Applying Data Mining in Telecommunications

Proefschrift
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op maandag 11 December 2017
klokke 10:00 uur
door
Dejan Radosavljevik
geboren te Skopje, Macedonië
in 1975

Promotor: Prof. dr. J.N. Kok	Universiteit Leiden
Co-promotor: Dr. P.W.H. van der Putten	Universiteit Leiden

Doctorate committee

Prof. dr. A. Plaat	Universiteit Leiden
Prof. dr. T. Bäck	Universiteit Leiden
Prof. dr. H.J. van den Herik (Secr.)	Universiteit Leiden
Prof. dr. C. Soares	Porto University
Dr. M. Baratchi	Universiteit Leiden
Prof. dr. M. Pechenizkiy	Technische Universiteit Eindhoven

This thesis is based upon work supported by T-Mobile Netherlands B.V.

Printing: Ridderprint BV | www.ridderprint.nl
ISBN: 978-94-6299-814-8

Contents

Chapter 1

Introduction

Everyone talks about rock these days; the problem is they forget about the roll.

Keith Richards

Due to the digital revolution in the last few decades, data has become abundant. The overwhelming presence of digital devices (e.g. computers, smart-phones) combined with platforms that enable generating and storing data have led to quantities of data that were difficult to imagine in the past. According to Marr (2015), more data has been created in the two years before 2015 than in the entire previous history of the human race. A more recent report from IBM (2017) has made this claim even more specific and shocking, stating that 90% of the data in the world today has been generated in the last two years[1]. Furthermore, 2.5 quintillion bytes of data is generated daily and this figure will likely grow, given the emergence of new technologies, devices and sensors (IBM, 2017). This raises the question of which part of that data is relevant or valuable. Getting value out of this abundance of data is of interest for both industry and academia.

Data mining is defined in brief as the process of discovering patterns in data (Witten and Frank, 2005). The patterns discovered must be meaningful in that they lead to some advantage, e.g. an economic advantage. Grant (2003) provides a more detailed definition of data mining as "an interdisciplinary field bringing techniques of machine learning, pattern recognition, statistics, databases, and data visualization to address the issues of information extraction from large databases." Furthermore, a process of data mining would lead to discovery of correlations, patterns and trends, going through large amounts of data stored in databases by applying pattern recognition techniques, statistical as well as mathematical techniques to analyze the gathered data. In short, as explained by Grant (2003), the analogy of mining suggests sifting through large amounts of low-grade ore (data) to find something valuable (information).

This thesis applies data mining in commercial settings in the telecommunications industry. The research for this thesis has been performed at T-Mobile Netherlands B.V. and the methods described in some of the chapters have been also applied in Deutsche Telekom subsidiaries in other countries. We had a rare opportunity to work on real commercial data sets and have the results of our research deployed in practice. Throughout this thesis we describe some of the challenges that data miners (or data scientists) meet when working on business problems and our solutions to these problems. The complex data sets we were analyzing contained in certain cases millions of records. In this research we were using simple methods combined in innovative ways to achieve results that were either an improvement on how the business was previously solving these problems or solving important business problems that were not addressed before in such detail. We address the stages of CRISP-DM (***CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining*), shown on Figure 1.1 (Wirth and Hipp, 2000), and our main focus is on the stages least covered in literature.

[1]The difference between these two reports can be seen as an indicator of the data explosion between their respective periods of publication

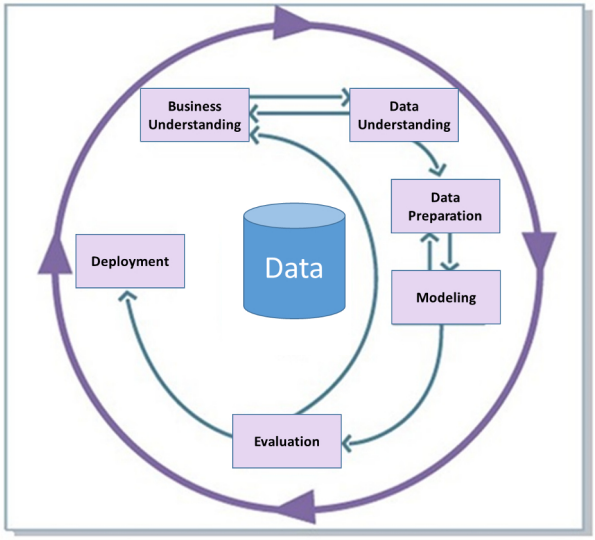

Figure 1.1: CRISP-DM Process Model for Data Mining

From a business understanding perspective, we had a unique opportunity to work on multiple business problems in mobile telecommunications. Our research stretches between the application areas of marketing, mobile network technology and finance. One can also see the respective departments of the operator as the end-users of the research. From a marketing perspective we address the problem of churn prediction. From a mobile network technology perspective, we are addressing the problem of forecasting mobile network load in order to identify sites which can potentially become overloaded in the future. From finance perspective we are addressing the process of service revenue forecasting. Working on each of these business problems required a lot of domain knowledge, which was provided by the experts working at the operator.

Data understanding and preparation is often only marginally addressed, mostly via dealing with outliers or missing values. However, in business settings understanding the data and preparing it for analysis takes a large part of the overall effort. The data that is necessary to address the problem is not even in the same database, let alone in a shape suitable for data mining. Given the scale of data sets in telecommunications, understanding and preparing the data is a task far from trivial. An example of this would be generating useful predictors from a call graph. Even setting the outcome variable is a matter of discussion, for example in the case of churn: do we take the moment the customer requested to be disconnected or the moment the customer was actually disconnected as the moment of churn? This can depend on the purpose of the model (e.g. churn campaign versus revenue forecasting).

Modeling and evaluation in literature are mostly covered by creating an algorithm that outperforms known approaches and using a standard performance metric (Root Mean Squared Error, Accuracy, Precision, AUC, Rank Correlation measures etc.). This is frequently the research focus. Instead, in our research we are using mostly standard and well known algorithms (e.g. linear or logistic regression) on large data sets. Our results show that these perform very well on commercial data sets where variance is a much more important problem than bias (van der Putten and van Someren, 2004). Algorithm tuning is out of scope of this thesis. While we did use standard performance metrics to evaluate performance, we also show how industries evaluate the overall success of a method.

The last step of the CRISP-DM process, deployment, is also important for industry. We find it essential to actually deploy the research developed, as things tend to function differently in a lab setting with artificial data sets than in the real world, where the data sets do not match the assumptions, parameters and distributions. Furthermore, we discovered that deployment choices can positively affect the acceptance of a data mining solution. Specifically, we created a scenario simulation platform for forecasting of mobile network load and service revenues, which generated use cases not originally foreseen. Additionally, we translated a churn model into a set of guidelines for optimizing the mobile network. Neither of these are standard deployment methods. On a personal note, we find it highly motivating and gratifying to see our research in action.

While research on algorithms definitely has its merits, focusing on the algorithm alone also has negative sides. The diffusion of an algorithm or a method depends on many factors other than just performance in terms of accuracy. In an article on implementation of data science in business organizations Veeramachaneni (2016) identifies reasons for failures of these projects in commercial settings: lack of data quality, lack of focus on business value of the model, focus on algorithm and tuning instead of translating the business problem into a machine learning problem etc. According to this research, machine learning experts were used to working with data already aggregated in useful variables. In our view, this practice has two drawbacks: first, some useful variables could have been omitted or not foreseen as useful by the creators of the data set; second, it is unclear to the machine learning experts how these values were created (e.g. in case of averages, how were missing values treated). The authors of this paper suggest, among other things, using simpler models and focusing on automation, which is in many respects similar to our approach.

The inspiration for this thesis is our belief that it is academia's responsibility not only to discover new ways of solving problems, but also to educate the business and government sectors on how to use these advances. This is a problem in the case of machine learning, where the giants of today (e.g. Facebook, Google or NSA) are utilizing cutting edge machine learning research on a large scale, while the rest of the companies are lagging behind. We do not want to criticize the devotion of researchers to new and improved machine learning algorithms, but to encourage them to go one

Table 1.1: Mapping of the Focus of the Thesis Chapters to the Stages of the CRISP-DM process

	Focus on Stages of Crisp DM					
	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Chapter 2	Key	Key	Strong	Weak	Strong	Weak
Chapter 3	Strong	Strong	Key	Key	Medium	Weak
Chapter 4	Strong	Strong	Strong	Weak	Key	Key
Chapter 5	Strong	Strong	Strong	Medium	Medium	Key
Chapter 6	Strong	Strong	Key	Strong	Key	Key

step further into deployment of these methods in practical circumstances (diffusion of their innovation), therefore realizing the full potential of their research.

For the reasons stated above, this thesis is focused on how to apply data mining in a real world setting. Even though we have focused on a single industry, i.e. telecommunications, other industries can also benefit, as the methods applied are easily transferable. Furthermore, most industries are concerned with how to keep their customers (reduce churn), how to improve the service they are offering (manage the network in telecom) and how to forecast their revenues. The latter two problems are transferable to governments or non-governmental organizations as well. We will also discuss the tools we used, as well as the deployment methods that helped our research to gain acceptance by the business.

1.1 Thesis Structure

This thesis is largely based on published papers. In this section we present the research questions, a brief overview of each chapter, as well as the focus of each chapter and the contribution with relation to the CRISP-DM process (see Table 1.1). In general, in almost every chapter our efforts in business understanding, data understanding and data preparation are a substantial part of the overall effort. However, for each chapter a different CRISP-DM stage is the key focus area. The values in Table 1.1 are given as guidelines of where we find our key contributions or what we see as a (interesting or unusual) solution for these stages of the process.

At the time when we began this research, data mining within T-Mobile Netherlands B.V. was not widely applied. Therefore, the overall research question of this thesis is:

How does one successfully apply data mining in telecommunications?

The chapters of this thesis are cases of applying data mining onto telecom problems. Each of these chapters will have a research (sub)question of its own. These will be listed in the next few paragraphs adjacent to the chapter overview.

Customer churn, i.e. losing a customer to the competition, is a major problem

in mobile telecommunications and many other industries. This is why we dedicate three chapters to this problem.

In Chapter 2 we discuss the impact of the experimental setup on prediction of prepaid churn in telecommunications. This chapter is based on our paper "The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter?" (Radosavljevik, van der Putten and Kyllesbech Larsen, 2010*a*). Prepaid customers in mobile telecommunications are not bound by a contract and can therefore change operators ("churn") at their convenience and without notification. This makes the task of predicting churn both challenging and financially rewarding. The chapter presents an exploratory study of prepaid churn modeling by varying the experimental setup on three dimensions: data, outcome definition and population sample. The research question for this chapter is:

Which one of the following variations in the experimental setup has the highest influence on the performance of prepaid churn prediction models: adding Customer Experience Management data, altering the characteristics of the sample or changing the outcome definition?

While adding more input variables and varying the sample did not influence the predictability of churn, a particular change in the outcome definition had a substantial influence. Here we emphasize that the problem formulation often is more important than the data or the method used to solve it. From the perspective of the CRISP-DM process, our main focus is on the importance of the business understanding and data understanding stages. This chapter also provides a method of automated data preparation. The algorithms that we are using here are quite simple and standard, hence the value "weak" for focus on modeling in Table 1.1. In this chapter we are also explaining the evaluation method and our choice of a performance metric that we are using in chapters 3 and 4.

In chapter 3, based on the publication "Combining Customer Attribute and Social Network Mining for Prepaid Mobile Churn Prediction" (Kusuma, Radosavljevik, Takes and van der Putten, 2013), we investigate the added value of combining regular tabular data mining with social network mining, leveraging the graph formed by communications between customers. Here we compared the performance of classical (tabular) prepaid churn models with models generated by mining the social network graph and two hybrid approaches: first, we enriched the data set used for the tabular models with features extracted from the communications graph and second, we created a propagation model using the scores of the tabular churn models as initial energy of each non-churner node (similar to boosting). The research question for this chapter is:

Do social network mining or attributes stemming from a social network graph add value in terms of model performance to traditional prepaid churn modeling in T-Mobile Netherlands?

From a CRISP-DM perspective, this is the only chapter where our main focus is on the modeling stage: we present two novel hybrid models. Transforming the call graph into features that can be used for data mining (data preparation) was also

substantial part of the work. However, our experiments showed that despite the high computational effort, the traditional tabular churn models scored best. This can be seen as an example application of the No Free Lunch Theorem (Wolpert and Macready, 1997), showing that Social Networks do not necessarily add value to every telecom churn prediction problem, opposite from many results shown from this period (see Dasgupta et al., 2008).

In the highly competitive and advanced telecommunications market in The Netherlands, network experience is crucial for the operators. T-Mobile Netherlands wanted to optimize the network experience for its customers in order to increase the satisfaction with the current customer base and attract new customers based on mobile network quality. This is why chapters 4 and 5 have two different ways of improving the network in their focus.

Chapter 4, based on the paper "Preventing Churn in Telecommunications: The Forgotten Network" (Radosavljevik and van der Putten, 2013), shows a different application of a churn model. While the problem of churn prediction is frequently addressed in literature, preventing customers from wanting to churn is not. Hence, the research question for this chapter is:

As a different method for model deployment, can a churn model be used to prevent churn by explaining its causes as opposed to using the predictions for targeting customers?

This chapter outlines an approach developed as a part of a company-wide churn management initiative. Our approach to churn prevention can also be seen as a bridge between the disciplines of marketing and mobile network technology, because we identify the technical drivers of churn. We are focusing on an explanatory churn model for the postpaid segment, assuming that the mobile telecom network, the key resource of operators, is also a churn driver in case it under-delivers to customers' expectations. The main focus of this chapter with regards to the CRISP-DM process is in the deployment and evaluation stages. The typical deployment method for a churn model is a retention campaign where customers are approached with an offer to continue their contract. In this case, there was no campaign. The model was used to generate a set of rules for network optimization in order to remove the key network related churn drivers and therefore prevent churn, rather than cure it. The insights generated by this model have caused a paradigm shift in managing the network of T-Mobile Netherlands. From an evaluation perspective, the evaluation of the model did not stop with just measuring performance on the test set. The real evaluation came months later, by measuring customer satisfaction after implementing the network optimization rules that were the generated from the model. This approach was later also used by a Deutsche Telekom operator in another country.

In chapter 5, based on "Large Scale Predictive Modeling for Micro-Simulation of 3G Air Interface Load" (Radosavljevik and van der Putten, 2014), we are trying to answer the following research question:

How can data mining be used to predict 3G mobile network interface load and simulate it under different scenarios?

In this chapter we describe how we built a large scale network load simulation tool. Forecasting mobile network load is typically used for planning network upgrades and budgeting purposes. However, focusing on the end user by using tools familiar to them resulted into applications of the model far beyond the original design. Looking at the CRISP-DM process, the key stage in this chapter is deployment. Our method of deployment enabled the end users that are not data miners to engage in scenario simulation activities of the complex network system. We created a micro-simulation system, which allowed testing the effect of changing the values of the input parameters according to scenarios envisioned by the users on the load of each cell in the mobile network. Even though the algorithm we used, linear regression, is not novel, our automated modeling process using wrappers for variable selection enabled us to generate models at a high pace, resulting in 30,000 models per day. After the initial success in T-Mobile Netherlands where the method was developed, the approach was also used by Deutsche Telekom operators in three other countries.

In chapter 6 we extend the method developed in chapter 5 onto the field of finance and service revenue forecasting. This chapter is based on "Service Revenue Forecasting in Telecommunications: A Data Science Approach" (Radosavljevik and van der Putten, 2017). The research question in this chapter is:

How can data mining be used to predict and simulate service revenues in telecommunications? In other words, does the approach developed in chapter 5 generalize to a different domain such as finance?

Revenue forecasting in general is one of the most important financial processes in any industry. For service based business such as telecommunications, timely and precise service revenue forecasts are essential, because they can drive important business decisions, such as when and where to intervene in order to accomplish the business targets. Here, we replicated the deployment method we discuss in chapter 5 using different tooling, as this was the end users' choice. By creating a simulation platform for service revenue forecasting the business can have a better idea of how different scenarios for changes in the input variables or the measures the business is planning could play out in practice. The key stages of CRISP-DM for this chapter are deployment, evaluation and data preparation. The data preparation stage is of interest, because we created our own data store using tools that are not originally designed for this purpose. The important aspects of deployment are similar to chapter 5: using tools familiar to end users and creation of a scenario simulation platform. From a modeling perspective this chapter is interesting because of the comparison between relatively simple and more complex models we provide, as well as the change in the way we modeled the outcome. From an evaluation perspective, we used mean absolute error as a performance metric to build the models. However, a different, problem specific performance metric was the only measure of success that was relevant to the business.

Finally, in chapter 7 we present the conclusions, lessons learned and the summary of this thesis.

Chapter 2

The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter?

Radosavljevik, D., van der Putten, P., & Larsen, K.K.

Based on a publication in Transactions on Machine Learning and Data Mining 3 (2), pp. 80-99 (2010)

Prepaid customers in mobile telecommunications are not bound by a contract and can therefore change operators ("churn") at their convenience and without notification. This makes the task of predicting churn both challenging and financially rewarding. This chapter presents an exploratory, real world study of prepaid churn modeling by varying the experimental setup on three dimensions: data, outcome definition and population sample. First, we add Customer Experience Management (CEM) data to data traditionally used for prepaid churn prediction. Second, we vary the outcome definition of prepaid churn. Third, we alter the sample of the customers included, based on their inactivity period at the time of recording. While adding CEM parameters and varying the sample did not influence the predictability of churn, a particular change in the outcome definition had a substantial influence.

2.1 Introduction

The problem of churn, or loss of a client to a competitor, is a problem facing almost every company in any given industry. This phenomenon is a major source of financial loss, because it is generally much more expensive to attract new customers than it is to retain and sell to existing ones. Therefore, churn is important to manage, especially in industries characterized by strong competition and saturated markets, such as the mobile telecom industry. Prepaid mobile phone subscribers are not bound by a contract; therefore, they can churn at their convenience and without notification, which makes the task of predicting the likelihood and moment of churn very important.

The first step in managing churn is identifying the customers with high propensity to churn. Published papers on churn modeling on private data sets in mobile telecommunications are relatively scarce, due to the commercially sensitive nature of the problem. This is even more evident for papers based on European mobile telecom operators' data. Most of the available research relates to postpaid churn (Au, Chan and Yao, 2003; Datta, Masand, Mani and Li, 2000; Ferreira, Vellasco, Pachecco and Barbosa, 2004; Hung, Yen and Wang, 2006; Hwang, Jung and Suh, 2004; Kim and Yoon, 2004; Lemmens and Croux, 2006; Lima, Mues and Baesens, 2009; Mozer, Wolniewicz, Johnson and Kaushansky, 1999; Neslin, Gupta, Kamakura, Lu and Mason, 2006; Wei and Chiu, 2002). Even fewer studies are available for prepaid churn in this industry (Alberts, 2006; Archaux, Martin and Khenchaf, 2004; Dasgupta et al., 2008). The majority of these assumed a fixed, single experimental setup in terms of outcome definition, population and data parameters. Their focus was mostly on the data mining algorithm used or algorithmic tuning. In order to understand prepaid churn modeling better, we decided to take an end-to-end view and test different experimental setups by varying on three dimensions: data, population sample and outcome definition (Radosavljevik, van der Putten and Kyllesbech Larsen, 2010*b*). We used standard data mining algorithms, decision trees (in their standard form) and logistic regression (Pegasystems, 2008; Witten and Frank, 2005), because it was

not our objective to determine the impact of the algorithm; research on data mining algorithms is abundant. Furthermore, from a business perspective, the output of either of these algorithms (the model) is very easy to interpret and communicate to parties that do not have extensive experience with data mining (e.g. business managers). This quality is even more evident in the case of decision trees, which have a very intuitive graphical representation. Finally, as our experiments have shown, the choice of algorithm is a minor factor of influence compared to the other dimensions mentioned.

Making new types of data available for modeling may improve model performance. We provide an overall framework for measuring customer experience, and in the experiments we investigate the added value of customer level service quality metrics (dropped calls, call setup duration, SMS delivery failure rate etc.).

The definition of population sample and outcome is primarily driven by the business objectives, i.e. how the model will be used. In general, operators discontinue the service and consider prepaid customers churned from an administrative perspective if they have not had an activity (e.g. made a call, sent an SMS, etc.) for a period of six months. However, from a churn prediction and marketing perspective such a long period is not very useful. Usually, the customer has churned long before that time. Therefore, it is much better to use a shorter period of inactivity as an outcome definition of churn. Our variations in population sample were based on the customers' inactivity period at the moment of recording. The dilemma here is, if customers have been inactive already for one month at the time of recording, are they retainable, or have they simply thrown away the SIM card? What should be the maximum period of inactivity allowed at the time of recording?

Our choice for the research question for this chapter was driven by the multiple possibilities and choices that exist when creating an experimental setup for prepaid churn modeling:

Which one of the following variations in the experimental setup has the highest influence on the performance of prepaid churn prediction models: adding Customer Experience Management data, altering the characteristics of the sample or changing the outcome definition?

The main contributions of this chapter are as follows. First, we provide a novel theoretical business framework for measuring customer experience in mobile telecommunications. Second, to our knowledge we are the first to investigate the added value of service quality oriented customer experience data for predicting prepaid churn. Third, we explore the impact of changing the criterion for the population sample. Fourth, we propose different outcome definitions of prepaid churn and measure the impact of changing the outcome definition on model performance. Finally, this research was conducted by one of the leading mobile telecom operators in Europe, driven by high priority business needs and using large amounts of customer data, which gives it a real world dimension. Given the highly competitive, confidential and strategic nature of mobile telecommunications churn management, real world results (based on the European telecom market) are not often available in literature.

The remainder of this chapter is structured as follows: In Section 2.2 we present how (prepaid) churn modeling has been performed in the past. Section 2.3 provides an overview of Customer Experience Management. In Section 2.4, the research setup and the experimental design are discussed, followed by results in Section 2.5. We end the chapter with a discussion (Section 2.6) and a conclusion (Section 2.7).

2.2 (Prepaid) Churn Modeling

In this section, we will present how modeling churn in mobile telecommunications has been addressed in previous research.

As mentioned in the introduction, the majority of papers on wireless churn assume a single experimental setup and deal with postpaid churn (Au et al., 2003; Datta et al., 2000; Ferreira et al., 2004; Hung et al., 2006; Hwang et al., 2004; Kim and Yoon, 2004; Lemmens and Croux, 2006; Lima et al., 2009; Mozer et al., 1999; Neslin et al., 2006; Wei and Chiu, 2002). Some of these papers did not explicitly state that they are related to postpaid churn; nonetheless, this conclusion can be made from the data sets they used, which contained demographic information and contract data. These parameters are unavailable for prepaid subscribers, as there is no contract in the real sense of the word. This is the key difference between prepaid and postpaid churn prediction, even though quite a few similarities exist.

Prepaid churn modeling is typically done by performing analysis on aggregated Call Detail Records (CDRs), but additional factors, such as subscription details (e.g. duration, subscription or discontinuation of services) and handset data, are often included. Parameters used in previous papers are as follows.

Archaux et al. (2004) took into account the following data: invoicing data, such as the amounts refilled by clients or amounts withdrawn by companies for the subscribed services and options; data relating to usage, such as the total number of calls, the share of local, national or international calls, consumption peaks and averages; data relating to the subscription, such as date of beginning (age of the subscription), the current tariff plan and the number of different plans the client used; data relating to application and cancellation of services; data related to the current and the previous profitability of the clients. Alberts (2006) selected data related to usage and billing information only: average duration of a single incoming and outgoing call, ratio between outgoing and incoming call durations, sum of incoming and outgoing revenues, the current and the maximum number of successive non-usage months, cumulative number of recharges, average duration of incoming and outgoing calls over all past months, the number of months since the last voicemail call, the number of months since the last recharge, the last recharge amount, the average recharge amount of all past months, etc.

The algorithm is the focus of many of the data mining efforts related to churn. Papers related to postpaid churn have investigated the performance of standard data mining algorithms, such as decision trees (Au et al., 2003; Ferreira et al., 2004; Hung

et al., 2006; Hwang et al., 2004; Lemmens and Croux, 2006; Lima et al., 2009; Mozer et al., 1999; Neslin et al., 2006; Verbeke, Martens, Mues and Baesens, 2011; Wei and Chiu, 2002), logistic regression (Hwang et al., 2004; Lemmens and Croux, 2006; Lima et al., 2009; Mozer et al., 1999; Neslin et al., 2006; Verbeke et al., 2011), neural networks (Au et al., 2003; Datta et al., 2000; Ferreira et al., 2004; Hung et al., 2006; Hwang et al., 2004; Mozer et al., 1999; Neslin et al., 2006), Bayesian classifiers (Neslin et al., 2006) and support vector machines (Verbeke et al., 2011), and compared them to novel approaches.

In prepaid churn modeling, Archaux et al. (2004) compared the performance of models built using neural networks and support vector machines, while Alberts (2006) compared models built using survival analysis with models built using decision trees. To summarize, in research literature there is a lot of emphasis on algorithm testing and tuning, whereas in real world practice this is a smaller piece of the puzzle with relatively modest impact. Hence, we decided to focus more on experimental setup.

The influence of Social Networks on (prepaid) churn has entered the focus of interest of both business and academic communities in the period after 2008 (Dasgupta et al., 2008; Richter, Yom-Tov and Slonim, 2010; Wang, Cong, Song and Xie, 2010). These papers strive to answer the question whether the decision of a subscriber to churn is dependent on the existing members of the community with whom the subscriber has a relationship. Dasgupta et al. (2008) showed that diffusion models built on call graphs (used for identification of Social Networks) have superior performance to their baseline model. At the time of publication, this approach was considered a very progressive novelty, with relation to both algorithm and data. However, it is our opinion that these results are biased by the fact that their baseline model is constructed on CDR data alone, without including other important factors, such as the handset, prepaid account balance or inactivity period. In our opinion, if these variables would have been added in their baseline model, the improvements of their approach would not have been as high.

As we will show in our experiments, combining past behavior (extent of usage) with current factors (e.g. prepaid account balance, phone type, price plan etc.) is crucial for predicting future behavior (churn). In general, there is a gap between traditional methods that do not take the social networks into account and social networks techniques that only mine the network. Therefore, in chapter 3 we investigate the value of hybrid approaches, combining these two methods.

In addition, we consider the origin of the data used in previous research. As telecom markets differ in various parts of the world, it is expected that the churners' patterns would differ as well. The US telecom market has been addressed by several papers (Datta et al., 2000; Lemmens and Croux, 2006; Lima et al., 2009). There is also research based on Asian markets (Au et al., 2003; Hwang et al., 2004; Kim and Yoon, 2004; Wang et al., 2010). Ferreira et al. (2004) analyzed the Brazilian market. Other researchers did not reveal the location of the operator (Dasgupta et al., 2008; Richter

et al., 2010) or used publicly available data sets (Lima et al., 2009; Verbeke et al., 2011). There is a visible lack of papers focusing on the European market (except Alberts, 2006).

Last, but not least, we would like to mention the issue of defining prepaid churn, or the lack of a clear definition of a churned prepaid customer. To the best knowledge of the authors only Alberts (2006) defined in detail which prepaid customers are marked as churned. We find the outcome definition to be of high importance and we address this issue in depth in Subsections 2.4.2 and 2.4.5, where we also propose our own definition(s).

2.3 Customer Experience Management

To put our churn management activities in perspective, we see them as part of a wider ranging company effort to manage and optimize the customer experience. Products and services (e.g. telecommunication services) tend to become commodities over time; therefore, managing the customers' experience can be a source of sustainable competitive advantage (Pine and Gilmore, 1999). Pine and Gilmore (1999) stated that experiences are as distinct from services as services are from goods, thus arguing that authentic experiences are the next evolutionary step in creating customer value. Smith and Wheeler (2002) claimed that branded customer experience drives customer loyalty and profitability. Customer Experience Management (CEM) is the process of strategically managing and optimizing these experiences across all customer touchpoints and channels, in interactions that are either customer initiated or company driven, direct or intermediated (Meyer and Schwager, 2007; Schmitt, 2003).

CEM is closely related to its predecessor Customer Relationship Management (CRM). The distinction is somewhat artificial, but one could argue that CRM focuses more on providing a 360 degrees view of the current relationship and managing customer processes efficiently, whereas CEM provides tools and methodologies to understand, improve and extend the relationship by optimizing the overall customer experience. For mobile telecommunication providers (and many other businesses) the key areas within an overall CEM strategy are customer acquisition, revenue stimulation for existing customers and customer retention management (churn management). Churn models, in particular, can be important components of overall strategies that make these predictions actionable to drive intelligent interactions and experiences.

2.3.1 Measuring the Customer Experience

The focus of this chapter is on building better prepaid churn models, both from a predictive power perspective, as well as from the point of view of predicting behavior that is actionable and makes business sense. For instance, predicting short term inactivity for a base that includes many already dormant accounts will result in a

powerful, yet not very useful model. Embedding the churn models in larger strategies to optimize experience in real time is more the scope of future research. From a CEM perspective, our topic of interest is measuring past customer experience as a potential input to churn models. We have created a theoretical business framework for measuring customer experience in mobile telecommunications in ideal circumstances (Figure 2.1).

To the best knowledge of the authors, this is a first attempt to create a framework of this kind and purpose. This framework provides a conceptual roadmap and direction for the coming years for the mobile telecom operator where this research was performed. In the short term, it is very challenging to measure most of these dimensions reliably. The framework contains three levels: Aspects, KPIs and Data Sources. The Aspects level represents the various aspects of customer experience in mobile telecommunications. The KPIs level defines possible measurements for the different aspects of customer experience. The Data Source level describes the potential location of various KPIs.

This framework has been designed for mobile telecommunications, but it can easily be customized for other business to consumer industries by replacing the industry specific aspects (Handset, Network Usage and Network Usability).

We are interested in the relative value of the CEM KPIs for general marketing use, as well as the business value of the software tools used to collect and analyze them. Specifically for prepaid churn prediction, the following six aspects of the CEM framework presented above can represent this added value, as they have not been used before for prepaid churn prediction: Brand, Network Usability, Customer Support, Direct Communication and Content. Network usage, Billing and Handset have been addressed by Alberts (2006), as well as Archaux et al. (2004). Peers (Social Networks) is the topic of many papers (Dasgupta et al., 2008; Richter et al., 2010; Wang et al., 2010). The Demographics aspect is not applicable to prepaid subscribers, because for most of them this data is not available.

For this research we want to particularly focus on the added value of the CEM tool data source. This tool measures service quality (dropped calls etc.) for operational purposes, but claims are made that this data is also of key importance for prepaid churn management. We want to validate whether these claims are warranted from a business case perspective, assuming that traditional usage and customer relationship data is already available.

2.4 Research Setup

In order to examine the influence of the CEM metrics, variations in population sample and outcome definitions, we constructed three separate experiments which are described in the subsections below.

A set of general rules applies to all three. The predictors were recorded at the end of March 2009 with one, three and six months of history. The random sample of

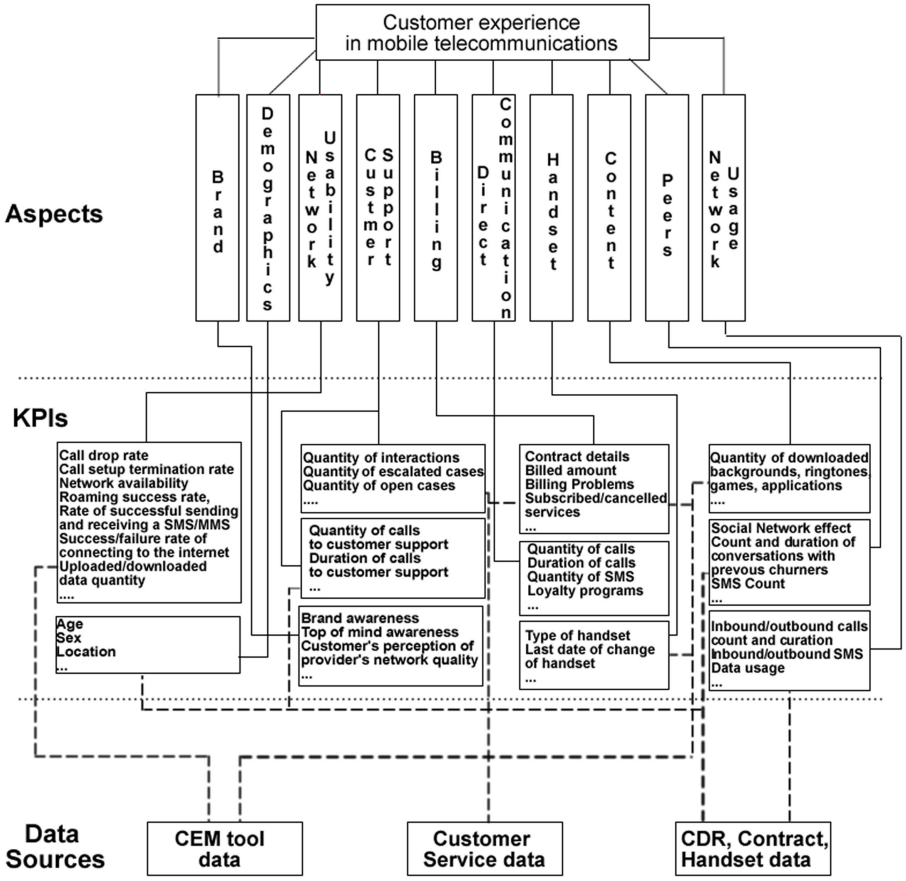

Figure 2.1: Customer Experience Framework for Mobile Telecommunications

around 10% of the prepaid customer base was divided into training, validation and test sets, using the 50:25:25 ratio. It was not necessary to oversample on churn, due to the choice of the software tool, which is able to handle unbalanced distributions of the outcome. The validation set was not used for developing the models in any automated way (e.g. model parameter setting or pruning using automated cross validation). It was used as an additional test set and for manual verification of the univariate performance of variables on an independent data set to avoid over-fitting. In other words, from an algorithm perspective it could have been called a test set, but from a methodological perspective we want to take a cautious approach and call it a validation set, because in theory the analyst himself could over-fit the data. The analyst does not have access to test set results in data preparation.

In our situation hold out validation rather than n-fold cross validation was both sufficient and more practical. If test data sets available are large enough, a simple holdout provides a true indication of future performance (see also Witten and Frank, 2005). Please note that we had ten times more instances available in the source data set. Hold out validation is also more practical given that it results in a model along with validation and test performance estimates. Cross validation at best results in a choice of an optimal modeling approach rather than a model. Hence, one would still need to build a single model, and likely require hold out validation to estimate the test set performance for the resulting model, because it might differ from the cross validation performance.

The next subsection will provide more technical details on the end to end data mining process followed.

2.4.1 End to End Data Mining Process

The commercial data mining tool Predictive Analytics Director (Pegasystems, 2008) was used for automated data preparation (attribute discretization, grouping and selection), modeling (logistic regression, decision trees) and model evaluation. All the steps in the data mining process, including the data preparation step, are actually decoupled from each other, with the goal of making the process more manageable and providing a factory approach to model building. On average, we do not expect this to have a negative impact on model performance or robustness. For instance, supervised discretization of continuous variables prior to modeling can sometimes improve the accuracy of the model (Dougherty, Kohavi and Sahami, 1995). As another example, wrapper based approaches for predictor selection (integrated predictor selection and modeling) are not inherently better than filter based approaches (predictor selection separated from modeling) (Tsamardinos and Aliferis, 2003).

The objective of data analysis is to transform the various attributes (a.k.a. variables, columns in the flat table), and identify the most informative ones among them at this stage, from a univariate perspective. A variable is considered informative if it has a certain level of influence over the outcome variable (which in this case is churn). Both statistical and graphical tools are used to establish this degree of influence. In

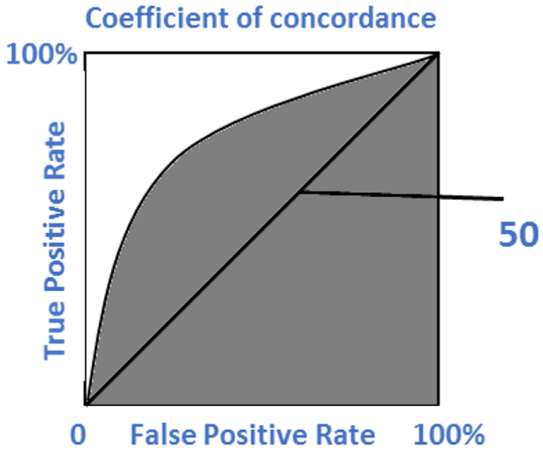

Figure 2.2: Coefficient of Concordance

this research, the chosen statistical criterion for evaluating the predictive performance of both variables and models with relation to churn is the Coefficient of Concordance (henceforth CoC), which is a rank order correlation measure, equivalent to Area under the ROC (AUC) and related to Kendall's tau (Kendall, 1938; Pegasystems, 2008). The major benefit of rank order correlation measures compared to basic measures such as accuracy is that these measures are not sensitive to skewed distributions of the outcome. For instance, assume a binary target for which one of the outcomes is very rare, e.g. churn. A majority vote model, which in practice is useless for selecting prospective churners, would have accuracy equal to the percentage of the majority class, i.e. 1 - churn_rate (e.g. if churn_rate=1%, accuracy=99%). However, in such a case, the values of CoC or AUC would only be 50 or 50%, respectively (equal to the values of random choice), which is the lowest value possible for real world models, as the prediction does not provide useful information to rank instances. One interpretation of the CoC measure is that in a scoring model it gives the probability that a randomly chosen positive case will get a higher score than a randomly chosen negative case. The CoC measures the gray area in the graph depicted on Figure 2.2 and can thus be translated to the Gini coefficient.

The data analysis process begins by discretizing continuous variables into a large number of bins. Numeric bins and symbolic values without statistically significant difference in churn rate are then grouped together. This balances the training set performance (many groups with varying churn levels) with out-of-sample (validation and test set) and out-of-time robustness (as many instances per group to provide robust estimates of churn for a group). Basically, this is a supervised, bottom-up approach to discretization of continuous variables and merging of numeric bins and

symbolic values into groups that display a significant difference in the outcome. The analyst can inspect the resulting histograms, and optionally change the parameters of the process, for instance the significance thresholds for merging bins and symbols.

After the initial data analysis, for the purpose of predictor selection, an automated procedure is used to group the variables that are correlated, independent of the target. A given predictor may have a high univariate performance, but also be correlated with other candidate predictors that are even stronger, hence not adding value to a model (subset predictor selection rather than univariate predictor selection). The user has three options when selecting predictors to be used for modeling: all predictors, only the best of each group, or manual selection. In our case we used the best predictors of each group as a starting point, and then experimented with minimal further manual selection, for instance by removing predictors from the bottom scoring groups altogether, or changing the parameters of the automated grouping process to force more or less groups. The main rationale for this was to force the use of CEM variables in order to enable a comparison.

The full data set consisted of more than 700 attributes, and the optimal number of predictors for final models typically ranges between 5 and 10 predictors. These volumes of attributes in the raw data versus the final model are quite typical for real world data mining, at least for marketing purposes; the full modeling table is reused as a basis for multiple modeling purposes, as it aims to capture all aspects relevant to customer experience (see also Figure 2.1). See the results section (Section 2.5) for more information of the types of predictors selected.

For logistic regression models, the raw predictor values are replaced with a normalized rank score concordant with the churn rate for the group (e.g. the predictor "age" is divided in various age ranges; if the group of 22-26 years shows higher churn rate on the training set, a higher score will be assigned to it). Logistic regression models are then built on the recoded data, to allow capturing of complex non-linear behavior, whilst using a stable low-variance modeler.

In our implementation, decision trees were forced to split between groups only, i.e. the grouped bins of discretized variables, thus not on the raw data (supervised discretization prior to induction (Dougherty et al., 1995)). Our motivation for this was to provide a level playing field for both algorithms (logistic regression and decision trees) and to ensure that only the "analyst approved" discretization was used. We used the CHAID splitting criterion, which selects the split points based on statistical significance as measured by the Chi Squared statistic (Kass, 1980). This criterion allows merging of both adjacent and non-adjacent bins of a discretized variable when making the splits.

For example, if a split on variable "age" is performed, and the bins 22-26 and 30-34 display similar level of significance, only one split for "age" will be created, containing both these intervals (i.e. age in 22-26, 30-34 and age not in 22-26, 30-34). The end result of our procedure is a binary tree. We are aware that CHAID and other decision tree induction methods are capable of producing n-ary trees (Perner, 2002),

which could have a somewhat better performance. However, it was not our goal to test algorithm performance, thus we used a standard binary decision tree.

The modeling process results in scoring models: a rank score concordant with the probability of being a churner is allocated to each of the instances. The CoC (AUC) measure is used to measure model quality (Kendall, 1938). In addition, we use gain charts as visual representation of model performance. On the y-axis, these charts show the captured proportion of the desired class (i.e. churners in selection divided by total number of churners) with increasing selection sizes (x-axis, from highest scoring to lowest scoring) (see Figure 2.4).

2.4.2 Definition of Prepaid Churn and Initial Sample

Defining prepaid churn is more complex than defining postpaid churn, as there is no concept of contract termination. Prepaid customers are deemed as churned if they are no longer "active" for a certain period of time.

The working definition of "activity" within the company where this research was conducted was used as operational definition. This definition is based on various aspects of usage of telecom services. On one hand, it is very detailed and captures the various aspects of activity within mobile telecom industry. On the other hand, it simplified the data acquisition.

Definition 1. Operational definition of activity: Outgoing (initiated) calls, top-up (recharge) of the account, sent SMS/MMS messages and received (incoming) and answered calls are considered as an activity. Received calls without picking up, received SMS/MMS messages and bonus credit top ups awarded by the company are NOT considered as an activity.

Involuntary prepaid churn in this company occurs after six consecutive months of no activity. After this, the customer can no longer use the services of the company via that particular SIM card, and the phone number may be reissued to another client after a certain period. However, six months of no activity is too long of a period to investigate voluntary prepaid churn. Most of the internal projects investigating prepaid churn within this company consider customers to have churned if they had not been active between two and three consecutive months. Therefore, we propose the following

Definition 2. Operational Prepaid Churn Definition: Customers are marked to have churned if they are registered to have two consecutive months of no activity, or more.

It was also necessary to set certain boundaries for the population taken into account for the modeling process. The population boundaries are set in Definition 3.

Definition 3. Population definition: The population consists of prepaid subscribers that meet the following two conditions:

1. They have at least one registered activity in the last 15 days.

2. Their first activity date is at least four months before.

The first condition enables avoiding subscribers who can already be classified as churners using the definition above. Arguably, to avoid this group, it would be sufficient to set the limit in condition 1 to 59 days, but this would make the prediction task trivial. The boundary was set to 15 days as a balance between excessive reduction of the sample and limitation of the information spillover (the subscribers inactive for 30 days have already begun to display churn behavior). The second condition is useful for avoiding frequent churners (it implies loyalty) or tourists, and for avoiding bias when measuring communication with previous churners.

In summary, for the particular data set constructed, all predictors were measured in the first trimester of 2009; churn was measured two months later; inactivity at recording was limited to maximum 15 days and first activity date had to be minimum four months prior to recording. This served as baseline data set.

2.4.3 Experiment A: Addition of CEM Parameters

Capturing the service quality oriented CEM metrics was performed by using a CEM tool deployed in the company. This tool suffered from two limitations. First, it contained only 40 days of user history. In order to give the CEM parameters a fair competing chance, we only used traditional parameters with one month history. Second, this tool did not cover the entire customer experience as depicted on the CEM Framework on Figure 2.1. It is more of a network experience measurement tool due to the fact that it is focused on the more hygienic aspects of customer experience (aspects noticed by customers only if they are absent or have low quality). This database contains only the Network Usability aspect (e.g. Call Success Rate, SMS Success Rate, Call Setup Duration, etc.) and the Content aspect (e.g. count of accessing company's website for downloading ring-tones, backgrounds, etc.) of our CEM framework.

However, this tool did not capture any data related to Customer Support; therefore, Customer Support data was added from the CDRs. Other aspects of the CEM framework, such as Network Usage, Handset and Peers, which were recorded from the company's CDRs and customer subscription data, were used for benchmarking purposes (to test the added value of parameters that are viewed exclusively as CEM KPIs). Therefore, they cannot be treated as potential contributions of CEM to the model's performance, but rather as traditional parameters as described in Section 2.2. Demographic information was not available for prepaid subscribers.

The two remaining aspects, Brand and Direct Communication, could not have been analyzed because the company did not have data appropriate for data mining (Brand), or did not communicate proactively to its prepaid customers. Therefore, only three aspects were left for analysis: Network Usability, Content and Customer

Support. Hence, the only added value stemming from using CEM KPIs on prepaid churn modeling in this research can exclusively originate from one of these three aspects of the CEM framework. In order to determine the added value of CEM, we compare models consisting only of "traditional" parameters, to models consisting of both "traditional" and CEM parameters.

2.4.4 Experiment B: Variations in Population Sample

In order to test impact of the variations in the population sample we changed the activity restriction into maximum 30 and 0 days of inactivity; therefore, we change condition 1 from definition 3 into condition 1a and condition 1b, respectively.

Condition 1a: Subscribers must have at least one activity in the last 30 days.
Condition 1b: Subscribers must have at least one activity in the last day before recording.

The reasons for varying the sample on maximum period of allowed inactivity are threefold. First, our intention was to inspect the impact of these changes on model performance. Second, we wanted to test the contribution of the CEM parameters under different circumstances. Third, the typical period of user inactivity, before they become unavailable for contact and retention, is disputable. Additional motivation for experiment B can also be found in the results of experiment A.

Furthermore, using zero days of inactivity can be seen as a way to avoid information spillover. We defined churn as uninterrupted inactivity for two months. It is clear that the best predictor of this is if a user has already been inactive for a certain period. In other words, users have already started to display churn behavior. This spillover cannot happen when the inactivity period at recording is limited to zero. Additionally, these subscribers are likely to still be available for contacting. We can expect that churn is harder to predict for this subgroup of subscribers.

2.4.5 Experiment C: Change in Outcome Definition

Since there is no general consensus on a practical definition of prepaid churn in mobile telecommunications, it is reasonable to experiment with different definitions. In practical terms, this is the first question that arises during a prepaid churn modeling project. A change in the churn definition not only affects the churn rate, but also the future retainability of prepaid consumers deemed as churners using that definition. Our motivation for changing the outcome definition was also to investigate the impact of this change on the performance of the models, while keeping the same sample and data set, and test the added value of the CEM parameters in this situation. For these reasons, we introduce a so-called "grace" period of 15 days, thus changing Definition 2 into

Table 2.1: Sample size, churn rate and CoCs in experiments A, B1a, B1b and C

Experiment	Inactivity Allowed	Sample size	Churn Rate (%)	Max CoC Train Set	Max CoC Validation Set	Max CoC Test Set
A	15 days	62565	3.88	87.9	85.6	87.5
B1a	30 days	67986	6.09	89.2	89.0	88.7
B1b	0 days	32104	0.7	85.3	77.1	79.5
C (Definition change)	15 days	62565	2.4	72.7	68.6	68.5

Definition 2b: Customers are marked to have churned if they are registered to have at least one activity in the first 15 days after recording, followed by 2 consecutive months of no activity, or more.

This outcome definition also resolves the information spillover issue discussed in Subsection 2.4.2. Namely, subscribers must have an activity in the first 15 days ("grace period") after recording, thus breaking out of their already displayed churn behavior. Subscribers inactive in this grace period, who continue to be inactive in the future, are labeled as non-churners, because they have already churned at the time of recording, and are of no interest. In addition, subscribers recognized as future churners using this definition are more likely to be available for contacting and retention, because we are aiming at subscribers who are first active for 15 days, and then inactive for two months or more.

2.5 Results

In this section we report the results of the experiments of adding CEM data, changing the population based on inactivity duration and changing the outcome definition, or experiments A, B and C, respectively (see Table 2.1 for highlights). As explained in Section 2.4, we are using CoC as the main quantitative measure to compare models. In addition, we present gain charts to provide a visual reference for the performance of the models on the training set. These charts are used for illustrative purposes only, namely to visualize CoC performance through percentage of detected churners in a given percentage of the population scores (e.g. 78% of churners within top 20% of population scores).

It is worth mentioning that performance-wise, two models can be compared only if they have been created under the same experimental setup (e.g. Models A_Excl_CEM and A_Incl_CEM can be compared, while Models A_Excl_CEM and C_Excl_CEM cannot).

During experiment A, it became immediately clear that certain aspects of the CEM framework will not be able to add value in the case of prepaid churn prediction.

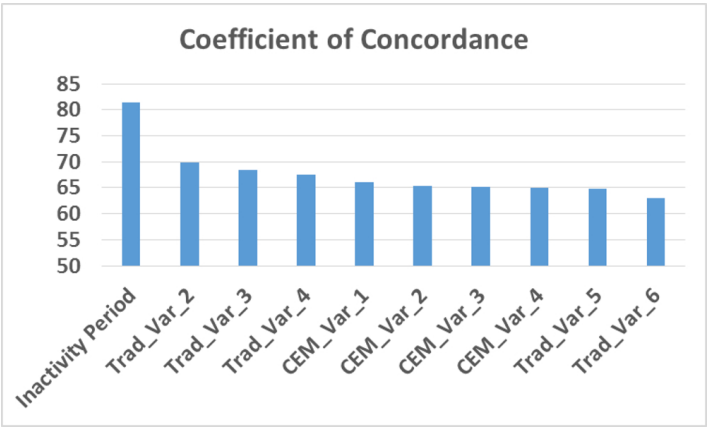

Figure 2.3: Coefficient of Concordance of predictors grouped in group 1 for experiment A

Namely, a very small percentage of the prepaid users had contact with Customer Service, or had downloaded any content from the provider's website. Therefore, none of the variables from these two aspects of the CEM Framework presented on Figure 2.1 could have been a good predictor. Furthermore, a very small percentage of these users used data services. Hence, Network Usability parameters related to mobile internet usage were also not good predictors. The only CEM predictors left to add value were the voice call and SMS related KPIs from the Network Usability of the CEM Framework. Model A_Excl_CEM contained only traditional parameters and was built on six predictors, containing only one month history. Adding parameters with longer history did not change the performance of the model substantially. The strongest predictor for churn was the inactivity period, followed by the remaining credit on the user's prepaid account. Other variables included were the handset and count and duration of calls. Model A_Excl_CEM had the following CoCs: 87.8 on the training set, 85.6 on the validation set, and 87.5 on the test set. Model A_Incl_CEM contains the same six variables, plus two CEM parameters (from the Network Usability Aspect of the CEM Framework), which were selected using a trial and error process based on their respective CoCs and the predictor group to which they belonged. Model A_Incl_CEM had the following CoCs: 87.9 on the training set, 85.6 on the validation set, and 87.5 on the test set. Thus, the results were reasonably consistent throughout the training, validation and testing sets. The only difference in the performances of these models was on the training set, which is not the best measure of model performance. Both these models were built using decision trees, but models built on the same variables using logistic regression performed just as well.

For visual reference only, we present the gain chart of the two models created for experiment A (on the training set) in Figure 2.4c. The difference in CoC of 0.1

Table 2.2: Grouping of variables of Model A_Incl_CEM

Predictors	CEM Framework Aspect	Group	Performance (CoC)
Inactivity Period	Network Usage	Group 1	81.1
Trad_var_x	Network Usage	Group 2	70.8
Trad_var_y	Network Usage	Group 2	70.6
Trad_var_z	Network Usage	Group 2	69.8
CEM_var_x	Network Usability	Group 2	68.9
Remaining credit	Billing	Group 3	67.4
CEM_var_y	Network Usability	Group 4	63.2
Handset type	Handset	Group 5	57.2

(A_Excl_CEM-87.8 vs. A_Incl_CEM-87.9) is not even visible on the gain chart. Their gain charts are identical up to 50% of the population. Both models were able to identify 78% of the churners within the top 20% of the population scores (a lift of 3.9 in the top 20% of population scores). This is representative of the test set performance as well, because the CoC on the test set of both models (87.5) is very similar.

The inability of CEM variables to substantially improve the base model is due to two reasons. First, a large number of these variables have low CoC (univariate performance). Second, even when the CoC is relatively high, in most groups there are traditional non-CEM predictors with CoC values higher than the CEM variables in the same group. Such is the case in the highest ranked group 1, depicted on Figure 2.3.

In order to illustrate the reasons for the weak effect of the CEM variables on model performance improvement, we isolated only the eight variables from model A_Incl_CEM, and ran the automatic grouping operation, with more strict grouping parameter settings than used previously, to force further grouping. The results of this exercise are presented on Table 2.2. One of the CEM parameters, CEM_var_x, has a reasonably high performance, but is in the same group as three other traditional variables, which means it has a degree of correlation with them, and has the lowest CoC in that group. This explains why CEM_var_x does not improve the model performance substantially. The second CEM variable in this model, CEM_var_y is in a class of its own, but it does not have a very high CoC; therefore, it cannot add substantial value to model performance. Please note that Table 2.2 and Figure 2.3 do not present the same groups of predictors.

Due to the strong influence of the inactivity period in experiment A, we decided to vary the population sample by changing the inactivity limit at recording to 30 and zero days (experiments B1a and B1b, respectively). Model B1a_Excl_CEM was built on seven non-CEM variables, very similar to the ones used in model A_Excl_CEM, using decision trees. B1a_Incl_CEM was built on the same seven variables plus two

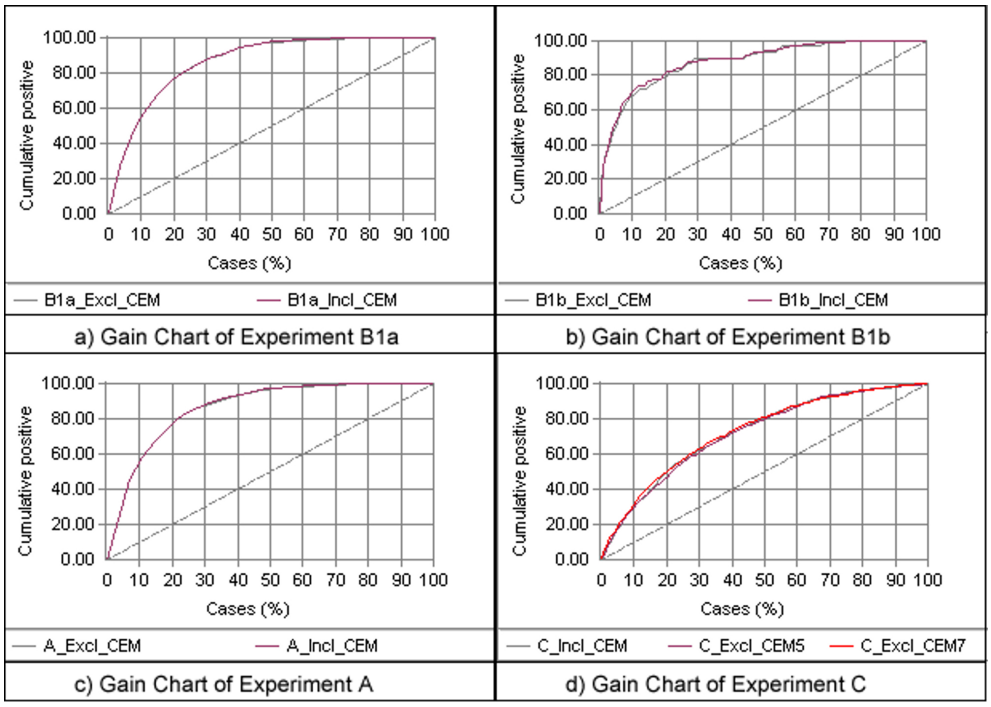

Figure 2.4: Gain chart of models for experiment A, B and C (training set)

CEM parameters (again from the Network Usability Aspect of the CEM Framework), selected based on the improvement they caused. Both models had the following CoCs of 89.2 on the training set, 89.0 on the validation set and 88.7 on the test set, which is a very stable performance. Once again, there was no visible performance difference (see Figure 2.4a). The influence of the inactivity period was even stronger, once again followed by the remaining credit. The gain chart on Figure 2.4a shows that both models identified about 78% of churners in the top 20% of the population scores.

In experiment B1b, the model B1b_Excl_CEM is built on only two non-CEM parameters (prepaid account balance and count of calls), because adding more variables caused overfitting (i.e. higher differences in CoC on the training, validation and test set). Model B1b_Incl_CEM was built on the same two variables, plus one more CEM parameter (call setup duration). This time, both models were built using logistic regression. In this situation, models built on decision trees were less stable across the three data sets, and their performance on the test set was lower, even though the performance on the training set was similar (overfitting). Nevertheless, even when using logistic regression, the models were not as stable as they had been in the previous cases.

Model B1b_Excl_CEM had the following CoCs: 84.9 on the training set, 77.1 on the validation set and 79.5 on the test set. Model B1b_Incl_CEM had the following CoCs: 85.3 on the training set, 74.0 on the validation set and 79.5 on the test set. Conclusively, adding CEM variables did not improve performance. On the contrary there is a performance drop on the validation set. Please note that the gain chart on Figure 2.4b is somewhat optimistic, because it was built on the training set. The performance on the test set is lower. Nevertheless, both models are able to identify 71% of the churners on the top 20% of the population scores in the training set, which is lower than the models in the previous experiments.

Finally, we present results for experiment C. The altered churn definition included the requirement of activity in the first two weeks of the outcome period, followed by a period of no activity of two months or more. Model C_Excl_CEM5 contained only traditional parameters and was built on five predictors. This model had the following CoCs: 72.2 on the training set, 68.5 on the validation set and 67.9 on the test set. Model C_Incl_CEM contained the same five variables, plus two CEM parameters. The CoCs of that model were 72.7 on the training set, 68.6 on the validation set and 67.9 on the test set. Both models were built using logistic regression, because these had a better performance on the test set when compared to models built on decision trees. There is an insubstantial 0.1 CoC improvement on the validation set, which is also achievable by adding two non-CEM parameters (model C_Excl_CEM7).

The gain charts of these models' performances on the training set are presented on Figure 2.4d. The maximum churners percentage achieved within the top 20% of population sample here is 50%. Note that this number would be even lower on the test set. Nevertheless, this is almost 30% identified churners less (in the top 20% of population scores) than what was achieved in experiments A and B1a.

The results of experiment C were less consistent throughout the three data sets when compared to experiments A and B1a, but somewhat more consistent when compared to B1b. It is interesting that the inactivity period was not a factor in these models, even though 15 days of inactivity were allowed at recording. The most powerful predictors were the remaining credit, the handset and the call count. Similarly to experiment A, the inability of CEM variables to improve the model performance in both experiments B and C is due to either low CoC or correlation with stronger traditional predictors.

2.6 Discussion and Future Research

We conducted three experiments in order to compare the influence of new CEM parameters, as well as changes in sample population and outcome definition, on prepaid churn model's performance.

CEM is advertised in literature to have an added value in predicting churn, but this was not the case in our experiments. Models without CEM parameters performed almost the same as models which included CEM parameters in all three

experiments. However, we only tested the CEM parameters with "hygienic" nature, which are only noticed when absent or unsatisfactory. The softer aspects of CEM remain untested. However, at this point we expect that it will be hard to find new non behavioral predictors with sufficient predictive power compared to the behavioral data. Even though the CEM data we had available contained only 40 days of history, it is unlikely that longer history would change the outcome, because the non-CEM parameters used by the models in most cases also had only one month history.

The rationale behind the low added value of CEM parameters for prepaid churn modeling may be found in several factors.

The prepaid customers themselves are the first factor. Prepaid customers that were subject to this research had average call duration of around one minute. A very low percentage of prepaid users used data services or have called the Customer Service in the research period. In other words, these events are rare, which limits the potential to become an interesting predictor. Next, prepaid users are mostly interested in controlling (reducing) their mobile phone expenses; otherwise, they could switch to using post-paid services that offer less expensive calling tariffs. The interest of prepaid users to control their mobile phone expenses may have been enhanced by the Global Financial Crisis of 2007/2008. To summarize, prepaid customers are more concerned with the quantity of experiences, which is measured by traditional predictors (Section 2.2) rather than the quality of their experiences, measured by the CEM parameters. The only exception is the handset which is a parameter that deeply influences the quality of the experiences, and is also regarded as a lifestyle product.

The second factor that could explain the low added value of CEM parameters was the high network quality. The percentage of customers experiencing network problems (negative experiences) is very small. However, the network quality cannot be seriously tested on average call duration of one minute.

The third explanation for the low added value of CEM parameters is that the quality parameters are correlated (to a degree) to their quantitative counterparts (e.g. number of dropped calls is correlated to a degree with number of calls).

Changing the population sample by varying the inactivity limit at the time of recording between 15 and 30 days also did not contribute to a substantial change in model performance. Models in experiments A and B1a had CoCs of about 88 and 89, respectfully (they identified between 78% and 79% of churners in the top 20% of the population). However, the churn rate did change drastically (there are almost twice more churners in B1a), while the sample size change was less than 10%, as presented in Table 2.1. These changes are even more evident at experiment B1b, where no inactivity was allowed at recording. Here, the sample size is twice smaller when compared to experiment A, while the churn rate is five times smaller when compared to experiment A, and even 9 times smaller when compared to experiment B1a. Due to the very low churn rate and the higher complexity of the task, the performance in this experiment was lower. The maximum CoC achieved on the test set was 79.5, which is nine CoC points less when compared to the other two

experiments; this results in a lower percentage of churners in the top 20% of the population (8% less when comparing training sets, but the difference is higher on the test set). The change of allowed inactivity period to zero also influenced the model stability. Note that we used a very small number of variables (two and three) in this experiment's models, in order to avoid overfitting. Once again it is important to emphasize that customers with zero days of inactivity at the time of deployment are more likely to be available for retention than customers with 15 or 30 days of inactivity.

The most dramatic change in model performance was shown when the outcome definition was changed and a so-called grace period of 15 days was included to mirror the inactivity period allowed at the time of recording. The model performance dropped by 20 CoC points on the test set, compared to experiments A and B1a (results in an almost 30% drop in identified churners in the top 20% percent of the sample on the training set, and even more on the test set). This does not mean that models built under experimental setup C are worse than the others. It merely implies the expected performance under such conditions. This steep decline is due to the complicated churn definition we deployed (we targeted an inactivity-activity-inactivity pattern). In this case, the inactivity period, that was a dominating variable in experiments A and B1a, was not a factor at all. The benefit of using such a definition is that upon deployment, the identified churners are likely to have an activity in the next 15 days, which makes them available for retention.

The focus of the research was on the impact of the experimental setup, rather than the algorithm used to create the model. Therefore, we used standard data mining algorithms, such as decision trees and logistic regression. Having said that, we would like to emphasize that in experiments A and B1a, there was barely any difference on model performance that could be attributed to the usage of the particular algorithms. In experiments B1b and C, there was a small difference in performance of algorithms, but it was not as substantial as changing the outcome definition or changing the allowed period of inactivity at recording to zero. In these two experiments, logistic regression had a more stable performance on the training, validation and test sets, compared to decision trees.

In terms of directions for future research, we would like to investigate a richer set of customer experience data, particularly around proactive communications and brand. Additionally, it would be worthwhile to investigate the relations between duration of inactivity and availability of subscribers for retention, by inspecting their presence on the network (regardless of making calls). Last but not least, we would like to take into account the feedback from retention campaigns, in order to focus on "retainable churners." After all, the end target of churn prediction is retention.

2.7 Conclusion

In this chapter, we presented how performance of prepaid churn models changes when varying the conditions in three different dimensions: data- by adding CEM parameters; population sample- by limiting the inactivity period at the time of recording to 15, 30 and zero days, respectively; and outcome definition- by introducing a so-called grace period of 15 days after the time of recording, in which customers must make an activity in order to be classified as churners.

Looking at the research question posed in section 2.1, the answer is clearly that changing the outcome definition has a much higher influence of the performance of the prepaid churn models than adding the CEM parameters or changing the characteristics of the sample based on inactivity at the time of recording. Adding the CEM parameters into the models did not add substantial value in terms of model performance under any of these conditions. Similarly, switching the population sample on the period of inactivity at the time of recording between 15 and 30 days did not influence model performance, only the sample size and churn rate. When we changed the population sample by disallowing inactivity at the time of recording, apart from the change in sample size and churn rate there was also a drop in performance and stability of the models. However, this drop in performance was not nearly as high as the one that occurred when changing the outcome definition by setting a grace period, thus making the behavior to be predicted more complex. This change obviously influenced the churn rate as well. Nevertheless, the latter two approaches should provide more time for retention.

Chapter 3

Combining Customer Attribute and Social Network Mining for Prepaid Mobile Churn Prediction

Kusuma, P. D., Radosavljevik, D., Takes, F. W., van der Putten, P.

Published in Proceedings of the 22nd Belgian Dutch Conference on Machine Learning (Benelearn), pp. 50–58 (2013)

Customer churn, i.e. losing a customer to the competition, is a major problem in mobile telecommunications. This chapter investigates the added value of combining regular tabular data mining with social network mining, leveraging the graph formed by communications between customers. We extend classical tabular churn data sets with predictors derived from social network neighborhoods. We also extend traditional social network spreading activation models with information from classical tabular churn models, which did improve their performance. Nevertheless, the traditional tabular churn models scored best.

3.1 Introduction

Churn, which is defined as the loss of customers to another company, is a crucial problem in the telecommunication industry. As the telecom market has matured and opportunities for growth are limited, retaining existing customers has become a higher priority. In order to minimize the churn rate, mobile telecom players have to form defensive strategies to identify and present the appropriate incentive to subscribers with high churn propensity. The conventional churn models that exploit traditional predictors, such as demographic information (e.g., age, gender or location), contractual details (e.g., package plan type, contract duration or price), usage facts (e.g., voice call duration, the frequency of sending text messages) and/or other service related information (e.g., number of interactions with customer service or number of dropped calls), are typically simple and have a good predictive accuracy (Ferreira et al., 2004; Hadden, Tiwari, Roy and Ruta, 2006). However, the predictive accuracy of these models cannot be guaranteed if there is less customer data available, namely in the prepaid segment of the telecommunication industry.

This chapter investigates the extent to which social network features derived from the graph formed by communications between customers can be exploited to improve churn prediction accuracy in the prepaid segment. Examples of such features include the number of neighbors of a customer and the number of interactions that a customer has with churned neighbors. The research question for this chapter is:

Do social network mining or attributes stemming from a social network graph add value in terms of model performance to traditional prepaid churn modeling in T-Mobile Netherlands?

This research study was conducted at one of the largest telecom providers in the Netherlands, and a data set containing 700 million call records was used to assess the quality of the various techniques discussed throughout the chapter.

We propose two novel models for churn prediction. The first is a hybrid tabular model, which combines both traditional predictors and social network features to predict churn, aiming to gain significant lift. Logistic Regression and the CHAID algorithm are utilized to derive the tabular models. These churn models, however, do not take into account the influential effect of an individual's decision to his/her social network. Dasgupta et al. (2008) have been able to address this problem by constructing a churn model based on a traditional social network mining technique,

i.e. spreading activation models. Their model propagated the negative churn influence from one subscriber to another in a cascade manner. Besides building hybrid tabular churn models using a combination of the traditional predictors and the social network features, we also propose a second approach, which extends the traditional propagation model to include the output by traditional churn models.

The rest of the chapter is organized as follows. Section 3.2 presents some related work within the field of churn prediction. Section 3.3 discusses the call graph and proposed algorithms. The research setup and the empirical models are introduced in Section 3.4. In Section 3.5, the experimental results and implications of all scenarios are presented. Finally, Section 3.6 summarizes the chapter and presents some suggestions for future work.

3.2 Related Work

Churn has been widely analyzed not only in the telecommunication industry (Ferreira et al., 2004; Hadden et al., 2006; Radosavljevik et al., 2010*a*), but also, among others, in the online gaming industry (Kawale, Pal and Srivastava, 2009) and banking (Prasad and Madhavi, 2012). Many machine learning techniques, such as decision trees, naive Bayes, logistic regression, neural networks and genetic algorithms, are often used to build the tabular churn prediction models.

Ferreira et al. (2004) utilized contractual and demographic information of a Brazilian mobile telecommunication provider to build several postpaid churn models using neural networks, decision trees, genetic algorithms and hierarchical neuro-fuzzy systems. Besides evaluating the predictive power, they also assessed the profitability value of those models, claiming that even the churn models with the worst performance are still able to save significant cost in the postpaid segment. Hadden et al. (2006) exploited provisions, complaints and repair interaction data to build churn models. They claimed that the regression tree model performed better than models built using neural networks or logistic regression. However, there is no further information regarding the performance comparison between the complaints-based model and the benchmark model based on demographic and contractual variables.

In chapter 2, we investigated the extent to which Customer Experience Management (CEM) data could improve prepaid churn prediction. Several Key Performance Indicators (KPI) of service quality combined with other subscriber data were used to train the decision tree models. Since the CEM data was always available, the constraint on lacking demographic information on the prepaid subscribers could be eliminated. Although some of the CEM variables were predictive, the empirical study showed that there was insufficient gain on this model's performance compared to the benchmark.

Several social network studies have been conducted by utilizing mobile call graph data to examine the structure and evolution of social networks (Backstrom, Huttenlocher, Kleinberg and Lan, 2006; Seshadri et al., 2008), the human mobility patterns

(Gyan, Hui, Zhi-Li and Jean, 2012) and their social interactions (Dasgupta et al., 2008). Dasgupta et al. (2008) analyzed the influential impact of the churned neighbors to their social circle by applying a spreading activation-based technique similar to trust metric computations (Ziegler and Lausen, 2004). Using call graph data, they were able to show that churn can be propagated through a social network. Although the study was limited to using social ties information only, reasonable predictive accuracy could still be achieved. Their analysis identified that the churn propensity of a subscriber correlates positively with the number of churned neighbors.

Kawale et al. (2009) conducted a similar study using social network data from a popular online gaming community. They proposed a new twist to the existing churn propagation model proposed by Dasgupta et al. (2008) by combining the social influence and user engagement in the game. The user engagement property, which refers to the length of the playing session during the observation period, can be classified as an intrinsic variable. This research showed that the models trained using a combination of social factors and this user engagement property performed better than traditional propagation models. Using collective classification techniques, Oentaryo, Lim, Lo, Zhu and Prasetyo (2012) were also able to demonstrate that the churn prediction accuracy could substantially be improved by utilizing the combination of traditional user profile and social features.

We applied ideas similar to the above mentioned works. A customer's decision to churn might not only depend on the social influences but also on how they perceive the products and services. Initially we found that the ratio of the immediate churned neighbors to the number of adjacent neighbors (degree) positively correlates to the churn behavior. When half of the neighbors have churned, the probability of a subscriber to churn is two times higher than the baseline churn rate. It implies to some extent that social behavior might have an impact on subscribers' churning decision. It could be that the hybrid models, which exploit both traditional predictors and social relationships, could outperform the simple social network and the tabular churn model built exclusively using traditional predictors. However, a question should be raised whether this adds actionable value over existing data. We suspected there may have been some element of publication bias: positive results get published more often, thus easier to find than non-significant or negative results, at least for trending topics. Hence, we decided to evaluate the business value experimentally. A call graph can be derived from raw data of communications between customers. This graph, further discussed in Subsection 3.2.1, is essentially a social network which can be leveraged in two ways. Classical "tabular" models are built on rectangular data sets, one row per customer with subscriber level information. This can be simply extended with attributes (columns) that contain information derived from the social network, as we will outline in Subsection 3.2.2. Likewise, a traditional approach to modeling social network dynamics is the spreading activation model, which can be used to model how customer behavior such as churn spreads over the network. Insights from traditional tabular models, more specifically churn scores, can be used

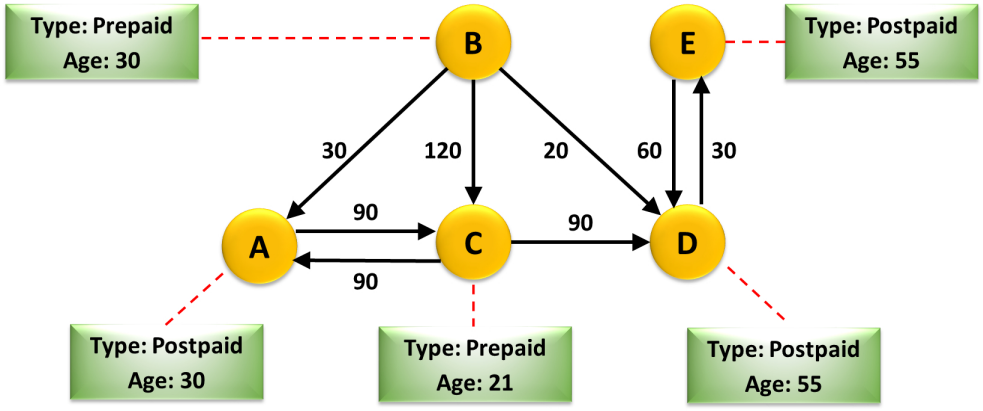

Figure 3.1: Telecom call graph.

to improve these classical social network models, a technique on which we will elaborate in Subsection 3.2.3.

3.2.1 The Call Graph

The *call graph* can be constructed from the Call Detail Records (CDRs) provided by the telecom provider. These CDRs contained detailed facts about mobile interactions, such as source phone number, destination phone number, the type of mobile communication, duration and a timestamp. This information is mapped to a directed social graph (Borgatti, 1994) $G = (V, E)$ as illustrated in the Figure 3.1.

In this call graph, *nodes* denote subscribers and an *edge* represents a mobile interaction between two subscribers. The *edge weight* can be calculated from one variable or a combination of interaction variables, e.g., voice call duration or SMS frequency. It could indicate the interaction intensity or the relationship strength between two nodes.

As several interactions could exist between the same pair of nodes, we treated duplicate edges between two nodes as a single edge, by aggregating the weight values. The aggregation method applied in this research is explained in Subsection 3.2.3.

We used Neo4j technology to store the graph structure and the content of graph elements. Neo4j differs from relational database management systems, as it is oriented to store semi-structured and network data, which makes it appropriate to store social graphs (Neo4j, 2012; Kusuma, 2013). It also provides an intuitive representation of the graph and it is easy to traverse through the graph's nodes and relationships. The scalability of this system presents a great advantage because its functionality can be easily extended to perform a large scale social network analysis.

Table 3.1: Social network features used in the extended tabular churn models.

CATEGORY	VARIABLE
CONNECTIVITY	Count of in/out-degree
	Sum & average of in-/out-weight
	Count & average of voice, SMS & voice+SMS to/from neighbors
	Total and average of edge weight*
	Total interaction frequency with neighbors*
	Total and average frequency with neighbors for voice & SMS separately*
	Degree, 2nd degree & 3rd degree count*
CHURNER CONNECTIVITY	Count of in/out-degree churners
	Sum & average of in/out-weight with churners
	Count & average of voice, SMS & voice+SMS to/from churners
	Total & average edge weight with churners*
	Total interaction frequency with churners*
	Ratio of in/out-degree churners to the total in/out-degree
	Ratio of in/out-weight churners to the total in/out-weight
	Ratio of in/out voice, SMS & voice+SMS frequency with churners to the total in/out-weight
	Ratio of churner weight to the total weight*
	Ratio of interaction frequency with churners to the total interaction frequency*
	Churner degree, 2nd & 3rd degree count*
	Ratio of churner degree to the total degree*
	Ratio of 2nd churner degree to the total 2nd degree*
	Ratio of 3rd churner degree to the total 3rd degree*

*direction is not taken into account

3.2.2 Extended Tabular Churn Models

Many tabular churn models generally exploit either subscriber profile information or social network statistics separately. The predictive power of churn models based merely on the traditional predictors might be reduced in case of many missing values. In our prepaid churn study, we only had access to limited demographic data because prepaid subscribers are not required to fill in their (accurate) personal information. On the other hand, the social network features might not be predictive enough to influence the churn decision. Neither the traditional models nor the models based exclusively on social networks can cover all aspects of churn on their own.

Therefore, we propose to combine both elements to predict churn, adding the features listed in Table 3.1.

When creating the extended tabular churn models we started with a model based on traditional predictors and added connectivity features from the social network call graph: the in-degree and out-degree, the number of second degree neighbors,

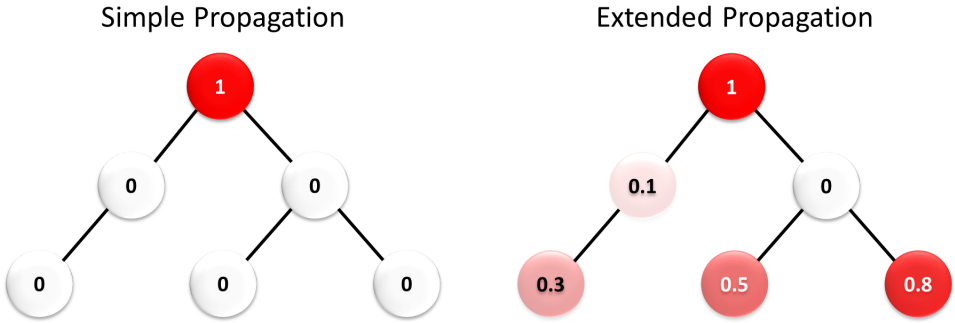

Figure 3.2: Initial energy of the simple and extended propagation technique.

sum and average of in-weight and out-weight calculated from duration of voice conversations, SMS and a combination thereof. We also added churn connectivity variables (in-degree and out-degree with churners, etc.), as well as the ratios of the total connectivity measures vs. the churners connectivity measures. A detailed overview of the added social network graph features is presented in Table 3.1. For a more detailed feature analysis, we refer the reader to Kusuma (2013).

3.2.3 Extended Social Propagation Models

In this subsection, we discuss an extension of the spreading activation model to measure how churn is diffused around telecom social network (Dasgupta et al., 2008). The churn propagation process begins by initialization of all nodes. In this study, we set the energy of *non-churners* using two different values (see Figure 3.2). For the *simple propagation approach*, the initial energy of non-churners was set to 0; for the hybrid *extended approach*, it was set to the churn score returned from the regular tabular models.

In the propagation process, for a node $x \in V$, the value of $En(x)$ represents the current amount of energy of a node, and the $En(x, i)$ represents the amount of energy or social influence transmitted to the node x via one or more of its neighbors at stage i (Dasgupta et al., 2008). After energy initialization, a set of previous churners (seeds) is activated. In stage 0, the current energy of the seeds $En(x)$ is used as initial spreading value. Therefore, the current energy value $En(x)$ becomes 0 and amount of energy in a node x at step 0 or $En(x, 0)$ becomes equal to 1.

In each consecutive stage i, the activated nodes transfer a portion of their energy to their neighbors and retain certain portion for themselves. The spreading factor $\delta \in [0, 1]$ controls the proportion of the transmitted energy, denoted by $\delta * En(x, i)$ and the amount of retained energy $(1 - \delta) * En(x, i)$. A spreading factor value of

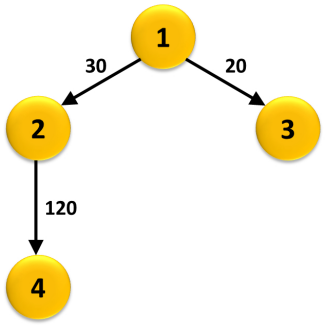

Figure 3.3: Spreading activation in a weighted graph.

$\delta = 0.8$ means that 80% of the energy is transferred to the neighboring nodes and 20% of the activated energy is retained by the node. This factor value could also be seen as a decay measure because the transferred energy will decline as it gets further away from the source. It implies that the direct neighbors will receive more influence than second degree neighbor and so on. The trust propagation study of Ziegler and Lausen (2004) has shown that people tend to trust individuals trusted by own friends more than individuals trusted only by friends of friends.

Since nodes can have multiple neighbors, the amount of the distributed energy from an active node to each neighbor depends on the tie strengths between the node pair. In Figure 3.3, for example, the amount of energy transferred from node 1 to node 2 might not be the same as the amount transferred from node 1 to node 3, because the edge weights are not equal.

Let y be a neighboring node of an active node x (with $x, y \in V$). We denote the amount of energy transferred from node x to node y in the i-th stage with $En(x, y, i)$. This amount depends on the relative edge weight of the paired nodes. This is determined by a transfer function $f(x, y)$, described in equation 3.3 below. The amount of energy transferred is then:

$$En(x, y, i) = \delta * En(x, i) * f(x, y) \tag{3.1}$$

The amount of energy of node x after the spreading computation is as follows:

$$En(x) = En(x) + (1 - \delta) * En(x, i) \tag{3.2}$$

There are multiple functions to determine the relative weight between two nodes. The simplest method is using linear edge weight normalization function (Ziegler and Lausen, 2004).

$$f(x, y) = w(x, y) / \sum_{(x,z)} w(x, z) \tag{3.3}$$

Here, $f(x, y)$ denotes the relative weight of the edge between x and y, $w(x, y)$ represents the weight of that corresponding edge, and $\sum_{(x,z)} w(x, z)$ represents the total weight of all edges connecting node x to its adjacent nodes.

We propagated the churn energy through both a directed and an undirected version of the graph. In the directed graph, energy is propagated only to outgoing edges, and in the undirected graph, both outgoing and incoming edges are used. For churn propagation, the remaining energy after termination ultimately determines the probability of a network member to churn. These churn probability scores are then distributed into score intervals. The upper interval groups contain more subscribers with high churn propensity behavior compared to the lower interval groups. Using the threshold score-based technique, the subscribers/groups with churn scores above a predefined threshold score can each be labeled as a "churner", and otherwise as a "non-churners". As an alternative, a cut-off point can also be determined by specifying the target group size.

3.3 Experimental Setup

This section describes our operational definition of churn in Subsection 3.3.1, after which the data set and weighting technique is discussed in Subsection 3.3.2. We then give an overview of the seven different scenarios that were used to construct the churn models, outlining our experimental setup in Subsection 3.3.3.

For our experiments, we used the software Predictive Analytics Director (Pegasystems, 2008) to automate variable discretization, variable selection and grouping, to train the scoring models and also to compare the models performance. The default evaluation statistic that is used to measure the performance of the predictors and models is Coefficient of Concordance (CoC) (Kendall, 1938). As explained in chapter 2, CoC measures the area under the ROC curve formed by the percentage of cases with positive behavior against the percentage of cases with negative behavior for each unique score (Harell, 2001).

3.3.1 Operational Definition of Churn

We constructed models for both prepaid and postpaid telecom segments. Although the definition of churn is different for each segment, we will only discuss the prepaid results because both studies have come to similar conclusions. Unlike postpaid subscribers, prepaid subscribers are not bound by a contract, which makes it easier for them to churn. Prepaid subscribers need to purchase a credit voucher before using any telecom service. If they do not have sufficient credit, they cannot initiate any calls, send SMS/MMS or connect to internet. They could re-enable the service by recharging or topping-up their credit.

A prepaid subscriber is disconnected from the network and he/she is marked as a churner after six consecutive months of inactivity. A prepaid activity could be translated to an outbound voice call, an inbound voice call, an outbound SMS, a data usage or a commercial voucher recharge, also known as top-up. As churn should be detected as early as possible, the disconnection date might not be an appropriate

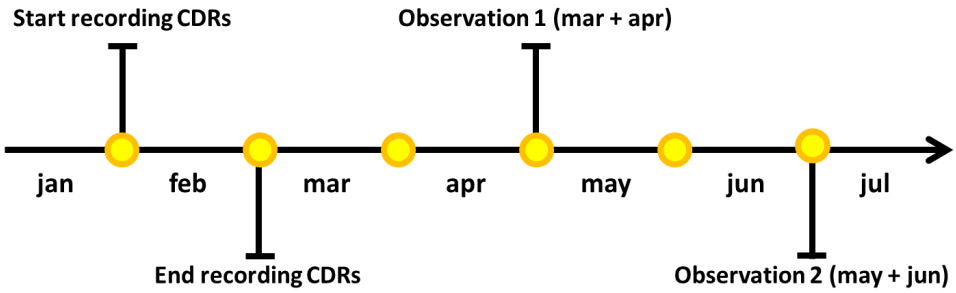

Figure 3.4: Call Graph Details.

churn date measure (Kraljevic and Gotovac, 2010). The prepaid subscribers might be long gone before they are actually disconnected from the network. Therefore, we define churn as two consecutive months of inactivity, or more. This is the same definition of churn we used in Chapter 2 (Definition 2). It was aligned with many internal studies that are conducted within the company.

3.3.2 Data Set

We used the CDRs from the whole month of February 2012, which is roughly about 700 million records, to construct the social graph. We included subscribers who have at least one call in February and we based our social network graph on the interactions that occurred in that month. The traditional predictors were also collected in this period.

We assumed that churn is also a social networking phenomenon, thus subscribers that communicated with people that have churned are more likely to churn themselves. Therefore, we labeled the nodes/subscribers that churned in the period before May 1, 2012 ('observation 1' in Figure 3.4) as seeds/churners of the propagation graph explained in Subsection 3.2.3. These are not the churners we are trying to predict.

The end goal was to use the traditional predictors as well as the social network information obtained in February 2012 to predict churn at the end of June 2012 ('observation 2' in Figure 3.4). This is a different experimental setup than the one described in chapter 2. In this research, we were trying to predict churn four months after the initial data recording. In Chapter 2, this period was shorter: it was set to two months.

In this research study, we only used the duration of voice calls in minutes and the count of text messages to construct the social graph. We could not explore mobile interactions utilizing the data connection, i.e. using over the top (OTT) services[1], due

[1] An over the top service is utilizing the telecom network to perform. However, it does not require any explicit affiliation with the network provider. Examples of over the top applications are WhatsApp, Skype or Viber.

to legal issues. Within the company, the postpaid cost of making one minute of a voice call was the same as one SMS. In the prepaid segment, one SMS was typically charged roughly the same as half of a minute of voice call. Therefore, we made the assumption that a text message is equivalent to a voice call of 30 seconds. Hence, we could generalize the edge weight $w(x, y, t)$ between a pair of nodes x and y at time t to include both types of mobile communication, voice calls and SMS and all interactions could all be measured uniformly in seconds. The identifier t represents the hourly timestamp at which the interaction starts, and is ranged from 1 until 29 February 2012.

$$w(x, y, t)' = w(x, y, t) * \begin{cases} 1, & \text{if voice call} \\ 30, & \text{if SMS} \end{cases} \tag{3.4}$$

Interactions that occurred outside working hours are assigned twice the weight to emphasize their importance. The underlying assumption here was that interactions within working hours mostly indicate communication of professional nature, whereas interactions outside working hours may involve communication of more personal nature (e.g., friends, family), which could have higher influence on the decision to churn. Motahari et al. (2012) showed that members of a family/friends social network are more likely to call each other on the weekend and the engagement ratio value within the family/friends network is at least twice as much compared to the rest of the population. Therefore, we introduce a weight scale $\rho(t)$, which is defined as follows:

$$\rho(t) = \begin{cases} 1, & \text{if t=weekdays (8-17)} \\ 2, & \text{otherwise} \end{cases} \tag{3.5}$$

$$w(x, y, t)'' = \rho(t) * w(x, y, t)' \tag{3.6}$$

We also assumed that a recent interaction should carry more weight than older ones. Therefore, the daily decay rate $\alpha = 0.2$ was manually selected. The weight value of an edge that is measured on a certain day exponentially decayed according to a predefined rate as follows:

$$w(x, y, t)''' = w(x, y, t)'' * e^{-\alpha * d} \tag{3.7}$$

Here, the symbol d corresponds to the gap measured in days between the interaction timestamp and the end of the observation period. At the end of the observation period, the weight values are aggregated. As a result, each node pair could only have maximum one edge in each direction, so two edges in total. The equation below formulates the aggregation process of the weight values.

$$w(x, y) = \sum w(x, y, t)''' \tag{3.8}$$

For an undirected graph, we could simply add up the weights for both directions together as follows:

$$w(x, y) = \sum w(x, y, t)''' + \sum w(y, x, t)''' \tag{3.9}$$

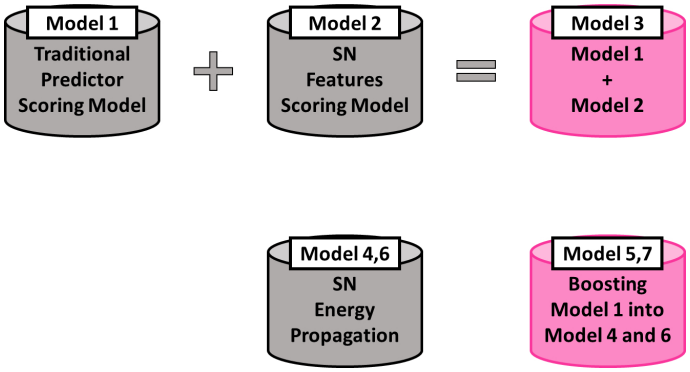

Figure 3.5: Implementation scenarios.

3.3.3 Churn Predictive Models

To investigate to which extent social network data could be used to predict churn and possibly improve churn prediction performance, we trained three tabular data mining models using scoring algorithms and four social network models using a spreading activation algorithm (see Figure 3.5).

Scoring Models

We applied logistic regression and a CHAID decision tree algorithm (Witten and Frank, 2005) to train our three scoring models:

- Model 1: simple scoring model
- Model 2: social network (SN) scoring model
- Model 3: extended scoring model

Model 1, a simple scoring model, was trained using traditional churn predictors, namely features such as prepaid credit, handset and usage information. We employed this model as the benchmark model. Model 2 was a social network scoring model, which focused solely on the social network attributes extracted from the call graph, such as the number of incoming and outgoing ties of the first and second degree neighbors. The extended scoring model, Model 3, was built by using the combined data set of the first and the second model. This hybrid model was trained using both traditional churn variables and social network features.

Propagation Models

The remaining four models were trained using energy propagation techniques based on the previously discussed spreading activation algorithm:

- Model 4: simple propagation model
- Model 5: extended propagation model
- Model 6: simple propagation model undirected
- Model 7: extended propagation model undirected

Churners in April 2012 were used as the source of the energy propagation. Each churned node was given an initial energy of 1. Model 4, which was a simple propagation model, set the initial energy of non-churners to 0. Model 5 was a hybrid model created by boosting of Model 1 into Model 4. It incorporated subscribers' intrinsic churn information into the propagation model. Instead of setting the energy of non-churners to 0, this model assigned the churn score obtained from Model 1 as the initial energy of the non-churner nodes. The intuition behind this idea is that a subscriber might already have a certain tendency to churn due to his/her experience with the provided service. Model 6 and Model 7 are similar to Model 4 and Model 5 respectively, except that those models were trained using an undirected instead of a directed graph.

The total energy value that remained after termination is assumed to be the probability of a network member to churn. To study the influential effect of churned neighbors in the social network, we then compared the propensity values of non-churners to the actual known churn class.

3.4 Results

In this section, we report the empirical result for each of the implementation scenarios (see Table 3.2 and Figure 3.6). We present and discuss only the scoring models based on decision trees, because these models had a slightly better predictive performance compared to the ones built using logistic regression. Moreover, we only include propagation models with the spreading factor that yield the best prediction results. The performance of any of these models cannot be compared with the models described in chapter 2, due to a difference in the experimental setup. As explained in Section 3.3.2, in this chapter we were trying to predict churn further into the future compared to the experimental setup in chapter 2, which makes the prediction task more difficult.

Model 3, which is the hybrid model that combined tabular churn predictors and social network variables derived from the social network graph, had the highest CoC score on the test set (64.98). Since it only slightly outperformed Model 1 (64.88), we can conclude that adding social network features on top of the traditional churn predictors did not appear to provide a substantial improvement for our scoring model. Model 2 built solely using social network predictors had the lowest predictive accuracy compared to the rest of the scoring models (56.57). By targeting the top 30% of the subscribers, Model 2 could find only 37% of the churners, while Model 1 and Model 3 were able to return about 50% of the churners. The lift chart shows that in

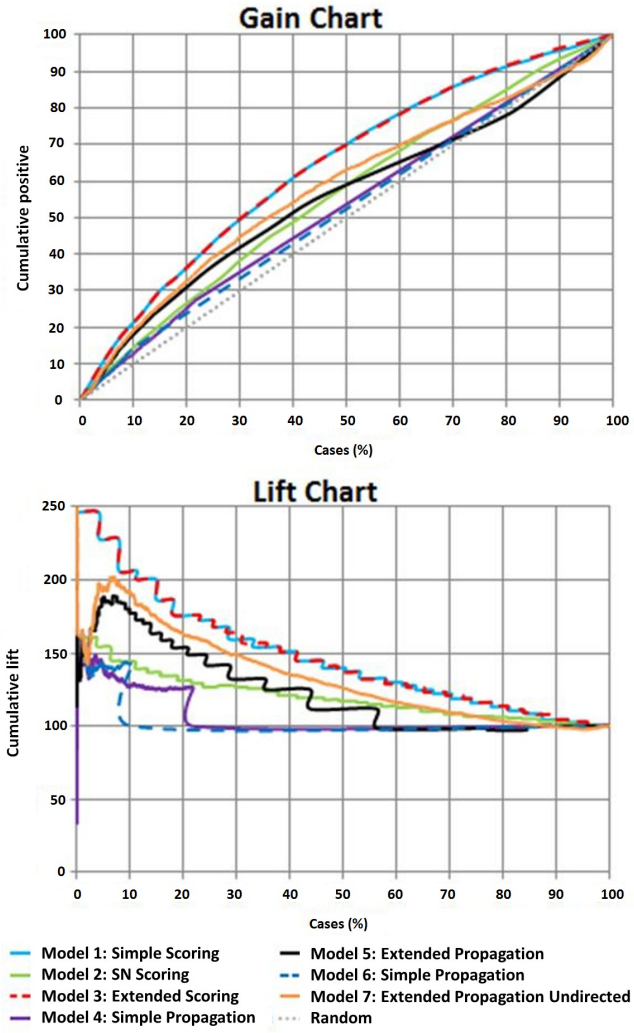

Figure 3.6: Gain and Lift chart of all models.

Table 3.2: Coefficient of Concordance of the scoring and propagation models.

Model Description	Performance on		
	Train Set	Validation Set	Test Set
Model 1- simple scoring model	65.48	64.47	64.88
Model 2- social network (SN) scoring model	57.93	56.72	56.57
Model 3- extended scoring model	65.65	64.45	64.98
Model 4- simple propagation model	53.34	53.43	53.04
Model 5- extended propagation model	55.26	54.58	55.24
Model 6- simple propagation model undirected	52.07	52.15	52.26
Model 7- extended propagation model undirected	58.39	57.66	58.30

the top 30% of the cases Model 2 had cumulative lift of 130%, whereas other scoring models had cumulative lift of 160%. In other words, the information derived from the social network was weakly predictive by itself and it failed to outperform the predictive power of the traditional predictors.

As expected, the extended propagation models (Model 5 and Model 7), which incorporated churn scores of the simple scoring model as the initial energy value in the propagation process, outperformed the traditional social network propagation models (Model 4 and Model 6). These extended or hybrid models provided better predictive accuracy than the simple propagation models for the directed and the undirected graph. By targeting 30% subscribers, Model 7 was able to correctly predict about 45% churners. It returned 5% less than the tabular churn models, Model 1 and Model 3. Although Model 7 incorporated the traditional predictor elements in the propagation process, the predictive power was still lower than that of the traditional tabular churn scoring models.

The simple propagation models that incorporated only the social neighborhood information, Model 4 and Model 6, had even lower performance compared to Model 2. Unlike Model 2, the simple propagation models used only the previous churner information within the social network without considering the individual churn propensity. This led us to believe that the churning behavior of neighbors does not have substantial influence on other members within a prepaid telecom subscriber social network. Traditional churn predictors apparently had a stronger influence on churn compared to social relationships.

3.5 Conclusions and Future work

Throughout this chapter we have investigated the extent to which social network information can be used to predict telecom churn, and how this information could potentially improve the predictive performance of conventional churn prediction

methods. We have assessed the performance of models constructed using classical tabular data mining, social network mining and hybrid models combining both techniques. The first hybrid model was built by extending traditional tabular churn predictors with social network variables extracted from the social graph. The second hybrid model was obtained by incorporating the results of a traditional tabular churn model into the social propagation graph, using them as initial energies of the non-churner nodes.

The performance of our models was verified using a large data set of 700 million call data records. Our initial observation showed that the churn probability was positively aligned with the number of churned neighbors. The regular tabular churn models constructed exclusively using social network information and the traditional social network models scored the least. This indicated that social network information alone is not sufficient to predict churn. Overall, the traditional tabular churn models had the best predictive performance. The added value of the social network variables to the tabular churn models was rather minimal. Although the second hybrid models were able to outperform the regular propagation models, they still could not beat the performance of the traditional tabular churn models. The contribution of traditional predictors to churn prediction was substantially higher than that of the social network behavior. Moreover, the performance gain of both hybrid models was not substantial enough to justify the computational costs. In a nutshell, the answer to the research question posed in section 3.1 is that social network mining and attributes stemming from a social network graph did not add substantial value in terms of model performance to traditional prepaid churn modeling in T-Mobile Netherlands. This was in in contrast to most of the statements made in research literature, but it did not come as a surprise to us because we were suspecting that in the other cases the models might have not been benchmarked well enough against standard models based on rich data, and that there may have been some instances of publication bias.

The current research study only explored the negative influential effect of previous churners within the social network. Future research could potentially be focused on removing this limitation. The influences from both churners and non-churners could be taken into account, as subscribers might spread messages based on how they perceive the product/service quality. Assuming bad news can have a stronger influential effect than good news, positive influence from non-churners to stay within the network might not be as strong as negative influence from churners. Since our energy propagation model was purely derived from node and neighborhood-based relationships, the spreading activation computations were done locally and subscribers did not have knowledge beyond their direct neighbors. Other algorithms, for example from the field of community detection, are capable to identify the role of subscribers within the social network, such as influencer or adopter. Rather than targeting all future churners, we can minimize our resources by focusing only on churners with high influential power.

Chapter 4

Preventing Churn in Telecommunications: The Forgotten Network

Radosavljevik, D., van der Putten, P.

Published in International Symposium on Intelligent Data Analysis, Springer Berlin Heidelberg, pp. 357–368 (2013).

This chapter outlines an approach developed as a part of a company-wide churn management initiative within T-Mobile Netherlands. We are focusing on an explanatory churn model for the postpaid segment, assuming that the mobile telecom network, the key resource of operators, is also a churn driver in case it under-delivers to customers' expectations. Typically, insights generated by churn models are deployed in marketing campaigns; our model's insights are used for network optimization in order to remove the key network related churn drivers and therefore prevent churn, rather than cure it. The insights generated by the model have caused a paradigm shift in managing the network of T-Mobile Netherlands.

4.1 Introduction

The phenomenon of churn, which denotes loss of a client to competitors, is a key problem across industries. New customers are difficult to find, especially in saturated markets, such as the European mobile communications market. Furthermore, it is far less expensive to retain existing customers than to acquire new ones. Retention is usually a process that identifies customers that are likely to churn, using various predictive modeling techniques, followed by approaching these customers with suitable offers that would persuade the customer into extending the contract. But, can the customer be prevented from even wanting to churn? Can the main churn drivers be mitigated beforehand?

This chapter is focused on a company-wide churn reduction initiative conducted in T-Mobile Netherlands. As explained above, churn/customer retention is typically a marketing based process. But, despite involving predictive analytics, this process is in its nature reactive, because the customer has already decided to churn and an action is being taken to stop this.

In this research we are taking a completely different approach: the model generated here is not to be used for campaigning. Our method attempts to tackle churn by identifying the key reasons why customers decide to churn in order to alleviate them, rather than identify prospective churners. Hence, the research question in this chapter is:

As a different method for model deployment, can a churn model be used to prevent churn by explaining its causes as opposed to using the predictions for targeting customers?

This approach is even more justified taking into account the current and future stringent European Data Privacy regulations, which limit operators' use of customer data for campaigning purposes. This is especially the case with Internet usage data.

The mobile telecommunications network is a key resource for telecom operators. It is the means of service delivery as well as the most frequent touch point with customers. Problems with ability to use the network (services) have been identified by surveys internal to the company, as well as in literature (Section 4.2), as one of the key reasons to churn. But, most of the time, customers are not experts and cannot pinpoint what exactly is going wrong. Most of the time, this is generalized

as "coverage problems". This research is taking a deep dive into various network problems and their relation to customer churn. The main objective here is to identify the problems that customers who have churned were experiencing, so that they can be corrected for the current customer base and reduce their likelihood of churn. In other words, rather than treating symptoms, we are treating the cause of the disease. This research and its outcome have caused a paradigm shift in managing the network with the operator where the research was conducted.

As opposed to chapters 2 and 3, in this research we are focusing on the postpaid customer segment. Even though these customers are bound by contract, which makes the task of churn prediction slightly less challenging, the revenues that are typically generated here are much higher than in the prepaid segment. Furthermore, postpaid customers' service usage is much higher, compared to the prepaid segment; therefore they would be more prone to experiencing network related issues which can potentially lead to churn. The combination of higher usage and revenues makes it easier to justify the network investments needed to remedy their problems.

The rest of the chapter is structured as follows. Section 4.2 describes the related work on telecom churn. Section 4.3 discusses the data set and methodology we used. Section 4.4 contains the results, their application. Limitations and future work are discussed in Section 4.5. Finally, we present our conclusions in Section 4.6.

4.2 Telecom Churn in Literature

Churn in various industries has been a growing topic of research for the last 15 years (Verbeke et al., 2011). According to Ballings and Van den Poel (2012), churn management consists of predicting which customers are going to churn and evaluating which action is most effective in retaining these customers. Retention strategies are in the focus of Hung et al. (2006). However, most often churn prediction and improving model performance were analyzed following one of these two strategies: adding/improving the data to mine and inventing new algorithms or improving the existing ones (Ballings and Van den Poel, 2012).

The remark above is certainly valid in the case of telecom churn literature. Many papers were trying to find the best algorithm that would outperform all others. Multiple works have examined Logistic Regression, Decision Trees, Neural Networks, evolutionary learning, discriminant analysis and Bayesian approaches (Au et al., 2003; Mozer et al., 1999; Neslin et al., 2006; Wei and Chiu, 2002). Other papers analyzed Support Vector Machines, Random Forest, Rotation Forest, Bagging and Boosting (Archaux et al., 2004; Idris, Khan and Lee, 2013; Lemmens and Croux, 2006; Verbeke et al., 2011). In our view, the value of this research is somewhat limited, at least for real world data mining, given the No Free Lunch theorem (Wolpert and Macready, 1997). In the period after 2008, an overwhelming theme in (telecom) churn research was Social Networks Analysis (SNA), claiming to largely improve on existing churn models (Dasgupta et al., 2008; Motahari et al., 2012; Nanavati et al.,

2006; Richter et al., 2010; Wang et al., 2010; Ngonmang, Viennet and Tchuente, 2012; Polepally and Mohan, 2012; Saravanan and Raajaa, 2012). However, as explained in chapter 3, this claim is not generally applicable, at least not in prepaid churn prediction on a European market. Most of the SNA research focuses on the Asian or US Markets.

Taking into account the data perspective, most of the literature, especially the one focusing on SNA, was using features extracted from Call Detail Records (CDRs). Contractual, demographic, billing, handset, customer service, market (competitor's offers), and customer survey data is often used in addition to CDRs (Archaux et al., 2004; Au et al., 2003; Hung et al., 2006; Mozer et al., 1999; Neslin et al., 2006; Wei and Chiu, 2002). Just a few of these papers take into account any network usage related problems as possible factors affecting churn. For instance, dropped calls were considered as potential churn influencers (Ahn, Han and Lee, 2006; Mozer et al., 1999). Service quality in general and innovativeness were marked as churn detractors by Malhotra and Malhotra (2013).

Predictive models trying to explain churn have not received as much attention in literature (Ballings and Van den Poel, 2012; Lima et al., 2009). Nevertheless, there are studies in industries other than telecom illustrating the need to gain insight into causes of churn (Anil Kumar and Ravi, 2008; Buckinx and Van den Poel, 2005). Furthermore, research based on customer surveys claimed that network coverage, mobile signal strength and voice call drops are reasons for customers to churn (Ahn et al., 2006; Birke and Swann, 2006; Malhotra and Malhotra, 2013; Min and Wan, 2009; Seo, Ranganathan and Babad, 2008). However, all these papers were based on survey data, thus perception of quality, not actual network measurements.

It is apparent that in most recent telecom churn research the physical telecom network, which is the means of delivering telecom services, has been largely neglected. At best, the (lack of) quality of voice call usage is considered. To the best of our knowledge, there is little or no research on how Internet usage on a mobile network and its quality parameters might affect churn. This is one of the key reasons why the topic of our research is an explanatory churn model for telecommunications with actual network quality usage parameters at its focus, instead of just the customers' perception of network quality. In addition, this model, unlike the models from related work, is not meant for retention campaigns; instead, it is concentrating on eliminating what we see as one of the crucial causes of telecom churn: poor experience using the services on the network.

4.3 Data Set and Methodology

In this section we will describe the process and the data set used in this research. As mentioned previously, this research was not started with retention campaigns in mind. It was a part of a cross departmental company-wide churn tackling initiative, executed in parallel with regular churn campaigns. Therefore, the objective of this

Table 4.1: List of contractual, demographic and CDR based features

Contractual and demographic features	Features Extracted from CDRs
Contract expiry	Amount of Voice Calls, SMS and
List of services/ products used	Internet Volume (MB) used,
Subscription fee	both local and roaming
Monthly Bill for each of services	Breakdowns of Voice Calls and SMS
Age, gender, zip code	onto national-international,
Handset	internal-external(competitors' network)

research was not to compete with churn models created for campaigning, but to detect whether there are telecom network quality related factors influencing churn and identify potential remedies.

4.3.1 Data Set

The results presented here are based on a random sample of 150,000 consumer post-paid subscribers of the operator from September 2012. This is just a fraction of the overall base. There was a limitation enforced on the data set related to contract expiry date: the sample was limited to subscribers whose contracts were expiring in three months or have already expired; thus only customers at risk of churn were taken into account. Churn was measured for the following two months, October and November 2012, combined.

The final data set consisted of 750 features, gathered by merging tables from CRM and Network databases. In addition to the attributes similar to what was described in Section 4.2 (see Table 4.1), we added features stemming from the CEM (Customers Experience Management) Framework (Radosavljevik et al., 2010*a*) we designed in Chapter 2 (see Figure 2.1), most importantly the Network quality or usability features (see Table 4.2). The features extracted from CDRs and the network quality features represent monthly aggregates. We also examined their respective three-month aggregates, as well as if there is a rising or declining trend in the past three months for any of these features and used these as potential predictors of churn.

4.3.2 Methodology

Our research setup is similar to what we have described in chapter 2. The data originally residing in various CRM and Network quality databases was collected into a single Oracle database (Oracle, 2011), which allowed easier manipulation and data cleansing. For Data analysis, Predictor Selection and Model Development and Assessment we used the commercial tool Predictive Analytics Director (Pegasystems, 2008).

Table 4.2: List of network quality features per category

General Network Quality	Voice and SMS quality	Internet quality
2G and 3G Coverage at home	Count of Dropped Voice Calls and SMS	3G and 2G Data Attempts
	Voice Call Setup Failures	3G and 2G Data Errors
Provisioning Errors	Voice Call and SMS drop rate	3G and 2G Success Rate
	Voice Call Setup Duration (Maximum and Average)	Ratio of 3G usage vs. 2G usage

We divided the sample into training, validation and testing set using the ratio 50:25:25. The validation set was used during the data analysis stage as a "pre-test" set, in order to verify the univariate performance of each predicting variable with relation to churn, established on the training set.

The performance measure used to evaluate the performance of each individual predictor, as well as the models, was Coefficient of Concordance (CoC). As explained in chapter 2, CoC is a rank correlation measure related to Kendall's tau, suitable for evaluating scoring models (Kendall, 1938; Pegasystems, 2008). It is a measure equivalent to the Area under the ROC (AUC). One interpretation of the CoC measure is that in a scoring model it gives the probability that a randomly chosen positive case will get a higher score than a randomly chosen negative case. The CoC value ranges from 50 to 100. Random choice has a CoC value of 50.

All models developed are scoring models, i.e. we calculated probabilities that someone will churn, without setting a cutoff point. As mentioned above, these models were not to be used for campaigning, but for network improvements, therefore setting a cutoff point to strictly classify whether an instance is a churner or not was not necessary. For this reason, using measures such as recall and precision were not applicable in our case.

During the data analysis stage, the continuous variables were discretized into bins. Bins without significant performance difference are then grouped together. Basically, this is a supervised, bottom-up approach to discretization of continuous variables. One of the advantages of this approach is that it can address non-linear effects of variables onto churn: namely, each separate bin got a score which is concordant to churn and this score was used for modeling. This process is similar for symbolic variables. Variables can be inspected via histograms and the discretization settings can be manually changed if deemed necessary. The next step in the process is predictor grouping which assists feature selection. Namely, variables that are correlated to each other are grouped together. A given predictor may have a high univariate performance, but also be correlated with other candidate predictors that are even stronger, hence not adding value to a model. We first used the best predictor of each group and then selected/deselected variables manually to develop the models

Table 4.3: Model Performance

Model Description	Number of Predictors	Performance on Training set (CoC)	Performance on Test set (CoC)
Campaign	3	76.0	75.9
Campaign_PlusNetwork	6	76.8	76.7
ContractEnd_PlusNetwork	5	75.1	74.7
Campaign_MinusContractEnd	5	68.7	68.1
PurelyNetworkBased	5	66.6	66.5

with a good performance, but also good explanatory value.

As explained previously, the topic of this research is not finding the next best algorithm. That is why we used standard algorithms, such as Logistic Regression and Decision Trees based on the CHAID splitting method (Witten and Frank, 2005). These methods also perfectly fit the explanatory nature of our research, because they are easy to interpret. This is an advantage in commercial settings, where people that need to make investment decisions based on the model and implement its results are not data miners.

The modeling process resulted in scoring models: each instance is allocated a rank score concordant with the probability of being a churner. The CoC (AUC) measure was used to measure model quality. In addition, we use gain charts as visual representation of model performance. On the y-axis, these charts show the captured proportion of the desired class (i.e. churners in selection divided by total number of churners) with increasing selection sizes (x-axis, from highest scoring to lowest scoring) (see Figure 4.1).

4.4 Results, Application and Discussion

Even though optimizing model performance is not the topic of this research, we deem it necessary to benchmark our network against the campaigning model. The performance (CoC) of the models we created is presented on Table 4.3.

It is worthwhile mentioning that all models presented here were built using Decision Trees with CHAID splitting criterion, which have an inherent characteristic of dealing with non-linear data. We also tested models using Logistic Regression, but they had somewhat worse performance (0.5 CoC points). Please note that due to the discretization process described in Subsection 4.3.2, this implementation of logistic regression is able to handle non-linear dependencies too. Furthermore, in order to test for non-linear interaction effects between a combination of two variables and churn we created close to 280,000 new predictors using two way combinations of all of the 750 variables. However, no strong non-linear effects were noted.

As can be seen on Table 4.3, adding network related features to a campaigning

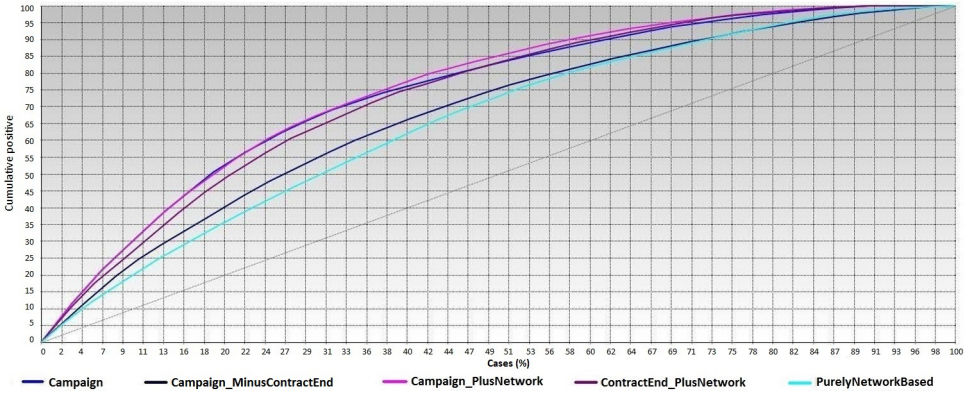

Figure 4.1: Gain Charts of Models Used

model (model Campaign_PlusNetwork) only marginally increases performance (1 CoC point), visible on Figure 4.1 only after the 40th percentile of cases ranked by churn, which confirms our result from (Radosavljevik et al., 2010*a*). However, campaigning wise, this has no meaning because rarely do campaigns address more than 40% of the base that is at churn risk.

The PurelyNetworkBased model, which is the topic of our research, had the weakest performance. Nevertheless, just for comparison reasons, we built a Campaign model without the strongest predictor - Contract End (Campaign_MinusContractEnd) and a model based on a combination of just the Contract End and Network Factors (ContractEnd_PlusNetwork). The Campaign_MinusContractEnd model performed only somewhat better than the Pure Network model (1.5 CoC on the test set in Table 4.3, or 5% more churners in the Top 20% of the scores on Figure 4.1), and the model ContractEnd_PlusNetwork performed only marginally worse than the campaigning model (1.3 CoC on the test set in Table 4.3, or 4% less churners in the Top 20% of the scores on Figure 4.1). The conclusion here is that, less the Contract End variable, the network quality parameters from our Purely Network Based model performed nearly as well as the other predictors.

However, performance was not the main topic of our research. The main aim was the explanatory value of our model. On Figure 4.1 it is shown that Purely Network Based Model could address the 35% of churners in the top 20% of scores, while the Campaign model addressed nearly 55% of all churners in the top 20% of scores. This may be interpreted as the Network factors being "responsible" for the 35% out of 55% of churners in the Top 20% of all scores and that correcting these parameters would mitigate at least a part of them[1]. The rest of the churn (the other 20%) is due to other reasons, e.g. a better competitor offer. Having this in mind, it was worthwhile analyzing the parameters that constitute this Purely Network Based model.

[1]In retention campaigns too, one cannot expect 100% acceptance rate

Table 4.4: Univariate performance of predictors (CoC)

Variable	Performance (CoC)
Contract End Date	73.1
Total Duration of Provisioning Errors in the past six months	62.5
Average Ratio of 2G and 3G Data Events in the past three Months	59.2
Count of 2G Data Events in the past three Months	57.5
Sum of Call Drops and Call Setup Failures in the past three months	56.8
Average Voice Call Setup Duration for the past three months	52.4

Due to confidentiality reasons we cannot disclose the exact numbers and weights of the parameters constituting our model. Nevertheless, we can disclose parameters of which our Network model was consisted, ranked by their individual performance (CoC): The Total Duration of Provisioning Errors in the past six months; The Average Ratio of 2G and 3G Data Events in the Past three months; The Count of 2G Data Events in the past three Months; The Sum of Call Drops and Call Setup Failures in the past three months and The Average Voice Call Setup Duration for the past three months. The individual influence (CoC) of each of these parameters onto churn is presented on Table 4.4. Just for comparison, we also show the performance of the best predictor, the contract end date, which has a superior prediction power. However, the purpose of these models was to investigate why customers churn from a network perspective and offer means of alleviating these reasons. In this case, the relationship with contract end date is secondary. When customers get closer to the end of their contract, there is a higher risk of churn. Moreover, customers out of contract for a longer period of time have proven to be loyal, as the other customers have left.

The influence of each of these parameters onto customer experience and therefore churn could be explained and was agreed upon by the company experts. First of all, it is interesting to note that the Sum of Call Drops and Call Setup Failures in the past three months was not a rate, but an absolute count. Namely, it was irrelevant if a customer dropped two calls out of 30 or out of a 100, the two dropped calls drove churn. The parameter Average Voice Call Setup Duration for the past three months implied that customers did not appreciate having to wait a long time to establish a voice call. Provisioning errors are errors where customers have not been enabled to use certain services on the network even though they have subscribed for them (e.g. not being provisioned to use Internet), or did not get the appropriate quality of service (e.g. being provisioned to used Internet at 1 Mbps when subscribed to 3 Mbps). These errors did not occur frequently but were deemed by experts to have a severely negative influence onto satisfaction even if they occurred once during the contract duration; therefore we summed up six months of these errors' history. It is interesting to see the growing influence of mobile Internet services onto churn, especially the

strong preference of customers to use the 3G network, which is by design much faster than the 2G network[2]. The low 2G speed was not deemed satisfactory, it could have been in fact perceived by the customers as not being connected at all. The influence of quality of Internet services onto churn was represented via the Number of 2G Data Events and the Ratio of 3G vs. 2G Data Events.

The added value of these parameters was that they denoted clear guidance for the technology department on which actions to take in order to prevent churn. Determining the exact thresholds of each parameter that led to churn was done by using the discretized variables. As explained in the methodology subsection of this chapter, each variable was separated into bins and each separate bin has gotten a score which is concordant to churn. Next, we were looking for thresholds in these parameters that, once crossed, pointed to higher churn probabilities. For example, let us assume that customers having four or more dropped calls in one month are two times more likely to churn than customers with less than four dropped calls in the same period: This would set the threshold of dropped calls per customer per month to four. This is just a theoretical example, as we are not at liberty to disclose the real figures.

Projects have been developed to maintain and correct these parameters and their respective critical values (increased churn risk thresholds). This also had a profound effect onto the mindset of the department maintaining the network: their focus has shifted from a network centric approach to a customer centric approach in managing the network. We will explain what this means using the example of Voice Call Drops. The network centric approach in managing this key performance indicator would be to just measure a network wide call drop rate and attempt to maintain it above a certain threshold by giving priority to fixing network sites with a large number of dropped calls. The customer centric approach in managing this parameter is to monitor the number of customers experiencing dropped calls and giving highest priority to network sites where most customers experience dropped calls. The customer centric approach allows addressing the problem of a higher number of customers, rather than focusing on network sites where only few customers experience a large number of dropped calls. It has already been implemented and has helped reduce the number of customers experiencing dropped calls in general, which resulted in improved satisfaction in customer surveys (internal to the operator), implicating that churn reduction should follow. Similar approaches were developed to address the other parameters from our model. Last but not least, the technology department in the company has set customer centric targets for managing the network. Using the theoretical example from the paragraph above, this would mean that the department would have set a target that no more than a small percentage of the base (e.g. 1%) should experience four or more dropped calls per month. It is worthwhile mentioning that the amount of customers which was dissatisfied with the operator's network

[2]3G networks could reach throughputs/Internet speeds of 21Mbps, while for 2G the maximum speed was only 64 Kbps

(according to customer surveys) was reducing in parallel with the reduction in the amount of customers experiencing the targeted number of dropped calls and similar customer centric network related targets.

It is possible that the solution applied to a given network site to reduce the number of customers experiencing dropped voice calls may also influence some of the other quality parameters, especially in a case of a 3G network site (e.g. increasing the coverage area or adding extra capacity to a 3G site might reduce both the number of customers experiencing dropped calls and prevent them from falling back to a neighboring 2G cell when using Internet). As an extension of this approach, it can be envisioned that sites where a high number of customers that are already at churn risk experience dropped calls are given priority, but this is subject to legal limitations with regard to data privacy[3].

To summarize, even though our churn model based entirely on network quality parameters had lesser performance compared to a normal campaigning model, it did have many other advantages: it addressed churn in a preventive manner, as it was not necessary to run retention campaigns with it; it provided guidance on which were the critical network parameters that needed to be corrected in order to address churn from a network perspective; and it created a mind-shift in the department managing the network into a customer centric perspective, which already resulted in increased customer satisfaction.

4.5 Limitations and Future Work

The first limitation we would like to address is the lack of coverage data per customer. We were only able to calculate (not measure) the coverage at home for each customer. Loss of coverage for each customer is impossible to measure from the network side. Having adequate coverage information could have improved our model. However, the Ratio between 2G and 3G data events does imply the influence of loss of 3G coverage or insufficient 3G capacity in certain areas onto churn.

Other limitations of this research are of legal nature. Namely, in most European countries stringent Data Privacy or Net Neutrality Laws (will soon) exist. This makes it impossible to look into individual consumption of different types of Internet use (e.g. browsing, streaming, messaging, VoIP etc.), which could provide even better insights into what type of service degradation leads to churn.

Next, as usage patterns change, so do the expectations from the service quality that the network provides. Therefore, in time we expect a change in the influence on churn of the various factors that we discussed which makes the model outdated. This will especially be the case after the introduction of 4G (LTE) networks, which allow much faster Internet speed (throughput). However, these issues can be addressed by remodeling.

[3]It involves storing the sites/locations of particular customers

As future work, we would like to go one step further, and investigate the benefits network experience measured directly on the phone, via a preinstalled app, of course with customers' permission. We believe that this would provide a 360 degrees view of customers' network experience and close the gap created by the data that is difficult to obtain due to technical or legal limitations. Measurements taken directly on the phone are the ultimate determinant of customer's network experience.

4.6 Conclusions

In this chapter we presented an atypical approach to churn management in commercial settings. We succeeded in explaining at least a part of churn via actual measurements of network quality. The main benefits of our approach are the following: First, we managed to build an explanatory churn model by sacrificing only a part of the performance. Second, our churn model was based on features that are extracted from actual network parameters rather than surveys (real network experience vs. perception). Third, this model generated insights on which network parameters are necessary to be corrected in order to reduce churn, which is a new way of churn reduction. The insights generated caused a shift from network centricity towards customer centricity in managing the telecom network. Using this approach, the churn mitigation process is no longer just a retention campaign: the churn efforts are no longer the responsibility of just the CRM teams, Marketing and Customer Service, but also the Technology department, which is responsible for the mobile network is involved. Referring to the research question stated in section 4.1, we have managed to use a different deployment form of a churn model in order explain and prevent churn rather than directly target customers.

Our research was deployed in T-Mobile Netherlands, part of one of the largest European telecom operators. It was used for setting department targets for managing the network and has already contributed to increased customer satisfaction, implicating that churn reduction should follow. In addition, another national operator from the Deutsche Telekom group has used the same approach.

Last but not least, we would like to point out the possibility of applying our research onto domains other than mobile telecom. Obviously, this approach can be mirrored onto fixed telecommunications and potentially into churn in other industries, but also in many other cases where prevention is more important than the cure, like certain medical research.

Chapter 5

Large Scale Predictive Modeling for Micro-Simulation of 3G Air Interface Load

Radosavljevik, D., van der Putten, P.

Published in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1620–1629 (2014).

This chapter outlines the approach developed together with the Radio Network Strategy and Design department of T-Mobile Netherlands, part of the Deutsche Telekom group, in order to forecast the Air-Interface load in their 3G network, which is used for planning network upgrades and budgeting purposes. It is based on large scale intelligent data analysis and modeling at the level of thousands of individual radio cells resulting in 30,000 models produced in one day. It has been embedded into a scenario simulation framework that is used by end users not experienced in data mining for studying and simulating the behavior of this complex networked system, as an example of a systematic approach to the deployment step in the KDD process. This system was a part of a standard business process in T-Mobile Netherlands for more than two years. This operator became a competence center for predictive modeling for micro-simulation of 3G air interface load for three other operators of the Deutsche Telekom group. A similar approach, based on different network parameters was also developed by the operator for 4G networks.

5.1 Introduction

This chapter reports on a deployed data mining application that has been developed by one of the largest European telecom operators and has been in continuous use for more than two years. In order to accommodate the continuing strong increase of mobile internet traffic, the operator's Radio Network Department had to continuously monitor and upgrade the 3G Radio Access Network. This required an Air-Interface load forecast for every radio cell in the network, including indications of denial or interruption of delivering service. However, such a detailed forecast was not readily available. Furthermore, there was a need to simulate different scenarios for different parts of the network. Given the complexity of the problem, the dimension of the network and the repetitiveness of the task, a manual approach was out of the question. Hence, the research question for this chapter is:

How can data mining be used to predict 3G mobile network interface load and simulate it under different scenarios?

In this chapter we present a fully automated approach that generates multivariate linear regression models on a grand scale, using primarily open source tools. The key business benefit of this research is that it solved a very complex and high impact business problem that could not be approached by using general planning approaches.

Traditional approaches for mobile network load forecasting have a number of practical issues. These are most often analytical or Monte Carlo based approaches (Mäder and Staehle, 2004). The load formula used is typically a general purpose analytical model, derived from physics knowledge and theory rather than from modeling on actual data, let alone being based on data from a specific operator. Furthermore, the inputs required are difficult to measure and forecast, or do not relate to changes in customer behavior such as call usage, which is easier to understand and

use for scenario purposes. Our approach makes it easier to translate customers and their usage into network load. To our knowledge, this is the first time a data mining approach has been used for Air Interface load prediction of a 3G mobile network.

In terms of business benefits, the exact return is confidential, but cellular network infrastructure forms a major part of an operator's investment budget, and this is a key system for tactical and strategic network investment decisions. In the group where this operating company belongs, up to 50% of wireless CAPEX investments are going into the radio access. For reference, operators worldwide invest more than 20 billion USD into cellular network infrastructure. Our methodology is first and foremost intended to ensure that capacity is added in time and at the right place, thus avoiding inefficient investments and poor customer experience due to traffic congestion, which can ultimately lead to churn. Last but not least, Internet access is recognized as a right by law in several countries, as a part of the rights to Freedom and Expression of Opinion. By adding capacity at the right places, operators provide a valuable social service, given the growing importance of communications and social media in everyday life. This is a data mining application in telecommunications that does not raise the usual privacy concerns; on the contrary it serves a social function by mediating high quality internet access.

An early version of our approach has been published in Radosavljevik, van der Putten and Kyllesbech Larsen (2012). Since then, our system has been rolled out in full use in T-Mobile Netherlands, the operator where it was developed. None of the other operator companies in the Deutsche Telekom group used a similar fine grained approach. Therefore, a more universal approach applicable to the other operators, too had to be developed. This involved dealing with complexities such as different network equipment vendors using different performance management systems and lack of certain measurements we have introduced in Radosavljevik et al. (2012). Nevertheless, the results of our new approach were very positive. Hence, this operator became a competence center for predictive modeling of 3G Air Interface load and it was performing this task for operators of the same telecommunications group in four countries.

Whilst the core intelligent data analysis algorithms used were not novel, we applied these on a large scale by modeling individual radio cells across a variety of dimensions (Section 5.3 motivates why we modeled at cell level). This has also been embedded into a simulation framework targeted at non-data miners using tools they are familiar with to enable them to run low level simulation scenarios. Hence, the goal is to provide a case example of an embedded, deployed intelligent data analysis system, dealing with real world aspects such as scale and having major business impact. Extensive simulations have been carried out by the operating companies using the system, and also novel use cases for scenario simulation analysis have been developed and applied.

As discussed, the technical novelty is not determined by the complexity of the base estimators used. We used simple linear regression models as data inspection

has shown that the behavior to be predicted is primarily linear, and experiments confirmed that complex algorithms actually performed worse given the high variance associated with these models. This is not uncommon in real world data mining problems (van der Putten and van Someren, 2004). What makes this problem out of the ordinary is the massive number of models. For each of cell in the network we create four models to predict different kinds of outcomes, resulting in a total of 30,000-100,000 models, depending on the amounts of cells in the network. Model parameters are estimated using ten-fold cross validation, which increases the number of models estimated to over 1 million. This process is repeated on a regular basis, given that the customer base and behavior, as well as the cellular network itself change constantly.

Finally, we did not just deploy the forecasted loads. The underlying regression formulas were provided by the data miners to the end user analysts as simple spreadsheets, which enabled them to tune various simulation and forecasting scenarios without further involvement from the data miners. This turned out to be not just a practical benefit, but a major opportunity for the business as a range of simulation use cases were explored that were not envisioned by the data miners up front.

We think that this approach, including the concept of decoupling data mining from forecasting and simulation processes, can easily be replicated and applied to problems from other industries. Examples of this are problems that require similar predictive models and simulation of networked systems on a large scale, such as for instance sensor networks, retail outlet planning, supply chain logistics and revenue predictions for products with a complex billing process (which we have already applied, see chapter 6).

The rest of the chapter is structured as follows. Section 5.2 describes the load parameters. Section 5.3 discusses the complex nature of network load and how to approximate it, including our motivation for modeling at the granular cell level. Section 5.4 describes the construction of the load formulas and forecasting of future network load using simulation, as well as other simulation scenarios. Limitations and future work are discussed in Section 5.5. Finally, we present our conclusions in Section 5.6.

5.2 Defining the Load Parameters

In this section we will describe how we measured the load for a cell, plus the underlying attributes that we used to predict future load. Both output and input parameters were measured per individual cell per hour.

5.2.1 Output Parameters

The communication between a network site/tower (radio network element that provides access to the mobile network) and a user's mobile device is separated into

downlink communication- directed from the site to the mobile device and uplink communication- directed from the mobile device to the site. Each physical site contains radio antennae that typically create three geographic cells of the mobile network. These cells share the physical resources of the network site.

Therefore, the Air-interface load for a cell consists of the Downlink Load (DL) and Uplink Load (UL). Multiple measures of both DL and UL can be devised. A cell is considered to be in overload if either the uplink or the downlink load is above a certain threshold. When a cell is in overload, customers demanding its radio resources cannot be served adequately. Obviously, all network sites containing cells in overload require an adequate upgrade.

Most of the background literature on telecom networks is related to network optimization or load control rather than load prediction (Geijer Lundin, Gunnarsson and Gustafsson, 2003; Muckenheim and Bernhard, 2001; Natalizio, Marano and Molinaro, 2005; Yates, 1995). In our previous research (Radosavljevik et al., 2012) we used the following measurements of load as output parameters: Count of RAB (Radio Access Bearer) Releases Due To Interference (Yates, 1995), Average Noise Rise (Geijer Lundin et al., 2003) and Average Noise Rise on Channels Dedicated to Release 99 Capable Devices (refers to lower data transfer speed up to 384 Kbps). Two additional uplink measures were considered: Count of RAB (Radio Access Bearer) Setup Failures and Count of RRC (Radio Resource Control) Setup Failures. These measurements were discarded at later stages of the process due to a very low number of models that could be generated because of too many zero-values.

In Radosavljevik et al. (2012), we used the following parameters as measures for downlink load: Percentage of Consumed Downlink Power (Muckenheim and Bernhard, 2001) and Count of "No Code Available" Situations (Natalizio et al., 2005).

However, some of these measures were specific to Nokia Data Warehouse (Nokia Siemens Networks, 2008), a performance management tool deployed at T-Mobile Netherlands where our research originated, or were not measured by other operators that were looking to use our system. Therefore, we used universal measurements, which are applicable to performance management systems of other vendors, such as Ericsson (Ericsson, 2013), Huawei (Huawei, 2013) or MyCom (MyCom, 2013). Therefore, in our new approach we picked measurements that are both compliant to the 3gpp Mobile Broadband Standard (3gpp, 1999) and universally defined and measured across the operators which are part of the Deutsche Telekom group.

Percentage of Uplink Load, also known as UL Carrier Load Percentage (3gpp, 1999) is the ratio between the total received power level on that carrier and the maximum acceptable level of interference.

$$UL\_LOAD = 100 * (1 - \frac{1}{10^{\frac{MeanRTWP - MinRTWP}{10}}}) \tag{5.1}$$

where MeanRTWP is mean Received Total Wideband Power per cell; MinRTWP is minimum Received Total Wideband Power per cell, used as the noise floor. In other

words MeanRTWP is mean of the power assigned to users, while MinRTWP is the power measured when no users are using cell resources.

Percentage of Downlink Load, also known as DL Carrier Load Percentage (3gpp, 1999) is the ratio between the total transmitter power level on that carrier and the maximum acceptable transmitter power.

$$DL\_LOAD = 100 * 10^{\frac{MeanTCP - MaxTxPower}{10}} \tag{5.2}$$

where MeanTCP is mean Total Transmitted Power per cell; MaxTxPower is maximum transmit power of the cell. In other words MeanTCP is mean of the power assigned to users, while MaxTxPower is the cell power capacity.

We used two additional measures for downlink load, based on code capacity, namely Code Utilization and Count of "No Code Available" Situations. Each cell has 256 codes that can be assigned to a mobile device for a voice call or a data session. The higher the downlink bandwidth required, the higher the number of codes will be assigned. For example, voice calls require 12.2 Kbps (translates into 2 codes), while Data Sessions can require up to 14.4 Mbps (which would consume all the codes of that cell).

Code Utilization Measures the fraction of codes used vs. codes available at the cell. It is averaged over an hour.

Count of "No Code Available" Situations -After all the codes have been assigned, the next device that requests a code from the cell, gets a "no code available" message and cannot use the cell resources. This variable measures the count of occurrences of this message per hour, and will be abbreviated as NCA.

5.2.2 Input Parameters

In Table 5.1 we provide a list of input parameters, as well as the description for each parameter we used for modeling. All these variables are measured per hour. Even though we included input parameters related to voice services, most of the input parameters are related to consumption of data services, because they require more cell resources. Forecasts for future values of the input parameters were available at the operator. In our earlier research on this topic (Radosavljevik et al., 2012), we used additional measures from the network management tool Nokia Data Warehouse. However, due to constraints mentioned in Subsection 5.2.1, namely different performance management tools from different vendors, not all of these could be measured. Therefore, in comparison with our previous research, we reduced the input parameter set by excluding the following measures: Average Soft Handover Overhead Area (measures the intersection of coverage of the particular cell with other cells), Average Proportion of Voice Traffic originated in that cell (as opposed to traffic originated in other cells and handed over to that cell), Average Proportion of Data Traffic originated in that cell, Average Voice Call Users, Maximum HSUPA users, Maximum HSDPA users and Total Active RABs, as they could not be measured.

Table 5.1: List of Input Parameters

Variable	Description
Average Count of Release 99 Uplink users	Average number of users that consumed uplink cell resources on a R99 capable device (up to 384 Kbps).
Average Count of Release 99 Downlink users	Average number of users that consumed downlink cell resources on a R99 capable device (up to 384 Kbps).
Average Count of HSUPA users	Average number of users that consumed uplink cell resources on a HSUPA (High Speed Uplink Packet Access) capable device (up to 5.76 Mbps).
Average Count of HSDPA users	Average number of users that consumed downlink cell resources on a HSDPA (High Speed Downlink Packet Access) capable device (up to 14.4 Kbps).[a]
Count of RRC attempts	Radio Resource Control (RRC) attempts are related to the signaling exchange between the mobile device and the network cells. There can only be one RRC connection open per mobile device at a time.
Count of Data Session RAB Attempts	Radio Access Bearer (RAB) is necessary to be assigned to a user in order to make voice call or a data session. Multiple RABs can be assigned to the same device. This variable measures the RAB attempts (not necessarily successful) for a data session in a cell in an hour.
Count of Voice Call RAB Attempts	This variable measures the RAB attempts (not necessarily successful) for a voice call in a cell in an hour. It is the only variable that addresses usage of voice services exclusively.
Average Downlink Throughput	Average per hour of the sum of downlink bandwidths consumed by all users served by the cell.
Average Uplink Throughput	Average per hour of the sum of uplink bandwidths consumed by all users served by the cell.

[a] Most of the current mobile devices are HSDPA capable. Theoretically, even higher speed scan be achieved for both HSUPA and HSDPA. But, an HSDPA device can also be assigned to a R99 downlink (slower) channel, if there are no HSDPA cell resources available.

5.3 Approximating the Load

Traditional approaches for mobile network load forecasting are most often analytical or Monte Carlo based approaches (Mäder and Staehle, 2004). However, the inputs required are difficult to measure and forecast, or do not relate to customer behavior such as call usage which is easier to understand and use for scenario purposes.

Most of the data mining literature on load forecasting is related to electrical networks. A good overview is presented in Feinberg and Genethliou (2010). Various methods have been deployed for this purpose: regression models, time series, neural networks, expert systems, fuzzy logic etc. The authors state a need for load forecasts for sub-areas (load pockets) in cases where the input parameters are substantially different from the average, which is a case similar to different cells in a mobile telecom network.

Related to mobile telecommunications, data traffic load (which is different than air interface load) focusing on a highly aggregated link has been forecasted in Svoboda, Buerger and Rupp (2008), comparing time series (moving averages and dynamic harmonic regression) with linear and exponential regression. Also, Support Vector Regression was used by Bermolen and Rossi (2009) for link load prediction in fixed line telecommunications.

In order to forecast the future load for each cell in the network, it is necessary to understand the relationship between the input parameters (causing the load situation) and the current load. The input parameters in case of the Air Interface load are all parameters which can be made accountable for the load situation in the cell (Section 5.2). Therefore, the load parameter (output) can be expressed as $L = f(x_1, x_2, , x_n)$. Ideally, the load of each cell x in a given time could be expressed as the sum of all users consuming resources of that cell at the particular time multiplied by the amount of resources they use plus the interference between that cell and all the other cells in the network (in practice limited to the neighboring cells):

$$L(x) = \sum_{i=0}^{m} \sum_{j=1}^{n} User_i * Resource_j + \sum_{y=1}^{Z} interference(x, y) \tag{5.3}$$

where m is the count of users that are using the resources of cell x, n is the count of resources of the cell x, z is count of all cells in the network and *interference(x, y)* is the interference measured between cell x and y. Unfortunately, there was no tool that would provide such a detailed overview.

In order to approximate the load function, we recorded the different load parameters (outputs) and input parameters described in Section 5.2, on an hourly basis during 1,5-8 weeks, depending on the operator. This provided approximately between 200 and 1,000 instances per cell or 20,000,000 instances in total on a network of 20,000 cells.

One of the choices to be made was whether a distinct formula for every cell shall be built or - alternatively - a common formula valid for all cells should be used. The

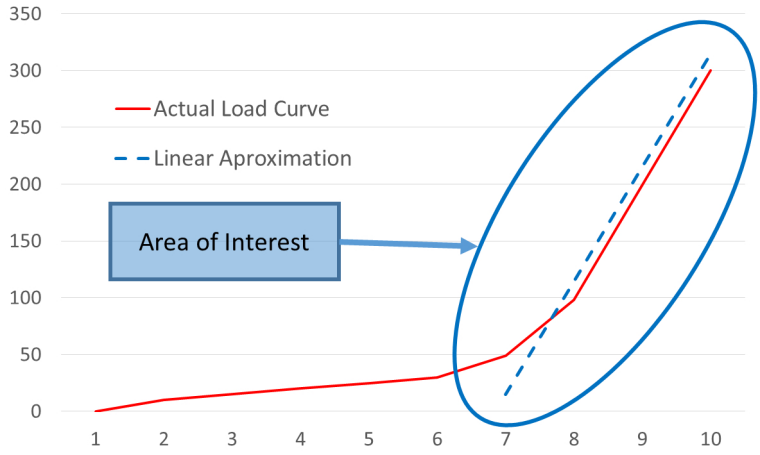

Figure 5.1: Actual Load vs. Linear approximation

approach where a model is created for each cell was chosen, due to the network experts' conviction that each cell is different, and a unified approach simply would not work, because some of the parameters influencing the load of each cell were immeasurable and unpredictable.

Next, the domain experts were intrinsically interested in being able to model cells that actually do not behave like other cells, especially when these are highly loaded. Furthermore, there would be a challenge in normalizing with respect to the varying capacity of the cells, i.e. what where the cell sized to handle. Finally, we hypothesized that not just model parameters could differ by cell, but also the optimal selection of features, similar to the concept of load pockets explained by Feinberg and Genethliou (2010).

The choice of linear regression (Witten and Frank, 2005) was made due to several reasons. First of all, even though the distribution of the values of each of the load measures we were trying to predict varied between close to linear and close to exponential, we were only interested in the higher values of the load curve, and this can be approximated quite well with linear regression, as shown on Figure 5.1. For this purpose, before constructing the regression formulas, we removed all zero instances. Furthermore, linear regression is a very fast algorithm compared to other methods, which is very useful when it is necessary to develop a large number of models in a short time. Even though it is imaginable that better results might be achieved by using non-linear regression, regression trees, or other algorithms, this might not be necessary in most cases (Figure 5.1).

Also, simple low variance methods such as linear regression frequently perform

much better in practice than more complicated algorithms, which can very often over fit the data (e.g. high variance algorithms such as neural networks). In other words, in real world problems variance is typically a more important problem than bias when it comes to data preparation and algorithm selection (van der Putten and van Someren, 2004). Trials on a smaller sample were already made with regression trees, but apart from the visibly increased time consumption, the accuracy did not improve. On the contrary, in some instances it decreased.

Last but not least, linear regression is easy to implement, easy to explain and its results and models are easy to export for other use. Exporting the models to Excel was of crucial value, as analysts would use them in order to predict the future load of each cell, by scaling the input parameters, based on internal forecasting models. In other words, this allowed non data miners to simulate future network load based on changes in the various types of network traffic, using simple tools they are familiar with.

5.4 Building the Load Formulas

In this section we will describe how the models were being generated and put to work. This includes the tools that were used, a detailed description of the approach, the results of this mass modeling process, the process of forecasting the future load and additional simulation scenarios.

5.4.1 Tools

The tools used in this research are either open source, or can be found in the IT portfolio of any telecom operator. These are the following.

Radio Network Performance Management System. As stated above, this research was using data from four different operators of the Deutsche Telekom group. Most of them had radio networks produced by different vendors, which meant that also different Radio Network Performance Management Systems were used for data collection of both the input and the output parameters. In this research, we used Performance Management Systems of Nokia (Nokia Siemens Networks, 2008), Ericsson (Ericsson, 2013), Huawei (Huawei, 2013) and MyCom (MyCom, 2013), depending on the operator. These software tools were already a part of the Network/IT infrastructure of the operators. They contain technical parameters related to the mobile network performance. The most important feature of these tools for our research was that they contained hourly aggregates of all the input and output parameters we used (Section 5.2). These are the only domain specific tools from our process.

Load Prediction and Simulation Data Mart. This is an Oracle Database 10g-64 bit v10.2.0.5.0 (Oracle, 2011) used for all our task specific data preparation and manipulation.

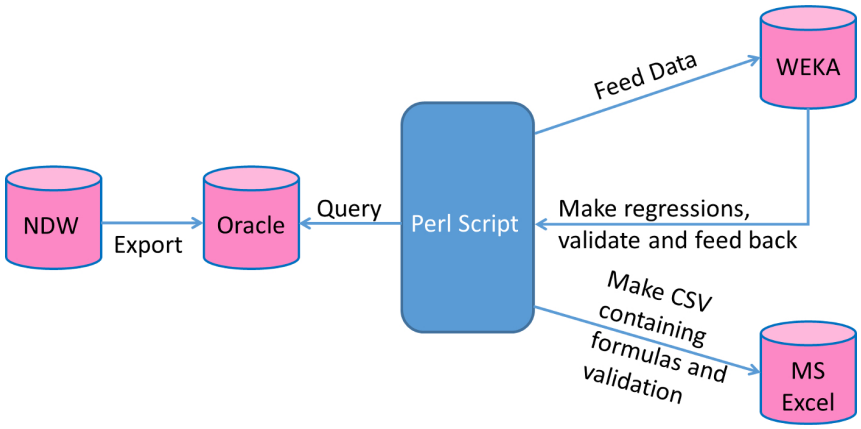

Figure 5.2: Communication Graph of the Tools used

Due to the fact that the necessary input and output parameters were stored at different tables in the respective Performance Management Systems, we needed a separate database where we could manipulate the data easier (e.g. merge tables, create indexes, and build the final flat table). In the case of Nokia Data Warehouse (Nokia Siemens Networks, 2008) this reduced the duration of the data collection and data preparation process from two weeks to one day by productizing data collection. Because we were rebuilding and rescoring models on a continuous and automated basis, this was a key improvement. Any other database platform (commercial or open source) could have been used. We opted for Oracle based on license availability.

WEKA 3.6.4 x64 (Hall et al., 2009), an open source data mining platform, was used for building the linear regression formulas and validating them. Of course, any other tool capable of deriving linear regression could also be used for this purpose. That said, our approach showed that even a research focused open source tool like WEKA can be used in critical commercial settings, at high complexity (e.g. 20.000 cells, 4 models each, around 1000 instances each).

Strawberry Perl for Windows v5.12.3 (Strawberry Perl, 2011) is an open source scripting language which we used in order to create the script that is the core of this approach. Our script created WEKA input files by querying the Oracle database, generated the regression models by executing calls to WEKA, and stored the regression formulas and the cross-validation outputs (Correlation Coefficient, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error, and Total Number of Instances used to build the model) in csv files.

MS Excel 2010 (Microsoft Corporation, 2010), part of MS Office 2010, was used to predict the future load of cells, using the regression formulas created by WEKA and extrapolations of the input values built using scaling factors based on handset/Internet usage developments (internal to the operator).

5.4.2 Process Description

A graph of how our approach used these tools to derive and store the regression models is presented on Figure 5.2. First, the data was extracted from the Network Performance Management Tools, e.g. Nokia Data Warehouse (NDW). The core of our approach is a Perl (Christiansen and Torkington, 2003) script that automated the derivation of regression formulas for each cell. This script executed calls to WEKA and queried the Oracle Database. It works in the following manner:

1. *Get list of cells from the database*
2. *For each cell*
   - 2.1 *Run a query on the database to isolate only the data related to that cell (all the input and output parameters).*
   - 2.2 *Make separate files for each of the load output parameters*
   - 2.3 *For each of the load output parameters*
     - 2.3.1 *Filter out all instances where the load is 0*[1].
     - 2.3.2 *Select only relevant variables for the regression formula of that cell, using a wrapper approach*
     - 2.3.3 *Build the linear regression formula and store it in a separate file.*
     - 2.3.4 *Use 10-fold cross-validation to validate the model.*
     - 2.3.5 *Store the formula, the number of instances used to build the regression formula, the correlation between the predicted and actual value for load, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) as reported from the cross-validation.*

While generating the models/regression formulas, we used a wrapper (Kohavi and John, 1997) approach. Wrapper approaches automatically select the best variables for predicting the outcome, taking into account the algorithm to be used, which in our case was linear regression.

Wrapper approaches do not necessarily perform better or worse than filter approaches (Tsamardinos and Aliferis, 2003). Our motivation to use the wrapper approach was to avoid human interaction with the model building process as much as possible, which obviously makes the process much faster.

It is worthwhile mentioning that the optimal feature and linear regression model selection were performed using 10-fold cross validation (Witten and Frank, 2005). This was done in order to balance between cells with large sample of non-zero instances and cells with a smaller sample. The reported correlation coefficient, MAE and RMSE are averages from the 10 repetitions. Using 10-fold cross validation already provided a good estimate of the accuracies of these formulas. Of course,

[1] We did not want noisy data. Cells/instances with no load are of no interest.

we did test them on completely new data sets, not only to confirm the accuracies achieved, but also to find out when is a good time to update the model. We expect that updates should be necessary every few months, because of the reconfiguration of the network, additions of new cells and upgrades to the existing ones.

5.4.3 Results and Discussion

Using this process we were able to run 30,000 regressions per day, by just one click. This does not necessarily result in 30,000 models, because in some cases it was impossible to derive a formula due to the large number of instances that were filtered out for zero load. But, in order to measure the load of a cell, it is sufficient that a model is generated for at least one output variable. Cases of cells where it was not possible to generate a model for any of the outcome variables were rare. Furthermore, cells that did not show any load by the means of the output variables were not of interest for our problem situation. For practical purposes, we will only present the modeling results for two of the four output variables we used to describe the air interface load in Section 5.2. We chose to present the results for the uplink and downlink load. All tables have the same structure. In the first column we list bands (intervals) of the output variables Downlink Load and Uplink Load, respectively. The second column contains the count of cells that fall into the respective bands. The third column presents the average count of non-zero instances (NZI) in each band. In other words, it presents the number of instances used to build the regression, because we only took non-zero output values into account. The fourth column presents the average Correlation Coefficient (CC) between the predicted and actual values of the variables in the particular band. These Correlation Coefficients are the result of the 10-fold cross validation. The last column presents the ratio between the number of formulas that were generated and the total count of cells in each band. Namely, for certain cells it was not possible to build the regression because of a very low number of non-zero instances.

The results of the Regression Modeling for Downlink and Uplink load for four different countries are shown in Tables 5.2-5.9. In Tables 5.2 and 5.3 we present the modeling results of T-Mobile Netherlands, the same operator published in Radosavljevik et al. (2012). However, the operator was still undergoing a full network swap during our research, which means every cell in the network was either already replaced or about to be replaced by a new one from a different network equipment vendor. At the moment of research, this operator was running both networks in parallel, which created an additional level of complexity. The results presented in Tables 5.2 and 5.3 are referring to the modeling process on the swapped part of the network using the new vendor's equipment. Hence, the total number of cells is smaller than reported in Radosavljevik et al. (2012). For this reason, and the fact that we are presenting different output variables in this chapter, the results of Radosavljevik et al. (2012) and these results should not be compared.

Table 5.2: Regression Modeling Results for Downlink Load (DL) for Country Operator 1

Downlink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$DL < 1$	642	/	/	/
$1 <= DL < 5$	132	504.2	0.906	99%
$5 <= DL < 10$	450	528.5	0.92	100%
$10 <= DL < 20$	2995	507.7	0.914	100%
$DL >= 20$	4120	511.1	0.955	100%

Table 5.3: Regression Modeling Results for Uplink Load (UL) for Country Operator 1

Uplink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$UL < 1$	643	579	0.254	1%
$1 <= UL < 5$	693	513.94	0.536	96%
$5 <= UL < 10$	2880	522.06	0.676	100%
$10 <= UL < 20$	3405	516.14	0.756	100%
$UL >= 20$	718	503.05	0.776	99%

Table 5.4: Regression Modeling Results for Downlink Load (DL) for Country Operator 2

Downlink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$DL < 1$	683	/	/	/
$1 <= DL < 5$	2379	155.4	0.777	80%
$5 <= DL < 10$	9406	247.9	0.824	96%
$10 <= DL < 20$	4550	271.7	0.846	95%
$DL >= 20$	1697	284.7	0.872	92%

Table 5.5: Regression Modeling Results for Uplink Load (UL) for Country Operator 2

Uplink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$UL < 1$	431	172.8	0.554	17%
$1 <= UL < 5$	138	247.5	0.649	84%
$5 <= UL < 10$	909	273.9	0.565	87%
$10 <= UL < 20$	9649	294.7	0.617	95%
$UL >= 20$	7816	288.8	0.801	98%

Table 5.6: Regression Modeling Results for Downlink Load (DL) for Country Operator 3

Downlink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$DL < 1$	387	/	/	/
$1 <= DL < 5$	2447	142.5	0.854	84%
$5 <= DL < 10$	2428	155.9	0.907	95%
$10 <= DL < 20$	2114	175.5	0.935	99%
$DL >= 20$	746	184.6	0.945	100%

Table 5.7: Regression Modeling Results for Uplink Load (UL) for Country Operator 3

Uplink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$UL < 1$	160	56.6	0.382	11%
$1 <= UL < 5$	715	120.6	0.482	77%
$5 <= UL < 10$	1909	146.4	0.546	92%
$10 <= UL < 20$	3195	162.4	0.677	98%
$UL >= 20$	2143	171.2	0.668	96%

Table 5.8: Regression Modeling Results for Downlink Load (DL) for Country Operator 4

Downlink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$DL < 1$	5	952.4	0.449	100%
$1 <= DL < 5$	604	949.7	0.705	100%
$5 <= DL < 10$	3801	958	0.776	100%
$10 <= DL < 20$	4802	957.9	0.807	100%
$DL >= 20$	1020	958.2	0.848	100%

Table 5.9: Regression Modeling Results for Uplink Load (UL) for Country Operator 4

Uplink Load	Count of Cells	Avg Count of NZI	Avg Correlation Coefficient	Models Built vs. Number of Cells
$UL < 1$	0	/	/	/
$1 <= UL < 5$	0	/	/	/
$5 <= UL < 10$	674	948.2	0.461	98%
$10 <= UL < 20$	4187	955.5	0.598	100%
$UL >= 20$	5371	957.2	0.669	100%

The results can be evaluated by using two criteria: the Average Correlation Coefficient and the Ratio of The Models Built (the last two columns in Tables 5.2 to 5.9). The Ratio of the Models built for both Downlink Load and Uplink Load grew alongside the number of non-zero instances for operators in all four countries, which was to be expected. We chose the Correlation Coefficient because it is a relative measure and therefore more intuitive than the Mean Average Error or the Root Mean Square Error. The Correlation Coefficient was also much easier to explain to the end users than the latter two error measures. As mentioned above, we report the Average Correlation Coefficient of each load band. The confidence intervals for the Average Correlation Coefficient at 95% confidence level were not wider than ± 0.02 for any operator in any of the Downlink or Uplink load bands, due to the relatively low standard deviations.

Furthermore, because of the choice we made at the beginning of the research, to focus on the higher levels of load and eliminate the zero values, the Average Correlation Coefficient between actual and predicted values also grows as the load is higher, both for Downlink and Uplink Load. In the lowest load bands, the performance of the models is not good. However, this was of no interest, as these were not the situations that we were trying to predict. These cells were not likely to be in overload in the foreseeable future.

However, when analyzing the Average Correlation Coefficient (ACC) between the predicted and actual values there is a visible difference between Downlink and Uplink load: The ACC for Downlink Load (Tables 5.2, 5.4, 5.6, 5.8) was much higher than the ACC for Uplink Load (Tables 5.3, 5.5, 5.7, 5.9). Uplink Load seems more difficult to predict using linear regression. There are two possible reasons for this: A crucial input parameter (predictor) may be missing; or the Uplink Load has less of a linear nature.

Last but not least, model performance across operators cannot be compared due to differences in network vendors, software versions, geography, population density and smartphone penetration rates (which cause higher network load). The importance of the smartphone penetration and population density was also confirmed by the automated feature selection, where variables such as the combined throughput (uplink or downlink), which is highly influenced by smartphones and the number of HSDPA/HSUPA users per hour (which are smartphone users) were the most often selected when building the respective load formulas.

5.4.4 Forecasting the Load

Once the load formulas have been derived it was possible to forecast the future load situation if the changes in the describing parameters are known. These changes of the input parameters were described by means of scaling factors. The scaling factors were calculated by using a traffic forecast model developed by the operator (out of

scope of this research). A cell was marked for upgrade[2] if any of the four output variables used as measures of Uplink or Downlink load, was above a predefined critical value.

This is done in the following way:

1. *For each output variable*

   2.1 *For each cell*

   2.3.1 *Select the top 100 instances of the output variable and its corresponding values for the input variables.*

   2.3.2 *Make averages of these input variables.*

   2.3.3 *Scale the input variables up or down, according to scaling factors developed by a traffic model.*

   2.3.4 *Feed the scaled values of the input parameters into the regression formula for the output variable for that cell*

   2.3.5 *If the resulting value is higher than the critical threshold for that output variable, the cell should be upgraded.*

The forecasting model provided better and more sophisticated forecasts and as such supported better network investment decisions, which account for the major part of the entire operator CAPEX cost. In simple terms, no money was wasted by investing in unnecessary network upgrades, providing two benefits: lower cost and redirecting investments into areas that had a higher impact on positive network experience for the customers.

In addition, note that this part of the process was performed in a tool as simple as MS Excel. This was a key driver for the business success of the solution. In our experience the importance of the Deployment step in the data mining process is generally underestimated. By providing not just the scores but also the underlying models in a format and tool that was immediately usable and tunable to end users who are not data miners, the solution was readily accepted and also used in new ways not necessarily intended by the data miners, for instance detailed simulation scenarios. In our view, this approach may be applicable to many other domains.

5.4.5 Applications- Simulation Scenarios

Initially, the only application of this research envisioned by the authors was the deployment scenario for forecasting future load and predicting necessary network upgrades due to "regular" traffic growth, as described in Subsection 5.4.4. This has already been used in four country operators belonging to the Deutsche Telekom group.

However, due to the flexibility of the approach, meaning using simple tools such as Microsoft Excel for implementation, the system developed a life of its own: the

[2]Technically speaking, the network site which generates the cell is upgraded, not the cell itself

end users in T-Mobile Netherlands, the operator where this research was originally developed, started creating simulation scenarios suited for different needs.

Step 2.3.3 of the algorithm described in the previous subsection mentions feeding scaling factors for the input parameters **based on a traffic model**. What if this traffic model was to be replaced by a different one? In that case, a new simulation scenario would be generated. Using our approach, all it takes to generate a new scenario is to change the values of the input parameters in MS Excel. The output of the model (Downlink and Uplink Load) would be automatically recalculated and the user could immediately see the effect. We will explain a few actual use case scenarios in the following paragraphs.

One of the first use cases generated was to predict future network load and evaluate network investments, based on proactive localized marketing campaigning. It was a co-operation between the Marketing and Network Technology Department. The Marketing Department provided their campaign description and expected benefits, namely new customers and increased service usage, which were trended in terms of the input parameters described in Section 5.2. These were fed into the model as described in Subsection 5.4.4, so the increased future load could be predicted and the necessary network improvements can be made, even before the marketing campaign was launched. A very similar scenario was in use for opening new stores, due to the fact that increased number of customers was expected when opening a brick-and-mortar store. This allowed for the network to be prepared to accept the new customers without impacting the experience of the existing ones.

Another very powerful application of this model was evaluating a business case for adding a new wholesale client- or an MVNO (Mobile Virtual Network Operator). This is an operator that does not own a network; instead a MVNO is renting the network of a bigger telecom operator in order to provide services. In this case, the localized traffic growth for predicting the future load was based on the location (or the evaluation of) the customers of the MVNOs and their respective service usage. These were then trended and fed into our model (via MS Excel) in order to evaluate the necessary network improvements, so that no degradation of service for the customers of the host operator would occur. However, these upgrades come at a certain cost, which is attributed to accepting the MVNO onto the host network. If the benefits (revenues) generated by accepting the MVNO are lower than the costs incurred, the business case is negative and therefore, rejected. This approach was used in the country where the research originated to reject a business case for adding an MVNO. Furthermore, a MVNO business case was evaluated for another operator from the same telecom group.

Last but not least, this approach was used as one of the criteria to determine the strategy for the network swap and deployment of LTE (4G) network in T-Mobile Netherlands B.V. As mentioned in Subsection 5.4.3, the operator was undertaking a major network infrastructure investment, namely replacing the entire radio network (every site) in order to modernize it and allow for deployment of 4G. Of course an

undertaking of this size cannot be performed all at once; hence clusters of cells were being planned for replacement at a certain time. Our load prediction method was one of the criteria used for giving priority to certain clusters, thus reducing the need of unnecessary investments into the "old" network. The underlying assumptions here were that the "new" generation radio network would have more efficient resource use and therefore could handle the load better, and that a certain amount of customers would start using 4G services, therefore offloading the 3G network.

5.5 Limitations and Future Work

The regression formulas developed by this approach can be used on a long term basis only if the mobile network stays the same (is frozen) over a longer period. But, this is not the case. The cellular network is a system of very complex dynamics. The many changes that occur, such as hardware and software updates, network reconfigurations and optimizations, as well as network upgrades and roll-out of new sites, which reduce the load of the existing ones, cannot be taken into account in advance. It is necessary to collect a new data set and rebuild the regression formulas, in order to incorporate all these changes into the model. This is why the process described in this chapter was scheduled for execution every 3-4 months.

Next, we intend to improve the predictions for uplink load. One method would be searching for additional input parameters to improve the performance of predicting uplink load using linear regression. Alternatively, we could look for a substitute for linear regression better suited for modeling uplink load on the cells where linear regression does not deliver. However, this algorithm should not substantially slow down the whole process and must be easily transferable to MS Excel, in order to keep the flexibility and the ease of building simulation scenarios.

Further evaluation of the quality of the derived load formulas of course also involves the comparison of the predicted load with the actually measured load in the future. However, it should be noted that there are a lot of factors impeding a direct comparison. As stated above, all changes to the settings of a cell within the forecasting timeframe affect the load formula, which means that after such changes the derived formula is - at least to some degree - no longer correct. For this reason it will be challenging to really quantify the accuracy of the predictive model. Developing a fair method of evaluation, which would incorporate the network changes, would be beneficial. In terms of the core algorithms, we do want to keep the benefit of using a simple, fast and robust low variance approach such as linear regression.

However, we do plan to explore a methodology that would allow us to combine a global network model with local models for each cell, for instance multitask or transfer learning (Caruana, 1997). In principle, we have almost infinite data available for most cells, so local models cannot be improved by a global model. Nevertheless, there could be an exception for a non-select small number of cells. Next, a clustering approach could be devised to group cells with similar formulas or levels of load,

thereby generating new knowledge for the telecom domain experts.

Furthermore, we do intend to investigate additional simulation scenarios for our approach, beyond those described in Subsection 5.4.5. Last but not least, this research has been implemented in four operators of this telecom group. Other operators from the same group are planned to follow, with their own use cases and applications.

5.6 Conclusions

In this chapter we presented a very simple yet effective approach of deploying data mining in commercial surroundings. Unfortunately, data mining is still seen as a black box in many industries, telecom not excluded. Even though some data mining activities are taken, typically in the Marketing/Customer Retention field, there is a myriad of other possibilities in business where data mining can be applied. In our opinion, it is better to start with simple methods, such as linear regression, because it is easier to understand them. Once these simple approaches gain acceptance, and familiarize companies with data mining, opportunities to apply more advanced techniques will arise.

In our result section we showed that it is easier to accomplish a target, if one is focused on it. Namely, with our approach we wanted to target cells where some load (non-zero load) occurs, in order to predict the part that really matters more correctly: the high end part of the load curve (the cells in overload). In other words, as the network load grew, so did the quality of the model's predictions. We willingly sacrificed the models' performances within cells with very low load, because they are of no interest.

Next, one of the key values of the approach is that a large number of regression models (close to 30,000 per day) were developed in a very short period of time with minimum human interaction. In order to do this, we deployed a simple algorithm such as linear regression, motivated by its speed and other benefits explained earlier, a wrapper feature selection, in order to avoid human interaction, and 10-fold cross validation which makes the models statistically sound. Manually, this task would be impossible. Obviously, the possibility to generate these formulas was crucial to the operator. At the moment, the commercial tools for this purpose offer only load predictions based on single variable regressions (MyCom, 2013), which is not as robust as our approach.

Typically, planning network upgrades is a reactive process. Our approach makes it proactive, which was acknowledged by the operator, who has fully integrated our approach into its network upgrade planning and budgeting activities. Of course, due to the fast pace network changes, the formulas would need to be upgraded every 3-4 months, but this was also scheduled as a part of the standard process. Due to confidentiality, we cannot disclose the exact return of this project, but given that the network is the key resource of an operator, the investments into its upgrades are quite sizeable. To our knowledge, this is the first time a telecom operator has applied

data mining in order to create a proactive network upgrade management process. This allowed the operator to manage network performance better and avoid extreme congestion situations, which can result in degraded customer experience and loss of reputation for the operator. As mentioned at the beginning, the research was performed at Deutsche Telekom, a large telecom operator group with branches in many European countries. Our research was used in four of the countries where this group is present.

Potentially the greatest benefit of our approach is the decoupling of the data mining process from simulation scenarios. This was accomplished by exporting the models into Excel sheets after they have been have been generated by our data mining process. Then, the end users, a team of radio network analysts who are not data miners, were able to use these formulas resulting from a data mining process for forecasting the future network load. This allowed them to simulate multiple traffic scenarios by scaling the current input parameters, which was as simple as changing values in their respective columns in Excel. These scenarios included "regular growth" scenarios, evaluations of network investments necessary to accommodate localized user growth due to targeted marketing campaigns, adding a new wholesale client (an MVNO- Mobile Virtual Network Operator) and prioritizing clusters for deployment of new technologies such as LTE/4G.

Next, we would like to point out the possibility of applying our research onto problems other than telecom network load. This approach would be applicable to any other industry where large scale regression models are necessary. This can be accomplished simply by replacing the data source, in this case the Radio Network Performance Management Tools, with a data source suitable for the industry that would like to apply our research. The decoupling of the data mining process from the simulation scenarios makes our approach more general to situations where detailed simulations are necessary, but the domain experts are not data miners. We already tested this approach for cluster based revenue predictions, which is a topic from the finance domain (see chapter 6).

Last but not least, perhaps one of the most interesting aspects of our approach is the extremely low cost. Given that we used the existing IT infrastructure (Server, Radio Network Performance Management Tools, Oracle, Excel) combined with open source tools (WEKA, Perl), the only costs that incurred were the Processing Time Cost (of the Server) and the labor cost of the employees in this project. Also, the Oracle Database that we used can be replaced with a less expensive or free database alternative in order to further reduce the cost, in case the potential user of our approach does not have an Oracle License. These amounts are insignificant compared to the actual investments being made into the network.

Addressing the research question posed in section 5.1, we have shown how data mining can be used to predict 3G mobile network interface load, and simulate it under different scenarios. In a nutshell, we have used relatively simple algorithms to create a large number of predictive models, therefore making possible predicting the load

on a cell level. We have used tools known to the end users to deploy these models, allowing them to use different scenarios for the input parameters. Acceptance was gained by decoupling the data mining process from the end users, but keeping the transparency that linear regression offers combined with tooling familiar to them.

Chapter 6

Service Revenue Forecasting in Telecommunications: A Data Science Approach

Radosavljevik, D., van der Putten, P.

Based on an extended abstract published in Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning. pp. 187–189 (2017)

In many industries revenue forecasts can drive important business decisions, such as when and where to intervene in order to accomplish business targets. For service based businesses such as telecommunications, timely and precise service revenue forecasts are of great importance. Furthermore, having a simulation platform for service revenue forecasting can provide the business with an idea of how these measures could play out in practice under multiple scenarios. This chapter discusses a real-world case of revenue forecasting in telecommunications. Apart from the method we developed, we will describe the implementation, which is the key factor of success for data science solutions in a business environment. We will also describe some of the challenges that occurred and our solutions to them. Last, but not least we will present the results of this process and explain our problem specific choice for error measure.

6.1 Introduction

In this chapter we focus on the postpaid segment of mobile telecommunications, where service revenues can be split into fixed (subscription fee which already contains certain services) and variable revenues (based on additional usage of services not included in the subscription). Furthermore, operators charge differently for using services while abroad (roaming) and for making international calls. Operators also charge other operators interconnect fees for incoming calls (a customer from operator B calls a customer from operator A), which results in costs and revenues not visible to the customer.

T-Mobile Netherlands, the operator where we developed and deployed this research, has about 2 million postpaid customers with more than 4000 combinations of differently priced rate plans and voice, SMS and Internet bundles. In order to forecast the revenue figure for one month, one has to account not only for the different usage patterns throughout the year, but also for the inflow of new customers, changes in contract of the existing ones and the loss of customers to competition. The systems that are used for actual customer billing are not built for simulation of revenues, as these are too embedded into operational processes. This makes the task of forecasting revenue across different scenarios far from trivial.

The standard approach for forecasting service revenues developed by the operator where we conducted this research is a very high level approach, comprising of many weeks of manual labor. It is executed via complex Excel (Microsoft Corporation, 2010) sheets and it delivers a non-transparent view of the forecasted service revenues, with no possibility of a breakdown per product (i.e. rate plan). Furthermore, due to the Excel based tooling, the process requires a lot of attention to detail in order to avoid errors and does not allow for efficient calculation of multiple scenarios, because it takes too long to generate the results.

Therefore, it was necessary to develop a new approach for service revenue forecasting in order to overcome the weaknesses of the existing one. The end objective

of this research, i.e. the task presented to us by the operator is to predict the service revenues for one year ahead on a monthly basis. As stated in the previous chapter, we wanted to adapt the work described there to a completely different domain. Hence the research question in this chapter is:

How can data mining be used to predict and simulate service revenues in telecommunications? In other words, does the approach developed in chapter 5 generalize to a different domain such as finance?

This chapter could be relevant to researchers interested in forecasting revenues in other industries, especially service based businesses. Furthermore, the simulation framework we have developed for testing multiple scenarios might be of interest to researchers in other domains as well, for example micro-simulation (Li and O'Donoghue, 2013).

The rest of this chapter is structured as follows. In Section 6.2 we present the related work. A short overview of our approach is given in Section 6.3. We address the process of data collection and understanding in Section 6.4. A detailed description of our end-to-end process is provided in Section 6.5. The results are presented in Section 6.6, while the discussion and the lesson learned are in Section 6.7. We address the limitations our research and future work in Section 6.8 and provide the conclusion in Section 6.9.

6.2 Related Work

Given the importance of the process of service revenue forecasting, it is surprising there is not a lot of research on this topic. There is literature available regarding revenue forecasting in the government sector: Fullerton (1989) compared time series with econometric models and composite models, while Feenberg, Gentry, Gilroy and Rosen (1989) tested the rationality of the methods used for this purpose. Weatherford and Belobaba (2002) addressed revenue management in the airline industry, while a comparison of forecasting methods for hotel revenue management is the focus of Weatherford and Kimes (2003). Time series were used by Trueman, Wong and Zhang (2001) in order to forecast revenues of online firms.

On the other hand, a good overview of applications of machine learning for supply chain forecasting was given by Carbonneau, Laframboise and Vahidov (2008), who compared the performance of various algorithms on the sales data of the Canadian foundry industry. According to this research, even though complex models such as neural networks and support vector machines performed somewhat better than a linear regression model in terms of Mean Absolute Error, the difference was not statistically significant. Heikkilä (2002) focused on supply chain forecasting in telecommunications, but this paper is mostly addressing the problem from a perspective of the relationship between suppliers of telecom network equipment and telecom operators.

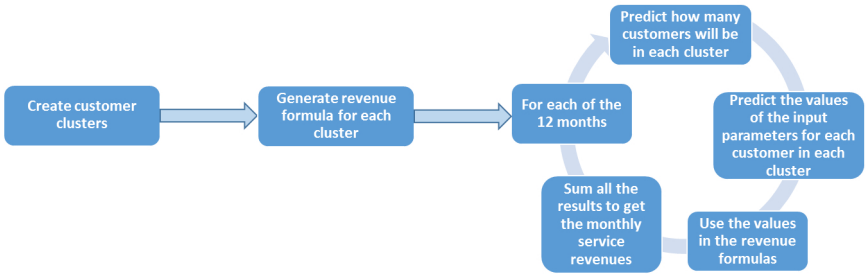

Figure 6.1: Overview of the Service Revenue Forecasting Process

However, none of these papers describe the process of service revenue forecasting in mobile telecommunications.

6.3 Overview of Our Approach

As stated above, the company where we developed this research required a much more timely, detailed and flexible approach for service revenue forecasting. In this section we provide an overview of the method we designed for this purpose. The ultimate goal of this approach was to forecast service revenues for one year on a monthly basis. A more detailed description of each step follows in Section 6.5. Nevertheless, the simplest description of our method is given on Figure 6.1: First, create clusters of customers; second, create revenue formula for each cluster based on certain inputs; third, predict how many customers will be in each cluster in each month; fourth, predict the values of the input parameters for each customer in each cluster for each month; last, use these inputs in the service revenue formulas generated and sum up to generate the monthly revenues.

As suggested in the previous chapter, we based our solution for service revenue forecasting on the method we developed for forecasting network load (Radosavljevik and van der Putten, 2014). Similar to the concept of load pockets (Feinberg and Genethliou, 2010) introduced in the previous chapter (i.e. network cells), we here used the concept of revenue pockets. Namely, we treated clusters of customer rate plans as revenue pockets.

The rationale behind clustering the rate plans is the following. It was not feasible to come up with a single service revenue model for all customers due to the many different rate plans that the operator offered (i.e. more than 4000 combinations of voice minutes, SMS and megabytes in a bundle). Furthermore, the usage habits and billing were quite different for customers in various rate plans. Despite the fact that all customers in a single rate plan could already be treated as a cluster due to their

similarity in usage and the way they are billed for services, a unique formula for each price plan was not feasible either, as some rate plans only contained a few customers. Therefore, some sort of rate plan based clustering for customers was necessary. We treated these clusters of rate plans as revenue pockets because we expected each of these clusters to contain customers with similar usage and similar billing methods.

After creating these clusters, we created service revenue models for each of them. These models were predicting monthly service revenues based on a set of input values (i.e. monthly values for usage parameters such as used voice minutes, SMS and megabytes) for each customer in a given cluster, and they were built using two months of historic data.

However, the clusters we created will not remain the same for the entire year and neither will the values for the input parameters. Therefore, we needed to make assumptions on both the quantity of customers per cluster per month (customer base changes) for the prediction period and how will the input values develop in the same period (input values changes).

In order to establish the monthly state of each cluster, we began with a snapshot of the customer base split onto rate plan clusters from the month before the prediction period. Next, we accounted for the monthly inflow and outflow of customers from each cluster, as well as customers renewing contracts and subsequently moving to a different cluster. This was performed via generating new customers for the inflow, removing certain customers for the outflow and migrating certain customers for the contract renewals, resulting in a new monthly cluster state.

After creating the monthly customer base snapshots for the full year ahead, we also needed to make assumptions for the input values for the rest of the year. This was done by extrapolating the development pattern for all of the inputs using a two year history per rate plan aggregated monthly.

Finally, after having calculated the monthly values for each input parameter for each customer in each cluster (including the incoming, outgoing and migrating customers), we used them in the revenue formulas we generated for each cluster. Summing up these values returned the monthly revenues.

It is worthwhile mentioning that we used generic off-the-shelf hardware for this purpose: a server with 128GB of RAM and 16 processing cores.

6.4 Data Collection and Understanding

Unlike in an academic research setting where seeking the best algorithm is the key to solving the problem and getting data is as easy as downloading a data set, in a business setting data collection and preparation represent a large chunk of the total work. Typically, the data is not available from a single data source, let alone structured in a format suitable for machine learning. In our case the data was spread across various Oracle (Oracle, 2011) and Teradata (Teradata Corporation, 2017) databases. The first thing we needed to do is unify the data (invoice data, usage data, interconnect

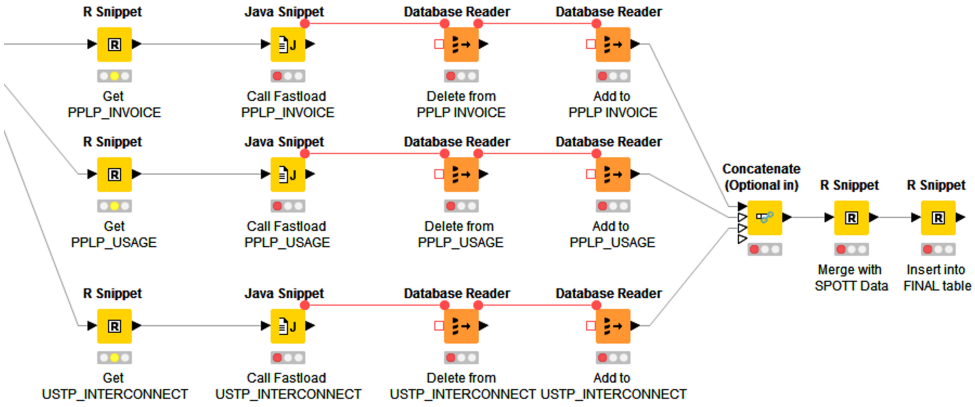

Figure 6.2: The ETL Process in KNIME using RJDBC

data, rate plans/bundles and inflow and outflow of customers) into a single sandbox which we can use to construct our flat table. We chose to use Teradata as a destination for our data set because we had the most administrative flexibility there. Any other database system would do. Also, we used the open source tools R (R Development Core Team, 2008) for moving data between data sources and KNIME (Berthold et al., 2007) for automating the process. The process is shown in figure 6.2.

Even though R is not typically used as an ETL tool, its RJDBC library (Urbanek, 2014) is very useful when necessary to export data from various database systems. For importing the data into our Teradata database we used its fastload functionality. KNIME is an analytics platform that enables using data residing in different databases and code written in different languages (Java, R, Python etc.) to be executed in a work flow fashion. It also contains a large implementation of various machine learning algorithms. During the ETL stage, we used KNIME as a kind of an orchestrator between R, Oracle and Teradata. Its visual interface makes debugging easier, as it is very clear to see which node in the diagram has failed.

After migrating the data to a single database we needed to restructure it, as it was recorded in a way corresponding to how it is presented to customers on an invoice. However, not all invoice items are service revenues (e.g. the handset fee and application purchases are not service revenues). Non service revenues are out of scope. Isolating the service revenues required a lot of expert help as each revenue type has a different code. Importing and fully aligning the data with what is officially billed and reported took almost 2 months of work. Next to this, distinction between usage that is billed vs. usage that is already part of the subscription was necessary.

The usage abroad was treated separately, because it is billed in a different manner than the usage in the home country. Last, but not least, the interconnect usage, which is not part of the customer invoice was added. Operators bill each other for the service of connecting calls between customers from a different operator. Even though these revenues are not directly billed to customers, they are part of the service revenues.

6.4.1 Data Set Description

In this subsection we describe the outcome variable and the predictors we used.

As the goal of the model is to forecast **Service revenues**, this is the obvious choice for the outcome variable. Service revenues can be further broken down into:

- **Monthly Recurring Cost**- Subscription cost paid by the customers regardless of usage (this is a constant)
- **Out of Bundle Revenues**- charged when customers use services after using up their subscription
- **Roaming Revenues**- charged when customers use services while abroad
- **Interconnect Revenues**- charged between operators when two customers of different operators call each other

Initially, the task given was to predict total service revenues without the breakdown listed above. Therefore, we used the Monthly Recurring Cost as an input.

In table 6.1 we list the inputs we used to predict the service revenues. It is worthwhile mentioning that usage abroad (roaming) is charged differently than usage in the home country (national usage). This is changing under EU regulations: all usage within EU countries is treated like national usage. Similarly, international calls made from the home country are charged differently than national calls.

We collected two months of monthly aggregated anonymized data for the outcome variable and the inputs in order to make our predictions. T-Mobile Netherlands, the operator where the research was performed, has two million customers with consumer postpaid contracts, so there were about 4 million records in the data set. Given the size of each cluster, ranging from minimum 500 customers (so 1000 data points) to more than 30,000 customers, this should be sufficient amount of data to model the dependencies between the inputs and the outcome per cluster. We only used two months of history for service revenue modeling because the pricing of services can change substantially over time, so the most recent data was the most relevant. The operator is under legal obligation to keep this data in order to be able to reproduce the invoices if necessary, for a period of three years. For trending the future values of the input parameters as they are crucial for predicting the future service revenues, we used data aggregated per cluster per month with history of 24 months.

Table 6.1: Inputs used for creating service revenue models

variable	Description
Monthly Recurring Cost	Subscription cost paid by the customers regardless of usage (this is a constant)
MBs within bundle	Amount of MB(megabytes) used within subscription
MBs outside bundle	Amount of MB(megabytes) used outside subscription
MBs in roaming within bundle	Amount of MB(megabytes) used abroad within subscription
MBs in roaming outside bundle	Amount of MB(megabytes) used abroad outside subscription
Voice Minutes within bundle	Duration of voice calls used within subscription
Voice Minutes outside bundle	Duration of voice calls used outside subscription
Voice Minutes International within bundle	Duration of international voice calls used within subscription
Voice Minutes International outside bundle	Duration of international voice calls used outside subscription
Voice Minutes in roaming within bundle	Duration of voice calls used abroad within subscription
Voice Minutes in roaming outside bundle	Duration of voice calls used abroad outside subscription
SMS within bundle	Count of SMS used within subscription
SMS outside bundle	Count of SMS used outside subscription
SMS in roaming within bundle	Count of SMS used abroad within subscription
SMS in roaming outside bundle	Count of SMS used abroad outside subscription
Incoming MMS count	Count of MMS messages received by the customer (this is service that is not used by many customers so the most frequent value is 0)
Incoming SMS count	Count of SMS messages received by the customer
Incoming Voice Calls Count	Count of voice calls received by the customer
Incoming Voice Calls Duration	Duration of voice calls received by the customer
Other usage	Count of usage of other services offered by the operators (e.g. Fax, MMS; services that are not used by many customers so the most frequent value is 0)

6.5 Clustering, Modeling and Deployment

In this section we will discuss our end to end approach to modeling service revenues. As stated in the introduction of this chapter, we have a five step approach to forecasting service revenues for a year. Therefore, this section is divided into five subsections. The first subsection describes how we clustered the customers. The second describes how we created revenue formulas for each cluster based on the inputs we listed in the previous subsection. The third subsection explains how we accounted for the monthly changes in the clusters with respect to how many and which customers remain, join or leave the cluster. In the fourth subsection we explain the method we used to extrapolate the values of the input parameters for each customer in each cluster for each month. Last, in the fifth subsection we explain how we combined the products of the first four steps in order to forecast the monthly revenues.

6.5.1 Clustering the Data

As stated in the introduction section, it was not feasible to come up with a single service revenue model for all customers due to the many different rate plans that the operator offered. Furthermore, the usage habits of customers are quite different for the various rate plans, as well as the way they are billed for the usage. However, a unique formula for each price plan was not feasible either, as some of these only contained a few customers. Therefore, we divided the entire customer base into 156 clusters, manually combining rate plans and bundles of similar characteristics (similar number of minutes, SMS and MB in the subscription), with a lower limit of 500 customers per cluster. Some clusters consisted of only one rate plan, in case when this rate plan already contained many customers (in certain rate plans there are more than 30,000 customers). Clustering algorithms were considered, however we had enough of clear structure in the rate plans to avoid this.

6.5.2 Generating the Service Revenue Formulas per Cluster

The approach we used is similar to the network capacity model described in chapter 5, where we discussed load pockets to simulate network service quality. Here, we treated each cluster as a revenue pocket, because each cluster consisted of customers who are billed in a similar manner for using services. Therefore, we created a model for each cluster of customers, taking various types of usage as input and service revenues as output, and saved it for scoring later. This process is automated using KNIME. A screen shot of the process is shown on Figure 6.3. For modeling revenues per cluster we used a data set containing two months history, which guaranteed a minimum of one thousand data points per cluster. As previously stated, we limited the data to two months of history because pricing of services can change substantially over time, so the most recent data was the most relevant in order to express the relationship between usage and service revenues.

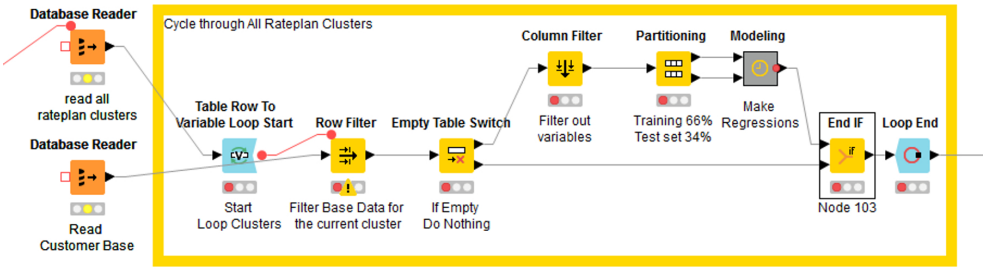

Figure 6.3: Modeling Workflow in KNIME

A KNIME workflow consists of nodes which perform certain tasks and pass on the computed result to another node. This process begins by firstly reading out the 156 clusters (node 'read all rate plan clusters' on Figure 6.3) and reading out all of the customer data (node 'Read Customer Base') including the cluster they belong in, the input variables (see Subsection 6.4.1) and the outcome (service revenues). At that moment, the data set consists of around four million data points. Next, we loop through the clusters and for each cluster the following is performed. First, only the data for customers belonging to the current cluster is isolated. The size of the data set per cluster ranges between 1000 and over 60,000 data points. Just for safety, we check if the result is empty, so that the rest of the code does not execute on an empty data set. If the resulting data set is not empty, we leave out certain variables that we do not use as inputs (e.g. name of the rate plan etc.). Next, we split the data for that cluster onto training and testing set with ratio 66%:34%, respectively. In the end, we run the modeling node. In the modeling node, a service revenue model is generated and saved using the cluster name, so we can use it for forecasting the service revenues.

We initially setup the process using linear regression as implemented in KNIME (Berthold et al., 2007) for modeling. However, we also experimented with other algorithms, such as support vector regression, random forests, regression trees, polynomial regression etc. KNIME allows for integration with WEKA (Hall et al., 2009), so we tested WEKA's implementation of linear regression combined with a wrapper as well. This is similar to what we used in the previous chapter for modeling network load. Last but not least, motivated by the No Free Lunch theorem (Wolpert and Macready, 1997) we made a so called 'winner' model, where each cluster was modeled by the algorithm that was the best performing on that particular cluster on the test set. From a methodological perspective, the results of the 'winner' model should not be compared to the rest, because we used the test set to select the winning model.

As stated in Subsection 6.4.1, service revenues can be further broken down into: Monthly Recurring Cost, Out of Bundle Revenues, Roaming Revenues and Interconnect Revenues. We had data for each of these variables, therefore we decided to look into modeling them separately. This was motivated by two factors: the business required a breakdown of revenues into components and we expected error reduction when modeling them separately, even when compounding the error.

We expected the model error to reduce due to the following. The Monthly Recurring Cost is independent of usage and can be treated as a constant. Therefore, it requires no modeling, which allows for error reduction. The other three components are much more linear in nature and completely dependent on usage. Therefore, they can be modeled using linear regression. Furthermore, the variable selection process is much more straightforward when modeling the components of service revenues. For example, when modeling the roaming revenues, we only use the input variables from Table 6.1 containing the word roaming in their name (Voice Minutes, SMS and MB in roaming, both within and outside bundle), as these are the only variables that can influence the roaming revenues. Similarly, all the variables from Table 6.1 with the word Incoming in their name (Incoming MMS, SMS, Voice call count and duration) are used for modeling the Interconnect revenues. All other variables from Table 6.1 are used for modeling the Out of Bundle Revenues.

The results of both modeling methods (modeling service revenues as a whole and modeling components of service revenues separately) are shown in Subsection 6.6.1.

6.5.3 Generating the Monthly State of the Clusters- Inflow, Changes in Contract and Outflow of Customers

Customer base growth or reduction substantially affects the total service revenues generated. Furthermore, customers changing their rate plans can also have influence on the revenues they generate. For the purpose of our service revenue forecasting process, we have received the absolute amount of incoming and outgoing customers, as well as the amount of contract renewals per month from the Marketing and Sales department.

For the incoming and renewing customers we also received the breakdown per rate plan cluster to be used in the forecasting process. It is worthwhile mentioning that these figures were based on the signing moment and not on the subscription activation moment. The difference between signing and activation can range up to four months. Revenue generation only occurs after the subscription has been activated. In order to bridge this, we created a translation between activation and signing using historic percentages of customers that activate their subscriptions within zero, one, two, three and four months.

For the outflow of customers, we created the breakdown per rate plan cluster using recent churn history per cluster and the totals acquired from the business. This part of our approach could be improved, but given the time limitation for the project

it was a good enough solution for forecasting purposes. Please note that modeling churn for budgeting purposes and modeling churn for campaigning purposes are processes that are very different. Modeling churn for budgeting purposes requires only the total expected amount of churners, while modeling churn for campaigning purposes requires churn scores for each customer so that they can be approached with an extension offer. Literature mostly addresses churn modeling for campaigning purposes (see chapters 2 and 3). One potential approach to using campaigning churn models for budgeting purposes would be to create the churn scores as true probabilities (not just scores) and then sum these up in order to get the absolute amount of churners. However, this was out of scope of this research for the moment.

We accepted the business plans for inflow, contract renewal and outflow of customers for two different reasons. First, these figures were also used in the standard budgeting process, so they would be useful for benchmarking purposes. Second, the inflow, renewals and outflow of customers are highly dependent on the business plans of the operator (e.g. planned discounts and marketing campaigns), so relying only on historic data without input from the business may not lead to best prediction results.

The monthly state of the clusters for each consecutive month is generated by taking the state of the cluster in the previous month, then adding the planned amount of new "imaginary" customers to the cluster (the customer inflow of that cluster for that month), removing randomly selected customers from that cluster for the case of outflow (the planned amount of churners for that cluster for that month) and also migrating randomly selected customers to different clusters (the planned migrations for that cluster for that month). For the new and the migrating customers, the business also provides the Monthly Recurring Cost, as it is part of their planning. It is worthwhile mentioning that the customers that are selected for churn are still generating revenues in the month they churn, so they were only removed once the revenues were calculated. In the month the churn occurs, they were just marked as churners to be removed from the cluster next month.

6.5.4 Generating the Values of Input Parameters- Scaling Factors

In order to predict the service revenues for one year, we also needed all the inputs listed in Table 6.1 to be extrapolated on a monthly basis for the same period, except for the Monthly Recurring Cost, which is a constant and does not change on a monthly level.

This part of the process was developed by the Mobile Network department of the operator. Forecasting the total usage of voice minutes, SMS and Internet services (total amount of megabytes) on network level is a standard part of their work. They use these forecasts to predict the network load, so that they can intervene before an overload occurs. However, we needed values per rate plan cluster, so they applied the same approach on the cluster data and all the input parameters listed in Table 6.1 (except for the Monthly Recurring Cost). For this purpose, a time series approach

(Hamilton, 1994) implemented in R using the packages tseries (Trapletti and Hornik, 2017) and forecast (Hyndman and Khandakar, 2008) was used. The results of these functions would return average value per month for each parameter and every cluster. We corrected these general monthly trends using the number of working days, weekends and holidays per month because we observed that the total daily traffic for the operator substantially differs for weekdays, weekends and holidays. For example, the total voice minutes per day is much higher on weekdays compared to weekends or holidays. Naturally, the number of days per month also plays a role in the monthly averages. We also addressed the seasonality in the roaming parameters (usage abroad): these parameters have much higher values during holiday periods, especially in the summer months. In the end of this step we created so-called scaling factors which give the ratios between the values of the input parameters of this month and their values in the previous month. These scaling factors were then used to generate the monthly usage of each customer and consequentially for forecasting the service revenues.

6.5.5 Putting it All Together

After generating the monthly values for the inputs and the monthly states of the customer base, they are used for forecasting the total service revenues for one year. This is achieved using the following process:

1. *Get the customer base and values of input parameters for the current month*
2. *Repeat 12 times*
   - *2.1 For each cluster*
     - *2.1.1 Create the customer base for next month by*
       - *2.1.1.1 Removing customers marked for discontinuation of contract in the current month*
       - *2.1.1.2 Adding customers that signed a contract ("imaginary customers") including their Monthly Recurring Cost*
       - *2.1.1.3 Changing the rate plan for customers who renewed their contract including their Monthly Recurring Cost*
       - *2.1.1.4 Marking customers to discontinue in the next month*
     - *2.1.2 Use the scaling factors to create the values for the inputs*
     - *2.1.3 For each input parameter the new customers are given the average value for their cluster*
     - *2.1.4 The input values are plugged into the revenue models corresponding to the cluster to predict the revenues for each customer in the cluster*
     - *2.1.5 Sum the values of all customers in the cluster*
   - *2.2 Sum the revenues of all clusters in that month and move to the next month*

Table 6.2: Algorithm performance

	Amount of Winning Models	Correlation Coefficient	MAE
Linear Regression KNIME	44	0.847	2.210
Linear Regression WEKA	66	0.855	2.105
Polynomial Regression	22	0.818	2.137
Random Forest	19	0.816	2.393
Simple Regression Tree	5	0.727	3.020
Winner		0.858	2.063

Please note that all customers added in step 2.1.1.2 of the procedure above are treated as average customers for that rate plan. The input parameters used for calculating their revenues are averages of the values for the input parameters of all other customers in that cluster in that month, as these are unknown customers and there is no information about their usage.

In the case of forecasting the service revenues as a whole (without a breakdown onto components), step 2.1.4 of this algorithm is performed only once for each cluster, using the service revenues model generated for that cluster. In the case of forecasting the components separately, we run the values of the input parameters through three models (the corresponding models for that cluster for Interconnect Revenues, Out of Bundle Revenues and Roaming Revenues, respectively) and add them to the Monthly Recurring Cost in order to calculate the service revenues:

$$\begin{aligned} TotalServiceRevenues(predicted) = MonthlyRecurringCost \\ +InterconnectRevenues(predicted) \\ +OutOfBundleRevenues(predicted) \\ +RoamingRevenues(predicted) \end{aligned} \tag{6.1}$$

6.6 Results

In this section we will present the results of the service revenue modeling process as well as the results of the forecasting process for one year.

6.6.1 Results of Service Revenues Modeling

As stated in Subsection 6.5.2 of this chapter, we used two different approaches to model the service revenues. The results of the process when modeling service revenues as a whole (not using the breakdown into components) are shown in Table 6.2. The values for the correlation coefficient between the actual values and predicted values, and the mean absolute error (in euros) are shown in the columns Correlation

Table 6.3: Modeling Service Revenue Components

	Correlation Coefficient	MAE
Interconnect Revenues	0.987	0.292
Out Of Bundle Revenues	0.871	1.384
Roaming Revenues	0.935	0.242
Total Service Revenues	0.931	1.918

Coefficient and MAE, respectively. Both were calculated using the actual and predicted value of the total service revenues on the test set using the same algorithm to model all clusters. From all the algorithms we used, the WEKA implementation of linear regression (using a wrapper for preselecting variables for modeling) performed the best in terms of Correlation Coefficient and Mean Absolute Error. It is interesting to note that using more complex algorithms did not improve performance in general, which is similar to the results of Carbonneau et al. (2008).

This is why we created the Winner model, using the best performing algorithm per cluster. In most cases, the WEKA implementation of linear regression is the winner of this contest (column Amount of Winning Models in table 6.2), followed by the KNIME implementation of linear regression, then polynomial regression, random forest regression and simple regression tree. From a methodological perspective, the results of the Winner model should not be compared to the rest, because we used the test set to select the winning model. Strictly speaking, in order to be able to compare the Winner approach with each of the individual models, one should use an intermediate test set (i.e. validation set) to select the best model per cluster and then check the performance of the Winner model on a test set which was not used for selecting it. However, the results of the Winner model we report in Table 6.2 are a good indication that this approach could bring an improvement.

We also tried to test with Tree Ensemble and Support Vector Regression. However, these algorithms did not contribute significantly in reducing the total error while substantially increasing the computation time (to hours per cluster in some cases).

Next, as explained in Subsection 6.5.2, we modeled the total service revenues by modeling each of components of service revenues separately.

The results of this experiment are shown in table 6.3. It is worthwhile mentioning that we calculate the total service revenues according to equation 6.1: by adding together the revenue components and the Monthly Recurring Cost (which is a constant for each customer).

Therefore, the Mean Absolute Error and the Correlation Coefficient are related to the three predictions added together, so it is not a sum of the errors separately. Treating the errors separately results in a Mean Absolute Error of 2.199. However, our objective is to predict the Total Service Revenues better. In order to make a fair comparison to the results in table 6.2, we should be comparing the error on Total Service Revenues from the first experiment with the total error of the second experiment

as shown in table 6.3, which shows an improvement over the first approach (MAE of 1.918 vs. MAE of 2.063).

Even though it is possible that deploying a "winner" approach similar to the one we refer to in table 6.2 onto the separate revenue components could further decrease the overall error, we opted to use linear regression. This was based on both the performance shown in table 6.3 as well as the readability of the linear models, which contributes to the transparency of the model and acceptance by the business.

6.6.2 Results of the Forecasting Process for a Full Year

The goal of the model was to forecast service revenues for a full year. As already stated in the first section of this chapter, in order to predict the service revenues for a full year, it was necessary to first forecast the monthly values of the input parameters as well as the changes in the customer base via the inflow, changes in contract and outflow of customers for the same period.

To validate our approach we used MAE, which is a typical prediction level measure. However, the key measure which was the only relevant measure for the end users of the model and the business altogether was how close the sum of our predictions is to the actual revenues. When predicting the revenues for the first four months of 2017 our model was only 0.3% off from the actual revenues, while the error of the standard budgeting process was 8 times higher. Due to confidentiality, we cannot include detailed numbers. This error measurement makes sense from a business perspective: the objective of the model was to predict total service revenues for one year. It was also the only feasible measure to compare the performance of the old and new process, as the old process could not measure MAE. As stated in Section 1 of this chapter, the old approach was very high level and it only reported a monthly total, without breaking the revenues onto clusters or customers.

We display the forecasting results in Qlikview (QlikTech International AB, 2014), providing drill down opportunities per rate plan, sales channel, voice minutes or Internet bundle etc. The business has gained better insights into the forecast than the actuals, resulting in a request to update operational reporting on the actuals in the same way.

The model is currently under review by the operator's financial controlling group. On top of the achieved performance in terms of accuracy it also substantially reduced manual labor in the revenue forecasting process. This model also provided an opportunity for testing multiple scenarios, which was previously not possible.

6.7 Discussion and Lessons Learned

One of the most difficult challenges we were facing during this research was to gain acceptance of the end-customer of the model. Several factors influenced this positively, but the key factor was the transparency of the model. Therefore, we

opened the source data, the forecast for the input parameters and the regression models to the end-users. Using linear regression was a key factor here, due to readability. If we were using more complex models, it would be almost impossible for the users to understand them in detail and get past the black-box phenomenon. Therefore, even though in some cases the more complex models lowered the MAE, we opted for a more uniform and transparent model. Furthermore, when comparing the performance between the old and new approach we used a simple error measure, which was the difference between the outcome of either of the methods and the actual results. Even though we agree this is not the most precise way of estimating the error, it was pivotal for the success of our method, because we were tasked to forecast the revenues for a full year. Of course, the monthly errors were also shown, so the users could see that there were no large fluctuations.

Most importantly, we used the same data sources and definitions of revenues as the financial department (the end user). However, we did create our own data repository where all the data was stored together. Therefore, we had to verify with the end users that the historic totals per month were matching. Transforming the data from the source systems into a format that was suitable for our analysis required a long time and a lot of domain expert knowledge, but it also helped us get the end users on board in the process. A data repository similar to ours can also be very useful for monthly financial reporting, hence a similar repository is now being built for operational purposes.

From a technical perspective, one of the lessons learned was to keep all tables in memory, instead of on disk. We used a server machine with 128GB of RAM memory (standard off the shelf hardware), which was sufficient for this purpose. This reduced the run time of the total modeling process from one hour to 10 minutes when using linear regression (when modeling the total service revenues by components we generate 468 models- three for each of the cluster). When using the more complex algorithms, the training for some clusters lasted longer than a few hours to calculate, especially when the sample size was large (more than 30,000 samples).

Trying to forecast the service revenues for a full year created a challenge from a run time perspective: our first run was too slow- 52 hours for predicting only four months ahead (so just one third of the task). We were able to reduce this to about 15 minutes per month of prediction (or approximately 3 hours of total runtime) by using R instead of KNIME for some data transformations, keeping everything in memory and only writing to disk once the full month was calculated, as well as optimizing the process flow of generating "new" random average customers for the inflow and contract changes and removing randomly selected customers from the base for the outflow. The main advantage of R versus KNIME was that table joins and aggregations using the dplyr package (Wickham and Francois, 2016) performed much faster. Ultimately, the whole process would run fastest if it was entirely performed in R. However, we are at the moment using the combination of R and KNIME for two reasons. First, there is no business requirement to further optimize performance, as

this process already replaces months of manual work with approximately 3 hours execution time. Second, this combination of tooling is much easier to debug and detect errors if they occur, given that KNIME displays the modeling pipeline in a graphical manner.

The short runtime and the flexibility of changing the inputs created opportunities for testing of multiple scenarios. For example, we used the inflow of customers, renewing contracts and outflow as they are provided by the business, using their default planning as a baseline. Furthermore, we used the historic data for the revenue models' inputs (the usage of services) to predict their values in the future. Our approach offers testing different scenarios for both the inflow, outflow and renewing contracts, as well as different levels of changes in the inputs for the revenue formulas. For example, instead of only relying on the historic data for making assumptions on how the inputs for the revenue models will develop in the future, the operator could now test the effect of increased or decreased usage onto service revenues. This particular functionality of our approach enabled the operator to test their assumptions stemming from the European regulations on abolishing roaming charges within the EU. Furthermore, one can also simulate different pricing levels (mostly via the Monthly Recurring Cost for new customers or customers migrating to a different cluster), which was not possible with the standard forecasting process. Being able to simulate on micro level across three different dimensions (changes in customer base, usage and pricing) created a versatile simulation platform that could be of interest to researchers or data mining practitioners in many different fields.

One interpretation of Pareto's rule (Pareto, 1964), especially popular in business, is that 80 percent of the result can be accomplished with 20 percent of the effort (Koch, 2011). In business circumstances, 80 percent is very often enough. Although the runtime of approximately 3 hours can be optimized further, this will only be done once the model is in full production use and if requested by the end users, keeping in mind that the runtime of the approach previously used for this purpose was measured in weeks.

Our approach can generalize to revenue simulation of any subscription based service with usage based pricing. Furthermore, it could be applicable to supply chain management, or the demand side of it, if applied to industries that have multiple products and many different customers purchasing them, for example fast moving consumer goods. One can imagine creating clusters of various products and forecasting their consumption and consequentially the revenues this would generate. Another potential application of this method in telecommunications would be adapting it to the supply chain management of mobile phone purchases, as each phone comes in multiple configurations (even colors). Storing these devices for a long time is expensive, as they lose their value over time when they are not sold, so creating a low level granularity model per device would be valuable.

6.8 Limitations and Future work

One of the limitations of this research is that we randomly selected customers from each cluster to exclude when modeling the outflow, using the quantities received from the business. This could be improved by calculating actual churn probabilities for each customer in a given cluster and excluding the ones with the highest rank, even if we were to accept the quantities provided by the business. The same scenario could apply for customers that are migrated to a different cluster (renewing customers).

An additional limitation of this research is accepting the total figures for inflow, outflow and migrations of customers per month as created by the business. We do not have detailed knowledge on how these are created. As part of our future work, we envision improving some of the business processes we took as business input in this research. For example, forecasting the amount of outgoing customers and the breakdown per rate plan can be improved by using survival analyses (Miller Jr, 2011) or fitting churn curves. The same applies for renewing customers. We already started an initial investigation into these processes.

Using these two approaches, we could improve the churn and renewal forecasting both from the perspective of total customers that will churn or migrate, as well as which customers would churn or migrate from each cluster.

As a part of our future work, we could also envision applying an approach similar to this to the field of supply chain management in mobile telecommunications, more specifically the mobile phone demand mentioned in the discussion section.

6.9 Conclusion

Addressing the research question from section 6.1, we have managed to generalize the approach we developed in chapter 5 to the financial domain, and use data mining to predict and simulate service revenues in telecommunications. We used generic hardware, open source tools for the majority of the process, and simple algorithms in a complex deployment flow to optimize a key business problem. The added value of this model is not only in enabling the business to have much better and much more timely insights in future revenues, assuming that the standard business plans for customer base growth are implemented. This model provides a scenario simulation platform giving the business an opportunity to test the potential measures designed to increase revenues at a much higher pace.

Chapter 7

Summary and Conclusion

In this thesis we researched various applications of data mining in telecommunications. We had a unique opportunity to do research in a real world setting, using commercial data sets. We worked on broad range of business problems in mobile telecommunications. Our research stretched between the areas of marketing, mobile network technology and finance. We addressed both segments in mobile telecommunications, prepaid and postpaid customers. In chapters 2 and 3 we focused on prepaid customer churn with a target to reduce it by identifying prospective churners. In chapter 4 we built a bridge between a churn model, which is a typically marketing problem, and the mobile network domain in order to prevent it. In chapter 5 we created a simulation platform for mobile network load and in chapter 6 we created a new approach to service revenue forecasting. In this chapter we will summarize the key content and lessons learned from the previous chapters in more detail. We will also address the research questions per chapter and the general research question of this thesis.

In chapter 2, we presented how performance of prepaid churn models changes when varying conditions in three different dimensions: data- by adding CEM (Customer Experience Management) parameters; population sample- by limiting the inactivity period at the time of recording to 15, 30 and zero days, respectively; and outcome definition- by introducing a so-called grace period of 15 days after the time of recording, in which customers must make an activity in order to be classified as churners.

From a theoretical perspective, we created a CEM (Customer Experience Management) Framework for Mobile Telecommunications. Unfortunately, adding the CEM parameters into the models did not add substantial value in terms of model performance under any of the experimental conditions. Similarly, switching the population sample on the period of inactivity at the time of recording between 15 and 30 days did not influence model performance, only the sample size and churn rate. When we changed the population sample by disallowing inactivity at the time of recording, apart from the change in sample size and churn rate there was also a drop in performance and stability of the models. However, this drop in performance was not nearly as high as the one that occurred when changing the outcome definition by setting a grace period, thus making the behavior to be predicted more complex. This change obviously influenced the churn rate as well. Nevertheless, the latter two approaches provided more time for retention. Therefore, looking at the research question posed in section 2.1, which one of the three variations in the experimental setup has the highest influence on prepaid churn modeling, the answer is clearly that changing the outcome definition had a much higher influence on the performance of the prepaid churn models than adding the CEM parameters or changing the characteristics of the sample based on inactivity at the time of recording.

Throughout chapter 3 we have investigated the extent to which social network information can be used to predict telecom churn, and how this information could potentially improve the predictive performance of conventional churn prediction

methods. We have assessed the performance of models constructed using classical tabular data mining, social network mining and hybrid models combining both techniques. The first hybrid model was built by extending the traditional tabular churn predictors with social network variables extracted from the social graph. The second hybrid model was obtained by incorporating the results of a traditional tabular churn model into the social propagation graph, using them as initial energies of the non-churner nodes.

The performance of our models was verified using a large data set of 700 million call data records. Our initial observation showed that the churn probability was positively aligned with the number of churned neighbors. The regular tabular churn models constructed exclusively using social network information and the traditional social network models scored the least. This indicates that social network information alone was not sufficient to predict churn. Overall, the traditional tabular churn models had the best predictive performance. The added value of the social network variables to the tabular churn models was rather minimal. Although the second hybrid models were able to outperform the regular propagation models, they still could not beat the performance of the traditional tabular churn model. The contribution of traditional predictors to churn prediction was substantially higher than that of the social network behavior. Moreover, the performance gain of both hybrid models was not substantial enough to justify the computational costs. In a nutshell, the answer to the research question posed in section 3.1 is that social network mining and attributes stemming from a social network graph did not add substantial value in terms of model performance to traditional prepaid churn modeling in T-Mobile Netherlands.

In chapter 4 we presented an atypical approach to churn management in commercial settings. Utilizing parts of the CEM Framework we created in chapter 2, we succeeded in explaining at least a part of the postpaid churn via actual measurements of network quality. The main benefits of our approach were the following. First, we managed to build an explanatory churn model by sacrificing only a part of the performance. Second, our churn model was based on features that were extracted from actual network parameters rather than surveys (real network experience vs. perception). Third, this model generated insights on which network parameters were necessary to be corrected in order to reduce churn, which is a new way of churn reduction. This model was built to explain churn and prevent customers from wanting to churn, rather than identifying prospective churners. The generated insights caused a shift from network centricity towards customer centricity in managing the telecom network, meaning focusing on sites where a large number of customers experience network problems rather than on sites where a high number of network problems occur (which could be caused by a malfunctioning of a single phone). Using this approach, the churn mitigation process is no longer just a retention campaign. The churn reduction efforts are no longer the responsibility of just the CRM teams, Marketing and Customer service, but also the Technology department. Managing the

network in a customer centric way is now part of the process and certain customer centric measurements were even set as targets for the Technology department. This resulted into a substantial reduction in the numbers of customers having poor network experience (e.g. the amount of customers experiencing more than 1 dropped call per week has been substantially reduced over two years). Finally, our research has already contributed to reduced dissatisfaction with the network, increased overall customer satisfaction and churn reduction (information proprietary to T-Mobile Netherlands). Therefore, referring to the research question stated in section 4.1, we have managed to use a different deployment form of a churn model in order explain and prevent churn rather than directly target customers.

In chapter 5 we had a goal of using data mining for prediction and simulation of 3G mobile network air interface load. We used a new way to deploy models resulting in a very simple yet effective approach of deploying data mining in commercial surroundings. Unfortunately, data mining is still seen as a black box in many industries, telecom not excluded. Even though some data mining activities are taken, typically in the Marketing/Customer Retention field, there is a myriad of other possibilities in business where data mining can be applied. In our opinion, it is better to start with simple methods, such as linear regression, because it is easier to understand them. Once these simple approaches gain acceptance, and familiarize the industries with data mining, opportunities to apply more advanced techniques will arise. Last but not least, for deployment we used tools that are already familiar to the end-users. This all resulted in a high acceptance of this model and users coming up with their own use cases, which were not originally intended. This model was deployed in multiple national operators of the Deutsche Telekom group. Addressing the research question posed in section 5.1, we have shown how data mining can be used to predict 3G mobile network interface load, and simulate it under different scenarios: we have used relatively simple algorithms to create a large number of predictive models, therefore making possible predicting the load on a cell level. We have used tools known to the end users to deploy these models, allowing them to use different scenarios for input parameters. Acceptance was gained by decoupling the data mining process from the end users, but keeping the transparency that linear regression offers combined with tooling familiar to them.

In chapter 6 we extended this approach onto the field of revenue forecasting. The added value of this model is not only in enabling the business to have much better and much more timely insights in future revenues, assuming that the standard business plans for customer base growth are implemented. This model provides a scenario simulation platform giving the business an opportunity to test the potential measures designed to increase the revenues at much higher pace. Addressing the research question from section 6.1, we have managed to generalize the approach we developed in chapter 5 to the financial domain, and use data mining to predict and simulate service revenues in telecommunications.

Some of the important lessons we learned can be generalized as follows.

In chapter 2 a key lesson is to spend more time defining the problem correctly than picking the right algorithm to solve it or looking for even more data to mine in. This can be a quite interesting finding for many companies who are constantly approached by vendors selling new platforms (more data to mine on) promising to reduce churn across all populations.

Similarly, in chapter 3 we learned that algorithms with high computational cost do not necessarily add predictive power, especially in cases when features traditionally used for the same purpose are already rich.

In chapter 4 we learned that it is worthwhile to try to formulate the problem differently or to look at it from another perspective. The value of an explanatory model can be very high as it can enable one to prevent the problem rather than cure it. Furthermore, the same parameters which did not improve the prepaid churn models in chapter 2, were very useful in designing the explanatory churn model for postpaid customers. This can be seen as a variation of the No Free Lunch theorem related to predictors: Just because certain features did not work on one part of a population (prepaid customers), does not mean they will not have high performance on a different population (postpaid customers). This is not contradictory to the lesson learned from chapter 2: we are just saying that adding more data (a new platform) to mine in does not improve performance on **ALL** problems and populations, but it still may be useful in certain situations.

In chapter 5 we learned that combining simple algorithms in complex deployment forms can reap great benefits. By extending the approach of chapter 5 in chapter 6 we learned that the answer to a company's most important questions can be found by reusing an approach developed for a completely different purpose. In both these chapters we learned that using predictions of inputs for forecasting the output variables created a powerful deployment platform for scenario simulation, resulting in use cases that were not originally foreseen.

In our research we were using a combination of open source software and commercial software. Generally, the commercial software used in some parts of our research can easily be replaced by open source counterparts. However, the automated data preparation (attribute discretization and grouping) as executed in the Predictive Analytics Director software (Pegasystems, 2008), did substantially reduce our workload. The hardware we used was mostly standard of the shelf servers or just a laptop. For deployment we used tools already familiar to the end-users. In general, we did not require additional hardware, software or data from what was already available. Metaphorically speaking, we cooked a meal with what was already in the house. This drives down the cost of applying these solutions, not only in the telecom industry.

From our perspective, the Business Understanding, Data Understanding and Data Preparation stages of the CRISP-DM process were very important. We have learned that in commercial settings the data is very often spread across various sources, so it is essential to unify it. In certain cases, we used existing data sources

and simply imported the necessary data, while in other we created a completely new data repository, which was also useful for operational purposes. Quite often, very rich features already existed it the data, but a so called data dictionary (what does every variable measure) did not. Therefore, domain expert help was crucial for both data preparation and understanding, and of course business understanding. These three stages in an industry setting represent a large part of the overall effort. Having these properly executed results in a rich data set where simple algorithms can perform very well. With regard to data preparation, apart from variable discretization, dealing with missing and extreme values, our choices here were ranging from how much history is relevant to the problem, to generating new variables via different levels of raw data aggregations (hourly, weekly, monthly) and taking into account ratios of these aggregates in order to capture changes in behavior over time.

The deployment step was also very important. This is especially visible in chapters 4, 5 and 6. In chapter 4, the result was not a classic churn model, but rather a set of guidelines for domain experts on what to improve. In chapters 5 and 6, the deployment was actually the key part of the process. Even though the modeling part resulted in a large number of models in a very short time, the method of deploying, which was to first generate new values for the inputs and then use them for forecasting is the part that brought the most of the value. The business now has opportunities to run micro level scenario simulations for two very important business processes. Furthermore, presenting the results using tools familiar to the users also helped the acceptance of the whole process. An overall lesson from a project management perspective was to consult the domain experts who are also the end users of the solutions every step of the way. This is very useful for both setting up the projects at the beginning (at the business understanding, data understanding and preparation stage), as well as for the end- user acceptance. In chapter 5, the users themselves were coming up with new scenarios (use cases)[1]. Last but not least, in business settings the evaluation of models normally does not stop by measuring performance on a test set (pre-labeled data): models are deployed in practice and the performance is benchmarked against actual measured values.

One interpretation of Pareto's rule (Pareto, 1964), especially popular in business, is that 80 percent of the result can be accomplished with 20 percent of the effort (Koch, 2011). One possible translation of this rule to data mining could be that 80 percent of the possible performance improvement can be accomplished with 20 percent of the effort (or time). The timeliness of the solution is often more important than ultimate accuracy. For example, while we are designing the perfect churn model a lot of customers can already be gone. In business, a preferred way is to deploy a solution that is good enough (performs better than the baseline) and improve it later. In some cases, the execution or scoring time of the model is also important: e.g. personalizing a website based on a predictive model: the model cannot cause a high delay in the time necessary to load the page, otherwise the customer might not be patient enough

[1]The approach from Chapter 6 is still under acceptance at the operator

to wait. Applying simple and fast algorithms on inexpensive hardware using open source software can help many organizations that struggle with budgets find at least a temporary solution, which is better than no solution at all (e.g. early detection of diseases).

Looking at the problems we were addressing in this thesis in chronological order, one may conclude that we were progressively addressing problems with higher business importance in each consecutive step. In chapters 2 and 3 we were addressing prepaid churn, which is an important business problem, but in large operators, such as the one where we were conducting our research, prepaid revenues are not nearly as substantial as the revenues from the postpaid segment. The good performance of our models (any of our models were better than the model deployed at the time- not stated in the chapter, as it was out of scope of the research) contributed to getting the task of explaining postpaid churn. After the successful delivery of this task, we were assigned to forecasting network load, which has a huge impact on the budgeting process of a telecom operator, as a large part of the budget is dedicated to network improvements. Last but not least, we got the task to forecast the revenues generated by the postpaid segment of the operator, which is one of the most important financial tasks in a company. One can see this as a journey to accepting data mining in business. Initially, we started with a problem that is important, but not on the top of the list. Solving each consecutive problem was gaining trust, so data mining gained acceptance as a solution to even most important business problems.

To summarize, we believe we have successfully answered the overall research question of this thesis, which was how does one successfully apply data mining in telecommunications? Our approach to this was to focus at the stages of CRISP-DM less covered in literature: business understanding, data understanding and preparation, and in particular deployment. We have used relatively simple algorithms, which performed well on these large datasets and intuitive performance measurements that were easy to explain to the business. We used hardware, software and data which were already available, or added open source software to keep the costs low. We created innovative deployment mechanisms using tools familiar to the end users and involved the users early on in the process. This all has led to the business accepting data mining as a solution to a much broader range of problems than before.

During this research we have paid due attention to the legal and privacy related aspects. European Data Privacy rules are very rigorous about what can and cannot be done with customer data. It is worthwhile mentioning that in chapter 2 and chapter 3 we were working on data from prepaid customers, which do not provide their private information (name, address etc.) to the operator. In chapters 3 and 4, the churn models generated were not used for campaigning. In chapter 3, the intention was to only analyze whether social ties add value to predicting churn. There was no added value, which was an additional reason to not deploy the model. In chapter 4, from the research setup onwards, we were looking for an explanation of churn, not another campaigning model. The results of the model were used to improve

customer experience. For the revenue forecast model in chapter 6, we only used the data that the operator has to store, as mandated by Law, in order to be able to reproduce customer invoices, if necessary. Furthermore, the data used for research in all chapters has been anonymized, except for chapter 5, where we were working on data from network cells, which does not contain any customer identifiers, as this data is already aggregated on a cell level. Unfortunately, in recent times data mining and machine learning techniques have a damaged reputation with relation to privacy (e.g. the case of Edward Snowden, PRISM and NSA; or Google's unification of user profiles for all services; or most recently, Facebook merging WhatsApp and Facebook profile information, for which they were fined €110 million by the EU). We have shown methods to gain valuable insights from customer data which are not in breach of any privacy rights or regulations, neither legally nor ethically.

From a generalization perspective, this thesis is applicable to many industries. Clearly, in almost any industry losing customers to competition (churn), managing the key resource and forecasting revenues are very important problems. Finding inexpensive means to address these issues is beneficial to many companies or governments. Using simple algorithms that can easily be explained and deployment methods preferred by end users helps acceptance. This is how we see data mining being applied and accepted in many fields outside telecommunications, helping more organizations become data driven. From a research perspective, we hope that we have shown that there are many other interesting problems to solve beyond building a better performing predictive model. Hopefully, by applying data mining in telecommunications, we will raise academic interest in finding better and more efficient ways of disseminating machine learning research.

Bibliography

3gpp (1999), 'RNSAP Cell Load Information Procedure and Message Contents'. Retrieved from `http://www.3gpp.org/ftp/tsg_ran/wg3_iu/TSGR3_07/Docs/Pdfs/R3-99c57.PDF`.

Ahn, J. H., Han, S. P. and Lee, Y. S. (2006), 'Customer Churn Analysis: Churn Determinants and Mediation Effects of Partial Defection in the Korean Mobile Telecommunications Service Industry', *Telecommunications Policy* **30**(10), 552–568.

Alberts, L. J. S. M. (2006), Churn Prediction in the Mobile Telecommunications Industry, Master's thesis, Department of General Sciences, Maastricht University.

Anil Kumar, D. and Ravi, V. (2008), 'Predicting Credit Card Customer Churn in Banks Using Data Mining', *International Journal of Data Analysis Techniques and Strategies* **1**(1), 4–28.

Archaux, C., Martin, A. and Khenchaf, A. (2004), 'An SVM Based Churn Detector in Prepaid Mobile Telephony', *International Conference on Information and Communication Technologies (ICTTA)* pp. 19–23.

Au, W., Chan, K. and Yao, X. (2003), 'A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction', *IEEE Transactions on Evolutionary Computation* **7**(6), 532–545.

Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X. (2006), 'Group Formation in Large Social Networks: Membership, Growth and Evolution', *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 44–54.

Ballings, M. and Van den Poel, D. (2012), 'The Relevant Length of Customer Event History for Churn Prediction: How long is long enough?', *Expert Systems with Applications* **39**(18), 13517–13522.

Bermolen, P. and Rossi, D. (2009), 'Support Vector Regression for Link Load Prediction', *Computer Networks* **53**(2), 191–201.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. and Wiswedel, B. (2007), KNIME: The Konstanz Information Miner, *in* 'Studies in Classification, Data Analysis and Knowledge Organization (GfKL 2007)', Springer.

Birke, D. and Swann, G. P. (2006), 'Network Effects and The Choice of Mobile Phone Operator', *Journal of Evolutionary Economics* **16**(1–2), 65–84.

Borgatti, S. (1994), 'A Quorum of Graph Theoretic Concepts', *Connections* **17**(1), 47–59.

Buckinx, W. and Van den Poel, D. (2005), 'Customer Base Analysis: Partial Defection of Behaviourally Loyal Clients in a Non-Contractual FMCG retail Setting', *European Journal of Operational Research* **164**(1), 252–268.

Carbonneau, R., Laframboise, K. and Vahidov, R. (2008), 'Application of Machine Learning Techniques for Supply Chain Demand Forecasting', *European Journal of Operational Research* **184**(3), 1140–1154.

Caruana, R. (1997), 'Multitask Learning', *Machine Learning* **28**(2), 41–75.

Christiansen, T. and Torkington, N. (2003), *Perl Cookbook*, 2 edn, O'Reilly, Sebastopol.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S. and Nanavati, A. A. (2008), 'Social Ties and Their Relevance to Churn in Mobile Telecom Networks', *Proceedings of the 11th International Conference on Extending Database Technology* pp. 668–677.

Datta, P., Masand, B., Mani, D. and Li, B. (2000), 'Automated Cellular Modeling and Prediction on a Large Scale', *Artificial Intelligence Review* **14**, 485–502.

Dougherty, J., Kohavi, R. and Sahami, M. (1995), Supervised and Unsupervised Discretization of Continuous Features, *in* A. Prieditis and S. Russell, eds, 'Proceedings of the Twelfth International Conference on Machine Learning', Morgan Kaufmann, San Francisco, CA, pp. 194–202.

Ericsson (2013), 'Ericsson Performance Management'. Retrieved from `http://www.ericsson.com/ourportfolio/products/performance-management`.

Feenberg, D. R., Gentry, W. M., Gilroy, D. and Rosen, H. S. (1989), 'Testing the Rationality of State Revenue Forecasts', *Review of Economics and Statistics* **71**, 300–308.

Feinberg, E. and Genethliou, D. (2010), Load Forecasting, *in* J. Chow, F. Wu and J. Momoh, eds, 'Applied Mathematics for Restructured Electric Power Systems', pp. 269–285.

Ferreira, J., Vellasco, M., Pachecco, M. and Barbosa, C. (2004), 'Data Mining Techniques on the Evaluation of Wireless Churn', *ESANN 2004 Proceedings – European Symposium on Artificial Neural Networks* pp. 483–488.

Fullerton, T. M. (1989), 'A Composite Approach to Forecasting State Government Revenues: Case Study of the Idaho Sales Tax', *International Journal of Forecasting* **5**(3), 373–380.

Geijer Lundin, E., Gunnarsson, F. and Gustafsson, F. (2003), Uplink Load Estimation in WCDMA, *in* 'IEEE Wireless Communications and Networking Conference', Vol. 3, pp. 1669–1674.

Grant, G. G. (2003), *ERP & Data Warehousing in Organizations: Issues and Challenges*, IGI Global.

Gyan, R., Hui, Z., Zhi-Li, Z. and Jean, B. (2012), 'Are Call Detail Records Biased for Sampling Human Mobility?', *ACM SIGMOBILE Mobile Computing and Communications Review* **16**(3), 33–44.

Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2006), 'Churn Prediction Using Complaints Data', *International Journal of Intelligent Technology* **13**, 158–163.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009), 'The WEKA Data Mining Software: An Update', *SIGKDD Explorations* **11**(1).

Hamilton, J. D. (1994), *Time Series Analysis*, Vol. 2, Princeton university press Princeton.

Harell, F. E. J. (2001), *Regressing Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York: Springer.

Heikkilä, J. (2002), 'From Supply to Demand Chain Management: Efficiency and Customer Satisfaction', *Journal of Operations Management* **20**(6), 747–767.

Huawei (2013), 'iManager M2000'. Retrieved from `http://www.huawei.com/en/products/oss/mbb-om-product/imanager-m2000/`.

Hung, S., Yen, D. C. and Wang, H. (2006), 'Applying Data Mining to Telecom Churn Management', *Expert Systems with Applications* **31**(3), 515–524.

Hwang, H., Jung, T. and Suh, E. (2004), 'An LTV Model and Customer Segmentation Based on Customer Value: A Case Study on the Wireless Telecommunication Industry', *Expert Systems with Applications* **26**, 181–188.

Hyndman, R. J. and Khandakar, Y. (2008), 'Automatic Time Series Forecasting: The Forecast Package for R', *Journal of Statistical Software* **26**(3), 1–22.
URL: *http://www.jstatsoft.org/article/view/v027i03*

IBM (2017), '10 Key Marketing Trends for 2017'. Retrieved from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN.

Idris, A., Khan, A. and Lee, Y. S. (2013), 'Intelligent Churn Prediction in Telecom: Employing mRMR Feature Selection and RotBoost Based Ensemble Classification', *Applied Intelligence* **39**(3), 659–672.

Kass, G. V. (1980), 'An Exploratory Technique for Investigating Large Quantities of Categorical Data', *Applied Statistics* pp. 119–127.

Kawale, J., Pal, A. and Srivastava, J. (2009), 'Churn Prediction in MMORPGs: A Social Influence Based Approach', *Proceedings of the 2009 International Conference on Computational Science and Engineering* **4**, 423–428.

Kendall, M. (1938), 'A New Measure of Rank Correlation', *Biometrika* **30**(1-2), 81–89.

Kim, H. and Yoon, C. (2004), 'Determinants of Subscriber Churn and Customer Loyalty in the Korean Mobile Telephony Market', *Telecommunications Policy* **28**(9–10), 751–765.

Koch, R. (2011), *The 80/20 Principle: The Secret to Achieving More with Less*, Crown Business.

Kohavi, R. and John, G. (1997), 'Wrappers for Feature Subset Selection', *Artificial Intelligence* **97**, 273–324.

Kraljevic, G. and Gotovac, S. (2010), 'Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services', *Automatika* **51**(3), 375–383.

Kusuma, P. D. (2013), Extending Traditional Telecom Churn Prediction Using Social Network Data, Master's thesis, Leiden University. Retrieved from http://www.liacs.nl/assets/2013-03PalupiKusuma.pdf.

Kusuma, P. D., Radosavljevik, D., Takes, F. W. and van der Putten, P. (2013), Combining Customer Attribute and Social Network Mining for Prepaid Mobile Churn Prediction, *in* 'Proc. the 23rd Annual Belgian Dutch Conference on Machine Learning (BENELEARN)', pp. 50–58.

Lemmens, A. and Croux, C. (2006), 'Bagging and Boosting Classification Trees to Predict Churn', *Journal of Marketing Research* **43**(2), 276–286.

Li, J. and O'Donoghue, C. (2013), 'A Survey of Dynamic Microsimulation Models: Uses, Model Structure and Methodology', *International Journal of Microsimulation* **6**(2), 3–55.

Lima, E., Mues, C. and Baesens, B. (2009), 'Domain Knowledge Integration in Data Mining Using Decision Tables: Case Studies in Churn Prediction', *Journal of the Operational Research Society* **60**, 1096–1106.

Mäder, A. and Staehle, D. (2004), Analytic Modeling of the WCDMA Downlink Capacity in Multi-Service Environments, *in* '16th ITC Specialist Seminar', pp. 229–238.

Malhotra, A. and Malhotra, C. K. (2013), 'Exploring Switching Behavior of US Mobile Service Customers', *Journal of Services Marketing* **27**(1), 13–24.

Marr, B. (2015), 'Big Data: 20 Mind-Boggling Facts Everyone Must Read', Forbes. Retrieved from `https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read`.

Meyer, C. and Schwager, A. (2007), 'Understanding Customer Experience', *Harvard Business Review* **85**(2), 116–126.

Microsoft Corporation (2010), 'Microsoft Excel'. Retrieved from `http://office.microsoft.com/en-us/excel/`.

Miller Jr, R. G. (2011), *Survival Analysis*, Vol. 66, John Wiley & Sons.

Min, D. and Wan, L. (2009), 'Switching Factors of Mobile Customers in Korea', *Journal of Service Science* **1**(1), 105–120.

Motahari, S., Mengshoel, O. J., Reuther, P., Appala, S., Zoia, L. and Shah, J. (2012), 'The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records', *In Proceedings of the 6th SNA KDD Workshop* pp. 9–17.

Mozer, M., Wolniewicz, R., Johnson, E. and Kaushansky, H. (1999), 'Churn Reduction in the Wireless Industry', *Proceedings of the Neural Information Systems Conference* .

Muckenheim, J. and Bernhard, U. (2001), A Framework for Load Control in 3rd Generation CDMA Networks, *in* 'In Proc of the IEEE Global Telecommunications Conference', Vol. 6, pp. 3738–3742.

MyCom (2013), 'NIMS-PrOptima Service & Network Performance Management'. Retrieved from `http://www.mycom-int.com/products/nims-proptima-service-and-network-performance-solution/`.

Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S. and Joshi, A. (2006), On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications, *in* 'Proceedings of the 15th ACM International Conference on Information and Knowledge Management', ACM, pp. 435–444.

Natalizio, E., Marano, S. and Molinaro, A. (2005), Packet Scheduling Algorithms for Providing QoS on UMTS Downlink Shared Channels, *in* 'IEEE VTC', Vol. 4, pp. 2597–2601.

Neo4j (2012), 'Neo4j: Community Edition (Version 1.8.M05) [Software]'. Retrieved from `http://Neo4j.org/`.

Neslin, S., Gupta, S., Kamakura, W., Lu, J. and Mason, C. (2006), 'Detection Defection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models', *Journal of Marketing Research* **43**(2), 204–211.

Ngonmang, B., Viennet, E. and Tchuente, M. (2012), Churn Prediction in a Real Online Social Network Using Local Community Analysis, *in* 'Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 282–288.

Nokia Siemens Networks (2008), 'Rel. RU10- System Library, v.1: RNC Counters - RNW Part'. Proprietary and Confidential.

Oentaryo, R. J., Lim, E. P., Lo, D., Zhu, F. D. and Prasetyo, P. (2012), 'Collective Churn Prediction in Social Networks', *In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* pp. 210–214.

Oracle (2011), 'Oracle Database Documentation Library'. Retrieved from `http://www.oracle.com/pls/db102/homepage`.

Pareto, V. (1964), *Cours d'Économie Politique*, Vol. 1, Librairie Droz.

Pegasystems (2008), 'Predictive Analytics Director (Version CDM 6.3) [Software]'. Retrieved from `http://www.pega.com/products/decision-management`.

Perner, P. (2002), *Data Mining on Multimedia Data, LNCS, vol. 2558*, Springer Verlag, Berlin.

Pine, B. J. I. and Gilmore, J. H. (1999), *The Experience Economy*, Harvard Business School Press, Boston, MA.

Polepally, A. and Mohan, S. (2012), Behavior Analysis of Telecom Data Using Social Networks Analysis, *in* 'Behavior Computing', Springer London, pp. 291–303.

Prasad, U. D. and Madhavi, S. (2012), 'Prediction of Churn Behavior of Bank Customers Using Data Mining Tools', *Business Intelligence Journal* **5**(1), 96–101.

QlikTech International AB (2014), 'QlikView Personal Edition (Version 11.2)'. Retrieved from `http://us-d.demo.qlik.com/download`.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Radosavljevik, D. and van der Putten, P. (2013), Preventing Churn in Telecommunications: The Forgotten Network, *in* 'International Symposium on Intelligent Data Analysis', Springer, pp. 357–368.

Radosavljevik, D. and van der Putten, P. (2014), Large Scale Predictive Modeling for Micro-Simulation of 3G Air Interface Load, *in* 'Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 1620–1629.

Radosavljevik, D. and van der Putten, P. (2017), Service Revenue Forecasting in Telecommunications: A Data Science Approach, *in* W. Duivesteijn, M. Pechenizkiy, G. Fletcher, V. Menkovski, E. Postma, J. Vanschoren and P. van der Putten, eds, 'Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning', pp. 187–189.

Radosavljevik, D., van der Putten, P. and Kyllesbech Larsen, K. (2010*a*), 'The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter?', *Transactions on Machine Learning and Data Mining* **3**(2), 80–99.

Radosavljevik, D., van der Putten, P. and Kyllesbech Larsen, K. (2010*b*), The Impact of Experimental Setup on Prepaid Churn Modeling: Data, Population and Outcome Definition, *in* I. Bichindaritz, P. Perner and G. Russ, eds, 'Advances in Data Mining, Workshop Proceedings', IBaI Publishing, Leipzig, pp. 14–27.

Radosavljevik, D., van der Putten, P. and Kyllesbech Larsen, K. (2012), Mass Scale Modeling and Simulation of the Air-Interface Load in 3G Radio Access Networks, *in* 'Advances in Intelligent Data Analysis XI', Springer Berlin Heidelberg, pp. 301–312.

Richter, Y., Yom-Tov, E. and Slonim, N. (2010), 'Predicting Customer Churn in Mobile Networks Through Analysis of Social Groups', *Proceedings of the SIAM International Conference on Data Mining* pp. 732–741.

Saravanan, M. and Raajaa, G. V. (2012), A Graph-Based Churn Prediction Model for Mobile Telecom Networks, *in* 'Advanced Data Mining and Applications', Springer Berlin Heidelberg, pp. 367–382.

Schmitt, B. H. (2003), *Customer Experience Management: A Revolutionary Approach to Connecting with Your Customers*, John Wiley and Sons, Hoboken, NJ.

Seo, D., Ranganathan, C. and Babad, Y. (2008), 'Two-level Model of Customer Retention in the US Mobile Telecommunications Service Market', *Telecommunications Policy* **32**(3), 182–196.

Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C. and Leskovec, J. (2008), 'Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions', *In Proceedings of the 14th ACM SIGKDD* pp. 596–604.

Smith, S. and Wheeler, J. (2002), *Managing the Customer Experience: Turning Customers into Advocates*, FT Prentice Hall, Harlow, USA.

Strawberry Perl (2011), 'Strawberry Perl'. Retrieved from `http://strawberryperl.com/`.

Svoboda, P., Buerger, M. and Rupp, M. (2008), Forecasting of Traffic Load in a Live 3G Packet Switched Core Network, *in* J. Chow, F. Wu and J. Momoh, eds, 'Proc. of 6th International Symposium on CNSDSP', pp. 433–437.

Teradata Corporation (2017), 'Teradata Online Library'. Retrieved from `http://info.teradata.com/HTMLPubs/DB_TTU_15_00/index.html`.

Trapletti, A. and Hornik, K. (2017), *Tseries: Time Series Analysis and Computational Finance*. R package version 0.10-40.
URL: *https://CRAN.R-project.org/package=tseries*

Trueman, B., Wong, M. F. and Zhang, X.-J. (2001), 'Back to Basics: Forecasting the Revenues of Internet Firms', *Review of Accounting Studies* **6(2–3)**, 305–329.

Tsamardinos, I. and Aliferis, C. (2003), 'Towards Principled Feature Selection: Relevancy, Filters and Wrappers', *Proceedings of the Ninth International Workshop on Artifcial Intelligence and Statistics* .

Urbanek, S. (2014), *RJDBC: Provides Access to Databases Through the JDBC Interface*. R package version 0.2-5.
URL: *http://CRAN.R-project.org/package=RJDBC*

van der Putten, P. and van Someren, M. (2004), 'A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000', *Machine Learning* **57**(1-2), 177–195.

Veeramachaneni, K. (2016), 'Why You're not Getting Value from Your Data Science', Harvard Business Review. Retrieved from `https://hbr.org/2016/12/why-youre-not-getting-value-from-your-data-science`.

Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2011), 'Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques', *Expert Systems with Applications* **38**(3), 2354–2364.

Wang, Y., Cong, G., Song, G. and Xie, K. (2010), 'Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks', *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 1039–1048.

Weatherford, L. R. and Belobaba, P. P. (2002), 'Revenue Impacts of Fare Input and Demand Forecast Accuracy in Airline Yield Management', *Journal of the Operational Research Society* pp. 811–821.

Weatherford, L. R. and Kimes, S. E. (2003), 'A Comparison of Forecasting Methods for Hotel Revenue Management', *International Journal of Forecasting* **19**(3), 401–415.

Wei, C. and Chiu, I. (2002), 'Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach', *Expert Systems with Applications* **23**, 103–112.

Wickham, H. and Francois, R. (2016), *Dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.
URL: *http://CRAN.R-project.org/package=dplyr*

Wirth, R. and Hipp, J. (2000), CRISP-DM: Towards a Standard Process Model for Data Mining, *in* 'Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining', pp. 29–39.

Witten, I. H. and Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edn, Morgan Kaufmann, San Francisco.

Wolpert, D. H. and Macready, W. G. (1997), 'No Free Lunch Theorems for Optimization', *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Yates, R. (1995), 'A Framework for Uplink Power Control in Cellular Radio Systems', *IEEE JSAC* **13**(7), 3141–3147.

Ziegler, C. N. and Lausen, G. (2004), 'Spreading Activation Models for Trust Propagation', *In Proceedings of the IEEE International Conference on e–Technology, e–Commerce, and e–Service* pp. 83–97.

Samenvatting

Vanwege de digitale revolutie in de afgelopen decennia is er veel data beschikbaar. De overweldigende aanwezigheid van digitale apparaten in combinatie met platformen die het genereren en opslaan van data mogelijk maken, heeft geleid tot hoeveelheden data die in het verleden moeilijk voorstelbaar waren. Volgens IBM (2017) is 90% van de vandaag beschikbare data gegenereerd in de afgelopen twee jaar. Daarnaast worden dagelijks 2,5 triljoen bytes aan data gegenereerd en dit cijfer zal waarschijnlijk groeien, gezien de opkomst van nieuwe technologieën, apparaten en sensoren. Dit roept de vraag op welk deel van die data relevant of waardevol is. Het verkrijgen van waarde uit deze overvloed aan data is van belang voor zowel de industrie als de academische wereld.

Datamining wordt in het kort gedefinieerd als het proces van het ontdekken van patronen in data (Witten and Frank, 2005). De ontdekte patronen moeten zinvol zijn en tot een voordeel leiden, bijvoorbeeld een economisch voordeel. Grant (2003), geeft een meer gedetailleerde definitie van datamining als "een interdisciplinair veld met technieken zoals machine learning, patroonherkenning, statistieken, databases en data-visualisatie om de problemen aan te pakken met betrekking tot informatie-extractie uit grote databases."

Dit proefschrift past datamining toe in commerciële omgevingen in de telecommunicatie-industrie. Het onderzoek voor dit proefschrift is uitgevoerd bij T-Mobile Nederland B.V. en de methoden beschreven in sommige hoofdstukken zijn ook toegepast in dochterondernemingen van Deutsche Telekom in andere landen. We waren in de gelegenheid om aan echte commerciële datasets te werken en de resultaten van ons onderzoek in de praktijk toe te passen. In dit proefschrift beschrijven we enkele van de uitdagingen die data miners (of data scientists) tegenkomen bij het werken aan bedrijfsproblemen en we geven onze oplossingen hiervoor. De complexe datasets die we aan het analyseren waren bevatten in bepaalde gevallen tot honderden miljoenen records. In dit onderzoek gebruikten we datamining methoden op een innovatieve manier om resultaten te bereiken die ofwel een verbetering waren van hoe het bedrijf eerder deze problemen oploste ofwel belangrijke bedrijfsproblemen oplosten die tot nu toe niet eerder waren behandeld op zo'n gedetailleerde manier.

De inspiratie voor dit proefschrift is onze overtuiging dat het de verantwoorde-

lijkheid van de academische wereld is om niet alleen nieuwe manieren te vinden om problemen op te lossen, maar ook om het bedrijfsleven en de overheid te leren hoe ze deze ontwikkelingen kunnen gebruiken. Dit is een probleem in het geval van machine learning, waarbij de technologiereuzen van vandaag (bijvoorbeeld Facebook, Google of de NSA) op grote schaal gebruik maken van modern machine learning-onderzoek, terwijl de rest van de bedrijven achterblijft. We willen de toewijding van onderzoekers aan nieuwe en verbeterde algoritmen voor machine learning niet bekritiseren, maar hen juist aanmoedigen om een stap verder te gaan met de implementatie van deze methoden in praktische omstandigheden (verspreiding in gebruik van hun innovatie), waardoor het volledige potentieel van hun onderzoek wordt gerealiseerd. Om de hierboven genoemde redenen is dit proefschrift gericht op het toepassen van datamining in de praktijk in de telecommunicatie branche.

Churn, d.w.z. het verliezen van een klant aan de concurrentie, is een groot probleem in mobiele telecommunicatie en vele andere industrieën. Daarom wijden we drie hoofdstukken aan dit probleem.

In hoofdstuk 2 bespreken we de impact van de experimentele opzet op de voorspelling van prepaid churn in telecommunicatie. Prepaid klanten in mobiele telecommunicatie zijn niet aan een contract gebonden en kunnen daarom zonder toestemming van operator wijzigen (”churn”). Dit maakt het voorspellen van churn zowel uitdagend als financieel lonend. Het hoofdstuk presenteert een verkennend onderzoek naar prepaid churn-modellering door de experimentele opstelling te variëren op drie dimensies: data, doelvariabele definitie en steekproef.

In hoofdstuk 3 onderzoeken we de toegevoegde waarde van het combineren van reguliere tabulaire datamining met social netwerkmining, waarbij gebruik wordt gemaakt van de graaf gevormd uit communicatie tussen klanten. Hier vergeleken we de prestaties van klassieke (tabulaire) prepaid churn-modellen en klassieke sociaal netwerk modellen met twee hybride modellen. Ten eerste hebben we de dataset, die voor de tabelmodellen wordt gebruikt, verrijkt met variabelen uit de communicatiegraaf. Ten tweede creëerden we een propagatiemodel met behulp van de scores van de tabulaire churn-modellen als initiële energie van elk niet-churner-knooppunt (vergelijkbaar met boosting).

In de zeer concurrerende en geavanceerde telecommunicatiemarkt in Nederland is netwerkervaring cruciaal voor de operators. Dit is de reden waarom de hoofdstukken 4 en 5 focussen op twee verschillende manieren om het netwerk te verbeteren.

Hoofdstuk 4 beschrijft een andere toepassing van een churn-model. Dit hoofdstuk schetst een benadering die is ontwikkeld als onderdeel van een bedrijfsbreed initiatief voor churn-management. Onze aanpak van churn-preventie kan ook gezien worden als een brug tussen de disciplines marketing en mobiele netwerktechnologie, omdat we de technische oorzaken van churn identificeerden. De typische implementatiemethode voor een churn-model is een retentiecampagne waarbij klanten worden benaderd met een aanbod om hun contract voort te zetten. In dit geval was er geen campagne. Het model werd gebruikt om een reeks regels te genereren voor net-

werkoptimalisatie om de belangrijkste netwerk-gerelateerde oorzaken van churn te verwijderen en daarmee churn te voorkomen in plaats van te genezen.

In hoofdstuk 5 presenteerden we een zeer eenvoudige maar effectieve benadering van het gebruik van datamining in een commerciële omgeving. Hiervoor passen we de uitkomsten van datamining, een grote verzameling van modellen, in een simulatieraamwerk, dat niet dataminers maar domeinexperts in staat stelt niet alleen de toekomst te voorspellen, maar ook om simulaties te doen onder verschillende scenario's en condities. Na het eerste succes in T-Mobile Netherlands, waar de methode werd ontwikkeld, werd de aanpak ook gebruikt door operators van Deutsche Telekom in vier andere landen.

In hoofdstuk 6 breidden we de methode die is ontwikkeld in hoofdstuk 5 uit tot het gebied van financiën en het voorspellen van de service inkomsten. Inkomsten voorspellen in het algemeen is een van de belangrijkste financiële processen in elke bedrijfstak. Voor op diensten gebaseerde activiteiten, zoals telecommunicatie, zijn tijdige en nauwkeurige voorspellingen voor service-inkomsten essentieel, omdat ze belangrijke zakelijke beslissingen kunnen sturen, zoals wanneer en waar men kan ingrijpen om de zakelijke doelstellingen te bereiken. Door een simulatieplatform voor eindgebruikers te maken voor het voorspellen van de bedrijfsinkomsten, kan het bedrijf een beter idee krijgen van de manier waarop verschillende scenario's voor input parameters of geplande maatregelen in de praktijk kunnen uitwerken. De belangrijkste aspecten van de implementatie zijn vergelijkbaar met hoofdstuk 5: gebruik van hulpmiddelen die bekend zijn bij eindgebruikers en creatie van een platform voor simulatie van scenario's.

Vanuit een generalisatieperspectief is dit proefschrift op vele industrieën van toepassing. Het is duidelijk dat in vrijwel elke sector het verliezen van klanten aan de concurrentie (churn), het beheren van de belangrijkste resources en het voorspellen van de inkomsten erg belangrijke problemen zijn. Het vinden van kostenefficiënte middelen om deze problemen aan te pakken, is gunstig voor veel bedrijven of overheden. Het gebruik van eenvoudige algoritmen die gemakkelijk kunnen worden uitgelegd en het gebruik van implementatiemethoden die de voorkeur hebben van eindgebruikers, helpt bij de acceptatie. Dit is hoe we datamining zien worden toegepast en geaccepteerd op veel gebieden buiten de telecommunicatie, waardoor meer organisaties data- en klantgedreven worden. Vanuit een onderzoeksperspectief hopen we dat we hebben laten zien dat er veel andere interessante problemen zijn om op te lossen dan alleen het ontwikkelen van een beter voorspellend model. Hopelijk zullen we door het toepassen van datamining in telecommunicatie de academische belangstelling vergroten om betere en efficiëntere manieren te vinden voor een snellere verspreiding van machine learning onderzoek.

Curriculum Vitae

Dejan Radosavljevik was born in 1975 in Skopje, Macedonia. After graduating from a BSc program in Computer Science at the University of Ss. Cyril and Methodius in Skopje in 2001, he has worked as a software developer for several Macedonian companies. In 2009 he completed a Master's degree in ICT in Business with cum laude distinction at Leiden University with a thesis on Prepaid Churn Modeling Using Customer Experience Management Key Performance Indicators. Since then he has worked in multiple positions related to artificial intelligence, data mining and data science at T-Mobile Netherlands B.V., in parallel to working on this PhD research at Leiden University. He currently holds the position of Lead Data Scientist within T-Mobile Netherlands.

List of Figures

List of Tables