Churn prediction models tested and evaluated in the Dutch indemnity industry

In fulfillment of the requirements for the degree of

Master of Science

Open University of the Netherlands
Faculty Management, Science and Technology

Author: Tom Siemes

First Supervisor: Dr. Jos Schijns MBA
Second supervisor: Prof. dr. ir. Remko Helms

November, 2016

Abstract

Due to global developments customer churn is getting a growing concern to the insurance industry. Technological improvements like the internet makes it much easier for customer to compare their policies, obtain new offers or even churn from one provider to another. The insurance industry therefore has become a heavily competitive market in which insurance companies have to compete to protect and expand their customer base in order to maintain or expand their market position. Thus, retaining customers is becoming more and more important and therefore finding customers who are most likely to leave is a central aspect. Many different techniques are available to identify customers who are most likely to leave, however which technique can be used best is often not clear. Research clarifies that the characteristics of the industry and/or dataset which is used are mostly assessing related to performance. In advance it is impossible to determine the best suited technique to use if previous research in which performance was tested has not been published. This study presents a data mining methodology in which the four most used prediction techniques in literature are tested and evaluated using a real life voluminous insurance company dataset to determine which technique performs best. Using the same dataset makes results comparable and clears out which technique performs best based on the insurance data domain characteristics.

Keywords: Data mining, indemnity insurance industry, customer churn, churn prediction, comparison, logistic regression, decision tree, neural networks, support vector machine, confusion matrix, ROC, AUC

Summary

Due to global developments customer churn is getting a growing concern to the insurance industry. Nowadays it is easy for customers to compare insurance providers on premiums and coverage by using the internet. As a result, the insurance industry is more transparent and has become a heavily competitive market in which insurance companies have to compete to protect and expand their customer base in order to maintain or expand their market position. Retaining customers can lead to profit boosts up to almost 100% by retaining only 5% extra (Reichheld and Sasser, 1990). Research showed that attracting and gaining new customers can is much more expensive then retaining the existing ones (Ng and Liu, 2000; Torkzadeh, Chang and Hansen, 2006). Losing a customer is not only a product less retained, potential future cash flows gained by cross- or upsell opportunities are lost (Gupta, Lehmann and Stuart, 2009). Long term relationships with customers generates higher profits during time, tend to be less sensitive for competitive marketing activities, become less costly to serve and may attract new customers through positive word-of-mouth, while dissatisfied customers perhaps spread negative word-of-mouth (Verbeke, Martens, Mues and Baesens, 2010).

In order to identify customers who are intending to leave the company, several data mining technologies (e.g. logistic regression, artificial neural networks, decision trees) have been widely used by researchers. Churn prediction in the indemnity insurance industry is barely present in literature. It is unclear which available prediction models are best performing on an insurance industry based dataset and which can provide insurance companies accurate insight into potential churners to generate specific marketing actions. There are no specific studies available where churn prediction models are compared/weighed like recommended by Neslin, Gupta, Kamakura, Lu, and Mason (2006). Where the increasing competition in the indemnity insurance industry results in the increasing need for accurate churn models, current available literature still has no answer specific to this industry.

The following main research question is defined as the problem statement of this study:

Which from the most used churn prediction models in literature performs best in the Dutch indemnity insurance industry?

Published literature clears out very obviously which prediction models are most used in the last few years. Specific literature research done for this study (see also appendix B) replenished with the papers from KhakAbi, Gholamian and Namvar (2010) and Tsai and Yu-Hsin Lu (2009), which were had focus on the usage of churn prediction models, extradite that the four techniques logistic regression, decision tree, neural network and support vector machine (which is increasing in the last few years) are the most used techniques. Available literature only provides five papers were the insurance domain

field is approached. Hur and Lim (2005), Risselada, Verhoef and Bijmolt (2010), Smith, Willis and Brooks (2000) and Günther et al. (2014) have used a dataset based on the insurance industry, were Soeini and Rodpysh (2012) are the only researchers who used questionnaire data for input instead of extracted data from an existing database. The four most used techniques are used in the insurance industry, but they are not executed on the same dataset, and they do not contain the same variables and they are not executed in the same insurance domain.

The most common performance measures (Fawcett, 2006) in predictive modelling are cleared out which were extracted from the confusion matrix or contingency table (Davis and Goadrich, 2006; Fawcett, 2006) and an inventory of the variables used in insurance company related papers was made to select a grounded collection of variables. To eventually build a structured and acknowledged methodology for this study, a structured framework of knowledge discovery in databases (KDD), like described by e.g. Fayyad, Piatestsky-Shapiro and Smyth (1996), is chosen to align the methodology of this study. In a five step process – which contain data selection, preprocessing, transformation, data mining and interpretation/evaluation – the answer to the main research question is acquired. The database of a leading Dutch indemnity insurance provider is used to collect a large, real life up to date dataset. SAS Enterprise Guide 6.1 is used to preprocess and transform the data, after which SAS Enterprise Miner 13.1 is used to build and execute the predictive models.

The outcomes of this study show that the performance in terms of the most common performance measures (Fawcett, 2006) of the most used models, unlike presented by e.g. Hur and Lim (2005) or Coussement and Van den Poel (2008), are not very different from each other. To answer the main research question, it depends from which perspective the results are examined. In other words, which goals or objectives – taking into account the research classification context – have to be considered (Labatut and Cherifi, 2012). A perfect measure does not exist (Stehman, 1997, Labatut and Cherifi, 2012) which makes it hard, whether or not impossible, to unanimous identify the best performing model. In the specific case of this study, predict as much churners as possible should be the main objective. Taking into account the main objective, the SVM with polynomial kernel should be used because this model showed the highest recall value using the 50:50 data distribution. If the ROC curve is taken as starting point, the decision tree with the 50:50 data distribution has the highest AUC value which is related to the best overall predictive performance. All the models tested in this study are predicting better than guessing at random and do have additional value (Fawcett, 2006). The lack of churn prediction papers related to the insurance industry, it is not possible to judge accurately whether the results presented in this study are the maximum attainable. Compared to the insurance related paper from Hur and Lim (2005), all of the models tested in this study reach higher or equal accuracy measures. Smith et al. (2000) show higher accuracy

measures compared to the results in this study. Increasing imbalance between the majority and minority class, in this study, leads to improving accuracy but much poorer recall. Nevertheless, by choosing the most suited model several preconditions can influence this choice. The ROC graph showed by Günther et al. (2014) seems very similar with the ones showed in this study, but the actual AUC value is not given. Despite of these similarities, until the ROC space is not in the upper-left-hand corner, potential improvements are still possible (Davis and Goadrich, 2006). Choosing this model is associated with a higher false positive rate, which might be a considered as an unwanted side effect, depending on for example the active customer retention management strategy. Also software availability, server performance or interpretation and communication could be reasons in the selection process of a model (Coussement et al., 2010; Günther et al., 2014).

Altogether, this research showed that a SVM with a polynomial kernel is able to identify the highest number of (possible) churners based on a dataset with the characteristics build by the used methodology. The results of this study suggest that an SVM with polynomial kernel is able to process the input data with best results. Published literature did not contain – unlike widely published in telecom- and retail industry – an extensive comparison based on the same insurance domain dataset like presented in this study, which makes results comparable and which clarifies the performance differences of the most used models based on an insurance related dataset. This study contributes to a more extended insight in the performance of the most used models by testing these models on the same, large, real world dataset which clears out which model performs best based on the insurance related data characteristics gathered from literature.

The limitations of this study also provide perspectives for further research. Due to the focus on the 'most used' models in literature, a bias is created which excludes other models, which are not or hardly not available in literature or models which are primary not suitable for customer churn prediction. The performance a prediction technique depends heavily on the characteristics of the data (King, Feng, and Sutherland 1995; Lim, Loh, and Shih 2000; Perlich, Provost and Simonoff, 2004). Because of the focus on insurance related papers, the range of variables chosen to predict customer churn might not be complete. Variables which have high(er) predictive power related to customer churn, but not directly related to the insurance data domain and therefore not included in this study, are excluded. A broader selection of variables might reveal other important variables with a high(er) measure of association to customer churn in the insurance industry and lead to better prediction performance. The fact is that all of the models score a lot better than guessing at random, but taking into account that the goal in ROC space is to be in the upper-left-hand corner (Davis and Goadrich, 2006), improvement in churn prediction performance is still possible in the insurance domain.

Table of content

Chapter 1

Introduction

This paper focuses on churn prediction in the insurance industry. Chapter 1 contains the introduction to the subject by first discussing and identifying the research area. Second, the problem statement and the research questions will be introduced and explained. Finally, in this chapter will be explained how this research will be executed by combining literature and field research.

1.1 Research area and research goal

Nowadays it is easy for customers to compare insurance providers on premiums and coverage by using the internet. As a result, the insurance industry is more transparent and has become a heavily competitive market in which insurance companies have to compete to protect and expand their customer base in order to maintain or expand their market position. Retaining customers, therefore is becoming more and more important. Reichheld and Sasser (1990) are the grounders of the 'zero defections' theory, which aims on retaining customers in order to prevent profit slump. They mention that companies can boost their profits by almost 100% by retaining only 5% more of their customers. Other researches have also shown that the cost of attracting new customers is more expensive then retaining the existing ones. Torkzadeh, Chang and Hansen (2006) show that it can be up to 12 times more expensive where Ng and Liu (2000) name 3 to 5 times more expensive. A customer has churned when he/she cancels all of his/her policies, either to switch insurance provider or because the insurance is no longer needed (Günther, Tvete, Aas, Sandnes and Borgan, 2014).

Losing a customer is not only a product less retained, potential future cash flows gained by cross- or upsell opportunities are lost (Gupta, Lehmann and Stuart, 2009). In addition, cash flow is directly lost and expensive acquisition has to be executed to maintain steady cash flows. Businesses gain from accurate predictive models if customer lifetime value (CLV) is used for marketing resource allocation (Venkatesan and Kumar, 2004). In earlier research accurate predictions of churn probabilities are an important element within CLV calculations (e.g. Donkers, Verhoef and de Jong, 2007; Blattberg, Malthouse and Neslin, 2009). Long term relationships with customers generates higher profits during time, tend to be less sensitive for competitive marketing activities, become

less costly to serve and may attract new customers through positive word-of-mouth, while dissatisfied customers perhaps spread negative word-of-mouth (Verbeke, Martens, Mues and Baesens, 2011).

Since research has shown that attracting new customers is much more expensive then retaining the existing ones, customer churn is an interesting topic for all businesses. Churn management can be identified as an important part of Customer Relationship Management (CRM), since they strive for establishing long-term relationships and maximizing the value of their customer base (Bolton, 1998; Lemon and Verhoef, 2004; Rust and Chung, 2006). Customer retention is mentioned by Ngai, Xiu and Chau (2009) as a central concern for CRM. Customer satisfaction, which refers to the comparison of customers' expectations with his or her perception of being satisfied, is the essential condition for retaining customers (Ngai et al., 2009). One-to-one marketing is , besides loyalty programs and complaints management, one of the important elements of customer retention. One-to-one marketing refers to personalized marketing campaigns which are supported by analyzing, detecting and predicting changes in customer behaviors (Ngai et al., 2009). Predict customer churn is one of the important elements of customer behaviors. Significant increase of profit can be achieved by only apply small improvements in customer retain management (Van den Poel and Larivière, 2004).

These elements concerning CRM mentioned above are important in many businesses, so also in the Dutch indemnity insurance industry. In addition to these elements which can be regarded as generic applicable in almost every business, there are also specific elements in the Dutch indemnity insurance industry which are decisive for the importance of customer churn prediction in the Dutch Insurance Industry which will be explained in the 'case study' section in appendix A about Centraal Beheer.

Situation and relevance

The research area shows us the increasing need to have highly accurate churn prediction models available in every industry. Recent research (Günther et al., 2014) provides only one model which investigates and predict customer churn in the indemnity insurance industry. The model presented is based on the logit-model extend with generalized additive models (GAM) which is also used by Coussement, Benoit and Van den Poel (2010) in more complex non-linear relationships. Despite of the good performance of the GAM, Günther et al. (2014) as well as Coussement et al. (2010) conclude that there are several reasons for not using GAM. First, the results of a fitted GAM are not easily interpreted of communicated. Second, when using GAM there will be always a danger of over-fitting the model. Therefore, Günther et al. (2010) suggest to apply GAM as a valuable tool in the model building process, rather than as the final model approach for predicting customer churn.

In other industries, such as the telecom- and banking industry, more research is published and more churn prediction models are used and developed with higher accuracy. Salazar, Vélez and Salazar (2012) present a comparison between Support Vector Machine (SVM), developed by Vapnik (2013), and Logistic Regression (LR). In the case of multivariate and mixture of distributions, SVM performs better than LR when high correlation structures are observed in the data. SVM also require less variables than LR to achieve a better (or equivalent) misclassification rate (MCR). Musa (2013) therefore concludes in an empirical comparison between SVM and LR that there is no significant difference in the overall performance measures. Coussement and Van den Poel (2008) show that – only when the optimal parameter-selection procedure is applied – SVM outperforms traditional LR. Research from KhakAbi, Gholamian and Namvar (2010) by reviewing 32 papers, shows a tendency towards techniques like Neural Networks, Decision Tree, Logit, Random Forests and Support Vector Machines. Neslin, Gupta, Kamakura, Lu, & Mason (2006) argue that comparing/weighing different models can improve customer churn prediction.

Churn prediction in the indemnity insurance industry is barely present in literature. It is unclear which available prediction models are best performing in the insurance industry and which can provide insurance companies accurate insight into potential churners to generate specific marketing actions. There are no specific studies available where churn prediction models are compared/weighed like recommended by Neslin et al. (2006). Where the increasing competition in the indemnity insurance industry results in the increasing need for accurate churn models, current available literature still has no answer specific to this specific industry.

Research goal

The research goal provides direction and delineation of this study. Based on the research area the following research goal is arisen:

The research goal of this study is to investigate which available churn prediction models found in literature are relevant and best performing in the Dutch indemnity insurance industry context.

By working towards the research goal, this study fulfils a gap in current literature by executing and comparing different churn prediction models in the (Dutch) indemnity insurance industry on accuracy and applicability. The results will give insight in which churn prediction models are best performing in this industry and the which are applicable in practice. Current literature contains only one specific paper (Günther et al., 2014) which is focused on the indemnity insurance industry by using GAM. However this technique showed good prediction results, it seemed not applicable in practice and it is not verified that this is the best performing model. This study will give insight in the

accuracy and applicability of different prediction models in the Dutch indemnity industry which can contribute to the selecting the most relevant prediction model.

1.2 Problem statement and research questions

Now the research goal is clearly formulated, the problem statement and research questions can be phrased. In order to work towards the research goal of this study, several research questions are formulated. The following main research question is defined as the problem statement of this study:

Which from the most used churn prediction models in literature performs best in the Dutch indemnity insurance industry?

In support of the main research question the following research questions are drawn up:

Research question one:
Which churn prediction models are most used in literature and which ones are best performing?

Research question two:
Which churn prediction models are applicable in the Dutch Indemnity Insurance Industry?

Research question three:
Which variables are most relevant in the current literature to generate highly accurate churn prediction models?

Research question four:
Which churn prediction model generates the most accurate churn prediction results in the Dutch indemnity insurance industry?

Research question five:
Which of the used churn prediction models is most suitable for actual practice taking into account the performance and applicability?

1.3 General approach

In order to answer the research questions, several steps have to be taken which are mentioned in a chronological sequence in this paragraph. First, this chapter (chapter 1) describes the research area and relevance for this study after which the problem statement is formulated. Furthermore the research area where this study is performed is

explained in the 'Case Study: Centraal Beheer' which is added in appendix A because data from this company will be used in this research. The relevance is gained by exploring existing literature using Google Scholar and the digital library provided by the Open University of the Netherlands. Chapter 2 shows an overview of the theoretical background according to the main subject customer churn prediction. The theoretical background actually is the fundamental basis for this study. The selected digital libraries are used to get an create an overview of the current literature focused on customer churn prediction. When the overview of the current literature is build, the research questions 1-3 can be answered. Another important role from this chapter is to create direction by grounding the expectations of this research. In chapter 3 the methodology of the empirical study will be explained. Based on the literature review a selection of churn prediction models will be executed in practice by using SAS Enterprise Guide (SAS EG)/Enterprise Miner (SAS EM) on data from Centraal Beheer. The results of the churn prediction analysis will be presented in chapter 4. A comparison between the outcomes of the churn prediction models as well as the applicability in practice will be evaluated. These results will provide the answers to research questions 4 and 5. Finally chapter 5 contains the conclusions from this research. This chapter also contains the discussion, implications and the limitations of this research replenished with directions for further research.

Chapter 2

Theoretical Framework

In this chapter, the theoretical background of this study is presented. This chapter generates an overview of the available research about customer churn prediction and will ground the aim of this study by explaining existing literature and identifying relevant gaps. Furthermore the theoretical framework will provide scientific literature which is used in this study to justify the chosen design and methods.

2.1 Customer churn and data mining

Firstly, customer churn is defined. In this research the same definition is used as Günther et al. (2010). This means that a customer is churned when he/she cancels all of his/her policy's because the customer switched to competitor provider or the need of insurance is no longer present (people diseased will be excluded). In order to prevent customer churn, many companies want to assess their potential churn population by using churn prediction modeling to retain or even cultivate the profit of potential of the customers (Tsai and Lu, 2010). To manage customer churn effectively, it is important for companies to have or build effective and accurate predictive model(s). To create effective and accurate models in literature, data mining techniques (e.g. logistic regression, artificial neural networks, decision trees) have been widely used by researchers.

2.2 Predictive models in churn prediction

Current literature provides many papers concerning customer churn prediction in different domains fields. This section presents a relevant overview of the existing literature in order to describe the current level of knowledge about customer churn prediction which can support the scientific relevance of this study.

To predict whether a customer will churn or not, a number of data mining techniques are applied for churn prediction, such as artificial neural networks, decision trees, logistic regression and support vector machines. Further on this paragraph the selected most used models are shortly explained. To create a quick overview of techniques which are commonly used in scientific researches, own literature research is combined with two earlier published papers (KhakAbi et al., 2010; Tsai and Yu-Hsin Lu, 2009) to create one

table which is published in Appendix B. The additional papers are gathered trough searching online databases which are listed in a table in appendix C. The table shows the specific techniques which are used, the names of the authors from the research and also the context/domain in which the technique(s) are applied are added.

Table 2.1: Overview of publicated prediction models listed by year.

		Year of publication														
		2000	2002	2003	2004	2005	2006	2007	2008	2009	2010	2012	2013	2014	2015	
no.	***Data mining technique***	**6**	**2**	**3**	**9**	**13**	**18**	**11**	**9**	**31**	**4**	**11**	**10**	**15**	**6**	***Total***
1	Decision tree	1	1	1	2	2	4	3	2	6	1	2	2	2	2	**31**
2	Neural Network	1			2	2	5	1	2	9		4	1	2	1	**30**
3	Logistic regression	1	1	1	2	3	5	2	1	2	1	3	1	3	1	**27**
4	Support Vector Machine	1			2	1	2	1	2	4	1		1	1		**16**
5	Random Forest	1				1		2	1		1					**6**
6	Bayesian Network					1							1	1	1	**4**
7	Survival Analysis							1				1			1	**3**
8	Naïve Bayes						1			2						**3**
9	Self Organizing Maps			1									1			**2**
10	Rough Set Theory							1						1		**2**
11	K-Nearest Neighbor									1				1		**2**
12	K-Means									1				1		**2**
13	Linear Regression					1				1						**2**
14	Association Rules									1			1			**2**
15	Classification tree	1												1		2
16	AdaCost											1				**1**
17	Gradient Boosting Machine								1							**1**
18	Linear Discriminant Analysis												1			**1**
19	AdaBoost												1			**1**
20	Tailor-Buttina													1		**1**
21	Time Series													1		**1**
22	ROCK						1									**1**
23	Regression Forests					1										**1**
24	Sequence Discovery									1						**1**
25	Markov Chains					1										**1**
26	Ensemble Methods				1											**1**
27	Z-score									1						**1**
28	Evolutionary Data Mining Algorithm									1						**1**
29	Linear Classifications									1						1
																148

According to earlier research (KhakAbi et al., 2010; Tsai and Yu-Hsin Lu, 2009) the top 3 of most common models seemed unchanged. The times a paper about churn prediction containing a Support Vector Machine has increased and has entered top 4 and clearly passed by the application of random forests (16 vs. 6 applications found). The four most applied techniques which are gathered during this literature review (29 different techniques where found, see table 2.1), are responsible for more than 70 percent of all the counted techniques. The tendency to techniques like Neural Networks, Support Vector Machine, Decision Tree, Logit and Random Forests described by KhakAbi et al. (2010), has persisted after 2010. Table 2.2 shows the percentages in usage including the cumulative percentage for each of the four techniques including the remaining techniques. Therefore these four techniques are defined as 'most used' and will be explored further on in this paragraph.

Table 2.2: Overview of insurance related papers

no.	*Data mining technique*	*Total count*	*Percentage*	*Cum. Percentage*
1	Decision tree	**31**	20,95%	20,95%
2	Neural Network	**30**	20,27%	41,22%
3	Logistic regression	**27**	18,24%	59,46%
4	Support Vector Machine	**16**	10,81%	70,27%
5	Remaining techniques	**44**	29,73%	100,00%
	Total	**148**	100,00%	

Appendix B gives extra information about the domain in which the research has been conducted. The Telecom Industry is by far the major domain field approaching churn prediction, mainly because there are a huge number of accounts wideband network in the world and the number of accounts is still increasing (Tsai and Lu, 2010). Available literature only provides five papers were the insurance domain field is approached. Hur and Lim (2005), Risselada, Verhoef and Bijmolt (2010), Smith, Willis and Brooks (2000), Soeini and Rodpysh (2012) and Günther et al. (2014) have used a dataset based on the insurance industry, were Soeini and Rodpysh (2012) are the only researchers who used questionnaire data for input instead of extracted data from an existing database. Other domains which are used in churn prediction are the banking industry, subscription services and other financial service industries. The four most used techniques are used in the Insurance Industry, but they are not executed on the same dataset, and they do not contain the same variables and they are not executed in the same insurance domain.

Table 2.3 gives an overview of the papers that focused on the insurance industry. The table directly clears out the differences and similarities.

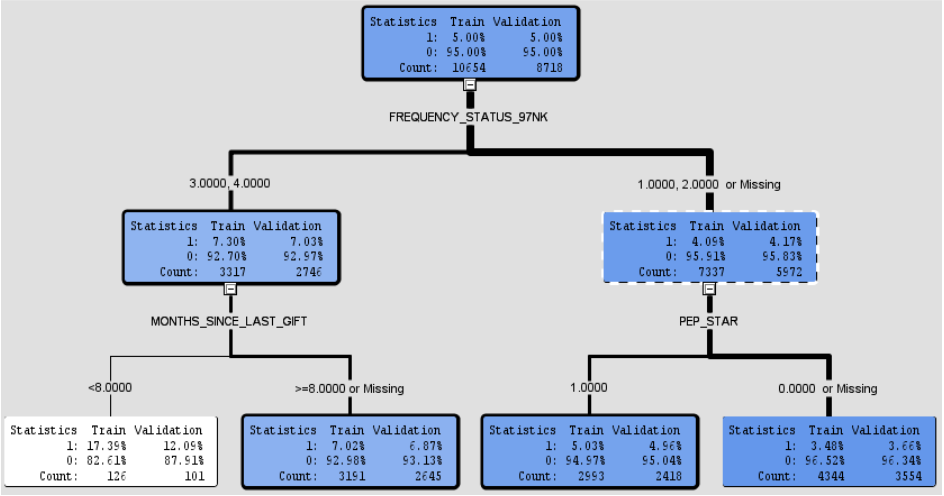
Table 2.3: Overview of insurance related papers

Author	Domain	Technique	# cases
Günther et al., (2014)	Indemnity insurance	GAM	127.961
Hur and Lim, (2005)	Car insurance	SVM, NN	13.200
Risselada et al. (2010)	Health insurance	LR, DT	1.789
Smith et al. (2000)	Car insurance	LR, DT, NN	20.914
Soeini and Rodpysh (2012)	Iranian insurance	DT	300

Decision tree

A decision tree is a tree that, according to this study, is the most applied technique in customer churn management. Osei-Bryson (2004) describes a decision tree (DT) as follows: "A DT is a tree structure representation of the given decision problem such that each non-leaf node is associated with one of the decision variables, each branch from a non-leaf node is associated with a subset of the values of the corresponding decision variable, and each leaf node is associated with a value of the target (or dependent) variable.". Figure 2.1 illustrates an relative simple decision tree constructed with SAS EM. This tree has seven nodes. The topmost node is called the root node, which splits up in branch nodes (any node that has child nodes) and finally ends in the child nodes (any node that does not have child nodes). The colors of the nodes are colored from light to dark, corresponding with the correctly classified observations.

Figure 2.1: Example of a Decision tree

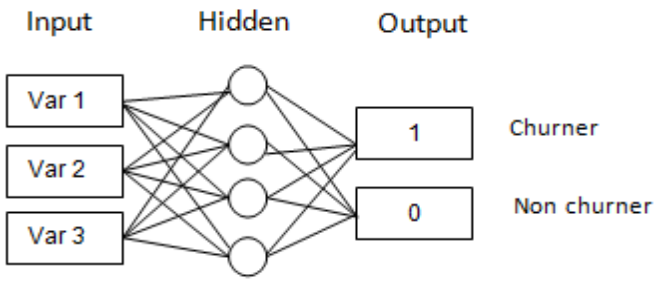


Classification trees (1) and regression trees (2) are the two main types of the decision tree, also known as CART (Tsai and Lu, 2010) developed by Breiman, Friedman, Olshen and Stone (1984), and are non-parametric statistical methods to construct a decision tree to solve classification (such as churn) and regression problems. Because decision trees are easily understood, they are widely used in many and different fields of research such as supplier selection and customer churn (Nie, Rowe, Zhang, Tian and Shi, 2011).

Artificial neural network

Neural networks are, according to this study, the second best most applied technique in customer churn prediction. Neural networks, also called artificial neural networks, are models which can be used for classification and prediction. Neural networks are based on a model of biological activity in the human brain, where neurons are interconnected and learn from experience. A neural network attempts to simulate biological neural systems which learns by changing the strength of the synaptic connection between neurons upon repeated stimulation by the same impulse (West, Dellana and Qian, 2005). Instead of trying to find different relationships by modifying the model, the neural network tries to learn relationships from the data by itself (Shmueli, Patel and Bruce, 2011). Figure 2.2 illustrates a visual example of a possible neural network. It's an interconnected group of nodes were each node represents an artificial neuron and a line represents a connection from the output of one neuron to the input of another.

Figure 2.2: Visualization of a possible neural network

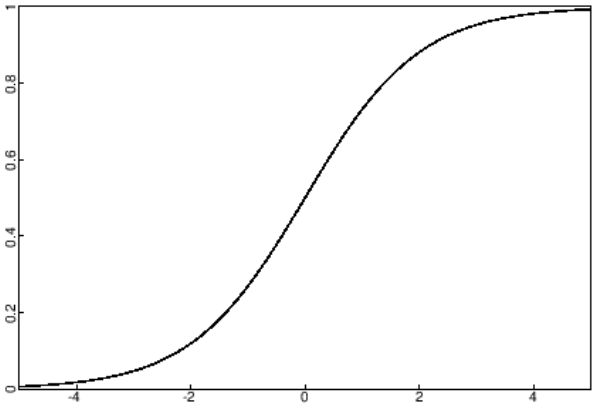


Researchers have studied many different neural network architectures, but the most successful application in data mining is the multilayer feedforward network. Due to different layers of nodes (also known as hidden layers) the neural network creates a fully connected network with a one-way flow and no cycles (Shmueli et al., 2011).

Logistic regression

Logistic regression, developed by Cox (1958), is a technique which can be used to predict or estimate the probability of a binary response (dichotomous, where it can only take two values, such as yes/no, win/lose, dead/alive, churned/retained) based on one or more predictor (or independent) variables, and therefore applicable for churn prediction modeling. By estimating probabilities using a logistic function, logistic regression can measure the relationship between a categorical dependent variable and one ore even more independent variables. Figure 2.3 shows a possible graphical outcome off the probability of churn (y-axis) versus one or more independent variables. It is a widely used statistical technique in many different field domains, also in customer churn management proven and showed by this study.

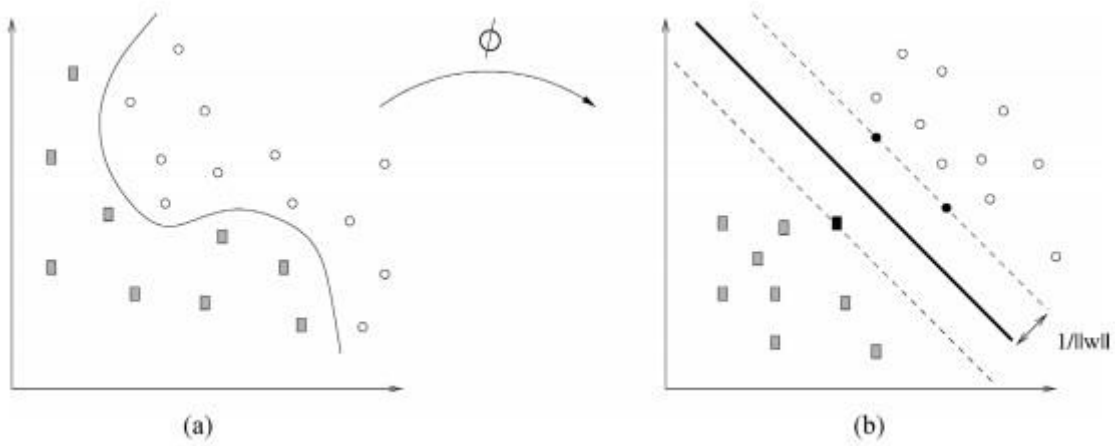
Figure 2.3: Possible S-curve logistic regression by positive impact on churn



Support vector machine

The support vector machine (Cortes and Vapnik, 1995) is a supervised machine learning classification and regression technique that combines computational algorithms with theoretical results. Both characteristics gave it a good reputation and have promoted the use of the technique within various areas (Salazar et al., 2012). In a classification problem like customer churn, the support vector machine searches for the hyperplane that lies furthermost from both classes, known as the optimal (maximal) margin hyperplane (Moguerza and Muñoz, 2006) like illustrated in figure 2.4. Non linearly dataset problems are dealt by the support vector machine by first projecting the data into a higher dimensional feature space and trying to find the linear margin in the new feature space (Farquad, Ravi and Raju, 2014).

Figure 2.4: Original data in (a) and mapped data in future space (b) (Moguera and Muñoz, 2006)



Previous research demonstrated that the support vector machine requires less variables than for example logistic regression to achieve an equivalent misclassification rate (Verplancke, Van Looy, Benoit, Vansteelandt, Depuydt, De Turck, and Decruyenaere, 2008). In other researches (e.g. Coussement and Van den Poel, 2008; Hur and Lim, 2005) the support vector machine proved to be highly accurate. Support vector machines

therefore also perform better than logistic regression when high correlation structures are observed in the data (Salazar et al., 2012). However, the main drawback of the support vector machine is that it creates a 'black box' model because it does not reveal the knowledge learnt during training in human comprehensible form (Farquad et al., 2014; Moguerza and Muñoz, 2006).

2.3 Performance of the most used models

The literature review so far made clear which techniques are most used in the available literature and which dataset domains are used. One of the main questions in this research is to gain insight in which of the most used models are best performing in the insurance industry. To evaluate the performance of churn prediction models several methods (e.g. confusion matrix) can be used. In binary decision problem like churn, where a classifier labels positive or negative, the decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table (Davis and Goadrich, 2006, Fawcett, 2006). A confusion matrix has four categories, as shown in table 2.4. The TP are examples of correctly labeled positives, the FP are negative examples incorrectly labeled as positive, the TN are negatives correctly labeled as negative and the FN are positive examples incorrectly labeled as negative.

Table 2.4: Confusion matrix

	Predicted churn	
Actual churn	**Churners**	**Non-churners**
Churners	True Positive (TP)	False Negative (FN)
Non-churners	False Positive (FP)	True Negative (TN)

The confusion matrix can serve as a basis for several common metrics (Fawcett, 2006) about the performance of the constructed model, which are listed below.

$$\text{fp rate} = \frac{FP}{}$$

$$\text{tp rate (recall, sensitivity)} = \frac{TP}{TP+FN}$$

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

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

$$\text{True negative reate/Specificity} = \frac{TN}{}$$

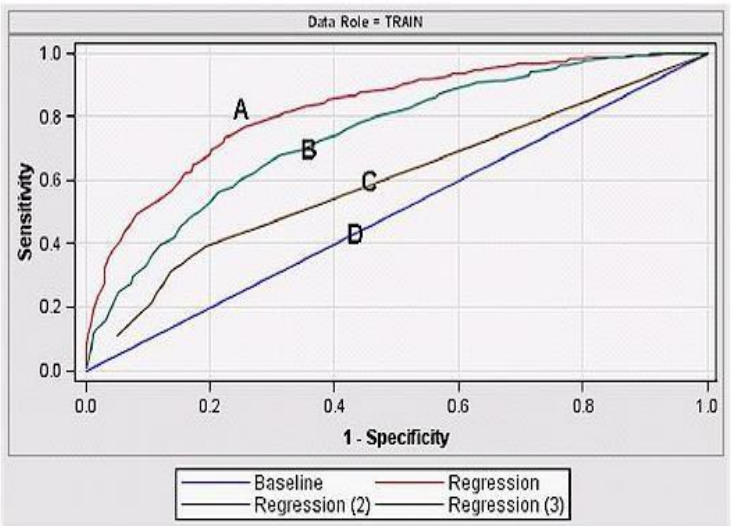
$$\text{F Measure (F1 score)} = \frac{2^{FP+TN}}{1 / \text{precision}+1 / \text{recall}}$$

fp rate: When it's actually no, how often does it predict yes?
tp rate (recall): When it's actually yes, how often does it predict yes?
precision: When it predicts yes, how often is it correct?

accuracy:	Overall, how often is the classifier correct?
Specificity:	When its actually no, how often does it predict no?
F measure:	Approximately the average of precision and recall.

These measures are used most often by users who are concerned with model performance over all submitted observations. However, Provost, Fawcett and Kohavi (1998) argued that for example accuracy measures alone can be misleading. Also mentioned by Yen and Lee (2009) when the data is unbalanced (majority and minority classes). In addition to the confusion matrix, the Receiver Operating Characteristics (ROC) graphs can be useful for organizing classifiers and visualizing their performance. A ROC graph is a two-dimensional graph in which the tp rate is plotted on the y axis and the fp rate is plotted on the x axis (Fawcett, 2006). ROC curves visualized, are very useful while assessing model accuracy in (binary) classification problems (Kaymak, Ben-David and Potharst, 2012). The area under a ROC curve – also called the AUC index – is one of the most commonly used scalars for ranking model performance. The bigger the area under the ROC the better the model performance. Figure 2.5 illustrates a possible ROC chart – extracted from the SAS EM manual – where the performance of three possible regression models is displayed. In this example, model A is outperforming the models B, C , there were the AUC from model A has a larger surface. Model D represents a (poor) model of random predictions, which appear as a flat 45 decree line and shows no discriminatory power. In short, the goal of the ROC space is to be in the upper-left-hand corner (Davis and Goadrich, 2006). This results in a higher AUC value which indicates a better performance. The AUC index is relatively simple to calculate (with SAS EM) and is easily interpreted, so models can be ranked by their AUC index for any given dataset (Kaymak et al., 2012).

Figure 2.5: Three models plotted in a ROC chart in SAS EM



Verbeke et al. (2011) already mentioned that, despite that churn prediction modeling has been extensively researched, there exists no general consensus about the performance of the different prediction techniques. For example they mention the papers of Mozer, Wolniewicz, Grimes, Johnson and Kaushansky (2000) and the papers from Hwang, Jung and Suh (2004) which both applied logistic regression and neural networks to predict customer churn in the telecom industry. In the first study they found neural networks to perform best while logistic regression was performing best in the second study. They also mention that broad benchmarking studies have not been published thus far, and widely varying methodologies and experimental setups impede to cross compare the results of different papers. This literature review underlines this statement, by identifying five papers focused on the insurance industry with five varying methodologies and specific areas of attention. Risselada et al. (2010) suggests that in prior marketing literature logistic regression and classification trees are commonly used by academics and practitioners and that both methods have good predictive performance but based on the papers they reviewed, a superior method has not been identified. The reviewed papers also vary in methodologies and experimental setups (Verbeke et al., 2011).

2.4 Variable selection

As mentioned before there are only five papers (Hur and Lim, 2005; Risselada et al., 2010; Smith et al., 2000; Soeini and Rodpysh, 2012; Günther et al., 2014) found in literature which have focus on the insurance industry. To create focus in application and use of specific variables which are being mentioned and discussed, this paragraph focusses on the variable selections within the four papers (the paper from Soeini and Rodpysh (2012) is excluded because variable selection was not specified) where the specific variables are mentioned.

The used/selected variables which are used in the four papers that focus on the insurance industry field domain are presented in table 2.5. Variables which were mentioned in a paper but which were excluded due to failed significance are not mentioned in table 2.5. In addition to table 2.5, appendix D contains an overview of the used variables including a screen print of the selected variables to add extra information according to the definition/operationalization of the selected constructs.

Table 2.5 shows that only one variable (age) is used in every paper about customer churn in the insurance industry. The variables premium and relationship length were used in three of those papers. Discount, gender and zip code are variables which were used in two of those papers. The variables, mentioned in table 2.5, where considered important by the researchers to predict customer churn. Risselada et al. (2010) and Günther et al. (2014) published results about the importance of the selected variables. In these papers the variables age, lifetime, package type and the fact that a customer rejoined the

company after cancelation where indicated as important predictors of customer churn. Despite of the fact that these four papers are aimed on the insurance industry domain, they have used different selections of variables. A possible cause is the fact that each paper has focus on a specific sub domain within the insurance industry. There was specific focus on health insurance (Risselada et al., 2010), online auto insurance (Hur and Lim, 2005), auto policy holders (Smith et al., 2000), Iranian insurance industry (Soeini and Rodpysh, 2012) and mixed insurance provider (Günther et al., 2014).

Table 2.5: Overview of selected variables in customer churn insurance related papers.

Authors	Smith, Willis and Brooks, 2000	Risselada, Verhoef and Bijmolt, 2010	Hur and Lim, 2005	Günther et al, 2014	Use count
Variable					
Age	V	V	V	V	4
Premium/Last premium	V		V	V	3
Relationship length	V	V		V	3
Discount		V		V	2
Gender	V			V	2
Zip Code	V		V		2
Age endorsement			V		1
MainInsurances				V	1
Car type			V		1
Credit/Debit			V		1
Deductible(auto)			V		1
Driver endorsement			V		1
Discount change				V	1
Family configuration		V			1
No. of home policies				V	1
Income		V			1
Medic. Expense			V		1
New Business	V				1
Number of airbag			V		1
Old insurer's quote (t+1)			V		1
Package type		V			1
Partner				V	1
Premium diff	V				1
Price of the car			V		1
Property Liability			V		1
Rating	V				1
Sum insured	V				1
Sum insured diff	V				1

Surcharge			V		1
Type of coverage			V		1
Vehicle age	V				1
Years on rating	V				1

Risselada et al. (2010) are the only researchers which mentioned specific CRM literature (Bolton, Lemon and Verhoef, 2004) which was used to ground the chosen variables covering the customer characteristics and relationship characteristics. The customer characteristics consist of sociodemographics (Mittal and Kamakura, 2001; Verhoef, 2003) which contains age and family configuration, and socioeconomics (Mittal and Kamakura, 2001; Verhoef, 2003) which contains income. The relationship characteristics consist of length (Bolton, 1998) expressed in the relationship duration and depth (Lemon, White and Winer, 2002; Bolton, Kannan, and Bramlett, 2000) expressed in insurance package type and the type of relationship (individual/collectively). These characteristics are proven to be relevant predictors in CRM literature, are commonly used and therefore also relevant in customer churn prediction.

2.5 Conclusions

The literature review, performed in order of this study, clearly showed the 'most used' churn prediction models. During to the reviewed literature, the decision trees (1), neural networks (2), logistic regression (3) and support vector machines (4) are the four most applied techniques in several field domains. Logistic regression is a widely understand, and a relatively easy method such as the decision tree method. Machine learning, such as artificial neural networks or support vector machines are also widely used and show good or better performance than the 'traditional' logistic regression technique. In churn prediction modelling, the insurance industry is not a common used domain in published literature. Only 5 papers focusing on the insurance domain were identified and they are all focused on other (specific) sub-domains, used different methodologies and different datasets.

A common used method to measure the performance of prediction models is the confusion matrix as mentioned in paragraph 2.3 which can be applied generically. The confusion matrix is a widely used structure in binary decision problem like churn containing four categories (TP, FP, FN and TN) which can support in measuring the performance of different prediction techniques. Based on the four categories server performance metrics can be calculated and the overall performance can be displayed. In addition to the confusion matrix, the ROC curve is commonly used to visualize and rank the performance of a predictive model. The AUC index (area under the curve) represents the performance by measuring the surface underneath the curve (the higher the surface

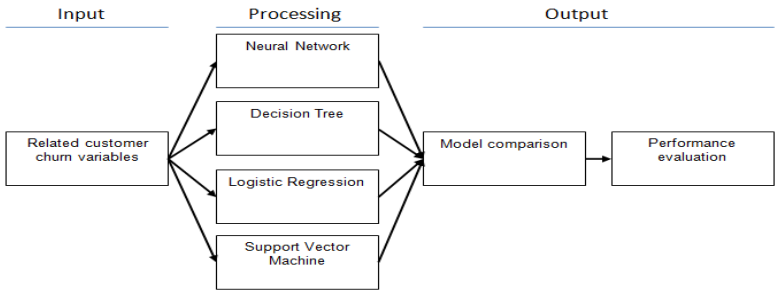
the better the performance). Although the prediction performance can be measured generalizable, in literature there is no consensus about the performance of the different prediction techniques (Verbeke et al., 2011). In prior marketing literature logistic regression and classification trees are commonly used by academics, have both good prediction results but a superior method was not been identified (Risselada et al., 2010). A lot of research is conducted but a vary in methodologies and experimental setups impede comparison (Verbeke et al., 2011).

The variables used in literature about churn prediction modelling in the insurance industry are differentiated. The five papers focusing on the field domain insurance industry, have different focus on sub-domains in the insurance industry and therefore only one variable (age) was used in all of the five papers. The variables age, lifetime, package type and the fact that a customer rejoined the company after cancelation where indicated as important predictors of customer churn (Risselada et al., 2010; Günther et al. ,2014).

2.6 Synthesis

A lot of research is done concerning customer churn prediction in different field domains. There are only five papers available in literature which are focused on churn prediction in the insurance industry. All of these papers are focused on other (specific) sub-domains, used different methodologies and different datasets and are therefore not comparable with each other. The literature review points out that logistic regression, neural networks, decision trees and support vector machines are most used when it comes to churn prediction. It is unclear which of these models are best performing in the insurance industry because available literature does not have the answer on this question. There is no general consensus about the performance of different prediction techniques (Verbeke et al., 2011), so the best practice for every situation cannot be generally pointed out. Hence, to answer the main research question, this research will test and evaluate the performance of the most used churn prediction models in the indemnity industry, which can serve as a base in developing customer churn prediction in the insurance industry field domain.

Figure 2.6: Research model



The research model in figure 2.6 gives an overview of the process which will be followed to identify the best performing model using an indemnity insurance related dataset with specific indemnity variables, chosen after widespread literature review. Furthermore, the influence of the chosen variables will be discussed, to identify possible drivers and their interaction of customer churn in the insurance industry.

Chapter 3

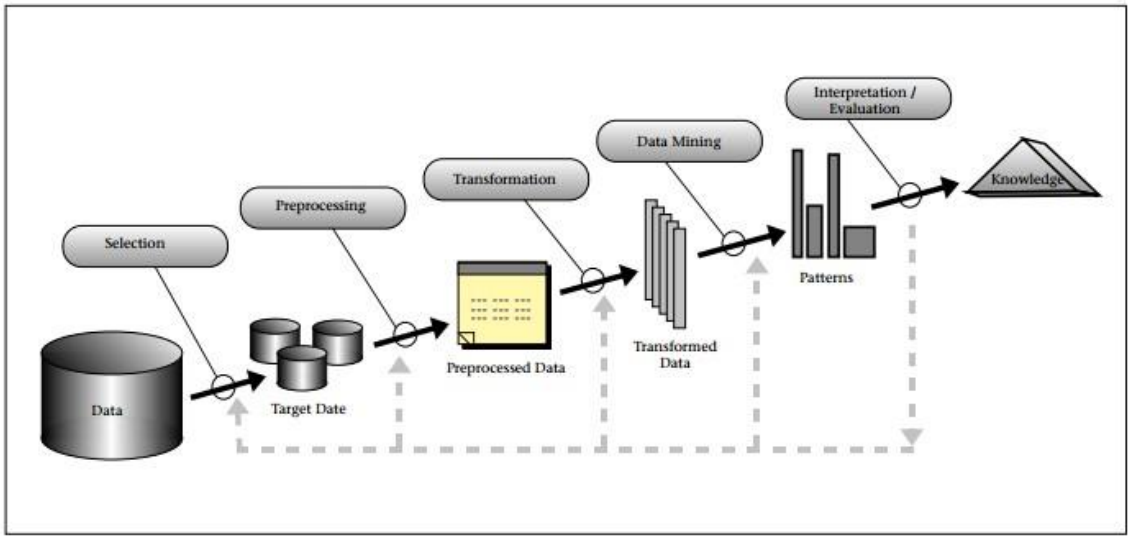
Methodology

In order to answer the research questions formulated earlier in chapter 1 and to accomplish the synthesis formulated in chapter 2, data analysis is required. To be able to execute data analysis, several steps need to be taken. This chapter describes the steps taken and the decisions made in order to execute data analysis.

3.1 Overall approach

To eventually build a structured and acknowledged methodology for this study, a structured framework of knowledge discovery in databases (KDD), like described by e.g. Fayyad, Piatestsky-Shapiro and Smyth (1996), is chosen to align the methodology of this study.

Figure 3.1 An overview of the steps that compose the KDD process. Adapted from *"From data mining to knowledge discovery in databases"* by U. Fayyad, G. Piatestsky-Shapiro and P. Smyth, 1996. *AI magazine*, *17*(3), 37.



Mapping low-level data (which are typically too voluminous to understand and digest easily) to something more abstract (e.g. a descriptive approximation or model of the process that generated the data), compact (e.g. short report) or more useful (e.g. a predictive model for estimation the value of future cases) is the basic problem addressed by the KDD process (Fayyad et al., 1996). The KDD process roughly consist of the

following steps (see figure 3.1): Selection (1), cleaning and preprocessing (2), transformation (3), data mining (4) and interpretation/evaluation (5). In the first step the data needs to be extracted from the database. In this study, the database of Centraal Beheer will be consulted to create a base for testing and evaluating the four selected customer churn models. SAS EG 6.1 is used to extract the data from the database. The second and third steps will prepare the data for the exact data mining activities. Some will be executed by using SAS EG, others by using SAS EM Client 13.1.The fourth step includes executing the data mining activities, which in this case means modeling and execute logistic regression, neural network, decision tree and a support vector machine by using SAS EM. In the last step the results of the executed models will be displayed and discussed. The last step will be elaborated in chapter 4 (Results).

3.2 Data collection

For this study, we consider a portfolio of private insured customers from Centraal Beheer, the largest indemnity insurance provider in the Netherlands. The company indicates six main types of insurance coverages: Car (1), home fire and theft (2), home (3), legal assistance (4), annual traveling (5) and third party insurance (6). In addition to the main type insurance, the company also offers several additional types of coverages, like listed in appendix E. Data from active customers on 01-01-2014 was extracted using SAS EG from the company's database. The two year period from 01-01-2014 to 31-12-2015 is used to follow the customer's behavior to identify whether the customer has churned (not active) or not (still active). For active and non-active customers the same definitions are used like Günther et al. (2014): A customer is defined being active if he has at least one coverage type in one of the six main insurance types or one of the additional insurance types. A customer has churned (non-active, = 1) during the selected period, if all the coverage types are cancelled and the customer has no active coverage types left within any of the insurance types.

Variable selection

Taking into account the literature review concerning the variable selection, only specific variables which were indicated important in literature are selected. The five papers that focus on the insurance industry, have a huge differentiation in the variables which were used, like displayed in table 2.4. The insurance domain in the publication of Günther et al. (2014) is nearly similar to the domain of Centraal Beheer. Therefore, this study includes the same type of variables like selected in the publication of Günther et al. (2014). Other possible relevant variables concerning the indemnity industry are added based on literature when they are used in at least two published papers concerning the indemnity industry.

The variables listed in table 3.1 where directly extracted from the company's database. The variables indicated with (a) are additionally joined (based on zip code) from external data which is integrated in the company's database. The external data is yearly purchased by the procurement department were it is tested and validated for applicability.

Finally this results in a selection of variables as listed in table 3.1.

Table 3.1: Variables directly selected from Centraal Beheer's database

No.	Variable	Description
1	Age	Age of the policy holder
2	Gender	Gender of the policy holder
3	Discount	Discount program
4	Home policies	Number of home policies
5	Number of main coverages types	Count of active main coverage types
6	Partner	Customer's partner has also a policy in the company
7	Premium	Last known yearly premium
8	Relationship length	Length of relationship in years
9	Returned customer	Customer rejoined with one or more coverage types
10	Education (a)	Policy holder's education class
11	Income (a)	Policy holder's income category
12	Social class (a)	Policy holder's social class
13	Number of additional coverage types	Count of active additional coverage types
14	Zip Code	Zip code of the policy holder

(A) = Added variable

The different coverage types are aggregated to one variable. In addition to the paper of Günther et al. (2014), additional coverage types are also included. The variables carcancelled (cancelation of car insurance), health and discount change are excluded from this research, because carcancelled was anticipated by the concerning insurance company and the effect was far from being significant (Günther et al., 2014), the health insurance is not a part of Centaal Beheer's portfolio and the discount change cannot be extracted from the database. In Dutch indemnity industry, it is common to profile customers by income, social class and education. Therefore these variables are added to the dataset. They are simplified into different categories from low to high. The specific values are listed in table 3.2. In summary, the variables 1-9 are added because they are included in the paper of Günther et al. (2014) and some of them are also used in other insurance related papers (see table 2.4). The variables 10-12 are added because these are common profile characteristics in Dutch indemnity industry. Because of a larger product portfolio the number of additional coverage types is also included and extracted from the database. Based on the use of at least two insurance related papers (see Table 2.5) the zip code of the policy holder is also included.

3.3 Data preprocessing and transformation

In real world, data is never totally complete. Data is incomplete, noisy and inconsistent because of not applicable, human- or computer error at data entry, shortcomings in data transmission or from different data sources, etc. Therefore, data cleaning, data integration, data transformation, data reduction and data discretization are major tasks in data preprocessing (Han, Kamber and Pei, 2011). This paragraph describes the steps taken in this study in accordance with the KDD process described by Fayad et al. (1996) to preprocess and transform the data before the actual data mining process begins.

Cleaning

To avoid incomplete, noise and/or inconsistent data, the following criteria are used by selecting the data in SAS EG:

- Customers with no complete information about premium, income, social class, education, number of main coverages types, number of additional coverages types, age, gender or zip code were excluded to avoid missing values;
- Customers who died during the 24 month evaluation period were excluded;
- In the Dutch indemnity industry customers have a 14 day trial period after they contracted an insurance, whether they can decide to keep or cancel their insurance. Insurances with a duration of 14 days or less (excluding temporary travel insurance) are excluded;
- Outliers in premium and age based using the boxplot procedure (Tukey, 1977) are excluded because they may greatly affect modeling results. The interquartile range (IQR) is Q3-Q1. The inner fences are Q1 – 1.5 IQR and Q3 + 1.5 IQR. Data outside these inner fences are identified as outliers and therefore are excluded from the dataset. Highly insured people and extremely elderly people are excluded with this action;
- The variables discount, number of main coverage types, number of additional coverage types, returned customer, partner, home policies and churn are set to 0 if they were not found (e.g. no partner also insured, not ever rejoined, no main coverage types, no additional coverage types, only 1 home policy, still active) in the database;
- Customers under 18 years of age are removed to exclude children.

After preprocessing and cleaning the data, the dataset contains over 867.598 observations. Each observation is tantamount to one customer including the variables as listed in table 3.1. Table 3.2 shows the variables including the specific meaning and/or range of each selected variable.

Table 3.2: Range and values of the selected variables

Variable	Value or range
Age	18 – 99 years
Gender	1 = male, 2 = female
Income (a)	1 – 6 (1 = minimum to 6 = 2.5 times average or above)
Social class (a)	1 – 6 (1 = wealthy to 6 = unskilled)
Education (a)	1 – 3 (1 = low, 2 = medium and 3 = high)
Discount	1 = discount program, 0 = no discount program
Premium	5 – 1438 EUR
Number of main coverage types	0 - 6
Number of additional coverage types	0 - 9
Zip code	1011 - 9999
Relationship length	1 - 41
Returned customer	1 = rejoined, 0 first period
Partner	1 = yes, 0 = no
Home policies	0 = one home policy, 1 = two or more home policies
Churn	1 = churned, 0 = not churned

The dataset containing these elements will serve as a base for the next phase in this study. In the next phase the dataset will be added to a project in SAS EM to prepare for and execute predictive churn analysis.

3.4 Data analysis

To be able to execute data analysis with SAS EM, several steps need to be taken which are mentioned below including an explanation.

1. Create a SAS EM project

The first step in SAS EM is to create a project. To prepare the project for data analysis, the dataset – as described in section 3.3 – is added to the project by defining a library and then add the dataset to the project. Furthermore a diagram is added to the project, which will be used to build the interactive process flow diagram. When the dataset is added to the project, the role (e.g. input, id, target) and level (e.g. interval, ordinal, nominal, binary) of the variables were defined. SAS EM identifies these characteristics automatically, which can be adjusted if necessary. Table 3.3 shows the characteristics which were assigned eventually. The variable numklt identifies every unique customer in the database. The 'Input' role determines if a variable will be included in the analysis related to the 'Target' variable.

Table 3.3: Characteristics role and level of the selected variables

Variable	Role	Level	Name variable dataset
Numklt	ID	Nomial	Numklt
Age	Input	Interval	Age
Gender	Input	Nominal	Gender
Income (a)	Input	Ordinal	Income
Social class (a)	Input	Ordinal	Soc_class
Education (a)	Input	Ordinal	Education
Discount	Input	Binary	Discount
Premium	Input	Interval	Premium
Number of main overage types	Input	Nominal	Count_product_a
Number of additional coverage types	Input	Nominal	Count_product_b
Zip code	Input	Nominal	Zip_code
Relationship length	Input	Interval	Relationship_length_years
Returned customer	Input	Binary	Rejoined_nuklt
Partner	Input	Binary	Partner
Home policies	Input	Binary	Two_homes
Churn	Target	Binary	Churn

2. Descriptive statistics

Once the project is created and the dataset is added, the first step is exploring the data by using the statexplore – and multiplot nodes to gain descriptive statistics from the dataset and visualize them. To include the interval variables in the statexplore node, the chi-square statistics option for interval variables is set to yes including a default number of 10 (visual) bins.

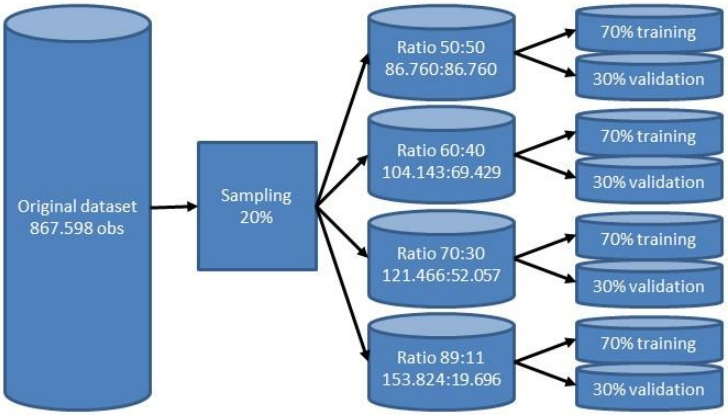
3. Sampling data

Data in real world applications can mostly divided in a majority – and minority class. These so called 'unbalanced data sets' can have undesirable impact when facing a classification problem. The minority class will probably be neglected by the model because it considers this class as noise instead of the most important group (Yen and Lee, 2009). The dataset available for this study also contains a highly unbalanced dataset with approximately 11,35% churners (98.481 observations across 799.117 non-churners). Therefore sampling is needed to manage unwanted effects from unbalanced data. There are many ways of handling imbalanced data such as random oversampling with replacement, random undersampling, focused oversampling, focused undersampling, oversampling with synthetic generation of new samples based on the known information, and combinations of the mentioned techniques (Chawla, 2005). Nevertheless, the most optimal application in each situation cannot be determined and is mostly very domain and classifier dependent, and is usually driven by empirical

observations (Burez and Van den Poel, 2009; Chawla, 2005; Chawla, Japkowicz and Kotcz, 2004). Because sampling methods are worth studying separately and the portfolio available for this study is large enough, the relatively standard sampling method down-sizing is used. This method consists of eliminating, at random, elements of the majority class until it matches the size of the other class which was indicated as a very effective method (Japkowicz, 2000). Because the portfolio for this study is very large and taking into account possible performance issues, the sample size for this study is set to 20 percent of the total dataset, which means that the tested datasets contain 173.520 observations. With this amount of observations this study has the largest portfolio compared to the five published papers which had focus on the insurance industry (see also table 2.4). Taking into account that the standard proportion of churners and non-churners is highly unbalanced and often not the best distribution for learning a classifier (Chawla, 2005), the distributions (non-churners:churners) 50:50, 60:40, 70:30 as well as 89:11 are tested in this study. The majority class and the minority class are randomly eliminated to create the different distributions using four different sample nodes in SAS EM with different level proportions and sample proportions.

To be sure findings are valid and can be generalized to enable predictions to be made about new data, the data is separated into a training (70%) and validation (30%) dataset by using the data partition node in SAS EM. Figure 3.2 illustrates a schematic overview of the datasets which were created following the process as mentioned above.

Figure 3.2: Schematic overview of sampling data to create tables ready for predictive modeling



4. Build predictive models

First the decision tree is built by adding a decision tree node to the diagram and connect it to the data partition node. The decision tree settings are extracted from the SAS website to automatically train and prune a decision tree to an optimal size. To automatically train and prune the decision tree, the maximum depth value is set to 10 (enables SAS EM to train a tree that includes up to ten generations of the root node), the

leaf size value is 8 (constrains the minimum number of training observations in any leaf to eight) and the number of surrogate rules property is set to 4 (enables SAS EM to use up to four surrogate rules in each non-leaf node if the main splitting rule relies on an input whose value is missing).

The second and third model added to the diagram are the regression node and the neural network node. All standard settings are used as preformatted in SASEM. An important aspect from Regression models and neural networks is that they ignore missing values. In the dataset for this study missing values do not occur because they are handled in the preprocessing stage of this study (paragraph 3.3).

The last model added to the diagram is the high performance support vector machine node (HPSVM). The HPSVM does have several tuning options which can influence the obtained accuracy. There are several guidelines available to choose the right settings for the support vector machine (e.g. Ben-Hur and Weston, 2010; Hsu, Chang and Lin, 2003), but there is no ultimate guideline to adjust SVM settings because they are data dependent so several kernels should be tried. Mainly there are two important choices to be made, namely: which kernel (linear, polynomial, radial basis function (RBF) and sigmoid) should be used and which kernel parameters should be used. The prediction performance depends on these two parameters (Hur and Lim, 2005, Ben-Hur and Weston, 2010; Hsu et al., 2003, Kim, 2003). For this study, the guidelines by Ben-Hur and Weston (2010) are used because they are the most actual found in literature. Their premise is to first use the linear kernel as a useful baseline, because in many applications it provides the best results. Gaussian and polynomial kernels often lead to overfitting and a linear kernel is way easier to tune since there is only one parameter (soft-margin constant) which affects performance (Ben-Hur and Weston, 2010). Their doesn't seem to be a way to indicate the right value for penalty parameter C except trying different values. Kim (2003) mentions that the parameter C value has to be between 1 and 100. This setting was adopted earlier by e.g. Hur and Lim (2005) and will also be leading in this study. The polynomial degree which is relevant in the polynomial kernel for the degree of flexibility (Ben-Hur and Weston, 2010) is default value of 2 (1 is equal to linear). After evaluating the results produced by the linear kernel, the other three non-linear kernels will be tried experimentally to see if performance can be improved.

5. Add model comparison nodes

At last, after the models are added to the process flow diagram, the model comparison nodes are added to the diagram. The model comparison node enables the user to compare the performance of competing models using various benchmarking criteria. In this study, the area under the curve (AUC) is selected as selection criteria. As a result, SAS EM will order the connected models by AUC sore in descending sequence. In the diagram build

for this study, different model comparison nodes are added. Two nodes are used to compare the different SVM type models. One node is connected to the remaining models (e.g. decision tree, linear regression and neural network) and de best performing SVM types.

Appendix F shows the SAS EM process flow diagram constructed for this study. At the left of the figure the dataset – prepared an preprocessed by using SAS EG – is specified as a SAS database file. Second, the sample nodes are added to the diagram with four different distributions as well as the descriptive statistics node. The next step is the data partition node which separates the sample datasets into training and validation sets. Once the dataset is separated, the different models are added, configured and connected to the data partition node. Finally the model comparison nodes are added to the grid and connected to the different models. In appendix F only the 50:50 distribution dataset is connected. Instead of modeling every model four times (for every distribution), the models are modeled once and manually connected to each sample node. After connecting the each sample node, the total process flow is executed an the results where extracted. The green check marks in appendix F appear when a node executes without errors or warnings. If an error appears, the program stops executing and a red cross appears in the top right bottom of the node instead of a green check mark. A warning is less severe than an error because the node will run completely. However, the warning is a sign that needs to be checked because it might reveal that things weren't processed as can be expected. A warning results in a yellow international recognized warning symbol in the top right corner of node. Both warnings and errors are not represented in the final process diagram as showed in appendix F.

Unlike mentioned in the fourth point of this paragraph, the RBF and sigmoid nodes are absent. These kernels are experimentally tried on the available datasets, however executing these kernels directly lead to errors due to insufficient memory. To execute these kernels the 'active set' options have to be configured in SAS EM, which is actually not suitable for large datasets (Vogt and Kecman; 2015). Therefore these kernels are not included in the final process flow and will not be discussed in chapter 4.

Chapter 4

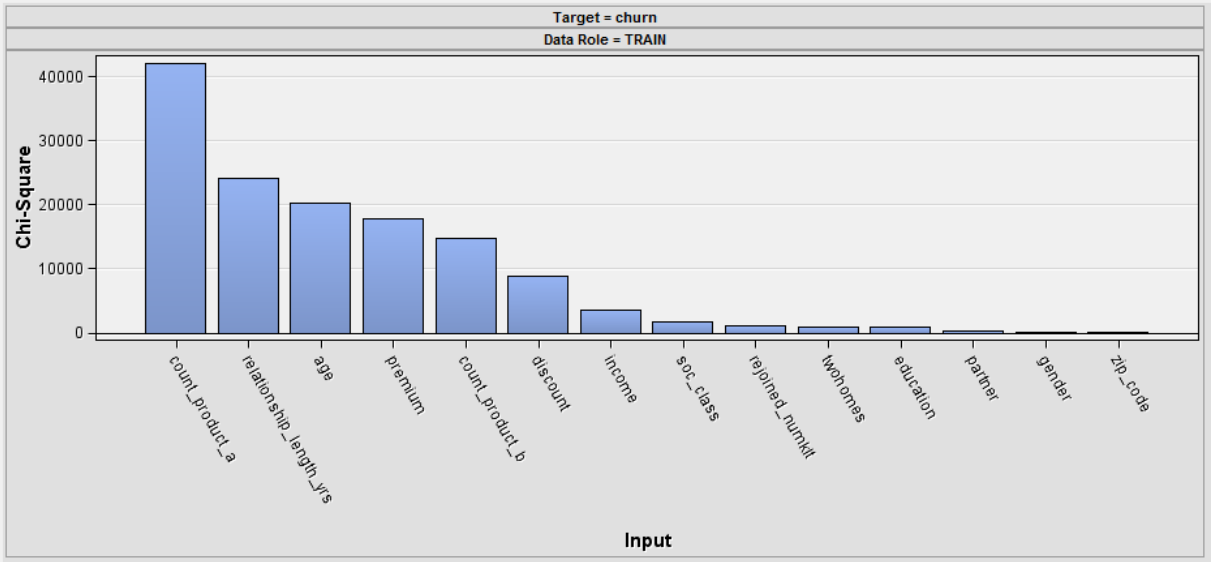
Results

Chapter 3 described the steps taken to prepare and eventually execute the data analysis on data based on the indemnity industry domain. This chapter is used to describe the results produced by following the described methodology. The first paragraph shows some general statistics from the dataset. The following paragraphs are dedicated to each predictive model to point out the generated results by following the methodology designed for this study. In the final paragraph the results of all models are compared.

4.1 Descriptive statistics

This paragraph shows some general statistics generated by using the statexplore node in the process flow connected to the complete database. Figure 4.1 shows the chi-square plot of every input variable, which represents the measure of association to the target variable, in this case churn.

Figure 4.1: Chi-square plot of every input variable



In summary, the most important variables related to churn are product ownership (count_product_a and count_product_b), relationship_length, age, premium and discount are indicated to have a higher measure of association related to the target variable. The remaining variables do have a much lower measure of association related to the target variable. Towards existing literature these outcomes are expected. They also found and

mentioned (huge) dependence between the target churn and these variables (Brockett, Golden, Guillen, Nielsen, Parner and Perez-Marin, 2008; Günther et al., 2014, Risselada et al., 2010).

Table 4.1 shows the summary statistics of the three interval variables included in the dataset. In order to determine if transformation of these variables is necessary to force variables to take on a fairly normalized distribution, the skewness and kurtosis statistics are important to investigate. Skewness and kurtosis values between -2 and 2 are considered acceptable in order to prove normal distribution (Goerge and Mallery, 2001) which means that transformation, to force variables to take on fairly normalized distribution, is not necessary.

Table 4.1: Summary statistics interval variables

Variable	Mean	Standard Deviation	Non Missing	Missing	Min	Med	Max	Skewness	Kurtosis
age	53,66	15,83	867598	0	18	54,0	99	-0,014	-0,687
premium	461,82	325,42	867598	0	5	394	1438,06	0,768	-0,124
relationship_length_yrs	11,20	7,27	867598	0	1	10	41	0,625	-0,436

4.2 Performance results

This paragraph gives an overview of the output results of the four tested models within every selected data distribution. The common performance parameters (Fawcett, 2006) false positive rate, recall, accuracy, precision, F1-measure and sensitivity are displayed. Many different measures exist, but yet, the perfect measure still does not exist (Stehman, 1997, Labatut and Cherifi, 2012). An appropriate measure must be chosen by considering the classification context and formulated objectives (Labatut and Cherifi, 2012). High accuracy on its own for example, does directly mean that it is a good classifier (Yen and Lee, 2009). In customer churn prediction, identifying the customers who are most likely to churn is important, because customer retention programs are aimed to customers who are most likely to churn. Therefore recall (when the model predicts yes, how often is it correct) is considered as the most important metric.

Decision Tree

Table 4.2 shows the performance results generated by the decision tree node in SAS EM for both the train as validation dataset.

Table 4.2: Confusion matrix results decision tree

Decision tree	Distribution			
Dataset: Train	50:50	60:40	70:30	Original
Recall	0,708	0,587	0,367	0,055
FP rate	0,311	0,213	0,089	0,004
Precision	0,695	0,648	0,637	0,615
Specificity	0,689	0,787	0,911	0,996
Accuracy	0,698	0,707	0,747	0,889
F-measure	0,701	0,616	0,466	0,102

Decision tree	Distribution			
Dataset: Validate	50:50	60:40	70:30	Original
Recall	0,703	0,580	0,361	0,049
FP rate	0,317	0,210	0,092	0,004
Precision	0,689	0,647	0,627	0,589
Specificity	0,683	0,790	0,908	0,996
Accuracy	0,693	0,706	0,744	0,888
F-measure	0,696	0,612	0,458	0,091

The results are clearly showing differences between the different data distributions which were tested. Considering recall as the most important metric, the 50:50 distribution is by far the best performing distribution to achieve the goal. Moving from 50:50 distribution closer to the original distribution makes that the minority class will be neglected by the model because it considers this class as noise instead of the most important group (Yen and Lee, 2009). The accuracy metric improves, but the recall metric diminishes faster. The F-measure – which is the metric that incorporates both precision and recall – also shows a trend downwards.

Logistic regression

Table 4.3 shows the performance results generated by the logistic regression node in SAS EM for the train and validation dataset.

Table 4.3: Confusion matrix results logistic regression

Logistic regression	Distribution			
Dataset: Train	50:50	60:40	70:30	Original
Recall	0,704	0,551	0,333	0,006
FP rate	0,343	0,221	0,100	0,001
Precision	0,673	0,625	0,589	0,561
Specificity	0,657	0,779	0,900	0,999
Accuracy	0,681	0,688	0,730	0,887
F-measure	0,688	0,586	0,426	0,013

Logistic regression	Distribution			
Dataset: Validate	50:50	60:40	70:30	Original
Recall	0,700	0,546	0,330	0,006
FP rate	0,344	0,215	0,098	0,001
Precision	0,670	0,628	0,591	0,476
Specificity	0,656	0,785	0,902	0,999
Accuracy	0,678	0,689	0,730	0,886
F-measure	0,685	0,584	0,423	0,011

The logistic regression results show the same trend as the results generated by the decision tree. The 50:50 distribution is by far performing best, taking into account recall as the most important metric. The performance in the other distributions is worse in comparison to the decision tree performance.

Neural network

In table 4.4 the performance results generated by the neural network node in SAS EM for the train and validation dataset and every chosen data distribution.

Table 4.4: Confusion matrix results neural network

Neural Network	Distribution			
Dataset: Train	50:50	60:40	70:30	Original
Recall	0,717	0,544	0,365	0,034
FP rate	0,332	0,202	0,100	0,003
Precision	0,684	0,642	0,610	0,555
Specificity	0,668	0,798	0,900	0,997
Accuracy	0,693	0,696	0,740	0,887
F-measure	0,700	0,589	0,457	0,064

Neural Network	Distribution			
Dataset: Validate	50:50	60:40	70:30	Original
Recall	0,714	0,538	0,357	0,031
FP rate	0,331	0,200	0,102	0,004
Precision	0,683	0,642	0,601	0,518
Specificity	0,669	0,800	0,898	0,996
Accuracy	0,691	0,696	0,736	0,887
F-measure	0,698	0,586	0,448	0,058

The neural network results are showing the same trend as the decision tree and logistic regression with reference to the decreasing performance when the data distribution shifts towards the original distribution in the dataset. The highest recall level is reached in the 50:50 distribution, while the highest accuracy level is reached in the original distribution combined with a very poor sensitivity level.

Support vector machine

This section shows the results generated by using the support vector machine node (HPSVM) in SAS EM. The operation procedure is slightly different compared to the other models, because the SVM node is tested by using different settings. To select the best performing SVM models in order to compare them to the other models, the first selection is made based on the recall. As mentioned before in paragraph 3.4, only the linear and polynomial kernels are executed. The RBF and sigmoid kernel are unfortunately not executable with several parameters due to insufficient server memory. Therefore, these kernels are not included in these performance results.

Table 4.5 explores the recall results of the linear SVM kernel for each of the C values and for every data distribution.

Table 4.5: Recall linear SVM for different C values and data distributions

SVM lineair recall	**Distribution**			
Dataset: Train	50:50	60:40	70:30	Original
SVM lin 1	0,708	0,554	0,184	0,000
SVM lin 25	0,651	0,554	0,184	0,000
SVM lin 50	0,710	0,554	0,184	0,000
SVM lin 75	0,653	0,554	0,184	0,000
SVM lin 100	0,712	0,554	0,184	0,000

SVM lineair recall	**Distribution**			
Dataset: Validate	50:50	60:40	70:30	Original
SVM lin 1	0,708	0,548	0,184	0,000
SVM lin 25	0,654	0,548	0,184	0,000
SVM lin 50	0,709	0,548	0,184	0,000
SVM lin 75	0,655	0,548	0,184	0,000
SVM lin 100	0,711	0,548	0,184	0,000

The linear kernel with C parameter set to 100 produces the highest recall with a 50:50 data distribution. The other data distributions generate much lower recall values just like the previous models which were tested. The linear SVM kernel produces the same percentages of recall for each parameter setting in the distributions which are not equal to 50:50. The C parameter has hardly no effect in the distributions other than 50:50. The figures in the confusion matrix are slightly different, but not enough to change the percentage value. In the original distribution the model does not indicate one single churner, which is useless if the goal is to identify as much churners as possible. The best performing distribution, which is 50:50 just like the other models, will be used in the comparison of the different models.

The results of the polynomial SVM are recorded in table 4.6. For every test distribution and every C parameter the results of sensitivity are displayed.

Table 4.6: Recall polynomial SVM for different data distributions and C parameters.

SVM polynomial recall	**Distribution**			
Dataset: Train	50:50	60:40	70:30	Original
SVM pol 1	0,730	0,546	0,275	0,000
SVM pol 25	0,730	0,545	0,274	0,001
SVM pol 50	0,730	0,546	0,275	0,001
SVM pol 75	0,730	0,546	0,275	0,001
SVM pol 100	0,730	0,546	0,275	0,001

SVM polynomial recall	**Distribution**			
Dataset: Validate	50:50	60:40	70:30	Original
SVM pol 1	0,728	0,539	0,271	0,000
SVM pol 25	0,728	0,539	0,269	0,000
SVM pol 50	0,728	0,539	0,270	0,000
SVM pol 75	0,728	0,539	0,270	0,000
SVM pol 100	0,728	0,539	0,270	0,000

Changing the C parameter using the polynomial kernel does not seem to have much effect. The confusion matrix values show small differences, but the percentages of sensitivity are exactly the same. Therefore, in the comparison with all the other models the default setting of C parameter 1 is admitted.

4.3 Models compared

Paragraph 4.2 showed the performance results of every tested model in this study. In this paragraph the best performing settings for every model are selected and plotted next to each other, to evaluate and compare the performance results of the tested models. Based on the previous results for each model, taking into account the recall values, the 50:50 distribution is used to compare the selected models in this study. Unfortunately there is no such thing as the perfect measure (Stehman, 1997; Labatut and Cherifi, 2012) and therefore several metrics will be included in this comparison, to eventually choose the 'best performing' model. Labatut and Cherifi (2012) advice to use the simplest measures whose interpretation is straightforward. They suggest that distinguishing classes in terms of importance is possible with complex as well with simpler methods. Complex measures are based on different combinations are difficult or impossible to interpret correctly. Despite chance correction is needed for their purpose, no existing estimation for chance correction seems relevant. They recommend, if the

focus has to be made on one specific class, using the true positive rate (recall) and positive predictive value (precision), or a meaningful combination such as the F-measure.

The ROC curve for each of the models, including the diagonal line which represents guessing randomly, is plotted in figure 4.2. The ROC curve gives a clearer view of the overall prediction performance of each model (Fawcett, 2006). The curve shows the tp rate (sensitivity) plotted against the fp rate (specificity) for all possible cut-offs (0-1). The larger the area under the curve (AUC) is, the better the model performs in terms of prediction.

Figure 4.2: Receiver operating characteristic (ROC) curve for each model based on 50:50 distribution training and validation sample.

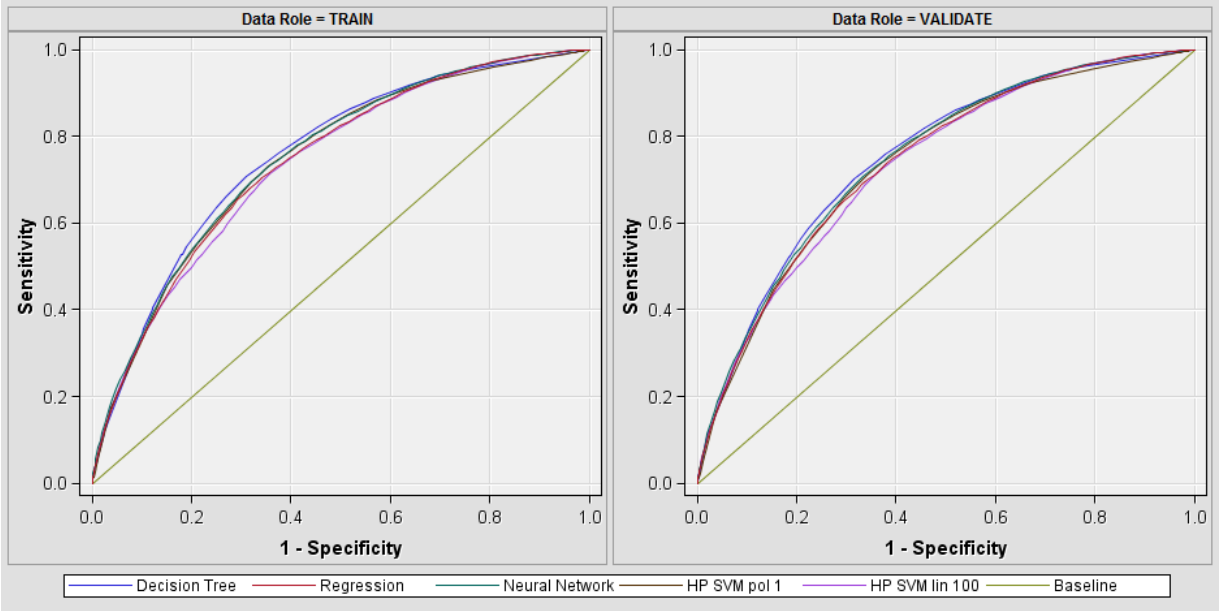


Figure 4.2 shows obviously that all of the selected model perform way better than simply guessing at random. Using one of the models definitely leads to more customers being correctly classified as churned compared to random selection of customers. Based on the visual curves, it is hard to directly choose the 'best performing' line because the differences do not seem that large. To enhance the intelligibility of the differences between the selected models, the AUC index for each model is presented in table 4.7. Taking into account that the larger the AUC value is the better the model performs in terms of prediction SAS EM selects the models with the largest AUC value.

Table 4.7: Area under curve (AUC) for each model based on 50:50 distribution training and validation sample in descending order.

Selected model	Model	AUC train	AUC validate
Y	Decision Tree	0,755	0,752
	Neural Network	0,750	0,749
	HP SVM pol 1	0,747	0,741
	Regression	0,741	0,740
	HP SVM lin 100	0,736	0,735

In addition to the ROC curve and AUC value, the metrics generated from the confusion matrix are recorded in table 4.8. For this overall comparison several extra measures extracted from the confusion matrix to generate an overview as complete as possible.

Table 4.8: Confusion matrix results for every selected model based on 50:50 data distribution for training and validation dataset.

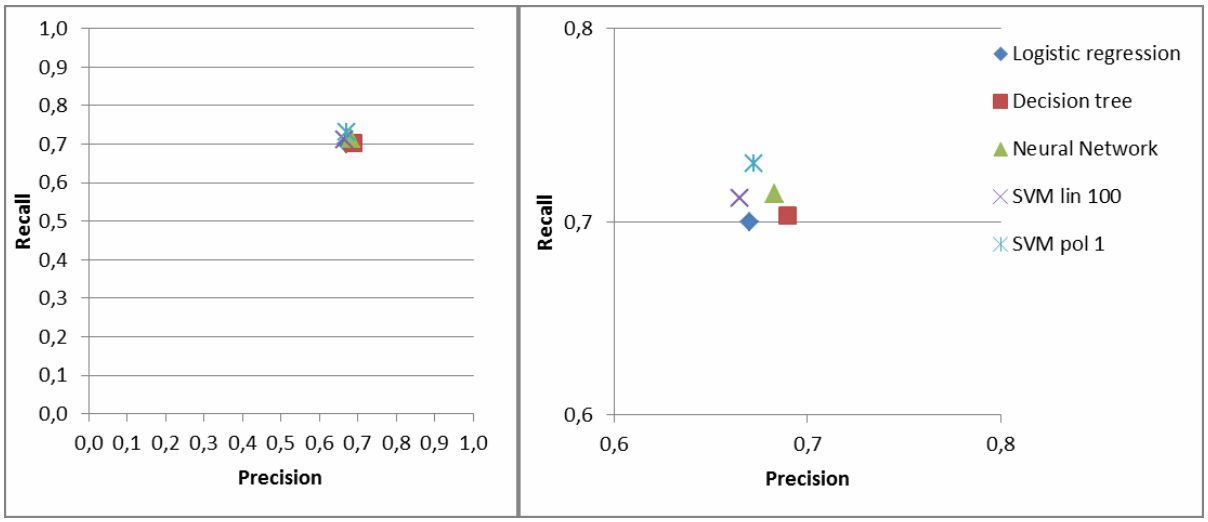
Measure	Model				
Dataset: Train	Decision tree	Logistic regression	Neural Network	SVM lin 100	SVM pol 1
Recall	0,708	0,704	0,717	0,712	0,730
FP rate	0,311	0,343	0,332	0,359	0,355
Precision	0,695	0,673	0,684	0,665	0,673
Specificity	0,689	0,657	0,668	0,641	0,645
Accuracy	0,698	0,681	0,693	0,677	0,687
F-measure	0,701	0,688	0,700	0,688	0,700

Measure	Model				
Dataset: Validate	Decision tree	Logistic regression	Neural Network	SVM lin 100	SVM pol 1
Recall	0,703	0,700	0,714	0,712	0,730
FP rate	0,317	0,344	0,331	0,359	0,355
Precision	0,689	0,670	0,683	0,665	0,673
Specificity	0,683	0,656	0,669	0,641	0,645
Accuracy	0,693	0,678	0,691	0,677	0,687
F-measure	0,696	0,685	0,698	0,688	0,700

Comparing these numbers, it is hard to directly point out a winner which is best performing. Figure 4.3 takes a closer look in the comparison of the different models. Recall and accuracy are plotted for each model in the 50:50 data distribution. Increasing recall seems to lead to an increasing fp rate as well. In the plot on the left hand of the figure the differences between the different models are hard to observe because the differences are not large. The plot on the right hand of the figure zooms between the values 0.6 and 0.8 from recall and precision, which makes the difference between the

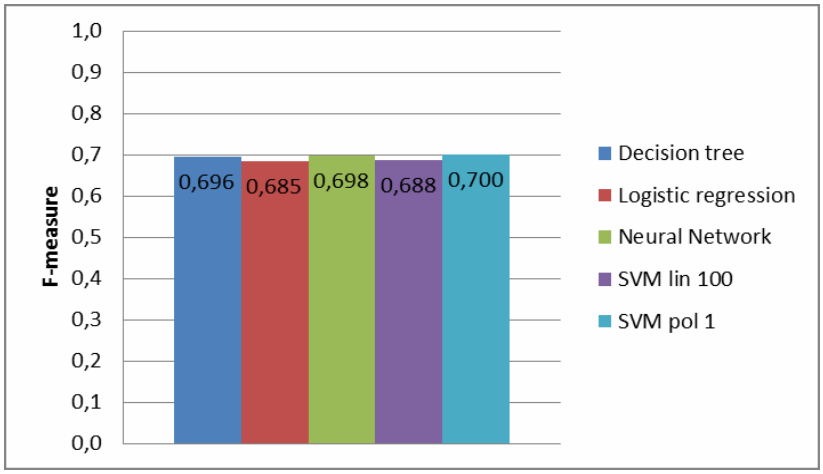
models more obvious. Nevertheless the differences in performance between the models are not very large.

Figure 4.3: Recall vs Accuracy plot for each of the selected models based on 50:50 validation data distribution.



Labatut and Cherifi (2012) suggest that the classic overall success rate or marginal rates should be preferred for the comparison of classifiers. In the confusion matrix tables admitted in this chapter the F-measure is already present. In addition to these tables and to visualize the differences of the F-measure by using a bar chart, figure 4.4 is added. The figure accentuates the fact that the differences between the models – expressed in a meaningful combination of recall and precision – are very close to equal.

Figure 4.4: F-measure plot for each of the selected models based on 50:50 validation data distribution.



Chapter 5

Conclusions and discussion

Nowadays customer churn is a huge concern to many companies, because the churning possibilities are increasing due to changing markets and new technological developments like internet and mobile communication. The indemnity industry is also an industry which is facing increasing churn rates due to global developments (e.g. internet, mobile devices) which simplify the possibilities for customers to actually churn. This chapter is used to formulate an answer to the research questions drawn up for this study. The answers found by executing data analysis fills a gap found in literature concerning the performance of churn prediction models in the indemnity insurance market.

5.1 Conclusions and discussion

After extensively plotting and evaluating the performance results of every model, associated with the theoretical framework built for this study, in this paragraph is used to formulate an answer for every research question which will be discussed simultaneously. This paragraph shows the research questions as mentioned in paragraph 1.2 of this study, starting with the regular research questions and closing with the main research question.

Research question one:
Which churn prediction models are most used in literature and which ones are best performing?

Published literature clears out very obviously which prediction models are most used in the last few years. Specific literature research done for this study (see also appendix B) replenished with the papers from KhakAbi et al. (2010) and Tsai and Yu-Hsin Lu (2009), which focused on the usage of churn prediction models, extradite that the four techniques logistic regression, decision tree, neural network and support vector machine (which is increasing in the last few years) are the most used techniques. The tendency to these techniques, described by KhakAbi et al. (2010), is thereby confirmed by the literature research performed in order to this study.
The answer to the question which models are best performing is a bit harder to explain, because literature shows some contradictions at this point. Verbeke et al. (2011) already

mentioned that, despite that churn prediction modeling has been extensively researched, there exists no general consensus about the performance of the different prediction techniques. For example they mention the papers of Mozer et al. (2000) and the papers from Hwang et al. (2004) which both applied logistic regression and neural networks to predict customer churn in the telecom industry. In the first study they found neural networks to perform best while logistic regression was performing best in the second study. Another point of interest is that broad benchmarking studies have not been published thus far, and widely varying methodologies and experimental setups impede to cross compare the results of different papers. This study underlines this statement concerning the indemnity industry, by identifying five papers focused on the insurance industry with five varying methodologies and specific areas of attention (see also table 2.3). It seems that in prior marketing literature logistic regression and classification trees are commonly used by academics and practitioners and that both methods have good predictive performance but based on the papers they reviewed, a superior method has not been identified (Risselada et al., 2010). In addition to this, the performance of a technique depends on the characteristics of the dataset (King, Feng, and Sutherland 1995; Lim, Loh, and Shih 2000; Perlich, Provost and Simonoff, 2004). The same technique can lead to different performance results if it is applied in a different (data)domain with different characteristics. Basically this is not a strange thought, because every domain has its own singularities, regulations, contract types etcetera which have an effect on data characteristics.

To asses which model performs best, an important methodological choice about the measure which is used to rank the algorithms needs to be taken. Labatut and Cherifi (2012) published a paper whether they examine the most popular accuracy measures and discuss their properties. There are numerous accuracy measures proposed among the years but many of them turn out to be equivalent. Despite the equivalence, they can lead to interpretation problems and may be unsuitable to the purpose of Labatut and Cherifi (2012). Reaching a 99% accuracy in this study would be excellent, but if no single (potential) churner is identified the practical suitability is worthless. Authors very often select an accuracy measure by relying on the tradition or consensus observed in their field. The overwhelming number of measures, the fact that measure properties are not always clear and the heterogeneity of these measures, choosing the most suitable one is a difficult problem (Labatut and Cherifi, 2012). The performance of a model might be influenced depending which performance measure is chosen. In this study a transparent and extensive way of measuring is applied which contributes to understanding and gives insight for further research.

Research question two:

Which churn prediction models are applicable in the Dutch Indemnity Insurance Industry?

In fact, every model – suited for binary classification problems – is applicable in Dutch Indemnity Insurance industry to predict customer churn. Literature looked up for this study does not contain any restrictions nor limitations for the applicability of these models in a specific industry or specific dataset. All models are generalizable applicable regardless to the concerning data domain.

Research question three:

Which variables are most relevant in the current literature to generate highly accurate churn prediction models?

This study is focused on the indemnity industry. Literature review made clear that only five published papers had specific focus on the insurance industry. Table 2.4 shows the variables used in these papers and are used as a basis for this research extracted from insurance industry related papers. Figure 4.1 shows the chi-square plot of every input variable, which represents the measure of association to the target variable churn. In this study, the most important variables related to churn appeared to be product ownership (count_product_a and count_product_b), relationship_length, age, premium and discount. Because these variables have the highest measure of association to the target variable. The other variables do have a much lower measure of association to the target variable churn. The papers from Günther et al. (2014) and Risselada et al. (2010) are the only two insurance related papers which expatiate about the relevance of the used variables in their study. Variables used in this study with a high measure of association to the target variable, were included in their researches and found relevant by the authors in one way or another related to customer churn. This study did not reveal different outcomes concerning which variables are (most) important in customer churn prediction in the insurance industry, but, confirmed the importance of the mentioned variables as reported in literature.

Research question four:

Which churn prediction model generates the most accurate churn prediction results in the Dutch indemnity insurance industry?

This study is aimed to find the best performing and most suitable churn prediction model in the (Dutch) indemnity industry. Based on literature, the most used models are tested and evaluated on a dataset based on the insurance industry domain. The results displayed in chapter 4 of this study, show that, the results are very close to each other independent from which approach they are examined. Taking into account the performance based on the ROC curve, the decision tree generates the highest AUC value (see figure 4.2 and table 4.8). All models perform way better than guessing at random, which means that they have additional value (Fawcett, 2006). From the most common measures extracted from the confusion matrix (Fawcett, 2006), the SVM with the

polynomial kernel has the highest recall value and the decision tree has the highest precision value (see figure 4.8). The model which scores best on the harmonic mean of these two, the F-measure, is the support vector machine with the polynomial kernel. From the point of accuracy the decision tree generates the highest value. Because of the fact these measures, from nearly every perspective, are very close to each other it is hard to identify the fair winner in this comparison. Unfortunately there is no such thing as a perfect measure, an appropriate measure must be chosen according to classification context and objectives (Labatut and Cherifi, 2012). If accurateness is the basic though – and the meaning of accurate is extracted from the dictionary as conforming exactly, errorless or faithfully representing or describing the truth – the measure which represents this the best should be decisive. The model with the highest sum of recall and specificity represents best score in correctness and accurateness. Taking into account the objective in this study, which is identifying as much churners as possible, recall is the most important measure to assess which model predicts best. The SVM with the polynomial kernel is able to predict the highest number of churners. Of course, business goals and side-effects of for example false positives have to be considered carefully, for example relating to the active/invented customer retention strategy.

The comparison of the performance results of this study and other insurance related studies is difficult whether or not impossible, because lots of details are missing and insurance related churn papers are hardly available in literature. In addition, different datasets are used with have their own specific area of attention in the insurance domain with datasets with other characteristics. A SVM executed by Hur and Lim (2005) for example, is not one-to-one comparable to the SVM used in this study, because the different characteristics of the data do influence the performance (King, Feng, and Sutherland 1995; Lim et al., 2000; Perlich et al., 2004). Hur and Lim (2005) and Smith et al. (2000) are the only two insurance related papers which present more than only accuracy measures. Smith et al. (2000) reach a much higher accuracy rate and recall at the expense of a higher false positive rate. The cause of this effect is probably related to the dataset (auto insurance) which is used, but cannot be determined by missing detail information. Hur and Lim (2005) only present average prediction accuracy which is comparable or worse to the accuracy reached in this study using a 50:50 distribution dataset. Increasing imbalance between the majority and minority class, in this study, leads to – as expected – improving accuracy but much poorer recall. Recall is not mentioned by Hur and Lim (2005) using a 80:20 data distribution, so comparing the relation between these values cannot be determined. Compared to the study from Günther et al. (2014), the ROC graph seems quite equal, which means that the AUC value is also quite equal in both studies. Further performance results are not given, which makes it impossible to compare the rest in detail.

Research question five:

Which of the used churn prediction models is most suitable for actual practice taking into account the performance and applicability?

In practice, various considerations should be made to choose the most suitable churn prediction model. Aspects about the performance, applicability or transparency may affect the choice of which model should be used best. If the assignment is aimed to predict the highest number of (potential) churners, the SVM with the polynomial kernel is the best choice to reach this goal. If transparency in the process has the highest worth, maybe the traditional logistic regression is the best choice. A SVM is less suitable related to transparency, because a main drawback of the SVM is that it creates a 'black box' model because it does not reveal the knowledge learnt during training in human comprehensible form (Farquad et al., 2014; Moguerza and Muñoz, 2006). Applicability in terms of interpretation and communication is more often called as a reason to not use a model in practice (Coussement et al., 2010; Günther et al, 2014). Applicability could be an important aspect, if for example not every model is supported by the available software or – like encountered in this study – system resources are insufficient to execute the intended set up. Nowadays several software packages like SAS EM or R are broadly available but not in every business. These software packages simplify building and executing complex statistical models. Nevertheless, specific goals, objectives or limitations can influence the choice of which model could be used best.

Main research question:

Which from the most used churn prediction models in literature performs best in the Dutch indemnity insurance industry?

This study showed that the differences between the most common performance measures (Fawcett, 2006) from each tested model are very similar to each other. In consensus with earlier published papers (Hwang et al., 2004; Levin and Zahavi 2001; Neslin et al. 2006) the differences between performance are rather small, in contradiction to other research (e.g. Hur and Lim, 2005, Salazar et al., 2012) which showed (huge) differences. Due to these contradictions in literature it was not possible to formulate an unambiguous answer in advance. To answer the main research question, it depends from which perspective the results are examined. In other words, which goals or objectives – taking into account the research classification context – have to be considered (Labatut and Cherifi, 2012). A perfect measure does not exist (Stehman, 1997, Labatut and Cherifi, 2012) which makes it hard, whether or not impossible, to unanimous identify the best performing model, especially when the results are very close to each other. In the specific case of this study, predict as much churners as possible should be the main objective. Taking into account the main objective, the SVM with polynomial kernel should be used because this model showed the highest recall value. Nevertheless, by choosing the most suited model several preconditions can influence this choice. Choosing this model is

associated with a higher fp rate, which might be a considered as an unwanted side effect, depending on for example the active customer retention management strategy. Also software availability, server performance (like encountered in this study) or interpretation and communication could be reasons to whether choose a model or not (Coussement et al., 2010; Günther et al., 2014).

Altogether, this research showed that a SVM with a polynomial kernel is able to identify the highest number of (possible) churners based on a dataset with the characteristics build by the used methodology. The differences are not as big as presented by e.g. Hur and Lim (2005) by comparing an SVM, neural network and logistic regression, but gives insight in performance differences using an insurance related dataset build and based on published literature. It is valuable to know which model(s) performs best concerning the specific data characteristics of a specific industry, to make proper choices in business. The results of this study suggest that an SVM with polynomial kernel is able to process the input data with best results. Every model should be able to predict a 'perfect' result, probably depending on the data characteristics used as input. Theoretical review showed several contradictions towards performance (e.g. Wolniewicz et al., 2000; Hwang et al., 2004), which support the fact that there is no general consensus about the performance of the different prediction techniques (Verbeke et al., 2011). Testing and evaluating available techniques on specific data characteristics contributes to theoretical knowledge which can support practical choices or further research. Published literature did not contain – unlike widely published in telecom- and retail industry – an extensive and transparent comparison based on the same insurance domain dataset like presented in this study, which makes results comparable and which clarifies the performance differences of the most used models based on an insurance related dataset. This study contributes to a more extended insight in the performance of the most used models by testing these models on the same, large, real world dataset which clears out which model performs best based on the insurance related data characteristics gathered from literature.

5.2 Implications

Various implications result from this study. From theoretical perspective it was unclear which of the most used models would perform best based on an insurance domain based dataset, because extensive comparison was never been made within this domain. Chapter 2 showed contradictory results about the prediction performance in literature. Earlier research showed rather small differences in performance (Hwang et al., 2004; Levin and Zahavi 2001; Neslin et al. 2006) but also (huge) difference in performance (e.g. Hur and Lim, 2005, Salazar et al., 2012) when different models were compared to each other. This study adds a comparison with rather small differences in performance.

An important point supported by literature (King et al., 1995; Lim et al., 2000; Perlich et al., 2004) is that the performance of a particular method heavily depends on the characteristics of the data. Not only the effect of every variable is important, also the size of the dataset, the degree of normalization and the number of categorical variables are determined (Perlich et al., 2004).

The results from this study suggest that the differences between the tested models are not large, but may serve as a guide for practical use when a same similar dataset is used. From practical perspective, this research showed the performance of each model which can support the choice of which model should be used, taking into account the goals and objectives to achieve, the intended customer retention strategy and the available resources. Due to the impact of data characteristics on prediction performance, a generalizable conclusion cannot be written down that easily. The results of this study suggest that an SVM with polynomial kernel, based on a dataset with the characteristics build by the described methodology, is able to predict the highest number of churners. Odds are, that executing the same analysis on a dataset with the same insurance related characteristics, should approximately produce the same results. Changing data characteristics might lead to different results, because operationalization of variables, sample size and the degree of normalization can be determined in prediction performance (King et al., 1995; Lim et al., 2000; Perlich et al., 2004). It is possible that expanding the selection of variables with characteristic which are less insurance domain specific, may lead to better prediction performance.

This research confirmed the importance of specific variables used to predict customer churn in the Dutch indemnity insurance industry compared to at least two insurance related papers (Günther et al., 2014; Risselada et al., 2010). Taking a closer look into the effects of the most important variables might reveal effects which can be helpful to reduce customer churn. Understanding the specific characteristics of customers who might leave the company is useful input to change or adjust the current customer retention strategy/program(s). The variables having a high measure of association to the target variable are important to investigate. Increasing competition due to global development and the financial potential reached when retaining customers (e.g. Reichheld and Sasser, 1990; Torkzadeh et al., 2006; Ng and Liu, 2000), enterprises are getting more and more interested in customer retention instead of acquiring new customers (KhakAbi et al. 2010). Knowing which aspects need attention to extend/enlarge customer loyalty is essential to develop customer retention programs.

Coussement and Van den Poel (2008) show that – only when the optimal parameter-selection procedure is applied – SVM outperforms traditional LR. The most common settings used in this study showed that SVM can also predict more (potential) churners than other methods. Considering this statement suggests that improvement potential

applying a SVM should still be possible. A closer look into the SVM settings and or data characteristics, might lead to increasing prediction performance using the same dataset.

This research underlines the theoretical premise that accuracy of a model is not enough (Provost et al., 1998; Yen and Lee, 2009) because class imbalance could lead to high accuracy but low recall. Besides accuracy other metrics should be considered in practice, to weigh performance results compared to business goals and objectives. From the invested insurance related papers in this study, only one paper (Smith et al., 2000) is transparent about other metrics than accuracy by displaying the whole confusion matrix. This study contributes to a more transparent overview of the tested models, taking into account more than only accuracy.

5.3 Limitations and implications for further research

This study was aimed to find the best performing and most suitable churn prediction model for the Dutch indemnity based on the 'most used' models in churn prediction literature. By aiming on models used in literature a bias is created because, other models, which are not or hardly not available in literature are excluded from this study. This also counts for models which are primary not suitable for churn prediction. How these models would act and perform by using a dataset build from a company in the insurance domain is not clear. Adding these models in comparison may lead to different outcomes and conclusions.

Because of the fact that the models used in this study are based on the indemnity industry, this specific setup is not directly generalizable to other industries because data characteristics do differ from each business domain. "A general conclusion is that performance of a particular method depends heavily on the characteristics of the data", concluded by e.g. Risselada et al. (2010). The models themselves are generalizable because they are used in many different areas of research (see appendix B). The lack of churn prediction papers related to the insurance industry, it is not possible to judge whether the results presented in this study are the maximum attainable. Compared to the insurance related paper from Hur and Lim (2005), all of the models tested in this study reach higher or equal accuracy measures. Smith et al. (2000) show higher accuracy measures compared to the results in this study. The fact is that all of the models score a lot better than guessing at random, but taking into account that the goal in ROC space is to be in the upper-left-hand corner (Davis and Goadrich, 2006), improvement in churn prediction performance is still possible in the insurance domain.

Variable selection was based on customer churn at insurance companies related literature. Because of this focus, the range of variables chosen to predict customer churn might not be complete. Variables which perhaps have high(er) predictive power related

to customer churn, but not directly related to the insurance data domain and therefore not included in this study, are excluded. A broader selection of variables might reveal other important variables with a high(er) measure of association to customer churn in the insurance industry and can contribute to higher prediction performance. The predictive power of the selected models might be influenced positive or either negative when another variable portfolio is used. Testing and evaluating insurance related datasets with another variable structure could have impact on the results, because performance prediction depends heavily on the characteristics of the data which is used.

In this study the most common settings of the selected predictive models are used. More detailed in depth analysis may lead to higher prediction performance for one or more models, which can lead to different conclusions when comparing the performance results an maybe affects the choice which model should be used best. Methods to find the best settings for each model and combining different (churn) predictive models or boosting techniques could have an effect on the prediction performance, either positive or negative.

Because of the insufficient memory on the server which was used, unfortunately not every SVM setup could be executed. Therefore this study does not contain the results of the prepared setups through which these results are not included in the comparison of models presented in this study. Adding these results might influence the results and conclusions from this study.

Bibliography

Achmea. (n.d.). *Achmea in ahnéén oogopslag*. Retrieved from https://www.achmea.nl/SiteCollectionDocuments/Factsheet-Achmea-in-een-oogopslag.pdf

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. *Life Science Journal*, *11*(3), 75-81.

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10), 552-568.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines.*Data mining techniques for the life sciences*, 223-239.

Blattberg, R. C., Malthouse, E. C., & Neslin, S. A. (2009). Customer lifetime value: Empirical generalizations and some conceptual questions. *Journal of Interactive Marketing*, *23*(2), 157-168.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth. *Belmont, CA*.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.*European Journal of Operational Research*, *164*(1), 252-268.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626-4636.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., & Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection?. *Journal of Risk and Insurance*, *75*(3), 713-737.

Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, *17*(1), 45-65.

Bolton, R. N., Kannan, P. K., & Bramlett, M. D. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the academy of marketing science*, *28*(1), 95-108.

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: a framework and propositions for future research. *Journal of the Academy of Marketing Science*, *32*(3), 271-292.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In*Data mining and knowledge discovery handbook* (pp. 853-867). Springer US.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1-6.

Clemente, M., Giner-Bosch, V., & San Matías, S. (2010). Assessing classification methods for churn prediction by composite indicators.*Manuscript, Dept. of Applied Statistics, OR & Quality, Universitat Politècnica de València, Camino de Vera s/n*, *46022*.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, *37*(3), 2132-2143.

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

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

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

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.

De Bock, K. W., & Van den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications*, *39*(8), 6816-6826.

Donkers, B., Verhoef, P. C., & de Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*, *5*(2), 163-190.

Prasad, U. D., & Madhavi, S. (2012). Prediction of churn behavior of bank customers using data mining tools. *Business Intelligence Journal*, *5*(1), 96-101.

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application.*Applied Soft Computing*, *19*, 31-40

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., Borgan, O. (2014). Modeling and predicting customer churn from an insurance company. Scandinavion Acturial Journal, Vol. 2014, No. 1, 58-71.

Goerge, D., Mallery, P. (2001). SPSS for Windows step by step: a simple guide and reference 10.0 update. *Allyn and Bacon, Toronto*.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of marketing research*, *41*(1), 7-18.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Hu, X. (2005). A data mining approach for retailing bank customer attrition analysis. *Applied Intelligence*, *22*(1), 47-60.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524.

Hur, Y., & Lim, S. (2005). Customer churning prediction using support vector machines in online auto insurance service. In *Advances in Neural Networks–ISNN 2005* (pp. 928-933). Springer Berlin Heidelberg.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications, 26, 181–188.

Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). A Multi-Layer Perceptron Approach for Customer Churn Prediction. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(7), 213-222.

Japkowicz, N. (2000, July). Learning from imbalanced data sets: a comparison of various strategies. In AAAI workshop on learning from imbalanced data sets(Vol. 68, pp. 10-15).

Kaymak, U., Ben-David, A., & Potharst, R. (2012). The AUK: A simple alternative to the AUC. Engineering Applications of Artificial Intelligence, 25(5), 1082-1089.

KhakAbi, S., Gholamian, M. R., & Namvar, M. (2010, January). Data mining applications in customer churn management. In *Intelligent systems, modelling and simulation (ISMS), 2010 International Conference on* (pp. 220-225). IEEE.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, 994-1012.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1), 307-319.

King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, *9*(3), 289-333.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*,*27*(2), 277-285.

Labatut, V., & Cherifi, H. (2012). Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276-286.

Lemon, K. N., White, T. B., & Winer, R. S. (2002). Dynamic customer relationship management: Incorporating future considerations into the service retention decision. *Journal of marketing*, *66*(1), 1-14.

Lessmann, S., & Voß, S. (2009). A reference model for customer-centric data mining with support vector machines. *European Journal of Operational Research*, *199*(2), 520-530.

Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing*, *15*(2), 2-22.

Li, G., & Deng, X. (2012). Customer churn prediction of china telecom based on cluster analysis and decision tree algorithm. In *Emerging research in artificial intelligence and computational intelligence* (pp. 319-327). Springer Berlin Heidelberg.

Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, *40*(3), 203-228.

Liou, J. J. (2009). A novel decision rules approach for customer relationship management of the airline market. *Expert systems with Applications*, *36*(3), 4374-4381.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2014). A customer churn prediction model in telecom industry using boosting. *Industrial Informatics, IEEE Transactions on*, *10*(2), 1659-1665.

MedCalc (2015, November 27). *Logistic regression*. Retrieved from https://www.medcalc.org/manual/logistic_regression.php

Mittal, V., & Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of marketing research*, *38*(1), 131-142.

Moguerza, J. M., & Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, 322-336.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *Neural Networks, IEEE Transactions on*,*11*(3), 690-696.

Musa, A. B. (2013). Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, *4*(1), 13-24.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273-15285.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Ng, K., & Liu, H. (2000). Customer retention via data mining. *Artificial Intelligence Review*, *14*(6), 569-590.

Kim, K., Jun, C. H., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, *41*(15), 6575-6584.

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

Pendharkar, P. C. (2009). Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. *Expert Systems with Applications*, *36*(3), 6714-6720.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, *4*(Jun), 211-255.

Prinzie, A., & Van den Poel, D. (2006). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM.*Decision Support Systems*, *42*(2), 508-526.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998, July). The case against accuracy estimation for comparing induction algorithms. In *ICML* (Vol. 98, pp. 445-453).

Qi, J., Zhang, L., Liu, Y., Li, L., Zhou, Y., Shen, Y., ... & Li, H. (2009). ADTreesLogit model for customer churn prediction. *Annals of operations research*, *168*(1), 247-265.

Reichheld, F. P., & Sasser, W. E. (1990). Zero defeciions: Quoliiy comes to services. *Harvard business review*, *68*(5), 105-111.

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198-208.

Rust, R. T., & Chung, T. S. (2006). Marketing models of service and relationships. *Marketing Science*, *25*(6), 560-580.

Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate?. *Revista Colombiana de Estadística*, *35*(2), 223-237.

Sharma, A., Panigrahi, D., & Kumar, P. (2013). A neural network based approach for predicting customer churn in cellular network services. arXiv preprint arXiv:1309.3945.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley and Sons.

Shao, H., Zheng, G., & An, F. (2008, October). Construction of Bayesian Classifiers with GA for Predicting Customer Retention. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on* (Vol. 1, pp. 181-185). IEEE.

Smith, K. A., Willis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study.*Journal of the Operational Research Society*, 532-541.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, *62*(1), 77-89.

Soeini, R. A., & Rodpysh, K. V. (2012). Applying Data Mining to Insurance Customer Chu rn Management. *International Proceedings of Computer Science and Information Technology*, *30*, 82-92.

Song, G., Yang, D., Wu, L., Wang, T., & Tang, S. (2006, December). A mixed process neural network and its application to churn prediction in mobile communications. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on* (pp. 798-802). IEEE.

Song, H. S., Kim, J. K., Cho, Y. B., & Kim, S. H. (2004). A personalized defection detection and prevention procedure based on the self-organizing map and association rule mining: Applied to online game site. *Artificial Intelligence Review*, *21*(2), 161-184.

Torkzadeh, G., Chang, J. C. J., & Hansen, G. W. (2006). Identifying issues in customer relationship management at Merck-Medco. Decision Support Systems, 42 (2), 1116_1130.

Tsai and Yu-Hsin Lu, C. F. (2009). Data Mining Techniques in Customer Churn Prediction. *Recent Patents on Computer Science*, *3*(1).

Tukey, J. W. (1977). Exploratory data analysis.

Obiedat, R., Alkasassbeh, M., Faris, H., & Harfoushi, O. (2013). Customer churn prediction using a hybrid genetic programming approach. *Scientific Research and Essays*, *8*(27), 1289-1295.

Osei-Bryson, K. M. (2004). Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*, *31*(11), 1933-1945.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1-9.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of marketing*,*68*(4), 106-125.

Verbraken, T., Verbeke, W., & Baesens, B. (2014). Profit optimizing customer churn prediction with bayesian network classifiers. *Intelligent Data Analysis, 18*(1), 3-24. doi:10.3233/IDA-130625

Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F., & Decruyenaere, J. (2008). Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Medical Informatics and Decision Making*, *8*(1), 56.

Vries, C.G. de., Schoenmaker, D., Streppel, J.B.M., Verhoeven, H.B.A. (2015). *Nieuw leven voor verzekeraars Rapport van de Commissie Verzekeraars.* Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2015/03/05/rapport-commissie-verzekeraar.pdf

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354-2364.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development.Journal of marketing, 67(4), 30-45.

Vogt, M., & Kecman, V. (2005). Active-set methods for support vector machines. In Support vector machines: theory and applications (pp. 133-158). Springer Berlin Heidelberg.

Wang, Y. F., Chiang, D. A., Hsu, M. H., Lin, C. J., & Lin, I. L. (2009). A recommender system to avoid customer churn: A case study. Expert Systems with Applications, 36(4), 8071-8075.

Wang, Y., Satake, K., Onishi, T., & Masuichi, H. (2015). Improving Churn Prediction with Voice of the Customer.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, *32*(10), 2543-2559.

Xie, Y., & Li, X. (2008, July). Churn prediction with Linear Discriminant Boosting algorithm. In Machine Learning and Cybernetics, 2008 International Conference on (Vol. 1, pp. 228-233). IEEE.

Xu, Y. (2012, May). Predicting customer churn with extended one-class support vector machine. In Natural Computation (ICNC), 2012 Eighth International Conference on (pp. 97-100). IEEE.

Xu, E., Liangshan, S., Xuedong, G., & Baofeng, Z. (2006, December). An algorithm for predicting customer churn via BP neural network based on rough set. In Services Computing, 2006. APSCC'06. IEEE Asia-Pacific Conference on (pp. 47-50). IEEE.

Yan, L., Fassino, M., & Baldasare, P. (2005, July). Predicting customer behavior via calling links. In Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (Vol. 4, pp. 2555-2560). IEEE.

Yan, L., Wolniewicz, R. H., & Dodier, R. (2004). Predicting customer behavior in telecommunications. Intelligent Systems, IEEE, 19(2), 50-58.

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. In Advanced data mining and applications (pp. 300-306). Springer Berlin Heidelberg.

Zan, M. O., Shan, Z., Li, L. I., & Ai-Jun, L. I. U. (2007). A Predictive Model of Churn in Telecommunications Based on Data Mining. In 2007 IEEE International Conference on Control and Automation (pp. 809-813).

Zhang, G. (2007, September). Customer Retention Based on BP ANN and Survival Analysis. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on* (pp. 3406-3411). IEEE.

Zhang, Y., Qi, J., Shu, H., & Cao, J. (2007, October). A hybrid KNN-LR classifier and its application in customer churn prediction. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on* (pp. 3265-3269). IEEE.

Zhao, J., & Dang, X. H. (2008, October). Bank customer churn prediction based on support vector machine: taking a commercial bank's VIP customer churn as the example. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on* (pp. 1-4). IEEE.

Zhao, X., Shi, Y., Lee, J., Kim, H. K., & Lee, H. (2014). Customer Churn Prediction Based on Feature Clustering and Nonparallel Support Vector Machine. *International Journal of Information Technology & Decision Making*,*13*(05), 1013-1027.

Appendix A: Intro to Centraal Beheer

This appendix gives a briefly overview of the organization Centraal Beheer Achmea, the company which provided the information to conduct this research. This appendix also describes the competitiveness of the indemnity insurance market.

Case study: Centraal Beheer

Centraal Beheer is a Dutch brand which belongs to Achmea. Achmea is an insurance company operating in the Dutch insurance industry (Achmea, n.d.). They hold a leading position in the Dutch insurance market. From this leading position Achmea also positions itself internationally as an innovative player in selected markets, including Turkey, Greece, Slovakia, Ireland and in Australia. Achmea offers Health, Life and non-life insurances, Income Protection insurance products and pensions. Figure 1 (Achmea, n.d.) shows an overview of the international brands of Achmea.

Figure 1.1. Summary interational brands Achmea. Adapted from Achmea website, by Achmea, 2015, retreived from https://www.achmea.nl/SiteCollectionDocuments/ Factsheet-Achmea-in-een-oogopslag.pdf.

	NUMBER OF CUSTOMERS	NON-LIFE	HEALTH	LIFE
achmea	9.161.500	✓	✓	✓
EUREKO SIGORTA	1.224.912	✓	✓	
INTERAMERICAN PART OF ACHMEA	853.204	✓	✓	✓
Union	689.199	✓	✓	✓
Friends First	276.500			✓
achmea australia	333	✓		

In this study the focus is on a Dutch brand from Achmea, named Centraal Beheer, one of the Dutch brands which are presented in Figure 1.2 (Achmea, n.d.). The Dutch Achmea is the umbrella brand which represents the twelve single brands. In total, the Dutch brands represent more than 9 million customers. Centraal Beheer is listed in the top 3 brands with the most active customers. Zilveren Kruis (Health insurances) is the largest label with approximately 3,5 million customers, followed by Interpolis (Health-, life and non-life insurances) with approximately 1.7 million customers. Centraal Beheer (non-life insurance , income protection insurance, banking products and pensions) is placed third with approximately 1.4 million customers.

Figure 1.2. Overview dutch brands Achmea. Adapted from Achmea website, by Achmea, 2015, retreived from https://www.achmea.nl/en/brands/Paginas/default.aspx

The brand Centraal Beheer offers non-life insurances, income protection insurances, banking products and pensions to private consumers and businesses. The different products are accommodated in several business units. For this study the focus is on the indemnity insurances. Firstly, because this part of the industry experiences a highly competitive environment and the generic elements mentioned in chapter one are relevant in this industry. Secondly, because current literature barely focused on the indemnity industry and thirdly, because my current position allows me to access data which is needed to perform churn prediction analysis. Due to several compliance regulations in the Netherlands it is not allowed to mix and match between for example health- and indemnity insurances. Data cannot be exchanged between the different business units because it can lead to foreknowledge and misuse. The indemnity insurances from Centraal Beheer can roughly categorized in four main streams, namely: Car -, fire and theft -, third party –, travel and legal insurances. Every category counts several coverages with which every customer can adapt in their own (current) situation.

As mentioned before, Centraal Beheer operates in a highly competitive industry. De Vries, Schoenmaker, Streppel en Verhoeven (2015) mention various characteristics from the Dutch indemnity industry in their yearly Dutch insurance report which are important for the competition in the market. They mention it is common to close year contracts. This means that a customer can cancel her/his insurances after one year. Mostly they can cancel their insurances monthly after the first contract year. Furthermore the insurance products are relatively simple and very homogeneous, which makes it easier for customers to compare – nowadays mainly by using the comparison websites on the internet – insurances from different providers and churn.

Appendix B: Prediction survey

Authors	Domain datset(s) used	Prediction techniques	Published	Literature Review	In Tsai and Yu-Hsin Lu (2009)	In KhakAbi, Gholamian and Namvar, 2010
Adwan, Faris, Jaradat, Harfoushi and Ghatasheh, 2014	Telecom Industry	Neural Network	2014	V		
Ahn, Han and Lee, 2006	Korean Telecom Industry	Logistic Regression	2006			V
Au, Chan and Yao, 2003	Credit card and PBX databases	Decision tree, Neural Network	2003		V	
Buckinx and Van den Poel ,2005	Retail industry	Logistic regression, Neural Network, Random forests	2005		V	V
Burez and Van den Poel, 2007	Pay-TV company	Logistic regression, Markov chains, Random forests	2007		V	
Burez and Van den Poel, 2008	Pay-TV company	Random forests Survival analysis	2008		V	V
Burez and Van den Poel, 2009	Banking industry, Telecom Industry, Subscription services, PayTV, Retail industry	Support vector machine, Gradien boosting Machine	2009			V
Chiang, Wang, Lee and Lin, 2003	Network banking	Association Rule/Sequence Discovery	2003		V	V
Chu, Tsai and Ho, 2007	Taiwan telecom company	Decision tree, Growing Hierarchical Self-organizing map	2007		V	V
Clemente, Giner-Bosch, San Matias, 2010	Spanish Retail Industry	Neural Network, Logistic Regression, Classification trees, Random Forest, AdaBoost	2010	V		
Coussement and De Bock, 2013	Online gamgling industry	Decision trees, Logistic Regression	2013	V		
Coussement and Van den Poel, 2008	Belgian newspaper publishing company	Support vector machines, Random forests, Logistic regression	2008		V	V

Coussement and Van den Poel, 2008	Belgian newspaper publishing company (subscription services)	Logistic regression	2008		V	V
Coussement and Van den Poel, 2009	Subscription services	Logistic regression, Random Forests, Support Vector Machines	2009			V
Davi Prasad and Madhavi, 2012	Banking Industry	Decision tree	2012	V		
De Bock and Van den Poel, 2012	Six real-life churn prediction projects from large European Companies	Logistic Regression	2012	V		
Farquad, Vadlamani Ravi and Bapi Raju, 2014	Banking Industry	Support Vector Machine, Decision tree	2014	V		
Glady, Baesens and Croux, 2009	Belgian financial service	Neural Network (MLP), Decision tree, Logistic regression	2009		V	V
Günther, Tvete, Aas, Sandnes, and Borgan, 2014	Insurance company	Logistic regression with generalized additive models	2014	V		
Hu, 2005	Financial service industry	Neural Network, Decision tree, Bayesian Network	2005			V
Huang, Kechadi and Buckley, 2012	Telecom Industry	Logistic Regressions, Linear Classifications, Naive Bayes, Decision Trees, Multilayer Perceptron Neural Networks, Support Vector Machines, Evolutionary Data Mining Algorithm	2012	V		
Hung, Yen and Wang, 2006	Wireless telecom services	Decision tree, Neural Network, K-means	2006		V	V
Hur and Lim, 2005	Online auto insurance	Support Vector Machine, Neural Network	2005	V		
Ismail, Awang, Rahman and Makhtar, 2015	Telecom Industry	Neural Network, Logistic Regression, Multiple Regression	2015	V		
Kim and Yoon, 2004	Korea mobile carriers	Binomial logit model	2004		V	
Kim and Yoon, 2004	Credit card and PBX databases	Decision tree, Neural Network	2004		V	

Kim, Jun and Lee, 2014	Telecom Industry	Logistic Regression, Neural Networks	2014	V		
Larivie`re and Van den Poel, 2005	Belgian financial services	Random forests, Regression, Forests, Linear regression, Logistic regression	2005		V	V
Larivière and Van den Poel, 2004	European financial services	Hazard models, Survival analysis	2004		V	V
Lemmens and Croux, 2006	Telecom Industry	Decision trees (Bagging, boosting)	2006	V		
Lessmann and Voß, 2009	Australian/German credit, Data Mining Cup (DMC)	Decision tree, logistic regression, Support Vector Machine	2009			V
Li and Deng, 2012	Telecom Industry	Cluser Analysis, Decision Tree	2012	V		
Liou, 2009	Airline industry	Rough Set Theory	2009			V
Lu, Lin, Lu and Zhang, 2014	Telecom Industry	Logistic Regression	2014	V		
Luo, Shao and Liu, 2007	Personal Handy-phone System Service	Neural Network, Decision tree	2007		V	
Mozer, Wolniewicz, Grimes, Johnson and Kaushansky, 2000	Wireless Telecom Industry	Neural Network, Logistic regression	2000		V	
Musa, 2013	nvt	Support Vector Machine, Logistic Regression	2013	V		
Nie, Zhang and Shi, 2006	Charge Email	Decision tree	2006		V	
Obiedat, Alkasassbeh, Faris andHarfoushi, 2013	Telecom Industry	K-Means, Classification Tree	2013	V		
Pendharkar , 2009	Telecom Industry	Neural Network	2009			V
Pendharkar, 2009	Cellular wireless network services	Genetic Algorithm + NN, Zscore classification model	2009		V	
Prinzie and Van den Poel, 2006	International Financial Service Provider	Logistic Regression, Time Series, Tailor/Butina	2006			V
Qi, Zhang, Liu, Li, Zhou, Shen, Liang and Li, 2009	Telecom Industry	Decision tree, logistic regression	2009			V

Risselada, Verhoef and Bijmolt, 2010	Internet Service Provider, Insurance Company	Logit, Classification tree	2010	V		
Shao, Zheng, and An, 2008		Bayesian Network	2008			V
Sharma, Panigrahi and Kumar, 2013	Subscription/Telecom Industry	Neural Network	2013	V		
Smith, Willis and Brooks, 2000	Insurance company	Neural Network, Logistic regression, Decision tree	2000		V	
Soeini and Rodpysh, 2012	Iran insurance industry	K-means clustering, CART Decision Tree	2012	V		
Song, Kim, Cho, and Kim, 2004	Online game site	Decision tree, Self organizing maps, Association rules	2004			V
Song, Yang, Wu, Wang and Tang, 2006	Chinese Telecom Industry	Neural Network	2006			V
Tsai and Lu, 2009	American Telecom Companies	Hybrid Neural Network,	2009		V	
Vafeiadis, Diamantaras, Sarigiannidis and Chatzisavvas, 2015	Telecom Industry	Neural Networks, Decision Trees, Support Vector Machines, Naïve Bayes, Logistic Regression	2015	V		
Verbeke, Dejaeger, Martens, Hur and Baesens, 2012	Telecom Industry	Decision trees, Ensemble methods, Neural network, Nearest neighbors, Rule induction techniques, Statistical classifiers, Support Vector Machine	2012	V		
Verbraken, Verbeke and Baesens, 2014	Telecom Industry	Bayesian Network	2014	V		
Wang, Chiang, Hsu, Lin and Lin, 2009	Wireless network company	Decision tree	2009			V
Wang, Satake, Onishi and Masuichi, 2015	Telecom Industry	Support Vector Machine	2015	V		
Wei and Chiu, 2002	Taiwan mobile company	Decision tree	2002		V	

Xie and Lie, 2008	Banking Industry	Neural network, Decision Tree, Support Vecter Machine, Linear Discriminant Analysis, Adaboost	2008			V
Xie, Li, Ngai and Ying, 2009	Chinese banking	Neural Network, Decision tree, Support vector machine, Random Forests	2009		V	V
Xu, 2012		Support Vector Machine, Neural Network, Decision Tree	2012	V		
Xu, Liangshan, Xuedong and Baofeng, 2006	NB	Neural Networks	2006			V
Yan, Fassino and Baldasare, 2005	Telecom Industry	Neural Network, Decision Tree, ROCK (Robust Clustering using linKs)	2005			V
Yan, Wolniewicz and Dodier, 2004	Telecom Industry	Neural Networks	2004			V
Zan, Shan, Li and Ai-jun, 2009	Telecom Industry	Decision tree	2009			V
Zhang, 2007	Telecom Industry	Neural Network, Survival analysis	2007			V
Zhang, Qi, Shu and Cao ,2007	Census income, Australian credit approval, Telecom, Medical Wisconsin Breast Cancer	Neural Network, Decision tree, Logistic regression, K-nearest Neighbor	2007			V
Zhao, Li, Li, Liu and Ren, 2005	Wireless Industry	ANN, Decision Tree, Naïve Bays, Support Vector Machine	2005	V		
Zhao, Shi, Lee, Kim and Lee, 2014	Banking Industry	Support Vector Machine	2014	V		
Zhaoa and Dang, 2008	Banking Industry	Neural network, Decision Tree, Logistic regression, Bayesian Network	2008			V

Appendix C: Search engines

This appendix gives an overview of the search engines used for this study.

For this study the Digital Library provided by the Open University of the Netherlands, which is connected to the databases listed in figure C.1, and google scholar (https://scholar.google.com) are used.

Figure C.1. List of connected databases. Adapted from The Open University website, by Open University, 2015, retreived from http://bibliotheek.ou.nl/

Academic Search Elite (EBSCO)
ACM Digital Library
Business Source Premier (EBSCO)
Cambridge University press
DOAJ - Directory of Open Access Journals
EBSCO Host
E-Journals (EBSCO)
Emerald [management plus]
ERIC (EBSCO)
Google Scholar / Google Wetenschap
GreenFILE (EBSCO)
HeinOnline
IEEE Digital Library
JSTOR Business, Biological, Mathematics & Statistics Collection
Kluwer Navigator **!! kies: inloggen met studentenaccount !!**
Lecture Notes in Computer Science
Legal Intelligence
Library, Information Science & Technology Abstracts - LISTA (EBSCO)
NARCIS - the Gateway to Dutch Scientific Information
Nature : international weekly journal of science
OpMaat Premium
Overheid.nl
Oxford Journals
PsycArticles (EBSCO)
Psychology and Behavioral Sciences Collection (EBSCO)
PsycINFO (EBSCO)
PubMed
Regional Business News (EBSCO)
SAGE Journals Online
Science
ScienceDirect (Elsevier)
SpringerLink
Taylor & Francis Group
Web of Science
Wiley Online Library
WorldCat

Appendix D: Variables insurance industry

Figure D.1. Descpription of available explanatory variables. Adapted from "Modeling and predictiong customer churn from an insurance company" by C.C. Günther, I.F. Tvete, K. Aas, G.I. Sandnes and O. Borgan 2014. *Scandinavian Acturial Journal*, Vol. 2014, No. 1, 58-71.

Explanatory variable	Description
`Premium`	Yearly total premium in NOK (range [0–50,000]).
`Age`	Age of customer (range 18–75).
`Gender`	Gender of customer (0 = Female, 1 = Male).
`Partner`	Customer's spouse or partner has also a policy in the company (0 = No, 1 = Yes).
`Discount`	Discount programme, ({1,2,3,4,5}, 5 denotes no discount programme).
`Car`	Customer has car insurance (0 = No, 1 = Yes).
`Home`	Customer has home insurance (0 = No, 1 = Yes).
`HomePolicies`	Number of home insurance policies (range 0–28).
`Health`	Customer has health insurance (0 = No, 1 = Yes).
`Lifetime`	Registered duration (in years) of continuous customer relationship (range [0–12.72]). If a customer exits and later returns, the value is set to 0 at the point of return.

Figure D.2. Variables used in the customer retention analysis. Adapted from "An analysis of customer retention and insurance claim patterns using data mining: A case study." by K.A. Smith, R.J. Willis and M. Brooks 2000. C.C. Günther, I.F. Tvete, K. Aas, G.I. Sandnes and O. Borgan 2014. *Journal of the Operational Research Society, 532-541.*

Table 1 Variables used in the customer retention analysis

Variable	*Data type*	*Status*	*Transformation*
Post code	Categorical	Used	Grouped into 10 bins
New business	Binary	Used	
Vehiele age	Continuous	Used	Grouped into 4 bins
Rating	Categorical	Used	Grouped into 2 bins
Years on rating	Continuous	Used	
Previous company	Categorical	Rejected	
Car category	Categorical	Rejected	
Policy holder age	Continuous	Used	Grouped into 5 bins
Gender	Binary	Used	
Premium	Continuous	Used	Log transformation
Premium diff	Continuous	Used	
Sum insured	Continuous	Used	Log transformation
Sum insured diff	Continuous	Used	
Claim history	Binary	Rejected	
Years on policy	Continuous	Used	
Terminated	Binary	Used	

Figure D.3. Summary statistics. Adapted from "Customer churning prediction using support vector machines in online auto insurance service." by Y. Hur and S. Lim 2005. *Advances in Neural Networks-ISNN* 2005, p. 928-933.

Table 1. Summary statisics

Feature name	Mean	Std. Deviation	Minimum	Maximum
Age of Driver	39.14621	9.11668	19	95
Zip Code	3.637727	1.891552	1	7
Car Type	2.112273	1.404095	0	5
Price of Car	607.2974	550.3214	1	9600
Credit/Debit	66.69394	24.31643	40	200
Surcharge	2.347197	6.212417	0	50
Driver Endorsement	2.920076	1.080609	1	5
Age Endorsement	1.981894	0.13334	1	2
Number of Airbag	0.434394	0.656068	0	2
Property Liability	3258.258	2628.364	0	100000
Medic. Expense	5396.856	5945.052	0	20000
Deductible(auto)	4.092197	8.421139	-1	50
Last Premium(t)	145764.1	245559.1	0	2005720
Old Insurer's Quote(t+1)	442528.5	245609.1	0	2504970
Type of Coverage	1.216439	0.481858	1	3

Figure D.4 Parameter estimates of the single logit models (insurance data). Adapted from "Staying power of churn prediction models." by H. Risselada and P.C. Verhoef and T.H. Bijmolt 2010. *Journal of interactive Marketing*, 24(3), 198-208.

Table 5
Parameter estimates of the single logit models (insurance data).

Variable	Period		
	2004	2005	2006
Age (years)	−.0104 **	−.0004	−.0144 **
Relationship length (years)	−.3625 **	−.2248 **	−.1116
Package type (ref. cat. '0')			
1	.0955	.8462	−.5464
2	−.0216	.4145	.5371
3	−.7110 **	−.2046	.2680
4	−.9320 **	−.1590	.2495
5	−.9182 **	−.2868	.6868 *
6	−.8702 **	−.2811	.5635 *
7	−.9738 **	−.0779	.7244 *
8	-1.1030 **	.1386	.4352
Family configuration (ref.cat. 'single')			
No kids	.1086	−.1515	.8324 **
Kids	.1337	−.0048	−.3571
Family1	−.1053	−.3456	.3325
Family2	.2331	.0938	.2395
Unknown	.6019 **	.3903 **	.9138 **
Income (ref. cat. 'unknown')			
>2 times standard	.3756	.0737	−.0448
Standard-2 times standard	−.0415	.0770	.1938
Standard income	−.0917	.0553	−.2693
Minimum–standard income	−.1248	−.0318	−.4017 *
Minimum	.0497	−.0822	−.7447 *
Variable	−.3054	−.1121	.0525
Collectively insured	−.1912	−.1767	−.4183 *
Moved	-3.7922 **	−.0730	−.5359 **

* $p<.05$.
** $p<.01$.

Appendix E: Product overview Centraal Beheer

Main coverage types (A-product)	Remaining coverage types (B-product)
Annual travel insurance	Accident passenger insurance
Car insurance	Boat insurance
Home: building insurance	Caravan insurance
Home: contents insurance	Garden insurance
Home: third party insurance	Disciple/trailer insurance
Legal assistance (basic) insurance	Home: accident insurance
	Home: all risk insurance
	Home: expensiveness out-of-doors insurance
	Home: garden insurance
	Home: glass insurance
	Home: student insurance
	Home: technical home assistance insurance
	Invalids insurance
	Legal assistance capital and fiscal insurance
	Legal assistance consumer insurance
	Legal assistance divorce insurance
	Legal assistance labor and income insurance
	Legal assistance traffic and medical insurance
	Mobile home insurance
	Moped insurance
	Motorcycle insurance
	No-claim protector insurance
	Roadside assistance insurance (EUR)
	Roadside assistance insurance (NL)
	Temporary travel insurance

Appendix F: SAS Enterprise Miner diagram