlated predictors, as the information is kept even when one feature is removed or permuted, leading to the importance appearing lower for these features than if only one of them was included (Hooker & Mentch, 2019). A caveat with interpreting PFI is that when there are interaction effects between multiple predictor variables, the combined effect of both of them is measured rather than the marginal (Molnar, 2020). Finally, a disadvantage with PFI is that only the magnitude of the effect of a certain feature is shown, not the direction of the effect, i.e. if it increases or reduces the probability of churn, or whether this relationship is monotonic or not. One way to visualize the direction of change is through ICE and PDP.

### 2.3.5.2 ICE and PDP

ICE and PDP are global model agnostic interpretation tools, which plot the effect of the output variable as a function of one or more predictors. ICE plots the estimated conditional expectation of  $Y$  as a function of a set of chosen predictor variables for each individual observation. The algorithm works by first choosing a set of predictor variables to be examined,  $X_S$ , and defining  $X_C$  as the rest of the variables, and  $\hat{f}$  the response function. The ICE plots  $\hat{f}$  as a function of  $X_S$  for each individual observation and displays them all together in one plot. For each observation,  $X_{C,i}$  is fixed, and  $X_{S,i}$  is varied across its values over all observations, producing a curve estimating the effect of  $\hat{f}$  as a function of only  $X_{S,i}$ . An example of an ICE plot can be seen as the gray curves figure 2.17 (Goldstein, Kapelner, Bleich, & Pitkin, 2015).

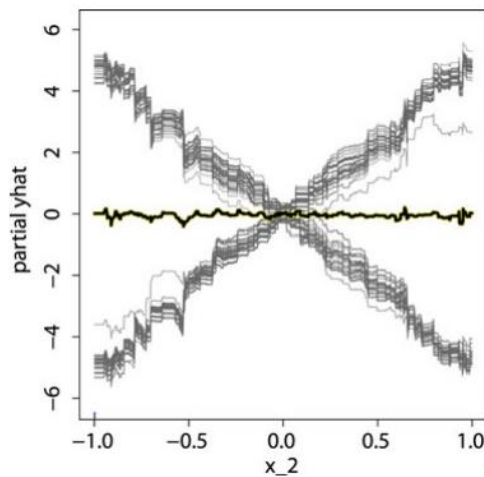


Figure 2.17 Example of ICE and PDP (Goldstein et al., 2015 p. 48).

The PDP can then be calculated as the average ICE across all observations, leading to a measure of the overall effect of a predictor on the model response. The PDP is highlighted in figure 2.17, as the filled black curve. A problem with PDP is that for some predictors, the aggregation of ICE values may hide the actual effect of the feature. In the example shown in figure 2.17, the predictor seems to have a negative effect on the outcome variable for a subset of the observations, and a positive effect on the rest. Averaged, the net effect is 0, and hence the PDP shows a constant zero value which could lead to the wrong conclusion of the predictor having no effect. This said, PDP can still be useful to see the overall effect of the predictor and is often displayed together in the ICE plot (Goldstein et al., 2015).

As ICE and PDP are calculated by replacing features with values of other observations, it suffers from the same drawbacks regarding correlated features as PDP does (Apley & Zhu, 2020).

### 2.3.5.3 SHAP

While the previously mentioned interpretation models are global models, SHAP is a local surrogate model, where the model prediction function  $\hat{f}$  is described in the neighborhood of a certain observation. SHAP is based on the Shapley value from game theory, which comes with clearly defined axioms and interpretations, and hence satisfies many of the desired properties of interpretable models (Molnar, 2020). SHAP estimates these values by modeling the prediction as the sum of a bias and single feature contributions and marginalizes over every feature to extract how the model behaves in its absence (Burkart & Huber, 2021). The originally proposed KernelSHAP is model agnostic but computationally expensive. For tree-based methods, TreeSHAP approximates KernelSHAP and is faster and handles correlated features better, though it can sometimes produce unintuitive feature attributions (Molnar, 2020).

Even though SHAP is a local model that explains a single observation at a time, the values can be aggregated to provide a global explanation as well. With SHAP values calculated for all or a large subset of the observations, estimations can be made on global feature importance, partial dependence, feature dependence, and summary plots (Lundberg et al., 2020). An example of a global feature importance, and local explanation summary plot derived from SHAP values for a US mortality dataset are shown in figure 2.18.

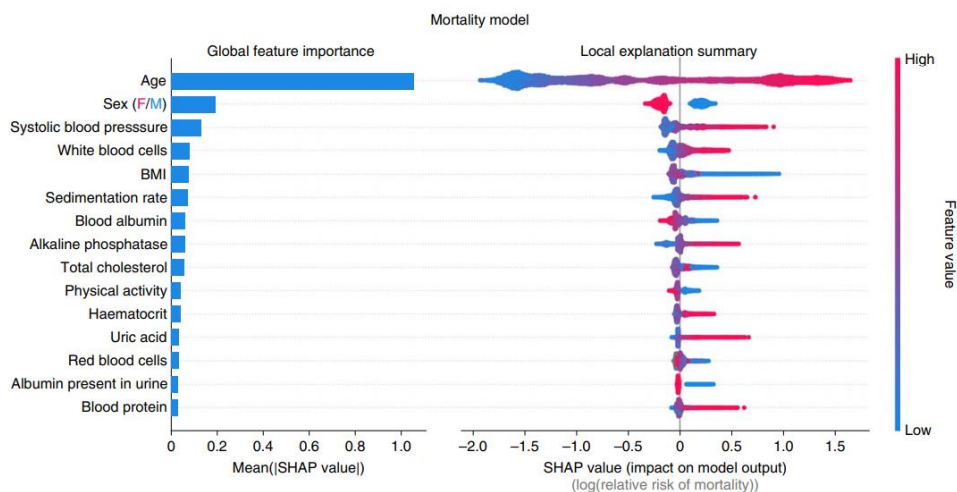


Figure 2.18 Global feature importance and summarized local explanation of SHAP values (Lundberg et al., 2020, p. 61).

# 3 Methodology

*This chapter describes the methodology of the thesis, first covering how to identify at-risk churners, and secondly actions to increase retention at Insurtechs.*

In figure 3.1, an overview of the methodology is presented.

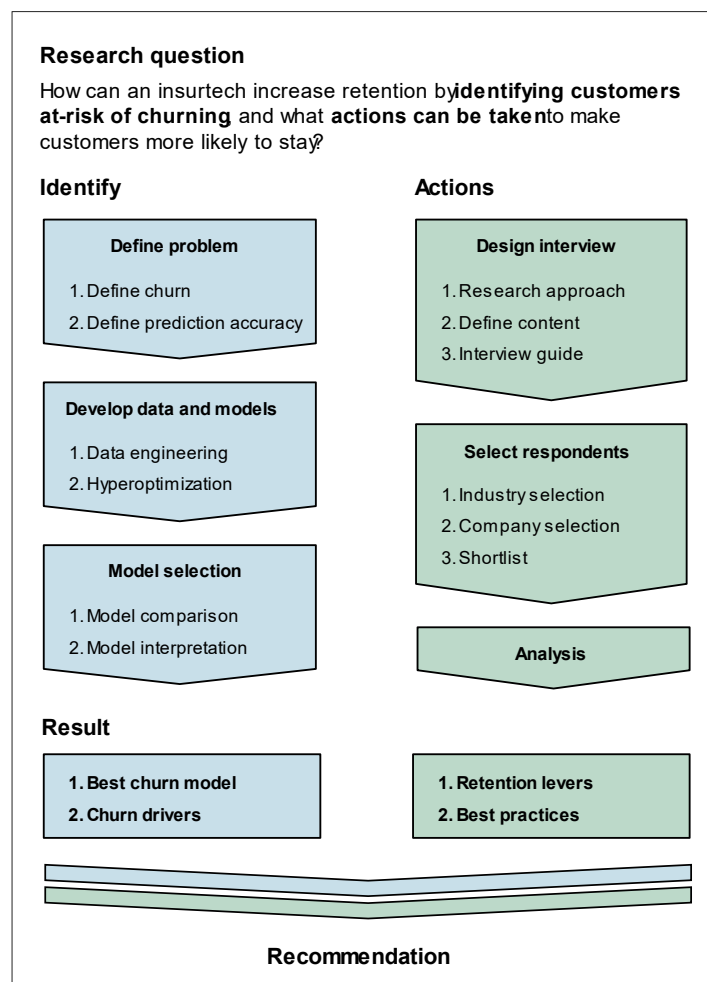


Figure 3.1 Overview of the methodology of this thesis.

## 3.1 Identify

In order to identify at-risk churners using a machine learning model, the following steps must be undertaken, which are covered in the following sections:

1. First, the **goal of prediction** must be decided, and for this purpose churn must be defined.
2. In order to evaluate the performance of a model or set of predictions and compare them, a **performance metric must be decided**.
3. **Data must be extracted and engineered** in order to train models to classify customers according to the definition.
4. Using the performance metric and data, a set of models are **optimized and compared**.
5. Finally, the **best model is selected** from the attempted models.

### 3.1.1 Prediction Goal

In the prediction task, we are approximating  $\hat{f}(X) = Y$ , and in order to do this, we must first define the outcome variable  $Y$  and a relevant set of observations  $(X, Y)$ .

### 3.1.2 Considerations

#### 3.1.2.1 Outcome Variable

The choice of outcome variable  $Y$  defines what problem the model is trying to solve. For example, in the churn context,  $Y$  could be whether a given customer will churn or not in the 3 months succeeding the prediction date, or it could be in the next 12 months for example. These two choices measure different churn behaviors. The former definition tries to predict accurately the timing of a customer churning, so that short term, or more last-minute retention measures can be employed. The latter definition is a measure of customer loyalty and is more useful to identify customers that should be targeted with more long-term loyalty efforts. The  $Y$  variable needs to be defined as churning within some span of time, as all customers will churn eventually (at the latest when the insurance company ceases to exist).

In some applications such as retail, churn can be hard to define, as there are many possible ways to measure if a customer is active or not. If a customer is considered inactive if no purchase is made within 3 months for example, customers who only purchase once a year but do so consistently and who consider themselves active may be considered churners. For Insurtechs with a subscription business model, this



definition is more straightforward, as customers with an active, non-cancelled subscription may be considered active.

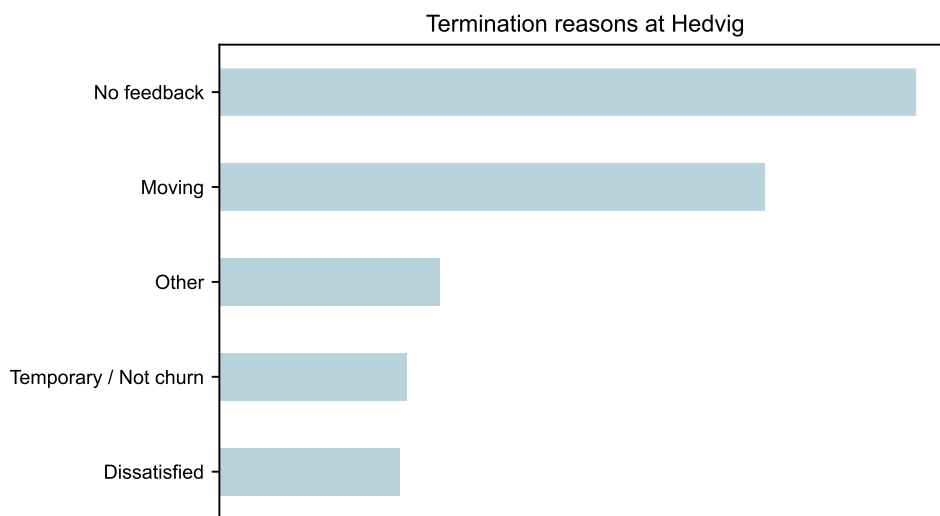
### 3.1.2.2 *Sampling*

In addition to defining the  $Y$  variable, one must select which customers to include. One consideration to be made is whether to include all historical data of previous terminations, or only the more recent ones. As Insurtechs are often growing rapidly and the state of the company developing, terminations years ago may not be as relevant for the current company as more recent ones. In addition, including historical terminations may alter the structure and balance of the dataset. For example, including all terminations, the churn rate in total may be 20%, but looking at only the customers who churned in the last month, the churn rate may be 1%. If the model is to be evaluated on the current customer base, e.g. in the context “which of our customers are going to churn in the coming  $X$  months”, training the model with data representing “which customers have historically churned” may not produce the best results. Rather, an argument could be made that the model should be trained with data of a similar structure. At the same time, only training the model with terminations made in a particular month may cause it to overfit to the characteristics of that month in particular and miss seasonal trends. For example, maybe many students churn in August because they move into new housing as the semester starts, but this model may not generalize to other months.

Depending on the goal of the prediction, considerations must also be made at what point in time customers should be considered. As the model is predictive, modeling if the customer is to churn within a period of time, the current state of the customer base cannot be used directly as we do not yet know which of the customers will churn in the future. As a result, customers must be sampled at a particular point in time. If churn within 3 months is to be predicted for example, the customers could be sampled as a snapshot of the customer base 3 months ago, excluding customers who signed up after that point in time, and excluding all customers who churned before that point in time. Depending on the goal of the prediction, other samples could be made. For example, if the purpose of the model is to predict which customers will churn within 3 months of signing up, all customers could be sampled at the point of signing up. This approach could be used e.g. in order to determine which customers should be prioritized for acquisition. A disadvantage of the approach however is that no information about how the customer uses the service is taken advantage of, so few predictive attributes are available. An alternative approach is to sample customers who have been on the service for example 6 months and predict which customers will stay 6 more months. In this case there is more data, and the result could be used to predict which customers will stay more long-term. Finally, one consideration that must be made as a result of having to sample customers at an earlier point in time is that, as Insurtechs have growing user base, if churn is to be predicted long in the future, e.g. 12 months, there exists significantly less data since a large share of the user base signed during the last 12 months.

Finally, there may be reasons to exclude some termination reasons. For example, students who have been a large segment for Hedvig may cancel their insurance during the summer months as they move back in with their parents and are covered by their insurance, but plan to activate it once the semester starts again. If these customers are included, the churn class may include customers who still consider themselves customers. Additionally, depending on the application, one may only want to include churn events that can be prevented, as predicting non-preventable events in that case may not serve a clear purpose in terms of increasing retention.

In figure 3.2, the churn reasons collected by the Hedvig customer service team when customers choose to terminate their contract is shown. As seen in the plot, the largest reason is moving, either that the customer moves or that another person moves in with the customer and the customer opts to use the other person's insurance instead of Hedvig. A significant number of termination reasons fall into the temporary or not churn category, which e.g. encompasses customers who cancelled because of a temporary move or were terminated e.g. because they were suspected of fraud. Note that, as seen in the plot, no feedback was received for most of the customers, which may bias the other categories as not responding may correlate with some reasons.



**Figure 3.2 Distribution of termination reasons at Hedvig. Note that the grouping of reasons was chosen by the authors of this thesis, and that the x-axis is removed for confidentiality.**

### 3.1.2.3 Approach

As a company tries to increase retention by identifying at-risk churners, many variants of the considerations above could be beneficial and evaluated as they are better for certain types of churn or interventions. In this paper, because of

delimitations, we have opted to define churn and hence  $Y$  as whether a customer will churn within the next 3 months, long enough to be actionable and preventable, and short enough to be relevant for targeting the customers with retention campaigns. For this purpose, the training set is constructed by backtracking the customer base to its state three months from the prediction date, and  $Y$  set to whether the customer has terminated in the subsequent period. For the termination reasons, we removed all samples which were labeled as temporary or no churn, in order to focus on cases where measures can be put in place.

### 3.1.3 Performance Evaluation

In order to evaluate the performance of a prediction model one needs 1) a metric that defines performance, 2) a methodology to measure that metric, and 3) a benchmark to compare the result to.

#### 3.1.3.1 *Performance Metrics*

Since the result of a model prediction is a list of predictions, one for each customer, an aggregate level of performance must be computed in order to effectively compare models. This metric can be defined in many ways depending on the purpose of the project. For example, the costs of miss-predicting a customer as churning or non-churn may be different, and well-selected metric can capture such differences. In addition, the gains of finding churning customers may be different for different types of customers, for example high or low value customers. For this purpose, a custom metric could be defined, for example the “gained aggregate customer lifetime value” of identifying and targeting the churners. In order to limit the scope of this project, we have opted to use the standard AU-ROC metric, complemented by the PR curve, as explained in section 2.3.2 *Performance Evaluation*.

The PR curve is a good complement to AU-ROC, as it measures the percentage of communicated customers who will actually churn when implementing a campaign. The recall then measures the share of customers who will churn that are on the list, and precision measures what percentage of customers on the list who will actually churn. For direct and targeted one-to-one campaigns, such as calling customers, capacity for a small company is often low and hence contacting around 100 of the top 100 most probable churners is feasible. In this context, the precision for low levels of recall is the most relevant, and a model that is especially accurate on this segment may be preferred over one that is slightly more accurate overall. As such, as AU-ROC measures the performance when contacting all customers for a campaign, the PR curve of small levels of recall, e.g. the top 100 as this is a realistic number of customers for Hedvig to call at this stage per month, is considered.

### 3.1.3.2 *Cross-validation*

As explained in section 2.3.1.1 *Training*, the data must be split into a training and test set in order to an unbiased estimate of the performance of the models, as they should be evaluated on unseen data. Additionally, as there are hyperparameters in the models, these also need to be evaluated on unseen data in order to not bias the results, and hence validation must be used.

Since the customer data is in a time series format, and can be extracted at different points in time, it is important to consider at what point in time the training, testing, and cross-validation datasets are sampled from. In these applications, validation should be made out-of-sample, but in order to get an accurate estimate of the long run performance of the model, and not overfit on a certain point in time, out-of-time samples could be used. Gregory (2018) for example formed their training set from January, cross-validation from February, and testing from March a particular year. In this project, in order to limit its scope, the performance is evaluated out-of-sample but not out-of-time. That is to say, the model is trained on customers at the same point in time, but the customers who are to be predicted are not included in the training set. For dividing the data set into training and testing, we chose 20% of the data set to be in testing and the remaining 80% of the data in training. This ensures that the models receive enough data during training to be able to fit the data, especially important in a startup setting where the number of observations are low, while getting a good estimate of the actual performance in the test set. In addition, hyper optimization is done using 5-fold cross-validation, again balancing enough data for training, and getting a good estimate of the performance.

### 3.1.4 **Dataset**

After defining the prediction goal, the next step is to extract and process data in order to train and evaluate the models. The data retrieved is in a tabular format, consisting of one customer per row, and information about each customer as columns. The exact number of customers in the dataset is not discussed for confidentiality reasons, however it can be said that the dataset consists of over 10,000 customers, and the type of data retrieved is shown in table 3.1.

#### 3.1.4.1 *Data Retrieval*

The data on Hedvig's customers is stored in an in-house database consisting of tables representing different characteristics. From this database, information can be extracted from queries, written in SQL (Structured Query Language). Using SQL, the desired customers can be sampled, and their information extracted.

#### 3.1.4.2 *Outcome Variable Y*

In the dataset, there is sometimes a discrepancy between the date a customer cancels their insurance and when the contract expires. This can happen for example when a customer asks to terminate in 6 months since they will be included in another insurance at that point. In this case, the date the customer cancels should be used as this is the event that can be prevented. Hedvig has not recorded the date that the termination was requested until recently, only the date of contract termination. As such, the requested date is not available for all customers. This could be solved in a multitude of ways, including substituting the missing values with for example the mean offset between the request and contract cancellation. In this project we opted to go for the simplest solution: using the requested date when available, else the contract cancellation date.

#### 3.1.4.3 *Prediction Variables X*

In the case of Hedvig, the problem is not an overwhelming amount of information, where only the relevant features must be extracted, rather the information about each customer is relatively sparse. This is because of 1) Hedvig being a relatively new company, and 2) Hedvig offering home contents insurance, a service that is fairly low engagement with little usage data or interaction points. As such, the approach of this project is to use all available predictors.

The available data consists of the following categories:

- **Demographic** data, for example age.
- **Customer** data, for example for how long they have been customers, from what channel they signed up and if they received any promotional offers such as a discounted premium.
- **Insurance contract and home** data, for example insurance premium, size of home and how many people are covered by the contract.
- **Claims** data, for example when they claimed and what the payout was. For some claims, the customer has submitted a rating on the scale 1-6.
- **Message** data: communication with customer service is done via chat, where the messages sent by the customer are available in the data set. As analysis of the contents of messages in an automated way is a project in itself, only the timing and number of messages are considered.
- **Referral** data: Hedvig has a referral program where a customer receives a discount for each person they sign up to the service with their own unique code. In order to limit the scope of the project, information about the customers that were referred is not included, rather the timing and number of referrals is considered.

- **App** data: Hedvig has an app where the user can communicate with customer service or submit a claim or check details about their insurance. As a delimitation this data is not included, as it requires additional extraction and processing.

The full list of variables recorded for each customer and a description of the data is shown in table 3.1.

**Table 3.1 Customer data used for prediction.**

<i>Feature name</i>	<i>Description</i>
<i>Demographic data</i>	E.g., age as an integer
<i>Acquisition date</i>	Date when the customer signed up at Hedvig
<i>Acquisition source</i>	The channel through which the customer entered, e.g. marketing & organic, referral, through a partner, or comparison site
<i>Discount</i>	Information about potential discounts that the customer has, when these start and end, and how large they are
<i>Line of business</i>	One of Renter, Student-renter, BRF, Student-BRF. BRF means owning the apartment, referring to the housing cooperative
<i>Insurance details</i>	Including the insurance premium, the size, and number of co-insured
<i>Claims</i>	List of claims where the date and payout of each claim is provided. Some claims also include the customer's rating of the claim
<i>Referrals</i>	List of dates where the customer has referred someone
<i>Messages</i>	List of dates when the customers has written in the Hedvig app chat
<i>Insurance history</i>	Whether the customer has had any previous insurance with Hedvig, and which edits they have done to their insurance

From this raw data, features are extracted and processed such that they can best be used in the models. In addition, different sets of these features are chosen to be fed into the models as a feature selection step. As explained in 2.5 *Class Balancing*, the accuracy of models may be improved if the data is resampled, which will be applied to the models in training.

### 3.1.5 Model Selection

In order to select the best model for the purpose of predicting churn, each machine learning model is presented in 2.6. *Machine Learning Classification Models for Churn* - including LR, DT, RF, GBT, SVM and FFNN - are evaluated and compared. In order to maximize the performance of each model, its hyperparameters are first optimized. In this context, not only the machine learning model hyperparameters, such as regularization factor, are tested, but also feature selection, preprocessing and balancing is considered a part of the hyperparameter optimization process.

After each model is optimized, its cross-validation AU-ROC and PR curve is evaluated, and the best performing model selected. Finally, the best model is inspected by looking at its performance in terms of AU-ROC and PR on the test set, as well as the relationship between the predicted churn probability and actual churn.

#### 3.1.5.1 Benchmark

To serve as a benchmark to the developed models, one of the simplest machine learning models, logistic regression will be used. This model is trained using the most basic customer data concerning the customer and insurance, excluding e.g., claims, messages, and referrals. In addition, the benchmark model is fed with the raw data, without any preprocessing, feature selection or balancing.

### 3.1.6 Churn Drivers

As the goal of this paper is not only churn prediction of individual customers but also an understanding of the drivers of churn, a methodology for this is presented. As churn drivers at a company can be quite specific for its service and customer base, a general set of drivers is hard to present. Rather, this thesis presents a methodology to extract churn drivers, and applies them to the Hedvig case. For this purpose, one could try to manually find patterns in the data, between those who churn and not and what characteristics they have. However, the machine learning models that have been developed are specifically designed for this purpose: trying to find patterns in the data to extract characteristics that customers who churn have (Molnar, 2020). As such, churn drivers are extracted by inspecting the models using the model interpretation tools presented in section 2.3.5 *Model Interpretation*, including PFI, ICE, PDP and SHAP. These tools are applied to the best performing model, as this model has proven to understand the churn dynamics the best.

## 3.2 Actions

### 3.2.1 Research Approach

There exists limited literature on how to target customers that are at-risk of churning and on best practices for retention work for Insurtechs. In order to understand what levers can be used to increase retention, and what the best practices are within these, a qualitative study with industry expert interviews will be conducted. The interviews will be conducted with industry professionals working with customer retention both in Insurtechs, as well as in adjacent industries, with a focus on companies with traits similar to Insurtechs. We will mainly be interested in interviewing professionals that actively work with customer retention, hence mainly occupied with e.g. CRM, marketing, or product development, and having a rather operational focus. It is also of interest to talk to professionals involved in the respective analytics departments, who have insight into the data availability and structure as well as knowledge about the existence or potential of analytical tools for customer retention work. From these interviews, we will identify a number of levers that can be used to increase retention. Then, the insights and best practices identified in the interview study will be categorized according to these levers and synthesized into actions that are relevant for Insurtechs and Hedvig. Additionally, we will assess trends and what impact working proactively with retention can have on lowering the churn rate in subscription businesses. The interview study of this thesis is conducted in an inductive manner, interviewing industry professionals on how to increase retention and structuring their answers around levers and best practices. These results are then compared to previous literature in the discussion 6.4 *Retention Management*. Another approach to gain insight on the impact of retention action could have been to carry out interviews with customers to get their perspectives. However, as this is a very time-consuming task, and a large number of interviewees are needed to ensure any generalizability of the results, we aim to get the customers' perspective through the data collected on the customers as described in section 3.1 *Identify*, while the interview study will focus on the industry professionals' points of view.

### 3.2.2 Interview Content and Format

Based on the research question regarding what action can be taken in order to increase retention, a more detailed set of topics, questions and sub-questions are developed. The interviews cover questions regarding how the companies work with retention in terms of e.g. how churn is measured, the organizational set-up and processes, and KPIs and responsibilities. Further we discuss what levers and concrete initiatives the organizations have taken in order to increase retention, their results and background. We discuss what the drivers of churn have been and what



tools, including but not limited to machine learning, have been used to identify customers or customer groups at-risk of churning. Additionally, we discuss whether the companies have the necessary structures and processes in place to collect the relevant data for these types of analyzes. Finally, we discuss what impact organizations have seen by working on retention and aim to discover best practices within the identified levers. In order to narrow the scope of the project, we focus our interview and discussion on retention levers related to predictive modeling, however other possibilities or levers that we discover in the process will be included in the result.

For the full interview guide, see Appendix B. The interview content is grouped to form four main overarching questions that will be discussed with the industry professionals:

1. **“How do you work with retention at your company?”**. This question will open up the discussion and get the respondent's first thoughts on retention work within their organization.
2. This discussion point will be followed by a more concrete question: **“What initiatives have you taken to increase retention?”**, in order to get into the details about specific levers and best practices at the company.
3. After this, we shift the discussion to the topic of identification of at-risk customers: **“Have you worked in a structured way to identify potential churners and take targeted initiatives to increase retention?”**. Here, we want to learn more about what analytical tools, including machine learning, are used in the companies of interest. We discuss if these types of analytical tools are deployed, what the impact has been, or what the reasons or barriers have been for not leveraging this type of technology. In addition to discussing the tools, we ask about how the output of the models are used to increase retention.
4. Finally, we conclude with a more general question about potential learnings from previous positions or experiences: **“Do you have experience from retention work from other companies? What does it look like in the industry?”**. The purpose here is to gain further insights on a broader, industry level, such as industry trends or what types of companies are the most successful in their customer retention work.

### 3.2.3 Interview Considerations

The interviews follow a semi-structured format, as we do not want to limit ourselves by strictly following a fixed, predefined set of questions. However, an interview guide with questions and sub-questions is used to guide the discussions, but are aimed to be more open-ended, allowing a more informative, investigative

discussion. The interview questions are therefore allowed to, at times, be altered or branched out, depending on the interview situation and the interviewee's background and knowledge.

In order for the interview responses and insights to be as credible and reliable as possible, we aim to conduct a large number of interviews with professionals in different industries, companies, and functions, with different backgrounds, experiences, and goals. By doing this, we aim to get a comprehensive account on retention work in subscription companies. Additionally, the information we receive will be triangulated and compared between the various perspectives, in order to increase the reliability and strengthen the conclusions of the study, as the bias will be reduced both by this approach, as well as, by the large number of respondents. Finally, the questions or discussion points are predetermined, even if they may be adjusted to a particular interview, and we generally aim to ask open-ended questions that does not bias the interviewee to respond in a particular way.

All interview participants will be informed of the option to anonymize their responses, or parts of them, in order for them to be able to talk more freely, and to protect the interests of the participants.

### **3.2.4 Industry and Company Selection**

In order to identify suitable industries and companies to interview, we deploy both an outside-in and an inside-out perspective. First, we identify a shortlist of industries where the business model is fairly similar to Hedvig's, and industries that share, in our opinion, important attributes and industry dynamics with the insurance industry. Then, we look for companies within these industries that have similar traits as Hedvig, in terms of e.g., life cycle and digitalization progress. Additionally, we combine this selection with an inside-out perspective, where we get Hedvig's take on the selection, and get their view on what industries companies could be fruitful to gain learnings from, based on their experience in the industry and in customer retention.

After discussions with Hedvig, from our literature review, and discussions internally between the authors of this thesis, we have created a framework for the industry and company selection for our interviews. The idea is to figure out what industry traits and dynamics are important when aiming to transfer learnings to Insurtechs, and Hedvig in particular. Additionally, important company attributes were considered in the same manner.

#### **3.2.4.1 Industry Attributes**

First, relevant industries must be selected for the interview study. For this, we have identified five relevant parameters: 1) Subscription business model, 2) Low

product/service differentiation, 3) Low customer engagement, 4) Similar churn/switching occasions, and 5) Comparison sites are used.

#### 1. Subscription business model

As insurance is based on a subscription business model, and the definition of churn is different for a subscription company compared to e.g. a company with frequent purchases, this is the main characteristic that we consider when deciding on relevant industries.

#### 2. Low product/service differentiation

Insurance is characterized by low differentiation and price competition. In industries with a higher degree of differentiation one would expect there to be different drivers of churn. Companies in a more differentiated industry have offerings that attract different customer groups to varying degrees, while insurance is rather based on trust, habit, and customer experience. Therefore, one would expect customer retention in companies within less differentiated industries to be more similar to that of insurance.

#### 3. Low customer engagement

In industries with higher customer engagement, there are first of all more frequent decision points where a customer can reconsider their subscription, compared to insurance. Additionally, less engaged customers are plausibly less attached to a particular company or brand, affecting the churn dynamics of the industry.

#### 4. Similar churn/switching occasions

Different industries have different switching occasions, and in order to learn as much as possible from the interview objects, companies in industries with similar churn occasions, e.g. moving to a new home, are of high interest. Insurance tends to have a long binding period, at least traditionally, which also affects the dynamics of the business.

#### 5. Comparison sites are used

The existence of comparison sites relates both to differentiation, customer engagement and switching occasions. The fact that comparison sites are widely used in an industry is a clear sign that the products or services are very similar apart from a price difference. Therefore, comparison sites could be a good signal of an industry sharing significant traits with insurance.

#### 3.2.4.2 *Company Attributes*

In addition to focusing on relevant industries, relevant companies must be selected within each industry. For this, we have identified three relevant parameters 1) Stage in life cycle, 2) Level of digitalization, and 3) Payment frequency.

### 1. Stage in life cycle

The focus on customer retention and the way companies operate vary to a high degree during a company's life cycle, and in order to examine what actions Hedvig could take in order to increase retention, we are mainly interested in interviewing companies that are in a similar stage in the company's life cycle as Hedvig. On the other hand, discussions with mature companies can be interesting to understand best practices, and which direction to focus on long term.

### 2. Level of digitalization

Hedvig is a disruptor in the insurance industry, not the least due to the digitally focused interactions with its customers. Interviewing companies that share this attribute would give us a more focused perspective of what actions are feasible for a digital Insurtech. This goes, to a large extent, hand-in-hand with the types of customer segments that the company is targeting.

### 3. Payment frequency

Lastly, the payment frequency is important. Hedvig practices monthly payments in an industry where the norm has been annual payments. In order for the learnings to be transferable to an as large extent as possible, the companies should preferably charge in monthly intervals.

#### 3.2.4.3 *Shortlist*

Based on this framework, we created a shortlist of industries, and companies within these industries, that could be of interest for our interviews. Insurance is clearly of interest, but there are other industries that share some of these attributes. In decreasing order of interest, we have identified the following industries: 1) Insurance, 2) Consumer electricity, 3) Private loans, 4) Telecommunication, and 5) Streaming services. We have chosen not to disclose the shortlisted companies in order to protect the integrity of the participating respondents.

We contacted companies based on this list and based on availability. Additionally, we broadened this set of companies in order to also get a comprehensive account on the topic of retention actions, as even a company in a less relevant industry could have interesting insights, e.g. due to a successfully developed analytics department modeling churn or other consumer behaviors.

In total, 28 interviews were conducted, both with decision makers at four Insurtechs, as well as with other subscription companies as proxies, such as two traditional insurance companies, three electricity providers, six telecom providers, six media and streaming services, as well as a couple of other companies. For the list of participating interviewees see Appendix A. The respondents have been anonymized to the extent that company names have been encoded with a name indicating the industry and type of company, the positions have been made more generic, and names have been adjusted, all to ensure that the interviews are treated with sufficient

confidentiality while also allowing the reader to follow and get some context about the interviewees industry and position. Throughout the thesis, when referring to a specific detail from an interview, the respondent is referred to this anonymized name as described in Appendix A.

### **3.2.5 Qualitative Analysis**

The responses of the interviews are synthesized into levers and best practices by recording and iteratively analyzing the information provided by the experts. From the responses of different experts in different industries, levers are constructed by generalizing the different approaches to retention management into more overarching principles. These levers are developed iteratively and adjusted as more interviews show that some approaches may not fit within the current framework, or that levers can be further generalized or are best split into more parts. After constructing the finalized levers, the responses of all the interviews are fit into the framework, and the best practices identified from those approaches that suggest the largest impact on retention.

# 4 Data Analysis

*This chapter describes the processing of the data, transformation of features, and comparison of different models.*

## 4.1 Constructing the Dataset

After the raw data was retrieved, a relevant time was drawn. Next, the data went through feature engineering in order to extract relevant features. After feature engineering, the resulting features were explored before modeling.

### 4.1.1 Sampling

The data retrieved is in a tabular format, consisting of one customer per row, and information about each customer as columns. The exact number of customers in the dataset is not discussed for confidentiality reasons, however it can be said that the dataset consists of over 10,000 customers. As the prediction goal is to predict which customers will churn within 3 months from a prediction date, the current status of the customer cannot be used to train models, as we do not yet know which of the current customers will churn in the future. Rather, the state of a customer 3 months ago, or longer, from the current data must be reconstructed. This was done by assigning a prediction time point to each customer and removing all events that happened after that time point. These events include e.g. switching insurance contracts when moving, making claims, or writing messages in the chat. Next, as the prediction goal is to predict 3 months forward from the current customer base, the status of the customer base at the time must also be reconstructed. This includes removing all customers who signed up after the prediction date and removing all customers who had already churned at that date.

## 4.1.2 Feature Engineering

Feature engineering involves extracting relevant numerical attributes from the data that can be used to successfully fit models and predict churn.

### 4.1.2.1 *Feature Extraction*

The first step of feature engineering is to extract the relevant features from the dataset. This step can involve both extracting new features from the data, as well as defining derived features that can be reproduced from other existing features.

For some customer attributes, e.g. age, no more features can be extracted and only derived features such as whether the person is older than 30 can be added. For many of the attributes of this dataset however, there is no single feature defining the information, rather such features must be extracted from the information at hand. These attributes include insurances, claims, referrals, messages, and discounts. The process of feature engineering can always be extended, as more features could be extracted or derived. In this project, the feature engineering was done in an iterative process, where features were tried, tested, and improved on. The feature engineering of each of the attributes are described below:

#### 4.1.2.1.1 Insurances

With insurances, the current, as well as the history of insurance contracts that the customer has had with Hedvig is referred to. For each contract, there exists data on e.g. what kind of insurance the customer has and at what premium, information about the home that is covered such as its size, as well as information about the start and end dates of the contract.

From this information, the most straightforward extraction is to include information about the current insurance contract, which we included. In addition to this, there may be more predictors of customer loyalty related to churn regarding the customer's insurance history. For example, the person may have edited their insurance with Hedvig, either because they moved to a different home, or because the amount of people living there changed when someone moved in or out of the home. For example, having stayed with Hedvig after moving, rather than switching to a competitor, may be an indication of loyalty and lower churn risk. Hence, the number of insurance edits at Hedvig is included as a predictor. Another example of where past insurances may be predictive of churn is when a customer terminates their contract and then comes back. This could be a predictor of loyalty, where the person has maybe tried some other insurance and then returned, indicating higher satisfaction with Hedvig. On the other hand, this could be an indicator of higher churn risk, as this person has switched insurance providers recently and may do so again. Another feature that can be extracted is whether or not the customer has an upcoming insurance starting at a future date. This could for example be when a customer requests their insurance to change in a couple of months as they will move

at that date. Having an upcoming insurance may be a strong predictor of retention, as people who have done such a change may be unlikely to quit before their new insurance has taken effect, as they have probably already decided on Hedvig for their next home. As some customers who sign up for Hedvig are still bound by their current insurance until a future date, they are not insured by Hedvig until that date even though they have signed an insurance contract. To capture the effect this might have on churn, whether or not the customer received insurance at the sign date is included.

#### 4.1.2.1.2 Claims

For claims with Hedvig, the data specifies what the claim was about and its result, e.g. in terms of payout. From this data, the result of the last claim is extracted, as the process of making a claim may have a strong correlation with churn if the customer is dissatisfied, or loyalty if the person is satisfied with the service and outcome. One potential indicator of churn is if the customer has had any claims at all or how many claims the customer has had in total. After completing the claims process, the customer receives an email asking to rate the claim on a 1-6 scale, and this information can be extracted for the customers that have chosen to respond. A claims rating could be a good way to capture the sentiment of the customer after their claim, and low scores may be indicative of churn while high scores may be indicative of loyalty.

#### 4.1.2.1.3 Referrals

As referrals lead to a premium discount per referral, giving customers a lower price which is hard for a competitor to directly match, it is probably correlated with loyalty. In addition, the action of referring someone reveals engagement from the customer, meaning they may be more likely to be satisfied with the product and stay. For this project, only the number and timing of referrals are included. For a more advanced analysis, one could study the network of referrals, i.e. look at the customers that were referred and their behavior. For example if one of the referred customers churns, the person that referred this customer may be more likely to churn as well.

#### 4.1.2.1.4 Messages

Messages in the Hedvig app are used to make claims and connect with customer service to e.g. make changes to the insurance or ask questions. In this project, in order to limit its scope, we only extract the timing and number of messages sent. For more advanced analyzes, the content of the messages could be studied, e.g. through sentiment analysis. If the customer seems dissatisfied in the interaction, it may be an indicator of churn. Since messages, only considering their timing, are associated with many different behaviors, they may be a weak predictor. With this methodology, messages regarding e.g. asking about coverage is treated the same as e.g. messages saying how satisfied the customer is. Since messages are often



clustered in time according to different events or interactions, the number of interactions is extracted by looking at the number of times that a customer has written messages more than 2 weeks apart.

#### 4.1.2.1.5 Discounts

Discounts are given out to some members as part of campaigns or promotions. Discounts are often in the form of e.g. giving away one or two months free, or six months with a 20% discount. In the data, discounts are represented by a starting date, an end date, and a percentage of monthly premium discount. From this, many features can be extracted. For example, whether the person has ever received a discount may be relevant as well as if they have an active discount. In order to find customers whose discount soon runs out and may churn because of that, the time until the discount ends are extracted. In addition, customers who have an upcoming discount may be less likely to churn as they may want to take advantage of it before leaving.

#### 4.1.2.2 *Feature Encoding*

After extracting relevant features, some of them need to be further encoded as the machine learning models attempted in this project can only handle numerical inputs. The data types that have to be encoded are categorical and time series features.

##### 4.1.2.2.1 Categorical Features

Some features, such as the insurance type, is categorical, and these categories need to be encoded into numbers. The two most common ways of doing this is either encode each category as a unique number (e.g. in the case of insurance type, Student-Rent could be 0, Student-BRF could be 1, etc.), or by using dummy variables. Dummy variables are indicator variables indicating that a customer belongs to a certain category, where the categorical column is replaced by as many new columns as the number of categories. In this case, a 1 in the first column for example and a 0 in all others may encode Student-Rent, while a 1 in the second column with a 0 in all others may encode Student-BRF. As one column is redundant in this case (the person is Student-rent for example if there is a 0 in all other categories), one of the columns is dropped, and hence the effect of being in any other category is measured in relation to the dropped column.

Some of the models used, e.g. logistic regression, are generally best coupled with dummy encoded attributes, while tree-based models work with both. As none of the tested models are best coupled with the number encoding, dummy variables are used.

##### 4.1.2.2.2 Time Series Features

Features in the dataset that are not only representing a single value, but rather a time series of events, such as messages or referrals, need to be further encoded as the

models tested cannot be used with dates. These need to be encoded into a numerical format, by for example counting the number of events in a time interval, the frequency of them, or the time since the last occurrence.

For messages, referrals, claims and insurances, the total number of events was defined as a feature. In addition, the number of events in the last 3 months was included, as more recent behavior may be more indicative of churn. The frequency of events, defined as the number divided by the number of years a customer has been with Hedvig, was also included. This helps decrease the correlation between the number of events and the time as a customer, as customers who have been with Hedvig for a longer time period have had more time to e.g. refer friends. Finally, the time since the last event was included. If the customer referred someone or moved last month, they may be less likely to churn, while if they have not interacted in a while, they may be more likely to churn. For claims, where there is more information about the event, information about the last event is included, e.g. its payout and rating.

#### 4.1.2.3 *Missing Values*

Some features in the dataset have missing values. This includes for example information regarding the last claim; if the person has had no claims, then there is no last claims payout to speak of. Most models tested, the exception being tree-based models, do not have any way of encoding a value as missing, instead a numerical value must be provided.

For some features, choosing this value can be difficult as there is not always any straightforward or single way to encode them. If the customer has no claims, should this be equivalent to a claim payout of zero? This value may be problematic since this value does not distinguish between a customer receiving no payout since their claim was not covered - which may lead to disappointment and churn - and the base case of not having a claim. At the same time, if a positive amount is set for each customer with a missing value, should it be equivalent to a large or small claim? Finally, using a negative value to indicate that there was no last claim may be problematic for the models that assume a monotonic relationship such as logistic regression, as this implies an opposite effect as having a large claim payout. For some features, for example the discount percentage of monthly premium, setting a default value is easier, since in this case a 0 value is equivalent to not having received a discount.

In general, the missing values were set such that the effect on the output variable for the customers with missing values was in accordance with the general effect for having that value. For a discount for example - not having received a discount is considered equivalent to having had a discount that expired a long time ago.

#### 4.1.2.4 Transformations

Some of the tested models, including logistic regression and SVM, work best with standardized data, that is that the scale of each feature is the same and that each feature is zero centered. In addition, standardized features make weights more easily interpretable. For example, if age varies from 0-100 years, while the number of referrals only varies from 0-10 in the dataset, a higher weight for the referral variable may not indicate that this feature has more weight on the output variable than for age, just that the scales are different. There are many ways to normalize the parameters. One of the more straightforward and simple ones, which is used in this project, is to subtract each feature with its mean and divide with its standard deviation. Using this method, the standardized value for a given parameter  $k$  for each customer  $i$  is calculated by:

$$X'_{ik} = \frac{X_{ik} - \mu_k}{\sigma_k}$$

Here,  $X'_{ik}$  is the standardized value,  $\mu_k$  sample mean of feature  $k$ , and  $\sigma_k$  sample standard deviation.

In addition to standardization, some models such as logistic regression generally perform better when the predictors are approximately normally distributed (Kuhn & Johnson, 2013, p .282). As some predictors are not normally distributed, such as how long customers have been at Hedvig, performance may be improved by transforming the values to better fit a normal distribution. There are many ways to normalize the predictors depending on the distribution of the parameters. One general approach is fitting a Box and Cox transformation (Kuhn & Johnson, 2013, p. 32-33) for each predictor. For this project, since the number of predictors was small, the distribution of each feature was inspected manually, and a log-transformation was applied to the relevant features. For some features, where 0 is a possible value, such as the number of months since the last claim,  $\log(x + 1)$  was applied.

### 4.1.3 Feature Exploration

After engineering the features, the features were inspected and iterated on. As a first step, the features were inspected by plotting and inspecting their distribution to see that the calculations were correct and in line with expectations. As Hedvig does not want the data to be published since it is sensitive to competitions, these plots are not included in this report. After inspecting the feature distributions, their correlation and relationship to the output variable was studied.

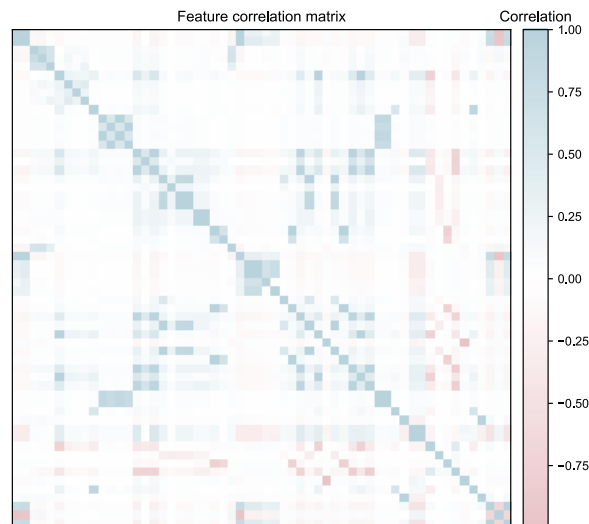
#### 4.1.3.1 Feature Correlation

As many features were extracted from the data, and many of them related or described the same events, some of them may be correlated. Measuring this correlation is of interest both to understand the data, and because too many or too correlated variables may be problematic for some models.

Heavily correlated variables can be problematic when modeling for two reasons. The first reason is that many models assume independent variables and generally perform better when this assumption is true (Hastie et al., 2009, p. 400). The second reason is that correlated variables may make model interpretation harder. If two variables are heavily correlated, the importance of them are not as straightforward. Depending on the interpretation method used, they may both get a low importance - as removing one of them results in the same performance since the correlated variable can be used instead - or share the importance among them, making them seem less important (James et al., 2013 p. 243). As such, understanding the correlation among variables in the dataset is important, and reducing correlation by e.g. choosing one of several highly correlated variables may be favorable. In order to study the correlation between the variables in the dataset, the pairwise Pearson correlation coefficients were calculated between each pair combination of the parameters (James et al., 2013 p. 70-71):

$$Cor(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

These correlation coefficients were then plotted in a heatmap, shown in figure 4.1.



**Figure 4.1** Correlation heatmap of the engineered features. Note that the feature names have been removed for confidentiality reasons.

From this heatmap, it is clear that some predictors have high correlation. One group of predictors that are highly correlated are features concerning the home and insurance contract, such as age, square meters and premium. This relationship is likely since younger people tend to live smaller, and smaller homes tend to have a lower premium due to Hedvig's pricing model. In addition to this group, there is a high correlation between features describing the same events such as different features extracted from discount information or extracted features from e.g. claims or referrals such as the total number of claims and the number of claims during the last 3 months. The features concerning messages are also correlated with e.g. edits and claims, since the chat is used for these purposes.

#### 4.1.3.2 *Feature Relationship to Output Variable*

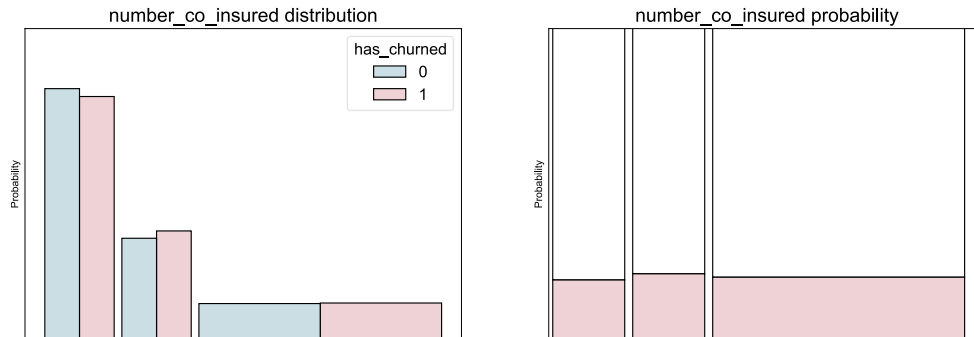
In addition to inspecting the distribution of features, and their relationship to each other, the relationship between the features and the output variable was studied. This process helps identify which features are good predictors of churn and which are not, and what values of the features correspond to a high or low churn rate. For each feature, the distribution of that feature for the churners and non-churners was plotted and compared, as well as the churn rate for different values of the features. As this data is sensitive, the full set of plots is not included in this report, instead a couple of examples are illustrated.

In general, the relationship to the output variable can be divided into two types:

1. Low or no signal in the feature to predict the output variable
2. A distinct signal and relationship between the feature and the output variable. In some cases there is a clear signal, but the feature values where the churn rate is significantly different is uncommon in the dataset, and hence only applies to a limited number of customers.

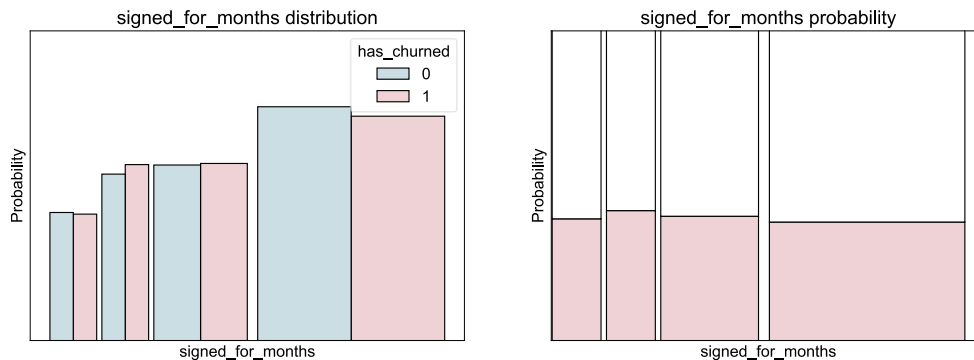
In the plots of features shown below, the features have been binned in order to better show the relationship with less noise than showing each data point. The bin thresholds were chosen manually by a combination of natural breaks in the data - for example no debts, one debt or multiple debts - and inspecting the distributions to look for bins that capture its structure. In addition, the y-axis ticks have been removed in the plots, and the scales are different on all the plots in order to not reveal the true distribution of the data, as requested by Hedvig.

An example of the first category with little or no distinct relationship that can be seen in the data is the number of co-insured, as seen in figure 4.2. Living alone, with one person, or even three or more does not appear to have a significant effect on churn.



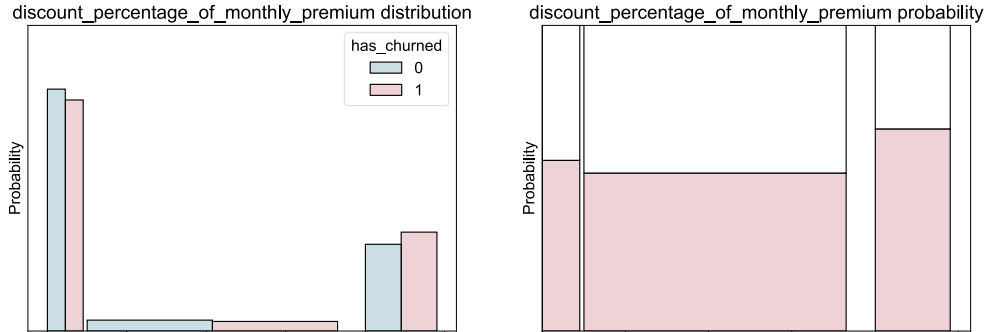
**Figure 4.2** Distribution plot (left) and percentage churn in each bin (right) for the number of co-insureds included in the insurance contract.

Looking at the time a customer has been at Hedvig, measured as the number of months since the first insurance contract was signed, a trend of increasing retention is expected, as longer time users probably are more satisfied with Hedvig, which was found to be true in the insurance company Gjensidige by e.g. Günther et al. (2014). Looking at the plot in figure 4.3, this relationship is not very strong. The churn risk seems to decrease for longer time customers who have been at Hedvig for a long time, however the difference to new users is not very large.



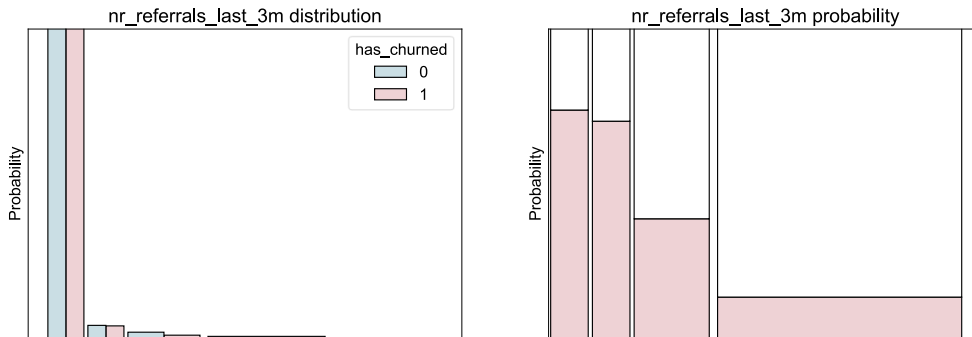
**Figure 4.3** Distribution plot (left) and percentage churn in each bin (right) for the number of months customers have been at Hedvig. Note that the x-axis has been removed for confidentiality reasons.

An example of a feature that does have a clear signal is the percentage of the monthly premium given to a customer, as seen in figure 4.4. In this case, users given free months are more likely to churn than those given no or a low discount.



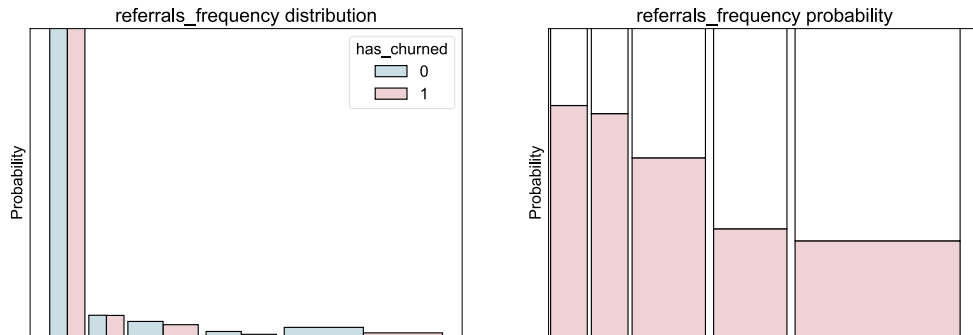
**Figure 4.4** Distribution plot (left) and percentage churn in each bin (right) for the percentage of monthly premium given to customers.

Another feature that is clearly linked to retention, is the number of referrals. In figure 4.5, the number of referrals in the last three months is plotted, and the churn probability clearly declines with the number of referrals. A potential issue with using this feature as a predictor for churn however is that the feature values that are associated with retention are uncommon and hence they cannot be applied to a majority of the customers. This is something that can be seen in many features, another example being the last claims ratings - if a low rating is given, the risk of churn is higher, but the number of customers who have a claim, chose to rate it, and rated it a low rating is low.



**Figure 4.5** Distribution plot (left) and percentage churn in each bin (right) for the number of referrals done in the last 3 months.

In the case of referrals, using the frequency of referrals per year still has a signal, and the higher values are more common as can be seen in figure 4.6. This transformation can however not be done to all features, and hence their predictive power may be limited.



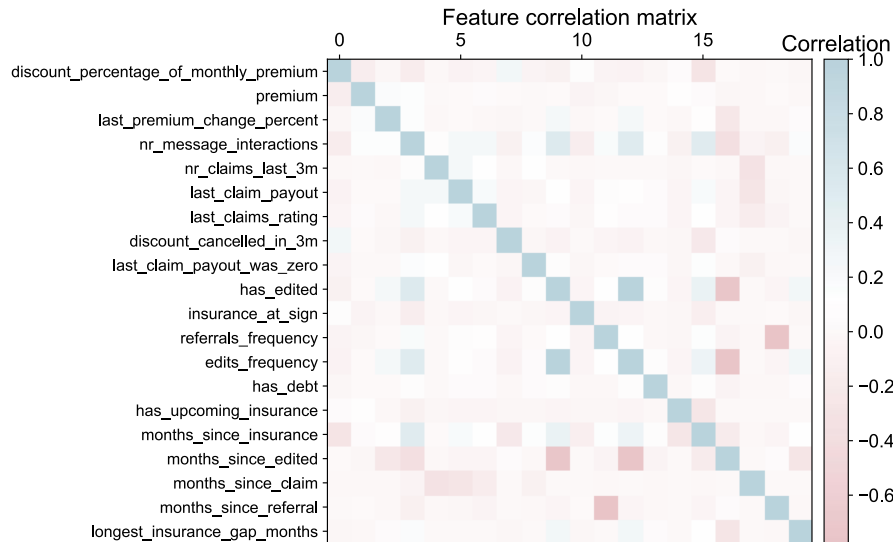
**Figure 4.6** Distribution plot (left) and percentage churn in each bin (right) for the referral frequency, measured as the number of referrals divided by the number of years the customer has been insured at Hedvig.

#### 4.1.3.3 Feature Selection

After extracting, processing, and exploring the features, the final step is selecting which features to use in the models. Depending on the model, including features with no signal may either do nothing as they are ignored, or it may even have a negative effect on performance. In this project, feature selection is treated as a hyperparameter and optimized for each model separately. In order to select the features, two methods are used: ranking features according to ANOVA F-score and picking the top percentile, or manually picking a subset of the features. In addition to feeding the models with a selected subset of features using these methods, many of the models themselves select or prioritize features through regularization.

For the first method, the ANOVA F-score of each feature is calculated to see which features have the strongest relationship to the output variable. The F-score measures how much the output variable varies depending on the feature value, and ranks features according to how large an effect they appear to have on the output variable (James et al., 2013 p. 75-78). The hyperparameter to tune is then the percentile of features to use - ranging from 0 meaning no features should be used, 50% meaning the top F-scoring half of the features should be used, to 100% meaning all features. In addition to this method, a manual subset of features was selected, based on the correlation and relationship with the output variable plots. For clusters of correlated features, e.g. all features related to referrals, only one or a few features were selected, based on those which ones appeared to have the strongest relationship to the output variable. A correlation plot between the features in this subset can be seen in figure 4.7. As seen in the plot, there are still correlations between some variables, such as whether the person has any edits and the edit frequency, however the number of correlated variables is significantly reduced.





**Figure 4.7 Correlation heatmap of a manual subset of the engineered features.**

## 4.2 Modeling

After extracting data and engineering features, models are trained to make predictions. In this project, a number of models are trained and compared, in order to select the one that is best suited for this problem in particular. This model is then further inspected to understand the churn drivers that it has extracted from the data.

### 4.2.1 Hyperparameter Optimization

Before evaluating and comparing the models, each model has to be optimized in terms of hyperparameters, in order to maximize the performance of each model. These hyperparameters come in two general categories. The first category is the hyperparameters of the mathematical model or algorithm, such as the regularization coefficient in logistic regression, or the maximum tree depth in the tree models. The second category includes the processing steps described in this report, which includes transforming features, balancing the data for training and feature selection.

In this section, each model will be optimized by sampling random combinations of these parameters and evaluating them. Because of computing constraints, not all combinations or values of parameters can be tried, which means that a suboptimal solution may be reached. To extend this project in the future, more rigorous optimization could be done to potentially increase performance further.

Classification models can be viewed as a pipeline of steps - first processing the data and features and then feeding this into a model with a set of hyperparameters. This pipeline, and the possible parameter values of each step, is illustrated in table 4.1.

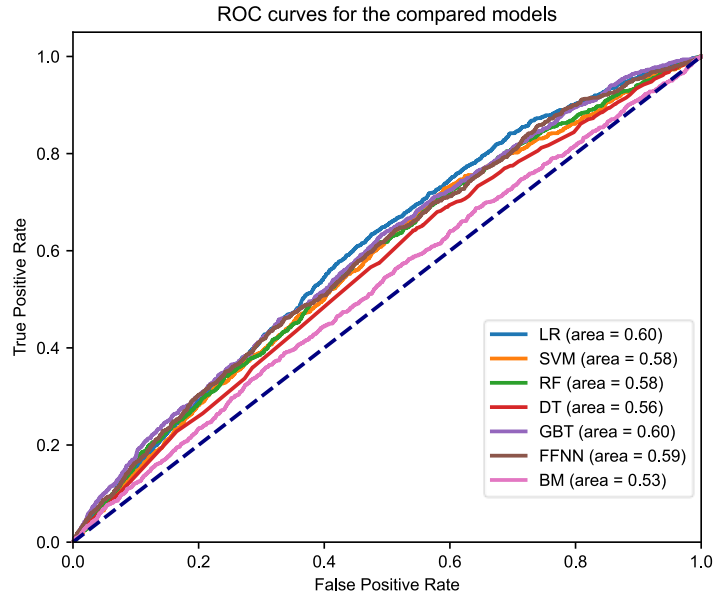
**Table 4.1 Processing hyperparameters for the models.**

<i>Step</i>	<i>Possible parameter values</i>
<i>Preprocessing</i>	<ul style="list-style-type: none"> <li>• No transformation</li> <li>• Transform the data as explained in 4.1.2.4 <i>Transformations</i>.</li> </ul>
<i>Feature selection</i>	<ul style="list-style-type: none"> <li>• Use all features</li> <li>• Selecting the best percentile of features according to F-score as described in 4.1.3.3 <i>Feature Selection</i>, varying the percentile. The tested percentile values are 10, 30, 50, 60, 80, 90 and 95.</li> <li>• Using the manual feature set according to 4.1.3.3 <i>Feature Selection</i></li> </ul>
<i>Balancing</i>	<ul style="list-style-type: none"> <li>• No balancing</li> <li>• Random undersampling</li> <li>• SMOTE</li> <li>• Model specific balancing</li> </ul>

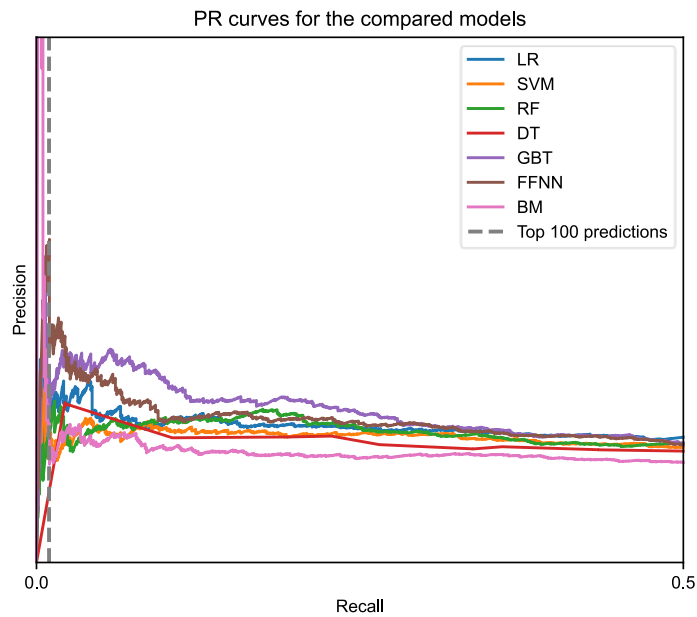
For each model, different combinations of these parameters are tried, and combined with the hyperparameters of the model, and evaluated with a 5-fold cross-validation. The best model is then selected among the best performing classifiers in the population. In order to reduce this process sensitivity to random chance, i.e. one set of parameters showing higher performance than the rest because of the randomness in the data and cross-validation process, the effect of each parameter is studied as well to choose the best values. The trial hyperparameter values were chosen in order to cover the range of possible values of the parameter, and were iterated on as optimization sessions were run, adding values close to the best performing values in previous sessions. The conducted hyperparameter optimization is further described in Appendix C.

#### 4.2.2 Model Comparison and Selection

After optimizing each model individually, their performance can be compared in order to choose the best model. In figure 4.8 below, the ROC curves of the optimal models are plotted, and the PR curves are shown in figure 4.9. The area under the curve ROC curve is provided in the plot legend and summarized in table 4.2. These curves are determined through 8-fold cross-validated prediction, meaning that the training data of a particular fold is used to predict the test data in the fold.



**Figure 4.8** ROC curves for the optimized models with area under the curve shown in the legend parenthesis.



**Figure 4.9** PR curves for the optimized models. Note that the y-limit is not 100% and that no y-axis is shown in order to disguise the churn rate of Hedvig. Recall corresponding to the 100 customers with the highest predicted churn probability is shown as a vertical line.

**Table 4.2 AU-ROC for the compared models. Highest values annotated as bold.**

<i>Model</i>	<i>AU-ROC (8-fold cross-validated on the training set)</i>
<b>LR</b>	<b>60%</b>
DT	55%
RF	58%
<b>GBT</b>	<b>60%</b>
SVM	58%
FFNN	59%
BM	53%

As seen in the plots, the difference between the models is relatively small. The worst performing models on both metrics are DT and BM, the rest of the models showing a very similar level of performance at ~60% AU-ROC. Even though the AU-ROC scores are determined by cross-validation, the scores vary depending on the random seed when splitting the data in a training and test set, making it harder to decide between models with similar performance. The best performing models according to AU-ROC are LR and GBT, with the same values in the comparison table. However, looking at the PR curves in figure 4.9, GBT has considerably higher precision for low levels of recall, e.g. at the top 100 prediction line shown in the curve. As such, for this purpose, GBT provides up to double the precision than LR, given that the recall required is low enough. On the other hand, LR is considerably more interpretable in itself, and simpler to implement. As such, for a pure predictive task, the GBT model offers better performance, especially for low levels of recall, and is hence chosen as the best prediction model for this project. As logistic regression offers a similar level of performance and intrinsic interpretation, it is considered as well.

### 4.2.3 Churn Drivers

As GBT performs the prediction task the best, inspecting this model will mean that the conclusions from the interpretation are the closest to the actual drivers in the data set. Hence, this model will be inspected using the tools described in section 2.3.5 *Model Interpretation*. In addition, since logistic regression performed on a similar level, and is easily interpreted, it is inspected as well to get another

perspective and further understand the churn drivers. In addition to the logistic regression model with  $L^2$  regularization, a  $L^1$  regularized, sparse LASSO model is presented and interpreted. The analysis of the churn drivers is conducted in Appendix D.

It is important to note that the predictive accuracy of both the GBT and LR model is low, and hence they are not able to fully model the dynamics of churn in the data set. This fact limits the conclusions that can be drawn from inspecting the models, as they themselves have not fully understood or modeled the churn. However, the models do pick up some signal in the data, as their predictive accuracy is non-zero and better than random guessing. Hence, some knowledge can be extracted from the models, acting as indicators of churn drivers. Because of the low performance score, only the overall drivers are examined, rather than detailed examination and inspection of e.g. interactions between features.

# 5 Results

*This chapter presents the results of the churn modeling and interview study.*

## 5.1 Identify

The best model to predict churn at Hedvig was concluded by randomly sampling different hyperparameters for all models including preprocessing, feature selection and balancing. From this process, the best model was decided as a Gradient Boosted Tree model, using the XGBoost implementation. For this model, categorical variables were dummy encoded, and all features standard scaled. Further, no feature selection was done, and balancing was achieved via the built-in *scale\_pos\_weight* parameter of the model rather than a sampling algorithm such as SMOTE.

Even though this Gradient Boosted Trees model achieved the best performance in the comparison, it is worth noting that performance very similar to this model was achieved via simpler models such as logistic regression with a ridge penalty, and even a sparse LASSO model using only five features.

### 5.1.1 Model Performance

The ROC and PR curves for the best model on the test set are shown in figure 5.1 and 5.2 below, with the benchmark model as baseline.

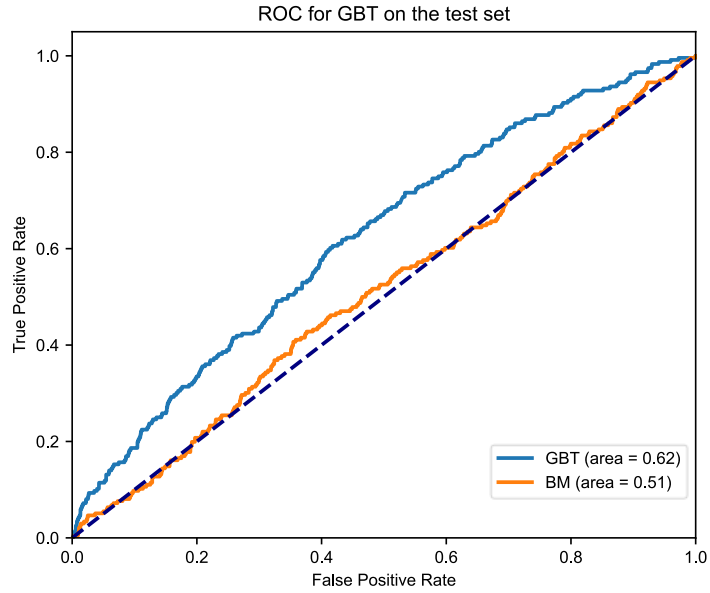


Figure 5.1 ROC curves for the best model and benchmark model, with area under the curve shown in the legend parenthesis.

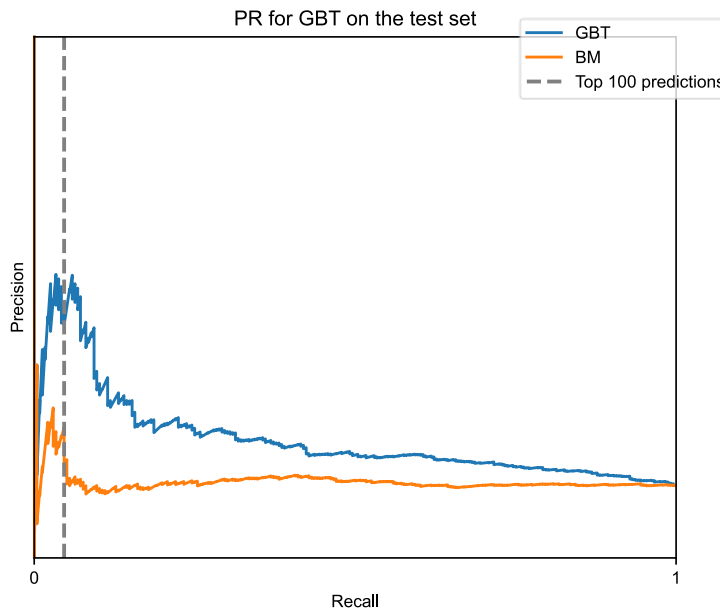


Figure 5.2 PR curves for the best model and benchmark model. Note that the y-limit is not 100% and that there is no y-axis, to disguise the churn rate of Hedvig. Recall corresponding to the 100 customers with the highest predicted churn probability are shown as a vertical line.

As seen in the ROC plot legend, the **best model achieved an AU-ROC of 62%**, compared to the **benchmark model with 51%**. Looking at the PR curve, the best model has a higher precision for all levels of recall. Further, looking at the top 100 churn scoring customers, the best model has about twice the precision of the benchmark, meaning that twice the customers on the list of top 100 customers at-risk of churning will actually churn compared to the benchmark. Note however that the **precision of both models is low, well below 30%**.

In figure 5.3 below, the histogram of churn probability output from the model is shown, comparing the customers who actually churned and those who did not. In the plot, it is evident that customers who are classified as higher probability to churn from the model are more likely to churn, as customers over ~55% estimated churn probability are quite likely to churn and that customers with less than ~40% churn probability are unlikely to churn. However, the classes are not fully separated by the model, rather there is a big middle ground that most customers fall into, where model output is not a significant predictor of churn.

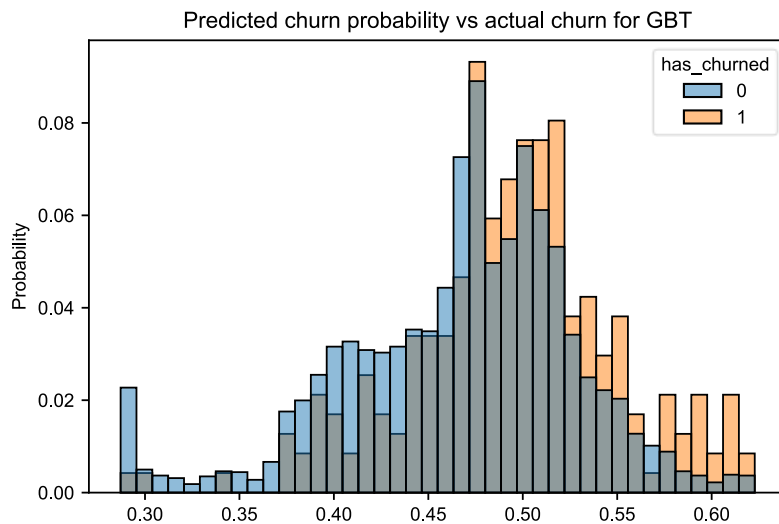


Figure 5.3 Distribution of churn probability output vs actual churn for the best model.

### 5.1.2 Churn Drivers

In order to understand the drivers of churn at Hedvig, the best model was interpreted using PFI, ICE, PDP and SHAP, as described in Appendix D. In addition, the coefficients of two logistic regression models were analyzed.

The most significant driver according to all methods was the line of business model of the customer, renters having the highest churn rate, while customers who owned their own home via a BRF had significantly lower churn rate. Next, the premium



seems to have a large effect on churn, where a premium around 100 is associated with the highest retention and a higher or lower premium appears to have an increasing churn risk. As premium encompasses many things and is correlated with e.g. line of business this is hard to interpret. The number of message interactions also correlates with churn, with the churn risk increasing with the number of interactions. As messages include a myriad of different things such as claims, edits or general insurance questions, the reason for this is unclear and needs to be studied further. Further, retention appears to be increasing with the time a customer has been at Hedvig, indicating loyalty over time. Retention is not only affected by the time a customer has been at Hedvig, but also when they signed up for the service. The most significant difference is if the customer signed up in July, where retention is comparably higher. Another factor that influences churn is the acquisition source, how the customer signed up to Hedvig. Specifically, if the user signed up through a referral, they are more likely to stay. Finally, the timing and number of referrals that the customer has done has a significant impact on retention. The more referrals a customer has, and the more recent their last referral was, the less likely they are to churn. With insurance edits, the relationship is more complex as retention increases with edits at first but decreases when the edit frequency is very high.

As a final note regarding the churn drivers, the models that were inspected to extract these drivers were not very accurate predictors of churn at Hedvig. As such, the models do not fully understand the relationship between the features and churn, and their interpretation should hence be read with caution.

## 5.2 Actions

In total, 28 interviews were conducted, both with decision makers at other Insurtechs, as well as with other subscription companies as proxies. For the list of participating interviewees see Appendix A, and for the full interview guide see Appendix B. Based on these interviews, we have examined the impact of retention work and trends for subscription companies. Further, we have identified seven levers to increase retention in subscription companies, followed by a discussion of the relevance for Insurtechs and specifically for Hedvig.

### 5.2.1 Impact of Retention Work

Depending on the industry, proactive retention work, combined with reactive save desk operations, can decrease churn by between 25-50% according to interview testimonials. However, this does not mean that all their revenue is saved, as retention work, including personnel costs and potential discounts are needed, and therefore one needs to test and see what level of retention is profitable. Furthermore, as

discussed in section 1.1.3 *Churn and Retention*, an increase in customer retention by just 5%, from 85% to 90% can give rise to an increase in net present value profits from 35% to 95%, shedding light on the critical role of retaining customers (Reichheld & Sasser, 1990).

According to Malin, customer loyalty executive at Audio streaming service B, one needs to understand why people stay, finding the customer group with a need for the service, who may not always know that they have this need. Additionally, one needs to understand decision points when customers consider changing service providers or cancelling their contracts, and to measure customer satisfaction in connection to an interaction or claim, and then follow-up if something goes wrong. A common way to do this is by following detractors in Net Promoter Score (NPS) surveys. NPS surveys ask customers how likely they are to recommend the service to a colleague or friend, and measure customer satisfaction.

Retention expert Camilla at Loyalty Factory emphasizes that there is a difference between loyalty and retention, where loyalty is the customer's satisfaction over time and retention is about saving customers who are at-risk of churning. Companies need to work with both, and for a subscription-based service without a binding period, one needs to focus even more on long term loyalty. Robert, ex-executive at Top 4 Swedish insurance company B, agrees and says that traditional insurance companies with one year binding period put all their efforts during the last three months before contract expiration.

Hugo, customer loyalty executive at Major Swedish media company A, says that retention work is extremely important in their business, as it is an industry that is not growing. For other companies, where the industry is growing rapidly, such as video streaming services, having a large churn is not as critical as it can be balanced out by a high customer acquisition rate from the growing market. In traditional television, there are no additional new tv subscriptions, and if Swedish media company A would not work with retention, customer loyalty executive Hugo says he would expect to see about twice the outflow of customers, and without the save desk probably even more than this. Comparing TV with insurance, insurance is not a shrinking market, but it is also not growing rapidly, indicating that retention actions can have a significant effect on retention.

#### 5.2.1.1 *Trends in Retention*

The importance of retention management has, according to Malin, customer loyalty executive at Audio streaming service B, increased over the last couple of years, as it is becoming more expensive to acquire new customers, they have more complex needs, and they need to make frequent decisions every month to stay or to leave. In 2021, the spend on existing customers is, according to Loyalty Factory, expected to increase by 30%, partially due to new technology and data availability. These investments are often spent on loyalty programs and new digital technology such as marketing automation systems, in order to better handle the customer relationships,

and about half of the companies surveyed plan to invest in marketing automation. Marketing automation refers to the use of software platforms and other technology in the automation of repetitive tasks and effectively communicating with customers through multiple channels. It is a useful tool when scaling up marketing efforts from ad-hoc campaigns to elaborated marketing processes.

Another factor that will affect this area is the availability of personal customer data on an individual level. In general, because of increased regulation on privacy, as well as public opinion shifting on this subject, high quality customer data is harder to acquire and store. For example, according to a Product Manager at Electricity provider startup A, there is a concern in the industry regarding the privacy changes in Apple’s new iOS14 update, allowing users to block the IDFA (Identifier for Advertisers) at the app level, making it possible for users abstain from sharing their data. This will make it more difficult for advertisers to target the right customer groups. In addition, the ‘quality’ of the acquired customers will decrease, driving customer churn and possibly making retention studies increasingly important.

### 5.2.2 Levers to Increase Retentions

Based on industry interviews, we identified seven levers that can be used to increase retention, around which we based our discussions with our interview study respondents. These levers are 1) Understanding Churn, 2) Customer Intake, 3) Product Improvement, 4) Lock-in, 5) Targeting at-risk Churners, 6) Save Desk, and 7) Organizational Setup, and are visualized in figure 5.4 in relation to the customer journey.

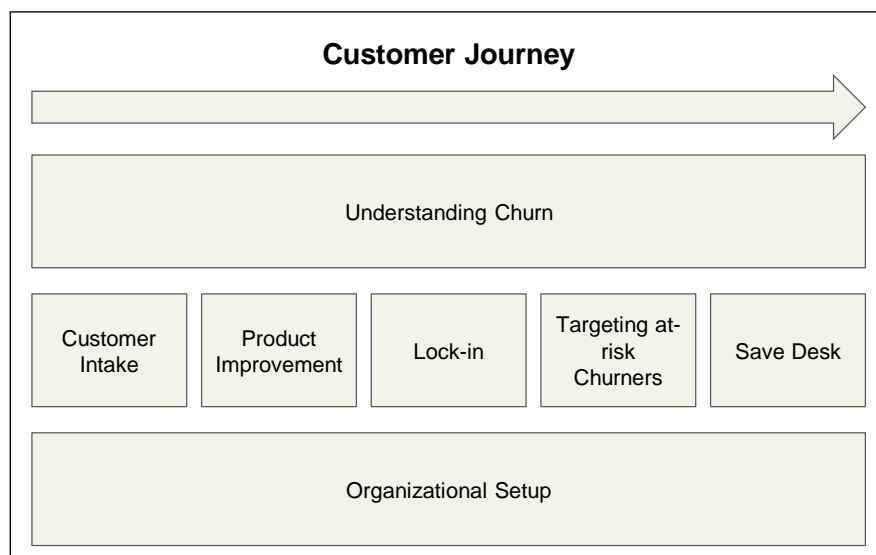


Figure 5.4 A framework of seven levers to increase customer retention.

#### 5.2.2.1 *Understanding Churn*

This lever relates to the *Identify* part of our thesis. Spending the same number of resources on all customers for a retention purpose is ineffective, and sometimes not even feasible. Therefore, understanding churn and the underlying drivers is a clear lever for increasing retention. This lever also includes best practices in predictive model development and as well as how the right data collection can be ensured.

#### 5.2.2.2 *Customer Intake*

The customers that a company acquires in the first place can have a large impact on their retention. If customers for example sign up for the service because of a limited time heavily discounted offer or trial period, many may choose to cancel their subscription after the discount expires. In addition, some customer segments may be less loyal and switch more frequently than others.

#### 5.2.2.3 *Product Improvement*

Having a great product has a clear impact in customer retention as customers who are happy often choose to stay on the service. Having a good product includes meeting the customers' demands, having the right features, and can extend to personalizing the offering and the content that each customer receives.

#### 5.2.2.4 *Lock-in*

There are ways to create barriers for the customers to leave, which therefore could be a lever to increase retention. This could be everything from binding periods to upselling customers with a second product or service or introducing some type of loyalty program. Having an engaged customer base could also serve as an emotional lock-in effect, keeping customers for a longer period of time.

#### 5.2.2.5 *Targeting at-risk Churners*

This lever includes the ways a company can target customers that have been identified as more likely to churn in a given period. This includes all the actions that can be taken to influence these customers to stay longer at the company and is the main focus of this project as it relates to using the churn prediction models discussed in this thesis.

#### 5.2.2.6 *Save Desk*

As a final measure, one can attempt to influence customers determined to leave to instead stay at the company. This is commonly addressed by calling customers to understand why they want to cancel and convincing them to stay by e.g. giving them a discount or something else of value.

#### 5.2.2.7 *Organizational Setup*

In order to be successful in increasing and maintaining a high level of retention, the right organizational setup is key. As retention is affected by the work of almost all functions within a company, from customer acquisition, on-boarding, product development, communication, and eventually retention or win-back team, having the right organizational setup and ownership of retention can be tricky. Therefore, it is of interest to work with retention in a structured way, having clear responsibilities and overall a good set-up and processes to follow-up the retention efforts.

### **5.2.3 Best Practices to Increase Retention**

The accounts of best practices that were mentioned in the 28 conducted industry interviews are summarized into the seven proposed retention levers in Appendix E. In this study, the focus has been on the levers and actions that relate to targeting at-risk churners that have been identified e.g. by using machine learning. Best practices are summarized in table 5.1, and described in more detail in Appendix E.

**Table 5.1 Summary of Best Practices for the seven identified levers.**

<i>Lever</i>	<i>Summarized Best Practices</i>
<i>Understanding Churn</i>	<ul style="list-style-type: none"> <li>• Collect feedback from all parts of the organization including churn data, NPS, customer service interactions etc.</li> <li>• Collect <i>triggers</i> for churn from internal or external sources.</li> <li>• Understand customer segments and their churn drivers.</li> <li>• Investigate descriptive models, looking at response to retention actions rather than churn.</li> </ul>
<i>Customer Intake</i>	<ul style="list-style-type: none"> <li>• Take advantage of knowledge of churn when choosing customers to target in campaigns.</li> <li>• Beware of risk of increased churn as a consequence of campaigns, e.g., steep discounting.</li> </ul>
<i>Product Improvement</i>	<ul style="list-style-type: none"> <li>• Match the right product with the right customer using multiple product tiers.</li> <li>• Develop product with churn drivers in mind.</li> <li>• Develop engaging features with value-add for the customer.</li> </ul>
<i>Lock-In</i>	<ul style="list-style-type: none"> <li>• Experiment with offering a voluntary binding period.</li> <li>• Offer an ecosystem of products and services.</li> <li>• Build a community around the product, resulting in <i>emotional attachment</i>.</li> <li>• Introduce a loyalty or referral program.</li> </ul>
<i>Targeting at-risk Churners</i>	<ul style="list-style-type: none"> <li>• Develop trigger-based campaigns using marketing automation system and A/B testing.</li> <li>• Communication via e.g., texting or calling to influence at-risk churners.</li> <li>• Communicate and motivate discounts to at-risk churners.</li> <li>• “Let sleeping dogs lie”.</li> <li>• Be present in channels connected to churn occasions e.g., moving.</li> </ul>
<i>Save Desk</i>	<ul style="list-style-type: none"> <li>• Staff specialized and highly skilled salespeople that know the specifics of coverage and competition.</li> <li>• Engage save desk staff in the process of improving the save desk practices.</li> </ul>
<i>Organizational Setup</i>	<ul style="list-style-type: none"> <li>• Dedicated retention manager and team, at minimum there should be a responsible manager and team for existing customers e.g., the CRM team.</li> <li>• Break down retention metric into actionable sub metrics tracked by different teams.</li> </ul>

# 6 Discussion

*This chapter discusses the results of the thesis, its limitations, recommendations to Hedvig, and suggests areas of future research.*

## 6.1 Summary and Conclusion

### 6.1.1 Identifying at-risk Churners

The best machine learning model to predict churn at Hedvig was an ensemble decision tree model, where 100 different decision trees with a depth of at most 6 were trained in order to vote on a churn prediction for a certain customer. The method used in this project does not fit these trees independently of each other, rather the model consists of a chain of tree models, where the next tree in the chain attempts to correct the errors of the previous one. The method is referred to as gradient boosted trees, and specifically the XGBoost implementation was used.

To train this model, a dataset was constructed using customer data related to their home, claims, referrals, messages, and insurance history at Hedvig. A series of processing steps were attempted in this thesis including transforming and scaling the data, feature selection, and class balancing. For the best performing model, each feature was scaled in order to have a zero mean and one as standard deviation. All features were then fed to the model and balancing was performed as an integrated step in the model rather than externally through sampling.

To evaluate the performance of the model, the Receiver Operating Characteristic (ROC) curve, as well as Precision Recall (PR) curves were studied. The values of these metrics were calculated on a part of the dataset that was unseen to the model during training, corresponding to 20% of the entire data set. The XGBoost model achieved a 62% area under the ROC curve, which can be compared to 50% which corresponds to random guessing, 51% which was achieved by the benchmark model, while 100% corresponds to perfect accuracy. Even though the model performs better than random guessing, an AU-ROC of 62% is considered low when taking actions on the predictions according to the machine learning professionals that were interviewed. Looking at the PR curve, the precision - corresponding to the percentage of customers which are considered at-risk of churning by the model that

will actually churn - is higher than the benchmark, and about twice as high for low levels of recall. To illustrate this, if one would consider the 100 customers that were considered the most at-risk of churning by the model, approximately 2X as many of these would actually churn compared to the benchmark. However, as the churn rate in home insurance is very low, this precision is indeed also low, and lower than 30%. As such, if one were to call the 100 customers that were considered most at-risk of churning by the model, well less than 30% of them would actually churn and a clear majority were probably not even considering it. **As such, customers at-risk of churning at Insurtechs like Hedvig cannot accurately be identified by the models attempted in this project.**

The reasons why churn at Hedvig was not predicted successfully can be summarized in the following points:

1. A **home insurance company has very little information** on their customers' satisfaction with their insurance as they have very few interaction and feedback points, compared to e.g. a music streaming service which can see their customers' day-to-day usage on the platform.
2. The information that Hedvig has, and the data set extracted from it that was used in this thesis, **does not include any strong predictors of churn**. Even though some characteristics may be associated with churn, they are often very uncommon, and the difference in churn probability between the classes is low.
3. The act of churning at a home insurance company is often triggered by **an external event** such as moving, which is hard to predict from the perspective of the Insurtech. The dataset does not include any strong triggers of customers being on their way to churn.

The reasons for the low performance of the models, and ways it could be improved are further discussed in 6.2 *Model Performance*.

#### 6.1.1.1 *Churn drivers*

Using the methods described in this paper, the only strong drivers of churn identified at Hedvig are 1) their line of business, e.g. renter or homeowner, 2) the insurance premium 3) how many message interactions the customer has, and 4) the timing and frequency of referrals and edits. The rest of the features considered in that data set which were expected to be strong predictors of churn e.g. claims, if the discount was recently canceled, and how long a customer has been at Hedvig did not offer any significant increase in predictive accuracy.



### **6.1.2 Actions to Increase Retention**

Based on the 28 interviews with retention professionals, we have concluded that proactive retention work, combined with reactive save desk operations, can have a large effect on retention, and can decrease the churn rate by between 25-50%. Additionally, there is a consensus that the importance of structured churn work is increasing, due to new technological possibilities, increased data availability, and the fact that acquiring new customers is becoming more expensive. The focus and spend on retention efforts is expected to increase and many companies consider investing in loyalty programs and new digital technology such as marketing automation to better handle the customer relationships. Furthermore, the seven levers we have identified that can increase retention in subscription companies are: 1) Understanding Churn, 2) Customer Intake, 3) Product Improvement, 4) Lock-in, 5) Targeting at-risk Churners, 6) Save Desk, and 7) Organizational Setup. Overall, we have not identified any ‘silver bullets’ to increase retention for Insurtechs. Rather, it is about continuously working on and improving all these aspects, which can collectively increase retention in the customer base. The conclusions regarding best practices for each lever is provided in Appendix E, and a recommendation for Hedvig provided in 6.5 *Recommendation to Hedvig*.

## **6.2 Model Performance**

As discussed in the conclusion under section 6.1.1. *Identifying at-risk Churners*, customers at-risk of churning could not be accurately identified by the models attempted in this project. In this section, we will discuss why the performance of the model is poor, as well as how it could be improved.

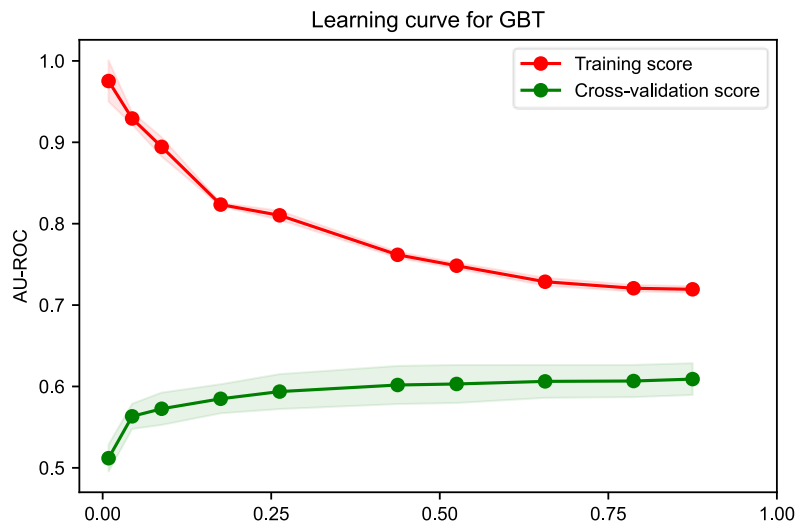
### **6.2.1 Explaining Performance**

By comparing this project to other papers and theses, as well as talking to data analysts in different industries, we have tried to understand the overall reasons why the predictive performance of the models used in this paper is low.

First, one can note that the classification models or hyperparameters used does not seem to be the issue, as none of the candidate models achieved an AU-ROC score over e.g. 65%. Further, when testing this broad range of models, building on very different assumptions and structures, their predictive performance was all approximately the same, as can be seen in table 4.2. This points to the data set being the problem rather than the modeling. The fact that the predictive performance of the tested model is low is not inconsistent with the machine learning theory

presented in this paper, as the different algorithms come with assumptions around the data, which appear not to be satisfied in this dataset.

Another potential reason why performance is poor is that the data set is too small, and that more observations are required to model churn at Hedvig. This hypothesis can be explored by looking at the learning curve of the final model, which is shown in figure 6.1 below. This curve is constructed by considering randomly sampled subsets of the data with an increasing number of observations from the data set. In the leftmost part of the plot, only a small fraction of the data set is considered, while at the rightmost part of the plot the full data set is considered. For each subset, the performance of the training set, as well as 8-fold cross-validated performance is plotted. This curve indicates whether the model would benefit from more observations in the data set, as an increasing slope could imply that more observations would increase performance further.



**Figure 6.1 Learning curve for the best model. Note that performance on the training set is better, indicating that the model is slightly over-fit.**

As seen in the plot, the learning curve has a relatively flat slope at 100% of the data, indicating that the number of observations does not seem to be the problem. This leads us to the conclusion that the reason for low accuracy is that the features in the dataset that are not strong enough predictors to accurately identify churn. The overall problems with the dataset that we have identified are the following: Few interaction points, weak predictors, and external events.

#### 6.2.1.1 Few Interaction Points

Compared to many other products, a home insurance company has very few interaction points with their customers. Looking at streaming services for example,

these companies have access to usage data of what kind of content their customers consume, at what frequency, time of day etc. Such a company can for example identify users who used the service last month but did not use it this month - which may be a reason for the customer to want to cancel their subscription. Even looking at traditional, low-engagement industries such as telco, they have access to information about how their customers call or usage data and can for example identify users who recently moved to an area with poor reception as potential churners. For home insurance, almost the only interaction the company has with the user is when they move or want to make changes to their insurance, or if they have a claim. Both these events happen very rarely, at most once or twice a year, meaning that there are few data points from the insurance company's perspective on what the customer is doing and whether they are satisfied or not. This can be compared with for example car insurance, where the insurer may get more data on e.g. how much the car that is insured is used. In addition to the few data points, home insurance is something that has a long decision time horizon and something few people consider every month, meaning that the decision to leave can happen months or years after the last interaction.

#### 6.2.1.2 *Weak Predictors*

Looking at the information that a home insurance provider such as Hedvig does have on its customers, the data does not appear to have a strong signal or relationship with churn. Most features in the data set have a maximum of 10% difference in churn rate compared to the overall churn rate. Compare this to a feature where everyone who has a certain characteristic does churn, and all who does not stay. In a Master's thesis done on churn at Spotify for example, some of the streaming usage features had a threshold above which only staying customers existed (Dinis, 2017). The problem with weak predictors is that if for example 10% of customers who recently had their discount expire choose to cancel, this also means that 90% do not churn, and it is hard to distinguish between those who do and those who do not in the prediction. In the data set, two customers could look almost exactly the same and still one of them churn and the other one not.

One reason explaining why predictors used for churn in this project are not strong enough to accurately predict churn is that most of them are risk factors rather than triggers. In this project, the outcome variable is defined as "which customers will churn within three months". As such, in order to accurately predict churn, the model needs to not only find customers who have risk factors for churning, but also indications of timing, that they will churn now. The features used in this data set are mostly risk factors which the model has identified, e.g. being a renter, or not having any referrals. Some features in the dataset act as triggers, e.g. the months since claim, attempting to capture customers who churn after a claim experience, and the time signed up, attempting to capture customers who e.g. switch every year, however these are rare in the dataset and not very strong triggers features.

Further, Hedvig has few predictors and a small data set compared to a more established insurance company such as Gjensidige, for which Günther et al. (2014) managed to accurately predict churn. Compared to Hedvig, Gjensidige has more customers, and more insurance products. In the Gjensidige paper for example, having two homes insured with the company is a significant predictor in the model, a case that does not yet exist at Hedvig.

### 6.2.1.3 *External Events*

In addition to the points mentioned above, what makes churn hard to predict at a home insurance company is that a switch is often triggered by an external event. As seen in figure 3.2, moving is the most common reason for customers terminating their contract at Hedvig. From Hedvig's perspective, moving is very hard to predict, as the customers rarely give any indication of moving before doing so. According to Cecilia, Marketing Manager at Top Swedish electricity provider, this is the reason why they are not successful in predicting churn. This does not only apply to moving, rather most other reasons for churning such as getting poached by a competitor, are also hard to predict using the data available for the insurance company.

Looking at performance of the different models, it can be noted that very simple models such as the LASSO, achieved accuracy close to more complex models such as GBT. This could be an indication that the relationships in the data are not very complex. Rather, from Hedvig's point of view, looking at simple risk factors such as the customer being a renter has the same predictive power as looking at 10+ factors. This may be because the churn event is very "random" and hard to predict from Hedvig's point of view with the current data set and can therefore not be successfully predicted outside of a few risk factors.

Finally, it is worth noting that churn in general can be difficult to predict for many companies. According to our interviews, none, or few of the major companies in insurance or adjacent industries such as consumer electricity or telco appear to have implemented successful churn models that they use. The companies that report that they have working churn models generally have more interaction points, stronger predictors and are less driven by external events, e.g. in audio streaming. Looking at previous churn prediction papers in insurance, Günther et al. (2014) achieved adequate performance of ~70% AU-ROC, but Mau et al., (2017) achieved an AU-ROC in line with this project for the majority of customers, those who had not requested online quotes.

## 6.2.2 **Improving Performance**

Learning from other projects and the interview respondents, we have identified a couple of ways that the accuracy for predicting churn at Hedvig may be improved

in the future. This discussion is limited to the churn classification approach presented in this paper.

The major improvement is to include more features who act as *triggers*, that are indications that a customer is in the process of considering their insurance and potentially terminating their contract. As moving is a common churn reason, including features that are indicators of a person moving would most likely help the model identify at-risk churners more accurately. This could include collecting information on how long customers expect to stay in their current housing, or other factors that are related to moving, e.g. when students expect to graduate. In addition, external data sources could be used, such as *Adressändring* which notifies when a customer has moved or is planning to move.

Another avenue worth exploring is looking at what behaviors customers who churn show during the weeks or months before. The Loan fintech that we interviewed for example noticed that customers checking their loan balance is a trigger for switching. This could perhaps be applied to Hedvig, e.g. when a customer checks their premium this may be indicative of them comparing their current premium to a competitor offer. Extending on this, linking individuals who have requested quotes at the Hedvig website or through comparison sites with current customers may lead to higher predictive performance, as shown by Mau et al. (2017).

Finally, an attempt could be made to increase the number of interactions and feedback points made with customers. More interactions that can be measured and related to churn could perhaps be achieved by including more engaging features in the Hedvig app, similarly to the car battery feature mentioned by the Car Insurtech we interviewed. In addition, more feedback could be collected from the interactions that already occur with customers, such as when a customer wants to edit their insurance.

### 6.3 Churn Drivers

The drivers of the final model are discussed in the results in section 5.1.2. *Churn Drivers*. The drivers that had a large effect on churn at Hedvig are 1) line of business e.g. renter or homeowner, 2) insurance premium, 3) number of message interactions, and 4) timing and frequency of referrals and edits.

Overall, the most significant driver was the line of business model of the customer, where renters had the highest churn rate, while customers who owned their own home via a BRF had significantly lower churn rate. This raises a question of whether customers who own their own apartment or house are in fact more loyal to Hedvig, or if this rather is because renters in general tend to move more often and hence reconsider their insurance more often. Answering this question could be very

important for understanding their churn and could have very different implications in terms of what actions should be used to retain the customers to a higher degree. Several of the other predictors in the data set involve similar questions, as to what actions can be taken if one knows that this driver is a strong predictor. Several of the drivers are not very actionable, however, largely due to the fact that the data mainly consist of customer characteristics and insurance contract details, rather than usage data that for many companies can be more insightful in terms of actions of how to tackle the churn.

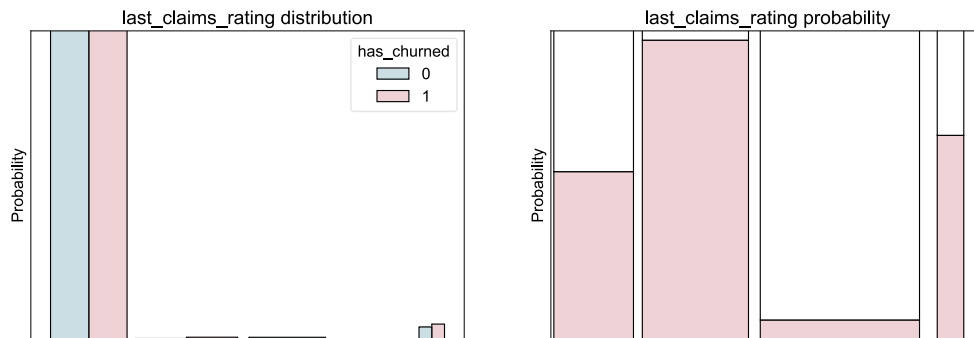
Additionally, retention seems to be higher for customers that have been with Hedvig a longer time. However, this driver is not very strong, which is in line with what we have heard in the interviews. Both Robert, ex-executive at Top 4 Swedish insurance company B and Cecilia, Marketing Manager at Top Swedish electricity provider says that there is a clear relationship that retention goes up during a customer's lifetime, however, that this trend is also first clearly noticeable and strong when some time has passed. Some of the reasons for this is that the decision-making process for changing insurance is rather long, as well as the fact that customers that e.g. already moved with Hedvig several times are plausibly less likely to churn. As a startup, none of the customers in Hedvig's customer base have been customers for more than 5 years, so this case does not exist in the data set. However, if one were to do this same analysis in 10 years' time, one would expect that the customers that have been with Hedvig until then would be considerably more likely to stay for another three months. With this in mind, the group of individuals that for a more mature insurance company would be comparably easier to predict does not exist at Hedvig, making the prediction task more difficult.

Furthermore, referrals seem to have a strong correlation with retention, both signing up through a referral and having referred other customers. This is in line with what we would expect to see and reinforces the success of their referral program in this aspect. Signing up through a comparison site did in contrast not have the significant effect that we initially expected. This could nevertheless again be because of the early stage of Hedvig's operations; a large share of the customers that have signed up through comparison sites have not been customers more than a year and may therefore not yet have evaluated their insurance decision again since signing up.

When Günther et al. (2014) analyzed churn at the insurance company Gjensidige they found that customers who had a discount last month but not any more were much more likely to churn than a customer who still has a discount. Even though there is a correlation between a canceled discount and churn risk, the effect at Hedvig is not strong enough to be a significant predictor considered by the models.

Another variable that was believed to have a significant impact was the rating that the customer had given on their last claim, which is believed to give an up-to-date indication of their satisfaction with Hedvig's service. Indeed, there seems to be a

strong relationship, where customers that had rated their last claim with a 1 or a 2 are significantly more likely to churn, as shown in figure 6.2.



**Figure 6.2** Distribution plot (left) and percentage churn in each bin (right) for a customer's last claims rating. Note that customers with no claims rating is encoded as 0.

Even though this relationship appears strong in the plot, this variable is however not considered significant by the model. The reason for this could be that there are simply too few customers that have both had a claim and then also rated the claim, generating too little additional information to the model to be considered compared to all other features. Only a small fraction has made a claim in the first place and then only one in four rated at least one of their claims. A way to increase the performance could therefore be to influence customers to be more likely to rate their claims and thereby collect more data of this potentially interesting attribute.

Inspecting churn drivers from the model can yield insights about customer behaviors. However, as the model did not fit the data well and inaccurately predicted a clear majority of customers, one should be careful about reading too much into these drivers.

## 6.4 Retention Management

This section discusses the findings of our interview study and compares the results to previous studies on retention management, including those presented in section 2.1 *Retention Management*.

The interview study conducted in this thesis was designed independently of previous retention frameworks, in order to get an “outside-in” perspective on churn management from practitioners in the industry. The levers that we present have many similarities with the retention frameworks for retention management presented by Winer (2001) and Ascarza et al. (2018). Table 6.1 compares the levers presented in this paper with the frameworks from previous studies.

**Table 6.1 Comparison between the levers presented in this paper and the framework presented in the theory.**

<i>Framework in literature</i>	<i>Corresponding Lever in this thesis</i>	
<i>Figure 2.1 (Winer, 2001)</i>	Customer Service	Product Improvement
	Frequency/Loyalty Programs	Lock-in, <i>Loyalty Programs</i>
	Customization	Product Improvement, <i>Personalization</i>
	Rewards Program	Lock-in, <i>Loyalty Programs</i>
	Community Building	Lock-in, <i>Emotional Attachment</i>
<i>Figure 2.2 (Ascarza et al., 2018)</i>	Who is at Risk?	Understanding Churn
	Why at Risk?	Understanding Churn
	Who do we target?	Targeting at-risk Churners
	When do we target?	Targeting at-risk Churners
	With what incentive?	Targeting at-risk Churners
	So What did we gain?	<i>No corresponding lever</i>
<i>Figure 2.3 (Ascarza et al., 2018)</i>	Acquisition	Customer Intake
	Pre-emptive	Product Improvement, <i>Lock-in</i>
	Proactive	Targeting at-risk Churners
	Reactive	Save Desk
	Win-back, Post Win-back	Extension of Save Desk



As seen in the table, our levers cover most of the factors presented in the previously presented frameworks. Some similarities and differences include:

- The framework presented in Winer (2001) dates back 20 years and is less focused on the current state of technology with its possibilities to target customers individually in campaigns and identify them using machine learning. Hence, our framework provides a more updated account on retention management. The factors presented in Winer (2001) are encompassed in our framework, in particular within the levers *Product Improvement* and *Lock-in*, as we consider e.g. *Customer Service* to be a part of the product experience.
- Our thesis presents *Organizational Setup* as a lever to improve retention, which is not presented in any of the previous frameworks. We believe that this lever is especially important for early-stage companies such as Insurtechs, as they tend to have a flexible organizational setup continuously under design, where retention should be kept in mind. This factor was mentioned by Hedvig in our initial interviews when designing the interview study.
- Ascarza et al. (2018) sheds light on the *Pre-emptive* campaigns, describing how a company should work with customers before they are at-risk of churning, an effort that is more general than the corresponding levers in our framework. On the other hand, *Product Improvement* and *Lock-in* presented in this thesis are more practical approaches of achieving this goal.
- Ascarza et al. (2018) highlights *Win-back* and *Post Win-back* as phases in retention management, which are only briefly covered in this thesis as an extension to the *Save Desk*. Robert at Top 4 Swedish insurance company B was the only person that we interviewed that mentioned such operations, mentioning that this could be relevant particularly for the insurance industry, as regulations allow companies to store customer data for 10 years after the customer has terminated the contract.
- To “let the sleeping dogs lie” is something Ascarza et al. (2018) emphasizes, and something that was mentioned often in our interviews discussed as part of the lever *Targeting at-risk Churners* in our framework.
- Measuring the effect of a particular campaign on retention is discussed in Ascarza et al. (2018) under the question of “So What did we gain?” is outside of the scope of this thesis and hence there is no corresponding lever.
- The need to integrate different campaigns and having a collected effort in increasing retention is emphasized in Ascarza et al. (2018), however this was not emphasized by the respondents in our interview study.

In summary, the seven levers presented in this paper have many similarities to those presented in Winer (2001) Ascarza et al. (2018). The differences are mostly regarding the emphasis on different factors, where this thesis is more practical in its nature and more focused on Insurtechs. In addition to presenting a framework, this thesis presents best practices and practical recommendations to an Insurtech to increase retention, as well as an updated account of trends in retention and impact of retention work at an Insurtech. Compared to Winer (2001), the framework presented in this paper is more focused on the options that companies have with today's technology.

Finally, the methodology of this paper can be compared to the framework shown in figure 2.2 (Ascarza et al., 2018). The *Identify* section of our method relates to the "at risk" components of the framework by Ascarza et al. (2018), and the *Action* section relates to the "target" and "incentive" components.

## 6.5 Recommendations to Hedvig

As some of the companies interviewed have inherently different churn dynamics, it is evident that only a subset of the identified actions is relevant to be employed at Hedvig. In addition, the relevance for Hedvig is highly dependent on what previous efforts have already been attempted at the company. Therefore, based on interviews with decision makers, our understanding of Hedvig's current situation, and discussions with Hedvig, we have identified our view of what actions to increase retention are currently the most promising for Hedvig. We have organized these recommendations in the framework of the levers that we have identified in this project.

### 6.5.1 Understanding Churn

As previously explained, understanding churn is an important first step in order to improve retention, and this has been the main focus of this thesis. In the context of understanding the reasons why customers churn, we believe that Hedvig is already doing a good job of asking customers why they choose to do so in customer service interactions, and by conducting churn surveys. Extending on this, Hedvig could improve their understanding of which customers churn, and identify at-risk churners through data-driven analysis, as attempted in this project. As explained in 6.2 *Model Performance*, accurately predicting churn at Hedvig will most likely be difficult because of the nature of the data available to a home insurance provider. However, we believe two avenues are worth exploring that may lead to higher predictive performance.

#### 6.5.1.1 *Trigger Features*

The first recommendation is to develop more *trigger* features in the data set that correlate with a user being on their way to terminate their insurance. A concrete suggestion that we have thought of is asking customers who are renters how long they expect to live in that home, as renters often have a lease with a fixed contract length. The time until that contract expires could then be used in the model as an indication of the customer being on their way to move and hence being at-risk of switching insurance providers. Knowing this date in advance can also be used in targeted campaigns, communicating with these customers ahead of time in order to be top-of-mind when they move, or providing an offer ahead of time. This method could be applied to students as well, which correspond to a large share of Hedvig's customer base, by asking them when they expect to graduate, after which many customers are expected to move and potentially switch insurance. Indications of people moving may also be acquired through third parties, such as *Adressändring*, which the Top Swedish electricity provider we interviewed mentioned taking advantage of.

In addition to working to identify signals of customers moving, the behaviors of customers on their way to terminate their insurance should be further studied. This could include e.g. seeing if there is any correlation between a customer checking their price or coverage and then subsequently churning, as the customer may check these things to compare their current insurance with a competitor offer. This approach was very successful at the Loan fintech we interviewed, which found a clear correlation between checking the loan balance and churning. In order to create stronger signals of churning, the price or coverage details could be placed "deeper" in the app or website page hierarchy, such that those who visit those pages are truly checking the details of their current insurance and not just exploring the app. In addition, this analysis should include identifying customers who have requested a new price at e.g. the Hedvig website, as this approach has been shown to be successful in identifying at-risk churners at a Swiss insurance company (Mau et al., 2017).

Finally, more feedback could be attempted to be collected from users in the interactions that they have in Hedvig. As discussed, a customer's rating on a claim may be a strong predictor of churn, but this rating is only received from a fraction of the customers who have a claim. In order to increase the response rate of the rating, it could be more integrated in the app and claims process rather than through a follow up email received after the claim has been processed. In addition to claims, ratings could be requested in connection with other interactions, such as rating the experience the customer had when moving or asking questions about coverage. More general feedback could also be requested e.g. by asking NPS questions outside of interaction occasions, e.g. at random times or when the user opens the app.

### 6.5.1.2 *Segment Prediction Models*

Another approach that could be taken by Hedvig in order to attempt to accurately predict churn is to develop predictive models for specific segments rather than the whole customer base. These subsets of the customer base corresponding to the segments may display more homogenous behavior when it comes to churning. One such segment could be for example students, where a strong predictor may be approaching graduation which can be found by the model even though this feature may not be strong enough for the general customer base to be included in a regularized model. Another potential segment would be customers who recently had a claim, e.g. within 3 months from the prediction date. In this segment, the claims rating, the time the claims process took, or claims payout may be significant predictors of churn. Using this segmented approach, more specific types of customers can be targeted which display less “random” or external behaviors when it comes to churn. Another advantage with the approach is that the message and targeting approach can be differentiated depending on the customer segment. For students for example, Hedvig may want to define an automatic campaign and send push notifications congratulating them on completing their first year of studies and reminding them that Hedvig is easy to move to their new home when it comes to moving out of student housing after graduation. For customers who were dissatisfied with their claim on the other hand, a more direct and personal approach of calling the customers and discussing their claim may be more appropriate. The idea for this segmented approach stems in the interviews with analytics experts at Top 4 Swedish telecom A, Major Swedish media company B and Audio streaming service B, all companies leveraging this approach. A risk with applying this approach at Hedvig is that the customer base may not be big enough yet for each segment to be big enough to successfully train models on.

As a final note regarding attempting to model and predict churn at Hedvig, achieving an accurate model will be hard for an Insurtech at Hedvig’s stage and customer relations, as previously discussed. This is evident as even some of the top telcos and electricity companies in Sweden have fail to do so accurately, as reported in our interviews. In addition, the model must be very accurate in order to be used in practice, in order to “let the sleeping dogs lie”, which indicates that predicting churn accurately enough to be used in practice at Hedvig will be challenging in the near future.

### **6.5.2 Customer Intake**

When dealing with customer churn, it is of high importance to analyze the customer intake, to make sure that one does not spend excess resources in the form of marketing, discounts, and onboarding on customers that are likely to leave Hedvig soon anyways. Like many insurance companies, Hedvig uses discounts as part of their campaigns to acquire new customers. Additionally, they give customers a

discount for each additional customer they refer, for as long as they remain at Hedvig. Discussions with industry professionals raised the question whether this is indeed effective or if these customers are more prone to churn sooner. However, as discussed in section 6.3. *Churn Drivers*, discount expiration is not a significant driver of churn, and referrals on the other hand seem to be a strong driver of retention. Therefore, our conclusion here is that both their discount structure and referral program seem to have desirable outcomes and could be continued further.

Furthermore, there is a pattern between the sign-up month and the likelihood of churning, where customers in July seem to be more loyal and are included in several of our model understanding analyzes as a significant feature of the models. When testing for statistical significance, the  $\chi^2$ -test p-value is 6.2%, and hence significant on a 90% confidence level. However, only based on this, it is hard to draw any clear conclusions, but it is something to keep in mind when tracking this further.

Additionally, homeowners have higher retention than rentals, and could therefore be prioritized in campaigns. If Hedvig was to successfully develop an accurate churn or CLV model in the future, it could be leveraged to prioritize customer intake when advertising digitally.

### **6.5.3 Product Improvement**

Improving the product and service is something that Hedvig already works with extensively, and hence there is less to say on this lever. One thing that could be kept in mind is to make sure that all the input from the retention work, customer service, and potentially save desk is collected and made use of in the product development to increase retention.

### **6.5.4 Lock-in**

Hedvig is already working towards offering more types of insurance such as pet or car insurance. One way to accelerate the process of having multiple products as a lock-in, is launching extension products on their current home insurance. This could be smaller insurance products or add-ons, such as extended travel insurance or extended coverage of e.g. sports equipment or bicycles. On the other hand, some customers may leave due to the price, and here Hedvig could potentially offer a cheaper, lower tier product where the deductible is higher, or some of the coverage that is not relevant to the customer is removed. This down-grading adjustment to the insurance would not necessarily have to be available on their website, but possible rather a way for the save desk to win back customers that consider leaving Hedvig.

Another way to create a lock-in effect could be to connect their customers with a physical device in their home. This could for example be a connected IoT water leakage detector or similar, in a partnership with a hardware provider. This could both generate a lock-in effect for customers, be used for marketing purposes, as well as decrease the likelihood and frequency of water damage related claims.

### **6.5.5 Targeting at-risk Churners**

Assuming that Hedvig manages to develop a model or method to accurately identify customers at-risk of churning, the next step is to target these customers with specific actions in order to affect their decision and hopefully make them stay. The most straightforward way to incentivize customers to stay is to give them a discount. However, most interview respondents agree that giving customers a discount could be used to acquire new customers but is often a poor way to induce loyalty once a customer has considered leaving. Even if one were to consider a discount as a measure, this action is not straightforward. As discussed in section 5.2.2 *Levers to Increase Retentions* and Appendix E, giving customers a discount, or price decrease, can have the reverse effect, according to an insurance pricing expert. Based on his experience and studies carried out at Top 4 Swedish insurance company A, giving customers a price decrease resulted in a slightly higher churn rate. However, as Hedvig has a business model where 80% of the fees are to be paid out to customers with claims, Swedish insurance pricing specialist suggests that Hedvig could target customers at-risk with a small and very well-motivated price decrease. This could for example be to say that there is an excess left from the fees in your specific customer group or area, you have not had any claim and we believe in you as a customer, and Hedvig therefore wants to give you a lower fee in the next period. In order for this to be reliable, he says that this excess should be as specific as possible, e.g. “There is an excess, and therefore you will enjoy a 2.24% discount on your fee in the next period”.

Generally, the companies that are most successful seem to have the churn analysis drivers, triggers, or predictions feeding into a well-designed marketing automation system, automatically targeting the identified at-risk customers with well-timed and designed messages. However, in order for this to be successful, one has to be rather certain about what customers in fact are at a true risk of churning, and hence need to make accurate predictions about it in the first place. Otherwise, one risks waking up customers that never considered leaving before receiving a message, but then start to reevaluate their decision. It is therefore also important to A/B-test the marketing campaigns before implementing them into customers' automated communication flows in the marketing automation system.

Some of the churn reasons may in fact not be a sign of disaffection. As Hedvig's customer base is relatively young, customers tend to move around more, and hence, there are some natural temporary cancellation reasons, especially for students. This

could be that a student is moving abroad on an exchange semester, moving in with their parents over the summer, or similar, and actually planning to activate their Hedvig insurance again after this break. If the customer then in fact returns, is it then really a churning customer, or has the customer rather become even more loyal to Hedvig's flexible non-binding insurance offering? Making sure to capture these customers could be important, and clear communication about pausing alternatives, a smooth return or reactivation could be key.

Finally, we discussed that if a customer visits certain pages on the website or in the app, it could be a sign that the customer considers leaving. Therefore, after identifying these triggers, a communication channel could be to make the chat appear directly to customers entering these pages, making it easy for customers to get help or to feel that they are being listened to.

#### **6.5.6 Save Desk**

A save desk can be a very impactful measure to reduce churn at the last minute. After talking to a large number of subscription companies, many have seen very successful results after implementing this. This is also true also for other insurance companies as discussed in 5.2.2 *Levers to Increase Retentions* and Appendix E. However, contacting them in the current communication channel, the Hedvig app, has clear limitations, and should rather be done over the phone. Due to the large success at other companies, it could be worth it for Hedvig to experiment with a save desk, at first on a smaller scale. In order to implement it successfully, it is important to have skilled employees dealing with these calls, and also to let them be part of setting up the program, in order for them to be motivated. Additionally, they should be incentivized to keep the customers with as little resources as possible, but still allowing some flexibility in tailoring the message to the specific customer at hand. A possible extension to a save desk is a "win-back desk", which would contact previous customers at Hedvig that have switched to a competitor, attempting to convince them to come back.

#### **6.5.7 Organizational Setup**

According to our interviews, who or which department is responsible for and has ownership of the retention metric and retention work differs between companies, and there is no clear organizational setup that is best suited for this. A best practice that we did find, however, is breaking down retention into sub-metrics that can more easily be linked to the efforts of a particular team or department, as reported by Marcus at the Home alarm startup that we interviewed. In Hedvig's case, this could be implemented as such: One person has ownership over the overall retention metric and leads the work in improving it.

Sub-metrics that can be tracked could for example be:

- Churn rate in the first 6 months, attributed to the growth team.
- Churn rate after a claim or customer service interaction, attributed to the claims and customer service team.
- Churn rate for customers who have been with Hedvig for e.g. 2 years or more, attributed to the CRM team.
- Churn rate for customers who are acquired from certain partnerships, attributed to the partnership team.

Additionally, the product team should be involved and have ownership of retention. As previously mentioned, when working to improve retention, it is important to collect feedback from customer touch points across the organization and customer journey. This information should then be studied together to get a holistic view of churn, and measures can then be decided and implemented in the relevant parts of the organization, e.g. the product or growth team.

## 6.6 Transferability

In this project, we have mainly focused on Hedvig as a case study, however the insights can in large be applied to other Insurtechs. As Hedvig currently only offers home insurance, we have focused on this insurance product and the dynamics and customer behaviors connected to a home such as moving, rather than e.g. discussing how customers switch cars for car insurance. Hence, some findings mainly apply to Insurtechs that offer home insurance as their only or main product. However, many of the findings made in this paper are general and can be applied to other products as well, e.g. working systematically with identifying and targeting at-risk churners. In some cases, the methodology presented in this paper may lead to better results, and a better prediction model, than we managed to develop for Hedvig. The Car Insurtech that we interviewed, for example, may have access to more data from the customer's usage, driving habits, location etc., which can serve as predictors for churn from the sensors that customers install in their cars. The methodology presented in this paper can be used for churn prediction and extracting churn drivers at any Insurtech.

## 6.7 Further Research

To further investigate the potential in identifying at-risk churners at Insurtechs, there are a number of further analyzes that could be carried out, some of these mentioned



in section 1.2.4. *Delimitations*. First of all, similar analyzes could be performed at other Insurtechs, e.g. with other coverage or geographies, to get a better understanding of the full Insurtech landscape. Additionally, one could attempt other types of models that are out of scope for this thesis, such as survival models, other time series approaches, a regression approach predicting remaining lifetime rather than churn in a given period, or prescriptive models such as uplift models. Another approach could also be modeling the entire customer lifetime value (CLV) which encompasses not only churn but also the future behavior of a customer in terms of buying more products or moving to homes with a different premium.

Another extension of our research could be to extract more advanced features from the Hedvig data set. This includes for example not only the number and frequency of chat messages, but further analysis of the contents and the sentiment of these interactions. Further, an analysis could be made on the referral network of a person, i.e. who they referred and who referred that person etc. From this analysis, one could investigate for example whether the event of a customer that you have referred churns, or the customer that referred you churns has an effect on your decision to stay. Another data source that could be leveraged is measuring customer engagement e.g. their opening frequency on membership emails, or whether or not they follow Hedvig on social media. A third extension could be to include external data sources. As moving is a frequent reason for churn, it could be interesting to include data sources such as students' expected graduation from the student card *Studentkortet*, or other external sources that could indicate that a customer is about to move. These data extensions are in line with the type of data that Ascarza et al. (2018) identifies as having the potential to increase prediction accuracy.

Finally, some of the approaches that were taken in this problem to solve certain sub-problems could be extended and improved. This includes for example more advanced balancing and sampling techniques, e.g. extensions of SMOTE (Fernández et al., 2018, p. 101-113), and more advanced hyperparameter search, e.g. Bayesian hyperparameter optimization (Ma et al., 2015). Finally, a more advanced sampling and validation approach could be employed, including churn examples from different months in a year rather than a single time period, and validation could be performed not only out-of-sample but also out-of-time. Finally, the prediction horizon could be experimented with, e.g. training different models for customers churning in 1, 3 or 6 months and comparing them.

## 6.8 Ethics

When handling customers' personal and usage data and making predictions on customers' future behavior, it is important to consider the ethical aspects regarding what data is collected, how the data is stored and used, and for what purpose it will

be used in the end. The customer data used in this thesis was stored and handled in a way that only the two authors of this thesis could access it. Additionally, the data was only handled throughout the extent of this project, and all data and models were deleted, after sending the relevant content to Hedvig for further development. Prior to receiving the data at the start of the project, Hedvig made sure to remove personal information on its customers that were not going to be used in this thesis, such as names and personal numbers. Hence, the data was at all times anonymized and there was no way for the authors of the thesis to access what information belongs to what customer.

When making predictions with customer data, and particularly demographical data, it is important to keep in mind what the end use of the model is and how certain variables are used in the predictions. Models that use this type of data to make generalizations about people's behavior risk including biases against certain demographic groups that could be problematic. Adding certain predictors could increase the performance, while it also be biased against certain groups of individuals and reinforce prior prejudices and lead to discriminatory classifications. For example, if the aim of the model was to model the fraud risk or to estimate customers' individual claims frequency or future payout amounts, e.g. to set an individual insurance price tag, one should avoid including general variables such as gender, nationality, or race, as, even if these could improve the performance of the prediction, it would lead to questionable and potentially illegal inferences. No such data is included in this project. Additionally, as the purpose of this thesis is to identify at-risk churners in order for Hedvig to be able to provide a better customer experience for these types of customers, the inference of this model is in its nature less sensible for potential biases than other use cases as those mentioned above.

# References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.
- Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), 1059-1086.
- Ascarza, E. (2018). Retention futility: *Targeting high-risk customers might be ineffective*. *Journal of Marketing Research*, 55(1), 80-98.
- Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., et al., (2018). In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1), 65-81.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Bolancé, C., Guillen, M., & Padilla-Barreto, AE. (2016). Predicting detection in non-life motor and home insurance. *Lectures on Modeling and Simulation*, 2, 107-120.
- Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, 17(1), 45-65.
- Boucher, J. P., & Couture-Piché, G. (2015). Modeling the number of insureds' cars using queuing theory. *Insurance: Mathematics and Economics*, 64, 67-76.
- Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., & Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75(3), 713-737.
- Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70, 245-317.
- Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2), 277-288.

- Cappiello, A. (2020). The Digital (R) evolution of Insurance Business Models. *American Journal of Economics and Business Administration*, 12(1), 1-13.
- 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.
- 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.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.
- Cooley, S. (2002) Loyalty strategy development using applied member-cohort segmentation. *Journal of Consumer Marketing*, 19(7), 550–563.
- Crosby, L. A., & Stephens, N. (1987). Effects of relationship marketing on satisfaction, retention, and prices in the life insurance industry. *Journal of marketing research*, 24(4), 404-411.
- Dagens industri. (2020). *ICA Försäkring utmanar storbolagen med helt ny typ av jämförelsetjänst*. [ICA Insurance challenges the industry with a new type of comparison service]. Retrieved February 4, 2021, from <https://www.di.se/brandstudio/ica-forsakring/ica-forsakring-utmanar-storbolagen-med-helt-ny-typ-av-jamforelsetjanst/>
- Daly, J. L. (2002). *Pricing for profitability: Activity-based pricing for competitive advantage* (Vol. 11). John Wiley & Sons.
- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, 233-240.
- De la Llave, M. Á., López, F. A., & Angulo, A. (2019). The impact of geographical factors on churn prediction: an application to an insurance company in Madrid's urban area. *Scandinavian Actuarial Journal*, 2019(3), 188-203.
- Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, 548, 497-515.
- Dickinson, B. (2015). *Insurance Is The Next Frontier for Fintech*. In *TechCrunch*. Retrieved 26 January, 2021, from <https://techcrunch.com/2015/08/05/insurance-is-the-next-frontier-for-fintech/>
- Dinis, C. G. J. (2017). Churn Analysis in a Music Streaming Service: Predicting and understanding retention.

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 11). Berlin: Springer.
- Dexe, J., Franke, U., Nöu, A. A., & Rad, A. (2020). Towards increased transparency with value sensitive design. *International Conference on Human-Computer Interaction*, 3-15. Springer, Cham.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44-65.
- Gramegna, A., & Giudici, P. (2020). Why to Buy Insurance? An Explainable Artificial Intelligence Approach. *Risks*, 8(4), 137.
- Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. *arXiv preprint arXiv:1802.03396*.
- Greis, N. P., & Gilstein, C. Z. (1991). Empirical Bayes methods for telecommunications forecasting. In *International Journal of Forecasting*, 7(2), 183-197.
- Grize, Y. L., Fischer, W., & Lützelshwab, C. (2020). Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry*, 36(4), 523-537.
- Günther, C. C., Tvette, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1), 58-71.
- Heno Madrigal, M. (2020). *Customer churn prediction in insurance industries: a multiproduct approach* (Doctoral dissertation, Universidad EAFIT).
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Hooker, G., & Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*.
- Hu, X., Yang, Y., Chen, L., & Zhu, S. (2020, April). Research on a customer churn combination prediction model based on decision tree and neural network. *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics*, 129-132. IEEE.

- 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.
- Ica Försäkring (2021). *Valet var givet – ICA Försäkring hade det bästa för oss*. [The choice was obvious – ICA insurance was best for us]. Retrieved 4 February, 2021, from <https://icaforsakring.realcontent.se/valet-var-givet-ica-forsakring-hade-det-basta-for-oss/>
- Insurance Sweden. (2019). *Insurance in Sweden 2019*. Retrieved 26 January, 2021, from <https://www.svenskforsakring.se/contentassets/368566060a79471bb2ec76bfe9bec920/insurance-in-sweden-2010-2019.pdf>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer.
- Kumar, V., & Reinartz, W., (2006). *Customer Relationship Management*, 179-206. 3rd ed. Berlin Heidelberg: Springer
- Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27(2), 277-285.
- Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2), 276-286.
- Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science*, 39(5), 956-973.
- Lindmark, M., Andersson, L. F., & Adams, M. (2006). The evolution and development of the Swedish insurance market. *Accounting, Business & Financial History*, 16(3), 341-370.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.
- Ma, S., Tan, H., & Shu, F. (2015). When is the best time to reactivate your inactive customers? *Marketing Letters*, 26(1), 81-98.
- Mau, S., Pletikosa, I., & Wagner, J. (2018). Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments. *International Journal of Bank Marketing*.

- Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, 83, 326-332.
- Molnar, C. (2021). *Interpretable Machine Learning*. Retrieved 6 March, 2021, from <https://christophm.github.io/interpretable-ml-book/>
- Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning--A Brief History, State-of-the-Art and Challenges. *arXiv preprint arXiv:2010.09337*.
- Morik, K., & Köpcke, H. (2004, September). Analysing customer churn in insurance data—a case study. *European conference on principles of data mining and knowledge discovery*, 325-336. Springer, Berlin, Heidelberg.
- 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.
- Puertas, A., O'Driscoll, C. Krusberg, M., Gromek, M., Popovics, P., Teigland, R., Siri, S. & Sundberg, T. (2017). *In The Next Wave of FinTech - Redefining Financial Services Through Technology*. Retrieved 27 January, 2021, from <https://www.hhs.se/contentassets/615a9c5cac064280877d07799d70e0d2/insurtechhregtechreportsse1.01.pdf>
- Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard business review*, 68(5), 105-111.
- Risselada, H., Verhoef, P. C. & Bijmolt, T. H. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, 24(3), 198–208.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
- Scriney, M., Nie, D., & Roantree, M. (2020, September). Predicting customer churn for insurance data. *International Conference on Big Data Analytics and Knowledge Discovery*, 256-265. Springer, Cham.
- SFS (2005). *Försäkringsavtalslag*. [Insurance contract law] Retrieved 26 January, 2021, from [https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forsakringsavtalslag-2005104\\_sfs-2005-104](https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forsakringsavtalslag-2005104_sfs-2005-104)
- Statista (2018). *Customer retention rate of businesses worldwide in 2018, by industry*. Retrieved 22 February, 2021, from <https://www.statista.com/statistics/1041645/customer-retention-rates-by-industry-worldwide/>

- Svensk Försäkring (2021). *Försäkringens historia*. [The history of insurance]. Retrieved 18 February, 2021, from <https://www.svenskforsakring.se/om-forsakring/forsakringens-historia/>
- 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.
- Verhoef, P. C., Franses, P. H., & Hoekstra, J. C. (2002). The effect of relational constructs on customer referrals and number of services purchased from a multiservice provider: does age of relationship matter? *Journal of the academy of marketing science*, 30(3), 202-216.
- 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.
- Winer, R. S. (2001). A framework for customer relationship management. *California management review*, 43(4), 89-105.
- Yu, X., Guo, S., Guo, J., & Huang, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, 38(3), 1425-1430.
- Zhang, R., Li, W., Tan, W., & Mo, T. (2017). Deep and shallow model for insurance churn prediction service. *2017 IEEE International Conference on Services Computing*, 346-353. IEEE.
- Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005, July). Customer churn prediction using improved one-class support vector machine. *International Conference on Advanced Data Mining and Applications*, 300-306. Springer, Berlin, Heidelberg.



# Appendix A Participating Respondents in Interview Study

*Note that the companies, positions, and names have been anonymized in order to keep the interview respondents and companies confidential.*

<i>Company</i>	<i>Position</i>	<i>Respondent</i>
Hedvig	Head of Insurance Operations	Kajsa
Pet Insurtech	Chief Operating Officer	COO
Car Insurtech	Head of Product	<i>Anonymized</i>
Content and car Insurtech	Chief Operating Officer	<i>Anonymized</i>
Top 4 Swedish insurance company A	Pricing expert	<i>Anonymized</i>
Top 4 Swedish insurance company B	Executive	Robert
Loyalty Factory	Retention expert	Camilla
Top Swedish electricity provider	Marketing Manager	Cecilia
Electricity provider startup A	Product Manager	PM
Electricity provider startup B	Founder	Founder
Low-price telecom	Product Manager	Oscar
Top 4 Swedish telecom A	Analytics expert	<i>Anonymized</i>
Top 4 Swedish telecom A	Customer loyalty executive	Joel
Top 4 Swedish telecom B	Customer loyalty executive	Christian
International telecom	Business Analyst	Svante
Major Swedish media company A	Customer loyalty executive	Hugo

---

Major Swedish media company B	Head of Streaming Service	Johanna
Major Swedish media company B	Analytics Expert	Magnus
Audio streaming service A	Data Analyst	John
Audio streaming service B	Customer loyalty executive	Malin
Audio streaming service C	Head of Product	<i>Anonymized</i>
Audio streaming service C	Data Analyst	Filip
Loan fintech	Customer loyalty executive	Kasper
Payment fintech company	Data Analyst	Kajsa
Home alarm startup	Co-Founder	Marcus
Car sharing service	Customer loyalty executive	Hanna
Swedish pharmacy	Customer loyalty executive	Erica
Recycling startup	Head of Sales	Martin

---

---

# Appendix B Interview Guide

*The interview questions were at times altered or branched out, depending on the interview situation and the interviewee's background and knowledge.*

## **How do you work with retention at your company?**

1. How prioritized is retention in your company?
  - a. Has it always been that way / do you think it will be in the future?
2. Who works with retention? How is the organization structured around retention? One who leads the work or a multifunctional project group?
3. How do you measure retention and churn?
4. What processes are there and how are they followed up? Which KPIs? Who is responsible or in charge?

## **What initiatives have you taken to increase retention?**

1. How did you come to choose these?
2. How did you identify which customers you would focus on?
3. What have they led to in terms of results? What was the return on investment?
4. When or in what phase of company maturity did you implement it?

**Have you worked in a structured way to identify potential churners and make targeted initiatives to increase retention?**

1. What do you think are the major driving factors for churn at your company? How did you identify these?
2. What tools have you used to identify customer groups or factors that drive churn?
3. Have you used machine learning or other analytical methods to predict which customers are likely to churn?

*If yes:*

- a. What methods have you used? With what type of data?
- b. How accurate are the models?
- c. What have you done with the output of the models?

*If no:*

- d. Is this anything you plan to use?
- e. Why have you not used it? Does it sound like something that would be interesting to implement in the future?
- f. Would you say that you have structures / working methods in place that enable the collection of relevant data for this type of analysis?
- g. Do you use analytical methods / machine learning for other tasks in your organization? How do you use it?

**Do you have experience from retention work from other companies? What does it look like in the industry?**

1. How does it vary between companies? What do you think is it that determines how much focus there is on retention in a company and how to work with it?
2. How has the view of retention and churn developed in recent years?
3. How do you think the view on retention will change in the coming years?

# Appendix C Hyperparameter Optimization

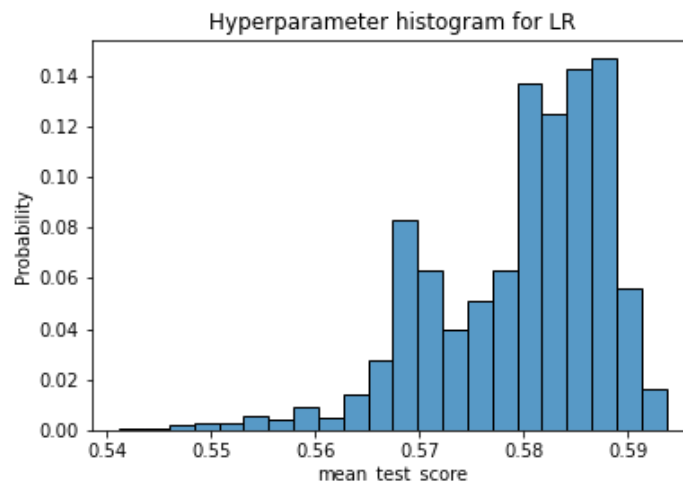
## C.1 Logistic Regression

The hyperparameters, and the tried values for each of them, in addition the processing parameters described in table 4.1, are shown in table C.1 below:

**Table C.1 Candidate hyperparameters for LR.**

Hyperparameter	Trial Values
Regularization type	$L^1$ or $L^2$
Regularization coefficient $C$	100, 10, 1.0, 0.4, 0.3, 0.2, 0.15, 0.1, 0.05, 0.01

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.1. In the plot, it can be seen that AU-ROC varies from 0.55 to 0.60 in the candidate population.



**Figure C.1 Distribution of the mean AU-ROC score for the different candidates for LR.**

Further inspecting the parameters, the model benefits from transforming and scaling, and performs the best on the manually selected feature set. Regarding regularization, Ridge regression has higher performance than the LASSO, with the best performing value for  $C$  given this processing being 0.4. The best performing hyperparameters for logistic regression are summarized in table C.2.

**Table C.2: Best performing hyperparameters for LR.**

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	Transforming and scaling
Feature	The manual subset of features
Balancing	Model specific
Regularization type	$L^2$
Regularization coefficient $C$	0.4

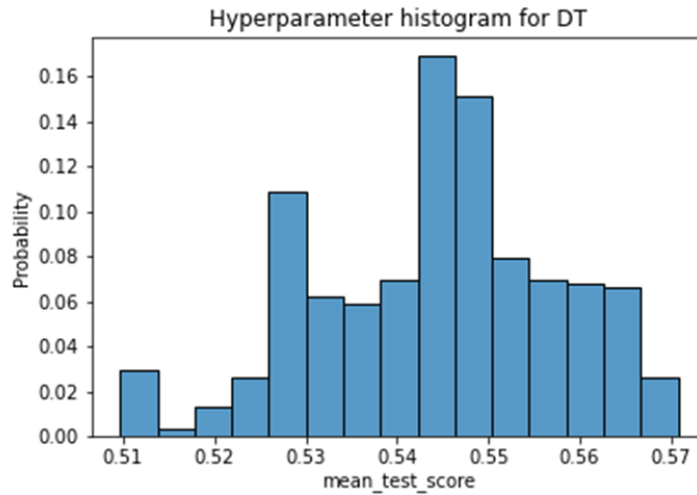
## C.2 Decision Tree

For the decision tree, the only model hyperparameter that is varied is the max depth, as shown in Table C.3 below:

**Table C.3 Candidate hyperparameters for DT.**

<b>Hyperparameter</b>	<b><i>Trial Values</i></b>
Max depth	2, 3, 4, 5, 6

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.2. In the plot, it can be seen that AU-ROC varies from 0.50 to 0.57 in the candidate population.



**Figure C.2** Distribution of the mean AU-ROC score for the different candidates for DT.

Further inspecting the parameters, the model performs the best with no transformation, SMOTE balancing, and no feature selection. The performance increases with the max depth in the trial range, probably because the model gets more flexible and can capture more of the churn dynamics, and that none of the trial values were big enough to lead to overfitting. The best performing hyperparameters for decision trees are summarized in table C.4.

**Table C.4** Best performing hyperparameters for DT.

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	No preprocessing
Feature	All features
Balancing	No balancing
Max depth	6

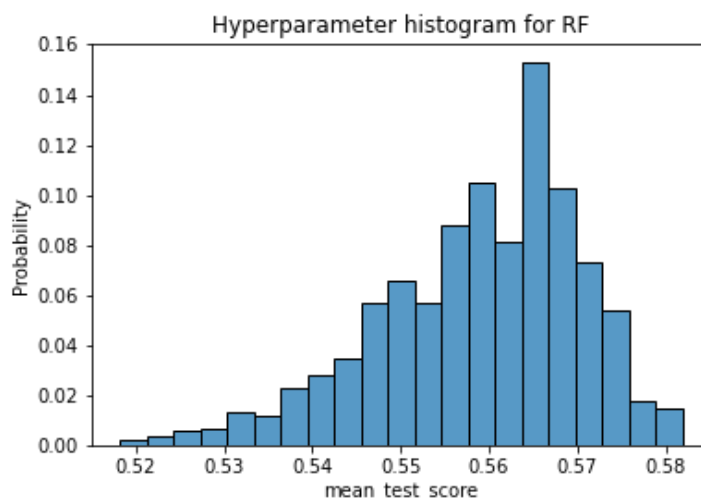
## C.3 Random Forest

The model specific hyperparameters for RF are shown in table C.5 below and explained in depth in Kuhn and Johnson (2013, p. 198-203).

**Table C.5 Candidate hyperparameters for RF.**

Hyperparameter	Trial Values
Number of estimators	10, 100, 200, 500, 1000, 1500, 2000
Max depth	No max, 2, 4, 6, 8, 10, 20, 30, 40, 50, 60
Min Samples Split	2, 5, 10, 20
Min Samples Leaf	1, 2, 4, 10

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.3. In the plot, it can be seen that AU-ROC varies from 0.52 to 0.58 in the candidate population.



**Figure C.3 Distribution of the mean AU-ROC score for the different candidates for RF.**

Further inspecting the parameters, the model does not benefit from transforming or scaling, which is consistent with the theory for tree-based methods. Feature selection does not have a big impact on performance, which is intuitive as the model has “built-in” feature selection when selecting features for splits in the algorithm, however slightly better performance is reached with the manually selected subset of features. Finally, the model performs best with SMOTE balancing. The best performing hyperparameters for random forest are summarized in table C.6.



**Table C.6 Best performing hyperparameters for RF.**

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	No preprocessing
Feature	The manual subset of features
Balancing	SMOTE
Number of estimators	500
Max depth	6
Min Samples Split	10
Min Samples Leaf	4

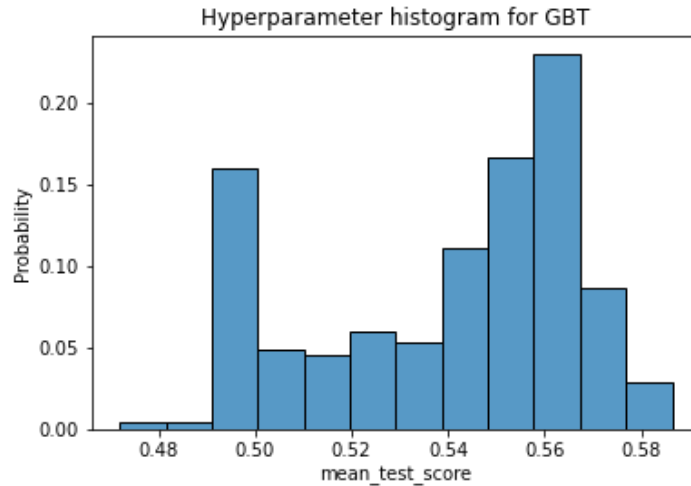
## C.4 Gradient Boosted Trees

The model specific hyperparameters for the XGBoost implementation used for GBT are shown in table C.7 below, and explained in depth in Chen and Guestrin (2016):

**Table C.7 Candidate hyperparameters for GBT.**

<b>Hyperparameter</b>	<b><i>Trial Values</i></b>
Number of estimators	100, 500, 1000
Max depth	3, 6, 8, 10, 12, 15, 20
Learning rate	0.001, 0.01, 0.1, 0.2, 0.3, 0.7
$L^1$ regularization term	0.1, 1, 5, 10, 50, 100
Scale weight (balancing)	1, 5, 10, 20, 25

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.4. In the plot, it can be seen that AU-ROC varies from 0.48 to 0.58 in the candidate population.



**Figure C.4: Distribution of the mean AU-ROC score for the different candidates for GBT.**

Further inspecting the parameters, the model does benefit from standard scaling, but not transforming. Further, balancing increases performance, with the higher performance achieved through the built-in scale weight parameter. Looking at feature selection, the model performance seems to perform best on the full feature set. Regarding the model hyperparameters, most of them do not appear to make any significant difference on performance, except the learning rate which has a clear effect. The best performing hyperparameters for gradient boosted trees are summarized in table C.8.

**Table C.8 Best performing hyperparameters for GBT.**

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	No preprocessing
Feature	All features
Balancing	Built-in balancing via the scale weight parameter
Number of estimators	500
Max depth	6
Learning rate	0.01
$L^1$ regularization term	100
Scale weight (balancing)	20

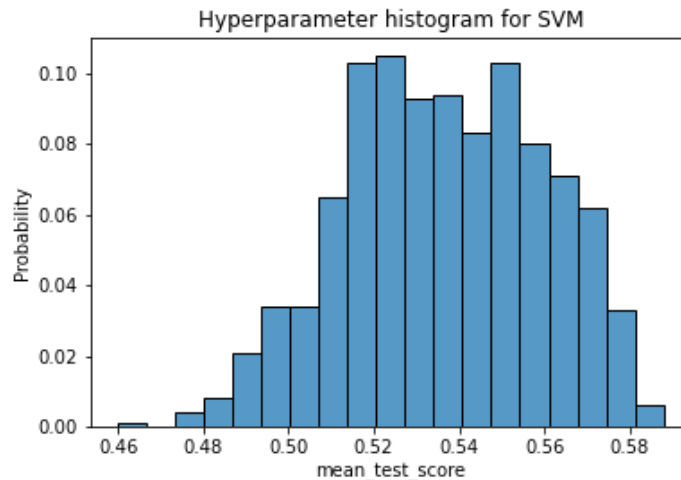
## C.5 Support Vector Machines

The hyperparameters, and the tried values for each of them, in addition the processing parameters described in table C.9, are shown in Table C.9 below:

**Table C.9 Candidate hyperparameters for SVM.**

Hyperparameter	Trial Values
Kernel	Radial Basis Function (RBF) or Polynomial
Degree (only for polynomial)	1, 2, 3
Regularization coefficient $C$	100, 10, 1, 0.4, 0.3, 0.2, 0.15, 0.1, 0.05, 0.01
Kernel coefficient	1, 0.1, 0.01, 0.001, 0.0001

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.5. In the plot, it can be seen that AU-ROC varies from 0.46 to 0.59 in the candidate population.



**Figure C.5 Distribution of the mean AU-ROC score for the different candidates for SVM.**

Further inspecting the parameters, SVM benefits greatly from the preprocessing and transforming as well as balancing. Regarding feature selection, the model seems to favor a limited number of features. The top performing model uses as few as 50% of the already limited subset of features. Further, almost all of the best performing models have a polynomial kernel with degree one. This is the simplest and linear model, analogous to logistic regression. The best performing hyperparameters for SVM are summarized in table C.10.

**Table C.10 Best performing hyperparameters for SVM.**

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	Transforming and scaling
Feature	The top 50% F-scoring features from the manual subset of features
Balancing	SMOTE
Kernel	Polynomial
Degree (only for polynomial)	1
Regularization coefficient $C$	0.3
Kernel coefficient	1.0

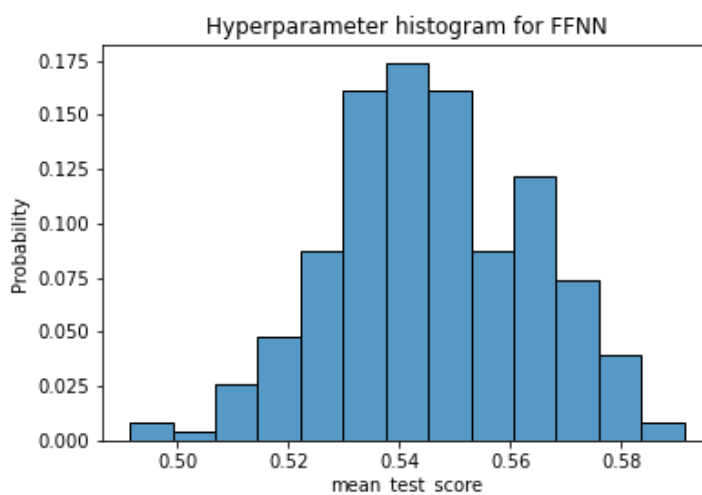
## C.6 Feed-Forward Neural Network

For the FFNN, a key hyperparameter is the network structure itself. This involves the number of layers, as well as their size. There is no one size fits all solution for network structure, or any structure that has been proven especially useful for churn prediction, instead the best structure will vary depending on the dataset. As such, different structures are tested in the random sampling in order to find the best one. Since the number of features in the dataset are relatively low, and because of computing constraints, only shallow networks with 1-3 layers are tested. Other hyperparameters of the model include the regularization parameter, learning rate and activation function, as explained further in Hastie et al. (2009), and Kuhn and Johnson (2013). The trial values for the FFNN hyperparameters are shown in table C.11.

**Table C.11 Candidate hyperparameters for FFNN. The network structure is encoded as a list showing how many layers the network has and how many neurons there are in each layer.**

Hyperparameter	Trial Values
Network structure	[100], [8, 8, 1], [10, 30, 10], [20, 20]
$L^2$ regularization parameter	0.01, 0.001, 0.0001, 0.00001
Learning rate	0.01, 0.001, 0.0001, 0.00005, 0.00001
Activation function	relu, logistic, tanh

The distribution of the mean AU-ROC score in the cross-validation folds for each of the candidates is shown in figure C.6. In the plot, it can be seen that AU-ROC varies from 0.49 to 0.59 in the candidate population.



**Figure C.6: Distribution of the mean AU-ROC score for the different candidates for FFNN.**

FFNN greatly benefits from preprocessing and transforming, however balancing appears to have a low impact on performance. Regarding feature selection, the FFNN model prefers a limited feature set, the best model selecting the top 80% F-scoring features. Regarding network structure, all attempted structures appear to be valid, and have similar performance. However, for this particular dataset, the three layer [10, 30, 10] structure performs slightly better than the others. The learning rate and activation have a significant effect on performance, where a logistic activation function, and the lower learning rates in the trial range, performs the best. The best performing hyperparameters for FFNN are summarized in table C.12.

**Table C.12 Best performing hyperparameters for FFNN.**

<b>Hyperparameter</b>	<b><i>Optimal Value</i></b>
Preprocessing	Transforming and scaling
Feature selection	The top 80% F-scoring features
Balancing	SMOTE
Network structure	10, 30, 10
L2 regularization parameter	0.1
Learning rate	0.0001
Activation function	logistic

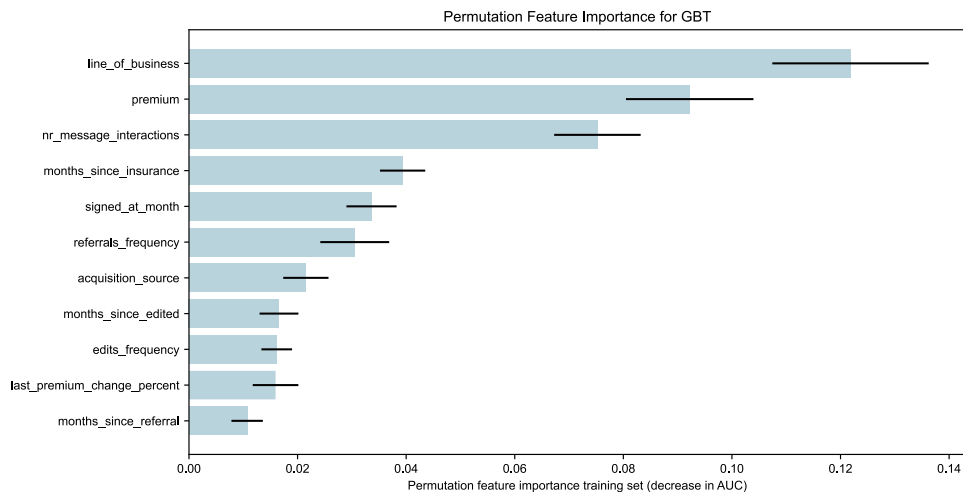
# Appendix D Churn Drivers

## D.1 Model Interpretation on GBT

The interpretation algorithms used in this project assume independence between the predictors in order to explain the model well, as importance is otherwise ‘shared’ between correlated predictors or in some cases calculated as a low value, as one of them could be used in place of the other. As such, the manually selected feature set was used with GBT in interpretation even though the final model uses the full feature set. In order to compensate for the lower number of predictors in the dataset, the regularization coefficient was adjusted, and the performance of this model is very similar to the original.

### D.1.1 PFI

First, in order to understand the importance of the drivers according to the model, the Permutation Feature Importance is computed, shown in figure D.1. In the figure, only the significant predictors, which contribute to at least 1% in AU-ROC are shown, the rest are considered unimportant for the model.



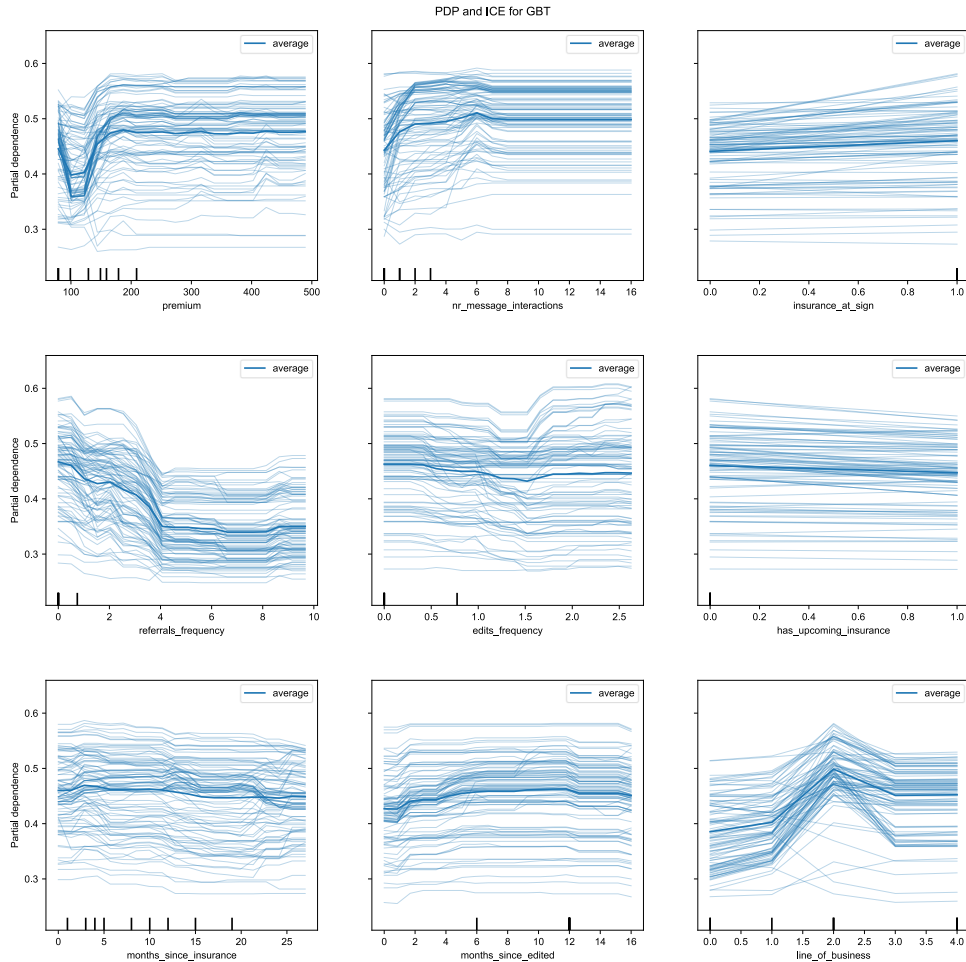
**Figure D.1 PFI for the significant predictors for GBT. The error bars represent a 95% confidence interval.**

According to PFI, the most important predictor is the line of business, corresponding to up to 14% AU-ROC. Next, the premium, the number of message interactions, the time being insured at Hedvig, and the month signed up to Hedvig is used for predictions. Finally, the acquisition source as well as frequency and recency of edits, and referrals are considered by the model, however these correspond to less than 2.5% AU-ROC.

### **D.1.2 ICE and PDP**

As the PFI only ranks the features of importance for predictions, based on random permutation, ICE and PDP can be used to further understand how the drivers affect churn. Using ICE and PDP, one can understand whether the feature leads to lower or higher churn predictions, for what values, if the relationship is monotonic or not, and if it affects all customers in the same way or not. In figure D.2, ICE plots are shown, together with the PDP calculated as the average ICE over all customers. The plot shows the subset of all features that have any effect on churn, since most of the features have a flat plot, with a constant conditional churn rate.





**Figure D.2 ICE plots for the significant predictors for GBT with PDP highlighted. The black bars in the bottom of the plots show deciles of the feature distribution.**

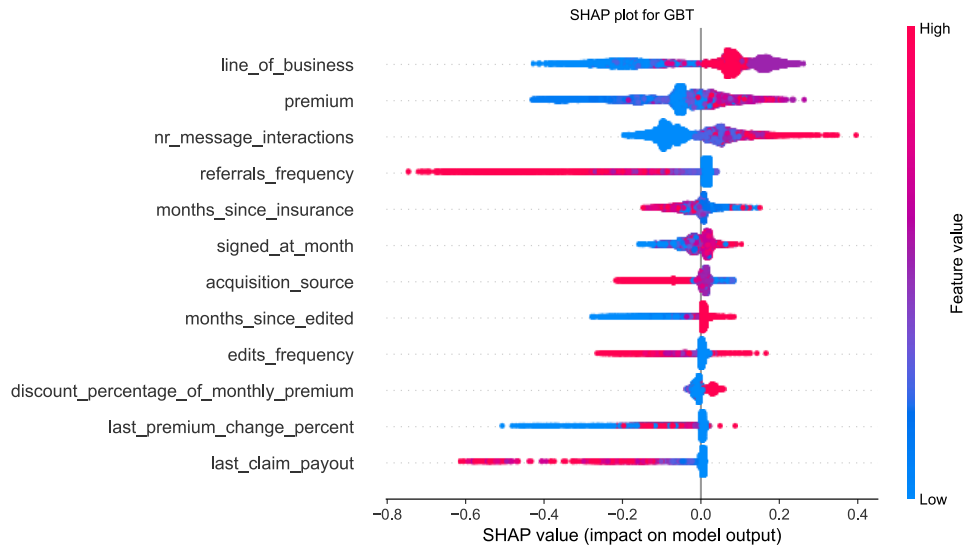
Some interesting things to note in the plots are:

- A premium around 100 is associated with the lowest churn risk. A higher or lower premium appears to have an increasing effect on the churn risk.
- The churn risk appears to increase with the number of message interactions the customer has with Hedvig.
- Whether or not the customer has signed up for Hedvig before their insurance with Hedvig starts does not have a homogenous effect on churn on all customers. Rather, the churn risk increases for some and decreases for some if they did not have insurance at the sign date. In general however, customers who did not have insurance immediately have a lower risk of churning, as indicated by the PDP.

- The referral frequency has a clear and positive monotonic relationship with retention.
- Having an edit frequency of 1, meaning that edits are made around once per year has a positive impact on retention compared to having no edits. However, as the number of edits exceeds one, meaning customers with multiple edits per year, churn increases according to some ICE plots. However, the PDP indicates that the general trend is that the churn risk decreases with the edit frequency.
- Having an upcoming insurance contract that has not been put into effect yet, and was decided beforehand, has a positive impact on retention in most of the cases.
- The time being insured by and being a customer at Hedvig does not have a straightforward positive effect on retention, however the average effect is a slightly decreased churn rate for longer time customers, as indicated by the PDP.
- Churn generally increases the longer it has been since the last edit. The same relationships can be seen for referrals, though not shown in the plot.
- Customers who are renters have the highest churn rate (encoded as 2 in the *line\_of\_business* plot), while BRF has the lowest churn rate.

### D.1.3 SHAP

To complement the PFI, ICE and PDP plot, a SHAP plot is provided in figure D.3. SHAP relies on an entirely different approach and assumptions and is therefore a good complement to the other methods. In addition, SHAP is considered more robust and clearly defined for feature importance, and its conclusions may therefore be considered with greater weight.



**Figure D.3 SHAP values for the most significant features for GBT. A higher SHAP value indicates a higher churn risk, and the feature values are encoded by color with a legend to the right. The height of the stacks indicates the density of points at a particular SHAP value.**

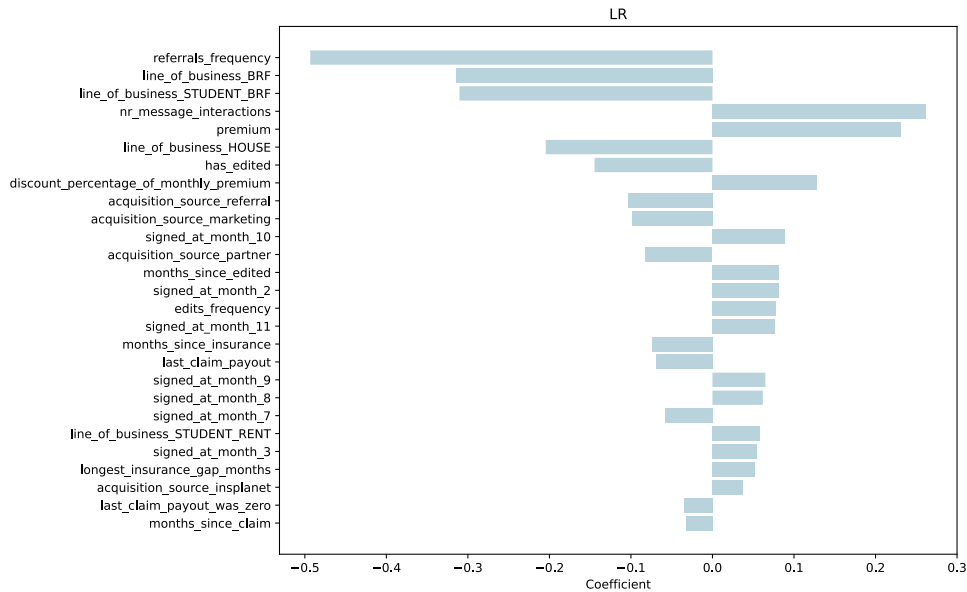
The following things can be noted from the SHAP plot:

- According to this model, line of business is the most significant predictor, with a large difference between the highest and lowest churning categories.
- Low premiums appear to have a positive impact on retention, while medium and high premiums are associated with higher churn. This is hard to interpret since premium is correlated with many other factors, e.g. being a student, or living alone.
- Having zero message interactions is associated with lower churn, while having one or more is associated with higher churn. This factor is hard to interpret since message interactions encompass many different things including claims and edits.
- Having a higher referral frequency is associated with significantly higher retention, with very high frequency leading to a very low churn risk.
- According to SHAP, customers who have been with Hedvig longer appear to have a lower churn risk.
- The month that a customer signs up can have an effect on the churn risk. This is probably linked to different campaigns happening in different months, e.g. students being acquired in August when the semester starts. It could also be linked to different types of customers switching insurance at different times during the year.

- The acquisition source has a significant effect on churn, where signing up through a referral is associated with the highest level of retention.
- If a customer recently edited their insurance, they are less likely to churn. This is intuitive as customers who recently did a change to their insurance probably won't cancel it right away afterwards.
- A high edit frequency has a non-straightforward effect on churn, sometimes associated with increasing or decreasing the churn risk.
- Customers given a high discount from the start are more likely to churn, but the effect is not very large.
- The last premium change has an effect on churn, where a decrease of premium since the last insurance appears to be associated with decreased churn risk, however the relationship is not straightforward.
- Customers given a high claim payout are associated with higher retention.

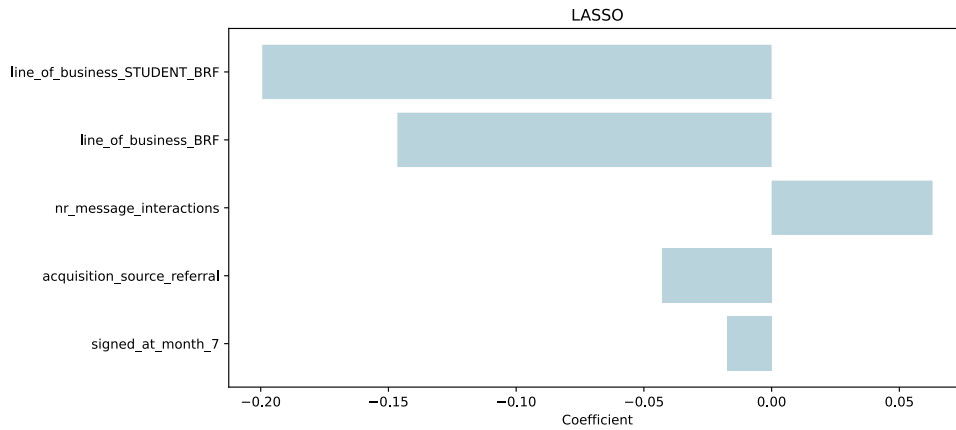
## D.2 Logistic Regression Coefficients

In figure D.4 below, the coefficients of the optimal logistic regression model, with a ridge penalty, is shown, excluding the coefficients with the lowest absolute value.



**Figure D.4 Coefficients of the optimal logistic regression model, excluding the coefficients with the lowest absolute value.**

As seen in the plot, the ridge regression model has a lot of non-zero coefficients, which makes it hard to get a clear and direct answer on what features are important for the model. As such, a LASSO model was trained as well, varying the regularization coefficient  $C$  until only a few predictors were included, while achieving a similar level of performance. The LASSO cross-validated AU-ROC score was 58%, not far away from the optimal LR model, and its coefficients are shown in figure D.5 below.



**Figure D.5 Non-zero coefficients of the LASSO. Note that the features are standard scaled which affects the magnitude of the coefficients. Hence, they cannot be interpreted directly as e.g. odds-ratios.**

As seen in the plot, by far the most significant driver is the line of business, the baseline value being renter. As such, customers owning their own home in a BRF or house have significantly lower churn. Whether this is because such customers are more loyal to Hedvig or move less often is not clear. The next two factors that the model identifies are the number of message interactions the customer has had, with a positive correlation with churn, and whether or not the customer was acquired via a referral, which also correlates with retention. Finally, the model considers at which month the customer signed up with Hedvig. According to the model, customers that signed up in July are less likely to churn than those who signed up in January.

Finally, it is worth noting that this sparse LASSO model, with only five coefficients, has an AU-ROC score similar to that of the top performing model, while being a lot less complex. This may indicate that the relationships that exist in the data are not very complex, but rather, a simple risk model is sufficient to capture the dynamics of churn in the data.

# Appendix E Best Practices to Increase Retention

## E.1 Understanding Churn

In order to increase retention, and implement programs and actions to combat churn, churn must first be examined and understood. This process involves collecting data from many sources such as customer interviews, churn surveys, as well as examining the actual churn data. In this process, it is important to understand which customers churn, why they do it, and what can be done to prevent it. In our interviews, we have seen that the companies that manage to generate accurate predictions tend to have a lot of data on their customers. This includes not only customer data, but also extensive usage data, and satisfaction feedback from many interaction points. In addition, the process is often hypothesis-based, working with hypotheses on signals that could affect churn and examining that relationship further. As such, achieving an accurate predictive model requires a solid understanding of the problem at-hand and substantial pre-processing in order to generate predictors that have a clear signal.

### E.1.1 Predictive Models

Before discussing predictive churn models, it is worth mentioning that there are many other modeling approaches that can be useful based on the same type of data. Magnus, analytics expert at Major Swedish media company B, highlights that one may also make use of segmentation algorithms, time series analysis, survival curves, or customer lifetime value models (CLV) in a retention context, and that it all depends on the specific problem at-hand and the data availability.

Many of the surveyed companies consider predictive churn models a useful tool in their retention work. In general, the start- or scale-ups we talked to are starting to use it or are excited about developing it once the necessary data is collected and in place. The more established companies, especially the streaming services, have sophisticated models in place for different uses, and have generally seen a higher predictive accuracy due to their larger amount of usage data. Camilla, retention expert at Loyalty Factory, regularly gives educational lectures on these topics to the

larger Swedish telecommunication and insurance companies, where churn risk and CLV modeling is commonly used. However, after talking to some of these companies, it is evident that even they are struggling with low predictive accuracy in some cases. Some companies, like Major Swedish media company A, use a multitude of different models. Their analytics team has developed around 30 different models to use for different purposes. These could be anything from simple regression models which are simple to communicate to the rest of the organization, to more sophisticated but less interpretable black box Neural Network models that provide more accurate predictions. Top 4 Swedish telecom A also makes use of multiple models. According to customer loyalty executive Joel, the company has developed one model to predict churn, and another model to predict how likely a given customer would be to accept an extension of the contract and binding period. Customer loyalty executive Christian at Top 4 Swedish telecom B explains how they include other types of models besides the aforementioned, such as models that take care of customers' needs, e.g., determining what customers want to buy and models related to hardware. They have a data management platform in place where they can track and connect existing customers with visits on their website. This can be very useful and could for example be used to see what customers have looked at a new phone release deal on their website, and then use this as input to predict and target customers who are likely to purchase.

Though most of the established companies that we talked to said that they have implemented churn models, these were not always used in practice, as accuracy is not considered good enough to be actionable. This is especially true for the industries that have churn dynamics and customer data similar to insurance, such as telco and home electricity. Joel at Top 4 Swedish telecom A mentioned that they have churn models in place, but that these are not used by the retention team for this reason. An analytics expert at the same telco provider also mentioned that they were struggling to develop a churn model for the Swedish market with enough accuracy and have made several attempts at the problem. Finally, Cecilia, Marketing Manager at Top Swedish electricity provider, said that they were not attempting to model churn, as it was too hard to predict. She mentioned that this stemmed from churn being highly correlated with moving, which is hard to predict as it is hard to identify indicators of customers considering or being in the process of moving. This situation is very similar to home insurance, where switches are often triggered by moving. In general, many companies report that a lot of the behavior around churn is random and not explainable, and therefore it is hard to construct a model that is accurate enough to make it actionable.

The companies that manage to accurately predict churn, including Major Swedish media company B and Audio streaming service C, often have usage data, or more interaction points than what is available in the home insurance context. For example, Malin, customer loyalty executive at Audio streaming service B, mentioned that they see clear triggers of churn related to e.g., the frequency of listening and how



much of the catalog the user has explored. Analytics expert at Top 4 Swedish telecom A, who has extensive experience from churn modeling, mentioned that though they were struggling to develop a churn model for the Swedish market with enough accuracy, while modeling churn for the same company in an emerging market proved successful. There, they had much more data available on more customers, resulting in more accurate models, higher significance on a larger number of predictive features, and more clear signals in the usage data overall. For this project, they applied a segmentation algorithm prior to fitting predictive churn models to the different segments. Through segmentation, distinct segments with distinct behaviors, needs and churn dynamics were identified. Different predictive models were then fit to each segment, which resulted in high accuracy. This attempt was successful for several reasons. Firstly, the different models had high predictive accuracy, explained by the fact that the different models on their own were able to pick up the different customer groups' behaviors, something that the collective model failed to do. Using a more advanced model such as Random Forest would be able to pick up similar signals as multiple separate models. However, by fitting a separate model to different types of customers, the predictions are more actionable in terms of targeting approaches, as the various models' drivers can be interpreted individually, and the communication be tailored to each of the customer groups or segments. In this regard, there is a trade-off, as with more models there are less observations per segment to fit the model on. In the emerging markets with about 60 million users, it was possible, and favorable, to fit multiple models and still get highly accurate predictions, while this has not been possible for them in Sweden due to the limitation of customers and data.

When dealing with predictive churn models is to understand the use case. Top 4 Swedish telecom B mainly uses marketing automation when targeting customers, so basically every single customer could be targeted without needing an increased capacity in the CRM team, limiting the impact of using churn models. However, if the customer would be targeted mainly over the phone where the cost of each conversation is higher, churn models would be much more useful. Something to bear in mind, however, is that for low engagement products such as insurance, the most profitable customers are generally the ones that just remain signed-up without really thinking about it, as the revenue is recurring, but no real marketing or support cost is attributed to these customers. Joel, customer loyalty executive at Top 4 Swedish telecom A, as well as pricing expert and ex-executive at Top 4 Swedish insurance company A, both emphasize this with the common saying that one should let sleeping dogs lie, referring to this idea that one should avoid engaging too much with customers that have forgotten about their subscription or have not considered making any changes to it. As such, targeting and communicating with customers that are not considering churning may lead to more harm than good, which emphasizes the need to carefully identify the at-risk churners rather than targeting a large portion of the customer base in the insurance industry.

Finally, Retention expert Camilla says that the next frontier of predictive modeling in a retention context is something called descriptive models. This includes uplift models, as mentioned in section 2.2 *Literature on Churn Prediction*. Some customers are inherently meaningless to target as they never would bite on a retention offer. Uplift modeling instead attempts to model the expected effect that a retention action would have on a given customer, and overall there are only a few large companies worldwide that do this effectively.

### **E.1.1.1 Triggers**

In order to accurately identify the timing, when customers are at-risk of churning soon, triggers that have a strong signal that indicate a churn behavior are key according to several of the surveyed companies used. Kasper, customer loyalty executive at Loan fintech, explains how they found that checking one's account balance online was a good trigger for a customer considering leaving. In order to switch to another provider, the customer first needs to know the current amount of debt that one has at the current provider in order to get a new quote, and something a common customer otherwise does not check regularly. Therefore, out of the customers that check this page online, a significant share went on to leave. Similarly, Top 4 Swedish telecom A considered an approach where customers visiting the subscription cancel page or checking details about their binding period or other contract details were more likely to churn. Telecommunication providers such as Top 4 Swedish telecom A and Top 4 Swedish telecom B use customer ratings to detect the same thing, and some continue by calling all customers that gave a low rating. Possibly a little more sophisticated, Audio streaming service B has hundreds of triggers set up, reacting to e.g. downward trends in usage. However, this is made possible thanks to their product being more engaging than e.g. insurance or telecommunications. In industries where triggers are not found from their own interaction with customers, such as when moving, external companies may be used to get triggers. Cecilia at Top Swedish electricity provider mentioned that they cooperate with *Adressändring*, which is a service used to make sure that all mail arrives at the new address when moving, to get a notification when customers may be moving, and therefore at-risk of churning.

### **E.1.2 Conclusion**

In conclusion it is important to understand churn, why people churn, and what the underlying drivers are. Here, the best model is to collect feedback from customers from all parts of the organization including churn surveys, looking at the churn data, NPS, customer feedback on products and customer service. Together this gives a full picture of which customers choose to leave and why. Next, in order to accurately predict churn, *trigger* behaviors that highly correlate with churn should be identified

and tracked. Churn prediction models can also be complemented by other models, including segmentation model, which can be useful to understand churn drivers and behaviors of different segments, as well as preventive actions. Finally, employing descriptive models such as uplift modelling may prove useful to identify customers who will react positively on a retention initiative, and help find customers which have churn behaviors that can be prevented.

## E.2 Customer Intake

Customers behave in different ways and depending on what type of customers that have been acquired, they will have different risks and reasons for churning. The type of customers that are acquired depends, among other things, on the marketing approach, the pricing, and the selection of target groups.

### E.2.1 Marketing Approach

Inherently, there is a trade-off between customer intake and customer retention, both in terms of budget restraints and marketing campaigns. For the latter, several companies see a clear causal relationship between large campaigns and increased churn rates. In the period following large campaigns and marketing efforts, companies typically see new customers signing up for their services. However, the retention rate in these campaign-acquired cohorts tends to be significantly lower, which has also been noticed by Hedvig. Likewise, by spending more resources on paid ads online, Product Manager at Electricity provider startup A explains, they have managed to grow their customer base, but at the cost of increasing the churn rate temporarily. For startup companies, keeping this in mind is of utmost importance, as the customer growth rate is directly affected by both factors. As privacy regulation tightens and the public opinion shifts, it may be increasingly hard to target the right customers, which could lead to a mismatch between the acquired customers and the target group. The Product Manager at Electricity provider startup A reflects on this as potentially heightening the importance of retention studies.

### E.2.2 Discounts

In many subscription services with low differentiation, discounts are used extensively to attract new customers, e.g. in electricity distribution, telecommunication, and insurance. In these industries, comparison sites are often used to compare prices between different actors, and companies give out discounts to new customers, sometimes running at a loss during the first year, hoping that the customers will stay longer. This leads to some customers, ‘switchers’, exploiting

these discounts by regularly using sign-up deals at different providers. Therefore, some companies including Electricity provider startup A, do not give out new customer discounts, as this would result in higher churn rates and losing money on ‘switchers’ instead. For other companies, such as TV and telecommunications providers, however, using discounts is the one most effective way to gain new customers and can hardly be restrained from if one is to grow its customer base in a price pressured industry.

### **E.2.3 Target Group**

If there is a good match between the product or service and the customer base’s needs, they are more likely to be satisfied with the provider and less prone to churn. Co-Founder Marcus of a Home alarm startup, a digital IoT connected home alarm, gives an example of this, where they have observed that the ‘alarm-users’, who primarily used the device for home alarm, scored lower on their NPS surveys. Marcus explains that Home alarm startup provides much more functionality than just a home alarm and is hence of lower value for a customer only interested in this particular feature than traditional alarm companies providing only this service. Hence, they started to focus their effort on another segment. Another aspect affecting the customer quality is exemplified by Kasper of the Loan fintech’s decision to market their service primarily toward creditworthy individuals of good financial standing. Even if this may not be an idea conceived to directly increase customer retention, it has influenced this metric as well. Furthermore, Data Analyst John at Audio streaming service A discusses an innovative approach to deciding what customers to spend the marketing budget on. Based on the data on its customers, they use machine learning to predict different customers’ Customer Lifetime Value (CLV). These factors are then used for the customer acquisition phase, where customers predicted to have a larger CLV are allocated a larger share of the marketing budget. Data Analyst John at Audio streaming service A mentions that using this method of prioritizing customer intake, customer acquisition can be focused on the customers that are less likely to churn to begin with.

### **E.2.4 Conclusion**

In conclusion, a company needs to consider what customers are acquired and through which channels, as some customers may be more or less loyal in the first place. In this work, the understanding of churn must be included, for example reconsidering campaigns that result in a high churn rate. Also, models for churn or more generally CLV, can be used to select which people should be targeted by marketing.

## E.3 Product Improvement

Having a great product obviously has a clear impact on customer retention. In fact, many of the companies that participated in our interviews, especially the earlier stage startups and scale-ups, had product improvement as their main retention strategy. This includes providing a great product with the right features, and the right price for the target customer. In addition, the product can be personalized, as preferences may vary between customers. Hence, depending on product or service type, personalization can be an important lever to increase customer satisfaction. Additionally, some companies use their churn analysis information systematically in improving their services.

### E.3.1 Product Features

There is not a lot to be said on product improvement that is both general enough to be transferable across companies, while still not being trivial. A better product, meeting the customers' needs to a higher extent will yield more satisfied customers and higher retention. Many of the surveyed subscription companies have their customer interaction at least partially through an app. However, when developing the app, it is important to make sure that this interaction is meaningful and not just another hasty product push. Product Manager at Electricity provider startup A emphasizes that many traditional electricity providers may historically only had one customer interaction point, being the invoice. At this point they may have released an app, but for the sole purpose of having the same "mossy" invoice but now in a digital format. It is important to really develop the app interaction from the customers' point of view, adding features that add value to the service. As an example, Head of Product at Car Insurtech, explains that during the Covid-19 pandemic noticed that many customers' cars have been discharged, as many cars are not being used as frequently as before. Therefore, as their product is already connected to the car, they introduced a feature where users could see their battery status in the insurance app, receiving a notification if the battery reached a critical level, which has been received very positively by its customers. In addition to providing value to its customers through good interaction points and value adding features, Malin, customer loyalty executive at Audio streaming service B, highlights the importance of ease of use. According to her, one of the key aspects, besides the actual content on the platform, is designing the app in a way that the customer can find its way around in the app among the various services, reducing the number of instances that the customer needs to search or ask for help.

### **E.3.2 Personalization**

Audio streaming service B among several other interviewed companies work extensively with A/B/C testing of app features, in-app communication, and content suggestions in order to match the offering better with what each individual customer wants to see. This is done by first running segmentation algorithms to group customers based on e.g. behavior, followed by testing what a given change in e.g. the front-page layout will have on churn. Additionally, altering the content in the app based on each customer's presumed preferences is a way to further personalize the offering. Retention expert Camilla at Loyalty Factory agrees that becoming more differentiated in one's message to different customers is key for reducing customer churn.

In addition to personalizing the content or interface, there is also a point to be made about personalizing the product itself, by introducing multiple product tiers. This is common to see in e.g. telecommunication offerings, but the Chief Operating Officer at Pet Insurtech also discusses this in an insurance context. When it comes to different socio demographic groups with different purchasing power, some customers may churn for reasons related to price. By introducing a lower tier product, and downgrading customers to what they can afford if they give indications of churning, they may be retained but at a lower monthly revenue. This could in an insurance context mean increasing the deductible or decreasing insurance coverage. Customer loyalty executive Christian at Top 4 Swedish telecom B mentioned that their main focus when it comes to CRM is to match the right product to the right customer, in order to give them the best possible experience and price.

### **E.3.3 Churn Analysis**

There are, besides A/B testing, other data driven approaches to improve the product and reduce churn. Incorporating data in a systematic way from churn analysis into product development, allows for better alignment between product attributes and customer needs. Hanna, customer loyalty executive at Car sharing service, explains how they make use of this by systematically collecting data such as end-of-journey feedback, responses from custom care and the sales organization, and other surveys, and reviewing this on a monthly basis and pushing it out to the respective part of the organization. Home alarm startup also uses these types of inputs and has good routines of collecting feedback, by combining insights from the customer support, NPS surveys, and other sources and transferring it to the product development team. Often, in case of negative feedback, a designer from the product team reaches out to the customer to learn more, both increasing the amount of relevant, qualitative feedback and making the customer feel that they are cared about.

### **E.3.4 Conclusion**

In conclusion, the product is important for retention, as happy customers generally stay, and product improvement is therefore important. In this respect, it is important to match the right product with the right customer so that they get coverage and price that they want rather than switching to a competitor. This can be achieved by having multiple product tiers. In order to improve the product, the knowledge about churn and its drivers should be incorporated in the product development process. Finally, the product should include engaging features with value-add for the customer.

## **E.4 Lock-in**

Many products and services are easily replaceable for customers by switching to a competitor, and this is especially true in a low differentiation industry such as insurance, where acquisition deals and discounts are prevalent for new customers. Therefore, a way to make customers less likely to churn is to create barriers for switching to a competing provider. A common straightforward way to create this lock-in effect is to enforce binding periods for customers, but also to create a service ecosystem that creates additional value when having multiple products or services from the same provider. The lock-in effect could also take an emotional form, by engaging the customer base in social media, or by creating a successful loyalty program.

### **E.4.1 Binding Period**

A straightforward way to make customers stay longer at the company is to legally enforce this by instituting a binding period. This is commonly done in the insurance industry, however this approach has recently been challenged by Hedvig as the only home insurance in Sweden without a binding period. In other industries that have significant up-front investment costs in the customers, a binding period is very important. According to customer loyalty executive Hugo at Major Swedish media company A, distribution channel marketing and hardware cost can in many cases rise to several thousand SEK, while the ongoing customer relationship virtually contributes in its entirety to the profit margin.

A long binding period can on the contrary be off-putting and limit the inflow of new customers. Svante, Business Analyst, explains how the International telecom he works for offers a middle ground here, by not enforcing a binding period for the customers, nevertheless giving customers a discount to customers as compensation for voluntarily locking themselves in at the company for a longer time.

#### **E.4.2 Service Ecosystem**

If a customer only has a single subscription with a company, this contract is easily replaced. However, as soon as a customer has two services from the same company, the churn risk decreases dramatically. As an example of this, Svante mentioned that International telecom saw an 80% increase in retention over 12 months for customers with an additional service. Likewise, according to the COO of Pet Insurtech, the best proven way to increase customer retention is to sell them a second insurance product, and a good way to address this for an Insurtech is to find a cheaper insurance product that can be sold on-top of the original insurance, thereby increasing engagement and loyalty to the company. This idea can be taken even further, where the Product Manager at Electricity provider startup A, is a firm believer in capturing their customers in their ecosystem of products and services. They have for example seen a significant churn decrease for customers with an IoT device connected to their app, as this creates an emotional connection to the customers and increases the value of the service in the eye of the customer. As a customer adds additional products and functionality to the service, the total value becomes larger than the sum of the individual components on their own. Similarly, Top Swedish electricity provider has, according to Marketing Manager Cecilia, observed an increased retention for customers with electricity contracts that also have a product connected, such as solar panels or car charger installed by them.

#### **E.4.3 Emotional Attachment**

Besides the concrete lock-in effects of binding periods and being signed up to multiple services at the same provider, there are other less tangible forms of lock-in effects based on habitual behavior and emotional attachment. This type of emotional lock-in based on engagement comprises increased customer interactions and well-designed loyalty programs.

Increasing the number of customer interaction points in a meaningful way could entail personalized emails, a well-thought-through app experience, or creating an engaging social media community. The COO of Pet Insurtech mentions their retention efforts in creating emotional bonds to the pet owners by growing their community on social media such as Instagram and Facebook, where customers post pictures of their pets, comment on their experience with the insurance, sharing special offers, or inviting them to competitions. To get a sense of the extent of this community, Pet Insurtech has over between 100,000-200,000 likes on Facebook, while If Skadeförsäkring and Folksam have around 25,000, Trygg-Hansa about 40,000, and Länsförsäkringar and Hedvig both have between 1,000 and 2,000. Additionally, creating an app that is engaging and not just a digital piece of paper is a key factor in retaining customers according to Product Manager at Electricity



provider startup A, and the more customers interact with the app, the less likely they are to churn, according to Chief Operating Officer at Content and car Insurtech.

#### **E.4.4 Loyalty Program**

A loyalty program is a marketing process that rewards customers based on level of engagement with the brand or the frequency of purchases and exists for the purpose of increasing the spend and relationship with the company. Loyalty programs saw a surge in the 1990s, building on the idea that it is cheaper to market to existing customers than to acquire new ones (Kumar & Reinartz, 2006, p. 182-186). There are a multitude of examples where these types of programs are used, possibly most commonly in the airline industry. Hedvig's referral program, Hedvig Forever, could be considered a loyalty program, giving customers a monthly discount after referring a new customer, as long as this new customer stays with Hedvig. Another example is Loan fintech engaging their customers in their loyalty program, by giving customers lower interest if they pay on time, and while also taking the opportunity to communicate with the customers.

#### **E.4.5 Conclusion**

In conclusion, retention can be increased through lock-in effects, e.g. offering multiple insurance products or add-ons to create an ecosystem of products and services. These products could include new verticals such as pet or car insurance, or smaller add-on products to the existing home insurance, e.g. extended travel insurance or lowered deductible. Other methods of increasing retention include implementing a voluntary binding period e.g. in exchange for a reduced premium, building a community around the product, and introducing a loyalty or referral program.

### **E.5 Targeting at-risk Churners**

After customers that are at-risk of churning have been identified, they must be targeted or influenced in some way in order to prevent them from cancelling their subscription.

#### **E.5.1 Discounting**

Most companies interviewed do not consider discounting as a viable option to keep customers in the long term. It can in many cases be used to acquire new customers,

but price is often not what makes customers stay. It is also not economically viable to give out discounts to existing customers, according to the COO of Pet Insurtech as some customers are supposed to leave and as discounts give birth to expectations of future discounts. In contrast, International telecom uses discounts in their campaign to keep customers identified to be at the verge of leaving. They use an RFM model (Recency, Frequency, Monetary Value) and focus on at-risk customers in four to five countries at the time. The COO of Pet Insurtech, instead thinks that you should give something else of value to the customer, that does not cost you a lot. Analytics expert at Top 4 Swedish telecom A agrees on this and says that if they had the churn prediction model right, it could be used to give at-risk customers something else of value, such as a headset for example, and customer loyalty executive at Major Swedish media company A reasons similarly but by e.g., giving away their streaming service for half the price if the customer agrees on a longer binding period.

Further, insurance pricing expert suggests that a discount, or price decrease, can have the reverse effect. Intuitively, if you give customers a discount, they would be more likely to stay as they are getting the same product for a lower price. However, based on his experience and studies carried out at one of the largest Swedish insurance companies, giving customers a price decrease resulted in a slightly higher churn rate. If one were to give out discounts, he argues, the price decrease has to be motivated and presented equally well as a price increase would be. His idea would be to try to think a bit more creatively, by e.g. giving customers at-risk of churning a small discount that is properly motivated and presented.

### **E.5.2 Communication**

There are countless different ways to influence customers through targeted communication, and customer loyalty executive Christian at Top 4 Swedish telecom B says that only the imagination sets limits on how one can pick up on various customer behaviors and tackle them either digitally or by phone. In order to be successful at targeting customers at-risk of churning, without spending excess resources per customer, having a clear CRM and loyalty strategy is key according to retention expert Camilla at Loyalty Factory. This strategy should comprise what technology and channels should be used, where marketing automation is fundamental. It is all about automating these processes and leveraging analytics to tailor the message to the specific customer. These messages or campaigns are usually trigger-based and should be developed even further to rather focus on clever, well-thought-through processes, than ad hoc campaigns. Camilla emphasizes that there is a lot of work behind a successful, integrated communication strategy, in order to find the right messages and the right level of interference. Robert, ex-executive at Top 4 Swedish insurance company B, agrees with this statement and argues that it is important to get as much information as possible on the customers

in order to work with specialized and targeted messages, and to build a relationship over time.

As discussed in section E.1 *Understanding Churn*, Loan fintech found that checking one's account balance online was a good trigger for a customer considering leaving. When targeting these customers, of course, at times it helps to give a discount. However, the sole act of reaching out to them over the phone and establishing a contact had a large positive effect on retention. Sometimes, the reason could be a changed need that the customers did not know that they could fill, or simply a misunderstanding about the actual cost of service when comparing loans with different conditions or maturity from different banks. Having a dialogue with these customers, educating them about the product is often enough to convince these customers to stay.

In other subscription services, contacting customers by phone is not a viable option. Audio streaming service C instead focuses on algorithm triggered notifications to get customers to interact more with the service. The notifications are sent out based on customers' different behaviors and machine learning is used to determine what type of notifications works for what types of customers.

In insurance, a common cancelation reason is that a customer moves to a new address. This is also true for electricity distribution companies, and both Electricity provider startup A and Top Swedish electricity provider try to help their customers in this process to keep as many of these customers as possible. It is a critical moment, and carried-out successfully, there is large potential to keep many of the churning customers.

### **E.5.3 Conclusion**

In conclusion, targeting at-risk churners can be effective, and should be done by A/B-testing different trigger-based campaigns through a marketing automation system. There is no general answer to what kind of campaign works the best, rather it should be experimented with which types of customers respond to certain actions. Communication can be done through many channels including notifications and e.g. texting or calling the customer and can include everything from reminding a potential mover that it is easy to move their insurance, to explaining the specifics of different coverages compared to competitors. Additionally, channels connected to churn occasions such as moving, by partnering with external services, can be used to convince customers to stay with the company. Besides communication, customers can be targeted via offering a discount, however this should be done with caution as price decreases may themselves lead to churn in the insurance context, and hence discount must be properly communicated and motivated. Finally, one must consider that caution must be taken, as targeting customers who did not

consider churning in the first place may instead increase their risk of churning, letting “the sleeping dogs lie”.

## E.6 Save Desk

As a final measure, one can try to convince customers who have already decided that they want to terminate their contracts to instead stay at the company. This is commonly addressed by forcing customers who want to cancel their insurance to call a specific number where the customers talk to a team specializing in convincing customers to stay. When considering the group of interviewed professionals, there is a strong faith in the save desks where they are implemented, and several of them have seen significant effects on the churn rate. Robert, ex-executive at Top 4 Swedish insurance company B, explains how his previous employer made use of the save desk to keep their insurance customers, where they had a specialist team with their best salesmen trying to convince the customers over the phone to stay. This team had more flexibility in terms of giving discounts or added services, and worked on commission, receiving a higher compensation if they could convince the customers to stay with less resources. Sometimes, this could be achieved by just going through the details and educating the customers, and overall, Robert says that up to a third of customers can be saved with the right salespeople and the right setup. Kasper at Loan fintech also saw that the sole act of reaching out to the customers over the phone and establishing a contact had a large positive effect on retention, and that educating customers on their service and conditions compared to the competitors has had a good save rate. Interestingly, Top 4 Swedish insurance company B that Robert used to work for noticed that three years after saving these customers, they had a higher retention rate than those that never even reached the save desk previously, shedding light on the critical role of the customer interaction and contact that had been established. In order to maximize the potential of a save desk, Robert and Kasper at Loan Fintech both agree that the save desk employees have to be extremely capable and educated, as well as motivated. Therefore, it is key that they are involved in the design and planning of the save desk.

In addition to downright giving discounts, which is further discussed in section E.5 *Targeting at-risk Churners*, Major Swedish media company A uses methods such as giving away hardware in return for customers agreeing to prolong their binding period. However, there is a trade-off, as giving too spectacular deals just to please the customers momentarily tends to simply postpone the churn of these customers.

In addition to hindering customers from leaving, a save desk can also help companies understand why customers were churning and investigate further what can be done to improve the service. This approach can, as mentioned, also be a way to get valuable insights, and International telecom uses the information from these

types of calls by categorizing the churning customers by churn drivers, in order to be able to further develop their service.

Finally, both Pet Insurtech and Hedvig noticed that part of their churn was due to payment issues, wherefore they initiated a project of calling customers with a failed payment. For Hedvig, it was a successful first step in dealing with churn, as they saw a very high conversion, around 80%, of these called customers that linked a payment method to the service.

### **E.6.1 Conclusion**

In conclusion, save desk can be implemented as a last resort to persuade customers who plan to leave to stay. This practice has proven very effective in some companies and industries, quoting that up to half of the customers can be convinced to stay. However, it is hard to estimate its effect on Insurtechs or Hedvig. In order to be successful in implementing a save-desk, specialized and highly skilled salespeople should be staffed on the desk that know the specifics of coverage and different options compared to competitors in order to convince the customer to stay. Further, the save desk staff should be engaged in the process of improving the save desk practices in order to assure buy-in and motivation.

## **E.7 Organizational Setup**

Retention is affected by the work of almost all functions within a company, from customer acquisition, on-boarding, product development, communication, and eventually retention or save desk. In order to work with retention effectively, and follow-up on how the work is going, it is important to define what the key metrics are to track for the company, who is responsible for the progress, and how the work and responsibilities are divided in the organization. This is not only important in order to know how the work is going but can also be valuable as motivation for the different functions or departments. Retention expert Camilla at Loyalty Factory says that retention should be tracked by every company, but only around 50% of companies do actually measure and track this metric.

It is hard to generalize how the retention responsibility is split within the interviewed companies, as it varies vastly between organizations and stages of the company. The COO of Pet Insurtech highlights the importance of them having a dedicated product manager, whose role is to solely focus on retention. According to them, this is key in order to have a data driven approach and to ask the right questions. Without the right data, machine learning models will do nothing for retention, hence their main current focus is on collecting more relevant data points. Once this data is available,

they consider it very useful to develop a model showing a heatmap of customers that are more likely to churn, to be able to increase the interactions with these customers proactively. With this in one's arsenal, it makes sense to redistribute parts of the budget from acquisition to retention. However, without the full attention of a product manager, retention will not be prioritized enough to collect the right data and to start the right discussions within the organization. In contrast to this, ex-Product Manager at Audio streaming service C does not believe in a role such as Chief Retention Officer or similar, as she considers the responsibility too vague, as it is affected by basically everything the company does.

At startups like Home alarm startup, the responsibility appears to still be centralized, in their case on the desks of the founders and controller, reviewing the progress weekly in their meaning. At larger companies, like Telia, the retention responsibility is, according to retention expert Camilla at Loyalty Factory, typically split into different KPIs and managed by the CRM team or whoever is in charge of the existing customer relationships. Marcus at Home alarm startup opines that, in order to be more actionable for all teams, the metric should be split into sub-metrics that more closely correspond to the work of the different teams, such as splitting churn depending on how far into their customer journey they canceled their contract. As retention is such a coarse metric, Marcus at Home alarm startup argues that it is better to look at live-metrics, or drivers of churn, where the effect can be seen more quickly, e.g. the offline rate of alarms in their case.

### **E.7.1 Conclusion**

In conclusion, the right organizational setup is key to implement and develop these retention strategies, and as retention is something that is affected by many parts of the organization and customer journey, retention needs to be cross-functional. However, in our interview study there was no consensus regarding which organizational setups is best. For example, some companies had a dedicated retention manager, while respondents from other companies did not believe such a role should exist. As an overall metric, retention can be very coarse and not measure the performance of a single team or initiative. Therefore, the metric should be broken down into smaller pieces which different teams have ownership of. For example, the churn rate during the first 6 months of a customer could be tracked and owned by the growth team, whereas the churn rate after a claim could be owned by the insurance claims team.