Predicting customer churn in the healthcare insurance market

a case study

Hendrikse, K.C.M.

Award date:
2017

Technische Universiteit
Eindhoven
University of Technology

MASTER THESIS

In partial fulfillment of the requirements for the degree of

Master of Science in Operations Management and Logistics for Healthcare

Predicting Customer Churn in the Healthcare Insurance Market: A Case Study

K.C.M. (Kim) Hendrikse, BSc.
0766791

Supervisors:

Dr. A.M. (Anna) Wilbik (*TU/e, Information Systems*)
Dr. Ir. R.M. (Remco) Dijkman (*TU/e, Information Systems*)
Prof. Dr. Ir. U. (Uzay) Kaymak (*TU/e, Information Systems*)
E.J.A. (Ellen) Seelen (*Interpolis, Senior Marketeer*)

Eindhoven, August 16, 2017

TUE. School of Industrial Engineering.
Series Master Theses Operations Management and Logistics

Subject Headings: data mining, customer relationship management, churn prediction, Net Promoter Score, feature selection, health insurance, AUK, AUC

“Intellectuals solve problems, geniuses prevent them.”

- Albert Einstein –

Preface

This thesis is the result of my final project in order to fulfill the master degree in Operations Management & Logistics for Healthcare. My time at the Eindhoven University of Technology, and especially the past half year, has been a greet learning experience and I would like to take this opportunity to thank the people who made this possible.

First of all, I would like to thank the people from the TU/e that were part of my project: my first supervisor Anna Wilbik, my second supervisor Remco Dijkman, and my third assessor Uzay Kaymak. I would like to thank Uzay Kaymak for making time to be part of my assessment committee, Remco Dijkman for his feedback and Anna Wilbik, for sharing her knowledge with me, for providing me with feedback and for supporting me in the decisions I have made throughout the project.

Furthermore, I would like to thank the people from Interpolis for their support during my project. I would like to thank René Voets for giving me the opportunity to be part of his team during the project. My special thanks goes to Ellen Seelen for supporting me during the project on all levels. Ellen, not only where you there to answer all my questions, your enthusiasm really motivated me to get the best out of myself! Furthermore, I would like to thank Dennis Kroese for sharing his extensive knowledge, experience and enthusiasm about the subject and providing me with his feedback. Moreover, my thanks goes to all the other people from the marketing department and marketing intelligence department, especially to Arjan de Groof, Joop van Leenders, Bas Kimmel, Pernel Schriks and Karen Terpstra, for their contributions to my project.

Finally, as the end of this project means the end of my time as a student, I would like to thank all the people that were part of this great time. A special thanks goes to my friends, who have put a smile on my face so many times. Thank you for making these great memories with me! Furthermore, I would like to thank my parents, my sisters and the rest of my family for their unconditional support and believe in me. And last but not least, I would like to thank my boyfriend, for cheering me up and helping me to believe in myself. Your support was of great help to me!

Kim Hendrikse

Executive Summary

This report is the result of a research project about customer churn prediction in the healthcare insurance sector. The project was conducted at Interpolis, a Dutch insurance company.

Background

Previous research indicated that the cost of retaining an existing customer is five times lower than obtaining a new customer (Chiang et al., 2003). This means that customer retention is very important for a company. However, in the Netherlands, around 7% of the people switch to a different healthcare insurance provider at the end of the year. At first glance, this percentage might not seem very large, but expressed in absolute numbers it is around 1,2 million people each year (Nederlandse Zorgautoriteit, 2016). This indicates that switching is a structural problem for the healthcare insurance companies.

This switching can be avoided with good customer retention, which is one of the four dimensions of Customer Relationship Management (CRM). CRM tries to understand customers and the factors that have an influence on their loyalty (Bhat & Darzi, 2016). One important part of customer retention is churn analysis (Ngai, et al., 2009). Customer churn can be defined as: 'the percentage of the firm's customer base that leaves in a given time period' (Blattberg et al., 2008). Churn analysis and churn prediction provide insights into what type of customers leave and what factors contribute to this decision to leave. These insights and predictions can be used to retain customers more effectively.

Research Question

The aim of the project was to develop an accurate and relevant customer churn prediction model for the healthcare insurance of Interpolis. Accurate meaning that the predictive ability of the model is sufficient enough for the model to be used and relevant meaning that the model is able to support the marketing department in their decision making. The data that was available for building this model, consisted of customer characteristics, policy characteristics, customer webpage visits, customer touchpoints and the Net Promoter Score (NPS). This led to the main research question:

How can an accurate and relevant customer churn prediction model be created using data about customer and policy characteristics, webpage visits, touchpoints, and the NPS?

Methodology

In order to answer the research question, *logistic regression* models were developed that used different *types of predictors,* arising from the data about customer and policy characteristics, touchpoints, webpage visits and the NPS, and the following types of *feature selection* methods:

- Feature selection by *domain experts*: For the base model, features were selected by *domain experts* from the 'Consumer Marketing' department. This was done with the help of a survey. A feature was selected when more than half of the experts indicated that they expected this feature to have a significant effect on customer churn.
- *Significance* method: This method followed a backward approach. First all the variables were added to the model and subsequently the variables that were the least significant were dropped one by one, until a model was reached in which all variables were significant at the 0.05 level.

- *Filter* method: This method ranked the variables based on the Mutual Information criterion. After the ranking was done, the filter method started by building a model with the variable that had the highest Mutual Information value and then adding the variable with the second highest Mutual information value, and so forth. It was checked which number of variables led to the highest value for the 'Area under the ROC curve' (AUC).
- *Wrapper* method: In this case a forward search was done, in which first all possible models with one variable were build. Subsequently, the model that led to the highest AUC was chosen. After this the remaining variables were added one by one such that the best model with two variables was found, and so forth. This process was continued until the performance of the model did not further increase.
- *Embedded* method: This was the 'least absolute shrinkage and selection operator' (Lasso) method. This method penalizes the regression coefficients by a shrinkage parameter λ, such that the coefficients of some input variables might have shrunk to zero. In this study, a value for λ was chosen that maximized the AUC value. For this λ, the regression coefficients were checked and the variables that had coefficients larger than zero were selected for the model.

The evaluation criteria that were used in this study are: *accuracy, recall, precision*, the area under the ROC curve (*AUC*) and the area under the Kappa curve (*AUK*).

Results

The performance indicators of the various models that were developed, are shown in Table 0.1.

Table 0.1. Performance indicators of the developed models.

***Variables**	**1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
****Best FS Method**	1	2, 4 & 5	2 & 4	5	2 & 4	5	2, 3 & 5	4	2, 3 & 5	2, 3 & 4
Accuracy	0.803	0.799	0.800	0.746	0.787	**0.807**	0.706	0.805	0.706	0.779
Recall	0.275	0.289	0.287	**0.397**	0.324	0.301	0.289	0.307	0.289	0.394
Precision	0.183	0.184	0.185	0.136	0.184	0.260	**0.265**	0.258	**0.265**	0.215
Kappa	0.112	0.116	0.117	0.093	0.122	0.168	0.092	**0.168**	0.092	0.161
AUC	0.648	0.650	0.650	0.639	0.662	**0.670**	0.524	0.670	0.524	0.662
AUK	0.066	0.066	0.067	0.053	0.072	**0.092**	0.025	0.092	0.025	0.084

* 1: *Customer and Policy Characteristics;* 2: *1 + Logged in or not;* 3: *If logged in, 1 + webpage visits;* 4: *1 + Contacted and/or Complained or not;* 5: *If contacted, 1 + Contact info;* 6: *5 + Filled in NPS or not;* 7: *If complained, 1 + Complaint info;* 8: *7 + Filled in NPS or not;* 9: *If NPS was filled in, 5 + NPS Score*

** 1: *Domain Experts;* 2: *Significance;* 3: *Filter;* 4: *Wrapper;* 5: *Embedded method*

Conclusion and Recommendations for the Company

A relatively accurate and relevant customer churn prediction model could be created by using variables related to:

- Customer and Policy Characteristics: *Customer Age, Contract Duration, Type of Additional and Dental Insurance and Voluntary Excess Risk*
- Webpage visits: whether the customer has *logged in* on the secured domain of Interpolis at least once or did not log in at all
- Touchpoints: whether the customer has *contacted* the insurance company and whether the customer has *filed a complaint.* If a customer had contacted the insurance company, the

predictive ability could be further increased by adding: the *Call, Mail and Social Media Frequency*

Considering the different feature selection methods, it was found that the *filter method* often resulted in the *smallest feature subset* and the *highest accuracy*, while the *wrapper method* often resulted in the *largest feature subset* and the *highest values for AUC and AUK*. Moreover, the *significance* and *embedded methods* often obtained the *highest Kappa*, but this did not differ much from the wrapper method. When taking into account that churn models often need to deal with highly unbalanced data, which makes the Kappa, AUC and AUK important criteria, the *wrapper method* might be the best method when using logistic regression to predict customer churn.

Apart from giving insight into the variables that play an important role in predicting churn, it was checked whether the models that were developed, could be implemented at the 'Consumer Marketing' department of Interpolis. Ideally, this would lead to a scenario in which Interpolis is able to predict churn for each customer on a daily basis. When this is known, different retention actions can be done, such as giving the customer extra attention by calling him, sending an e-mail or a letter, or adapting the content of webpages, service mails and even phone calls. However, in order to get there, there are several important boundary conditions that need to be complied.

First of all, an *accurate model* is needed, that is able to retrieve many churners, without making too many mistakes considering the non-churners. The latter is important because of the unwanted side effect of making a retention effort for a customer who was not thinking about churning. These customers are now reminded of the possibility to churn. This effect is known as the 'wake up' effect.

Secondly, the model should be *robust* or i.e. the model should have high *staying power*. A model has high staying power when its predictive performance in a number of periods after the estimation period remains approximately the same (Risselada et al., 2010).

Thirdly, when an accurate and robust model is developed, the *data infrastructure* should make it possible that the data that is needed as input for this model, is easily accessible and available on a daily basis. This is necessary in order to automate the prediction.

Considering the models that were developed in this study, for now, only the model that used the customer and policy characteristics as predictors meets these boundary conditions. Although it should be noted that the 'wake up' effect needs to be considered. But besides this it was found that, when using customer age, contract duration, voluntary excess risk, additional insurance, dental insurance and collectivity discount:

A *quarter of the churners* can be reached when contacting approximately *10% of the overall population*

Therefore it is recommended that Interpolis uses this model in the upcoming campaign, making it possible to adjust the retention efforts by the likelihood of a customer churning. Meanwhile, Interpolis can work on a better data infrastructure, increase the quality of the data about webpages and contact moments, explore the importance of other variables and experiment with different prediction techniques. Moreover, Interpolis and Rabobank, which is the sales channel from Interpolis, can cooperate more closely in developing churn prediction models, such that the data from Rabobank, can be used to improve the model.

Table of Contents

Chapter 1 Introduction

The healthcare insurance market is going through a major change. Insurance companies need to change their business models such that long-term collaborative relationships with their clients can be reached, instead of short-term contractual ones (EY, 2015). There are several trends within the health care sector that provoke this need for change. One of these trends is that customers are taking more and more control over their health care decisions. This means that insurance companies need to be customer centric (EY, 2015; Yoder et al., 2012).

If insurance companies want to build long-term collaborative relationships, switching should be avoided. However, in the Netherlands, around 7% of the people switch to a different healthcare insurance provider at the end of the year. At first glance, this percentage might not seem very large, but expressed in absolute numbers it is around 1,2 million people each year (Nederlandse Zorgautoriteit, 2016). This means that this switching is a structural problem for the healthcare insurance companies. In order for insurance companies to retain their customers effectively, a first step is to know what type of customers leave and what factors contribute to this decision to leave. These insights can be gained through the development of a customer churn prediction model. Churning can be defined as the loss of customers.

The aim of this project was to develop such a churn prediction model for Interpolis, a Dutch healthcare insurance company. For this reason more information about the Dutch Healthcare Insurance System and Interpolis is given in the following section 'Case Study'. After this the research questions that were defined for this project are discussed. The remaining part of this chapter discusses, why this project was conducted in terms of its theoretical and practical contribution and outlines the remaining chapters of this thesis.

1.1 Case Study

This section gives more information about the Dutch healthcare insurance system in general and more specifically about the insurance company 'Interpolis', were this study was conducted.

1.1.1 The Dutch Healthcare Insurance System

The Dutch Healthcare system is based on three principles, namely: access to care for everyone, solidarity through a health insurance that is available and mandatory for everyone and good quality of care. This system is based on four different laws, namely: the 'Health Insurance Act', the Act for Long-term Care, the Social Support Act and the Juvenile Act. The 'Health Insurance Act' is the law that is the most important one for healthcare insurers and therefore this one will be explained into more detail (Ministerie van Volksgezondheid, Welzijn en Sport; 2016).

The 'Health Insurance Act' was introduced in 2006 and caused big changes in the healthcare system. Before 2006 the Dutch healthcare system was a combination of public and private insurance. When a citizen had an income below a certain amount, he had a public insurance. Citizens with a higher income were obliged to insure privately. After 2006, the public insurance was abolished and all citizens were obliged to insure privately.

However, the 'Health Insurance Act' still has public elements. The government compiles several public conditions, which ensure the social nature of the healthcare insurance (Ministerie van Volksgezondheid, Welzijn en Sport; 2016):

- Citizens are obliged to close a standard health insurance contract ('basispakket') and they are free to choose the insurance company of their preference.
- Insurance companies are obliged to accept citizens for the standard health insurance contract, regardless of the health of the patient.
- The premium of the basic insurance policy is the same for every individual, regardless of his health, age or background.
- Healthcare insurers have the duty to make the insured care available to all their clients.
- The content of the 'standard health' package is stated by the law.

Next to the obligatory 'standard health' package, insurance companies may provide supplementary care packages, for example for specialized dentist care. It is possible for a customer to have his basic insurance at one insurer and his additional or dental insurance at a different one. However, this is not very common due to the risk of insufficient or overlapping coverage. Moreover, additional packages are not mandatory (Ministerie van Volksgezondheid, Welzijn en Sport; 2016).

The government does not have a direct participation in the implementation of the 'Health Insurance Act'. In this way, the healthcare insurance companies and the healthcare providers have a lot of freedom to decide on how the implementation should be done. This makes it possible to create a market in which competition and market forces stimulate high quality and efficiency (Ministerie van Volksgezondheid, Welzijn en Sport; 2016).

Healthcare insurance companies can choose for themselves which healthcare providers they want to contract or not. In this way, they can give direction to quality and costs. To ensure the quality of care, there are several government institutions that monitor the care that is given (Ministerie van Volksgezondheid, Welzijn en Sport; 2016).

In the same way, citizens can switch between healthcare insurance companies if they are not satisfied with their current one. But they are always legally required to pick one. When a person wants to switch to another healthcare insurance company he needs to do this before the 31st of January. Switching to a different health insurer after the 31st of January , is only allowed in one of the following cases (Independer, 2017):

- In case a person turns 18 years old.
- In case a person is insured through the same policy as his/her partner and is going to a divorce.
- In case a person starts working for a new employer and wants to insure through the collectivity of his/her new employer. Collectivity means that a certain group has the same basic insurance.
- In case a person was insured through the Department of Defense while being in military service and quits this military service.
- In case a person immigrates to the Netherlands.

1.1.2 Interpolis

Interpolis merged with Achmea in 2005 and together with other brands like, 'Centraal Beheer', 'FBTO' and 'Zilveren Kruis' it is the biggest insurer in the Netherlands. Interpolis collaborates with Rabobank, a Dutch banking company. Rabobank is the only sales channel of Interpolis, meaning that if a customer wants to close an insurance contract, he needs to do this via Rabobank. This means that Interpolis primarily focuses on Rabobank clients. Interpolis has lots of different insurance policies, ranging from car insurance, to travel insurance and health insurance. With its slogan 'Glashelder' (crystal clear), Interpolis is striving to unambiguous and straightforward policies and conditions. Interpolis wants to do more than compensating claims. It wants to give people insights into their risks and show how prevention can decrease or eliminate risks (Interpolis, 2017).

This project was conducted on the 'Consumer Marketing' department, more specifically in the team that is responsible for the marketing related activities considering the healthcare insurance. Apart from this department, the project was conducted in close collaboration with the 'Marketing Intelligence' department, because the people in this team have a lot of knowledge about the data used for the model. Besides these people, a person from the Strategy, Marketing & Innovation team from Achmea gave his support during the execution of the project. This person developed churn prediction models for the healthcare insurance of 'FBTO' and 'Zilveren Kruis' and has therefore much experience when it comes to developing churn models for the healthcare insurance sector.

1.2 Research Questions

The aim of the project was to develop an accurate and relevant customer churn prediction model for the healthcare insurance of Interpolis. Accurate meaning that the predictive ability of the model is sufficient enough for the model to be used and relevant meaning that the model is able to support the marketing department in their decision making. The data that was available for building this model, consists of customer characteristics, policy characteristics, customer webpage visits, customer touchpoints and Net Promoter Score (NPS). This led to the main research question:

How can an accurate and relevant customer churn prediction model be created using data about customer and policy characteristics, webpage visits, touchpoints, and the NPS?

Answering this question, consisted of answering several sub-questions.
The aim of sub-questions 1 to 4 was to identify good predictors for customer churn. Sub-question 1 focuses on the customer and policy characteristics only, while sub-question 2, 3 and 4 are about the change in predictive ability of the model when data about customer web page visits, customer touchpoints, and the NPS are used additionally. Sub-question 5 focuses on how different feature selection methods influence the predictive ability of the model.

Sub-question 1: What customer and policy characteristics are good predictors for customer churn?

Sub-question 2: How does the inclusion of data about customer web page visits improve the predictive ability of a customer churn model?

Sub-question 3: How does the inclusion of data about other customer touchpoints (phone calls, emails and messages on social media) improve the predictive ability of a customer churn model?

Sub-question 4: How does the inclusion of data about the Net Promoter Score (NPS) further improve the predictive ability of a customer churn model?

Sub-question 5: How do different feature selection methods influence the predictive ability of a customer churn model?

While sub-questions 1 to 6 are mainly about the development of the model, sub-question 6 is about the implementation of the model. The aim of this sub-question was therefore to figure out how the model can be used in the marketing department.

Sub-question 6: How can the models support decision making in the marketing department?

1.3 Motivation

This section provides answers to why this project was conducted in terms of its contribution to research (theoretical) as well as its contribution for the company (practical).

1.3.1 Theoretical Motivation

The development of churn prediction models is a frequently examined subject. However, most of the existing prediction models are applied to data from the telecommunication industry. And, even when a model is applied to insurance data, it is often not health insurance data, but mostly motor insurance data. The reasons for a customer to churn might differ strongly from industry to industry, which makes it necessary to do more research about customer churning in the health insurance sector.

Using data from a Dutch health insurance company has several strong advantages that arise from the nature of the Dutch healthcare system. First of all, Dutch citizens are obliged to close a health insurance contract. This implies that when a customer churns, he is always going to a Dutch competitor, except when an insurance contract comes to an end because of decease of the insured person or when the customer leaves the Netherlands. Secondly, the Dutch healthcare insurance market is very competitive, so it happens quite often that customers switch to another healthcare insurance company. In 2015 the percentage of people that switched was equal to 7,3% of the insured, which is approximately 1,22 million insured people (Nederlandse Zorgautoriteit, 2016). Lastly, this switching can, in most cases, only occur once per year (at the end of the year), since the health insurance contract has a duration of one year. This might make the development of the churn prediction model more reliable, compared to churn prediction models in other markets.

Apart from expanding the churn prediction model to the healthcare insurance market, this project gives new insights into using different types of predictors for customer churn in the healthcare insurance sector. While previous studies focused mostly on rather basic customer and policy characteristics, the aim of this study was to expand these models by including data about customer touchpoints, customer webpage visits and the Net Promoter Score.

Finally, this study focuses on how different feature selection methods may lead to different results. Feature selection is when one selects a subset of features for the development of the model. In the academic literature it is well known that feature selection is often necessary to overcome the problem of the so called ‘curse of dimensionality’. This means that in high dimensionality, data becomes sparser, which negatively influences algorithms designed for low-dimensional space. Apart from this, the use of too many features often leads to overfitting on the training data, which may lead to poorer performance when using the model on new data. Feature selection is therefore an important step in the development of a model and there are many different methods to select

features (Li et al., 2016). This study compares several of these different methods and gives insights into what method leads to the best result in different cases.

1.3.2 Practical Motivation

From research, it is known that the cost of retaining an existing customer is five times lower than obtaining a new customer (Chiang et al., 2003). In order for insurance companies to retain their customers effectively, they first need to know which customers leave and what their reasons for leaving are. This is where the development of a churn prediction model might offer a solution. If it is known that a customer is likely to churn, the insurer can actively try to motivate the customer to stay.

Getting more insight into what type of customers are leaving and what factors contribute to this, is exactly what Interpolis wants. One of the reasons for this is that the customer churn rate of the health insurance policy of Interpolis is higher compared to the average on this market. Developing a churn prediction model could be the beginning of a series of actions to decrease this churn rate. Luckily Interpolis has a lot of data available that can be used for the development of such a churn prediction model. The Net Promoter Score (NPS), is for example, constantly collected after a customer has contacted the company. A variable such as the NPS might be a good indicator for churn, since a low NPS score might indicate that a customer is not happy with the company and wants to churn.

1.4 Thesis Outline

The structure of the thesis is as follows: Chapter 2 starts with explaining into more detail what customer churn is, what predictors for customer churn are found in the literature and explains important terms and techniques that are used in this study. By doing this, Chapter 2 provides the theoretical background that is needed to fully understand this study. Chapter 3, named 'Methodology' is about how the models, that are needed to answer the research questions, will be generated considering the experimental setup. After giving insight into the methodology, Chapter 4 describes the data that is available for the development of the model and explains how the preparation of this data is done. In Chapter 5 the different models that are developed and their performance according to several well-known evaluation criteria are discussed. After the different models are evaluated, Chapter 6, named 'Practical Implications' focuses on how the models can support decision making at the marketing department, answering sub-question 6. Chapter 7 gives conclusions of this study, discusses its limitations and gives recommendations for future research.

Chapter 2 Theoretical Background

This chapter provides the theoretical background of the study, by explaining important terms, techniques and findings from previous studies. It starts with defining customer churn and the broader research area of customer relationship management, followed by giving insights into variables that were found to predict customer churn. After this, two important terms are explained, namely customer touchpoints and the Net Promoter Score. These constructs are expected to predict customer churn. Finally, this chapter explains the techniques that are used in this study: Logistic Regression, Term Frequency - Inverse Document Frequency, different evaluation criteria and different feature selection methods.

2.1 Customer Churn and Customer Relationship Management (CRM)

Blattberg, Kim and Neslin (2008) define customer churn as: 'the percentage of the firm's customer base that leaves in a given time period'. According to them, customer churn can be divided into two broad types: voluntary and involuntary churn. With *involuntary churn*, it is the company that decides to terminate the contract with the client. This occurs, for example, when the customer does not fulfil his financial obligations as mentioned in the contract (Blattberg et al., 2008).

In the case of voluntary churn, it is the customer that decides to terminate the contract with the company. Voluntary churn can again be divided into two types: deliberate voluntary churn and incidental voluntary churn. In case of *deliberate voluntary churn*, the customer is not satisfied or has received a better offer from a competitor. *Incidental voluntary churn* means the customer does no longer need the product or service or has moved to a location where the company does not operate (Blattberg et al., 2008).

Customer churn analysis is part of the broader field of Customer Relationship Management (CRM). The main goal of CRM is to make it possible for companies to build long-term relationships with customers. In order to do this, CRM tries to understand customers and the factors that have an influence on their loyalty (Bhat & Darzi, 2016). CRM consists of four dimensions, namely: 1) Customer Identification, 2) Customer Attraction, 3) Customer Retention and 4) Customer Development. Churn analysis belongs to the customer retention dimension (Ngai, et al., 2009).

2.2 Predictors for Customer Churn

To get an idea about what variables can be used to predict customer churn in the insurance sector, a literature study was performed. This literature study compared eleven articles on the predictors that were used for predicting churn in the insurance sector. An overview of these articles can be found in Appendix A Table 1, together with an overview of the search terms (Table 2) and search engines (Table 3) that were used to find these articles.

It was concluded that there is a lot of variation between the predictors used in different studies. Which predictors were selected, depended strongly on which predictors were available and what the aim of the prediction model was. For example, some studies mostly focused on the characteristics of the customer, while others focused more on how cancelation of one policy affected the cancellation of other policies owned by that household.

However, there were several predictors that were found in almost all the studies that were analyzed. These predictors can be divided into three groups:

- Predictors related to *customer characteristics*, such as: customer age, customer gender and marital status.
- Predictors related to the *policy*: total premium, discount, change in premium, type of coverage, total cost of claims.
- Predictors related to the *relationship* between customer and company: total number of policies and duration of continuous relationship.

These frequently used variables were found in studies that developed a churn model for any type of insurance product. Among these studies, there were two studies that specifically focused on health insurance. Apart from the predictors mentioned above, these studies also used predictors about: the customers *family configuration*, the *income* of the customer, the *insurance package type*, whether the customer is *individually or collectively* insured and whether the customers spouse or *partner* was also insured at the company (Risselada et al., 2010; Günther et al., 2014).

2.3. Customer Touchpoints and the Net Promoter Score (NPS)

Zomerdijk and Voss (2011) described touchpoints as "moments of contact between the customer and the organization" (Zomerdijk and Voss, 2011). According to Halvorsrud, Kvale and Følstad (2016) there are several attributes that characterize a touchpoint. The first one is the *initiator*, which is the person who sets up the contact. This can be the customer, service provider or a complementary service provider involved in the service delivery. The second one is *time*, which specifies when the touchpoint took place. The third one is the *channel* by which the touchpoint was mediated. Examples of channels are telephone, webpages and e-mail. The fourth and last one is the *trace*, which is some physical or digital content that emerged as the result of the touchpoint, for example an entry in a database that keeps track of incoming calls. Defining touchpoint in this way, means that advertising and broadcast commercials are excluded from this definition (Halvorsrud et al., 2016).

The net promoter score (NPS) has become one of the most widely used marketing metrics, since 2003. It is measured through one simple question, namely: 'How likely is it that you would recommend this company to a friend or colleague?'. Customers answer this question on a scale from 0 to 10, with 10 meaning that the customer would strongly recommend the company. The respondents are grouped into three categories on the basis of their score. Customers who answer 9 or 10 are considered *promoters*, customers who answer 7 or 8 are called *passives*, and those who answer 6 or less are called *detractors*. The score is the percentage of promoters minus the percentage of detractors (Bendle & Bagga, 2016).

The NPS is expected to be a good indicator for customer churn. According to Bain & Company promoters are far more likely to remain customers than others. Passives may defect, if a competitor draws their attention and detractors have high rates of churn (Bain & Company, 2017). De Haan, Verhoef and Wiesel (2015) studied the predictive ability of different customer feedback metrics for retention. They found that NPS has a significant impact on retention at both the customer and firm level (De Haan et al., 2015).

Bain & Company defines three types of NPS, namely: *competitive benchmark* NPS, *relationship* NPS and *experience* NPS. Competitive benchmark NPS uses not only the feedback from a company's own

customers, but also feedback that is given to its competitors. On the contrary, relationship and experience NPS only look at feedback from a company's own customers. Relationship NPS data is collected regularly with a customer sample, while experience NPS is collected after selected experiences, transactions or episodes. For example, after a customer called to the company (Bain & Company, 2017).

2.4 Logistic Regression

Logistic regression attempts to fit a line (an intercept and slope) to the data, which makes it similar to linear regression. However, logistic regression can be used in case of a non-linear output variable. This is for example the case when predicting churn, since the customer either churns or does not churn (Sainani, 2014). Figure 1A demonstrates what happens if linear regression is used to predict a binary outcome variable (Y), like churn with a predictor X. It can be seen from this that the line does not have a good fit and is therefore not able to predict Y accurately.

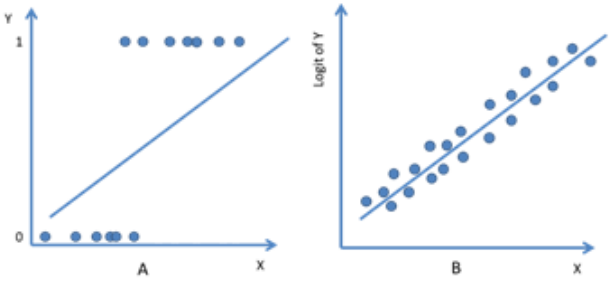


Figure 1. Scatterplot of a binary outcome (A) and a logit outcome (B) predicted by x.

So, instead of fitting a line to the binary outcome, Logistic Regression uses a transformation of the outcome, called a logit or log odds. In Figure 1B it can be seen that, after this is done, the line has a much better fit. Therefore, it is better able to predict Y with predictor x (Sainani, 2014). The formula of this logit is as follows:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n = \ln\left(\frac{P(y_i)}{1 - P(y_i)}\right)$$

In this formula, β_n are the coefficients of the predictors x_n and $P(y_i)$ is the probability of the event occurring. This makes the logit directly related to the probability of the outcome (Sainani, 2014).

Besides using logistic regression in this study, another technique is used, which is the 'Term Frequency – Inverse Document Frequency' (TF – IDF). In this study the TF – IDF technique was used for data transformation purposes, but it will now be first explained in general.

2.5 Term Frequency – Inverse Document Frequency (TF - IDF)

The Term Frequency – Inverse Document Frequency (TF-IDF) is a method that is well known in the field of information retrieval and text mining. It is often used to determine keywords in a given document by looking at both the frequency and the distinctiveness of a word. For example, the word 'and' might occur very frequently in a certain document, however it would be incorrect to define this as a keyword since it is not distinctive (i.e. it occurs frequently in most documents). On the other hand a word that occurs in only few documents, might be very distinctive, however when it only occurs very few times in a given document, defining it as a keyword would lead to poor information retrieval results.

The TF-IDF algorithm accounts for both these characteristics by using the following terms (Havrlant & Kreinovich, 2017):

- The *Term Frequency*, which is the number of times a word occurs in a document.
- The total number of documents, referred to as *N*.
- The *Document Frequency*, which is the number of documents that contain the given word.

The *Inverse Document Frequency* can be calculated by using the total number of documents (*N*) and the *Document Frequency*, as follows (Havrlant & Kreinovich, 2017):

$$IDF = \ln\left(\frac{N}{Document\ Frequency}\right)$$

After this the product of the *Term Frequency* and *Inverse Document Frequency* is calculated. The words with the highest TF-IDF are often chosen as keywords for a certain document (Havrlant & Kreinovich, 2017).

2.6 Evaluation Criteria

Evaluation criteria that are commonly used for testing the validity of a classification model are: *accuracy, recall, precision and the area under the ROC curve (AUC)* (He, and Garcia, 2009). Recently, however, a variant of the area under the ROC curve (AUC) is getting increasingly attention, this criterion is called the *area under the Kappa curve (AUK)*.

2.6.1 Accuracy, Recall and Precision

The accuracy, recall and precision are calculated with the use of the so called *confusion matrix* (He, and Garcia, 2009). This confusion matrix is shown in Table 1.

Table 1. Confusion Matrix.

		Predicted Class	
		0	1
Actual Class	0	True Negative	False Positive
	1	False Negative	True Positive

As can be seen from the confusion matrix a predicted value can be: a True Positive (TP), a False Positive (FP), a False Negative (FN) or a True Negative (TN). In case of a TP or a TN, the case is predicted correctly and in case of a FP or a FN the case is predicted incorrectly.

The *accuracy* of a model is simply the ratio of the correctly classified cases to the total number of cases. Therefore it can be calculated with the following formula (He, and Garcia, 2009):

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

The *precision* is the ratio of the correctly classified positive cases to the total number of cases that are predicted as positive. This means precision tells something about how often the model is correct, when it predicts a case to be positive. It can be calculated with the following formula (He, and Garcia, 2009):

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

On the other hand, *recall* is the ratio of the correctly classified positive cases to the total number of actual positive cases. This means recall is actually a measure of how complete the results are, i.e. how many of the positive cases does the model retrieve. The formula for recall is (He, and Garcia, 2009):

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

When the data is unbalanced, which is frequently the case in customer churn prediction, accuracy is not a very reliable performance measure (He, and Garcia, 2009). To clarify this, consider the example of a very poor model which simply predicts that no customer churns. This means the model correctly classifies all the people who do not churn, so the amount of true positives is very large. All the people who do churn are classified incorrectly, however this amount is so small compared to the people who churn that the accuracy of the model will still be very high.

However, when one uses the recall and precision as performance measures in this case, one will find that, since there are no true positives, precision and recall will both be 0. This is because recall and precision emphasize the importance of the model predicting the positive cases correctly, instead of predicting the negative cases correctly.

Thus, in a classification model there is a tradeoff between predicting the positive cases correctly and predicting the negative cases correctly. The latter can be measured with the specificity of a model. This tradeoff can be plotted with the Receiver Operating Characteristic (ROC) and the area under this curve is often used as a performance measure.

2.6.2 Area under the ROC curve (AUC) and Area under the Kappa curve (AUK)

Many modelling techniques, including Logistic Regression, give a probability estimate of a case belonging to the positive or negative class. By defining a threshold value, which is often 0.5, it can be determined whether the observation is predicted as occurring (1) or not occurring (0). While the accuracy, recall and precision are calculated for one chosen threshold, a *ROC curve* plots the complement of the specificity against the recall of a model for different thresholds (He, and Garcia, 2009). An example of a ROC curve is shown in Appendix B Figure 1. The ROC curve is independent of the class distribution (Kaymak et al., 2012). The dotted line in the figure in the appendix shows a random classifier. For a random classifier the area under the curve is equal to 0.5, while for a perfect classifier this value is equal to 1. This means that classifiers used in practice should have a value as close as possible to 1, but at least higher than 0.5 (Kaymak et al., 2012).

One of the drawbacks of the area under the ROC curve is that, due to the indifference for class distribution, it might be less suitable for highly unbalanced data sets (Kaymak et al., 2012), which is often the case when predicting churn. In order to overcome this drawback, Kaymak, Ben-David and Potharst (2012) came with an alternative for the AUC: the area under the Kappa curve (AUK). In this curve the complement of the specificity is plotted against the Cohen's Kappa values (Kaymak et al., 2012), of which an example is given in Appendix B Figure 2. Cohen's Kappa can be defined as follows (Kaymak et al., 2012):

$$Cohen's\ Kappa = \frac{Accuracy - P_{chance}}{1 - P_{chance}}$$

Here P_{chance} is the probability that a class is predicted correctly due to chance.

This means Cohen's Kappa is actually an indication for how good the model performs compared to chance. Just as the area under the ROC curve is a performance measure for the model, the area under the Kappa curve can also be a performance measure. However, in contrast to the AUC, AUK emphasizes the importance of predicting the positive classes correctly, which makes it a good performance indicator for models tested on unbalanced data (Kaymak et al., 2012).

2.7 Feature Selection Methods

In section 1.3.1 it was explained that feature selection is often necessary to reduce the curse of dimensionality. Different types of feature selection methods exist. Based on the search strategy that they use, they can be categorized into three different types. These three types are: *filter methods, wrapper methods and embedded methods* (Li et al., 2016).

2.7.1 Filter Methods

Filter methods rank the independent features of a model according to some feature evaluation criteria. This can for example be a correlation coefficient or the mutual information criterion. These criteria can be seen as measures for variable importance. After this ranking is done, the top x variables are selected for the model, while the other variables are filtered out. The ranking of the features is *independent of any learning algorithms*, which makes filter methods generally computationally efficient (Li et al., 2016).

Mutual Information is one of the widely used measures for the dependency between variables. When considering two discrete random variables, say x and y, their mutual information can be calculated when their probabilistic density functions, $p(x,y)$, $p(x)$ and $p(y)$ are known. The mutual information is then calculated as follows (Peng, et al., 2005):

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information thus shares the amount of information shared by x and y together, i.e. how much knowing one of the variables, reduces uncertainty about the other (Li et al., 2016).

2.7.2 Wrapper Methods

In contrast to filters, wrappers do not rank the variables based on a criterion. Instead, a wrapper method takes into account the predictive performance, which makes it *dependent on the learning algorithm*. A wrapper method consists of two steps: 1) it searches for a subset of features and 2) it evaluates the model that is built with the selected feature subset. These steps are repeated until some stopping criterion is met. The feature subset that leads to the highest performance, makes the final feature subset. A disadvantage of wrapper methods is that, since they are dependent on the learning algorithm and the feature search space is often large, they are often computationally expensive. To reduce this search space, different search strategies can be used that lead to local optimum performance, for example sequential search, hill climbing search and genetic algorithms (Li et al., 2016).

2.7.3 Embedded Methods

Embedded methods are a tradeoff between the filter and the wrapper methods. They are dependent on the learning algorithm, but are far more efficient than wrapper methods. The reason for this is that they do not evaluate each feature subset repeatedly. The regularization models are well known examples of embedded methods. The most widely used regularization model is the l_1 –

norm, or Lasso, regularization. Lasso stands for ‘Least Absolute Shrinkage and Selection Operator’ and penalizes the regression coefficients by a shrinkage parameter λ, such that the coefficients of some input variables might shrink to zero and can be eliminated. The goal of this penalizing is to reduce overfitting of the model (Li et al., 2016).

Chapter 3 Methodology

This chapter explains which models were developed to answer the main research question and sub-questions and how these models were developed in terms of the experimental setup. The development of these models was done according to the 'Cross-Industry Process for Data Mining' (CRISP-DM). More information about this process can be found in Appendix C.

In Model 0 and Model 1 variables about *customer and policy characteristics* were used. These models were built in order to answer sub-question one, which was about identifying *customer and policy characteristics* that are good predictors for customer churn.

In order to answer sub-question five, which was about how different feature selection methods influence the predictive ability of the models, different feature selection methods were used. As was explained in section 2.2, a literature study was done to get an idea on what type of variables play an important role in predicting churn. When collecting the data, it was strived to include as many of these variables as possible, such that in the end of this study it was possible to compare the findings of this study with the findings in the literature. This means that the data collection process already was a first step in the feature selection process. After the data was collected, different methods were used to check whether a smaller feature subset needed to be chosen. For the base model (model 0) predictors were used that were selected by domain experts, while for the other models different feature selection methods were used. These are discussed in section 3.1.

Model 0: Customer and Policy Characteristics selected by domain experts
Model 1: Customer and Policy Characteristics selected by different feature selection methods

The aim of sub-question two was to check whether the inclusion of data about *customer web page visits* improves the predictive ability of the model. This means that model 1 needed to be extended with data about the customer web page visits. This led to the development of two new models. The first one kept the entire customer database and only checked whether logging in on the 'My Interpolis' page was an indicator for churn. The second model only focused on the customers who had logged in at least once, meaning that their web page visits were known and making it possible to have a closer look on the predictive ability of different types of webpages.

Model 2: Features Model 1 + variable that indicated whether customer logged in or not
Model 2a: Features Model 1 + customer web page visits for customers who have logged in

The aim of sub-question three was to check whether the inclusion of data about *other customer touchpoints* (phone calls, emails and messages on social media) leads to an improved predictive ability of the model. Again, model 1 was extended, this time with data about these customer touchpoints. Model 3 kept the entire customer database and used variables that indicated whether a customer has contacted Interpolis or has filed a complaint. After this, the customers who had contacted Interpolis were selected and features about these contact moments were added. This led to Model 3a. The customers who had filed a complaint were selected for Model 3b, in which variables about these complaints were added.

Model 3: Features Model 1 + variables that indicated whether a customer has contacted Interpolis or has filed a complaint
Model 3a: Features Model 1 + touchpoint data for customers who have contacted
Model 3b: Features Model 1 + touchpoint data for customers who have complained

Finally, models were created to answer sub-question four, which was about the further improvement of the model when adding *NPS related variables* to the model. The NPS used in this study is of the type '*Experience NPS*', meaning that it was measured after a contact moment or complaint. Therefore, in this case the models that include the NPS data were not extensions of model 1, but of model 3a and 3b. First, a distinction was made between customers who filled in the survey and customers who did not, to check whether this was an indicator for churn. Subsequently, the customers who filled in the survey were selected and information about the NPS score that they gave was added to the models. This led to the development of four different models.

Model 4a: Features Model 3a + a variable that indicated whether customers who have contacted Interpolis, filled in the NPS survey
Model 4b: Features Model 3b + a variable that indicated whether customers who have filed a complaint, filled in the NPS survey
Model 4c: Output Model 3a + NPS data for customers who have contacted and gave a NPS score
Model 4d: Output Model 3b + NPS data for customers who have filed a complaint and gave a NPS score

Sub-question five was about how different feature selection methods influence the predictive ability of the customer churn models. In order to answer this research question, the models 1 to 4d were developed for each different feature selection method. This is explained into more detail in section 3.1.

Sub-question six was about how the model can support decision making in the marketing department. Therefore, answering this research question did not lead to the development of a new model, but was more about the comparison and implementation of the different models.

In summary, eleven different models were developed using datasets that included different amounts and types of customers. An overview of these different models and their relation to each other is shown in Figure 2.

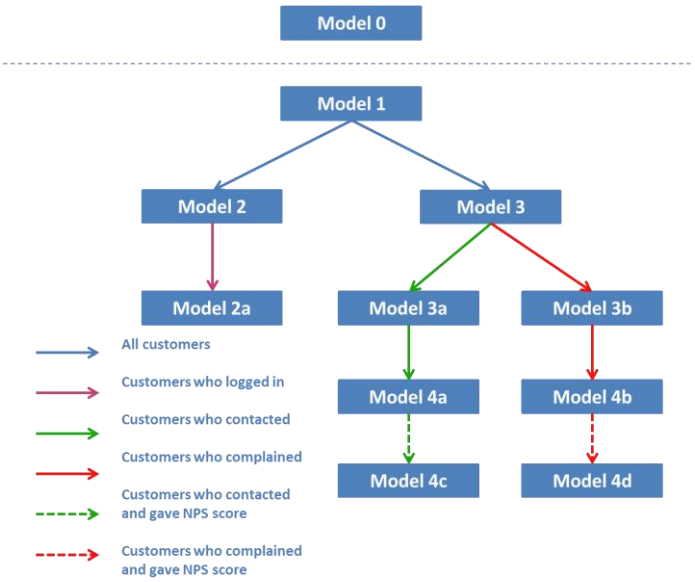


Figure 2. Overview of different models and their relation to each other.

The colors of the arrows in Figure 2 indicate on which datasets the models were build. In summary there were six different datasets, namely: one that included all customers, one that included customers who logged in, one that included customers who have contacted Interpolis, one that included customers who have filed a complaint, one that included customers who had contacted Interpolis and filled in the NPS survey and, finally, one that included customers who had filed a complaint and filled in the NPS survey.

The remainder of this chapter discusses the experimental setup that was followed in order to develop the models.

3.1 Experimental Setup

For the development of the models, *Logistic Regression*, with different types of feature selection, was used. This was done by using one of the well-known open source programs 'RStudio', which uses the R programming language for statistical computing and graphics.

Logistic regression was chosen because it has several strong advantages. First of all, it is simple to understand and easy to implement. Secondly, it gives insight into the importance and significance of the different variables. This is probably why logistic regression seems to be one of the most popular methods in a churn context (Günther et al., 2014). Finally, logistic regression is computationally fast, which makes it feasible to use feature selection methods that require the building of many different models.

For Model 0, four different train sets were used, namely: one in which the churn: non-churn ratio was 50:50, one in which this ratio was 20:80, one in which this ratio was 30:70 and one in which this ratio was 40:60. These train sets were created by randomly undersampling the non-churn population. The reason for this is that datasets that are used for predicting churn are often highly imbalanced and training the model on unbalanced data often reduces the performance of the model (Burez and Van Den Poel, 2009). The churn:non-churn ratio that led to the highest performance for Model 0, was chosen as train ratio for the other models.

In Model 0 features were selected by *domain experts* from the 'Consumer Marketing' department. This was done with the help of a survey, which is shown in Appendix D. A feature was selected when more than half of the experts indicated that they expected this feature to have a significant effect on customer churn.

In the remaining models features were selected by using four different types of feature selection methods. The first method focused on the *significance* of the variables. It was a backward approach in which all variables were added to the model and subsequently the variables that were the least significant were dropped one by one, until a model was reached in which all variables were significant at the 0.05 level.

The other feature selection methods that were used can be categorized into: a filter, a wrapper and an embedded method. All these methods had the same optimization criterion, namely maximizing the 'Area under the ROC curve' (AUC). This optimization criterion was chosen, because AUC takes into account correct classification of both the positive as well as the negative cases at different threshold values, as was explained in section 2.6.2. Moreover, AUC is better known than the 'Area under the Kappa Curve' (AUK).

In this case, the *filter* method ranked the variables based on the Mutual Information criterion. As was mentioned in section 2.7.1, mutual information is one of the widely used measures to define dependency of variables (Li et al., 2016). A reason for this might be that, in contrast to correlation, mutual information does not require linear associations, making it more generally applicable. In order to calculate the mutual information with RStudio, the variables needed to be discrete. The variables that were not categorical yet, were discretized with equal width binning and with the number of bins equal to three. After the ranking was done, the filter method started by building a model with the variable that had the highest Mutual Information value and then adding the variable with the second highest Mutual information value, and so forth. It was checked which number of variables led to the highest value for the AUC.

The aim of the *wrapper* method was similar to that of the filter method: to optimize the AUC of the model. However, as was explained in section 2.7.2, wrappers do not rank the variables based on a criterion, they immediately start with building different models. In this case a forward search was done, in which first all possible models with one variable were build. Subsequently, the model that led to the highest AUC was chosen. After this the remaining variables were added one by one such that the best model with two variables was found, and so forth. This process was continued until the performance of the model did not further increase.

The final method, which falls into the category *embedded* methods, is called the least absolute shrinkage and selection operator (Lasso) method. As was explained in section 2.7.3, this method penalizes the regression coefficients by a shrinkage parameter λ, such that the coefficients of some input variables might have shrunk to zero. In this study, a value for λ was chosen that maximized the AUC value. For this λ, the regression coefficients were checked and the variables that had coefficients larger than zero were selected for the model.

So in summary four different feature selection methods were used in this study: a method that focuses on the significance of the variables, a filter method, a wrapper method and an embedded method.

The experimental setup of the filter, wrapper and embedded methods is a little different from the feature selection method that focuses on the significance of the variables. The reason for this is that the filter, wrapper and embedded methods require optimization of a parameter, which is the number of features for the filter and wrapper method and the lambda for the embedded method. This optimization process requires, apart from a train and a final test set, a validation set. In order to get robust results 10 fold cross validation was used. This process is shown in Figure 3.

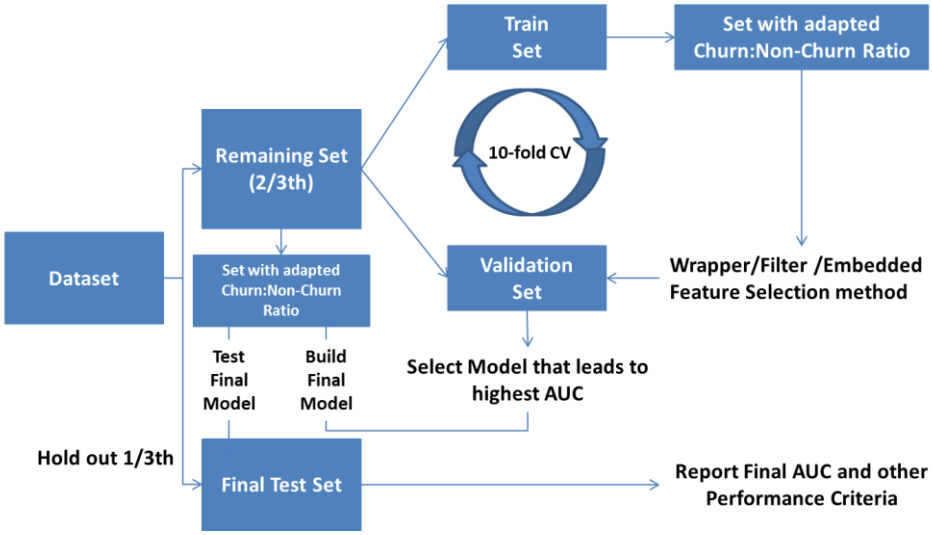


Figure 3. Overview of the experimental setup of the filter, wrapper and embedded feature selection methods.

Figure 3 shows that first a final test set was extracted from the dataset, which was equal to approximately 1/3th of the entire dataset. This is necessary in order to overcome overfitting and was done with stratified random sampling. The remaining set was split into a train set and a validation set, which was replicated with 10 fold cross validation. Tenfold cross validation was chosen because extensive tests have shown that this method results in the best error estimate(Witten et al., 2016). This means that ten train and ten validation sets were created. In this case, the validation sets were necessary to optimize the parameters. The train sets were then changed such that the ratio of customers who churned to customers who did not churn was the same as the ratio that led to the best performance for Model 0. In order to have a good representation on how a model performs on the final test set, the validation sets were kept into the original churn/non-churn ratio, just like the final test set. For each of the ten train sets, features were then selected with either the filter method, wrapper method or embedded method. For each method, these ten feature subsets were compared and a subset was chosen that was expected to lead to the highest AUC value. Chapter 5 gives more insights into how these subsets were chosen. The model was then trained on the entire dataset, except for the holdout set, in order to maximize the use of data for training. The churn/non-churn ratio of this large train set was also changed to the ratio that led to the best performance for Model 0.

Finally, the model was tested on the final test set. The evaluation criteria that were used in this study are: *accuracy, recall, precision, the area under the ROC curve (AUC) and the area under the Kappa curve (AUK)*. These criteria were explained in section 2.6. When the models were compared in this study, the most important criteria was the Cohen's Kappa value. The reason for this is that, in

agreement with domain experts, it was concluded that the aim of the model is to identify potential churners, but without getting too many false positives. This is because of the so-called 'wake-up effect', meaning that customers who did not have the potential to churn but of which the model predicts churn, are reminded of the fact that they can switch to a different insurer when retention efforts are done. This issue is explained into more detail in Chapter 6 'Practical Implications'. This makes both recall and precision important measures. However, since there is a tradeoff between these variables, it is sometimes difficult to choose the best model based on these two measures. Cohen's Kappa takes into account both these measures, but emphasizes correct prediction of the positive cases. This makes Cohen's Kappa a good criterion for this case. The AUC and AUK are also important because, in contrast to recall, precision and Cohen's Kappa, they measure the overall performance of a model and not just the performance at one threshold, as was explained in section 2.6.2.

The experimental setup of the feature selection method that uses the significance of the variables is shown in Figure 4. This experimental setup is less complex than the one of the filter, wrapper and embedded method, since cross validation is not needed here.

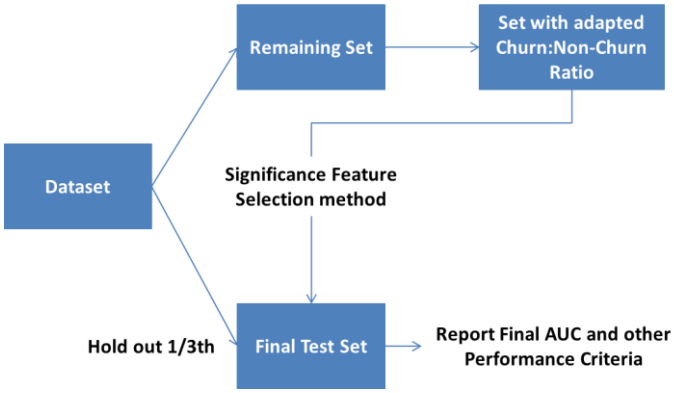


Figure 4. Overview of the experimental setup of the method that focuses on the significance of the variables.

One difficulty in extracting the final test set from the datasets, was that the models were built on sets that included different types of customers, which was indicated by the different colors of the arrows in Figure 2. This influences the comparison of the models, since the test sets of the different datasets cannot exactly include the same customers. For example, a customer who did not complain cannot be part of the 'Customers who complained' test, but he can still be in the test set that included all the customers.

In order to compare the models in a reliable way, it was important that the test sets were at least partly overlapping. If this would not be the case, the possibility existed that one model seemed to outperform another model, while this was just because the test set was accidentally working in favor of that one model.

This problem was solved by using the policy numbers of the 'All customer' test set to determine the test sets of the other datasets. For example, the policy numbers that appeared in the 'All customer' test set and in the 'Customers who have contacted' dataset were used to form the 'Customers who

have contacted' test set. This process was the same for the other datasets. This is shown in Figure 5, indicated by 'Test Set X' and 'Remaining Data X'.

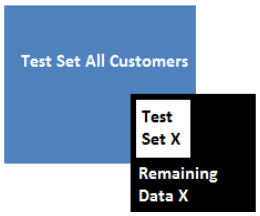


Figure 5. Representation of the different test sets.

For all these test sets, it was checked whether this led to a size equal to approximately 1/3th of the entire dataset. If this was by far the case customers were selected randomly and were added to the test set until the amount was equal to approximately 1/3th of the entire dataset.

Another option to overcome the problem of non-overlapping test sets was to divide the datasets based on the policy numbers. For example, to select all the policy numbers from 1 to 100 for the test set and use the policy numbers from 101 to 300 for the remaining set. However, in this case the policy numbers were strongly correlated with the start date of the contract, so this could lead to poor model performance. Therefore this option was rejected.

Chapter 4 Data Understanding & Data Preparation

In this chapter the 'Data Understanding' and 'Data Preparation' steps from the CRISP-DM Process are discussed. These steps were done prior to the modelling step.

4.1 Data Understanding

The 'Data Understanding' step started with the collection of the data. Data was mainly collected from the year 2015, however to determine which customers churned it was checked whether the policy numbers from 2015 still existed in the 2016 dataset. The datasets that included customer and policy characteristics and the webpage visits were delivered as comma separated value (csv) files by a data/web analyst from the 'Marketing Intelligence' department. The datasets that included the other customer touchpoints (phone calls, emails and messages on social media), the complaints and the NPS were delivered as csv-files by an expert from the 'Customer Insights' department. An overview of the different datasets is given in Table 2.

Table 2. Overview of the different datasets.

Dataset	Record/row	Year
Customer and Policy Characteristics 2015	Customer	21st of October 2015
Customer and Policy Characteristics 2016	Customer	31st of January 2016
Webpage visits	Webpage visit	2015
Other customer touchpoints and corresponding NPS	Contact moment	2015
Complaints and corresponding NPS	Complaint	2015

The remainder of this section focuses on the variables that were included in the different datasets.

4.1.1 Customer and Policy Characteristics

The variables that were included in the 'Customer and Policy Characteristics' datasets are given in Table 3. The datasets for the years 2015 and 2016 included the same variables, however for the 2016 dataset the only relevant variable was 'Policy number'. This variable was later used to determine whether a customer churned or not, as will be explained in section 4.2.5.

Table 3. Variables that were included in the 'Customer and Policy Characteristics' datasets.

Variable Name	Type of Variable	Description
Policy number	String	*A unique identification number that is given to every customer and his corresponding policy*
Basic insurance	Categorical (dichotomous)	*Whether a customer has a basic insurance policy at Interpolis (value 1) or not (value 0)*
Voluntary Excess Risk	Categorical	*The height of the voluntary excess risk a customer has. This can either be: €0, €100, €200, €300, €400 or €500.*
Additional insurance	Categorical	*The type of additional insurance a customer has. This can be: none, type 1, type 2, type 3 or 'other'.*
Dental insurance	Categorical	*The type of dental insurance a customer has. This can be: none, coverage up to €250, coverage up to €500, coverage up to €1,000 or coverage up to €1,250*
Collection period	Categorical	*Whether a customer pays his premium on a yearly (value 12) or monthly (value 1) basis*
Start Date	Date	*The date on which the policy started*
Date of Birth	Date	*The date on which the customer was born*
Premium	Numeric	*The actual amount of premium a customer pays in €*
Gross Premium	Numeric	*The amount of premium without discounts in €*
Collectivity Discount	Numeric	*The amount of collectivity discount a customer gets in €*

Considering the additional insurance, Interpolis has three main types of additional insurance and several additional insurances that are based on collectivity. The three main additional insurances differ in type of coverage and have a cumulative nature.

4.1.2 Webpage visits

The dataset that included the webpage visits consisted of a lot of different variables. Most of these variables had a very technical nature, for example 'Browser Width' and 'Browser Height' and were not expected to predict customer churn. Therefore only the most important variables are described in Table 4.

Table 4. Variables included in the 'Webpage visits' dataset.

Variable Name	Type of Variable	Description
Policy number	String	*A unique identification number that is given to every customer and his corresponding policy*
URL	String	*The webpage URL that was visited*
Previous webpage	String	*The webpage URL that was previously visited and linked to the visited URL*
Load date and Time	Date & Time	*The date and time on which the web page was loaded*

Considering the URL's, they included all the different URL's that Interpolis manages, ranging from the homepage to the page where a customer can invoice his healthcare costs.

4.1.3 Other customer touchpoints and corresponding NPS

Apart from the webpage visits, data from other touchpoints were available. This dataset included information about all the incoming phone calls, emails, messages on social media and a corresponding NPS if this was given by the customer. This means that the customer was always the *initiator* of the touchpoint. The NPS which is given in this case, was of the type '*Experience NPS*', as was discussed in section 2.3. Table 5 gives an overview of these different variables in more detail.

Table 5. Variables included in the 'Other customer touchpoints and corresponding NPS' dataset.

Variable Name	Type of Variable	Description
Policy number	String	*A unique identification number that is given to every customer and his corresponding policy*
Contact Date	Date	*The date on which the contact moment took place*
Contact Identification Number	String	*A unique identification number that is given to every moment of contact*
Contact Channel	Categorical	*The type of channel the customer used to contact Interpolis. This can be: phone, e-mail or a message on social media*
First Time Right Potential	Categorical (dichotomous)	*Whether the issue for which the customer has contacted Interpolis can be solved in this one moment of contact (value 1) or whether more contact moments are needed for this (empty cell)*
Not First Time Right	Categorical (dichotomous)	*In case the contact moment had the potential to be solved in that one contact moment, this variable indicates whether this was the case (empty cell) or not (value 1)*
NPS Survey: NPS Score	Categorical	*The Net Promoter Score the customer gave for the given moment of contact*

4.1.4 Complaints and corresponding NPS

The dataset 'Complaints and corresponding NPS' included different variables about the complaints that the customers have filed. All customers get a comprehensive survey about the handling of the

complaint. Because this survey was too large to include all questions in the model, the focus was on the NPS related variables, which were again of the type '*Experience NPS*' as was discussed in section 2.3. An overview of these variables is given in Table 6.

Table 6. Variables included in the 'Complaints and corresponding NPS' dataset.

Variable Name	Type of Variable	Description
Policy number	String	*A unique identification number that is given to every customer and his corresponding policy.*
Complaint Date	Date	*The date on which the customer has filed the complaint.*
Type of Complaint	Categorical	*The complaint can be of the type ' Regular' or ' Reassessment'. A reassessment takes place when the customer is not satisfied with the outcome of the regular complaint he had filed.*
Assigned	Categorical (dichotomous)	*Whether the complaint is assigned (value 1) or not (value 0).*
Handled Personally	Categorical (dichotomous)	*Whether the complaint is handled personally (value 1), meaning by phone or not (value 0), meaning by email or a different channel.*
Lead Time First Reaction	Integer	*The number of working days that passed before the customer got a first reaction about the filing of his complaint.*
NPS Survey: NPS score	Categorical	*The Net Promoter Score the customer gave for the process of handling the complaint.*
NPS Survey: Explanation NPS score	String	*An explanation for why the customer gave that specific NPS score.*

4.2 Data Preparation

In the 'Data Preparation' step the data was prepared in a way that it could be used for modeling purposes. 'Real life' data can contain all sorts of pollution, such as discrepancies, missing values, outliers or input errors, which will lead to poorer model performance in the end. Good data preparation enhances the chance of building models with good performance.

4.2.1 Duplicate values

The different datasets were each checked on duplicate values.
In the 'Customer and Policy Characteristics' dataset each row represented a customer with his policy number as a unique identifier. Therefore each policy number could only occur once, which was found to be the case after checking for duplicates.

In case of the 'Webpage visits' dataset each row represented a web page visit. However, these web page visits can be part of a session, in which multiple webpages were visited. In this case, these rows do not have one unique identifier, but the combination of policy number, load time and URL has to be unique in order to be valid. This seemed to be the case for all visits.

The 'Other customer touchpoints and corresponding NPS' dataset consisted of different contact moments. Each contact moment had an identification number, therefore this number can only occur once in the dataset. This identification number was checked for duplicate values and it was found that there existed 6.880 duplicate values for this variable. According to an expert, this is caused by switching to a different CRM system in 2015. Therefore for each duplicate, one value was removed from the dataset. However, the system that was used also allowed one moment of contact to have multiple rows. For example when a customer called about multiple subjects, it could happen that this was registered in two rows, which made it difficult to interpret the dataset. Even after removing

the duplicate values for the identification number, this still seemed to be the case. This issue is further discussed in section 4.2.5.

Finally, in the dataset that includes the complaints and corresponding NPS each row represents a complaint. A complaint does not have a unique identification number, therefore it was checked whether there exist rows in the dataset that have the same value for all variables as discussed in Table 6. This was not the case, so it was concluded that no duplicates existed.

4.2.2 Missing values

The datasets contained several variables that have missing values. The names of these variables and the corresponding amount of missing values is given in Table 7.

Table 7. Variables that have missing values and their corresponding amount of missing values.

Dataset	Variable name	Amount of missing values
Customer and Policy characteristics 2015	Voluntary Excess Risk	1.20%
	Date of Birth	0.05%
Complaints and corresponding NPS	Policy number	1.14%

The NPS survey is only filled in by a small fraction of the customers who have contacted or have filed a complaint. For the customers who have contacted Interpolis, 1,2% filled in this survey and for the customers who have complained, 12,7% filled in this survey. However, this problem was solved by developing different models for customers who filled in the survey as is shown in Figure 2.

For the remaining variables in Table 7, there was no reason to believe that these missing values have a non-random nature. Because of this and the fact that the amount of missing values is so small, the missing values were ignored. This means that the cases that have missing values for at least one of the variables in Table 7 were removed.

4.2.3 Case Selection

Considering the case selection, several decisions were made on which customers were included in the dataset and which customers were not.

The first decision dealt with the customers who have *no basic insurance* at Interpolis. The percentage of customers who only have additional or dental insurance, but no basic insurance is very small. According to experts, most of these customers have very specific or uncommon healthcare costs and therefore choose their additional or dental insurance very carefully. Their decision process to churn or not might therefore be very different from customers who do have basic insurance. Moreover including customers without a basic insurance makes the distribution of the variable 'Premium' and 'Gross Premium' very large, since in this case also very small premium amounts are possible. Because of causing these biases, the customers who do not have basic insurance were removed from the dataset.

The second decision for case selection deals with the *age* of the customers. In the Netherlands, no premium needs to be paid for children younger than 18 years old. Children are often insured at the same insurer as one of their parents and often do not make this decision themselves. Moreover, children will not be contacted for any retention attempts. For the purpose of this model, children younger than 18 years old were therefore not selected.

The third decision for case selection deals with customers of which their *policy* did not *start* in the month *January*. In section 1.1.1 "The Dutch Healthcare Insurance System", it was explained that in the Netherlands people can close a new healthcare insurance contract between the 1st and 31st of January. There are only some exceptions on this. People that did not become a customer in January, and thus met one of the exception criteria, might have a very different reason why they have chosen for Interpolis, than people whose contract started in January. Therefore these customers were not selected for the model. In order to remove these customers, the variable 'Start Month' was transformed into a variable named '*Regular Inflow*'. This variable had the value 1 when the policy of a customer started somewhere between the 1st and 31st of January and the value 0 when his policy started at a different date. After this all the customers with a value of 0 for 'Regular Inflow' were removed from the dataset.

The fourth decision for case selection is based on the variable *'Previous webpage'* from the 'Webpage visits' dataset. In some cases, it can happen that an employee from Interpolis visits the webpage on behalf of the customer in order to answer his questions. The variable 'Previous webpage' shows when this is the case. The webpage behavior of an employee might be very different from that of a customer and therefore these cases were not selected.

Finally, in the dataset 'Other customer touchpoints and corresponding NPS' the variable *'Contact Channel'* sometimes has the value *'Chat'*, which is not a valid channel for Interpolis. According to an expert, it is likely that these are input errors, since this value only occurs in 0,05% of the cases and employees fill in the value themselves. Therefore cases for which the 'Contact Channel' is 'Chat' were removed from the dataset.

4.2.4 Outliers

There is the possibility that there are outliers present in the continuous data. Therefore, the variables: 'Premium', 'Gross Premium', 'Collectivity Discount' and 'Lead Time First Reaction' were checked for outliers. This was done by creating a boxplot for each variable. These are shown in Appendix E Figures 1 to 4.

From the boxplots it can be seen that there were outliers present in these variables. However, an explanation for the large distribution of the variables 'Premium', 'Gross Premium' and 'Collectivity Discount' is that these variables are sometimes given on a monthly basis and sometimes on a yearly basis, depending on the collection period. These variables were therefore transformed such that they were all on a monthly basis. This was done by dividing the premium by 12, if the collection period was equal to 12 months and doing nothing if the collection period was equal to 1 month.

After these transformations, the boxplots were created over again. The new boxplots are shown in Appendix E Figures 5 to 7. Although there seem to be less outliers in the boxplots now, there are still several outliers present. According to experts, amounts between €73.12 and €174.65 were possible for the year 2015 (Interpolis, 2014). The premium is equal to €73.12 when a customer has only the basic insurance with maximum voluntary excess risk. In case a customer is 65 years or older, has no voluntary excess risk, has no collectivity discount and has both the additional and dental insurance with maximum coverage the amount is equal to €174.65.

The 11 cases that have a premium and gross premium higher than €800 are probably input errors and are therefore removed from the dataset. After this is done, the maximum amount of premium

and gross premium is equal to €162.54 and €174.65 respectively. The values below €73.12 were all customers who turned 18 in 2015, which means that they only pay premium in the months were they were 18. This explains how it is possible that these values are lower and therefore these outliers were not removed. The outliers for the variable 'Collectivity Discount' were also checked with the help of an expert and it was concluded that these were not input errors and were therefore not removed.

The outliers for the variable 'Lead Time First Reaction' were checked and it was concluded that these values still fall within the boundaries of possible values. When a customer experiences high lead times, this might be a probable reason for him to churn. Therefore it was decided that the outliers of these variables were not removed.

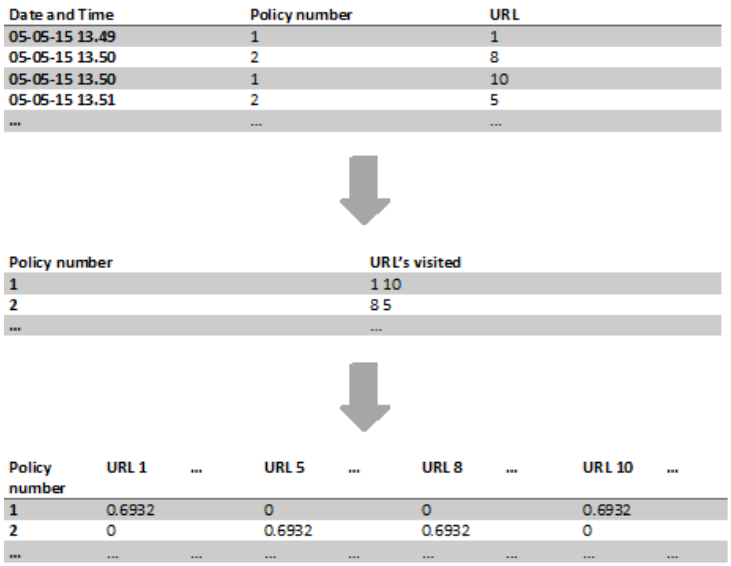
4.2.5 Transformation of variables

In order to merge the datasets they needed to be transformed in such a way that in each dataset a row represented a customer. The datasets could then be merged by the variable 'Policy number'. This means that the datasets that includes the webpage visits, other customer touchpoints and complaints needed to be aggregated to customer level. This was done by creating several new variables.

First, variables were created that indicate: 1) Whether the customer has logged in on the secured environment or not, 2) Whether the customer has contacted Interpolis or not and 3) Whether the customer has filed a complaint/reassessment or not. These variables were needed as input for model 2 and 3. Experts expected that the *timing* of the log in moment or contact moment might also be important when using these variables as predictors for churn. At the end of the year, when the potential moment to switch is getting closer and the campaign is starting, customers might think more about switching. The activities that will take place at this moment are therefore expected to have a bigger influence on churn. Therefore for the variables about logging in and contacting a distinction was made between 1) *Between January and October* or 2) *In November or December*. For the variable that indicated whether the customer had filed a complaint or reassessment this distinction was not made. The reason for this is that the fraction of customers who file a complaint is very small and dividing this data into two different time frames, would lead to even smaller datasets. Secondly, the timing of when a customer has filed a complaint is expected to be of less importance in predicting churn than the timing of logging in or contacting.

For models 2a, 3a and 3b more detailed information about the web page visits and other touchpoints was included. For the webpage visits, first a variable was created that indicated which webpages (URL's) a customer visited. For example a value of '1 5 64 5 73', means that the customer visited URL's 1, 64 and 73 one time and visited URL 5 two times. On these concatenated values the TF/IDF technique was used, which was explained in section 2.5. The reason why TF/IDF is used in this case, is because in this way the distinctiveness of webpages in predicting churn is emphasized. For example, when a customer visits webpage x many times and other webpages zero times, webpage x should have a different weight than it has for a customer who visits webpage x many times, but also visits the other webpages many times. For the first customer webpage x is a lot more distinctive than it is for the second customer. In a similar way, webpages that are visited by only a fraction of the customers might be more distinctive than webpages that are visited by all the customers.

This TF/IDF transformation led to a matrix, in which each row represented a customer and each column represented a URL with a TF/IDF value for that specific URL. Figure 6 clarifies this process in a simplistic way.



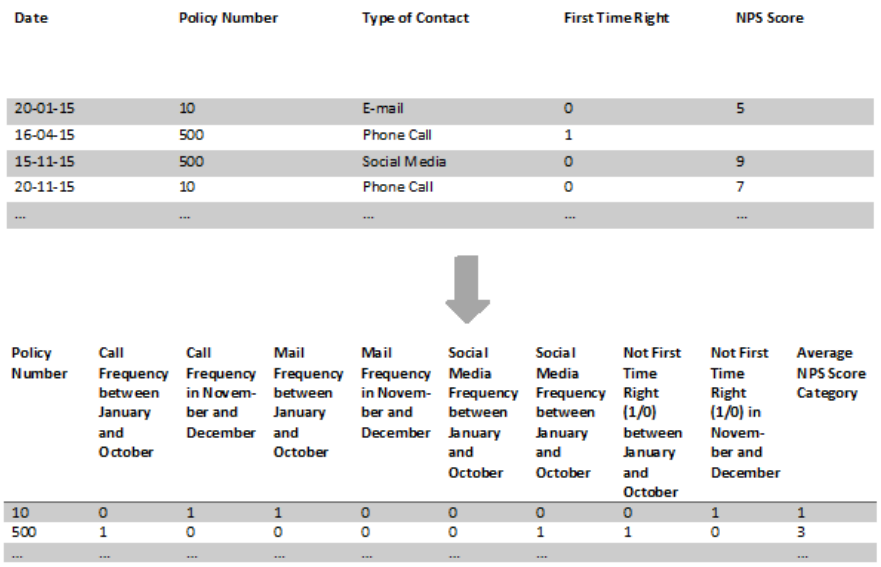
Date and Time	Policy number	URL
05-05-15 13.49	1	1
05-05-15 13.50	2	8
05-05-15 13.50	1	10
05-05-15 13.51	2	5
...	...	...

Policy number	URL's visited
1	1 10
2	8 5
...	...

Policy number	URL 1	...	URL 5	...	URL 8	...	URL 10	...
1	0.6932		0		0		0.6932	
2	0		0.6932		0.6932		0	
...	...	...	...	...	...	...	...	...

Figure 6. Process of TF/IDF transformation for webpage visits.

From the 'other touchpoints and corresponding NPS' dataset, the variables were created that were needed for Model 3a, 4a and 4c. These were categorical variables that indicated: 1) how often a customer called, 2) how often a customer emailed, 3) how often a customer has sent a message on social media, 5) whether the customer had a contact moment that was not first time right, and 6) the average NPS score. For the first four variables, again, a distinction was made between the two time frames: Between January and October, and November and December. A summary of these changes in the dataset is shown in Figure 7.



Date	Policy Number	Type of Contact	First Time Right	NPS Score
20-01-15	10	E-mail	0	5
16-04-15	500	Phone Call	1	
15-11-15	500	Social Media	0	9
20-11-15	10	Phone Call	0	7
...	...	...	...	...

Policy Number	Call Frequency between January and October	Call Frequency in Novem-ber and December	Mail Frequency between January and October	Mail Frequency in Novem-ber and December	Social Media Frequency between January and October	Social Media Frequency between January and October	Not First Time Right (1/0) between January and October	Not First Time Right (1/0) in Novem-ber and December	Average NPS Score Category
10	0	1	1	0	0	0	0	1	1
500	1	0	0	0	0	1	1	0	3
...	...	...	...	...	...	...			...

Figure 7. Changes in the 'Other touchpoint and corresponding NPS' dataset.

Before the categorical variables were obtained, the number of phone calls, e-mails and messages on social media were counted. However, as was mentioned in section 4.2.2, the dataset that contained the moments of contact was difficult to interpret. The reason for this was that the system that was used for logging the moments of contact, made it possible that one moment of contact could have multiple rows in the dataset. This problem was partially solved by categorizing these variables into three groups, which makes these variables less sensitive to the poor quality of the data. This categorization was done according to the distribution of the phone calls, e-mails and social media messages. As can be seen in Appendix F Figures 1 to 6 these distributions were very skewed. Therefore, for each channel and time horizon combination the weighted average was calculated and based on this, three categories were created for each of the six channel and time frame combinations. The first category consisted of the customers who used channel x in time frame t zero times, the second category consisted of the customers who used channel x in time frame t the weighted average plus and minus one (except when this was equal to 0) times, and the third category consisted of the customers who used channel x in time frame t more than the weighted average plus one times.

Apart from categorizing the call frequency, mail frequency and frequency of messages on social media, the average NPS score was categorized. As was explained in section 2.3, the original NPS scale groups the respondents into three categories: customers who answer 9 or 10 are considered *promoters*, customers who answer 7 or 8 are called *passives*, and those who answer 6 or less are called *detractors*. However, this categorization is questioned many times. First of all, because the respondents giving the rating six are part of the detractor group, even though they are above the midpoint of the rating scale. Secondly, the categorization might be dependent on cultural aspects (Kristensen and Eskildsen, 2014). Some experts suggest that there should be a European version of the Net Promoter Score, since Europeans often give less extreme values than Americans (Checkmarket, 2017). Therefore in this study the average NPS score is categorized as follows: category 1 includes the customers with a rounded average NPS score of 1 to 5, category 2 includes the customers with a rounded average NPS score of 6 or 7, and category 3 includes the customers with a rounded average NPS score of 8, 9 or 10.

From the 'Complaints and corresponding NPS' dataset, variables were created that indicated: 1) whether a customer has filed at least one reassessment or not, 2) whether at least one of a customer's complaint or reassessment was not assigned 3) whether at least one of a customer's complaint or reassessment was not handled personally, 4) whether for at least one of a customer's complaint or reassessment the first reaction was not given within the promised lead-time of 5 working days and 5) the average NPS score (categorical).

Date	Policy Number	Type of Complaint	Assigned	Handled Personally	Lead Time First Reaction	NPS Score
20-01-15	10	Regular	0	1	3	4
11-04-15	500	Reassessment	1	0	10	
05-11-15	500	Reassessment	0	0	12	7
01-12-15	10	Regular	0	1	5	6
...	...	...	...	...	...	...

Policy Number	Reassessment	Not Assigned	Not Handled Personally	Not within promised lead time first reaction	Average NPS Score
10	0	1	0	0	1
500	1	1	1	1	2
...	...	...	...	...	...

Figure 8. Changes in the 'Complaint and corresponding NPS' dataset.

It was chosen to emphasize the negative events, for example to look at whether a customer had a complaint that was not handled within the promised lead time, instead of looking at whether a complaint was handled within the promised lead time. The reason for this is that it is known from psychological studies that people tend to remember negative events better than positive events. Meaning that customer behavior, such as making the decision to churn, is far more likely to be affected by bad experiences than by good experiences (Lax, 2012).

Even though rows in the dataset 'customer and policy characteristics' already represented a customer, transformations needed to be done in this set as well.

One very important variable that was created was the dependent variable of the model, namely whether a customer churned or not. This variable was created by taking the policy numbers of the 2015 dataset as a base and checking for each policy number whether it was still available in the 2016 dataset or not. If it was still available it was concluded that the customer did not churn (value 0) and if it was unavailable it was concluded that the customer churned (value 1). This means that it was not possible to make a distinction between the type of churn, i.e. it was not possible to see whether the churn was involuntarily or voluntarily, as was discussed in section 2.1. Although, due to confidentiality, no information can be given about the churn:non-churn ratio, it was concluded that the classes were strongly unbalanced.

Another variable was created from the variable 'Date of Birth', namely the variable 'Age'. In this case, age is defined as the age in which the customer turned in the year 2015. Subsequently this variable was categorized and transformed into the variable 'Age Category'. The categories were defined as followed:

- Category 1: 18 up to and including 25
- Category 2: 26 up to and including 35
- Category 3: 36 up to and including 45
- Category 4: 46 up to and including 55
- Category 5: 56 up to and including 65

- Category 6: 65 and older

Another categorization was done with the variable 'Voluntary Excess Risk'. By exploring this variable, which is shown in Appendix G Figure 1, it was found that most customers choose to have either no voluntary excess risk or the maximum voluntary excess risk that is possible. Therefore the variable 'Voluntary Excess Risk Category' was transformed into three different categories. The first one includes the customers who do not have any voluntary excess risk. The second category includes the customers who have a voluntary excess risk between €100 and €400. Lastly, the third category included the customers who have the maximum voluntary excess risk of €500. Apart from this categorization the variable 'Voluntary Excess Risk' was also transformed into a binary variable, which indicated whether the customer had a voluntary excess risk (value 1) or not (value 0).

Another binary variable was created for the variable 'Collectivity Discount'. This variable, named 'Collectivity Discount Binary', had the value 1 when a customer was collectively insured and therefore got a discount and value 0 when the customer was individually insured and therefore did not get a discount. The reason why this was done, is because all customers get the same percentage of discount and it was therefore expected that the absolute height of the discount was of less importance.

In the 'Data Understanding' section it was mentioned that the variable 'Additional Insurance' has three main groups and several collectively additional insurances. Because the amount of customers who have a collectively additional insurance is rather small, these customers are grouped together and this additional insurance group is called 'Other'. Apart from this categorization, the variable 'Additional Insurance' was also transformed into a binary variable, that indicated whether the customer had additional insurance or not. The same was done for the variable 'Dental Insurance'.

The last variable that was created is the variable 'Contract Duration'. This variable was created from the variable 'Start year' and indicates how long the customer has had his/her healthcare insurance policy at Interpolis uninterrupted. This variable was created by subtracting the start year from 2015. It should be noted here that there is the possibility that customers with a start year of 2006, were already insured privately at Interpolis before the Healthcare Insurance Act was introduced. Subsequently this variable was categorized into groups, namely: 1) a group of customers of which their contract only started that year ('Contract Duration' equal to 0), 2) a group of customers of which their contract started at least 1 year ago, but at most 2 years ago ('Contract Duration' equal to 1), and 3) a group of customers of which their contract started more than 2 years ago ('Contract Duration' > 1).

4.2.6 Feature Selection

Although to a large extent, feature selection was done in the modelling step, the feature selection process already started with the collection of the data. As was explained in Chapter 3, it was strived to include as many variables as possible that were found to be important in predicting churn in the literature. However, as could already been seen in section 4.1.1, some relatively basic features, such as customer gender, could not be collected. The reason for this was the switch to a new information management system and the difficulty in accessing the old database. Moreover, the dependency on Rabobank, which is the sales channel of Interpolis, made it difficult to access data about products other than health insurance. Table 8 shows which variables were suggested from the literature and whether they were available for this study or not.

Table 8. Overview of suggested features from literature and whether they were available.

Feature suggested from literature	Available for this study
Customer Age	✓
Customer Gender	
Marital Status	
Total Premium	✓
Discount	✓
Change in Premium	
Type of Coverage/Insurance Package Type	✓
Total Cost of Claims	
Total Number of Policies	
Duration of Continuous Relationship	✓
Customer Family Configuration	
Customer Income	
Individually or Collectively Insured	✓
Partner Insured at Interpolis	

After the collection of the data was done, the feature selection process continued. For the base model (Model 0) features were selected by domain experts. This was done by sending a survey to the 'Consumer Marketing' team. In this survey the participants were asked whether they expected that a certain variable had a significant effect on whether the customer churned or not in 2015. This was asked for the variables that have a check mark in Table 8. The variable 'Type of Coverage/Insurance Package Type' was split into variables that indicated: whether the customer had additional Insurance and whether the customer had dental Insurance. Moreover, a variable was included that indicated whether the customer had a voluntary excess risk. A feature was selected when it was expected to have a significant effect on churn by more than half of the participants. This was the case for the variables: 'Customer Age', 'Additional Insurance', 'Contract Duration' and 'Premium'. The survey, together with its results can be found in Appendix D.

For Model 1, it was decided to exclude the variables 'Premium' and 'Gross Premium', because these variables depend only on: the age of the customer, the type of additional and dental insurance the customer has, whether the customer is individually or collectively insured and on what date the policy started. Because of this high correlation between these variables, including these variables will lead to high multicollinearity, which is bad for the performance of a Logistic Regression model (Lani, 2010).

4.2.7 Merging the datasets

The datasets were merged with an 'inner join' by the variable 'Policy Number'. After this was done, six different datasets were created, namely: one that included all the customers, one that included customers who logged in, one that included customers who have contacted, one that included customers who have complained, one that included customers who have contacted and have filled in the NPS survey and one that included customers who have complained and have filled in the NPS survey. This was done in order to develop the different models shown in Figure 2. Due to confidentiality reasons, no information is given about the sizes of these different datasets.

4.2.8 Dividing the data into Test, Train and Validation sets

In section 3.1 , called 'Experimental Setup', the process of dividing the datasets into a test set, a train set and a validation set was explained. This was done by using the policy numbers of the 'All customer' test set to determine the test sets of the other datasets. For all these test sets, it was

checked whether this led to a size equal to approximately 1/3th of the entire dataset. Table 9 shows the sizes of the different test sets in percentage of total customers for each particular dataset.

Table 9. Size and overlap of the different test sets.

Type of dataset	Percentage of Customers in Test set
All customers	33.33%
Customers who logged in	32.90%
Customers who have contacted	33.24%
Customers who have filed a complaint	33.53%
Customers who have contacted and have filled in NPS	34.42%
Customers who have filed a complaint and have filled in NPS	34.83%

From Table 9 it can be concluded that the proportion of test data is approximately 1/3th for all the datasets, as was desired. The test sets other than the 'All customer' test set are thus entirely subsets of the 'All customer' test set.

Chapter 5 Results

In this chapter the different models and their performance are discussed and compared. The structure of this chapter is based on Figure 2. Model 0 is discussed first, followed by Model 1, Model 2, Model 2a, Model 3, Model 3a, Model 3b, Model 4a, Model 4b, Model 4c and, finally, Model 4d.

5.1 Model 0: Customer and Policy Characteristics

As was explained in section 3.1 and 4.2.6, in Model 0, features were used that were selected by experts from the 'Consumer Marketing' department. This led to the selection of four features, namely: Age Category, Additional Insurance, Contract Duration and Premium. Therefore a logistic regression model was built which used these four features as predictors. Because the original dataset was highly imbalanced, considering churn and non-churn customers, four different train sets were used to build the model. These were: a train set in which the churn: non-churn ratio was 50:50, one in which this ratio was 20:80, one in which this ratio was 30:70 and one in which this ratio was 40:60. This resulted in four different models, of which the performance indicators are shown in Table 10.

Table 10. Different performance indicators for each train set for Model 0.

Criterion	50:50	20:80	30:70	40:60
Accuracy	0.5895	**0.8989**	0.8718	0.8029
Recall	**0.6370**	0.0024	0.1073	0.2747
Precision	0.1468	**0.4375**	0.2216	0.1830
Kappa	0.0891	0.0036	0.0843	**0.1120**
AUC	0.6479	0.6476	**0.6480**	**0.6480**
AUK	0.0661	0.0660	**0.0662**	0.0661

From Table 10 it can be seen that the model that is built on the 40:60 train set delivers the highest Kappa value and that the AUC and AUK values are very similar for the different models. Therefore the 40:60 ratio is chosen as the train ratio for the remaining models. The model summary of the model that is trained on the 40:60 set can be found in Appendix I Table 1.

5.2 Model 1: Customer and Policy Characteristics

For model 1, features about customer and policy characteristics were used, namely: Customer Age, Voluntary Excess Risk, Additional Insurance, Dental Insurance, Contract Duration and Collectivity Discount Binary. As was mentioned in section 3.1 four different methods were used to decide which features were selected and which were not, namely a method that uses the significance of the variables, a filter method, a wrapper method and an embedded method (Lasso).

First, variables were selected with the use of the method that takes into account the significance of the variables. All six variables were selected, although the variables 'Additional Insurance', 'Dental Insurance' and 'Voluntary Excess Risk' required changes in the way they were categorized. For the additional insurance, customers with the least extensive and with the most extensive additional insurance were grouped together with the customers who had no additional insurance, because there were no significant differences between these categories. For the same reason, the customers with the least extensive dental insurance were grouped together with the customers with no dental insurance. And, finally, the customers with a voluntary excess risk up to and including €400 were grouped together with the customers who had no voluntary excess risk. These adapted variables were also used in the following models.

The filter method first ranked the features based on the mutual information criterion. Features were then added to the model one by one and for each iteration the area under the ROC curve was calculated. The result of this is illustrated by the graph in Figure 9. The different colors of the lines represent the different folds that resulted from ten-fold cross validation.

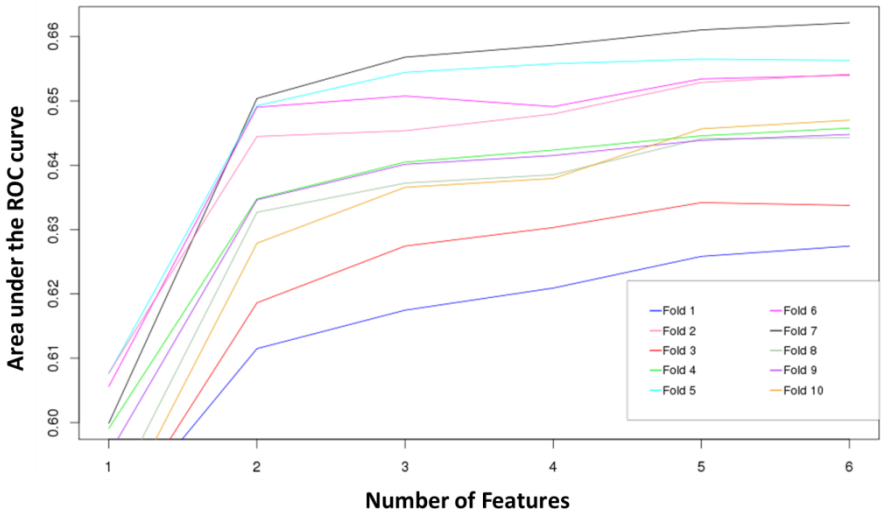


Figure 9. The area under the ROC curve vs. the Number of Features for each fold for Model 1.

Figure 9 shows that the increase in performance after the second feature is included, is only very small, namely around 0.01. Therefore only the two features with the highest mutual information value were selected by the filter method. This were the features 'Customer Age' and 'Contract Duration' as can be seen in the second and fourth column of Table 11.

Table 11. Feature selection that resulted from the different feature selection methods for Model 1.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Customer Age	0.0148	✓	✓	✓ (10)	✓
Contract Duration	0.0137	✓	✓	✓ (10)	✓
Voluntary Excess Risk	0.0031	✓		✓ (9)	✓
Additional Insurance	0.0022	✓		✓ (9)	✓
Dental Insurance	0.0015	✓		✓ (9)	✓
Collectivity Discount Binary	0.0003	✓		✓ (8)	✓

The wrapper method resulted in a feature subset for each fold. The number in brackets after the check mark in the fifth column in Table 11 shows in how many folds (out of 10) a feature was selected. When a certain feature was selected in five folds or more, it was used in the final model. This was the case for all six variables.

Finally, the last column in Table 11 shows, which variables were selected by the embedded feature selection method, which was the lasso regression. Lasso regression required optimization of the shrinkage parameter, λ. Ten-fold cross validation was used for this optimization process. Figure 1 in Appendix H shows the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis).

This figure shows the value of lambda that leads to the highest AUC value, which is the left vertical line in the figure, and the largest lambda value within one standard error of this best lambda value, which is the right vertical line. It can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen and the corresponding coefficients of the different variables were checked. A variable was selected when at least one of the categories had a coefficient that was not equal to zero.

Table 11 summarizes for each method which variables were selected. Using these variables, two different models were developed, of which the different performance indicators are shown in Table 12.

Table 12. Different performance indicators for Model 1.

Criterion	Significance/Wrapper/Embedded (40:60)	Filter (40:60)
Accuracy	0.7989	**0.8002**
Recall	**0.2885**	0.2747
Precision	**0.1840**	0.1798
Kappa	**0.1156**	0.1085
AUC	**0.6501**	0.6415
AUK	**0.0662**	0.0634

From Table 12 it can be concluded that the model resulting from the significance, wrapper and embedded methods outperforms the model resulting from the filter method, in almost all the performance indicators. Therefore the former is chosen to be the best Model 1. The model summary of this model can be found in Appendix I Table 2.

5.3 Model 2: Customer web page visits for all customers

For model 2, apart from the features that were used in Model 1, features about whether a customer logged in on the secured Interpolis domain or not were added. As was explained in section 4.2.5, a distinction was made between logging in between January and October and logging in in November or December. Again, the four different methods were used to decide which features were selected and which were not. This was done in the same way as in Model 1.

For the filter method, again, the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 10.

Figure 10 shows that, again, the increase in performance after the second feature is included, is only very small. Therefore only the two features with the highest mutual information value were selected by the filter method. This were the features ‘Customer Age’ and ‘Contract Duration’ as can be seen in the second and fourth column of Table 13.

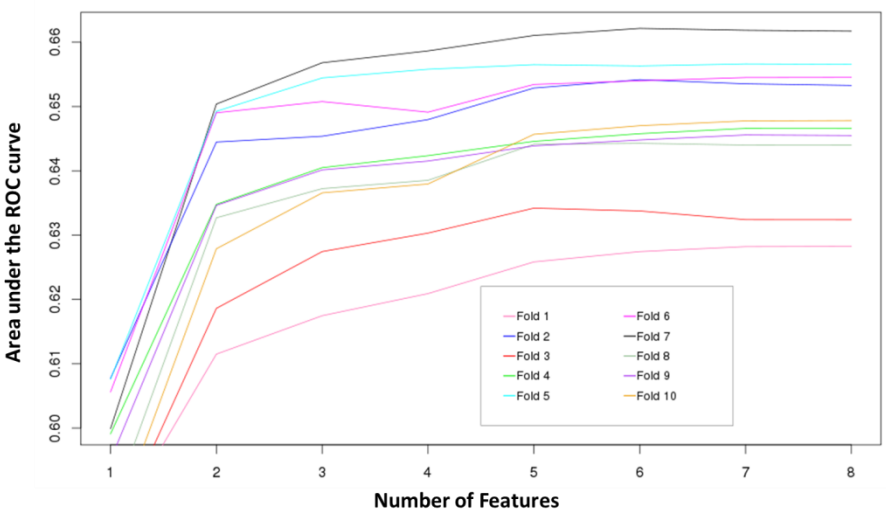


Figure 10. The area under the ROC curve vs. the Number of Features for each fold for Model 2.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 2 in Appendix H. Again, it can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 13.

Table 13. Feature selection that resulted from the different feature selection methods for Model 2.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Customer Age	0.0148	✓	✓	✓ (10)	✓
Contract Duration	0.0137	✓	✓	✓ (10)	✓
Voluntary Excess Risk	0.0031	✓		✓ (8)	✓
Additional Insurance	0.0022	✓		✓ (9)	✓
Dental Insurance	0.0015	✓		✓ (9)	✓
Collectivity Discount Binary	0.0003	✓		✓ (8)	✓
Logged in in November or December	4.7054×10^{-5}	✓		✓ (6)	
Logged in between January and October	3.0867×10^{-6}			(2)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 14.

Table 14. Different performance indicators for Model 2.

Criterion	Significance/Wrapper (40:60)	Filter (40:60)	Embedded (40:60)
Accuracy	**0.8004**	0.8002	0.7989
Recall	0.2872	0.2747	**0.2885**
Precision	**0.1852**	0.1798	0.1840
Kappa	**0.1167**	0.1085	0.1156
AUC	**0.6504**	0.6415	0.6501
AUK	**0.0671**	0.0634	0.0669

From Table 14 it can be concluded that, although the model resulting from the embedded method leads to the highest recall, the model resulting from the significance and wrapper method leads to the highest values for the other performance indicators. Therefore the latter is chosen to be the best Model 2. The model summary of this model can be found in Appendix I Table 3.

5.4 Model 2a: Webpage Visits for customers who logged in

For model 2a, the customers who logged in on the secured domain were selected. For these customers, apart from the features that were used in Model 1, features about which webpages were visited and about which webpages were visited before visiting one of the Interpolis websites. Again, a distinction was made between visiting a webpage between January and October and in November or December.

Since there are around 110 different URL's, the feature selection method that uses the significance of the variables was not feasible in this case.

Considering the filter method, the area under the ROC curve for the different number of variables is illustrated by the graph in Figure 11.

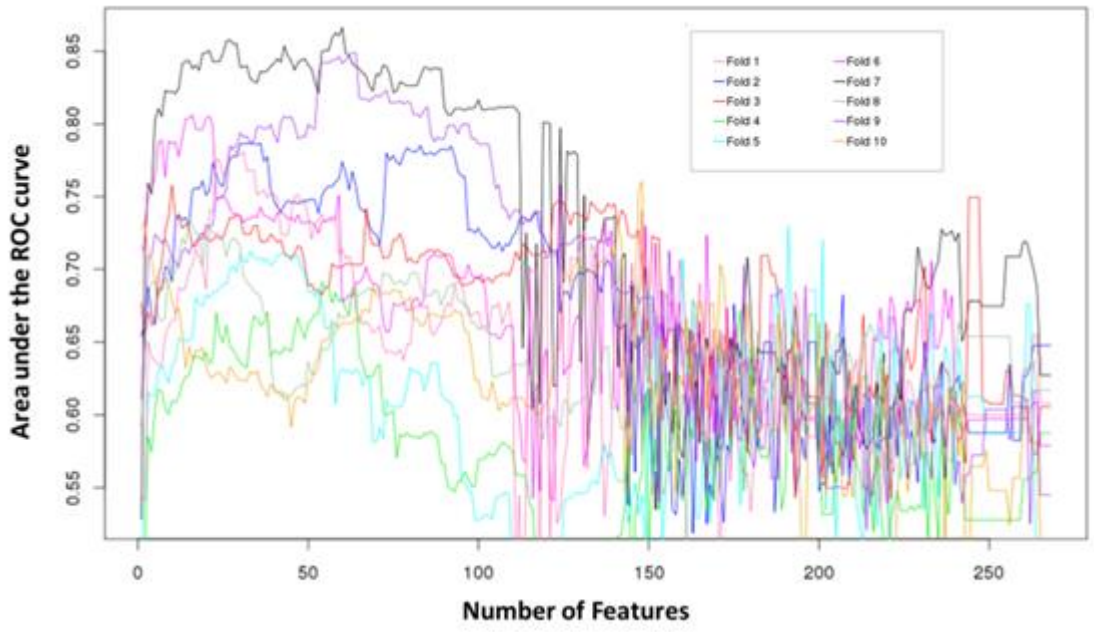


Figure 11. The area under the ROC curve vs. the Number of Features for each fold for Model 2.

Figure 11 shows that the performance of the models is very unstable, especially after hundred variables are selected. In order to make a better decision on how many variables to include, the graphs of the different folds were averaged. The graph resulting from this is shown in Figure 12.

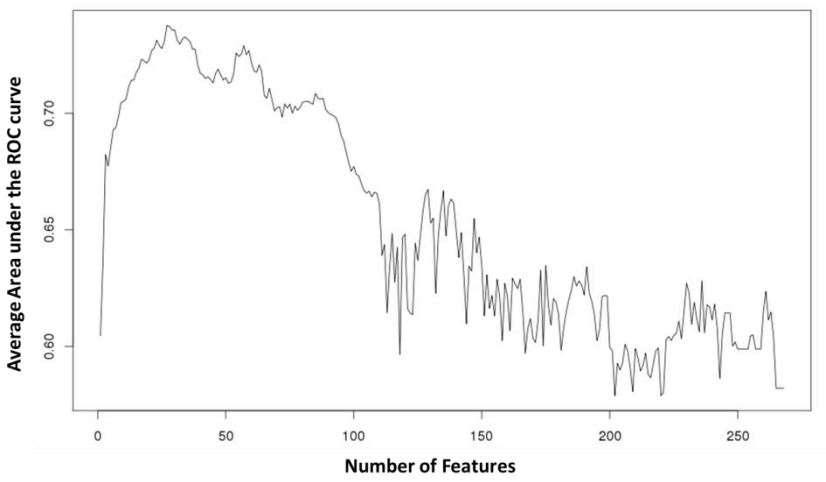


Figure 12. The average area under the ROC curve vs. the Number of Features for the different folds for Model 2.

It was found that the number of variables leading to the highest average AUC was equal to 27. Therefore the 27 variables with the highest mutual information were selected for the model. These variables were: customer age, contract duration, additional insurance, dental insurance and 23 webpages. The webpages were mostly about terminating the contract, about requesting information considering the insurance, such as coverage, care consumption, voluntary excess risk, about claiming, about problems with logging in, and about frequently asked questions.

The wrapper method resulted in a feature subset for each fold. Figure 13 shows how often a feature was selected 1 time, 2 times, 3 times, 4 times and 5 times. There were no features that were selected more than 5 times. This is another indication that the model is very unstable across the different folds. The features that were selected 4 or 5 times were used to build the final model. This was the case for 28 features. These variables were: contract duration, dental insurance and 26 URL's. The webpages were mostly about terminating the contract, modifying the policy, coverage, insight into care consumption and voluntary excess risk, information about damage and how to claim and information about paying.

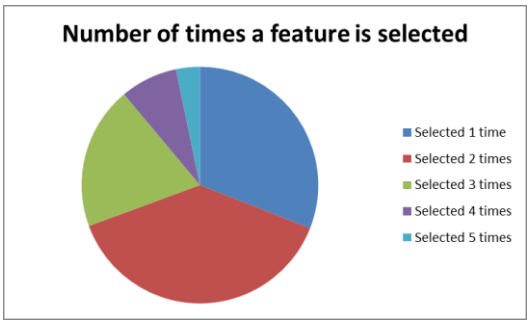


Figure 13. Circle diagram that shows how often a variable was selected x times.

For the embedded feature selection method, Figure 3 in Appendix H shows the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC)

(y-axis). Again, it can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen. This led to the inclusion of 16 variables. These variables were: customer age, contract duraction, additional insurance and 13 URL's. These URL's were, again, mostly about terminating the contract, about requesting information considering coverage and care consumption, about claiming and about problems with logging in.

The different feature selection methods selected different amounts and different types of variables. The performance of these different models is shown in Table 15. In this case, an extra performance criterion is used, namely 'Gross Recall'. The reason for this is that model 2a can only be used for the customers that have logged in. While recall shows the proportion of retrieved churners from the total amount of churners that have logged in, gross recall shows this proportion from the total amount of all churners, so also the ones that did not log in. Although, the aim of the models is to be used separately, meaning that for each type of customer the corresponding model is used, the gross recall is still useful for comparing the different models.

Table 15. Different performance indicators for Model 2a.

Criterion	Filter (40:60)	Wrapper (40:60)	Embedded (40:60)
Accuracy	0.7175	0.7330	**0.7461**
Recall	**0.4412**	0.3676	0.3971
Gross Recall	**0.0051**	0.0042	0.0046
Precision	0.1310	0.1214	**0.1357**
Kappa	0.0880	0.0690	**0.0926**
AUC	**0.6469**	0.6188	0.6393
AUK	**0.0602**	0.0409	0.0533

From Table 15 it can be seen that the model resulting from the filter method and the model resulting from the embedded method have quite similar scores for the different performance indicators. However, the model resulting from the embedded method leads to the highest Kappa, which was explained to be the most important performance indicator. Therefore this model is chosen to be the best Model 2a. The model summary of this model can be found in Appendix I Table 4.

5.5 Model 3: Touchpoint data for all customers

For model 3, apart from the features that were used in Model 1, features about whether a customer has contacted Interpolis between January and October, whether a customer has contacted Interpolis in November or December and whether a customer has filed a complaint were used. Again, the four different methods were used to decide which features were selected and which were not.

For the filter method the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 14.

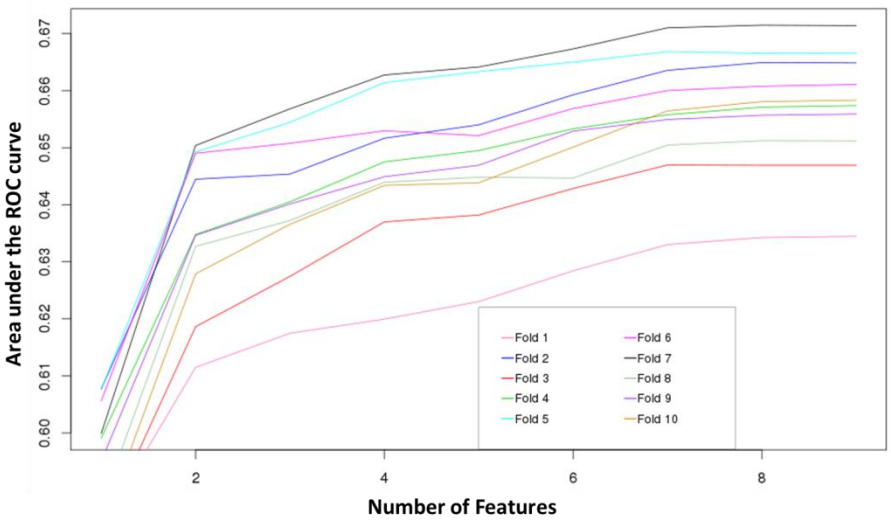


Figure 14. The area under the ROC curve vs. the Number of Features for each fold for Model 3.

Figure 14 shows that, again, the increase in performance after the second feature is included, is only very small. Therefore only the two features with the highest mutual information value were selected by the filter method. This were the features 'Customer Age' and 'Contract Duration' as can be seen in the second and fourth column of Table 16.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 4 in Appendix H. Again, it can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 16.

Table 16. Feature selection that resulted from the different feature selection methods for Model 3.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Customer Age	0.0148	✓	✓	✓ (10)	✓
Contract Duration	0.0137	✓	✓	✓ (10)	✓
Voluntary Excess Risk	0.0031	✓		✓ (9)	✓
Contact between January and October	0.0026	✓		✓ (10)	✓
Additional Insurance	0.0022	✓		✓ (10)	✓
Contact in November or December	0.0017	✓		✓ (9)	✓
Dental Insurance	0.0015	✓		✓ (10)	✓
Collectivity Discount Binary	0.0003	✓		✓ (8)	✓
Complaint	0.0001	✓		✓ (7)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 17.

Table 17. Different performance indicators for Model 3.

Criterion	Significance/Wrapper (40:60)	Filter (40:60)	Embedded (40:60)
Accuracy	0.7868	**0.8002**	0.7849
Recall	0.3242	0.2747	**0.3284**
Precision	**0.1843**	0.1798	0.1838
Kappa	**0.1220**	0.1085	**0.1220**
AUC	**0.6616**	0.6415	0.6611
AUK	**0.0717**	0.0634	0.0715

As can be seen from Table 17 the performance indicators of the model obtained by the significance and wrapper methods are very similar to those of the model obtained by the embedded method. The filter method, on the other hand, performs slightly worse. Since the AUC and AUK values of the model resulting from the significance and wrapper method are slightly bigger, the model resulting from this method is chosen. The model summary of this model can be found in Appendix I Table 5.

5.6 Model 3a: Touchpoint data for customers who have contacted

For model 3a, the customers who have contacted Interpolis were selected. For these customers, apart from the features that were used in Model 1, features about these contact moments were used. These features were: the call frequency, the mail frequency, the social media frequency and whether one of the contact moments were not first time right. For these variables, again, a distinction was made between the time horizons 'Between January and October' and 'November or December'.

By selecting the variables with the significance method, it was found that the variables 'Mail Frequency between January and October', 'Social Media Frequency between January and October' and 'Social Media Frequency in November and December' required different categorizations. For the variable 'Social Media Frequency in November and December' the first two categories were grouped together, while for the variables 'Mail Frequency between January and October' and 'Social Media Frequency between January and October' the first and the third category were grouped together. These adapted variables were also used in the following models.

For the filter method the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 15.

Figure 15 shows that, again, the increase in performance after the second feature is included, is only very small. Therefore only the two features with the highest mutual information value were selected by the filter method. This were the features 'Mail Frequency in November and December' and 'Customer Age' as can be seen in the second and fourth column of Table 18.

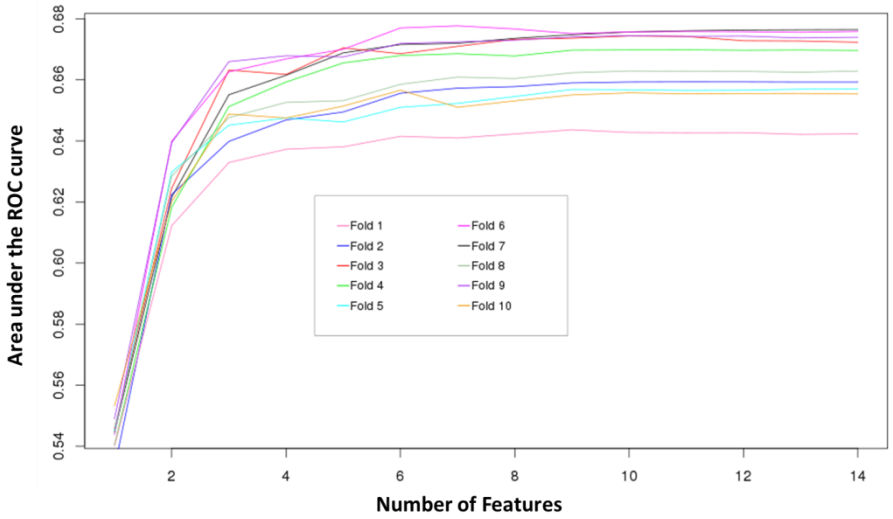


Figure 15. The area under the ROC curve vs. the Number of Features for each fold for Model 3a.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 5 in Appendix H. Again, it can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 18.

Table 18. Feature selection that resulted from the different feature selection methods for Model 3a.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Mail Frequency November/December	0.0178	✓	✓	✓ (10)	✓
Customer Age	0.0150		✓	✓ (10)	✓
Contract Duration	0.0092	✓		✓ (10)	✓
Voluntary Excess Risk	0.0022	✓		✓ (9)	✓
Additional Insurance	0.0020	✓		✓ (7)	✓
Dental Insurance	0.0016	✓		✓ (9)	✓
Call Frequency November/December	0.0014			✓ (5)	✓
Mail Frequency Between January and October	0.0008	✓		✓ (6)	✓
Call Frequency Between January and October	0.0004	✓		✓ (9)	✓
Social Media Frequency Between January and October	0.0004	✓		✓ (8)	✓
Not First Time Right in November/December	0.0002			(3)	
Social Media Frequency November/December	8.4482×10^{-5}	✓		✓ (5)	✓
Not First Time Right Between January and October	6.0164×10^{-5}			(3)	
Collectivity Discount Binary	1.7815×10^{-5}			✓ (5)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 19. Again, 'Gross Recall' is used, because model 3a can only be used for the customers that have contacted Interpolis.

Table 19. Different performance indicators for Model 3a.

Criterion	Significance (40:60)	Filter (40:60)	Wrapper (40:60)	Embedded (40:60)
Accuracy	0.8237	**0.8720**	0.8068	0.8074
Recall	0.2496	0.1059	**0.3011**	0.3007
Gross Recall	0.1112	0.0472	**0.1341**	0.1340
Precision	0.2703	**0.4308**	0.2588	0.2597
Kappa	0.1597	0.1274	0.1675	**0.1682**
AUC	0.6465	0.6300	**0.6699**	**0.6699**
AUK	0.0827	0.0709	**0.0923**	**0.0923**

Table 19 shows that the model resulting from the embedded method has the highest Kappa value and, together with the model resulting from the wrapper method, has the highest AUC and AUK values. Therefore this model is chosen to be the best model 3a. The model summary of this model can be found in Appendix I Table 6.

5.7 Model 3b: Touchpoint data for customers who complained

For model 3b, the customers who have filed a complaint were selected. For these customers, apart from the features that were used in Model 1, features about these complaints were used. These features were about whether a complaint was not assigned or not handled personally, whether the lead time of handling the complaint was within the promised lead time and whether the customer wanted to have a reassessment of the complaint. Because the dataset of the customers who have filed a complaint was relatively small, it was chosen to use the output of model 1 as input variable, instead of using the features from model 1 separately. This was done in order to reduce the 'curse of dimensionality'.

For the filter method the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 16.

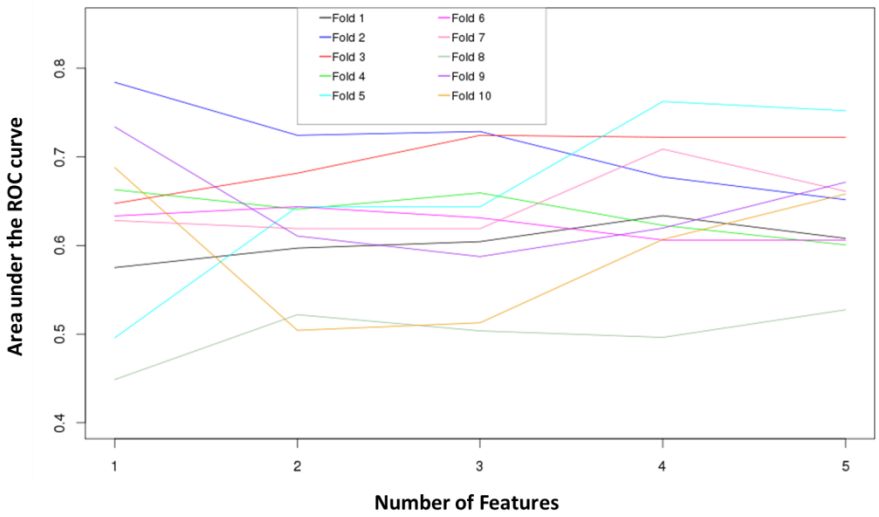


Figure 16. The area under the ROC curve vs. the Number of Features for each fold for Model 3b.

Figure 16 shows that the optimal number of features differs per fold. Therefore it is hard to decide which number will lead to the model with the best performance. It seems that there is not really an increase in performance after the first variable is added. Therefore the filter method only selected the variable with the highest mutual information value, which was the outcome of Model 1 as is shown in Table 19.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 6 in Appendix H. In this case, there seems to be quite some increase in performance between the right and left vertical line. Therefore, this time, the value of lambda that corresponds to the left vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 20.

Table 20. Feature selection that resulted from the different feature selection methods for Model 3b.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Outcome Model 1	0.0304	✓	✓	✓ (6)	✓
Not Assigned	0.0012			(3)	
Not Handled Personally	0.0006			✓ (5)	
Reassessment	0.0001			(3)	
Not within promised lead time 1st reaction	5.7901×10^{-7}			(4)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 21. Again, 'Gross Recall' is used, because model 3b can only be used for the customers that have filed a complaint.

Table 21. Different performance indicators for Model 3b.

Criterion	Significance/Filter/Embedded (40:60)	Wrapper (40:60)
Accuracy	**0.7056**	0.6926
Recall	0.2889	0.2889
Gross Recall	0.0022	0.0022
Precision	**0.2653**	0.2500
Kappa	**0.0922**	0.0748
AUC	**0.5236**	0.5071
AUK	**0.0245**	0.0153

Table 21 shows that, except for the recall and gross recall, all the performance indicators of the model resulting from the significance, filter and embedded methods are higher than those of the model resulting from the wrapper method. Therefore the model resulting from the significance, filter and embedded method is chosen to be the best model 3b, even though this means that only the outcome of model 1 is used as an input variable. The model summary of this model is given in Appendix I Table 7.

5.8 Model 4a: NPS data for customers who have contacted

For model 4a, the customers who have contacted Interpolis were selected. For these customers, apart from the features that were used in Model 3a, a feature was added that indicated whether the customer filled in the NPS survey or not.

For the filter method the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 17.

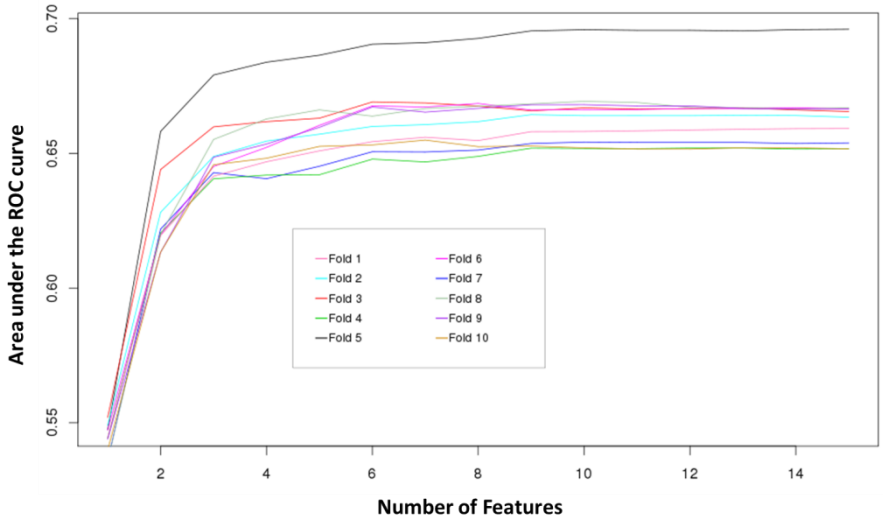


Figure 17. The area under the ROC curve vs. the Number of Features for each fold for Model 4a.

Figure 17 shows that, again, the increase in performance after the second feature is included, is only very small. Therefore only the two features with the highest mutual information value were selected by the filter method. This were the features 'Mail Frequency in November and December' and 'Customer Age' as can be seen in the second column of Table 22.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 7 in Appendix H. Again, it can be seen that there is no much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 22.

Table 22. Feature selection that resulted from the different feature selection methods for Model 4a.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Mail Frequency November/December	0.0178	✓	✓	✓ (10)	✓
Customer Age	0.0150		✓	✓ (10)	✓
Contract Duration	0.0092	✓		✓ (10)	✓
Voluntary Excess Risk	0.0022	✓		✓ (8)	✓
Additional Insurance	0.0020	✓		✓ (8)	✓
Dental Insurance	0.0016	✓		✓ (9)	✓
Call Frequency November/December	0.0014			(2)	✓
Mail Frequency Between January and October	0.0008	✓		✓ (8)	✓
Call Frequency Between January and October	0.0004	✓		✓ (5)	✓
Social Media Frequency Between January and October	0.0004	✓		✓ (6)	✓
Not First Time Right in November/December	0.0002			(1)	
Social Media Frequency November/December	8.4482×10^{-5}	✓		✓ (7)	
Not First Time Right Between January and October	6.0164×10^{-5}			✓ (5)	
Collectivity Discount Binary	1.7815×10^{-5}			✓ (5)	
Filled in NPS	1.6122×10^{-7}			(4)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 23. Again, 'Gross Recall' is used, because model 4a can only be used for the customers that have contacted Interpolis.

Table 23. Different performance indicators for Model 4a.

Criterion	Significant (40:60)	Filter (40:60)	Wrapper (40:60)	Embedded (40:60)
Accuracy	0.8237	**0.8720**	0.8050	0.8075
Recall	0.2496	0.1059	**0.3067**	0.2995
Gross Recall	0.1112	0.0472	**0.1367**	0.1335
Precision	0.2703	**0.4308**	0.2580	0.2595
Kappa	0.1597	0.1274	**0.1684**	0.1677
AUC	0.6465	0.6300	**0.6697**	**0.6697**
AUK	0.0827	0.0917	**0.0922**	**0.0922**

Table 23 shows that the model resulting from the wrapper method has the highest Kappa value and, together with the model resulting from the embedded method, has the highest AUC and AUK values. Therefore this model is chosen to be the best Model 4a. The summary of this model can be found in Appendix I Table 8.

5.9 Model 4b: NPS data for customers who complained

For model 4b, the customers who have filed a complaint were selected. For these customers, apart from the features that were used in Model 3b, a feature was added that indicated whether the customer filled in the NPS survey or not.

For the filter method the area under the ROC curve was calculated for the different number of variables. The result of this is illustrated by the graph in Figure 18.

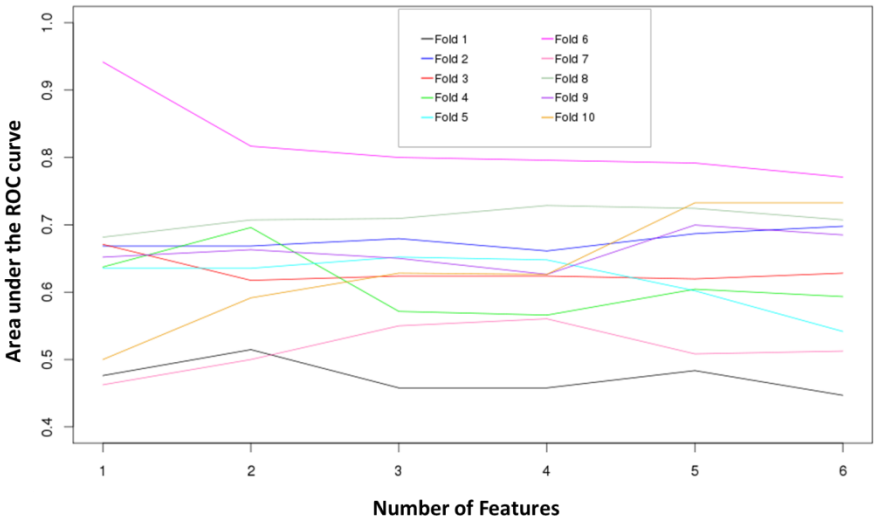


Figure 18. The area under the ROC curve vs. the Number of Features for each fold for Model 4b.

Figure 18 shows that the optimal number of features differs per fold. Therefore it is hard to decide which number will lead to the model with the best performance. It seems that there is not really an increase in performance after the first variable is added. Therefore the filter method only selected the variable with the highest mutual information value, which was the outcome of Model 1 as is shown in Table 24.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 8 in Appendix H. In this case, there seems to be quite some increase in performance between the right and left vertical line. Therefore, this time, the value of lambda that corresponds to the left vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 24.

Table 24. Feature selection that resulted from the different feature selection methods for Model 4b.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Outcome Model 1	0.0304	✓	✓	✓ (7)	✓
Filled in NPS	0.0061			(4)	
Not Assigned	0.0012			✓ (6)	
Not Handled Personally	0.0006			(2)	
Reassessment	0.0001			(4)	
Not within promised lead time 1st reaction	5.7901×10^{-7}			(2)	

The performance indicators of the models resulting from the different feature selection methods are given in Table 25. Again, 'Gross Recall' is used, because model 4b can only be used for the customers that have filed a complaint.

Table 25. Different performance indicators for Model 4b.

Criterion	Significance/Filter/Embedded (40:60)	Wrapper (40:60)
Accuracy	**0.7056**	0.7013
Recall	0.2889	0.2889
Overall Recall	0.0022	0.0022
Precision	**0.2653**	0.2600
Kappa	**0.0922**	0.0863
AUC	0.5236	**0.5256**
AUK	0.0245	**0.0255**

As can be seen from Table 25 the Kappa of the model resulting from the significance, filter and embedded method has the highest Kappa value. Therefore this model is selected as the best model 4b. The model summary of this model can be found in Appendix I Table 9.

5.10 Model 4c: NPS data for customers who have contacted and gave NPS score

For Model 4c the customers who have contacted Interpolis and who filled in the NPS score after a contact moment were selected. Because the proportion of customers who have contacted Interpolis and filled in the NPS survey is very small, the output of model 3a was used as input, instead of using the features of Model 3a separately. Since Model 3a with the embedded feature selection gave the highest performance, this model is chosen as input variable.

For the filter method, this time, no graph was created, because the feature selection was only done for two variables. Instead, it was checked in how many folds one variable led to the highest AUC and in how many folds both the variables led to the highest AUC. In 7 out of the 10 folds the model with two variables led to the highest AUC. Therefore the filter method selected both variables.

Considering the embedded method, the graph that plots the logarithm of the different values for λ (x-axis) and their corresponding values for the area under the ROC curve (AUC) (y-axis) can be found in Figure 8 in Appendix H. There is not much increase in performance between the right and left vertical line. Therefore the value of lambda that corresponds to the right vertical line is chosen.

A summary of the variables that were selected for each method, can be found in Table 26.

Table 26. Feature selection that resulted from the different feature selection methods for Model 4c.

Feature	Mutual Information Value	Selected in Final Model - Significance	Selected in Final Model - Filter	Selected in Final Model - Wrapper	Selected in Final Model - Embedded
Outcome Model 3a Embedded	0.0368	✓	✓	✓ (10)	✓
NPS Score	0.0230	✓	✓	✓ (7)	

From Table 26 it can be concluded that both variables are selected by all the feature selection methods. The performance indicators of the model that includes both these features are given in Table 27. Again, 'Gross Recall' is used, because model 4c can only be used for the customers that have contacted Interpolis and have filled in the NPS survey.

Table 27. Different performance indicators for Model 4c.

Criterion	Significance/Filter/Wrapper (40:60)	Embedded (40:60)
Accuracy	**0.7791**	0.7627
Recall	0.3939	0.3939
Gross Recall	0.0044	0.0044
Precision	**0.2149**	0.1985
Kappa	**0.1608**	0.1405
AUC	**0.6617**	0.6493
AUK	**0.0838**	0.0690

From Table 27 it can be concluded that the model that is obtained by the significance, filter and wrapper method performs better on all performance criteria, except recall and gross recall which is the same for both models. Therefore this model is chosen as the best Model 4c. The model summary of this model can be found in Appendix I Table 10.

5.11 Model 4d: NPS data for customers who complained and gave NPS score

The aim of Model 4d was to select the customers who have filed a complaint and who have filled in the NPS score after they had filed this complaint and to include this NPS score into the model. The dataset of customers who have filed a complaint and have filled in the NPS score was only a very small fraction of the entire customer base. After changing this data to a 40:60 Churn:Non-Churn ratio, the size of the dataset was even smaller. Therefore, unfortunately, it was decided that this dataset was too small to build any models with.

5.12 Comparison of the Models

Now that all the models are developed and that for each model a decision is made which feature selection methods leads to the best result, all these best models can again be compared based on their scores on the performance indicators. This overview can be found in Table 28.

Table 28. Comparison of the performance indicators of the different models.

Criterion	Best Model 0	Best Model 1	Best Model 2	Best Model 2a	Best Model 3	Best Model 3a	Best Model 3b	Best Model 4a	Best Model 4b	Best Model 4c
Feature Selection Method	Domain Experts	Sign., Wrapper and Embedded	Sign. and Wrapper	Embedded	Sign. and Wrapper	Embedded	Sign., Filter and Embedded	Wrapper	Sign., Filter and Embedded	Sign., Filter, Wrapper
Accuracy	0.8029	0.7989	0.8004	0.7461	0.7868	**0.8074**	0.7056	0.8050	0.7056	0.7791
Recall	0.2747	0.2885	0.2872	**0.3971**	0.3242	0.3007	0.2889	0.3067	0.2889	0.3939
Gross Recall	-	-	-	0.0046	-	0.1340	0.0022	**0.1367**	0.0022	0.0044
Precision	0.1830	0.1840	0.1852	0.1357	0.1843	0.2597	**0.2653**	0.2580	**0.2653**	0.2149
Kappa	0.1120	0.1156	0.1167	0.0926	0.1220	0.1682	0.0922	**0.1684**	0.0922	0.1608
AUC	0.6480	0.6501	0.6504	0.6393	0.6616	**0.6699**	0.5236	0.6697	0.5236	0.6617
AUK	0.0661	0.0662	0.0671	0.0533	0.0717	**0.0923**	0.0245	0.0922	0.0245	0.0838

From Table 28 it can be concluded that Model 3a, resulting from the embedded method leads to the highest accuracy, AUC and AUK. The highest recall is reached by Model 2a, resulting from the embedded method and the highest precision by Model 3b and Model 4b, resulting from the significance, filter and embedded methods. The highest Kappa is obtained by Model 4a, resulting from the wrapper method.

Chapter 6 Practical Implications

In this chapter an answer is given to the sixth research sub-question, which was about how the models can support decision making at the marketing department of Interpolis. First, the general insights about the relationship between different predictors and customer churn is discussed. This is followed by an investigation on whether the models fulfill the boundary conditions to be implemented at the marketing department. Finally, recommendations are given to the marketing department of Interpolis.

6.1 Gained Insights

First of all, the models give insight into which variables play an important role in predicting customer churn, which is one of the advantages of using logistic regression. This leads to new information that the people from the marketing department can take into account when making decisions on a daily basis. First of all it was found that customers with the highest risk to churn were customers with the following characteristics:

- They are *aged* between *26 and 36*
- They have their health insurance at Interpolis for the *first year*
- They have the *maximum voluntary excess risk* of €500
- They have the *type 2 additional insurance*
- They have *no dental insurance* or the *least extensive* dental insurance
- They are *not collectively insured*

Although this is the combination of characteristics that leads to the highest churn probability, the variables all have their own contribution to this probability. The variables customer age and contract duration are the most important characteristics.

Apart from these customer and policy characteristics, it was found that customers who *logged in* on the Interpolis domain in *November or December* had a lower probability to churn. On the other hand, customers who had *contacted* Interpolis or who had *filed a complaint*, had a higher probability to churn. Although this depended partly on the channel that was used for the contact and when this contact moment took place. Having *contact per e-mail* in November or December leads to the highest probability for a customer to churn. However, when a customer had sent two or more *messages on social media* in November or December this was an indicator for a lower probability to churn. Finally, as was expected, customers who gave a *higher NPS score*, had a lower probability to churn.

Apart from giving insight into the variables that play an important role in predicting churn, it was checked whether the models that were developed, could be implemented.

6.2 Model Implementation

Ideally, model implementation would lead to a scenario in which Interpolis is able to predict churn for each customer on a daily basis. This would make it possible to create a dashboard in which the probability to churn for each customer can be requested. Or, even one step further, to implement a system that gives a notification when a customer is likely to churn, for example when his probability is higher than 0.5. When this is known, different retention actions can be done, such as giving the customer extra attention by calling him, sending an e-mail or a letter. But besides giving these

customers 'extra attention', the content of webpages, service mails and even phone calls could be adapted to the churn probability of a customer, which would make the model relevant on a daily basis. However, in order to make this work, there are several important boundary conditions that need to be complied. These boundary conditions are first discussed in general and are then applied to the models that were developed in this study.

6.2.1 Boundary Conditions

First of all, an *accurate model* is needed, that is able to retrieve many churners, without making too many mistakes considering the non-churners. The latter is important because of the unwanted side effect of making a retention effort for a customer who was not thinking about churning, but was selected as churner by the model. Even though these customers were not thinking about churning before the retention action took place, these customers are now reminded of the possibility and, in worst case, might even consider to churn. This effect is known as the 'wake up' effect.

In general, factors that play an important role in the performance of a model are the amount of available data, the quality of the data, the way in which the data preparation is done, the modelling technique that is used and the predictive ability of the predictors that were used.

Secondly, the model should be *robust* or i.e. the model should have high *staying power*. A model has high staying power when its predictive performance in a number of periods after the estimation period remains approximately the same (Risselada et al., 2010). Risselada et al. (2010) studied the staying power of customer churn prediction models for internet service providers as well as for insurance companies and found that this staying power was rather low for various methods, including logistic regression (Risselada et al., 2010). Therefore this is an important issue to take into account.

Thirdly, when an accurate and robust model is developed, the *data infrastructure* should make it possible that the data that is needed as input for this model, is easily accessible and available on a daily basis. This is necessary in order to automate the prediction.

6.2.2 Meeting the Boundary Conditions

Considering the models that were developed in this study, only the models that used the customer and policy characteristics as predictors (Model 0 and Model 1) meet the third requirement. Retrieving the webpage, contact and complaint data, was not very easy, since these datasets were administered by different people and there was no good infrastructure for retrieving this data on a daily basis. Data about the customer and policy characteristics, however, is easier to use, since, in contrast to the webpage, contact and complaint data, this data is not changing much over the year. Therefore it is unnecessary for this model to be updated on a daily basis.

According to the first condition, the model should be able to retrieve many churners, without making too many mistakes considering the non-churners. When comparing the best Model 0, for which features were selected by domain experts, and the best Model 1, for which features were selected by the significance, wrapper and embedded methods, it was found that Model 1 performs better on all performance criteria except the accuracy. Moreover, as was explained in section 4.2.6, Model 0 is expected to have high multicollinearity, since it uses 'Premium' as a predictor. Therefore it is not recommended to use Model 0.

The Kappa, AUC and AUK values of Model 1 are mediocre, when taking into account that the models try to predict human behavior, which is difficult. It should be noted that, since the precision of the model is rather low, the *wake up effect* needs to be taken into account when considering different retention activities. This is important for both the channel as well as the message that is chosen for the retention activity. Moreover, for customers who have a moderate risk to churn (around 0.6 or 0.7) the wake up effect is more present than for customers that already had a high risk to churn. For example, for customers that do not have a very high chance to churn, it might be better to use a more careful approach, such as changing a webpage, instead of calling the customer.

In order to check whether Model 1 fulfilled the third condition, the model was tested on data from 2016. The performance indicators resulting from this are given in the third column of Table 29.

Table 29. Different performance indicators for Model 1 tested on data from 2015 and 2016.

Criterion	Best Model 1 tested on 2015 data	Best Model 1 tested on 2016 data
Accuracy	0.7989	0.8285
Recall	0.2885	0.2546
Precision	0.1840	0.1859
Kappa	0.1156	0.1213
AUC	0.6501	0.6462
AUK	0.0653	0.0620

From Table 29 it can be seen that the different performance indicators of the model when tested on data from 2015 are very similar than when the model was tested on data from 2016. This means that the model has high staying power, which means that the chances that this model is performing similar on data from 2017 are also high. In order to give even more insight into how the model performs on data from 2016, a cumulative gain chart is created, which is shown in Figure 19. This chart shows how much percent of the churners can be reached when contacting x percent of the total population.

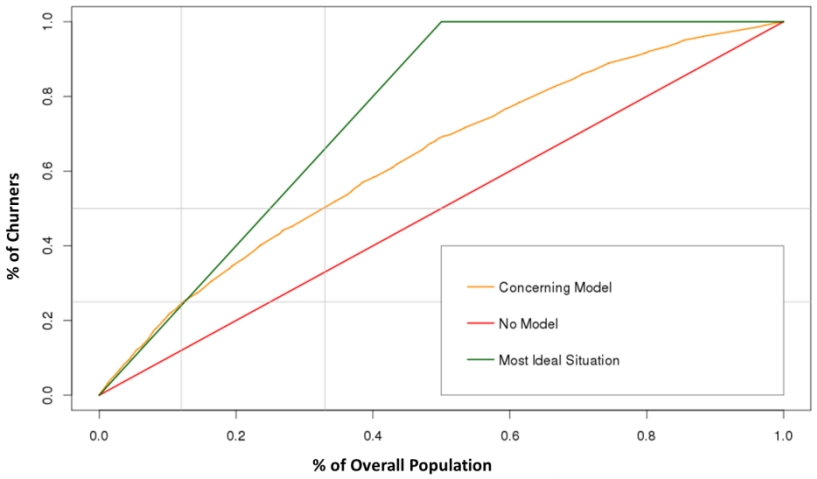


Figure 19. Cumulative Gain Chart for Model 1 tested on Data from 2016.

From Figure 19 it can be concluded that, when using customer age, contract duration, voluntary excess risk, additional insurance, dental insurance and collectivity discount on the data from 2016, a *quarter of the churners* can be reached when contacting approximately *10% of the overall population*. The red line in this figure shows the situation when no model would be used, in this case

a quarter of the population needs to be contacted in order to reach a quarter of the churners. The green line shows the ideal situation, from which it can be concluded that the model performs especially well when contacting less than 20% of the customers.

6.3 Recommendations

It is recommended that Interpolis starts with using Model 1 for the upcoming campaign, making it possible to adjust the retention efforts by the likelihood of a customer churning. Meanwhile, Interpolis must work on a better data infrastructure, increase the quality of the data about webpages and contact moments, explore the importance of other variables and experiment with different prediction techniques such as neural networks. Other variables that could be important in predicting churn, are: the number of other insurance products a customer has at Interpolis, the customer's income, the area where the customer is living, the family configuration of the customer, whether the rest of his family is also insured at Interpolis and the consumption of care. Although, concerning the ethical aspect, Interpolis also needs to consider how far it wants to go in using this sensitive data for prediction purposes. Moreover, Interpolis and Rabobank, which is the sales channel from Interpolis, can cooperate more closely in developing churn prediction models, such that the data from Rabobank, can be used to improve the model. Such that, in the end, the scenario discussed at the beginning of this chapter is no longer just a scenario, but the daily way of working.

Chapter 7 Conclusions, Limitations and Future Research

This chapter starts with drawing conclusions by answering the research questions. Just like all studies, this study has its limitations. These are discussed in the second paragraph of this chapter. Finally, recommendations for future research are given in section 7.3.

7.1 Conclusions

The goal of this study was to develop a customer churn prediction model for the healthcare insurance of Interpolis and to answer the main research question, which was stated as follows:

How can an accurate and relevant customer churn prediction model be created using data about customer and policy characteristics, webpage visits, touchpoints, and the NPS?

In order to answer this question, logistic regression models were developed that used different types of predictors and different types of feature selection methods.

It was found that the customer and policy characteristics *'Customer Age'* and *'Contract Duration'* are very important features for predicting churn, which is in line with findings from previous studies as was discussed in section 2.2. Moreover, although not as important as those variables, the variables related to: the type of *additional or dental insurance* a customer has and the amount of *voluntary excess risk* a customer has, were also found to predict churn. The variable that indicated whether a customer was insured individually or collectively, and thus whether the customer got a discount on his additional insurance or not, was rather unstable across the different models and different feature selection methods. Therefore the predictive ability of this variable is doubtful.

Apart from the customer and policy characteristics, it was found that including data about customer *web page visits* could slightly improve the predictive ability of the model. However, this was only the case when a variable was added that indicated whether a customer had logged in on the 'My Interpolis' domain in November or December. This led to slightly higher values for the accuracy, precision, Kappa, AUC and AUK of the model. When more detailed variables about the different webpages that a customer visited were included, this did not lead to an increased predictive ability of the model. On the contrary, this led to lower values for all the performance indicators, except for the recall, which increased significantly.

As regards the inclusion of data about other *customer touchpoints*, a distinction was made between 'regular' contact moments and, more specifically, complaints. First, variables were added that indicated: whether the customer had contacted Interpolis between January and October, whether the customer had contacted Interpolis in November or December, and whether the customer had filed a complaint. When adding these variables to the model, the predictive ability improved moderately in terms of the recall, precision, Kappa, AUC and AUK values.

After this, the customers who had contacted Interpolis were selected and more detailed variables about these contact moments were added. This led to an even further increase in the predictive ability of the model. These detailed variables included information about the call frequency, the mail frequency and the frequency of messages on social media for the two time frames: Between January and October, and November and December.

However, selecting the customers who had filed a complaint and including more detailed information about these complaints, did not improve the predictive ability of the model. On the contrary, all the performance indicators decreased, except for the precision, which increased slightly. A reason for this might be the small size of the dataset, since the proportion of people who file a complaint is only 0.35%. This issue is further discussed in section 7.2.

Finally, the inclusion of data about the Net Promoter Score did not further improve the predictive ability of the model. The only exception here is that for the customers who had contacted the insurance company and who had filled in the NPS survey, the recall increased when the NPS score was added to the model. A reason for why this did not lead to an increased predictive ability, might again be the small size of the datasets, which is caused by the fact that only a small proportion of the customers fill in the NPS survey.

The use of the different feature selection methods, led to different feature subsets and thus to differences in the predictive ability of the models in almost all cases. In most cases, the filter method resulted in the smallest feature subset, while the wrapper method resulted in the largest feature subset. The filter method often led to the highest accuracy, while the wrapper method led to the highest values for the AUC and the AUK in most cases. The highest Kappa value was often obtained by the significance or the embedded methods, although this did not differ much from the wrapper method. The method that led to the highest recall or precision, differed from time to time.

In summary, a relatively accurate and relevant customer churn prediction model can be created by using logistic regression with variables related to: the customers *age*, the *contract duration*, the type of *additional insurance*, the type of *dental insurance* and the *voluntary excess risk*. The predictive ability of this model can be further increased when variables are included about: whether the customer has *logged in* on the secured domain of the insurance company, whether the customer has *contacted* the insurance company and whether the customer has *filed a complaint*. For customers who had contacted the insurance company, the predictive ability could be further increased by including information about *the call frequency*, the *mail frequency* and the *frequency of messages on social media*. Considering the different feature selection methods, it is hard to tell which method leads to the best model, because it depends on which performance indicator is considered to be the most important. However, when taking into account that churn models often need to deal with highly unbalanced data, which makes the Kappa, AUC and AUK important criteria, the *wrapper method* might be the best option when using logistic regression to predict customer churn.

Although some of the findings in this study are specific for the health insurance sector, such as the importance of additional and dental insurance in predicting churn, most of the other findings are expected to be applicable for predicting churn in other sectors as well. For example, customer age, contract duration and the number of complaints, are expected to be just as important for predicting churn in other sectors. Moreover, the findings on the different feature selection methods are expected to be relevant and useful for other sectors and perhaps even for applications other than churn prediction.

7.2 Limitations

One of the main limitations of this study arises from the fact that logistic regression was used as modelling technique. Although, logistic regression has many advantages, such as giving insight into

the importance and significance of the different variables and being computationally fast, it also has several drawbacks. One of the most important drawbacks is that it relies on several assumptions. Although many of the assumptions are met in this study, the assumption that states that there should be little or no multicollinearity between variables is, especially for the webpage data, a little bit doubtful in this case. The reason for this is that it is expected that a customer who is looking for certain information on the internet will probably visit more than one webpage that contains information about this subject. This makes it very probable that webpages about the same subjects are correlated, which could be an explanation for the rather poor performance of the model that uses the webpage visits. Furthermore, it could be the case that the relation between the independent variables and the log odds is not linear, which decreases the performance of a logistic regression model (Lani, 2010). In order to overcome these issues, other methods that do not make this assumptions, like neural networks or fuzzy inference systems, could be used.

Another limitation is caused by using mutual information as the ranking criterion for the filter method. First of all, mutual information is a univariate ranking criterion, which does not take into account correlation between pairs of variables. Secondly, as was explained in section 3.1, mutual information requires categorization of the variables. Although, many of the variables already were categorical variables, some variables required discretization in order for the mutual information criterion to work. This was the case for the different webpage variables, which were the TF/IDF values, and when the outcome of one model was used as input for the other. Within the scope of this study, only one method for discretization was used, which was equal width binning with the number of bins equal to three. The reason for why this was chosen is that the data is very skewed, since the majority of cases have values equal to 0. This means that equal frequency binning is not feasible, since all the values bigger than 0 need to be binned in order to get somewhat equal frequency bins. But when this way of binning would be done, the added value of the TF/IDF transformation is lost. Therefore equal width binning seems to make more sense. But again, considering the data distribution, if the number of bins is made too large, this will lead to empty bins. Therefore it would be good to have a look at other discretization techniques, for example logarithmic binning, because the discretization method might influence the ranking of the variables and therefore the performance of the model. Moreover, other ranking criteria, especially multivariate ones need to be considered as well.

Another limitation of this study was that the dataset that included information about the complaints that the customers have filed was very small in size. This made it difficult to develop good models with the use of these variables. The dataset that included the webpage visits was not very small in absolute numbers, but because of the large amount of variables that were included in this dataset, it was too small to really find reliable results. Apart from the data quantity, another issue that is worth mentioning is about the quality of the data. The data that was available considering the customer characteristics was very limited, due to the switch to a new information management system and the difficulty in accessing the old database. Moreover the dependency on Rabobank for accessing certain data, made it difficult to include variables other than variables about the health insurance.

7.3 Implications for Future Research

One interesting subject for future research might be to compare a filter, wrapper or embedded method that optimizes the AUC, with a filter, wrapper or embedded method that optimizes the AUK. It would be interesting to see how the selected features and performance of the models differ when

using these two optimization criteria. It is expected that, especially for highly unbalanced data, optimizing the AUK might lead to better results. However, the use of the AUK, is still rather limited.

Moreover, a comparison of the different feature selection methods can be done when using modelling techniques different from logistic regression. A wrapper approach might for example be feasible for logistic regression, but for modelling techniques that need more computation time, like neural networks or fuzzy inference systems, this approach might be too time consuming. For these techniques, embedded methods might provide a better solution.

Furthermore, for the wrapper method different search algorithms can be compared. While in this study a forward sequential search was done, other methods, like hill climbing or genetic algorithms could also be used as search algorithms. In the same way, for the filter method different ranking criteria can be used. In this study mutual information was used, but other, multivariate criteria, might lead to different results (Li et al., 2016).

Finally, in recent times, the question arises whether it is ethical to use data for all kinds of predictions. Some people think that companies and the government go too far in using sensitive data and that this violates their right to privacy. Others are willing to give up a bit of their privacy, as long as this means better service or improved safety. Therefore it is good to carefully consider this aspect in the near future and to reflect on both these perspectives.

References

Bain & Company. “Measuring Your Net Promoter Score”. Retrieved on March 15th 2017 from: http://www.netpromotersystem.com/about/measuring-your-net-promoter-score.aspx

Bain & Company. “Three Types of Net Promoter Scores”. Retrieved on March 15th 2017 from: http://www.netpromotersystem.com/about/three-types-of-scores.aspx

Bendle N, Bagga CK. “The Metrics That Marketers Muddle”. *MIT Sloan Management Review* 2016; 57*;* p. 73-82.

Burez J, Van Den Poel D. “Handling class imbalance in customer churn prediction”. *Expert Systems with Applications* 2009;36; p. 4626–4636.

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. “CRISPDM 1.0”. Retrieved on 2nd March 2017 from: https://www.the-modeling-agency.com/crisp-dm.pdf

Bhat S.A, Darzi MA. "Customer relationship management: An approach to competitive advantage in the banking sector by exploring the mediational role of loyalty". *International Journal of Bank Marketing* 2016;34; p.388-410.

Blattberg RC, Kim B, Neslin S.A. Database Marketing: Analyzing and Managing Customers. New York; 2008 .

Bolancé C, Guillén M, Padilla-Barreto AE. “Predicting Probability of Customer Churn in Insurance”. In: León R, Muñoz-Torres MJ, Moneva JM (Eds.), *Modeling and Simulation in Engineering, Economics and Management*, vol.254.Switzerland; 2016. p. 82–91

Brockett P L, Golden LL, Guillen M, Nielsen JP, Parner J, Perez-Marin AM. “Survival Analysis of a Household Portfolio of Insurance Policies: How Much time Do You Have to Stop Total Customer Defection?”. *The Journal of Risk and Insurance* 2008*;*75*;* p.713-737*.*

Checkmarket. “Why there needs to be a European variant of the Net Promoter Score”. Retrieved on 23th May 2017 from: https://nl.checkmarket.com/blog/nps-eu/.

Chiang DA, Wang YF, Lee SL. “Goal oriented sequential pattern for network banking churn analysis”. *Expert System with Applications* 2003;25; p. 293-302.

EY. “The future of health insurance – A road map through change”. Retrieved on 10th February 2017 from: http://www.ey.com/Publication/vwLUAssets/EY-the-future-of-health-insurance/$FILE/EY-the-future-of-health-insurance.pdf

Guelman L, Guillén M, Pérez-Marín AM. “Random Forests for Uplift Modeling: An Insurance Customer Retention Case”. In: Engemann KJ, Gil-Lafuente AM, Merigó JM (Eds.), *Modeling and Simulation in Engineering, Economics and Management*, vol.115.Berlin; 2012. p. 123-133.

Guillén M, Nielsen JP, Scheike TH, Pérez-Marin A. “Time-varying effects in the analysis of customer loyalty: A case study in insurance”. *Expert Systems with Applications* 2012;39;p. 3551–3558.

Günther C, Tvete IF, Aas K, Sandness GI, Borgan Ø. “Modelling and predicting customer churn from an insurance company”. *Scandinavian Actuarial Journal* 2014; 1;p*.* 58-71.

Haan, de E, Verhoef PC, Wiesel T. “The predictive ability of different customer feedback metrics for retention”. *International Journal of Research in Marketing* 2015;32;p. 195–206.

Halvorsrud R, Kvale K, Følstad A. “Improving service quality through Customer Journey Analysis”. *Journal of Service Theory and Practice;*2016;26;p. 840-867.

Havrlant L, Kreinovich V. “A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)”. *International Journal of General Systems* 2017;46;p. 27-36.

He H, Garcia EA. “Learning from Imbalanced Data”. *IEEE Transactions on Knowledge and Data Engineering* 2009;21;p. 1263-1284.

Hur Y, Lim S. “Customer Churning Prediction Using Support Vector Machines in Online Auto Insurance Service”. In: Wang J, Liao X, Yi Z (Eds.), *Advances in Neural Networks – ISNN 2005,* vol.3497.Berlin;2005. p. 928-933.

Independer. “Tussentijds overstappen zorgverzekering”. Retrieved on 10th March 2017 from: https://www.independer.nl/zorgverzekering/info/tussentijds-overstappen-zorgverzekering.aspx

Interpolis. “Interpolis ZorgActief® Premietabel 2015”. Retrieved on 19th March 2017 from: https://www.interpolis.nl/~/media/files/premietabel-zorgactief-2015.pdf

Kaymak U, Ben-David A, Potharst R. “The AUK: A simple alternative to the AUC”. *Engineering Applications of Artificial Intelligence* 2012; 25; p. 1082–1089.

Kristensen K, Eskildsen J. “Is the NPS a trustworthy performance measure?”. *TQM Journal* 2014;26; p. 202-214.

Lani J. “Assumptions of Logistic Regression”. Retrieved on 20th July 2017 from: http://www.statisticssolutions.com/wp-content/uploads/kalins-pdf/singles/assumptions-of-logistic-regression.pdf

Lax H. “Bad is Stronger than Good: Lessons for Customer Loyalty & Experience”. Retrieved on 21st July 2017 from: http://customerthink.com/bad_is_stronger_than_good_lessons_for_customer_loyalty_experience_by_howard_lax/

Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. “Feature Selection: A Data Perspective”. Retrieved on 7th July 2017 from: https://arxiv.org/abs/1601.07996

Ministerie van Volksgezondheid, Welzijn en Sport. “Het Nederlandse zorgstelsel”. Retrieved on 16th March 2017 from: https://www.rijksoverheid.nl/onderwerpen/zorgverzekering/documenten/brochures/2016/02/09/het-nederlandse-zorgstelsel

Morik K, Köpcke H. “Analysing Customer Churn in Insurance Data – A Case Study”. In: Boulicaut J, Esposito F, Giannotti F, Pedreschi D (Eds.), *Knowledge Discovery in Databases: PKDD 2004,* Vol. 3202. Berlin; 2004. p. 325-336.

Nederlandse Zorgautoriteit (NZa). “Marktscan Zorgverzekeringsmarkt 2016”. Retrieved on 20th March 2017 from: https://www.nza.nl/1048076/1048181/Marktscan_Zorgverzekeringsmarkt_2016.pdf

Ngai EWT, Xiu L, Chau DCK. “Application of data mining techniques in customer relationship management: A literature review and classification”. *Expert Systems with Applications* 2009:36*:* p. 2592–2602.

Peng H, Long F, Ding C. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005;27; p. 1226-1238.

Risselada H, Verhoef PC, Bijmolt THA. “Staying Power of Churn Prediction Models”. *Journal of Interactive Marketing* 2010;24; p. 198–208.

Sainani KL. (2014). “Logistic Regression”. *American Academy of Physical Medicine and Rehabilitation* 2014;6; p. 1157-1162.

Schena F. “Predicting Customer Churn in the Insurance Industry: A Data Mining Case Study”. In: Petruzzellis L, Winer RS (Eds.), *Rediscovering the Essentiality of Marketing.* Switzerland; 2016. p. 747-751.

Sundarkumar GG, Ravi V. “A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance”. *Engineering Applications of Artificial Intelligence* 2015;37; p. 368–377.

Witten IH, Frank E, Hall MA, Pal CJ. Data Mining : Practical Machine Learning Tools and Techniques. Massachusetts: Cambridge;2016.

Yoder J, Rao A, Bajowala M. “Insurance 2020: Turning change into opportunity”. Retrieved on 10th February 2017 from: http://www.pwc.com/gx/en/insurance/pdf/insurance-2020-turning-change-into-opportunity.pdf

Zomerdijk LG, Voss CA. “NSD processes and practices in experiential services”. *Journal of Product Innovation Management* 2011;28; p. 63-80.

Appendices

Appendix A – Background information on Literature Review

Table 1. Overview of the scientific articles used in the literature review.

Title	Authors
Churn Management	Blattberg, Kim & Neslin
Predicting Probability of Customer Churn in Insurance	Bolancé, Guillen, Padilla-Barreto
Survival Analysis of a Household Portfolio of Insurance Policies: How much time do you have to stop total customer defection?	Brockett, Golden, Guillen, Nielsen, Parner
Random Forests for Uplift Modeling: An Insurance Customer Retention Case	Guelman, Pérez-Marín & Montserrat Guillen
Time-varying effects in the analysis of customer loyalty: A case study in insurance	Guillen, Nielsen, Scheike, Perez-Marin
Modelling and predicting customer churn from an insurance company	Gunther, Tvete, Aas, Sandnes & Borgan
Customer churning prediction using support vector machines in online auto insurance service	Hur & Lim
Analysing customer churn in insurance data – A case study	Morik, Kopcke
Staying Power of Churn Prediction Models	Risselada, Verhoef, Bijmolt
Predicting Customer Churn in the Insurance Industry: A Data Mining Case Study	Schena
A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance	Sundarkumer, Ganesh & Vadlamani

Table 2. Search terms that were used to find the articles in the literature review.

Index	Keyword	Synonyms & Variants	Source
1.	Predict*	Forecast*, estimate, estimating, estimation, analysis, analyze	http://www.thesaurus.com/
2.	Churn*	Migration, retention, defection	
3.	Insurance	Insurer	

Table 3. Search engines that were used to find the articles in the literature review.

Search Engine	Content
Web of Science	• 100+ years of abstracts • 90+ million records covering 5,300 social science publications • 800 million+ cited references • 8.2 million records across 160,000 conference proceedings
IEEE Xplore Digital Library	• 170+ journals • 1,400+ conference proceedings • 5,100+ technical standards • +/- 2,000 eBooks • 400+ educational courses
Elsevier/Science Direct	• 3,800+ journals

	• 35,000+ books • 14+ million peer-reviewed publications
SpringerLink	• 3,300+ journals • 227,100+ books • 5,900+ book series • 45,000+ protocols • 700+ reference works
ACM Digital Library	• 1,400+ journals • 24,000+ proceedings • 158,000+ books

Appendix B – ROC and Kappa Curves

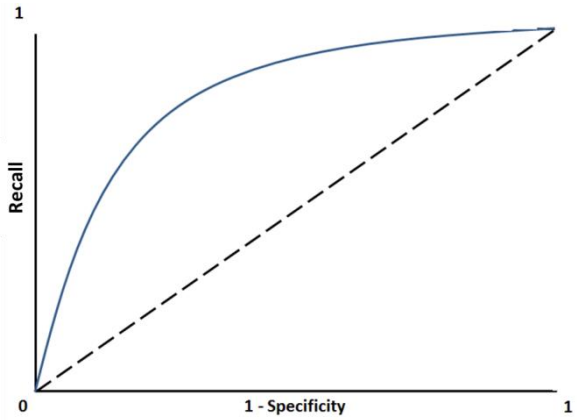


Figure 1. Example of ROC curve.

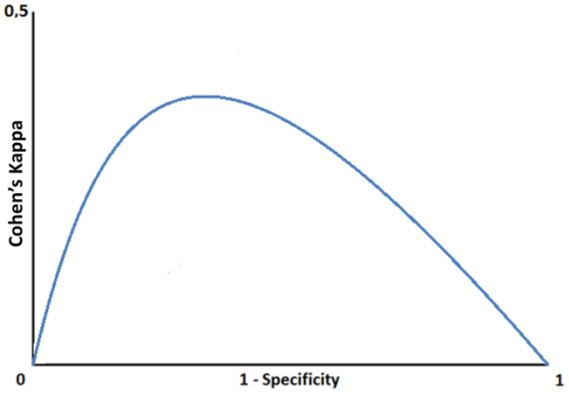


Figure 2. Example of Kappa curve.

Appendix C - The Cross-Industry Process for Data Mining (CRISP-DM)

CRISP-DM stands for cross-industry process for data mining. As the name suggests, this framework can be used for any type of data mining process. It consists of six steps and the methodology is robust and well-proven (Chapman et al., 2000). An illustration of the CRISP-DM framework is given in Figure 1.

Figure 1. The Cross-Industry process for data mining (CRISP-DM).

From Figure 1 it can be seen that the six steps are: 'Business understanding', 'Data understanding', 'Data preparation', 'Modeling', 'Evaluation' and 'Deployment'. The arrows between the different steps show that the process is iterative, allowing for iterations between the different steps (Chapman et al., 2000).

The first step, 'Business understanding', is to understand what the business wants in terms of the objectives of the project. These business needs are than translated into a problem definition and a project plan (Chapman et al., 2000).

The next step is 'Data understanding' and starts with the collection of the data. After the data is collected, the aim is to become familiar with the data. This can be done by making a data description, which describes all the different attributes and by gaining some first insights into the data. Apart from this, the quality of the data is checked during this step (Chapman et al., 2000).

Once the data is collected and the meaning of the different variables are understood by the modeler, the data needs to be prepared. This is the third step in the CRISP-DM framework. The input of this step consists of one or more raw data sets, which cannot yet be used for developing the model. The data first needs to be merged, transformed and cleaned. Apart from this, this step also includes record and attribute selection. The output is a final dataset, which can directly be used for the development of the model(s). It is found that this step often takes the most time (Chapman et al., 2000).

The actual development of the model is the fourth step in the framework. Here choices need to be made about what data mining techniques are going to be used. It is common that different techniques are used, of which some have specific requirements according to the data that is used. Therefore, there are often iterations between the 'data preparation' and 'modeling' steps (Chapman et al., 2000). Evaluation of the model from a data analysis perspective is also done in this step, with the use of several performance measures. According to García et al. (2016) the evaluation criteria that are most common for testing the validity of a model are: accuracy, sensitivity, specificity, the area under the ROC curve, lift chart and loss function (García et al., 2016).

In the 'Evaluation' step of the project the different models are further evaluated. While part of the evaluation is already done in the modeling step, this step has a broader focus. It does not only evaluate the model from a data analysis perspective, but also checks if the model properly achieved the business objectives (Chapman et al., 2000).

The final step 'Deployment' focuses on the implementation of the model and the aim here is to find ways in which the model can be used in daily business.

Appendix D – Feature Selection Survey Marketing Team

Results Survey - 23-05-2017

Do you expect that the age of a customer had a significant influence on whether that customer churned or not?	
	Total
Yes	82.35%
No	17.65%
N=	**17**

Do you expect that the fact that a customer has additional insurance or not had a significant influence on whether that customer churned or not?	
	Total
Yes	52.94%
No	47.06%
N=	**17**

Do you expect that the fact that a customer has dental insurance or not had a significant influence on whether that customer churned or not?	
	Total
Yes	29.41%
No	70.59%
N=	**17**

Do you expect that the fact that a customer has a voluntary excess risk or not had a significant influence on whether that customer churned or not?	
	Total
Yes	41.18%
No	58.82%
N=	**17**

Do you expect that the number of years a customer has his health insurance at Interpolis had a significant influence on whether that customer churned or not?	
	Total
Yes	64.71%
No	35.29%
N=	**17**

Do you expect that the height of the premium a customer pays had a significant influence on whether that customer churned or not?	
	Total
Yes	70.59%
No	29.41%
N=	**17**

Do you expect that the collectivity discount a customer gets had a significant influence on whether that customer churned or not?	
	Total
Yes	41.18%
No	58.82%
N=	**17**

Appendix E – Boxplots of variables 'Premium', 'Gross Premium', 'Collectivity Discount' and 'Lead Time First Reaction'.

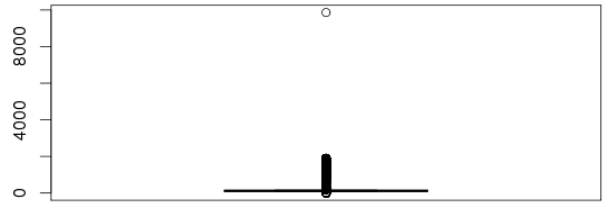


Figure 1. Boxplot for the variable 'Premium'.

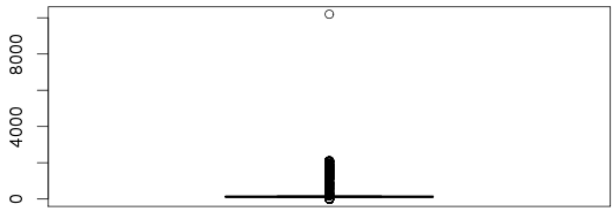


Figure 2. Boxplot for the variable 'Gross Premium'.

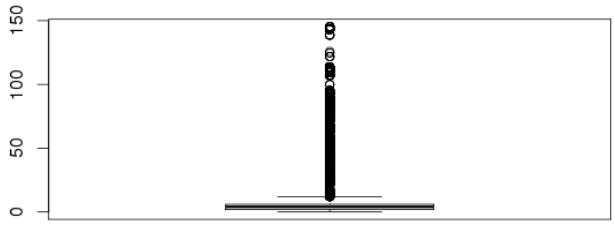


Figure 3. Boxplot for the variable 'Collectivity Discount'.

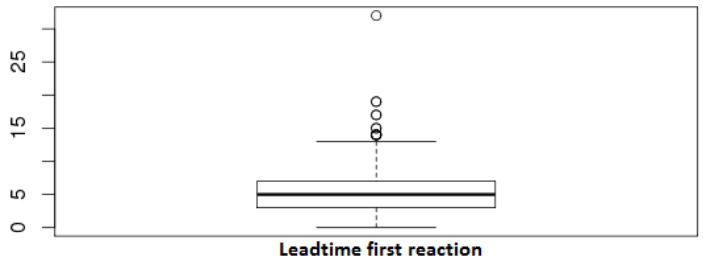


Figure 4. Boxplots for the variable 'Leadtime first reaction'.

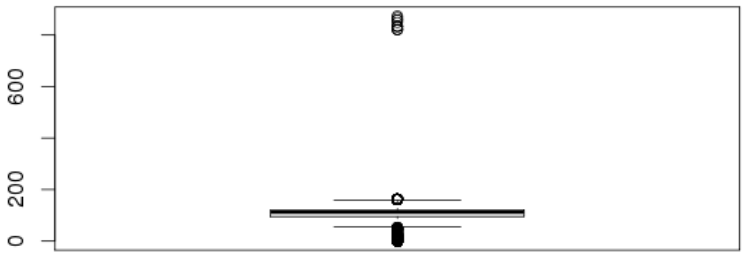


Figure 5. Boxplots of the transformed variable 'Premium'.

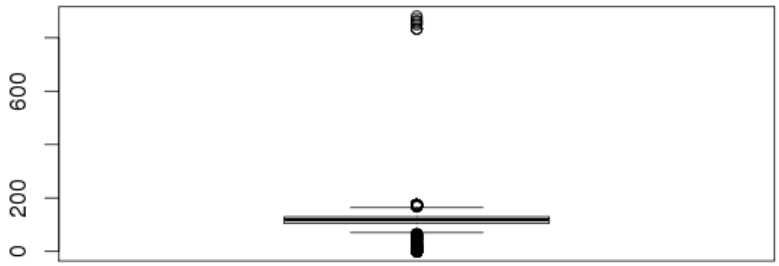


Figure 6. Boxplots of the transformed variable ‘ Gross Premium’.

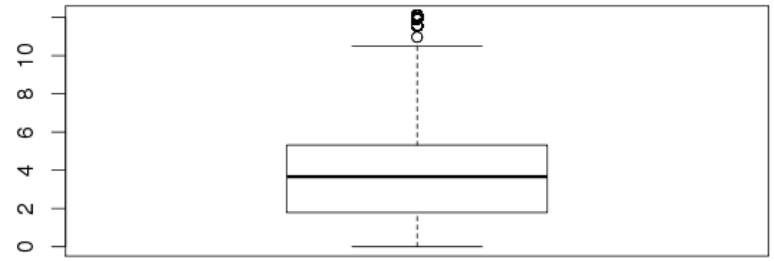


Figure 7. Boxplots of the transformed variable ‘Collectivity Discount’.

Appendix F – Distribution Contact Data (Masked because of confidentiality)

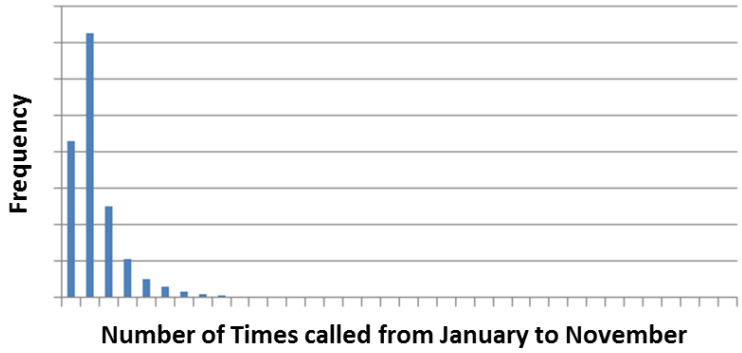


Figure 1. Distribution of variable 'Number of Times called between January and October'.

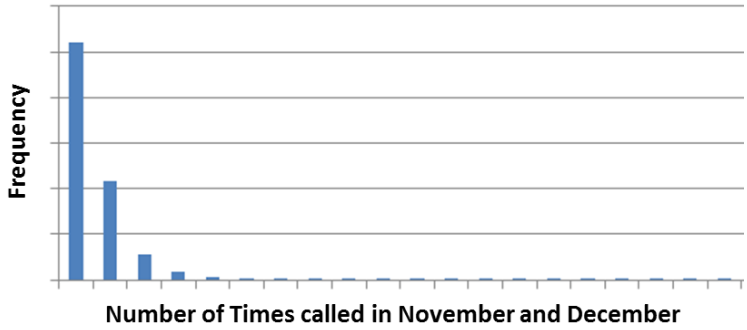


Figure 2. Distribution of variable 'Number of Times called in November and December'.

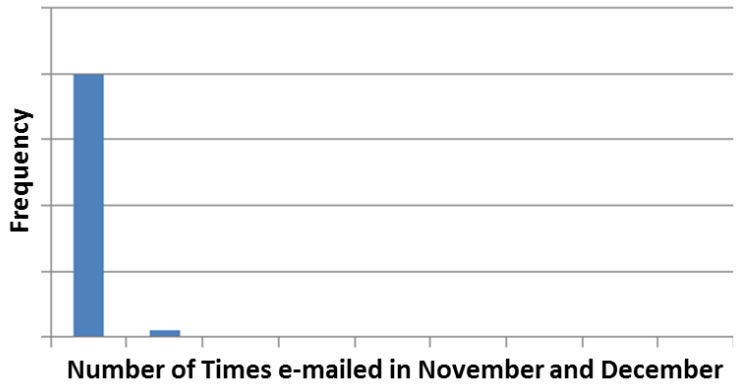


Figure 3. Distribution of variable 'Number of Times e-mailed in November and December'.

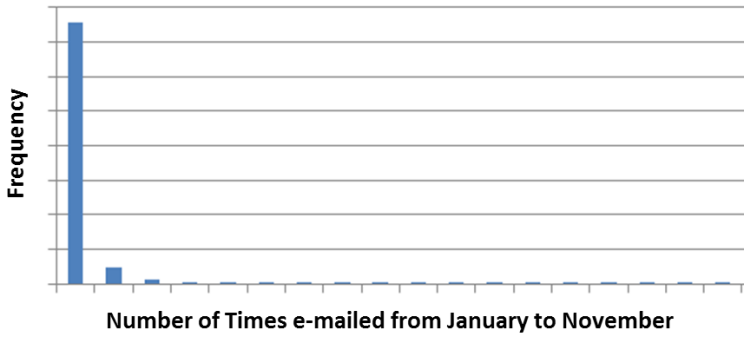


Figure 4. Distribution of variable 'Number of Times e-mailed between January and October'.

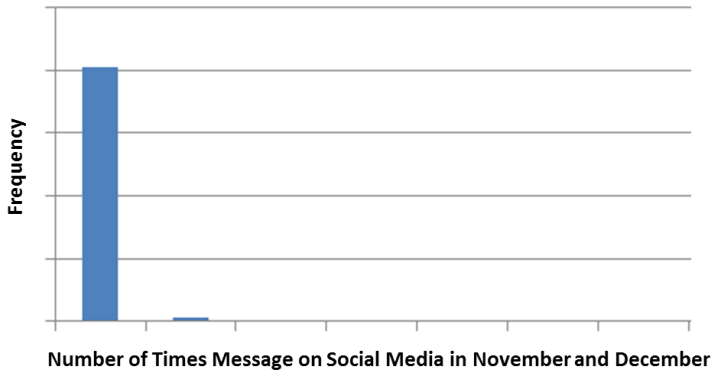


Figure 5. Distribution of variable ‘Number of Times Message on Social Media in November and December’.

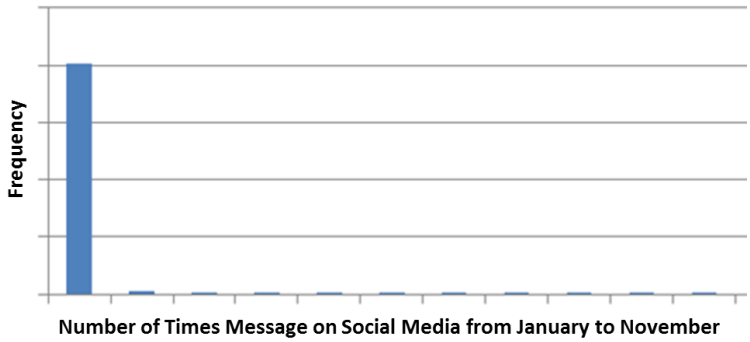


Figure 6. Distribution of variable ‘Number of Times Message on Social Media between January and October’.

Appendix G – Distribution Voluntary Excess Risk (Masked because of confidentiality)

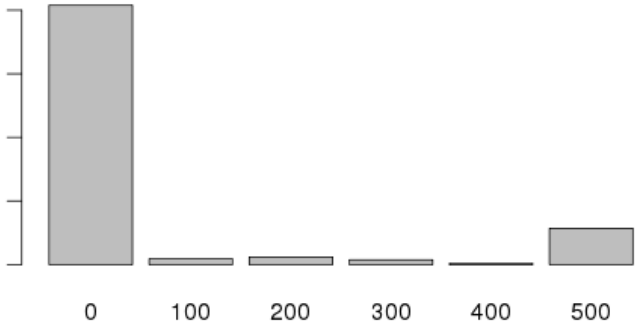


Figure 1. Distribution of the variable 'Voluntary Excess Risk'.

Appendix H – Lambda Optimization Curves for Embedded Method

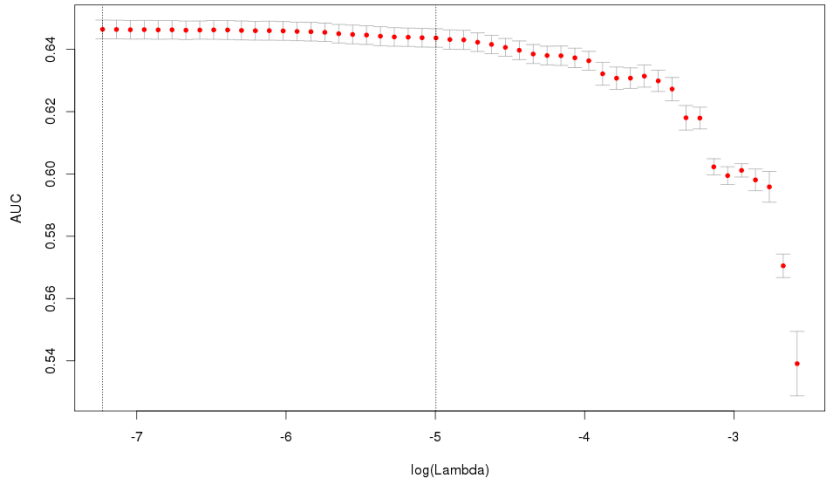


Figure 1. The logarithmic of Lambda vs. the Area under the ROC curve for Model 1.

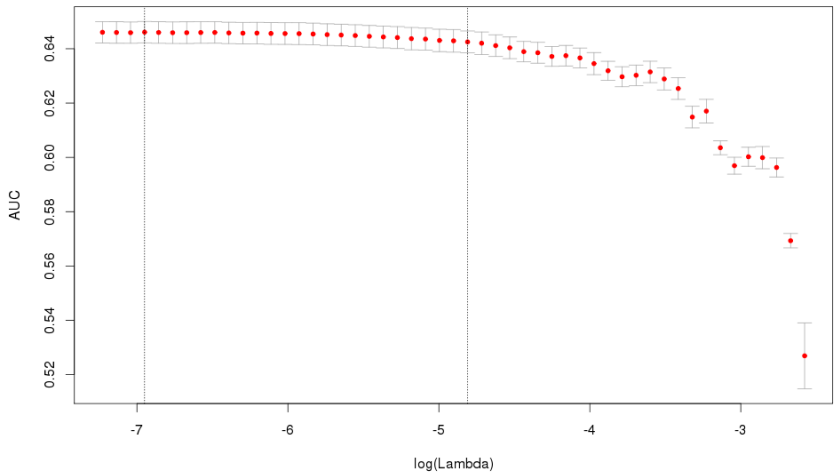


Figure 2. The logarithmic of Lambda vs. the Area under the ROC curve for Model 2.

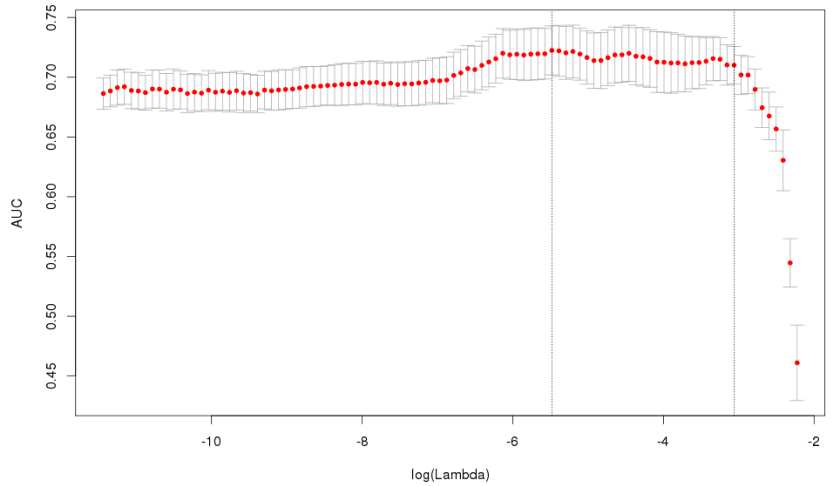


Figure 3. The logarithmic of Lambda vs. the Area under the ROC curve for Model 2a.

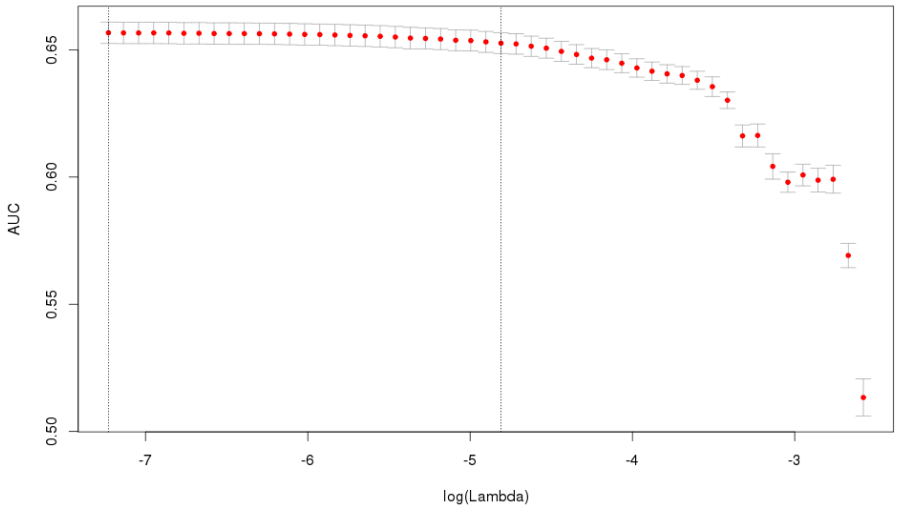


Figure 4. The logarithmic of Lambda vs. the Area under the ROC curve for Model 3.

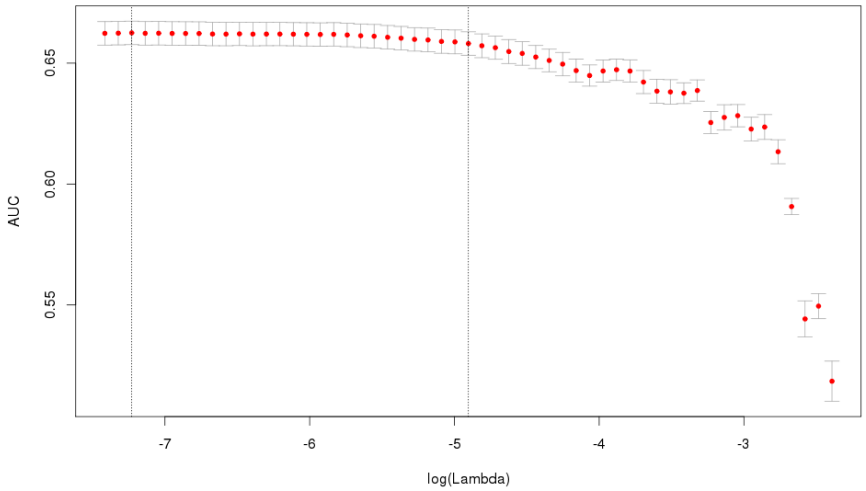


Figure 5. The logarithmic of Lambda vs. the Area under the ROC curve for Model 3a.

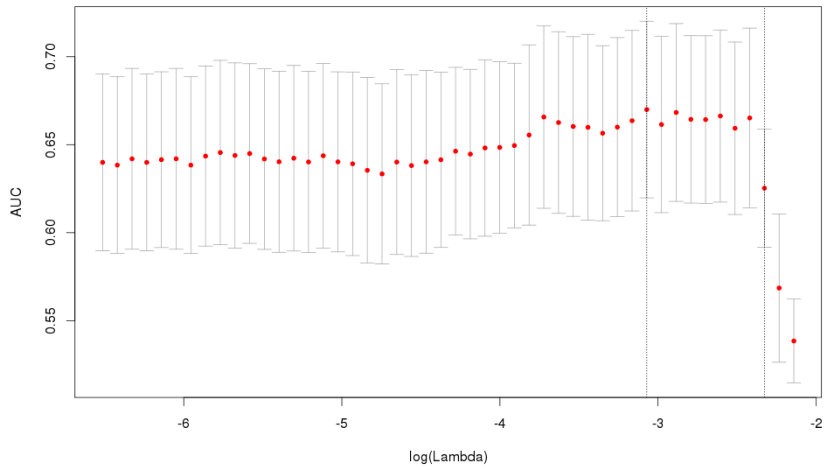


Figure 6. The logarithmic of Lambda vs. the Area under the ROC curve for Model 3b.

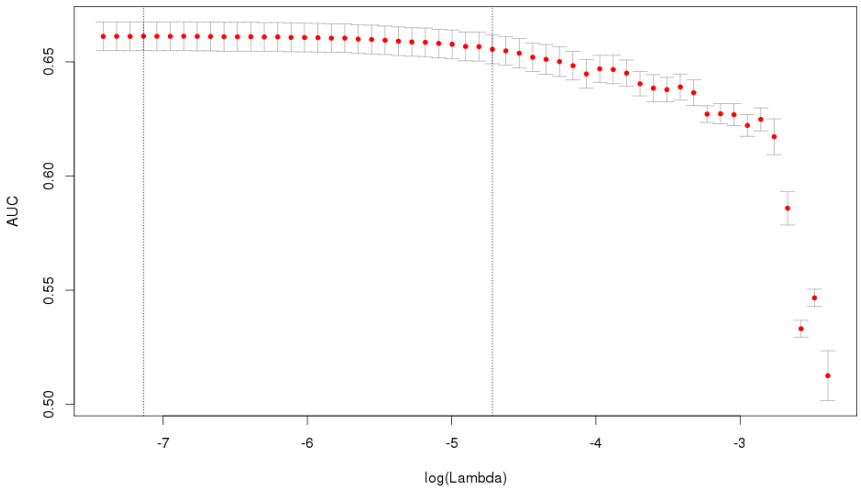


Figure 7. The logarithmic of Lambda vs. the Area under the ROC curve for Model 4a.

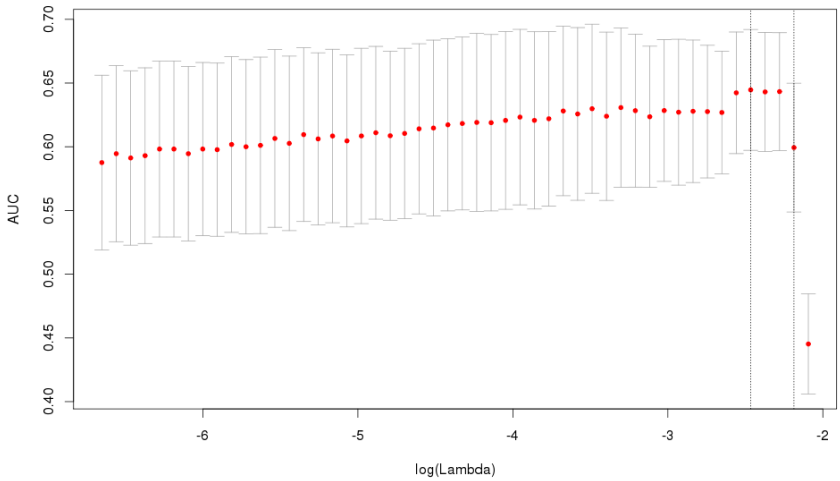


Figure 8. The logarithmic of Lambda vs. the Area under the ROC curve for Model 4b.

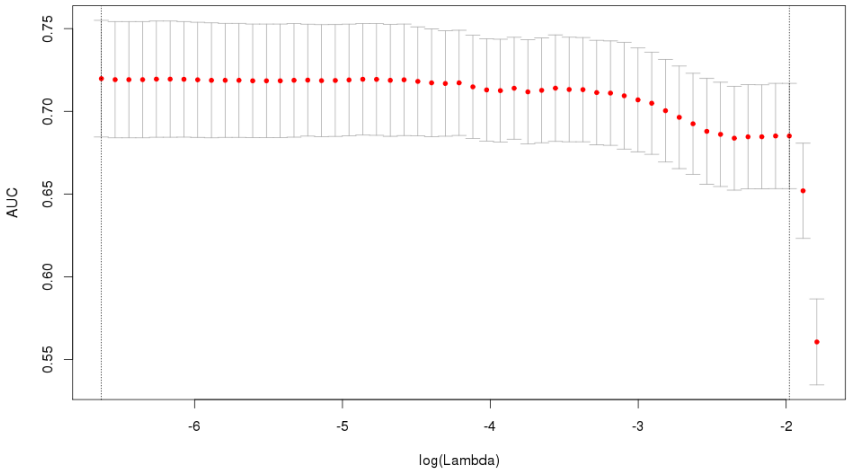


Figure 9. The logarithmic of Lambda vs. the Area under the ROC curve for Model 4c.

Appendix I – Model Summaries

Table 1. Model summary of the best Model 0. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	1.5570	0.0840	18.5440	< 2e-16***
Age Category 2	0.0473	0.0407	1.1620	0.2453
Age Category 3	- 0.3629	0.0435	- 8.3390	< 2e-16***
Age Category 4	- 0.2725	0.0440	- 6.1950	5.81e-10***
Age Category 5	- 0.5223	0.0477	- 10.9590	< 2e-16***
Age Category 6	- 0.9927	0.0491	- 20.230	< 2e-16***
Additional Insurance 1	0.1137	0.0333	3.4130	0.0006***
Contract Duration Category 2	- 0.5955	0.0497	- 11.9810	< 2e-16***
Contract Duration Category 3	- 0.8765	0.0317	- 27.6210	< 2e-16***
Premium	- 0.0098	0.0008	- 11.7510	< 2e-16***

Table 2. Model summary of the best Model 1. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.5727	0.0528	10.8440	< 2e-16***
Additional Insurance Category 2	0.0882	0.0302	2.9200	0.0035**
Additional Insurance Category 'Other'	0.4094	0.0505	8.0990	5.54e-16***
Dental Insurance Category 2	- 0.1005	0.0291	- 3.4580	0.0005***
Dental Insurance Category 3	- 0.4096	0.0654	- 6.2640	3.75e-10***
Dental Insurance Category 4	- 0.4183	0.1942	- 2.1540	0.0313*
Voluntary Excess Risk Category 2	0.3227	0.0332	9.7280	< 2e-16***
Age Category 2	0.1176	0.0433	2.7130	0.0067**
Age Category 3	- 0.2933	0.0472	- 6.2120	5.24e-10***
Age Category 4	- 0.2045	0.0476	- 4.2930	1.76e-5***
Age Category 5	- 0.4307	0.0512	- 8.4180	< 2e-16***
Age Category 6	- 0.8967	0.0525	- 17.0660	< 2e-16***
Contract Duration Category 2	- 0.6020	0.0497	- 12.1010	< 2e-16***
Contract Duration Category 3	- 0.8982	0.0320	- 28.0860	< 2e-16***
Collectivity Discount Binary 1	- 0.1269	0.0352	- 3.6080	0.0003***

Table 3. Model summary of the best Model 2. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.5769	0.0529	10.9150	< 2e-16***
Additional Insurance Category 2	0.0885	0.0302	2.9310	0.0034**
Additional Insurance Category 'Other'	0.4083	0.0506	8.0770	6.66e-16***
Dental Insurance Category 2	- 0.0999	0.0291	- 3.4380	0.0006***
Dental Insurance Category 3	- 0.4090	0.0654	- 6.2540	4.01e-10***
Dental Insurance Category 4	- 0.4185	0.1943	- 2.1540	0.0313*
Voluntary Excess Risk Category 2	0.3228	0.0332	9.7280	< 2e-16***
Age Category 2	0.1191	0.0434	2.7460	0.0060**
Age Category 3	- 0.2923	0.0472	- 6.1890	6.05e-10***
Age Category 4	- 0.2030	0.0477	- 4.2610	2.04e-5***
Age Category 5	- 0.4303	0.0512	- 8.4090	< 2e-16***
Age Category 6	- 0.8960	0.0525	- 17.0530	< 2e-16***
Contract Duration Category 2	- 0.6037	0.0498	- 12.1330	< 2e-16***
Contract Duration Category 3	- 0.9013	0.0320	-28.1540	< 2e-16***
Collectivity Discount Binary 1	- 0.1242	0.0352	- 3.5290	0.0004***
Logged in in November and December 1	- 0.1956	0.0772	- 2.5340	0.0113*

Table 4. Model summary of the best Model 2a. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.4550	0.5531	0.8230	0.4108
Age Category 2	0.3662	0.5451	0.6720	0.5018
Age Category 3	-0.6184	0.5673	-1.0900	0.2757
Age Category 4	0.2270	0.5499	0.4130	0.6798
Age Category 5	-0.8207	0.6469	-1.2690	0.2045
Age Category 6	-1.1810	0.6318	-1.8700	0.0616 .
Contract Duration Category 2	-0.010	0.4828	-0.0210	0.9830
Contract Duration Category 3	-1.1790	0.2815	-4.1870	2.83e-5***
Additional Insurance Category 2	-0.2203	0.2818	-0.7820	0.4343
Additional Insurance Category 'Other'	1.5480	0.6748	2.2930	0.0218*
URL_3.x	-7.5410	3.0800	-2.4490	0.0143*
URL_8.x	-2.9260	1.6830	-1.7380	0.0822 .
URL_inzicht.x	13.9700	11.1000	1.2590	0.2082
URL_30.x	-7.7890	5.7800	-1.3480	0.1778
URL_zorggebruik.x	5.5260	3.7820	1.4610	0.1440
URL_allesineen.x	6.0030	3.2640	1.8390	0.0659 .
URL_80.x	-197.7000	12100	-0.016	0.9870
URL_53.y	9.1130	2.9940	2.0700	0.0385*
URL_68	19.4700	9.9040	1.9660	0.0493*
URL_44.y	9.1130	4.1430	2.2000	0.0278*
URL_21.y	8.0530	3.6150	2.2270	0.0259*
URL_70.y	38.2400	16.3300	2.3420	0.0192*
URL_51.y	3.5220	1.9510	1.805	0.0710*

Table 5. Model summary of the best Model 3. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.4440	0.0536	8.2810	< 2e-16***
Additional Insurance Category 2	0.0795	0.0303	2.6200	0.0088**
Additional Insurance Category 'Other'	0.4082	0.0508	8.0370	9.17e-16***
Dental Insurance Category 2	-0.1200	0.0292	-4.1080	3.99e-5***
Dental Insurance Category 3	-0.4453	0.0659	-6.7630	1.35e-11***
Dental Insurance Category 4	-0.4523	0.1958	-2.3090	0.0209*
Voluntary Excess Risk Category 2	0.3485	0.0334	10.4390	< 2e-16***
Age Category 2	0.1074	0.0435	2.4680	0.0136*
Age Category 3	-0.2914	0.0474	-6.1450	8.00e-10***
Age Category 4	-0.2073	0.0479	-4.3320	1.48e-5***
Age Category 5	-0.4381	0.0514	-8.5250	< 2e-16***
Age Category 6	-0.9529	0.0529	-18.0060	< 2e-16***
Contract Duration Category 2	-0.6001	0.0500	-12.0070	< 2e-16***
Contract Duration Category 3	-0.8831	0.0322	-27.4540	< 2e-16***
Collectivity Discount Binary 1	-0.1454	0.0353	-4.1130	3.90e-5***
Contact in November or December 1	0.3337	0.0332	10.0470	< 2e-16***
Contact between January and October 1	0.2948	0.0271	10.8700	< 2e-16***
Complaint 1	0.4487	0.1867	2.4040	0.0162*

Table 6. Model summary of the best Model 3a. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.0129	0.0897	0.1440	0.8859
Age Category 2	0.1983	0.0676	2.9310	0.0034**
Age Category 3	-0.1222	0.0749	-1.6310	0.1029
Age Category 4	-0.0539	0.0750	-0.7190	0.4720
Age Category 5	-0.2621	0.0806	-3.2510	0.0011**
Age Category 6	-0.7783	0.0787	-9.8950	< 2e-16***
Contract Duration Category 2	-0.3684	0.0746	-4.9380	7.89e-7***
Contract Duration Category 3	-0.6787	0.0475	-14.2910	< 2e-16***
Additional Insurance Category 2	0.0833	0.0433	1.9220	0.0547
Additional Insurance Category 'Other'	0.3306	0.0752	4.3940	1.11e-5***
Dental Insurance Category 2	-0.1939	0.0420	-4.6190	3.86e-6***
Dental Insurance Category 3	-0.4130	0.0897	-4.6050	4.13e-6***
Dental Insurance Category 4	-0.9944	0.3004	-3.310	0.0009***
Voluntary Excess Risk Category 2	0.2779	0.0550	5.0570	4.26e-7***
Call Frequency in November and December Category 1	-0.0426	0.0471	-0.9020	0.3668
Call Frequency in November and December Category 2	-0.0468	0.1006	-0.4650	0.6416
Call Frequency between January and October Category 1	0.1397	0.0533	2.6200	0.0088**
Call Frequency between January and October Category 2	0.3582	0.0761	4.7050	2.54e-6***
Mail Frequency in November and December Category 1	1.8745	0.1003	18.6880	< 2e-16***
Mail Frequency in November and December Category 2	0.9691	0.3412	2.8410	0.0045**
Mail Frequency between January and October Category 1	0.2557	0.0743	3.4430	0.0006***
Social Media Frequency in November and December Category 2	-2.4730	1.2086	-2.0460	0.0407*
Social Media Frequency between January and October Category 1	0.2847	0.1354	2.1030	0.0355*

Table 7. Model summary of the best Model 3b. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.9703	0.5581	-3.5310	0.0004***
Output Model 1	4.1569	1.3959	2.9780	0.0029**

Table 8. Model summary of the best Model 4a. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-0.0412	0.0917	-0.4490	0.6533
Voluntary Excess Risk Category 2	0.2809	0.0555	5.0610	4.17e-7***
Additional Insurance Category 2	0.0755	0.0448	1.6860	0.0918 .
Additional Insurance Category 'Other'	0.3238	0.0767	4.2230	2.41e-5***
Dental Insurance Category 2	-0.2013	0.0433	-4.6440	3.42e-6***
Dental Insurance Category 3	-0.4207	0.0905	-4.6480	3.35e-6***
Dental Insurance Category 4	-1.0048	0.3006	-3.3420	0.0008***
Age Category 2	0.1956	0.0677	2.8910	0.0038**
Age Category 3	-0.1249	0.0750	-1.6650	0.0959 .
Age Category 4	-0.0572	0.0750	-0.7630	0.4457
Age Category 5	-0.2667	0.0807	-3.3060	0.0009***
Age Category 6	-0.7861	0.0786	-10.0010	< 2e-16***
Contract Duration Category 2	-0.3698	0.0746	-4.9560	7.21e-7***
Contract Duration Category 3	-0.6786	0.0476	-14.2670	< 2e-16***
Collectivity Discount Binary 1	0.0300	0.0578	0.5200	0.6033
Call Frequency between January and October Category 1	0.1704	0.0460	3.7060	0.0002***
Call Frequency between January and October Category 2	0.4167	0.0810	5.1420	2.71e-7***
Mail Frequency in November and December Category 1	1.8933	0.0988	19.1630	< 2e-16***
Mail Frequency in November and December Category 2	0.9890	0.3410	2.9000	0.0037**
Mail Frequency between January and October Category 1	0.2854	0.0743	3.8400	0.0001***
Social Media Frequency in November and December Category 2	-2.4785	1.2214	-2.0290	0.0424*
Social Media Frequency between January and October Category 1	0.3003	0.1345	2.2330	0.0255
Not First Time Right between January and October	-0.0989	0.0787	-1.2570	0.2086

Table 9. Model summary of the best Model 4b. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.9703	0.5581	-3.5310	0.0004***
Output Model 1	4.1569	1.3959	2.9780	0.0029**

Table 10. Model summary of the best Model 4c. '***' means significant at 0 level, '**' at 0,001 level, '*' at 0,01 level, and '.' At 0,1 level.

Coefficient	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.5929	0.4218	-3.7760	0.0002***
Output Model 3a Embedded	5.0405	0.8971	5.6180	1.93e-8***
Average NPS Category 2	-0.8957	0.3340	-2.6820	0.0073**
Average NPS Category 3	-1.0781	0.2950	-3.6540	0.0003***