Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/57982 holds various files of this Leiden University dissertation.

Author: Radosavljevik, D.
Title: Applying data mining in telecommunications
Issue Date: 2017-12-11

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

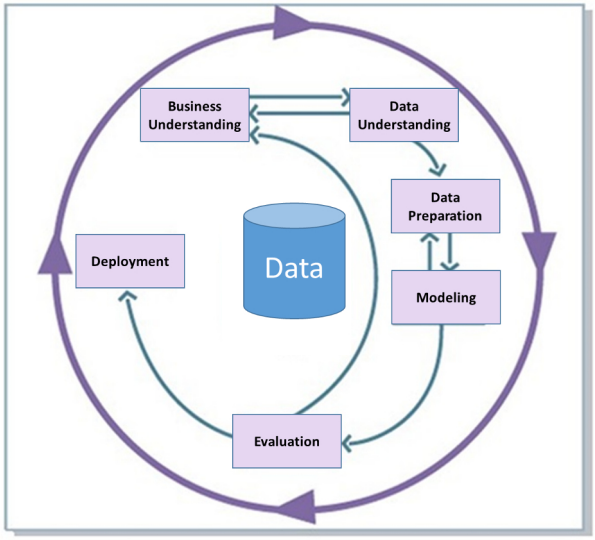

Figure 1.1: CRISP-DM Process Model for Data Mining

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.