Eindhoven University of Technology

Master Thesis

Customer churn prediction for an insurance company

Author:
Chantine Huigevoort

Supervisors:
Eindhoven University of Technology
dr. ir. Remco Dijkman
dr. Rui Jorge de Almeida e Santos Nogueira
CZ
Wouter Wester MSc

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

Information Systems
IE&IS

April 2015

"Believe you can and you are halfway there."

Theodore Roosevelt

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

Subject headings: data mining, customer relationship management, churn prediction, customer profiling, health insurance, AUK, AUC

Abstract

Dutch health insurance company CZ operates in a highly competitive and dynamic environment, dealing with over three million customers and a large, multi-aspect data structure. Because customer acquisition is considerably more expensive than customer retention, timely prediction of churning customers is highly beneficial. In this work, prediction of customer churn from objective variables at CZ is systematically investigated using data mining techniques. To identify important churning variables and characteristics, experts within the company were interviewed, while the literature was screened and analysed. Additionally, four promising data mining techniques for prediction modeling were identified, i.e. logistic regression, decision tree, neural networks and support vector machine. Data sets from 2013 were cleaned, corrected for imbalanced data and subjected to prediction models using data mining software KNIME. It was found that age, the number of times a customer is insured at CZ and the total health consumption are the most important characteristics for identifying churners. After performance evaluation, logistic regression with a 50:50 (non-churn:churn) training set and neural networks with a 70:30 (non-churn:churn) distribution performed best. In the ideal case, 50% of the churners can be reached when only 20% of the population is contacted, while cost-benefit analysis indicated a balance between the costs of contacting these customers and the benefits of the resulting customer retention. The models were robust and could be applied on data sets from other years with similar results. Finally, homogeneous profiles were created using K-means clustering to reduce noise and increase the prediction power of the models. Promising results were obtained using four profiles, but a more thorough investigation on model performance still needs to be conducted. Using this data mining approach, we show that the predicted results can have direct implications for the marketing department of CZ, while the models are expected to be readily applicable in other environments.

Management summary

This master thesis is the result of the Master program Operation Management and Logistics at Eindhoven University of Technology. This research project focuses on the design and application of a prediction model for customer churn which, providing insight in churn behavior in a case study for CZ (Centraal Ziekenfonds), a major Dutch health insurance company. The main research question of this research is defined as:

What are the possibilities to create highly accurate prediction models, which calculate if a customer is going to churn and provide insight in the reason why customers churn?

Previous literature acknowledges the potential benefits of customer churn prediction. The marketing costs of attracting new customers is three to five times higher than when retaining customers, which makes customer churn an interesting topic to investigate for businesses.

With literature analysis and expert interviews the characteristics for customer churn were identified. The most important churning characteristics found in this research are age, the number of times a customer is insured at CZ and health consumption. With the K-means algorithm four different customer profiles were identified with respect to churning behavior. The profiles are given below in the numeration. The first profile represents the averages of the population, the second and third profile represent non-churning customers and the last profile indicates a churning profile.

- Profiles which are comparable to the average of the population.
- Older customers, who have no voluntary deductible excess and consume more health insurance than average.
- Young customers which do not pay the premium themselves and have a group insurance.
- Young customers, who consume less health insurance than average and pay the premium themselves.

To discover which churn prediction techniques are widely used in the literature, a literature study was performed. The four most used techniques in the literature are logistic regression, decision tree, neural networks and support vector machines. When implemented on pre-processed and cleaned datasets, the logistic regression and neural networks techniques showed the best performance. The training sets were corrected for imbalanced data, by artificially including more churners without resorting to oversampling or undersampling. The logistic regression technique showed the best results with a balanced data set between churners and non-churners. Neural networks performed best on a 70:30 (non-churn:churn) distribution.

The lift charts of logistic regression and neural networks displayed the best performance. Approximately 50% of the churners can be reached by contacting 20% of the population. When applied to data from different years, the models showed similar behavior and results, indicating the generality of the constructed prediction models. When the churning possibilities (predicted with logistic regression or neural networks) are ordered from high to low, and 20% of the customers with the highest churning possibility are contacted, it is expected from a cost-benefit analysis that no net costs are made. The neural network technique generates a benefit of €4,319, with only 5,000 cases in the sample set. To see if even better results could be generated, homogeneous profiles based on K-means clustering were used to create the churn prediction models. It was difficult to conclude which model performed best based on the used performance parameters. A possible reason for this can be that the K-means cluster sizes, were to small.

The main conclusion of this research is that it is possible to generate prediction models for customer churn at CZ with good prediction characteristics. By combining a research-based focus with a business problem solving approach, this research shows that the prediction models can be used within the CZ marketing strategy as well as in a general academic setting.

Recommendation for the company

The results were investigated with lift chart, cost-benefit analysis and the models were tested on data of 2014. The models from logistic regression and neural networks performed almost evenly well, but only the logistic regression model provides insights in the variables which are important to predict customer churn. For this reason it can be concluded that the logistic regression technique works best for the marketing department of CZ. It is recommended to investigate how the results can be implemented. Different possibilities are available, for example, the effect of contacting customers with a predicted high possibility of churning can be investigated. Additionally, a change in the assistance approach when customers contact CZ can be implemented when a customer with a high churn probability is identified.

Limitations identified during this research

- Data extraction is not checked by other SAS Enterprise Guide experts.
- Each technique is tested with a different sub-set of the original data set sample.
- For the cost-benefit analysis no real costs and benefits were applied.

Future research should concentrate on

- Investigation in variables which can be used for the representation of customer satisfaction.
- Model generation with most influencing variables identified in this research.
- Further elaboration on the performance parameters for imbalanced data sets.

Acknowledgements

This thesis is the result of 7 months of hard work on my master thesis project in order to fulfill my master degree in Operations Management and Logistics at Eindhoven University of Technology. This thesis project was carried out from October 2014 to April 2015 at CZ. I realize that this thesis was only possible with the help and guidance of others. I would like to take this opportunity to thank some people who surrounded me and who motivated me during my master and during my master thesis project.

First of all, I would like to thank my supervisors from the university. My first supervisor Remco Dijkman provided me with useful feedback and asked questions which resulted in interesting insights and brought my thesis to a higher level. I would also like to thank my second supervisor Rui Jorge de Almeida e Santos Nogueira. He always managed to set me at rest when I panicked and thought that I could not solve the problems I was facing.

Secondly, I would like to thank my supervisors from CZ, espacially Wouter Wester for his commitment to the project and feedback. As a result I had contact with a wide range of people and a good feeling about the research problem. I would also like to thank Liesan Couwenberg, who has coached me during my master thesis project. She made sure that I was able to collect the data in time and really supported me with my project management.

Finally, I would like to thank my family and friends. They never lost their patient and supported me throughout my whole master. A special thanks goes to my boyfriend, Bas Rosier, he was always there and supported me with asking the right questions.

I want to conclude with the fact that I really enjoyed my time at the University. It has been an unforgettable period in my life.

Chantine Huigevoort
April 2015

Contents

Chapter 1

Research introduction

This research project focuses on the design and application of a prediction model for customer churn which, providing insight in churn behavior in a case study for CZ (Centraal Ziekenfonds), a major Dutch health insurance company. As a formal introduction, Chapter 1 discusses the research area, research goals and research design. The research starts with an identification of the research area and the central problem definition (Section 1.1). With the problem definition the research questions and project goals are formulated, which are discussed in Section 1.2. How the research project will be executed is discussed in Section 1.3.

1.1 Research area and churn context

To describe the research area first the research field, problem outline and relevance are discussed. The research area and problem outline will be discussed in the context of a health insurance company with a case study.

Research field

Customer Relation Management (CRM) is concerned with the relation between customer and organization. In the twentieth century academics and executives became interested in CRM [54]. CRM is a very broad discipline, it reaches from basic contact information to marketing strategies. Four important elements of CRM are: *customer identification, customer attraction, customer development* and *customer retention* [51]. An example of *customer identification* is customer segmentation, e.g. based on gender. *Customer attraction* deals with marketing related subjects such as direct marketing. An important element of *customer development* is the up-selling sales technique. Finally, *customer retention* is the central concern of CRM, and is linked to loyalty programs and complaints

management. Customer satisfaction, which refers to the difference in expectations of the customer and the perception of being satisfied, is the key element for retaining customers [51]. Customer retention is about exceeding customers expectations so that they become loyal to the brand.

When customer expectations are not met, the opposite effect can occur, i.e. customer churn. Customer churn is the loss of an existing customer to a competitor [9]. In this research a competitor is a different brand, which can result in a churning customer although the customer stays at the same company [34]. To manage customer churn first the churning customers should be recognized and then these customers should be induced to stay [2].

The marketing costs of attracting new customers is three to five times higher than when retaining customers [49], which makes customer retention an interesting topic for all businesses. For example, health insurance companies in the Netherlands are particularly concerned with customer satisfaction and retention, because the required basic health insurance package is generally the same for each company. This creates a highly dynamic and competitive environment, in which customers are able to quickly switch between health insurance companies. Major companies often serve millions of customers, making it difficult to extract useful data on customer switching behavior and to predict changes in customer retention.

A useful approach to deal with large amounts of information is data mining. Data mining is a technique to discover patterns in large data sets. There are multiple modelling techniques that can be used in data mining, such as clustering, forecasting and regression. Data mining deals with putting these large data sets in an understandable structure. Data mining is part of a bigger framework, named Knowledge Discovery in Databases (KDD) [2, 67]. An overview with the process of KDD is shown in Figure 1.1.

Before data mining is applied data selection and pre-processing activities are necessary. Pre-processing activities are needed to create a high quality data set. If the data set does not have a high quality level, the results of the data mining techniques are also not of high quality. Data sets are often incomplete, inconsistent and noisy, which creates the need of data pre-processing [2]. Data pre-processing tasks are e.g. data cleaning, data integration, data transformation, data reduction and data discretisation [2]. Good data pre-processing activities are key to produce a valid and reliable model. When the data set is of sufficient quality, the data mining activities can be applied, as shown in Figure 1.1.

Which data mining technique is used to create the prediction model depends on the goal for which the prediction model is used and the data in the data set. The model in this

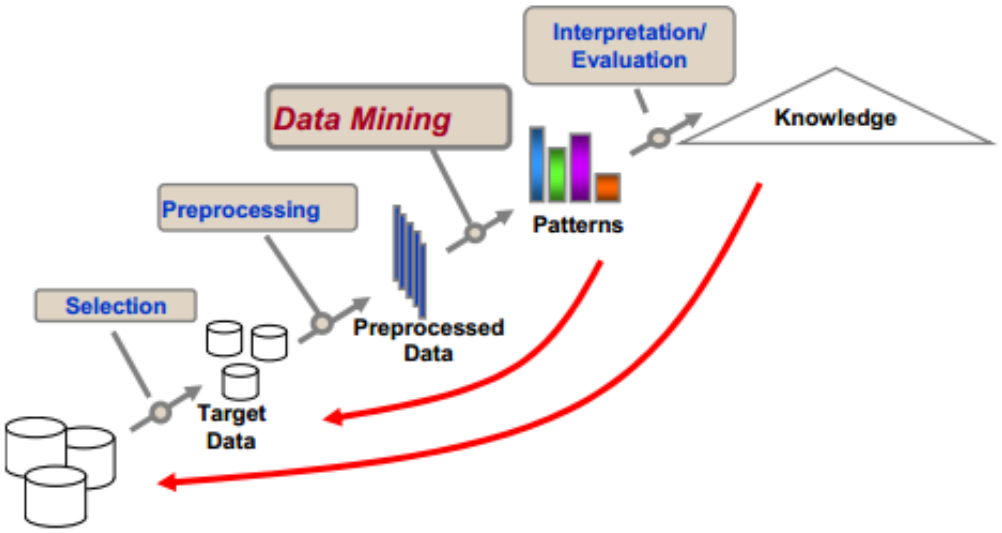


FIGURE 1.1: An overview of the knowledge discovery process in databases [2].

research project should be able to predict customer churn. The prediction models can be calculated with multiple modeling techniques e.g. decision trees and neural networks. When the prediction models are generated the results can be analysed to discover new insights and knowledge.

Case study: Centraal Ziekenfonds

CZ is a health insurance company and the core activity is the supply of the mandatory insurance for health costs. Its mission is to offer good, affordable and accessible health care. CZ was founded in 1930 in Tilburg, and provides health insurance policies for three major health insurance brands, CZ, OHRA and Delta Lloyd. This graduation project is performed at CZ so the other two brands are not taken into consideration.

The product portfolio of CZ consist of general insurance policies and additional insurance policies. The product portfolio contains three general insurance policies and six additional packages for extra reimbursements. The differences in the general insurance policies are the percentage of reimbursement for non-contracted care providers and the number of deductible levels. The additional packages are split up in three phases of life and basic, plus and top policies.

The long term strategy of CZ is to realize the best health care possible and to provide stable low premium health insurance policies. Currently, CZ employs roughly 2500 people in various departments [16].

The market in which CZ operates

A major health insurance reform took place in the Netherlands on January 1, 2006. Before the reform there were private and public insurance policies. The public health

care was organized by the government which decided what was covered in the insurance. Table 1.1 shows the differences between health insurance before and after 2006.

Before 2006		After 2006
Private insurance policy	**Public insurance policy**	**Basic insurance policy**
Earnings >€ 33,000	Earnings <€ 33,000	-
Market based premium	Premium set by government	Market based premium
Voluntary	Compulsory	Compulsory for everyone
Market based included care	Government based included care	Government based included care
Additional insurance policies		**Additional insurance policies**
Market based premium		Market based premium
Market based included care		Market based included care

TABLE 1.1: Differences in health insurance before and after 2006.

A major difference is that it is mandatory for everyone after 2006 to have a basic insurance. Before 2006 people earning more than € 33,000 were not obligated to have a health insurance. Nowadays everyone is obligated to have a basic health insurance and the premium is market based. The coverage of the basic health insurance is determined by the government. There are no major changes for the additional insurance policies.

Today there are four major health insurance companies which have a combined market share of almost 90% in 2014 [53], which has been stable for years. Achmea has a market share of 32% and is the largest insurance company, VGZ has a market share of 25%, while CZ and Menzis have 20% and 13% respectively. Health insurance policies can roughly be divided into individual and group insurances [15]. The number of group insurances increases slightly over the years 2010-2014 (with 2% over the years 2010-2013 and for the year 2014 with 1% [53]). In 2014 over 70% of all customers insured in the Netherlands have a group insurance. A reason for this is that with a group insurance the customers receives a discount of approximately 5% [53].

Problem outline and relevance

As discussed in Section 1.1 the government determines what will be covered in the basic insurance policies. In such a strictly regulated market, a unique competitive environment is evident. The government does not interfere with the additional insurance policies and this combination creates a dynamic and competitive environment.

There is a decrease in customer churn from 8.3% in 2013 to 6.9% in 2014, but this still encompasses 1.2 million customers. The outflow of 2013 contains switches in group insurances which is reflected in the high churn percentage in that year [53]. According

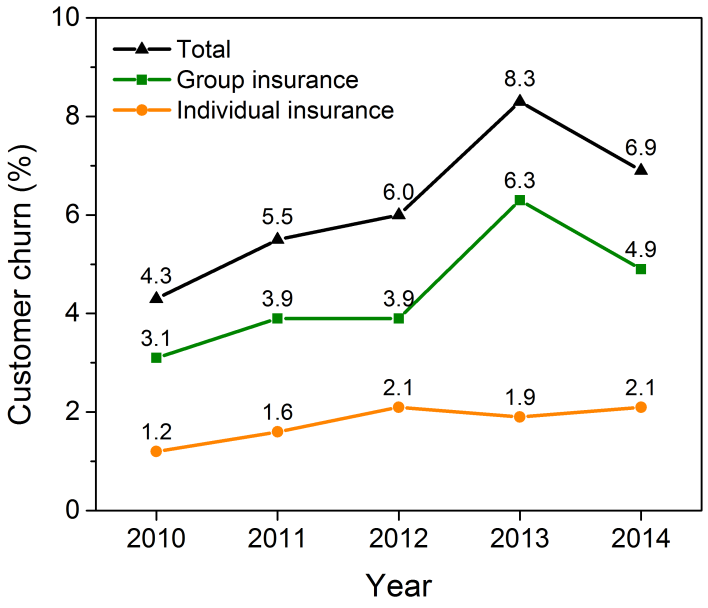


FIGURE 1.2: Percentage of customers which change to another health insurance company per year. Adapted from NZa [53].

to a survey by the National Health Authority (Nationale Zorgautoriteit, NZa) the price level of the health insurance is the number one reason of customer churn [53]. However, the exact reasons for customer churn are unclear, and they did not reach a significant conclusion. Figure 1.2 indicates customer churn percentages in 2010-2014.

The research to find the reason to stay at a health insurance company received enough responses to create an overview. The following ten reasons cover 75% of the given reasons to stay [53]:

- Satisfied with the coverage of the total health insurance.
- I am member of this health insurance company for a long time.
- Satisfied with the service of my health insurance company.
- Satisfied with the coverage of the basic health insurance.
- Satisfied with the discount of my group health insurance.
- Satisfied with the quality of organized healthcare.
- Satisfied with the coverage of the supplementary health insurance.
- I know what I can expect from my health insurance company.
- Satisfied with the hight of the total premium.
- The effort was too large to search for a new health insurance company.

To get an overall indication of churning customers and non-churning customers, the NZa measured a number of characteristics, shown in Table 1.2. A churning customer in this measurement is a customer which has switched for three or more times between health

insurance companies. As can be seen in Table 1.2, churners have less insurance costs than non-churners and the average age is lower for churners. These characteristics makes churners an attractive group to focus on.

Characteristics	Non-churners	Churners
Percentage female	51%	52%
Average age	47 years	33 years
Costs per customers in 2011	€2,206	€1,345

TABLE 1.2: Characteristics of churning customers versus non churning customers. Adapted from NZa [53].

We can conclude that there is a dynamic and competitive environment in which CZ operates. While there are some indicators for non-churning behavior, the precise reasons behind churning behavior remain unclear. Insights into churning behavior can be of vital importance to CZ to gain key advantages over the competition. We define the main problem as follows:

Problem statement
The recent increase in the dynamic and competitive environment of health insurance companies results in switching behavior of customers. It is unclear what the indicators are of switching behavior and which customers switch to a competitor.

1.2 Research goal and questions

With this problem statement the goal of the research and the research questions can be formulated. With answering the research questions the goals are automatically reached.

Research goal

The problem statement can be translated in a research goal. When the research questions are answered, the research goal also should be achieved. The research goal is as follows:

The research goal is to predict which customers are going to switch and understand why these customers switch. The prediction model should be relevant and applicable for the marketing department

Research questions

The research questions which are derived from the goal are represented in a main research question and four sub-research questions. The results of this research project will not

only be practically useful for CZ, but will also contribute to the applications of data mining techniques in academic literature.

Main research question
What are the possibilities to create highly accurate prediction models, which calculate if a customer is going to churn and provide insight in the reason why customers churn?

Sub-research question 1
Which customer characteristics and behavior aspects are key to predict customer churn behavior?

Sub-research question 2
Which techniques can be used to generate the best churn prediction models?

Sub-research question 3
Which customer profiles should be analysed separately and what is the difference between the profiles?

Sub-research question 4
Which model generates the best results, comparing on accuracy and interpretability?

1.3 Project strategy and research design

This research is based on the combined strategy of Van Aken *et al.* which combines a business problem solving approach with a research-based focus [66]. This research started with an identification of the research area and the research goal and questions and was discussed in Chapter 1. Figure 1.3 shows the actions and results of the remaining chapters.

In Chapter 2 the variables that are needed to create a good prediction model are selected. These variables are identified by means of a thorough literature study and interviews conducted with key experts within the company. The combined results will give an indication of which variables are key to describe a churning profile. Furthermore, the created data set is prepared for model generation with the identification of normality,

missing values, extreme values and variable transformation, while imbalanced data set problems are tackled. From the relevant data set of CZ the data is collected, which is stored in SAS Enterprise guide. To create a complete data set the zip codes of deprived areas are collected (CBS). The purity level and urbanity level of a neighborhood is also collected from the CBS, which is combined in this research to a level of urbanity per zip code.

Using the selected variables, a data analysis is performed in Chapter 3, while customer profiles are identified. With the identification of these profiles sub-research question 1 is answered. The data set is statistically compared with the population of the Netherlands and statistical tests to test for significant differences between churning and non-churning customers. Chapter 3 answers the question whether churning customers significantly differ from non-churning customers. After the model generation the findings will be verified by investigating which variables influence the model the most (Chapter 6).

In Chapter 4 data mining techniques from the literature are reviewed. The literature is selected based on the research strategy of Jourdan *et al.* [36]. First the selection strategy is explained, then the selected literature is categorized for a clear presentation of the results. Chapter 4 will provide an answer on sub-research question 2, i.e. which technique generates the best churn prediction model.

Based on these findings, Chapter 5 applies the identified techniques to the pre-processed data set, resulting in a prediction model. Two profiling methods and four prediction techniques are applied and their performance analysed with four performance parameters. The performance parameters that are used are Area Under the Cohen's Kappa curve (AUK), Area Under the ROC-Curve (AUC), precision and sensitivity. How the AUK and AUC relate to each other with imbalanced data is investigated. With the results of the profiling techniques sub-research question 3 is answered.

The best performing models of Chapter 5 are used in Chapter 6 to interpret the found results in four different ways. First, lift charts are analysed to see how many churners can be reached with which part of the population. Second, the robustness of the created model is checked, using a test set comprised of data from 2014. Third, to see if the models generate benefits for CZ a cost-benefit analysis is applied. Chapter 6 will conclude with a Section on the use of homogeneous profiles in the prediction models. With a combined interpretation of these results sub-research question 4 can be answered.

The research will conclude with Chapter 7, in which the sub-research questions and the main research question will be answered. Besides this the results are generalised and the recommendation for CZ, limitations and further research are discussed.

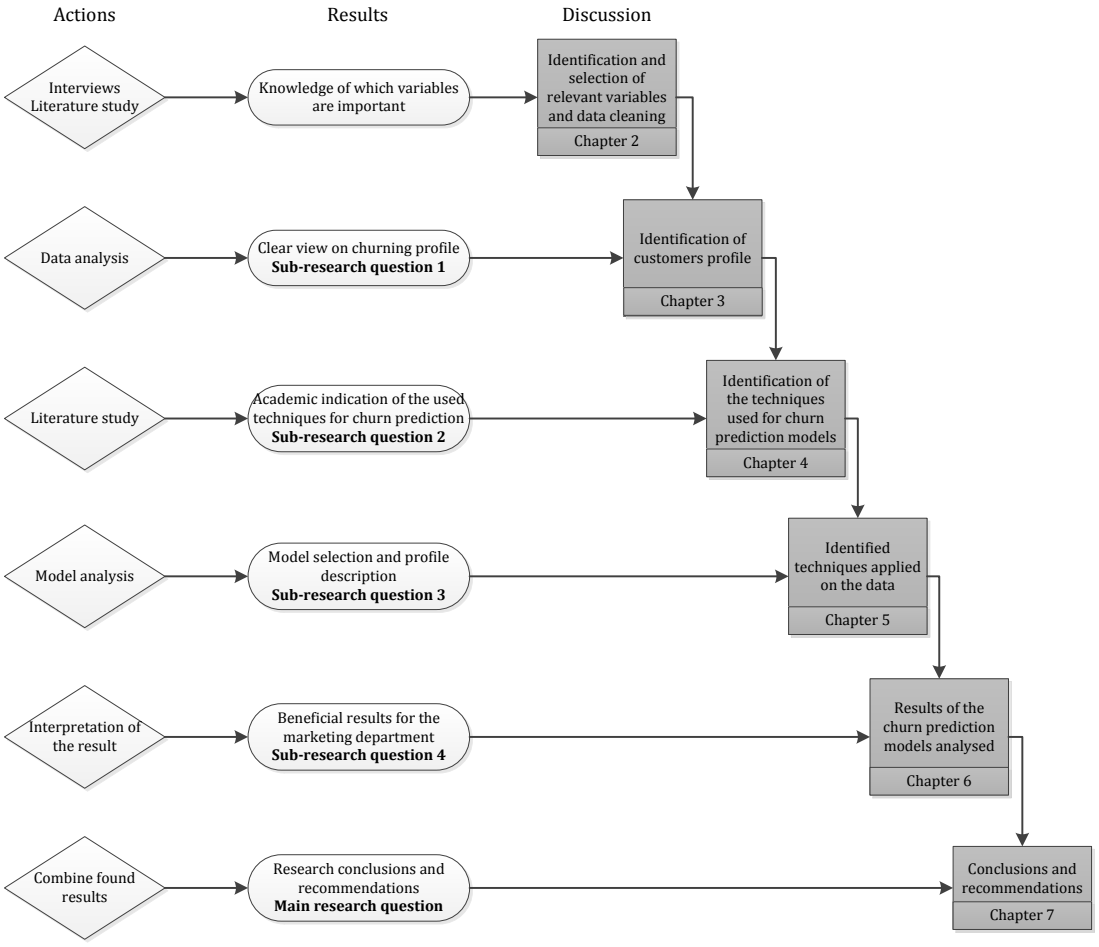


FIGURE 1.3: Schematic overview of the actions and expected results per chapter.

Chapter 2

Identification and selection of relevant variables

The data selection procedure starts with identifying the variables which can influence churn of a customer. The selection starts with the identification of variables used in the literature. The literature selection and results are discussed in Section 2.1. The findings in the literature are a guidance for the interviews with experts. Which experts are interviewed is discussed in Section 2.2. The findings in Section 2.1 and 2.2 are combined to select the variables which are discussed in Section2.3. In Section 2.4 the extraction of the data from SAS is discussed. After extraction of the data from SAS Enterprise Guide, the data set is prepared for creation of a prediction model. Which preparations are done is mentioned in Section 2.5. This chapter will conclude with the problems which an imbalanced data set gives (Section 2.6).

2.1 Variable selected from the literature

For an academic substantiation the literature is considered. To generate a broad perspective, a short and simple search term is chosen:

```
Churn prediction variables
```

Google Scholar is used as search engine because it searches in a wide range of journals. The selection stopped when two new selected articles did not suggest new variables. The selection of an article was based on title. If the term *churn prediction* is included in the title, the article was scanned for new variables and when the research used new variables it was added to the variable list.

Socio-demographic variables	**Resources**
Identification number	[30, 64]
Age	[12–14, 19, 22, 30, 32, 38, 40, 58, 67]
Gender	[13, 14, 19, 22, 30, 32, 40, 48, 67, 74]
Location identifier (ZIP code)	[38, 48, 67, 73]
Network attributes	[19, 22, 38, 58]
Segment selected by the company	[13, 14, 22, 48, 64]
Educational level	[19, 74]
Income	[19, 58, 74]
Customer satisfaction	[38, 40]
Customer/company-interaction variables	**Resources**
Number of contact moments	[12–14, 30, 38, 40, 48, 67]
Elapsed time since last contact moment	[12–14]
Number of complaints	[13, 14, 38]
Elapsed time since the last complaint	[12–14]
Reaction on marketing actions	[12–14]
Number of declarations	[74]
Outstanding charges	[30]
Duration of current insurance contract	[12–14, 38, 40, 58]
Number of times subscribed	[12–14]
Product-related variables	**Resources**
Premium price	[22, 32, 38, 48, 53, 58, 64, 67, 73]
Discount	[22, 64]
Payment method	[30, 32, 48, 64, 67]
Type of insurance	[30, 58, 64, 67, 74]
Product usage	[32, 38, 40, 53, 74]
Brand credibility	[38]
Switching barrier	[38]

TABLE 2.1: Variables selected from the literature.

Table 2.1 shows the found variables with reference to the source papers. The variables are split in socio-demographic, customer/company-interaction and product-related variables. The socio-demographic variables describe the customer, the customer/company-interaction variables describe the relationship between the customer and the company, and the product-related variables include information of the health insurance of that customer.

The papers of Günther *et al.* and Risselada *et al.* are the only two papers which focus on the insurance market [22, 58]. There are six papers which focus on the telecommunication industry [30, 32, 38, 40, 67, 73], and three selected papers discuss the banking industry [19, 48, 74]. The fourth topic that is discussed is about the newspaper market, three papers predict churners in this subject [12–14]. The last topic is multimedia on demand, which is discussed by Tsai and Chen [64]. Variables used in these articles with no direct

application or use in the present research are not mentioned in Table 2.1. An example of a specific variable that is not useful for this research is call duration, which is important in the telecommunication industry.

2.2 Variable selection indicated by experts of CZ

For the selection of specific variables related to the health insurance market experts within the company are interviewed. With these interviews a better understanding of the market and customer interactions will be generated. To find all relevant variables, eight different divisions within the company are contacted, of these eight divisions eleven experts are interviewed. Table 2.2 shows all divisions and expert functions. These divisions are selected because together, they cover almost the entire company. All divisions which have customer contact are selected. The divisions marketing intelligence and business intelligence do not have direct contact with the customer. These divisions are selected because marketing intelligence performs multiple market researches and the business intelligence division has contact with all divisions in the company which result in basic knowledge of all divisions.

Division	**Experts**
1. Customer administration	Team leader customer and service
2. Declaration services	Manager declaration services & Manager medical reviews
3. Quality management	Manager quality management
4. Business intelligence	Member of the business intelligence team
5. Marketing & Sales	Manager market intelligence
6. Healthcare advice	Manager health advice
7. Customer service	Manager credit control & Data analyst customer service
8. Contact centre	Manager brand contact centre

TABLE 2.2: Interviewed experts.

The main focus of the interviews is on the interaction between customer and the company. The interviews always started with a short introduction into the research. Then the expert was asked what important contact moments are with respect to their expertise. All experts also indicated variables which are important from their personal perspective. Table 2.3 shows all variables mentioned by the experts, labeled with the division of the expert that mentioned the variable. The divisions are identified by their numbers from Table 2.2.

Socio-demographic variables	**Division**
Identification number	All experts
Age	All experts
Gender	All experts
Location identifier	7
Network attribute	4
Segment selected by the company	4
Education	7
Income	7
Customer satisfaction	2, 3
Life events	7
Customer/company-interaction variables	**Division**
Number of contact moments	2, 3, 5-7
Type of contact (email, call, etc.)	3, 7
Experience during contact	2, 3, 7, 8
Customers mention that they are going to switch	2, 3, 7
Number of complaints	2-7
Number of declarations	1, 2, 5, 7, 8
Outstanding charges	4, 7
Number of authorizations	2-8
Handling time of authorizations and declarations	2-4, 8
Duration of current insurance contract	2, 3, 8
Number of times subscribed	3
Automatic administrative changes not reported	1
Product related variables	**Division**
Type of insurance	All experts
Premium price	All experts
Discount	8
Deductible excess	1, 6-8
Payment method	1, 6, 8
Product usage	7
Contracted care	2-4, 8
Brand credibility	3

TABLE 2.3: Variables indicated by the experts.

2.3 Variables selected based on literature and expert knowledge

As shown in Tables 2.1 and 2.3 the suggested variables from literature and expert interviews are partially overlapping. A group of variables that were found in literature are not selected because they are not (completely) stored in the database of the company. Examples are customer satisfaction, life events, education, income, switching barrier and brand credibility. Table 2.4 shows all variables that will be taken into consideration in

this research. Tables A.1 and A.2 of Appendix A show all variables, and the rejected variables with rejection reasons.

Socio-demographic variables
Identification number
Network attribute
Gender
Age
Location identifiers
Segment selected by the company

Customer/company-interaction variables
Number of contact moments
Number of complaints
Number of authorizations
Number of declarations
Number of payment regulations
Duration of current insurance contract
Number of times subscribed

Product related variables
Type of insurance
Premium price
Discount
Pays premium
Voluntary deductible excess
Product usage
Usage deductible excess
Contribution

TABLE 2.4: Variables selected for the prediction of churn.

The variables that are selected from the literature and the expert interviews can be divided into two groups, namely time-variant variables and time-invariant variables. Time-invariant variables are variables which do not change during the year and are measured on 1 January and the second group of variables represents actions which take place during the year and are measured on 31 December. The dependent variable, if a customer churns or not, is measured on 1 January of the next year. For example, if the variables are selected for the year 2013 the dependent variable is checked on 1 January 2014. Figure 2.1 shows in a systematic way the variables ordered into time varying and non-time varying variables. The variables "type of insurance" and "location identifiers" both consist of three variables shown in Figure 2.1. The variables labeled with a green colour indicated that the variable is mentioned in the literature and by experts. The orange labeled variables are only mentioned by experts, these variables can be applied by every health insurance company that wants to investigated this churning problem.

17 variables are non-specific for the health insurance market, which results in a model generation which can be applied by a wide range of research areas.

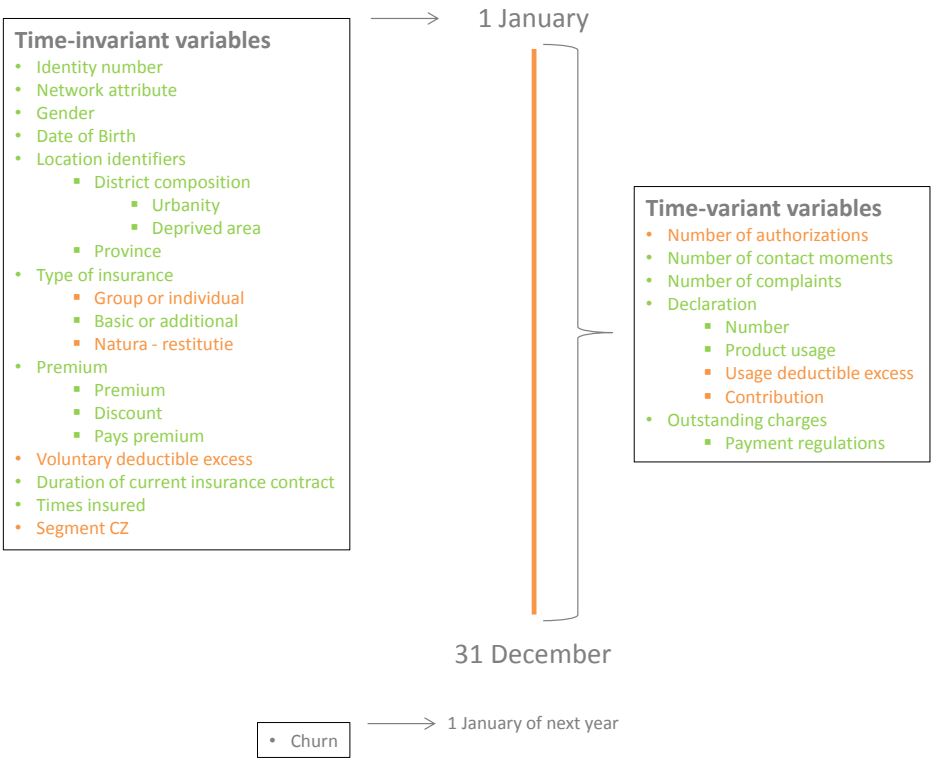


FIGURE 2.1: Variables ordered into time-variant variables and time-invariant variables. The variables in the left box are time-invariant and measured on 1 January. For the variables in the right box the measurement takes place on 31 December because they are time-variant.

2.4 Method to collect the data

The data is extracted from SAS Enterprise Guide. The year 2013 is taken as the measurement year. The database contains 2.8 millions customers which is too large to fully take into consideration. In the literature data sets extracted between 1,800 to 340,000 cases are generally found [22, 32, 67, 74]. In this research a random subset database is extracted containing information on 30,000 customers. Roughly between 10 to 35 variables are used in literature studies, e.g. Hung *et al.* and Günter *et al.* use 10 variables and Kim *et al.* 35 [22, 32, 40]. Table 2.5 shows how many cases and variables each paper uses. In this research 24 variables are selected, as discussed in Section 2.3.

The data is selected from SAS Enterprise Guide with help from the business intelligence division. The data set contained a wide range of diversity within variables which resulted

Literature	# Cases	Real churn percentage	# Variables
Coussement *et al.* [12]	134.120	11,95%	24
Coussement *et al.* [13]	90.000	11,14%	32
Coussement *et al.* [14]	12.764	18,5%	20
Farquad *et al.* [19]	14.814	6,76%	22
Günter *et al.* [22]	160.000	confidential	10
Huang *et al.* [30]	827.124	3,3%	7
Hung *et al.* [32]	160.000	8,75%	10
Keramati *et al.* [38]	3.140	15,7%	13
Kim *et al.* [40]	89.412	9,7%	36
Mozer *et al.* [48]	2.876	6,2%	134
Risselada *et al.* [58]	1.474	unknown	6
Tsai & Chen [64]	37.882	unknown	22
Verbeke *et al.* [67]	338.874	14,1%	22
Zhao *et al.* [73]	2.958	10,3%	171
Zhu *et al.* [74]	1.780	7,35%	11

TABLE 2.5: Sizes of data sets and number of variables used in the literature indicated per paper.

in the need to reduce and simplify the variables. For each situation multiple discission are made. An example is that a product's name can change in time, however if a customer did not change the health insurance, the old product is still in use by this customer. This resulted in 67 different labels for the type of insurance variable in the database. This was simplified to two cases: basic or additional health insurance. Furthermore, some socio-demographic groups are excluded to generate a data set which represents the main market for the company. The following list gives an overview of the excluded groups:

- Because of privacy reasons all police officers are excluded from the research.
- All foreigners are excluded because they can choose from other products than regular customers.
- Customers registered after 1 January are not taken into consideration to generate an equal measurement period for all customers [23].
- All included customers have a basic health insurance at the company.
- Customers who died during the measurement period are excluded.
- Customers who serve time in prison are excluded because there are different regulations for this group.

Some variables are selected but to collect these variables an assumption needs to be made. This is also the case in the example of the 67 different labels for the type of insurance variable. The following assumptions are made:

- The product types are simplified to basic and additional health insurances.

- The authorisations are counted in the year they are handled. When authorisations are reopened, are they counted as a new authorisation.
- Authorisations and complaints are not split up in acceptance and rejection because the number of authorisations and complaints is limited.
- There are three types of complaints: objections, disputes and regular complains. These three types of complaints all have a different procedure but are all stored under complaints.
- Payment regulations are counted in the year the regulation started.

There are two variables extracted from information of the CBS (Statistics Netherlands). We included if customers live in an urban area and a deprived area. These variables are the results of the location identifier variable, indicated by one of the experts of the customer service division. The urban area needs to be calculated. The level of urbanity (UA) is calculated according to Equation 2.1.

$$\overline{\text{UA}} = \sum_{N} \text{Pur} \times \text{Lev} \tag{2.1}$$

N	zip code	Pur N	Pur	Lev	Pur × Lev
1	6411	5	0.14	2	0.29
2	6411	6	0.17	5	0.86
3	6411	6	0.17	5	0.86
4	6411	6	0.17	4	0.69
5	6411	6	0.17	4	0.69
6	6411	6	0.17	4	0.69
Total		35			**4.06**

TABLE 2.6: AU calculation for the zip code 6411. Pur N indicates the neighbourhood purity with 6 as high and 1 as low. An urbanity level (Lev) of 5 indicates a high urban area and 1 a low urban area.

With N the neighbourhoods with the same zip code, Pur the normalised fraction of the neighbourhood in the selected zip code and Lev the urban area level per neighbourhood. Pur reaches from a high fraction (6) to a low fraction (1) indicated by CBS. The urban area level of the neighbourhoods is scaled from 5 (high) to 1 (low). Figure 2.2 displays the urbanity level per municipality in the Netherlands. When we zoom in to the zip code 6411 located in Heerlen (Figure 2.3) we see that the urbanity within this zip code differs. Table 2.6 shows the calculation of $\overline{\text{UA}}$ for zip code 6411. First all neighbourhoods of this zip code are selected, which results in 6 neighbourhoods (N = 6). Second, the corresponding fractions of the neighbourhoods are summed to calculate the normalised fraction per neighbourhood (Pur). This normalised fraction is multiplied by the urban area level of the total neighbourhood, and is summed resulting in the $\overline{\text{UA}}$ level of zip code 6411.

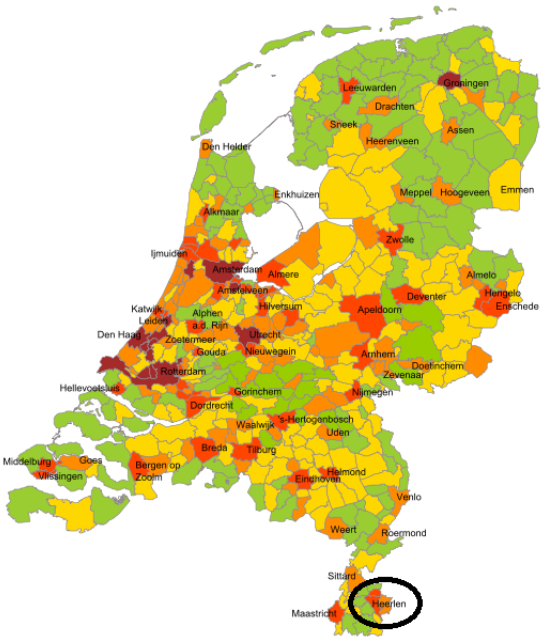


FIGURE 2.2: Urbanity level per municipality, with red indicating a high urbanity level and green a low level. Source: CBS.

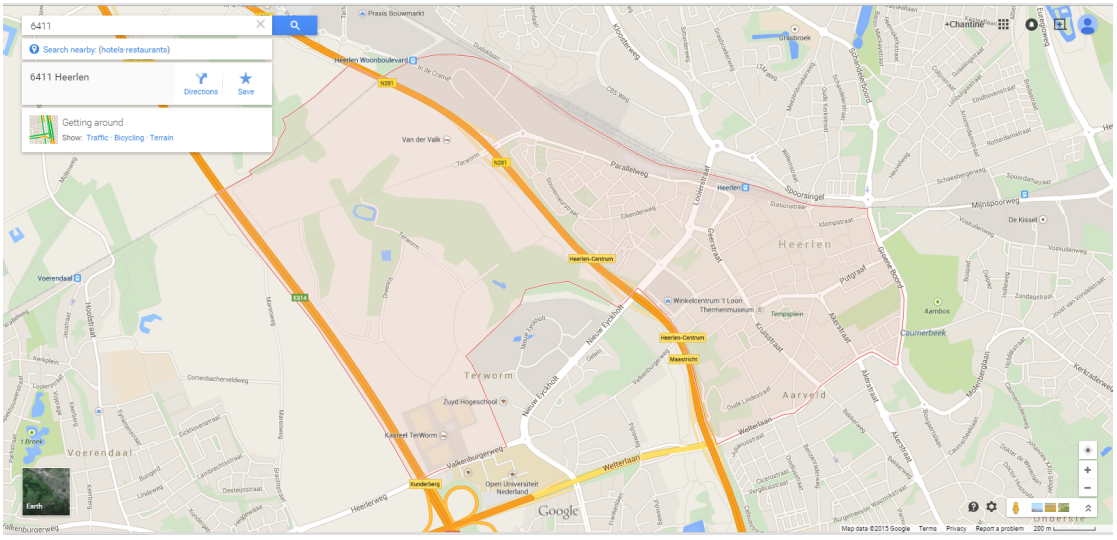


FIGURE 2.3: Area of the zip code 6411 marked in red (Source: Google Maps).

Deprived area is included as a dichotomous (i.e. true or false) variable and is based on the zip code of the customer.

2.5 Preparation of the data set for model generation

When all the variables are extracted from the data base of CZ, the data is examined on normality and if dichotomous variables are equally distributed. When this is known

the missing values and extreme values are analysed. This Section will conclude with the variables that are transformed.

Distribution of the data

Table 2.7 shows the Kolmogorov-Smirnov, the test is used to see if the variables are normally distributed. However all variables are significant for the test, which means that none of the variables represent a normal distribution. A drawback of these test is that a significance level is easily reached with a large data set [20]. In this research 10,000 cases are used which means that a large data set is used. To make a good informed conclusion the variables are also plotted (Appendix B). These plots also show that non of the variables represent a normal distribution, which supports the conclusion of the Kolmogorov-Smirnov test.

Variable	Kolmogorov-Smirnov Sig. level
Age	0.00
Premium	0.00
Discount	0.00
Consumption	0.00
Deductible Excess	0.00
Contribution	0.00
Urbanity	0.00
Nr. of complaints	0.00
Nr. of contacts	0.00
Nr. of declarations	0.00
Nr. of authorisations	0.00
Nr. of payment regulations	0.00
Nr. of times insured	0.00
Duration of contract	0.00
Family size	0.00

TABLE 2.7: The table shows the results of the Kolmogorov-Smirnov test to test for normality.

All the dichotomous and ordinal variables of the data set are distributed as expected, which is shown in Figure 2.4. The Y-as is not included because of confidential reasons, but the proportions represented in Figure 2.4 are representative for the variables in the data set. The variables represented in this Figure are churn (C) & non-churn (NC), gender, if customer pay the premium themselves (PP) or not (N-PP), if they are living in a deprived area (DA) or not (N-DA), if they have a group insurance (GI) or have an individual insurance (N-GI), what type of insurance they have and if they have an additional insurance (AD) or not (N-AD).

Missing data

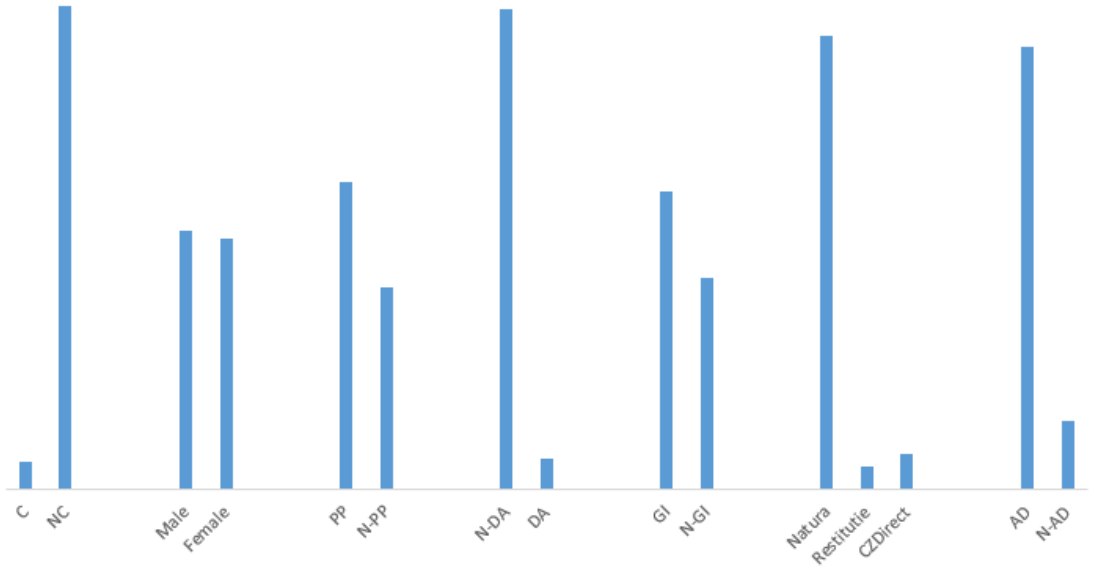


FIGURE 2.4: Differences in the dichotomous and ordinal variables.

During the extraction of the data from SAS Enterprise Guide not all variables are filled. If a variable is not completely filled this can be interpreted as that the customer did not use this service. All these "missing values" are replaced with a zero to make the data set complete. Missing values are most often seen with time-variant variables, while time-invariant variables are always completely stored in the database. There is no need to investigate these missing values on missing at random or missing completely at random [21], because a non present value can be interpreted as a non use of the service which is not a missing value.

Extreme values

In the data set no outliers are detected. It is important to recognize that these extreme values exist but are real, therefore no further actions are needed. An example of a variable with extreme values is age, within the data set are customers included who reach the age of 95 years old. Only eight customers included in the data set are 95 years old or older. During the data selection, discussed in Section 2.4, the selection is made which cases are excluded in the research. This resulted in an exclusion of the exceptional profiles.

Variable transformation

The total number of switch opportunities a customer has during his or her insurance at CZ is an important variable to consider when looking at churning profiles. The variable is closely related to age and this is especially apparent for older customers who often are insured for prolonged periods of time, which is visually displayed in Figure 2.5.

$$\text{Duration of contract} = \frac{\text{Switch opportunities}}{\text{Age}} \tag{2.2}$$

Therefore, the variable is corrected for the age of the customer by simply dividing the number of switching opportunities by the age (Equation 2.2). This new transformed variable is used throughout the rest of this study. All other variables are used as they are collected from the database.

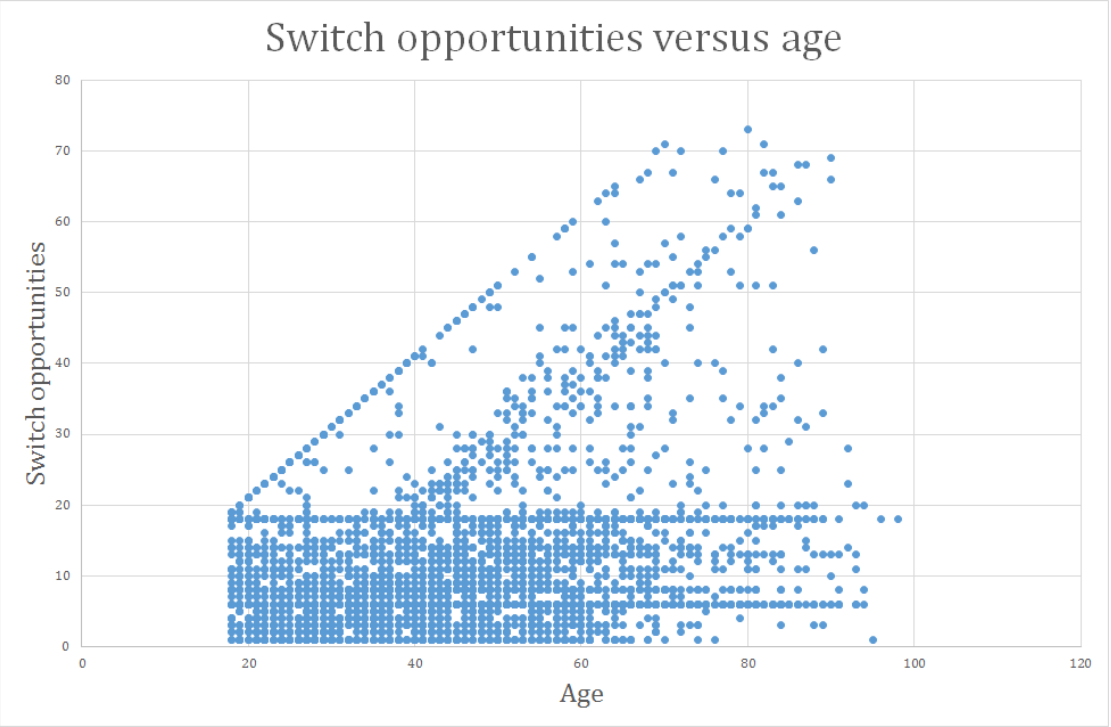


FIGURE 2.5: Switch opportunities compared with age to calculate normalised value for the duration of the current contract.

2.6 Imbalanced data set problems

Most techniques are driven to generate a high accuracy. A highly unbalanced data set will result in a model which is neglecting the minority class, because the accuracy will still be 95% or higher. However, the minority class is usually the more important class [72]. As shown in Table 2.5 the churning customers are always the minority group, which is also the case for this research. The model which is generated with the data set will reach an accuracy of 95% and mark the minority class as noise. To make sure that these churning customers are not seen as noise the possibilities of under- and oversampling are investigated.

In the literature several re-sampling strategies are discussed, such as random oversampling with replacement, random undersampling, directed oversampling, oversampling with informed generation of new samples and combinations of the above techniques [5, 7]. Some drawbacks exist for random sampling strategies, e.g. the random undersampling strategy can cut out valuable cases. With random oversampling the cases of the minority class are duplicated which can result in overfitting problems [41]. The directed strategies are comparable to the random strategies, however these strategies make an informed choice to duplicate or cut out cases. Chawla *et al.* , Kotsiantis *et al.* and Yen *et al.* conclude that oversampling with informed generation of new samples works better than random oversampling and prevents overfitting [5, 41, 72].

However, in this research a re-sampling strategy is chosen which does not duplicate minority cases or deletes majority cases, because it is possible to extract enough minor and major cases from the original data set to generate a balanced data set. It is unclear if the data set ratio should be a 50:50 learning distribution or that it should be another ratio [68]. Chawla *et al.* indicate that ratio is mostly empirically determined [7]. Therefore, in this research a wide range of ratios, (non-churn:churn) 50:50, 66:33, 70:30 and 80:20, are used for the training sets. The test data will be a random sample of the original data set.

Chapter 3

Comparative analysis of churning and non-churning profiles

Now the data set is collected, the information which is stored in the data set is analysed. In Section 3.1 the differences between the whole population of the Netherlands and the population of CZ are discussed. When these differences are known the differences between churners and non-churners of CZ are compared, discussed in Section 3.2.

3.1 Information stored in the data compared with the population of the Netherlands

For a better understanding of what the customer population of CZ looks like, a comparison is made with the overall population of the Netherlands. If the population of CZ is statistically similar to the population of the Netherlands, the models made specifically for CZ can be generalised and potentially used in other applications. Real percentages and values are not mentioned because of confidentiality. Because the sample size is large, small differences in populations can already lead to a statistically significant difference [20]. Therefore, the z-score is reported [21], which gives the absolute difference in means of two populations (μ_1 and μ_2), normalized for the standard deviation σ of the largest population:

$$z = \frac{|\mu_1 - \mu_2|}{\sigma^2} \tag{3.1}$$

Tables 3.1 and 3.2 and Figure 3.1 give an overview of the measured differences relevant for the insurance market, defined by the NZa [53]. For these dichotomous variables,

a one-sample binomial test was used and tested the H_0-hypothesis that the difference between Netherlands and CZ population is zero. Table 3.1 compares socio-demographic and product-related variables. This Table shows that all variables differ significantly, for a significance level of 5%. This is due to the large sample size which is used for this research. However, the z-score is small for all variables in the Table. This means that the population of CZ is comparable with the population of the Netherlands. The differences between churn and group insurance in both populations are slightly higher. The reason that the churn rate of CZ is slightly different is due to large group insurance switches between health insurance companies. CZ was not involved with these switches, which explains the difference. The group insurance rate of CZ also differs slightly, a reason for this can be that CZ does not have many special group insurances.

The significance level of the premium variable cannot be calculated because this is not a dichotomous variable. For this reason the one-sample binomial test cannot be applied. The Wilcoxen signed rank test can be used to calculate the significant difference in this case. However, the mean premium of the Netherlands is given by the NZa and not the median [53], which makes it impossible to use this test.

	Netherlands		Sig. level between CZ and NL	z-score
Churn	8.3%	(a)	0.000	0.14
Male	49.5%	(b)	0.010	0.02
Deprived area	4.7%	(a)	0.000	0.05
Premium	€ 1213	(a)	-	-
Group insurance	68%	(a)	0.000	0.19
Additionally insured	85.7%	(c)	0.001	0.03

TABLE 3.1: Differences between the population of the Netherlands and all customers of CZ. Significant levels and z-scores were calculated based on the average difference between populations. Sources: a: NZa, b: CBS, c: Vektis.

Table 3.2 shows that for all levels of the voluntary deductable excess significance differences are found. However, the z-scores for all variables are small, so the differences between the populations are relatively small.

Voluntary deductable excess	Netherlands	Sig. level between CZ and NL	as z-score
€ 0	90.3%	0.000	0.05
€ 100	1.4%	0.000	0.02
€ 200	1.1%	0.000	0.02
€ 300	0.7%	0.000	0.00
€ 400	0.2%	0.000	0.00
€ 500	6.2%	0.000	0.06

TABLE 3.2: Differences of voluntary deductable excess between the population of the Netherlands and all customers of CZ.

Figure 3.1 shows how the population is divided over the provinces. CZ operates mainly in the south of the Netherlands, especially in the provinces Noord-Brabant, Limburg and Zeeland. This is because CZ has a traditional origin, at first CZ was only responsible for the health insure of the population in the south of the Netherlands.

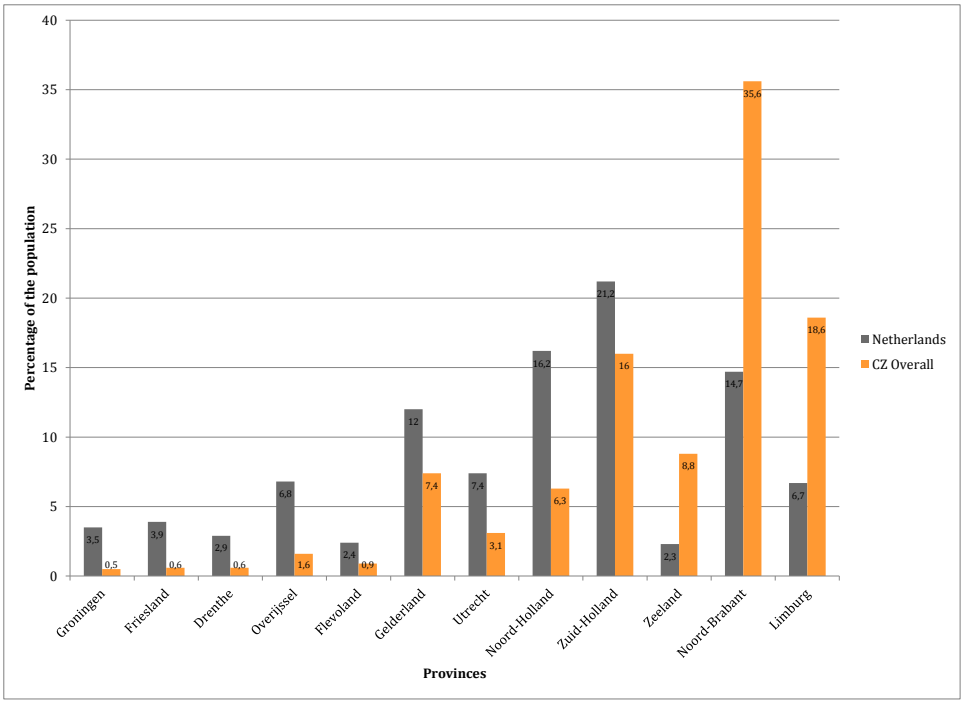


FIGURE 3.1: Differences in living area between the population of the Netherlands and all customers of CZ.

The population of CZ is significantly different from the population of the Netherlands however, the differences are very small. The only differences that could be considered, is that the market of CZ is mainly located in the south of the Netherlands.

3.2 Statistical differences between a churning and non-churning profile

To assess the differences between the churning and non-churning population at CZ, several statistical tests were performed. To compare the means of both populations, a Student's t-test can be performed. First however, equality of variances was checked using Levene's test, to determine which t statistic should be used [20]. Levene's test is used instead of an F-test, because the tested populations are strictly not normally

distributed (see Chapter 2). When the Levene's test is significant ($p<0.05$) than H_0 can be rejected:

H_0: The variances of the considered variable in the churning and non-churning population are equal.
H_1: The variances of the considered variable in the churning and non-churning population are unequal.

Significance levels of Levene's test on important variables are shown in Table 3.3. When significant, a t-test assuming unequal variances is conducted, when the significance level is above 0.05, equal variances are assumed. The corresponding significance levels for the t-test are represented in the second column of Table 3.3, based on the following hypotheses (with $p<0.05$):

H_0: The mean of the considered variable in the churning and non-churning population are equal.
H_1: The mean of the considered variable in the churning and non-churning population are unequal.

If the t-test is significant then there is a significant difference between churners and non-churners. The results are based on two-sided tests with significance level of 0.05. The variables age, urbanity, the number of times someone is insured and the number of declarations show the largest differences measured with the z-score.

Besides the continuous variables, the dichotomous variables are also investigated by comparing column proportions using the same H_0-hypothesis as for the t-test. The results are represented in the Tables 3.4 and 3.5.

There are no significant differences between churners and non-churners regarding gender and the deprived area variables. Churning customers do not have a group insurance and additional insurance as often as non-churning customers, the difference shown in the data set is significant (shown in Table 3.4).

For the voluntary deductable excess no remarkable differences are found. Table 3.5 shows that the there is no differences or the differences are small. We cannot conclude that the churning profiles really differ from non-churning profile due to the small z-scores regarding voluntary deductable excess.

Overall we can conclude that churning profiles significantly differ from non-churning profiles. The main differences can be found in the variables age, number of times insured, urbanity, group insurance and additional insurances.

Variable	Levene's test Sig. level	t-test for equality of means Sig. level	z-score
Age	0.000	0.000	0.48
Premium	0.004	0.000	0.20
Discount	0.009	**0.144**	-
Consumption	0.000	0.000	0.02
Deductible Excess	0.000	0.000	0.25
Contribution	0.034	0.022	0.05
Urbanity	**0.348**	0.000	0.32
Nr. of complaints	**0.960**	**0.984**	-
Nr. of contacts	0.000	**0.105**	-
Nr. of declarations	0.000	0.000	0.29
Nr. of authorisations	0.020	0.000	0.10
Nr. of payment regulations	0.000	0.001	0.13
Nr. of times insured	0.000	0.000	0.34
Duration of contract	0.043	**0.133**	-
Family size	**0.696**	0.038	0.09

TABLE 3.3: The table shows the results of the Levene's test for equality of variances and the t-test for equality of means. The table provides insight in the differences between churners and non-churning customers.

Variable	Sig. level between churn and non-churn	z-score
Male	No	-
Deprived area	No	-
Group insurance	Yes	0.30
Additionally insured	Yes	0.36

TABLE 3.4: The table shows for the dichotomous variables if the differences are significant. The z-score shows the magnitude of the difference.

Voluntary deductable excess	Sig. level between churn and non-churn	z-score
€0	Yes	0.17
€100	Yes	0.02
€200	No	-
€300	No	-
€400	No	-
€500	Yes	0.20

TABLE 3.5: Differences of voluntary deductable excess between churners and non-churners of the customers of CZ.

Chapter 4

Data mining techniques for churn prediction

To discover which churn prediction techniques are widely used in the literature, a literature study is performed. A structured research methodology is used based on the strategy of Jourdan *et al.* [36].

Accumulation of article pool

According to an article by Lawrence and Giles published in Science any one search engine is limited in covering all relevant literature [43]. The authors indicate that "*no single engine indexes more than about one-third of the "indexable Web", and by combining the results of six engines the results yield about 3,5 times as many documents on average as compared with one engine*". Lewandowski also advised that researchers should use multiple search engines [44]. Indeed, in this literature review also six search engines are used; ACM Digital Library, IEEE Xplore, Ingenta Connect, Science Direct, Springer and Web of Science.

Search engine	indexes
ACM Digital Library [1]	44 high impact journals 275 proceedings added each year
IEEE Xplore [33]	160 journals added each year 1.200 proceedings added each year
Ingenta Connect [35]	5 million articles in 10.000 publications
Science Direct [59]	2.500 journals 26.000 books
Springer [63]	2.200 journals 110.000 books, 8.400 added in 2013
Web of Science [70]	12.000 journals 150.000 conferences

TABLE 4.1: Indexing range of selected search engines.

The selected search engines index a broad range of literature. For example, IEEE Xplore focusses on proceedings, adding 1,200 each year, while Science Direct provides mainly access to journals and books. This provides an access to a wide range of relevant literature which is the goal of selecting search engines. The used search term is based on the relevant terms of CRM which is discussed in Section 1.1. There are no specific data mining techniques selected because during the literature selection no specialisation was preferred. From the selected literature, the most frequently used techniques should be discovered which results in this specialisation. The search term which is used to collect the literature is:

```
‘‘data mining’’ AND (‘‘customer loyalty’’ OR ‘‘customer retention’’
OR ‘‘customer churn’’ OR ‘‘customer behavior’’)
```

Preferably the title includes a keyword or combination of keywords. Most search engines give the option to only search in the title of papers. If there are less than five results or if this option does not exist in the search engine, the possibility to include the abstract as search field was applied. The results of the search engines were all checked on the following selection criteria:

- Papers are written in the English language.
- The paper is no older than 20 years.
- The keywords or related terms should be present in the title or the abstract.
- The result is no citation or patent.

The generated search results are reviewed on the title of the paper. The total number of selected papers based on the title is 27. Subsequently, the papers are reviewed in more detail, the abstract, introduction and conclusion is read from each paper. The selected papers are now reviewed and subjected to rejection criteria. The rejection criteria are the following:

- There is no access to the whole article.
- There is no application in *data mining* or *customer retention*.

Sixteen research papers and two literature reviews were accepted after applying the selection criteria, Appendix C gives an overview of the accepted papers.

Categorization by category

Which techniques are used in the research papers and literature reviews is shown in Table 4.2. The table is split in three columns, the two literature reviews selected during the selection and the last column represents a summary of the selected literature during this literature study. The top four most mentioned techniques of the literature reviews,

shown in Table 4.2, are selected and compared. The profiling techniques in Table 4.2 are used in the selected research papers to segment the customers. In the literature study performed for this research the techniques are used for prediction, classification and feature selection. In this research K-means and SOM are used to profile customers in homogeneous groups.

Data mining technique	Literature review by KhakAbi *et al.* [39]	Literature review by Tsai & Lu [65]	Literature study for this research
Prediction techniques			
Neural Networks	15	10	4
Decision Tree	13	9	9
Logistic Regression	13	8	2
Support Vector Machine	7	2	0
Profiling techniques			
SOM	2	1	2
K-means	1	1	3

TABLE 4.2: Used data mining techniques split up per category compared with the churn prediction methods of two literature reviews.

Techniques used in the literature to predict customer churn

A diversity of techniques are used in this research to not only make an accurate prediction but also to give more insight. The techniques given in the Table 4.2 are all applied to the data. This is done so that the best model is generated for the problem and results in more insight. In Chapter 5 are the profiling and prediction techniques discussed. For the prediction techniques the performance indicators are discussed in Section 5.2.1.

Chapter 5

Application of profiling and prediction techniques

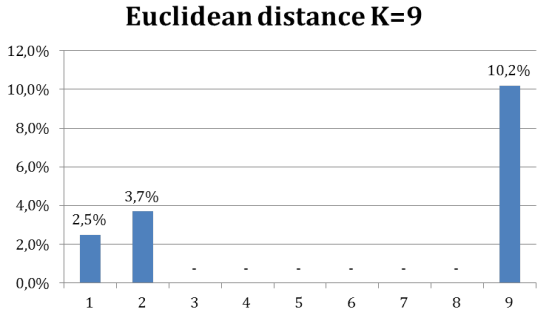
In Chapter 4 the techniques used in the literature are identified. In this chapter these techniques are applied on the data set of CZ. The data analytic tool Konstanz Information Miner (KNIME) is used. First the profiling techniques K-means and Self-Organizing maps are discussed (Section 5.1). Then is discussed how the prediction models were evaluated (Section 5.2.1) and finally the prediction models are discussed: models generated with Logistic Regression (Section 5.2.2), Decision Tree (Section 5.2.3), Neural Networks (Section 5.2.4) and finally with Support Vector Machines (Section 5.2.5).

5.1 Profiling of the selected customers

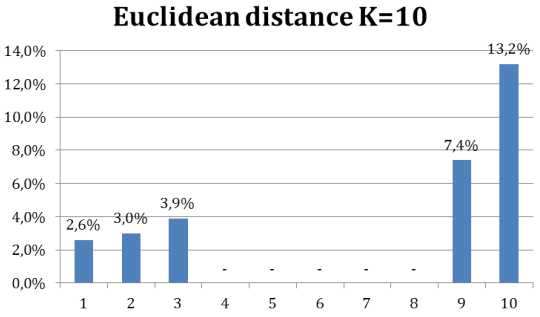
Two profiling methods are used to generate homogeneous profiles. In Section 5.1.1 K-means is discussed and in Section 5.1.2 the Self-Organizing Maps. In Appendix D the exact settings can be found. The settings which are changed are discussed separately per technique.

5.1.1 K-means

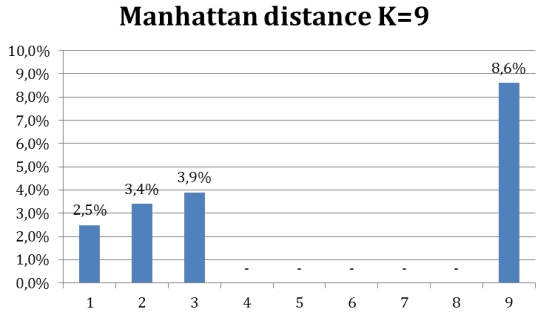
K-means clustering aims to partition the cases, for this research the customers, into K clusters. Each case is part of that cluster, which has its centroid closest to the case centroid. The centroid is a mean value for all the variables. For K-means the Euclidean distance and the Manhattan distance are used to calculate the centroid. The number of clusters ranges from $K = 2$ to $K = 10$. This range is chosen to see if the clustering



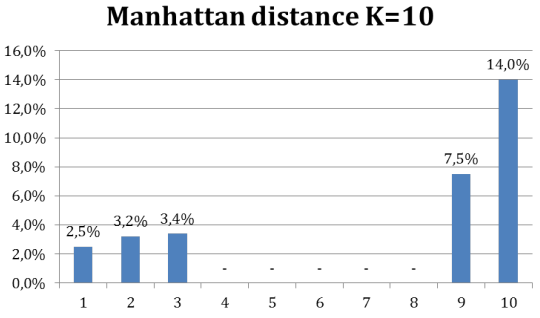
(A) K-means clusters with $K = 9$ and the Euclidean distance.



(B) K-means clusters with $K = 10$ and the Euclidean distance.



(C) K-means clusters with $K = 9$ and the Manhattan distance.



(D) K-means clusters with $K = 10$ and the Manhattan distance.

FIGURE 5.1: Representation of the clusters, clusters marked with **-** represent clusters with an average churning rate which is confidential.

algorithms classify the churning and non-churning customers in separate groups with homogeneous profiles. It is important to normalize the variables before the algorithm is applied. When the variables are not normalized, the weight of the variable with the largest variation is larger [29].

The K-means clustering technique is applied on the original data set. For each generated cluster within each setting the churn rate is calculated. Also are the averages per cluster analysed to see what the differences are between each cluster. With these values the different profiles are identified.

We found that for $K < 8$ all generated clusters do not differ much. With $K = 9$ and $K = 10$ more diverse profiles can be recognized. There are a couple of clusters that show average results on most variables, however there are also multiple clusters which show differences and have a higher (churn rate >7%) or lower percentage (churn rate <4%) of churners compared with the average of the population of CZ. Figure 5.1 shows the churn rates per cluster for $K = 9$ and $K = 10$. The clusters with an average churn rate are included in the graphs, however these are not representing values due to confidential issues. Also for this reason are the clusters not included in Tables 5.1 to 5.3. The

churn rate is determined by combining the churning variable per case after profiling and then the percentage of churning customers per cluster is calculated. Tables 5.1 and 5.2 show the six variables that differ most with the average rate. The variables that showed differences are age (A), the number of times a customers was insured (TI), if a customers pays the premium themselves (PP), customers with or without group insurances (GI), if they have a voluntary deductible excess (VDE) and consumption (C).

The most important finding from Table 5.1 is that there are four types of profiles. Older customers, who have no voluntary deductible excess and consume more health insurance than average, are mostly non-churning customers. Young customers, who consume less health insurance than average and pay the premium themselves do churn more often. Young customers which do not pay the premium themselves and have a group insurance do not churn as often compared to the average population. The fourth profile is not shown in Table 5.1, this profile is comparable with the average of the population. This profile is not considered because it does not give more homogeneous groups than the selected cases at random with respect to churn.

K = 9	A	TI	PP	GI	VDE	C
Cluster 1 (churn rate = 2.5%)	↑	↓	↑	↑	↓	↑
Cluster 2 (churn rate = 3.7%)	↓	↑	↓	↑	=	↓
Cluster 3 (churn rate = 10.2%)	↓	↑	↑	↓	↑	↓
K = 10	A	TI	PP	GI	VDE	C
Cluster 1 (churn rate = 2.6%)	↑	↓	↑	↑	↓	↑
Cluster 2 (churn rate = 3%)	↓	↑	↓	↑	=	↓
Cluster 3 (churn rate = 3.9%)	↑	=	↑	↑	↓	↑
Cluster 4 (churn rate = 7.4%)	↓	↑	↑	↓	↓	↓
Cluster 5 (churn rate = 13.2%)	↓	↑	↑	↓	↑	↓

TABLE 5.1: Profiles defined with K-means clustering, $K = 9$ and $K = 10$ and the Euclidean distance are used. The remaining clusters show an average profile and are therefore not mentioned. An arrow going up means that the value is higher than the average over all the clusters, an arrow going down the opposite.

Table 5.2 (with the Manhattan distance) shows that the profiles of the clusters are comparable with Table 5.1 (with the Euclidean distance). For this reason it does not make a difference to choose one or the other. However, when the clusters are compared on the coverage level of all the churners stored in the data set, the Manhattan distance works better (see Table 5.3). Using Euclidean distance, 12.1% and 18.6% of the total churner population is reached when all clusters with a churn rate above 7% are combined, for $K = 9$ and $K = 10$, respectively. For the Manhattan distance 20.1% and 29.7% of the churners is reached. K-means clustering is used in Section 6.4 to create clusters with a homogeneous profile before applying the prediction models.

K = 9	A	TI	PP	GI	VDE	C
Cluster 1 (churn rate = 2.5%)	↑	=	Yes	Yes	No	↑
Cluster 2 (churn rate = 3.4%)	↑	=	Yes	Yes	No	↑
Cluster 3 (churn rate = 3.9%)	↑	=	Yes	No	No	↑
Cluster 4 (churn rate = 8.6%)	↓	=	Yes	No	No	↓
K = 10	A	TI	PP	GI	VDE	C
Cluster 1 (churn rate = 2.5%)	↑	=	Yes	Yes	=	↑
Cluster 2 (churn rate = 3.2%)	↑	=	Yes	No	=	↑
Cluster 3 (churn rate = 3.4%)	↑	=	Yes	Yes	=	↑
Cluster 4 (churn rate = 7.5%)	↓	=	Yes	No	=	↓
Cluster 5 (churn rate = 14%)	↓	=	Yes	No	=	↓

TABLE 5.2: Profiles defined with K-means clustering, $K = 9$ and $K = 10$ and the Manhattan distance are used. The remaining clusters show an average profile and are therefore not mentioned. An arrow going up means that the value is higher than the average over all the clusters, an arrow going down the opposite. The yes and no values mean if the profile includes the representing variable or not.

	Coverage of the total amount of churners	
K = 9	Euclidean distance	Manhattan distance
Cluster 1	7.5%	6.5%
Cluster 2	4.5%	9.5%
Cluster 3	12.1%	8.0%
Cluster 4	-	20.1%
K = 10	Euclidean distance	Manhattan distance
Cluster 1	7.5%	6.5%
Cluster 2	3.5%	5.5%
Cluster 3	7.5%	9.5%
Cluster 4	7.0%	14.1%
Cluster 5	12.6%	15.6%

TABLE 5.3: Profiles defined with K-means clustering, $K = 9$ and $K = 10$ and the Manhattan distance are used. The remaining clusters show an average profile and are therefore not mentioned. The percentages represent the part of all the churners.

5.1.2 Self-Organizing Maps

Self-Organizing Maps (SOM) is an unsupervised learning technique, a process of self-organization, and aims at data reduction. The key advantage of this technique is the retention of topological information in the SOM. N-dimensional data is mapped onto a 2-dimensional grid, $\mathbb{R}^N \rightarrow \mathbb{R}^2$. The SOM process is iterative, with every iteration the map nodes are updated by one input node. The dimension of a map node is equal to the dimension of an input node. SOM can handle large amount of data, which is key for this research project. The number of cases which is extracted for the training set is 5,000. Another advantage of SOM is the natural start, which means that no random starting point has to be selected. This is advantageous, because with random selection two cases which belong to the same group can be selected as centroid [42].

There is no rule for the selection of the best training parameters, these parameters are selected by trail and error [42]. During the first trial the default settings of all parameters are used. The settings for the number of learning epochs is based on the method used by Kuo *et al.* , which tested 1000, 1500 and 2000 epochs [42]. When the computational time becomes to long, the number of epochs will be reduced. The learning rate is varied between 0.8 and 1.2 to improve the homogeneity of the profiles per cluster.

Unfortunately, the results do not show differences when the parameters differ, only the order of the clusters differ. The churn percentage differs slightly per cluster, two clusters have a slightly lower and two a slightly higher churn percentage compared with the average churn rate of the population. When the variables are compared, no large differences are shown to give an explanation for the differences. Because the differences are small between the clusters, clustering using SOM cannot be used to profile the customers.

5.2 Churn prediction model generation

For the generation of the prediction models five training sets were constructed, which differ in the ratio between non-churners and churners. This is done to overcome the imbalanced data set problem which was discussed in Section 2.6. Model 1 represents the training set with the original distribution, model 2 the 80:20 (non-churning:churning) distribution and model 3 to 5, 70:30, 66:33 and 50:50 (non-churning:churning) respectively. These training set distributions are always linked to the same model number for all techniques. Figure 5.2 gives an overview of how these training sets are generated. First three data sets of 10,000 cases are selected from the database of CZ. One data set which represents the original distribution between churners and non-churners, a set with only churning customers and one set with non-churning customers are extracted from the CZ database. Thereafter the training sets and the test set are created. The test set will be used to test all models.

The training sets are used in four different techniques to predict which customer is going to churn. Per technique different settings are used which gives different results. The settings which are changed during the model generation are discussed per technique. The overall settings can be found in Appendix D for reproducibility.

In this paragraph first the performance measurements are discussed which are used to qualify the models (Section 5.2.1). In Section 5.2.2 the models generated with logistic regression are discussed. Section 5.2.3, 5.2.4 and 5.2.5 include the models from the decision tree algorithm, neural networks and support vector machine.

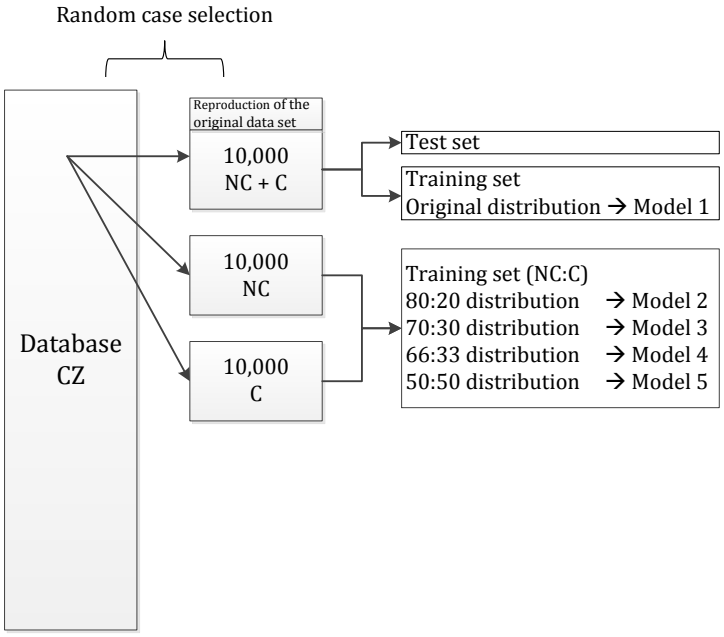


FIGURE 5.2: Creation of the training and test sets from the data base of CZ.

5.2.1 Performance measurements applied to the generated models

To measure the performance of the models with a binary target variable, a confusion matrix is used, shown in Table 5.4.

Real churn	Predicted churn	
	1	0
1	True Positive (TP)	False Negative (FN)
0	False Positive (FP)	True Negative (TN)

TABLE 5.4: Confusion matrix for binary classification.

With this table the quality of the model can be assessed. The quality measurements are classification accuracy, sensitivity and specificity (Equations 5.1 to 5.3). As discussed in Section 4.2 it is needed to deal with imbalanced data. The distributions which are discussed in the introduction of Section 5.2 are used as training set for the model generation. The testing set will always have a distribution equal to the distribution of the original data set. An imbalanced data set can give some problems when a model is assessed on accuracy, defined in Equation 5.1, it will classify all the cases as majority class. For example, if the the population consist of a positive to negative class ratio of 1:9 and the model classifies every customer with negative an accuracy of 90% is reached. Therefore, this quality indicator is too simplistic [10, 55].

$$\text{Classification accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

It is more interesting to see how many churners are correctly predicted. This is calculated with Equation 5.2, and represents the true positive rate. Equation 5.3 gives the true negative rate, and together these two equation give more insight in the purity of the predicted group.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5.2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.3}$$

Precision gives the accuracy over the cases which are predicted to be positive. This is an accuracy measurement which indicates how well the model predicts the cases which are labelled as positive [38]. In this research the positive cases are labeled as churners.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.4}$$

The Cohen's Kappa value [10] is used to identify the optimal threshold value (Equation 5.5). The Kappa value favors the correctly classified minority cases over the majority cases [37], which is also what is preferred in this research. The largest Kappa value is used to detect the optimal threshold value. The confusion matrix that belongs to the optimal threshold value is used for the calculation of precision and sensitivity. For each model generated in the Sections 5.2.2 to 5.2.5, the precision and sensitivity is calculated.

The Receiver Operating Characteristic (ROC) curve [4, 25] and the Area Under the Kappa (AUK) curve [37] are calculated with confusion matrices based on different threshold values. The ROC curve is a graphical illustration of how well a model predicts and the most popular tool which is used over the years to rank model performance [4, 6, 12, 25, 40]. The larger the area under the curve (AUC) the better the model is. When a model has a AUC value <0.5 than the model predicts worse than selection at random. Figure 5.3 shows three lines which indicate which curve is generated by a good prediction model. In an ideal situation, the curve quickly increases in sensitivity for low specificity and stays constant for higher specificity. The ROC curves often intersect each other, making the visual classification of the models difficult [37]. For this reason the AUC is calculated and compared for model classification.

Because the ROC-curve is applied in a wide range of research areas to measure model performance, the results for the AUC values are also checked in this research. However a disadvantage of the AUC value is that it does not consider the weigh of a false negative and false positive prediction. This is an important factor because false negative predictions are more likely to occur than false positive predictions with an imbalance data set [46]. Therefore, the Area Under the Kappa curve (AUK) is measured, this criteria takes the class skewness in the data into consideration [37]. For the Kappa curve the x-axis represents the rate of false positives and the y-axis represents the Cohen's Kappa value. Equation 5.5 calculates the Kappa value, which can be seen as a nonlinear transformation between the true positive rate and true negative rate [37]. The parameters of Equation 5.5 can be calculated with $p = TP + FN$, $n = FP + TN$, $t = \frac{TP}{p}$ and $f = \frac{FP}{n}$.

$$\kappa = \frac{2p(1-p)(t-f)}{p + (1-2p)f + p(1-2p)(t-f)} \tag{5.5}$$

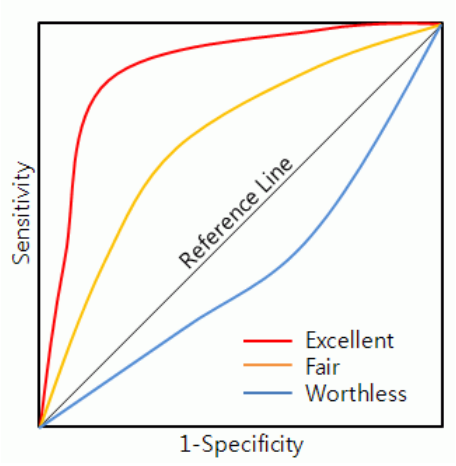


FIGURE 5.3: ROC curves which indicate when a model generates a good prediction versus a bad prediction.

With the four performance parameters AUK, AUC, precision and sensitivity the best model will be selected per technique and then compared with the performances of the other techniques.

5.2.2 Logistic Regression

Logistic regression is a specialized form of regression that can be used to predict or profile a binary, two-group, categorical variable. With logistic regression a probability is estimated how likely a case fits in a group Y, e.g. the chance that a case belongs to class 1, $P(Y = 1)$. Logistic regression has the advantage that it is less affected than discriminant analysis when the basic assumptions, particularly normality of the variables, are not

met. Discriminant analysis is another classical statistical technique that can be used for prediction or profiling. Discriminant analysis suits better in problems with three or more groups in the dependent variable [24]. Multiple linear regression is also similar to logistic regression. The difference with logistic regression is that multiple linear regression estimates a continuous value for a new observation [60]. During this research a binary value needs to be predicted, churn or non-churn, which make linear regression appropriate.

Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.058	0.062	0.063	0.063	**0.064**
AUC	0.602	0.717	0.735	0.733	**0.739**

TABLE 5.5: Performance parameters, AUK and AUC, of the Logistic Regression models.

Table 5.5 shows the performance parameters, with the five generated models which differ in training set distribution. Model 5, containing a training set of 50% churners and 50% non-churners performs best using logistic regression, with AUC and AUK values of 0.064 and 0.739, respectively. However, the differences in these values between Models 3-5 are not large. It was expected that the distributions with a higher churn rate would perform better [68].

LR	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.105	0.121	0.126	0.141	0.146
Threshold	0.1	0.4	0.5	0.5	0.7
Precision	0.114	**0.157**	0.131	0.141	0.147
Sensitivity	0.330	0.180	0.332	0.343	**0.335**

TABLE 5.6: Performance parameters, precision and sensitivity, of the Logistic Regression models.

Table 5.6 shows that model 5 also reached the highest sensitivity level, but the precision level is slightly lower compared to model 2. When the sensitivity level and precision levels are checked together, we can see that the sensitivity level of model 2 is very low compared to model 5. However the precision does not result in a big difference. This means that model 5 also with these performance measures, performs best.

5.2.3 Decision tree

The aim of the decision tree technique (DT) is to classify and label records and reduce dimensions [60]. It is often used because it is easy to interpret [31, 60]. The C4.5 algorithm developed by Ross Quilan is used to generate the decision tree [56]. At each node the most effective split is applied, the criteria of splitting is based on the information

gain. This information gain considers the relevance of an attribute. After creating the decision tree the Minimal Description Length pruning method is applied to identify the least reliable leafs of the tree. During the pruning process the least reliable leafs are replaced by nodes which were first split-nodes. With the C4.5 algorithm the pruning method is applied on the decision tree generated during the training set.

The models will be generated for different parameter settings, the minimum number of records per node will be 1% and 0.1% of the training set, 50 and 5 cases respectively. Also here the different training set distributions are applied. Tables 5.7 and 5.8 show the results of the five models with these two different parameter settings.

Minimum number of records set to 0.1%					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.032	0.051	**0.073**	0.058	0.052
AUC	0.520	0.564	0.595	0.638	0.682
Minimum number of records set to 1%					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.032	0.053	0.064	0.060	0.064
AUC	0.520	0.559	0.703	0.670	**0.718**

TABLE 5.7: Performance parameters, AUK and AUC, of the Decision Tree models.

When the minimum number of records is set to 0.1% the largest AUK value is generated with a training set distribution of 70:30. But the AUC value is rather low, and is much higher for model 5, with the minimum number of records set to 1%. However, the AUK value takes into account that the data set is imbalanced, which results in the conclusion that model 3 with a minimum of 0.1% records per node is the best model when the AUK and the AUC are used as performance measure.

DT 0.1%	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.064	0.122	0.151	0.117	0.089
Threshold	0.1	0.5	0.5	0.5	0.7
Precision	**0.245**	0.206	0.242	0.146	0.113
Sensitivity	0.049	0.127	0.150	0.210	**0.263**
DT 1%	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.064	0.118	0.145	0.135	0.124
Threshold	0.1	0.4	0.6	0.6	0.7
Precision	**0.245**	0.193	0.176	0.156	0.161
Sensitivity	0.049	0.131	0.221	0.247	0.191

TABLE 5.8: Performance parameters, precision and sensitivity, of the Decision Trees models.

When model 3 with a minimum of 0.1% records per node is investigated with the performance measures precision and sensitivity, it can be seen that the precision is the second highest for all models. The model with the highest precision is model 1 (with both

settings), but when we check the sensitivity, which is really low, it can be concluded that this model does not do very well. The sensitivity of model 3 with a minimum of 0.1% records per node is also low. However, all the model generated with the decision tree techniques do not show a high sensitivity level. It is concluded that also for these performance parameters model 3 with a minimum of 0.1% records per node is the best performing model of the decision tree technique.

5.2.4 Neural networks

Neural networks, also called artificial neural networks (ANN), are able to identify complex relationships within the data which is not possible with other classifiers [60]. Another advantage is that neural networks have a high tolerance to noisy data. Two disadvantages of this technique are, the difficult interpretation due to these complex relationships and the algorithm does not make a difference between the importance of predictors.

The RProp algorithm for multilayer feedforward networks is applied, with the possibility to change the parameters for the number of hidden layers and the number of hidden neurons per layer [57]. According to Heaton *et al.* the number of hidden layers is mostly chosen as one and the number of nodes per layer is mostly the mean of the input and output nodes [27]. With this as starting point, a prediction model with two and three hidden layers is also generated. The number of hidden neurons will also be set to the number of input nodes, and a sum of the input and output nodes based on the settings given in the other node descriptions for neural networks in KNIME. The variables are normalized prior to the model generation.

Number of hidden neurons set to 13					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.052	0.061	0.065	0.061	0.062
AUC	0.586	0.702	0.720	0.715	0.721
Number of hidden neurons set to 24					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.053	0.065	0.068	0.066	0.064
AUC	0.597	0.709	0.731	0.728	0.726
Number of hidden neurons set to 26					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.058	0.071	**0.069**	0.061	0.062
AUC	0.606	0.716	**0.732**	0.708	0.720

TABLE 5.9: Performance parameters, AUK and AUC, of the Neural Network models with one hidden layer.

From the first three models, shown in Table 5.9, we can conclude that the 70:30 training set distribution performs best (model 3). For all parameter settings, this distribution results in the best performance. The differences between the three models is small, but the model with 26 hidden neurons performs slightly better. The models with the original churning rate perform worst.

HN 13	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.103	0.125	0.157	0.129	0.132
Threshold	0.2	0.5	0.5	0.6	0.7
Precision	0.150	0.178	0.154	0.157	0.136
Sensitivity	0.146	0.159	0.368	0.209	0.347
HN 24	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.097	0.142	0.156	0.152	0.146
Threshold	0.1	0.5	0.6	0.6	0.8
Precision	0.110	**0.213**	0.167	0.185	0.172
Sensitivity	0.332	0.154	0.281	0.213	0.225
HN 26	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.124	0.170	0.171	0.139	0.149
Threshold	0.2	0.4	0.5	0.6	0.8
Precision	0.167	0.198	0.165	0.167	0.200
Sensitivity	0.174	0.237	**0.375**	0.213	0.181

TABLE 5.10: Performance parameters, precision and sensitivity, of the Neural Network models with one hidden layer. With the number of hidden neurons indicated with HN.

With the performance parameters precision and sensitivity can also be concluded that the models with the original churning rate perform worst. For the best performing model indicated by the AUK and AUC the sensitivity level is also the highest. However the precision of model 2 with 24 hidden neurons is higher, which means that less non-churning customers are labeled as churning compared to the other models. For this research it is more important to predict more churning customers, which is the minority class, than the purity of the prediction. Therefore, it is concluded that model 3 with 26 hidden neurons performs best, compared to all the models with one hidden layer.

For the models with two hidden layers, results shown in Table 5.11, the training set distribution of 70:30 (model 3) also performs best. The model with 13 hidden neurons resulted in the highest AUC and AUK. The model which predicts the worst with two hidden layers is the model with 13 neurons and the training set with the original churners versus non-churners rate, which was also the case with one hidden layer.

When the precision and the sensitivity are checked together (Table 5.12) none of the models generated with two hidden layers perform well on both parameters. For this reason the AUK value is leading which results in a 13 hidden neurons with two hidden layers model.

Number of hidden neurons set to 13					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.051	0.066	**0.069**	0.056	0.060
AUC	0.569	0.707	**0.732**	0.706	0.715
Number of hidden neurons set to 24					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.059	0.059	0.066	0.057	0.065
AUC	0.580	0.699	0.723	0.703	0.729
Number of hidden neurons set to 26					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.062	0.062	0.065	0.061	0.059
AUC	0.599	0.698	0.721	0.712	0.715

TABLE 5.11: Performance parameters, AUK and AUC, of the Neural Network models with two hidden layers.

HN 13	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.099	0.174	0.157	0.127	0.141
Threshold	0.1	0.5	0.5	0.5	0.8
Precision	0.123	**0.227**	0.155	0.131	0.172
Sensitivity	0.221	0.202	0.360	0.356	0.209
HN 24	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.111	0.120	0.154	0.116	0.157
Threshold	0.1	0.4	0.5	0.6	0.8
Precision	0.130	0.140	0.155	0.144	0.176
Sensitivity	0.245	0.237	0.336	0.202	0.252
HN 26	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.116	0.150	0.149	0.132	0.128
Threshold	0.1	0.4	0.5	0.6	0.8
Precision	0.126	0.163	0.147	0.163	0.147
Sensitivity	0.308	0.273	**0.490**	0.202	0.245

TABLE 5.12: Performance parameters, precision and sensitivity, of the Neural Network models with two hidden layers.

The results of the last models are shown in Table 5.13. The three hidden layer models also perform best with the 70:30 distribution (model 3), and perform worst with the original distribution. Model 5 with the 26 hidden neurons and three hidden layers result in the highest AUC and AUK value.

Also for the models with three hidden layers the precision and sensitivity are not very high. Especially when both parameters are checked together. Model 3 with 26 hidden neurons of Table 5.14 shows the second highest performance and a mean value for sensitivity. So it can be concluded that model 3 with 26 hidden neurons is the best performing model when the AUK, AUC, precision and sensitivity are considered.

There are two models which perform evenly well, based on the AUK and the AUC

Number of hidden neurons set to 13					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.057	0.056	0.060	0.060	0.061
AUC	0.587	0.689	0.712	0.714	0.712
Number of hidden neurons set to 24					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.050	0.063	0.056	0.063	0.062
AUC	0.571	0.696	0.704	0.718	0.724
Number of hidden neurons set to 26					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.054	0.064	**0.067**	0.057	0.062
AUC	0.584	0.701	**0.725**	0.710	0.724

TABLE 5.13: Performance parameters, AUK and AUC, of the Neural Network models with three hidden layers.

HN 13	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.109	0.119	0.130	0.131	0.127
Threshold	0.1	0.3	0.5	0.6	0.7
Precision	0.124	0.126	0.132	0.150	0.128
Sensitivity	0.280	0.356	0.380	0.245	0.399
HN 24	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.095	0.155	0.127	0.161	0.157
Threshold	0.1	0.4	0.5	0.6	0.8
Precision	0.116	0.164	0.129	**0.179**	0.167
Sensitivity	0.237	0.289	0.379	0.261	0.285
HN 26	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.102	0.124	0.171	0.121	0.118
Threshold	0.1	0.3	0.5	0.5	0.7
Precision	0.118	0.126	0.165	0.126	0.126
Sensitivity	0.277	**0.407**	0.371	0.356	0.383

TABLE 5.14: Performance parameters, precision and sensitivity, of the Neural Network models with three hidden layers.

values. Both models are trained with a 70:30 distribution, one model has one hidden layer and 26 hidden neurons and the other one has two hidden layers and 13 hidden neurons. However the model with one hidden layer and 26 hidden neurons performs better on precision and sensitivity level. So the model with one hidden layer and 26 hidden neurons bason on a 70:30 distribution is the best performing model generated by neural networks. Overall can we conclude that the 70:30 distribution performs best for neural networks. Interestingly, all the models which are trained with the original churners rate perform worst.

5.2.5 Support Vector Machines

Support vector machines (SVM) can be used for classification and regression analysis, in this research it is used for classification of churners and non-churners. SVM constructs hyperplanes in a multidimensional space to separate cases of different class labels, in this research churn and non-churning customers [11]. With the kernel functions a higher dimensional space is generated to rearrange the cases in the corresponding class.

The LibSVM algorithm is used because it runs faster than the SMO algorithm to build the SVM classifier, according to the node description of LibSVM in KNIME. A procedure introduced by Hsu *et al.* is adapted to select the parameter settings [29]. The kernel function that should be tried first is the radial basis function (RBF). With cross-validation the appropriate settings for the costs and the γ of the kernel function can be found. Unfortunately, cross-validation is not possible with two different data sets for training and testing. As starting point a $\gamma = \frac{1}{24}$ and C = 1 (costs) are used, based on the default in the node description. Different values for γ and C are applied to manually see what the best settings are. To generate a reference model the linear kernel function is used. The variables are normalized prior to the model calculation.

Costs parameter set to 0.8					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.053	0.049	0.046	0.060
AUC	0.500	0.561	0.592	0.598	0.707
Cost parameter set to 1.0					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.055	0.051	0.047	0.060
AUC	0.497	0.563	0.591	0.591	0.708
Cost parameter set to 1.5					
AUK	0.000	0.055	0.049	0.046	**0.061**
AUC	0.494	0.563	0.592	0.593	**0.711**

TABLE 5.15: Performance parameters, AUK and AUC, of the Support Vector Machine models with a γ of 0.04 for the RBF kernel function.

As shown in Tables 5.15 to 5.17, the 50:50 training distribution performs best with all settings. Another interesting finding is that the training set with the original churn rate performs even worse than random. Overall the linear kernel function performs slightly better with a costs rate of 0.8, however the differences are very small.

When the models generated with support vector machines are compared on precision and sensitivity the same results are found. The sensitivity of the models with the original data set is one or very close to one, but the precision is extremely low. The precision corresponds to the real churn rate stored in the data set. Due to confidentiality this

Costs parameter set to 0.8					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.055	0.053	0.049	**0.060**
AUC	0.500	0.564	0.601	0.610	0.711
Cost parameter set to 1.0					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.056	0.054	0.051	**0.060**
AUC	0.498	0.566	0.602	0.615	0.712
Cost parameter set to 1.5					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.054	0.053	0.056	**0.060**
AUC	0.500	0.564	0.603	0.628	**0.713**

TABLE 5.16: Performance parameters, AUK and AUC, of the Support Vector Machine models with a γ of 0.1 for the RBF kernel function.

Costs parameter set to 0.8					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.000	0.042	0.048	**0.062**
AUC	0.497	0.500	0.584	0.603	**0.718**
Cost parameter set to 1.0					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.00	0.035	0.041	0.047	**0.062**
AUC	0.500	0.597	0.584	0.591	0.716
Cost parameter set to 1.5					
Performance parameter	Model 1	Model 2	Model 3	Model 4	Model 5
AUK	0.000	0.023	0.042	0.047	**0.062**
AUC	0.499	0.575	0.584	0.598	0.716

TABLE 5.17: Performance parameters, AUK and UAC, of the Support Vector Machine models with a linear kernel function.

value is not included in the Tables 5.18 to 5.20. It can also be concluded that all cases are labeled as churners, which is an undesirable result.

The main conclusion from support vector machines is that the models with a 50:50 distribution perform best. The model that performs best is the model with a linear kernel and a cost rate of 0.8. Because the differences between these models are small, it can only be concluded that this model performs best for the sample set used for this analysis.

5.2.6 Selection of the model

In the literature study for this research, it is found that most papers use the decision tree technique (Chapter 4). In the literature reviews it is shown that neural networks are used the most to predict customer churn [39, 65]. What we can see from the results

C 0.8	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.104	0.089	0.098	0.137
Threshold	0.9	0.2	0.3	0.6	0.7
Precision	-	0.136	0.106	0.139	0.152
Sensitivity	1.000	0.190	0.336	0.156	0.271
C 1.0	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.108	0.101	0.112	0.138
Threshold	0.9	0.2	0.5	0.6	0.7
Precision	-	0.140	0.216	0.170	0.153
Sensitivity	0.992	0.191	0.092	0.141	0.271
C 1.5	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.108	0.092	0.118	0.140
Threshold	0.9	0.2	0.5	0.6	0.7
Precision	-	0.141	0.127	0.185	0.153
Sensitivity	0.985	0.187	0.176	0.137	0.282

TABLE 5.18: Performance parameters, precision and sensitivity, of the Support Vector Machine models with a γ of 0.04 for the RBF kernel function.

C 0.8	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.107	0.112	0.116	0.132
Threshold	0.9	0.2	0.6	0.6	0.7
Precision	-	0.139	0.180	0.174	0.146
Sensitivity	1.000	0.195	0.130	0.145	0.279
C 1.0	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.109	0.107	0.106	0.133
Threshold	0.9	0.2	0.6	0.6	0.7
Precision	-	0.139	0.183	0.165	0.147
Sensitivity	0.996	0.198	0.118	0.134	0.279
C 1.5	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.105	0.117	0.112	0.131
Threshold	0.9	0.2	0.5	0.6	0.7
Precision	-	0.136	0.157	0.170	0.145
Sensitivity	1.000	0.195	0.176	0.141	0.279

TABLE 5.19: Performance parameters, precision and sensitivity, of the Support Vector Machine models with a γ of 0.1 for the RBF kernel function.

in this research is that logistic regression, neural networks and support vector machines perform well when all performance parameters are considered. However, when only the AUK value is taken into account, the decision tree technique performs best, but this technique has low AUC and sensitivity levels. This implies that it is important to make an informed decision with multiple performance parameters. A visual inspection of the graphs in Figure 5.4 show that only for logistic regression and the decision tree technique the maximum of the ROC-curve (point most oriented to the upper left corner) and the Cohen's Kappa curve are the same. For neural networks and support vector machines is this not the case, and is the corresponding threshold value different. Due to time issues

C 0.8	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.000	0.080	0.084	0.135
Threshold	0.9	0.1	0.4	0.3	0.7
Precision	-	-	0.114	0.101	0.149
Sensitivity	0.992	1.000	0.179	0.397	0.282
C 1.0	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.059	0.081	0.081	0.137
Threshold	0.9	0.3	0.5	0.5	0.7
Precision	-	0.152	0.114	0.169	0.150
Sensitivity	1.000	0.061	0.179	0.179	0.286
C 1.5	Model 1	Model 2	Model 3	Model 4	Model 5
Kappa	0.000	0.047	0.083	0.092	0.131
Threshold	0.9	0.2	0.4	0.6	0.7
Precision	-	0.078	0.116	0.156	0.146
Sensitivity	0.996	0.435	0.179	0.114	0.275

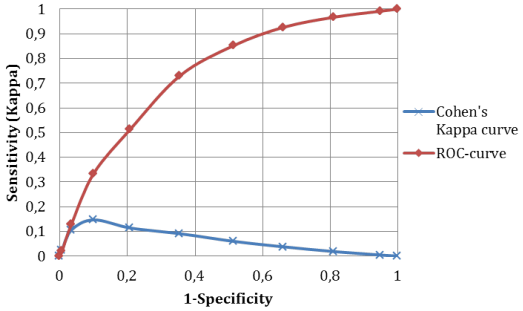
TABLE 5.20: Performance parameters, precision and sensitivity, of the Support Vector Machine models with a linear kernel function.

the point closest to the upper left corner could not exactly be determined and is not known what the best threshold value would be if the ROC-curve is used. It can only be concluded due to visual inspection the optimal threshold value sometimes is equal between the two measurements and sometimes it is not. This supports the conclusion that it is important to use multiple performance measures when an imbalanced data set is used.

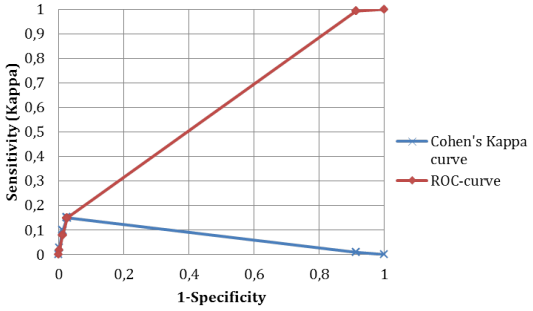
For more insight in how well the models predict for the marketing department of CZ, lift charts will be compared in Section 6.1 for each technique with the best performing model. Another reason why the lift charts will give more insight is because the marketing department of CZ cannot contact the full customer database. As a result they would like to know how many churning customers are reached when they contact a subset of the full population. The robustness of the models is tested with a test set from 2014 (Section 6.2) and if models generate beneficial results is analysed with a cost-benefit analysis (Section 6.3).

According to Visa & Ralescu neither a 50:50 training set, nor the original training set would perform best [68]. This is what we have found in this research as well for decision tree and neural networks. However during this research, logistic regression performs best with a 50:50 training set distribution.

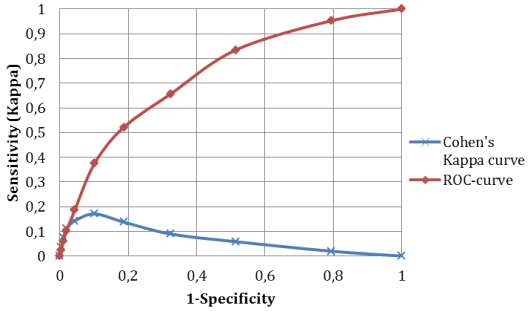
The profiles which are found with the profiling algorithms are used in Section 6.4 to see if this results in better prediction models. This is tested on the best performing models of decision tree, neural networks and logistic regression. According to the research



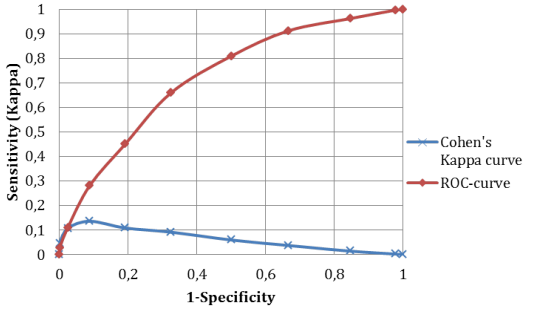
(A) Graph representing the AUK and AUC curves for Logistic Regression with a 50:50 training set distribution.



(B) Graph representing the AUK and AUC curves for Decision Tree with a 70:30 training set distribution.



(C) Graph representing the AUK and AUC curves for Neural Networks with a 70:30 training set distribution.



(D) Graph representing the AUK and AUC for Support Vector Machines with a 50:50 training set distribution.

FIGURE 5.4: Comparison between the maximum point in the ROC curve and Cohen's Kappa Curve. The best performing model is used per technique.

performed by Ng & Liu and Ng *et al.* this would generate better results, because of the homogeneous profiles [49, 50].

Chapter 6

Interpretation of churn prediction models

For a more in-depth analysis the best performing models of Chapter 5 are used. In Section 6.1 the results generated with the models of Chapter 5 are compared using lift charts. Section 6.2 tests if the prediction techniques perform evenly well on data of different years. From these Sections the two best performing models are compared with a cost-benefit analysis (Section 6.3). Additionally, in Section 6.4 K-means clustering algorithm is applied to create homogeneous profiles before a prediction model is generated.

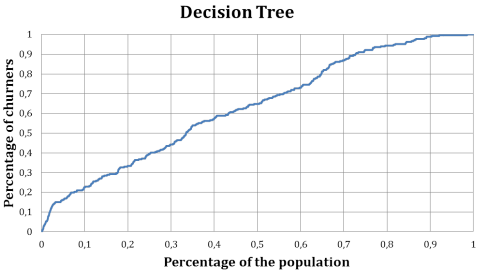
6.1 Analysis of the results for the marketing department of CZ

For each technique the best performing model is selected to create lift charts which give insights in how many churners are reached when a sub-set of the population is contacted. From the models, customers are assigned a churning possibility. The population is sorted, from high to low, based on that churning possibility. A lift chart is then created by plotting the actual churn percentage in a sub-set of the sorted population against the percentage of that sub-set of the entire sorted population. When 100% of the sorted population is used, 100% of the churners are reached. For the random case, the chance of encountering a churner is 50% and therefore the lift chart simply follows $y = x$. Figures 6.1a to 6.1d represent the lift charts of the models discussed in Chapter 5.

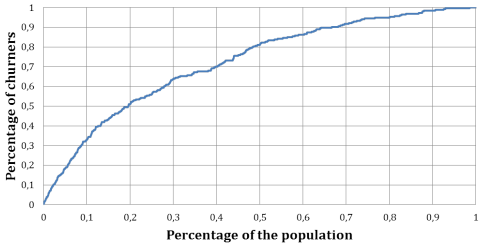
The lift charts of logistic regression and neural networks show the best performance. Approximately 50% of the churners can be reached by contacting 20% of the population.

Support vector machines also perform well, with 20% of the population approximately 45% of the churners is reached. With decision tree less than 35% of the churners are reached, which makes decision tree the least interesting technique.

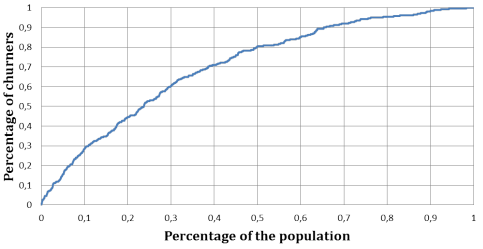
(A) Lift chart for logistic regression model 5.



(B) Lift chart for decision tree model 3 with a minimum of 0.1% records per node.



(C) Lift chart for neural networks model 3 with one hidden layer and 26 hidden neurons.



(D) Lift chart for support vector machine model 5 with linear kernel and C = 0.8.

FIGURE 6.1: Lift chart for the best performing models per technique.

For the marketing department of CZ the logistic regression technique is most interesting. This is because it predicts very well and for each variable is known what the influence is on the prediction. The exact influence of the ten most influential variables is given in Equation 6.1. Table 6.5 represents the corresponding variables per predictor α.

$$f(\alpha) = \frac{1}{1 + e^{-(\beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \cdots + \beta_{10}\alpha_{10})}} \tag{6.1}$$

Equation 6.1 and Table 6.1 shows that the number of declarations, number of times a customer is insured, age and the number of contact moments result in the most influencing variables. It already was expected that the number of declarations, number of times insured and age have high influence on the prediction. This is because during the profiling in Section 5.1 these variables also gave more insight in a churning and non-churning profile. During the profiling the consumption of a customer was important which can be linked to the number of declarations.

α	β	Variable
α_1	-4.06	Number of declarations
α_2	3.88	Number of times insured
α_3	-3.61	Age
α_4	3.21	Contact moments
α_5	2.95	Discount
α_6	-2.18	Premium
α_7	-1.26	Duration of contract
α_8	-0.99	Payment regulations
α_9	0.90	Contribution
α_{10}	0.41	Urbanity

TABLE 6.1: The ten variables with the most influencing variables from logistic regression are linked to the corresponding predictor α.

6.2 Model created for 2013 tested on the data of 2014

In Section 6.1 the lift charts of the models are reviewed, which resulted in more insight in the prediction quality of the models. For a better understanding of how well the models will predict in a different setting, data from 2014, instead of 2013, is used for the test set. The same performance parameters are used to test which technique performs best.

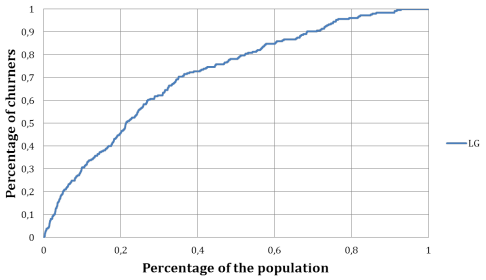
The worst performing technique is decision tree, with this technique all cases are predicted with the same churn possibility. Resulting in the bad performance parameters displayed in Tables 6.2 also no lift chart is represented due to the bad performances.

Performance parameter	LR	DT	NN	SVM
AUK	0.063	0	0.057	0.067
AUC	0.719	0.5	0.692	0.731
Kappa	0.129	0	0.118	0.162
Threshold	0.9	0.1	0.5	0.7
Precision	0.202	-	0.171	0.186
Sensitivity	0.140	1	0.159	0.241

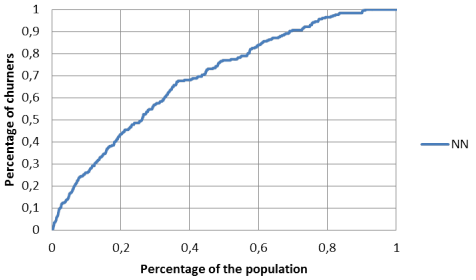
TABLE 6.2: Performance parameters for the models tested on the data from 2014. Precision is not included for DT because of confidentiality.

Support vector machine and logistic regression perform best on the test set of 2014, as can be seen in the performance parameters of Table 6.2. The AUK value of support vector machine is even higher than the generated results with a test set generated with data from 2013. When the values of Table 6.2 are compared between each other, none of the techniques show better results.

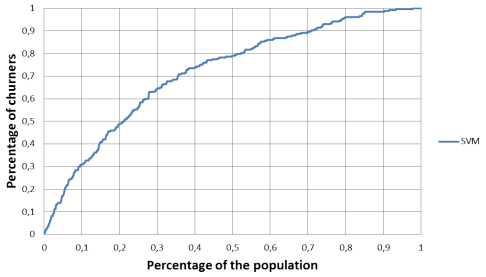
The results of the lift charts in Figure 6.2 show that support vector machine, neural networks and logistic regression perform well. With all techniques would approximately



(A) Lift chart for logistic regression model 5.



(B) Lift chart for neural networks model 3 with one hidden layer and 26 hidden neurons.



(C) Lift chart for support vector machine model 5 with linear kernel and C = 0.8.

FIGURE 6.2: Lift charts for the best performing models per technique.

50% of the churners be reached when 20% of the population is contacted. These results are comparable with the results from Section 6.1.

We can conclude that logistic regression and neural networks are the most interesting technique to predict customer churn. Both techniques perform well on test data from 2013 and 2014 which means that the generated models are applicable on data of multiple years.

6.3 Cost-benefit analysis applied on different models

To see which technique results in the highest benefits a cost-benefit analysis (CBA) is executed on test data from 2013. The CBA is applied on the best models generated with logistic regression and neural networks.

The costs and benefits [55] are divided in the the four categories of the confusion matrix which is discussed in Section 5.2.1. For this CBA it is assumed that all true positive predictions will not churn after interference. The exact benefits in this CBA are difficult to estimated, therefore the following benefits are assumed:

- A *false positive* occurs when the model assigns a churning label to a customer, but the customer was not going to churn. The preparation and contacting costs for

the marketing actions is set to €10. The benefit in this case is negative: FP = -10.

- A *false negative* is a customer who was predicted to not churn, but will churn without interference. In this case, no money is spent or gained. The benefit in this case is zero: FN = 0.
- A *true positive* is a customer who was predicted to churn and was going to churn when it was not contacted. The benefit for this customer is set to the average benefit of a customer calculated over three years (Table 6.3). The benefit in this case is positive: TP = 101.
- A *true negative* is a customer who is predicted as a non-churner and was not going to churn. There is no benefit nor costs, which results in a benefit of zero: TN = 0.

The marketing costs of CZ to attract new customers were €18 per customer in 2013 [18]. According to Ng and Liu these costs are three to five times higher than when customers are retained [49]. This would result in a cost between €3.60 and €6 for retaining customers. In this research is chosen for marketing costs of €10 because this would be the first time that CZ would apply this marketing strategy (contacting customers with a high possibility to churn).

	2014	2013	2012	Average
Benefits in millions	€315	€191	€518	€341
Average benefit per customer	€91.95	€56.15	€154.68	€100.93

TABLE 6.3: Average benefit of CZ calculated over three years [17].

In Figure 6.3 the CBA is shown for the models of logistic regression and neural networks. On the x-axis the different threshold values are represented, and on the y-axis the benefits. When all the customers are contacted, in this research 5,000 customers are selected in the test set, neural networks with a threshold value of 0.5 generates the highest benefits. All customers with a churning possibility of 0.5 or higher are contacted and results in a benefit of €4,785.

When the predicted churning possibilities are ordered from high to low, and 20% of the customers with the highest churning possibility are contacted no costs are made (Figure 6.4). When the benefits are used to select the best performing model, the neural networks would be selected. This model generates a benefit of €4,319, with only 5,000 cases in the test set. By adding more cases the graphs will be spreading and show a higher benefit level.

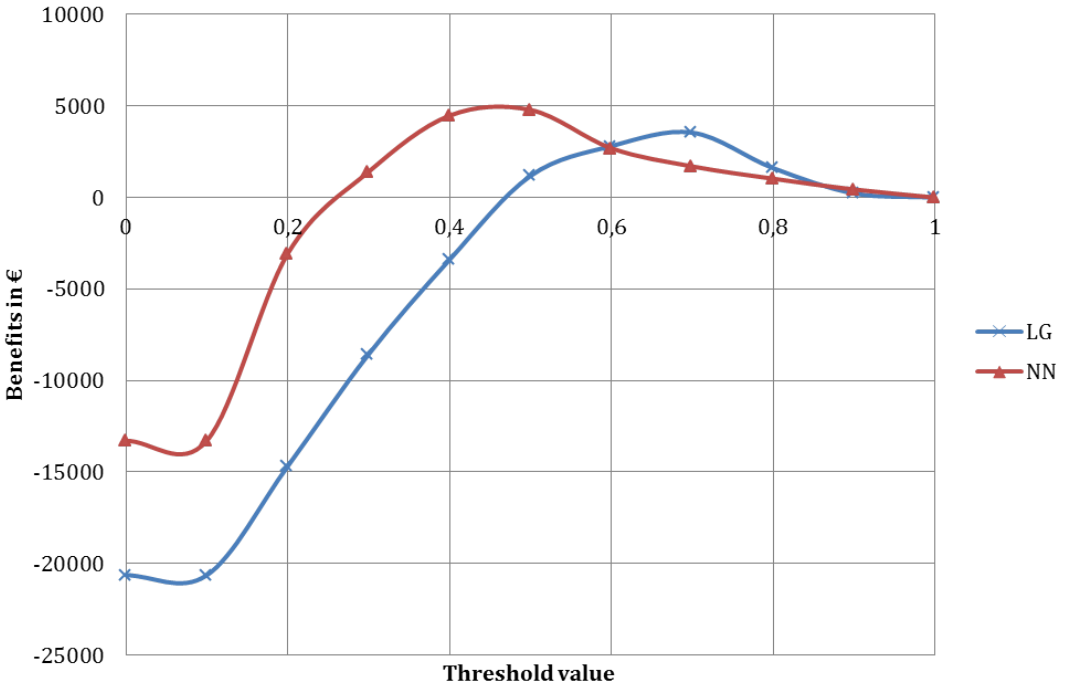


FIGURE 6.3: Cost-benefit analysis over all the cases in the data set.

Overal we can conclude that the generated models can result in benefits for the company. This CBA shows that it will be beneficial when 20% of the customers with the highest churn possibility are contacted.

6.4 Model generation on homogeneous profiles

To see if even better results could be generated, the homogeneous profiles are used to create the churn prediction models. Lin *et al.* found promising results when a clustering algorithm was used before the real prediction model was generated [45]. The new models are generated with the settings of the best performing models of logistic regression and neural networks. The generated models are compared on the performance parameters discussed in Section 5.2.1.

The K-means clustering technique is used to create the training and test sets. The number of clusters is set to 9 and 10, and the Euclidean distance and the Manhattan distance are used because these settings showed the best results in Section 5.1.1. From the results the clusters with a churning profile, marked as such in Section 5.1.1, are selected. These clusters contain young customers, who consume less healthcare than average and pay the premium.

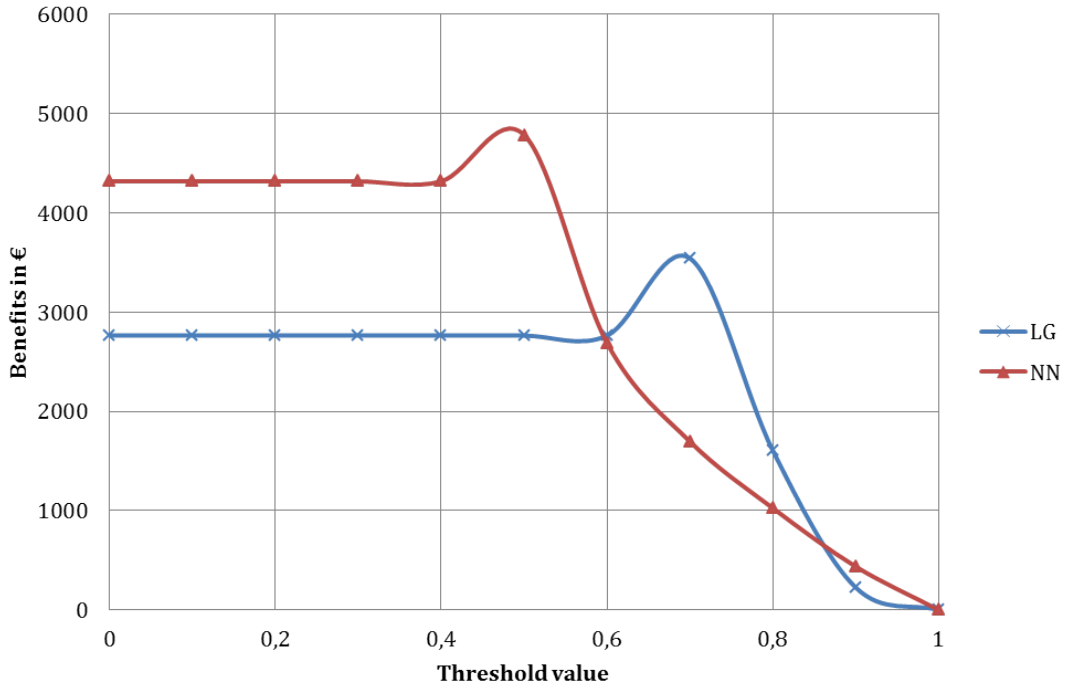


FIGURE 6.4: Cost-benefit analysis over 20% of the cases which have the highest prediction rate in the data set.

Performance parameter	LR E9	LR E10	LR M9	LR M10
AUK	0.055	0.018	0.066	-
AUC	0.577	0.525	0.618	-
Performance parameter	NN E9	NN E10	NN M9	NN M10
AUK	0.016	0.034	**0.096**	0.070
AUC	0.512	0.533	**0.629**	0.592

TABLE 6.4: Performance parameters, AUK and AUC, of the Logistic Regression (LR) and neural network (NN) models. With different distance measurements (Euclidean distance (E) and Manhattan distance (M)) and different values for K ($K = 9$ and $K = 10$)

The results of Table 6.4 show that the Manhattan distance performs best for all models according to the performance parameters AUK and AUC. Especially the model generated with neural networks and $K = 9$. From these nine clusters only the clusters which fit in the churning profile are selected. This means that there were 2 clusters selected resulting in 433 cases. Unfortunately, the logistic regression did not run for the clusters selected with the Manhattan distance with 10 clusters. The exact reason for this is unclear, but it is most likely that it did not run due to insufficient information. A reason for this can be that the sample size was too small. Table 6.5 shows that the logistic regression model with Euclidean distance and 10 clusters generates the best results however, the differences are small. Also here the clusters with an average churning profile are selected resulting in 159 cases.

Performance parameter	LR E9	LR E10	LR M9	LR M10
Kappa	0.147	0.090	0.179	-
Threshold	0.3	0.4	0.5	-
Precision	0.240	0.211	0.444	-
Sensitivity	0.300	0.333	0.133	-
Performance parameter	NN E9	NN E10	NN M9	NN M10
Kappa	0.028	0.073	0.191	0.201
Threshold	0.8	0.1	0.1	0.2
Precision	0.200	0.207	0.157	0.258
Sensitivity	0.150	0.250	0.276	0.258

TABLE 6.5: Performance parameters, precision and sensitivity, of the Decision Tree models.

The model generated with neural networks, Manhattan distance and $K = 9$ performs best of all models. The AUK, AUC and the sensitivity is the highest of all models. Table 6.6 shows the performance parameters of the best performing models of Chapter 5. It can be concluded that the generated model with homogeneous clusters shows promising results. The AUK value is even higher than the AUK value of the model of Chapter 5. The same conclusions can be made for the logistic regression model with the Euclidean distance and $K = 9$. When the models are generated on larger homogeneous clusters the real benefit could be determined. Further research should investigate the differences between model performance of homogeneous clusters and with the correction of imbalanced data.

Performance parameter	LR	NN
AUK	0.064	0.069
AUC	0.739	0.732
Kappa	0.146	0.171
Threshold	0.7	0.5
Precision	0.147	0.165
Sensitivity	0.335	0.375

TABLE 6.6: Results of the best performing models discussed in Chapter 5.

Chapter 7

Conclusions and recommendations

A major problem for every company is a high number of churning customers. This is hard for health insurance companies in the Netherlands because of the dynamic and competitive environment. The indicators for customers switching to a competitor are unclear, however it is known that churning customers use less health which makes them an interesting group to focus on. In this research, a case study was conducted at CZ, one of the four major health insurance companies in the Netherlands. Using data mining techniques, predictions models were constructed to categorize and predict customer churn. In this last Chapter the main research question and sub-questions will be answered. Furthermore, the company recommendations, generalisation, research limitations and issues for further research are discussed (Sections 7.2 to 7.5).

7.1 Revisiting the research questions

To answer the main research question four sub-questions were defined of which the answers will be discussed here. This Section will conclude with the answer on the main research question.

Sub-research question 1

Which customer characteristics and behavior aspects are key to predict customer churn behavior?

Customer characteristics and behavior aspects are mentioned by experts and found in the literature which would influence churning behavior (Chapter 2). In this chapter a high similarity is found between the variables mentioned by experts and the found

variables in the literature. The variables can be divided in socio-demographic variables, customer/company-interaction variables and product related variables. Furthermore, there is a difference made between time-variant and time-invariant variables. Time-invariant variables are variables which do not change during the year and time-variant variables represent actions which take place during the year. Table 7.1 gives six examples of time-variant and time-variant variables with respect to the group they are divided in.

Variable type	Time-variant	Time-invariant
Socio-demographic variables		
Age	×	
Gender	×	
Customer/company-interaction variables		
Duration of contract	×	
Number of contact moments		×
Product related variables		
Premium	×	
Consumption		×

TABLE 7.1: Sub-set of the variables for the indication of the variable types.

Chapter 3 answers the question, which variables are different for churning and non-churning customers. This results in the overal conclusion that churning profiles significantly differ form non-churning profiles. The main differences can be found in the variables age, number of times insured, urbanity, group insurance and additional insurances.

Sub-research question 2

Which techniques can be used to generate the best churn prediction models?

With a literature review it is investigated which techniques would be best suitable for churn prediction. Table 7.2 shows the four most found techniques applied to predict customer churn. These four techniques are applied to generate the best results. Not only it was found that these techniques were the best performing, also research were using profiling techniques for the generation of homogeneous clusters. These were used in combination with the prediction technique to generate more accurate results. According to the literature the SOM and K-means profiling techniques were most applied for the creation of homogeneous clusters.

The techniques that are used for churn prediction all try to generate a high accuracy level. In this research the churning customers are the minority group in the data set, which results in a prediction that none of the customers is going to churn and an accuracy level of 90% or higher is still reached. For this reason the training sets are corrected, and five different churning:non-churning distributions are used (original, 80:20, 70:30 66:33 and 50:50).

Data mining technique	Literature review by KhakAbi *et al.* [39]	Literature review by Tsai & Lu [65]	Literature study for this research
Prediction techniques			
Neural Networks	15	10	4
Decision Tree	13	9	9
Logistic Regression	13	8	2
Support Vector Machine	7	2	0
Profiling techniques			
SOM	2	1	2
K-means	1	1	3

TABLE 7.2: Used data mining techniques split up per category compared with the churn prediction methods of two literature reviews.

Sub-research question 3

Which customer profiles should be analysed separately and what is the difference between the profiles?

Two profiling techniques are used to identify homogeneous profiles, K-means and SOM. This is done because the literature indicated that this results in better prediction models. Section 5.1 resulted in more insight in the profiles which could be predicted and clustered in homogeneous groups. The most important findings are found with the K-means clustering technique, which is that there are four types of profiles which should be analysed separately. In the list given below the first profile represents the average profile of the population, the second and third profile represent non-churning customers and the last profile indicates a churning profile.

- Profiles which are comparable to the average of the population.
- Older customers, who have no voluntary deductible excess and consume more health insurance than average.
- Young customers which do not pay the premium themselves and have a group insurance.
- Young customers, who consume less health insurance than average and pay the premium themselves.

The profile which represents the churning customers is used in this research to generate better prediction models. The results found with these models were very promising however, the sample size was small. For a determination for the real benefits a larger sample size is needed.

Sub-research question 4

Which model generates the best comparing on accuracy and interpretability?

The accuracy of the models is compared on AUK, AUC, precision and sensitivity, the logistic regression and neural networks technique show the best performance. Lift charts are used for more insights and show that for both techniques approximately 50% of the churners is reached when 20% of the population is contacted (Section 6.1). To investigate the robustness of the models a test set with data from 2014 is used (Section 6.2). The results from this test set are comparable to the results of the models tested on data from 2013. The models of 2013 show slightly better results a reason for this can be that the declarations were not fully collected yet. The benefits of these techniques are calculated with a cost-benefit analysis for the year 2013 (Section 6.3), which resulted in the highest benefit for the neural network (€4,785).

When the marketing department only would check what would result in the highest benefits, the neural networks model 3 with one hidden layer and 26 hidden neurons was the best choice. However, for the marketing department of CZ the logistic regression technique is most interesting. This is because it predicts very well and for each variable it is known what the influence is on the prediction. This gives more insights which results in a better strategy to make sure that a customer is not going to churn. An important insight that is given by the logistic regression model is that the churning possibilities of customers becomes higher when they contact CZ. The focus already is on a good customer care during contact moment, but further research could investigate which contact moments influence the churning possibilities.

Main research question

What are the possibilities to create highly accurate prediction models, which calculate if a customer is going to churn and provide insight in the reason why customers churn?

To create a highly accurate prediction model, the imbalanced data set problem is solved. For each technique a different training set distribution is best. During this research the 50:50 training set distribution resulted in the best results when the logistic regression technique is used (Section 5.2.2). A 70:30 distribution worked best for the neural network technique (Section 5.2.4). From this can be concluded that each technique works best with a different training set distribution. However, the overall conclusion which can be made is that the original distribution always performs worst.

The neural network and logistic regression model are tested on different levels to see which model would work for the marketing department of CZ (Chapter 6). The results are investigated with lift chart, cost-benefit analysis and the models are tested on data of 2014. The models performed almost evenly well, but only the logistic regression model provides insights in the variables which are important to predict customer churn. For this reason it can be concluded that the logistic regression technique worked best for this problem.

Promising results are found when the models are created for homogeneous churning profiles. However, more research is needed to see if this results in a higher benefit than when the imbalanced data set problem is solved by combining more churning cases. The main conclusion of this research is that it is possible to generate prediction models for customer churn at CZ with good prediction characteristics. This results in a fulfillment of the project goal.

7.2 Recommendations for the company

For CZ it is important to win as broad a support as possible. The profiles of churning customers show that they are not using much healthcare which makes them an interesting group. To win their support it would be good to tell this group what CZ is doing for the premium they are paying. The focus is now mainly on what CZ does for health consuming customers.

Another area they could focus on is, why the churning possibility rises when customers contact more often. It should be investigated if there are chances for CZ and why are these churning customers contacting. An approach can be, when a customer with a high churning possibility contacts a different assistance approach will be applied. If CZ is considering contacting these customers themselves the sleeper effect should be considered. This effect means that customers renew their contract without any actions taken. When these customers are woken up by CZ their behavior could chance.

7.3 Generalisation of the prediction model

As discussed in Section 2.3 the variables that are used to predict customer churn for a health insurance company are applied in a wide range of areas. They are used from research in the telecommunication market to the bank industry. Six variables are used, are typically applicable for the insurance market in the Netherlands. Because these variables are based on the Dutch health insurance market, the generalisation is only applicable to this market.

When other health insurance companies of the Netherlands want to use this model, they should investigate what the differences are in population between them and CZ. CZ is a company which mostly represents customers in the south of the Netherlands, for which should be corrected. But the main predicting variables are not linked to the location where a customer is living. In the logistic regression model generated in this research the nine most influencing variables are not based on location. From this can be concluded

that it is most likely that other health insurance companies in the Netherlands can use the model.

7.4 Limitations of the research

Most limitations are a result from the selection and collection phase of this project. The choices made during the data selection are done based on available knowledge. However when the internal process were known in more detail more informed decisions could be made. An other issue during the data collection is that there was no second check on the extraction method from SAS Enterprise Guide.

Limitation that is applicable for the used techniques is that with a more in depth analysis of one specific technique the best settings could be found. The data sets used in this research are not corrected for the higher representation of the southern provinces in the Netherlands and it is unclear what the exact effect is on the prediction models.

7.5 Issues for further research

By answering the research questions, new research questions arise. For the homogeneous profiles found with the K-means clustering method more cases should be extracted to see what the real benefits are. Also it would be interesting to know if homogeneous clusters predict better than a data set that is corrected for the minority group.

For a proper analysis of the models created with imbalanced data, the performance parameters (AUK, AUC, precision and sensitivity) should be investigated in more depth. In this research a start is made but it would be interesting to know how these parameters perform on other data sets.

In this research all variables are used to predict the possibility of customer churn. In further research it should be investigated if the models would perform better when only the most influencing variables are used.

Bibliography

[1] ACM Digital Library. http://librarians.acm.org/digital-library, 2014.

[2] Yas A. Alsultanny. Database Preprocessing and Comparison between Data Mining Methods. *International Journal of New Computer Architectures and their Applications (IJNCAA)*, 1(1):61–73, 2011.

[3] Wai-Ho Au, Keith C.C. Chan, and Xin Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7(6):532–545, December 2003.

[4] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[5] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer US, Boston, MA, 2010.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and Philip W Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[7] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[8] Mu-Chen Chen, Ai-Lun Chiu, and Hsu-Hwa Chang. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4):773–781, May 2005.

[9] Bong-Horng Chu, Ming-Shian Tsai, and Cheng-Seen Ho. Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, 20(8):703–718, December 2007.

[10] Jacob A. Cohen. A coefficient of agreeent for nomial scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[11] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[12] Kristof Coussement, Dries F. Benoit, and Dirk Van den Poel. Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, 37(3):2132–2143, March 2010.

[13] Kristof Coussement and Dirk Van den Poel. 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, January 2008.

[14] Kristof Coussement and Dirk Van den Poel. Integrating the voice of customers through call center emails into a decision support system for churn prediction, April 2008.

[15] CZ. http://www.cz.nl/english/health-insurance. November 2014.

[16] CZ. Maatschappelijk Verslag - CZ groep 2013. Technical report, 2014.

[17] CZ. Maatschappelijk verslag Kerncijfertabel 2014. Technical report, 2014.

[18] CZ. http://www.cz.nl/over-cz/nieuws/2013/wat-besteedt-cz-aan-marketing, 2015.

[19] Mohammed Abdul Haque Farquad, Vadlamani Ravi, and Surampudi Bapi Raju. Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*, 19:31–40, June 2014.

[20] Andy Field. *Discovering statistics using PSS*. SAGE Publications Inc., 3 edition, 2009.

[21] Andy Field. *Discovering Statistics using IBM SPSS Statistics*. SAGE Publications Ltd, fourth edition, 2013.

[22] Clara-Cecilie Günther, Ingunn Fride Tvete, Kjersti Aas, Geir Inge Sandnes, and Ø rnulf Borgan. Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1):58–71, February 2014.

[23] Özden Gür Ali and Umut Arıtürk. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, 41(17):7889–7903, December 2014.

[24] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate Data Analysis: A Global Perspective*. Pearson Education, 7 edition, 2010.

[25] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982.

[26] Yue He, Zhenglin He, and Dan Zhang. A Study on Prediction of Customer Churn in Fixed Communication Network Based on Data Mining. *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, 1:92–94, 2009.

[27] Jeff Heaton. *Introduction to neural networks in java.* Heaton Research, Inc, 2 edition, 2008.

[28] Wu Heng-liang, Zhang Wei-wei, and Zhang Yuan-yuan. An Empirical Study of Customer Churn in E-Commerce Based on Data Mining. In *2010 International Conference on Management and Service Science*, pages 1–4. IEEE, August 2010.

[29] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A Practical Guide to Support Vector Classification. *Technical Report, Department of Computer Science National Taiwan University*, (1):1–16, 2010.

[30] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, January 2012.

[31] Chung-Fah Huang and Sung-Lin Hsueh. Customer behavior and decision making in the refurbishment industry-a data mining approach. *Journal of Civil Engineering and Management*, 16(1):75–84, January 2010.

[32] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, October 2006.

[33] IEEE Xplore. http://www.ieee.org/publications_standards/publications/ subscriptions/publication_types.html, 2014.

[34] Independer. http://www.independer.nl/zorgverzekering/info/marktcijfers/aantal-zorgverzekeraars.aspx. November 2014.

[35] Ingeta Connect. http://www.ingentaconnect.com/about/researchermenu, 2014.

[36] Zack Jourdan, R. Kelly Rainer, and Thomas E. Marshall. Business Intelligence: An Analysis of the Literature 1. *Information Systems Management*, 25(2):121–131, March 2008.

[37] Uzay Kaymak, Arie Ben-David, and Rob Potharst. The AUK: A simple alternative to the AUC. *Engineering Applications of Artificial Intelligence*, 25(5):1082–1089, August 2012.

[38] Abbas Keramati, Rouhollah Jafari-Marandi, Mohammed Aliannejadi, Iman Ahmadian, Mahdieh Mozaffari, and Uldoz Abbasi. Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24:994–1012, November 2014.

[39] Sahand Khakabi, Mohammad R. Gholamian, and Morteza Namvar. Data Mining Applications in Customer Churn Management. *2010 International Conference on Intelligent Systems, Modelling and Simulation*, pages 220–225, January 2010.

[40] Kyoungok Kim, Chi-Hyuk Jun, and Jaewook Lee. Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, 41(15):6575–6584, November 2014.

[41] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. Handling imbalanced datasets : A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[42] R. J. Kuo, L. M. Ho, and C. M. Hu. Integration of self-organizing feature map and K -means algorithm for market segmentation. *Computer & Operations Research*, 29:1475 – 1493, 2002.

[43] Steve Lawrence and C. Lee Giles. Searching the world wide Web. *Science (New York, N.Y.)*, 280(5360):98–100, April 1998.

[44] Dirk Lewandowski. The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, 64(6):915–937, October 2008.

[45] Wei-Chao Lin, Chih-Fong Tsai, and Shih-Wen Ke. Dimensionality and data reduction in telecom churn prediction. *Kybernetes*, 43(5):737–749, April 2014.

[46] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, March 2008.

[47] Niyoosha Jafari Momtaz, Somayeh Alizadeh, and Mahyar Sharif Vaghefi. A new model for assessment fast food customer behavior case study: An Iranian fast-food restaurant. *British Food Journal*, 115(4):601–613, 2013.

[48] Michael C Mozer, Richard Wolniewicz, David B Grimes, Student Member, and Eric Johnson. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Transactions on Neural Networks*, 11(3):690–696, 2000.

[49] Kiansing Ng and Huan Liu. Customer Retention via Data Mining. *Artificial Intelligence Review*, 14(6):569–590, 2000.

[50] Kiansing Ng, Huan Liu, and HweeBong Kwah. A Data Mining Application: Customer Retention at the Port of Singapore Authority (PSA). *ACM SIGMOD Recod*, 27(2):522–525, 1998.

[51] Eric W.T. Ngai, Li Xiu, and Dorothy C.K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2):2592–2602, March 2009.

[52] Guangli Nie, Lingling Zhang, Xingsen Li, and Yong Shi. The Analysis on the Customers Churn of Charge Email Based on Data Mining Take One Internet Company for Example. In *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, pages 843–847. IEEE, 2006.

[53] NZa. Zorgverzekeringsmarkt 2014 - Weergave van de markt 2010-2014. Technical report, Nederlandse Zorgautoriteit, 2014.

[54] Adrian Payne and Pennie Frow. A Strategic Framework for Customer Relationship Management. *Journal of marketing*, 69(4):167–176, 2005.

[55] Foster Provost and Tom Fawcatt. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc., 2013.

[56] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc, 1993.

[57] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *IEEE International Conference on Neural Networks*, 16:586–591, 1993.

[58] Hans Risselada, Peter C. Verhoef, and Tammo H.A. Bijmolt. Staying Power of Churn Prediction Models. *Journal of Interactive Marketing*, 24(3):198–208, August 2010.

[59] Science Direct. http://www.sciencedirect.com, 2014.

[60] Galit Shmeuli, Nitin R. Patel, and Peter C. Bruce. *Data mining for business intelligence: concepts, techniques, and applications in microsoft office excel with xlminer.* John Wiley and Sons, second edition, 2011.

[61] Kate A. Smith, Robert J. Willis, and M Brooks. An analysis of customer retention and insurance claim patterns using data mining : a case study. *Journal of the Operational Research Society*, 51(5):532–541, 2000.

[62] Hee Seok Song, Jae Kyeong Kim, and Soung Hie Kim. Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, 21(3):157–168, October 2001.

[63] Springer. http://www.springer.com/gp/about-springer/company-information/what-we-do, 2014.

[64] Chih-Fong Tsai and Mao-Yuan Chen. Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications*, 37(3):2006–2015, March 2010.

[65] Chih-Fong Tsai and Yu-Hsin Lu. Data Mining Techniques in Customer Churn Prediction. *Recent Patents on Computer Science*, 3(1):28–32, February 2010.

[66] Joan van Aken, Hans Berends, and Hans van der Bij. *Problem Solving in Organizations: A Methodological Handbook for Business and Management Students.* Cambridge University Press, New York, 2007.

[67] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, April 2012.

[68] Sofia Visa and Anca Ralescu. Issues in Mining Imbalanced Data Sets - A Review Paper. *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, pages 67–73, 2005.

[69] Fudong Wang and Meimei Chen. The Research of Customer's Repeat - Purchase Model Based on Data Mining. *2009 International Conference on Management and Service Science*, pages 1–3, September 2009.

[70] Web of Science. http://thomsonreuters.com/web-of-science-core-collection/, 2014.

[71] Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23(2):103–112, August 2002.

[72] Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, April 2009.

[73] Yu Zhao, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren. One-Class Support Vector Machine. In *Advanced data mining and applications*, pages 300–306. Springer, 2005.

[74] Bing Zhu, Jin Xiao, and Changzheng He. Proceedings of the Eighth International Conference on Management Science and Engineering Management. *International Conference on Management Science and Engineering Management*, 280:97–104, 2014.

Appendix A

All accepted and rejected variables

Product related variables	Accepted	Source	Reason of rejection
Premium price	Yes	Lit and E	
Discount	Yes	Lit and E	
Deductible excess	Yes	E	
Payment method	Yes	Lit and E	
Type of insurance	Yes	Lit and E	
Product usage	Yes	Lit and E	
Brand credibility	No	Lit and E	Not stored in the data base
Switching barrier	No	Lit	Not stored in the data base
Contracted care	No	E	Not stored per customer

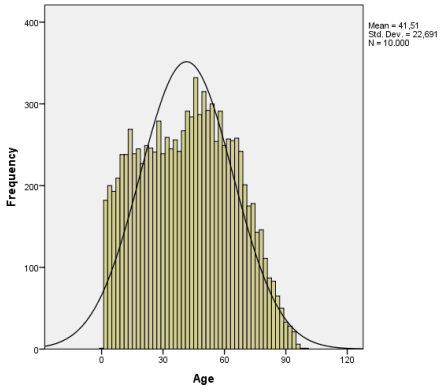
TABLE A.1: Product related variables selected from the literature (Lit) and experts (E). Indicated if the variable is accepted or rejected. And if the variable is rejected the reason of rejection.

Customer/company-interaction variables	**Accepted**	**Source**	**Reason of rejection**
Number of contact moments	Yes	Lit and E	
Number of complaints	Yes	Lit and E	
Number of declarations	Yes	Lit and E	
Outstanding charges	Yes	Lit and E	
Duration of current insurance contract	Yes	Lit and E	
Type of contact (email, call, etc)	Yes	E	
Number of authorizations	Yes	E	
Handling time of authorizations and declarations	Yes	E	
Elapsed time since last contact moment	No	Lit	Not enough time to collect all the information
Customer mentioned that they are going to switch	No	E	Not completely stored in the data base
Experience during contact moment	No	E	Is not completely stored in the data base
Elapsed time since the last complaint	No	Lit	Not enough time to collect all the information
Reaction on marketing actions	No	Lit and E	Not stored in the data base
Number of times subscribed	No	Lit and E	Not enough time to collect all the information
Socio-demographic variables	**Accepted**	**Source**	**Reason of rejection**
Identification number	Yes	Lit and E	
Age	Yes	Lit and E	
Gender	Yes	Lit and E	
Location identifier (ZIP code)	Yes	Lit and E	
Network attributes	Yes	Lit and E	
Segment selected by the company	Yes	Lit and E	
Educational level	No	Lit and E	Not stored in the data base
Income	No	Lit and E	Is not stored in the data base
Customer satisfaction	No	Lit and E	Not stored in the data base
Life events	No	E	Not completely stored in the data base

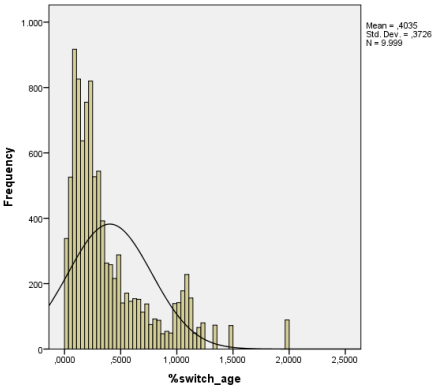
TABLE A.2: Customer/company-interaction and socio-demographic variables selected from the literature (Lit) and experts (E). Indicated if the variable is accepted or rejected. And if the variable is rejected the reason of rejection.

Appendix B

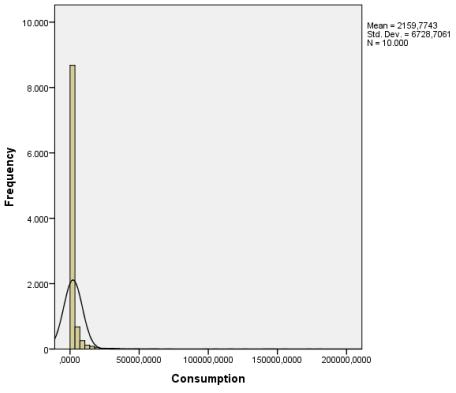
Graphical examination of the data



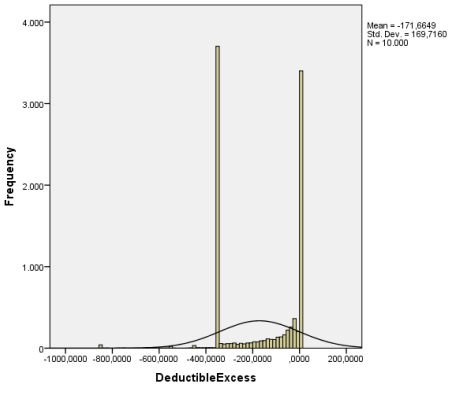
(A) Age.



(B) Duration of contract.

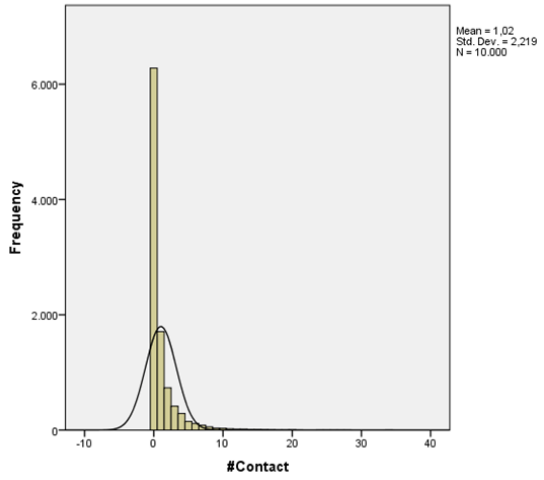


(C) Consumption.

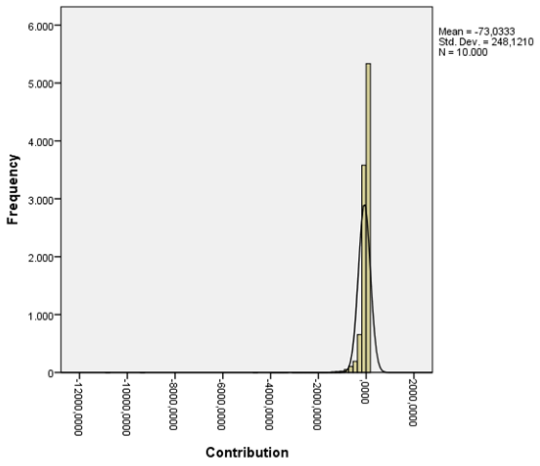


(D) Deductible excess.

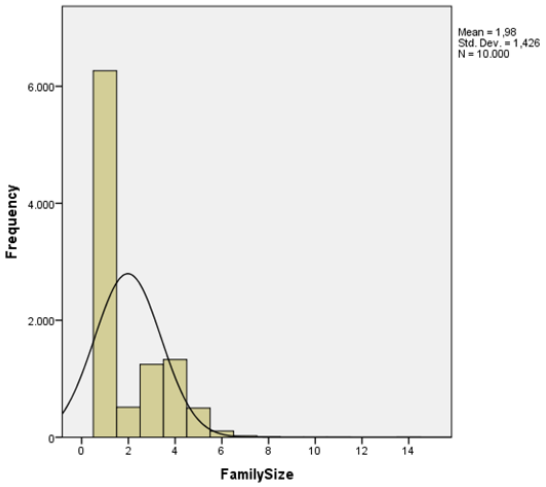
FIGURE B.1: Part 1: A visual insight of the interesting variables in the data set.



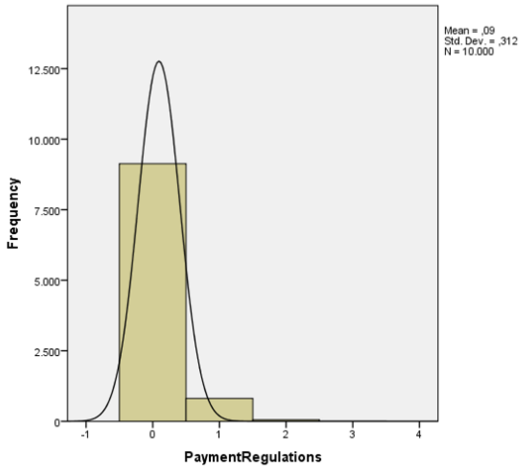
(A) Number of contact moments.



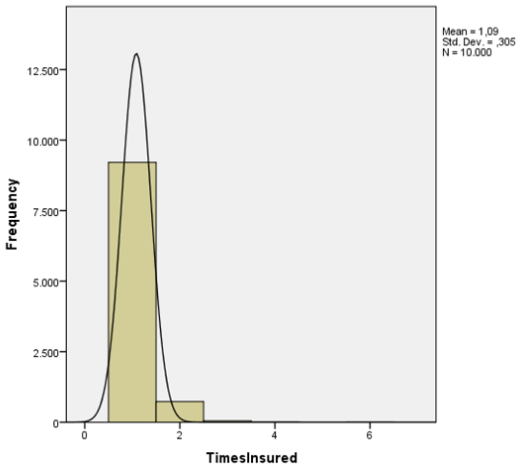
(B) Contribution level.



(C) Family size.

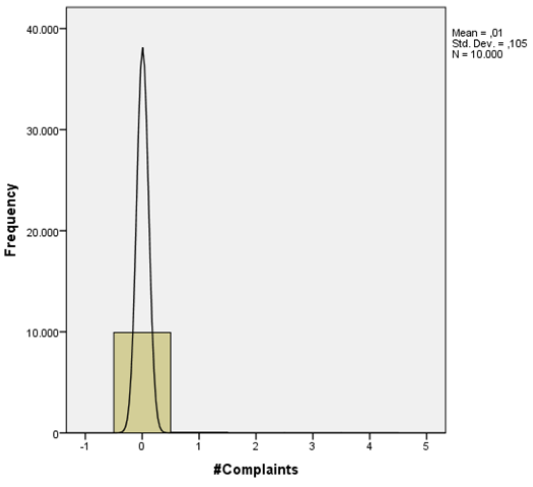


(D) Number of payment regulations.

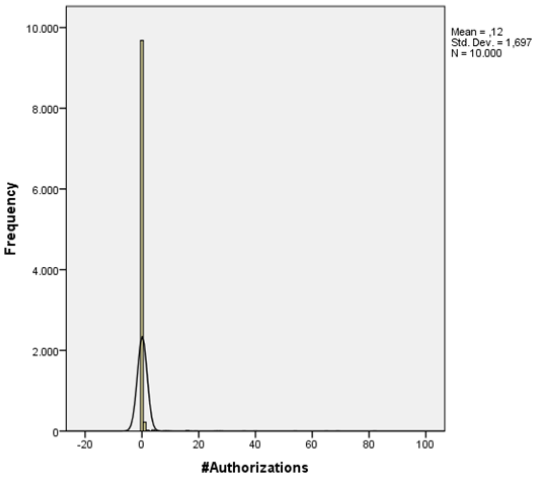


(E) Times insured.

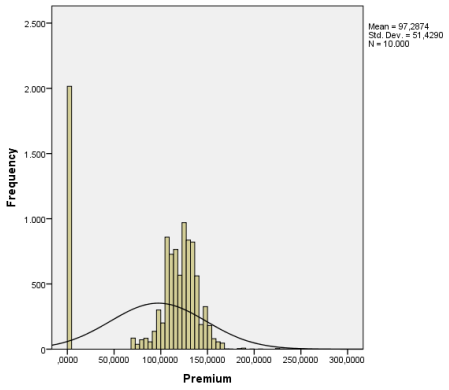
FIGURE B.2: Part 2: A visual insight of the interesting variables in the data set.



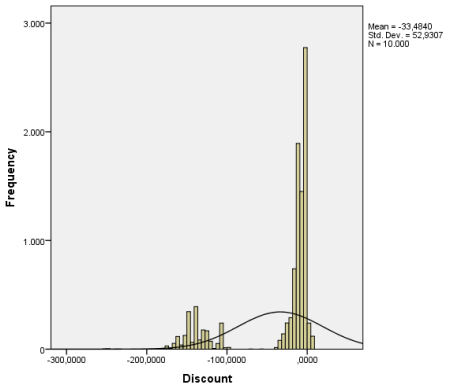
(A) Number of complaints



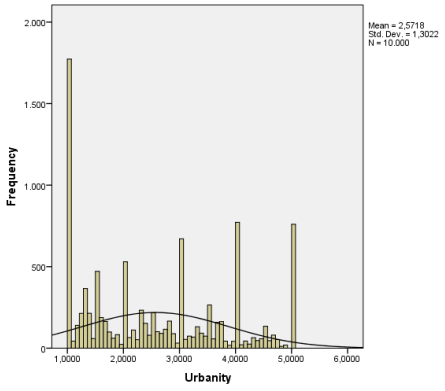
(B) Number of authorizations



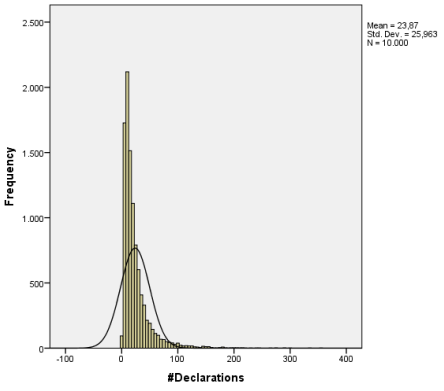
(C) Premium.



(D) Discount.



(E) Urbanity.



(F) Declarations.

FIGURE B.3: Part 3: A visual insight of the interesting variables in the data set.

Appendix C

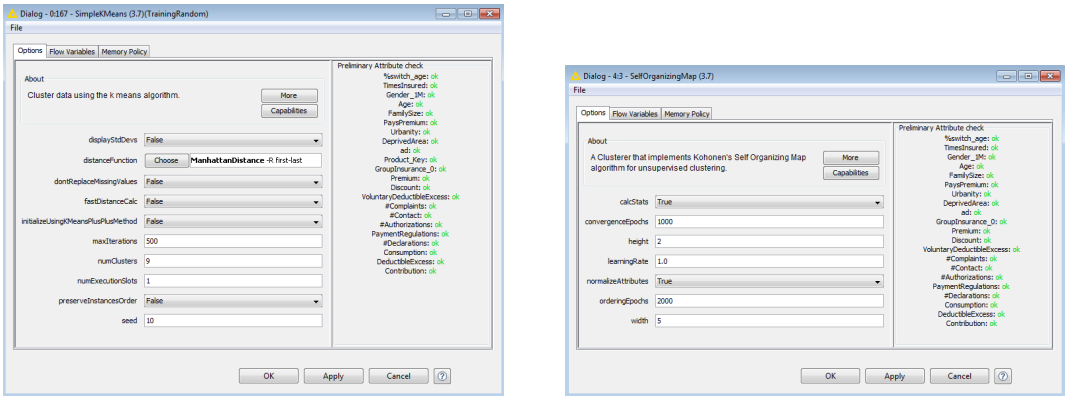
Accepted literature for identification of the used techniques

1. W. Au, K. Chan, X. Yao. *A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction* [3].
2. M. Chen, A. Chiu, H. Chang. *Mining changes in customer behavior in retail marketing* [8].
3. B. Chu, M. Tsai, C. Ho. *Toward a hybrid data mining model for customer retention* [9].
4. Y. He, Z. He, D. Zhang. *A Study on Prediction of Customer Churn in Fixed Communication Network Based on Data Mining* [26].
5. W. Heng-liang, Z. Wei-wei, Z. Yuan-yuan. *An Empirical Study of Customer Churn in E- commerce Based on Data Mining* [28].
6. C. Huang, S. Hsueh. *Customer behavior and decision making in the refurbishment industry-a data mining approach* [31].
7. S. Khakabi, M. Gholamian, M. Namvar. *Data Mining Applications in Customer Churn Management* [39].
8. W. Lin, C. Tsai, S. Ke. *Dimensionality and data reduction in telecom churn prediction* [45].
9. N. Momtaz, S. Alizadeh, M. Vaghefi. *A new model for assessment fast food customer behavior case study: An Iranian fast-food restaurant* [47].
10. K. Ng, H. Liu. *Customer Retention via Data Mining* [49].
11. K. Ng, H. Lui, H. Kwah. *A Data Mining Application: Customer Retention at the Port of Singapore Authority (PSA)* [50].

12. G. Nie. *The Analysis on the Customers Churn of Charge Email Based on Data Mining Take One Internet Company for Example* [52].
13. K. Smith, R. Willis, M. Brooks. *An analysis of customer retention and insurance claim patterns using data mining: a case study* [61].
14. H. Song, J. Kim, S. Kim. *Mining the change of customer behavior in an internet shopping mall* [62].
15. C. Tsai, Y. Lu. *Data Mining Techniques in Customer Churn Prediction* [65].
16. F. Wang, M. Chen. *The Research of Customer's Repeat - Purchase Model Based on Data Mining* [69].
17. C. Wei,I. Chiu. *Turning telecommunications call details to churn prediction: a data mining approach* [71].

Appendix D

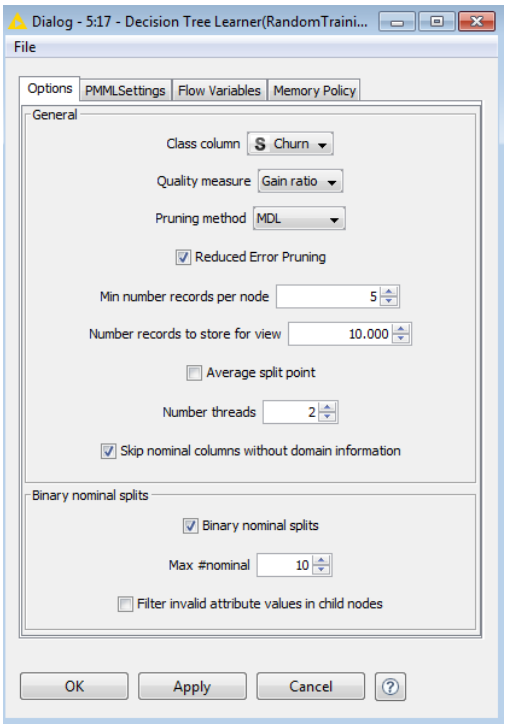
General settings used during profiling and prediction model generation.



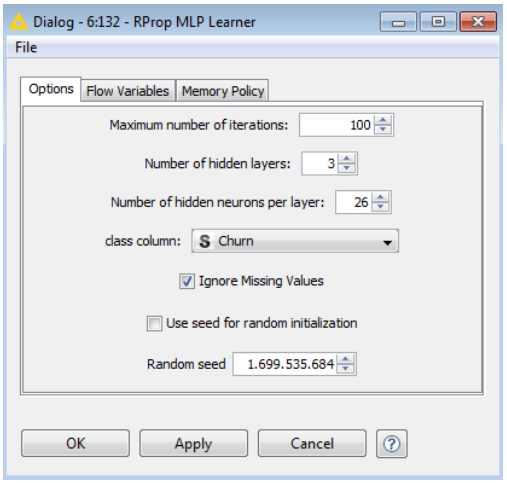
(A) General settings used for the K-means profiling technique.

(B) General settings used for the SOM profiling technique.

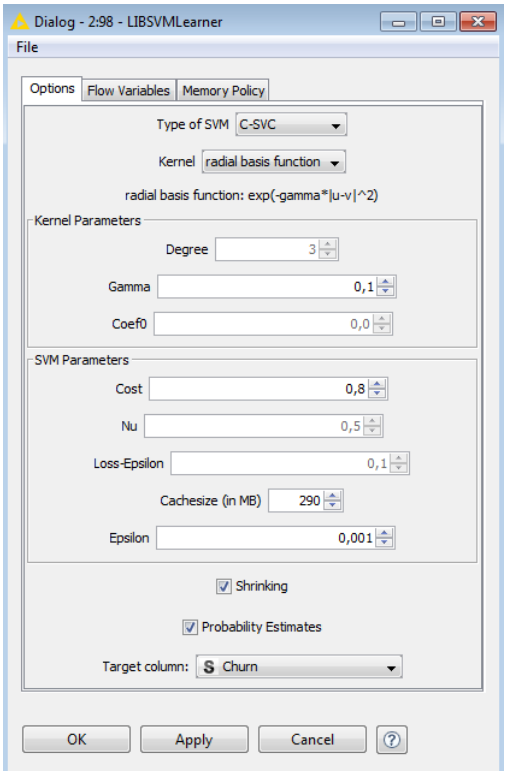
FIGURE D.1: General settings for the profiling techniques.



(A) General settings used for the DT prediction technique.



(B) General settings used for the NN prediction technique.



(C) General settings used for the SVM prediction technique.

FIGURE D.2: General settings for the prediction techniques