# Introduction

The telecommunication market is a highly competitive market. Customers are able to choose between a variety of suppliers and it is relatively easy to switch between them. This results in a very high churn rate and represents a serious problem for the suppliers as the costs for attracting new customers exceed the costs of customer retention (e.g., Lu, 2002). An important and necessary task for the telecommunication companies is to predict the probability of customers to cancel the contract in the near future. This information can be used then to start a loyalty initiative such as offering a discount or a new mobile phone to the customer. To gain valuable knowledge out of the data telecommunication providers have about their customers, data mining techniques can be used.

This working task deals with such a customer churn problem. The provided real-world data of 100,000 customers of an anonymous telecommunication provider are described by 173 attributes, such as customer characteristic and usage behaviour. The observed customers were with the company for at least six month and they were sampled during July, September, November of 2001 and January of 2002. At the time a customer was observed the input variables were calculated based on the previous four months. If a customer churned within a period of 31-60 days after the observation time, the data provides the information of “churn=1”, otherwise “churn=0”. The data set is divided into two parts of equal size (training and test set). One set represents the training set including the information if a customer churned as explained above. This set will be used to build our model. Our goal is then to predict the churn probabilities for each customer in the test set. The higher the probability the higher the risk that a customer will churn within the period of 31-60 days after the observation time. An accurate prediction gives the company sufficient time to react to the information and retain the customer.

In doing so we proceed in accordance with the KDD process (knowledge discovery in databases) (e.g., Fayyad, et al., 1996). The first step of selection is already done as our target data is selected (see above). In chapter 2 we will some exploratory data analysis which is part of the preprocessing. The following steps of the KDD process will be divided into two chapters (chapter 3 and 4). Our approach is to run two iterations during the model building process. Within the first iteration we will focus on rather simple methods (for example for data cleaning and reduction). We will start with data cleaning such as missing value and outlier handling. Next within transformation we will focus on data reduction (through principle component analysis and feature selection) and encoding. Within the next step of data mining we will apply data mining algorithms to search for patterns in our data. In our case the data mining model is classification and four methods will be applied: Logistic regression, Naïve Bayes Classifier, Random Forest and Artificial Neural Networks. To increase the predictive accurary of the individual models we will also apply Ensemble Methods (do we? If yes, which ones? Random Forest is already an Ensemble Algorithm). At the end of chapter 3 we will interpret the results and evaluate which approaches worked well and which should be improved. In the second iteration we will handle the needs for improvement identified in the first iteration and focus on rather complex methods if required. The structure of the second iteration equals that of the first iteration based on the KDD process. During this procedure we will always keep in mind that the KDD process is not a strict sequence, but rather an iterative process.

* What is the problem?
* Motivation: why the chosen topic is interesting and relevant?

# Feature Selection

The data set consists of 172 input variables and probably not all of them are relevant for the prediction. One part of vertical data reduction is the feature or variable selection where we want to find a subset of relevant variables. Out of a variety of feature selection approaches we decide within the first iteration to use a simple Filter Approach. The approach is to pre-screen the variables prior to model building and only use those variables as predictors that pass a certain statistical criterion. In our case we use the rfe function out of the caret package (http://topepo.github.io/caret/filters.html). Through an underlying ANOVA test it is tested for each variable if the mean is different between the two classes “churn=1” and “churn=0” (null hypothesis). If the null hypothesis can be rejected the variable is regarded as relevant and therefore should be considered in the prediction model. There are several other approaches for feature selection that are more effective, for instance the Wrapper Approach or using models with built-in feature selection. The latter have the advantage that feature selection is included in the objective function that is optimized which is not given when feature selection is separated from all the other steps (like in the Filter Approach). Furthermore a disadvantage of the Filter Approach is that it is used in our case in a univariate manner, so that redundancy and interactions of variables are not taken into account. Nevertheless considering that we have high-dimension data we decide in the first iteration for this approach because of its much lower computational costs. As a result of the filter we get 68 variables that can be viewed as relevant thus will be used as input variables for the data mining in the following.

Now it makes sense to check how many missing values the selected variables had before imputation. It would not make much sense to include variables that have a too high missing value rate because the information content would be low. But within our selected variable subset only two variables have more than 3% missing values, namely the variables “lor” (length of residence) and “adults” (number of adults in household). From a business perspective these two variables do not seem to be relevant for the churn problem and thus are excluded as well. Hence we have now 66 predictors for further analysis.

# References (Introduction)

Lu, J. (2002). Predicting customer churn in the telecommunications industry––An application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, 114-27.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.