Business Analytics & Data Science   
Winter 2015/2016  
Special Working Task (SWT)

# Scope & Objective

An important learning objective of the module is to familiarize students with standard procedures and tools used in the field of business analytics and data science. To achieve this, the SWT exposes students to a challenging real-world business problem. To solve the problem, students are required to apply a number of different skills and to employ several diverse analytical techniques. To reflect this complexity and the importance of the task for achieving key learning objectives, the SWT has to be completed successfully to pass the module. More specifically, passing the SWT by achieving a minimal grade of 4.0 is a mandatory prerequisite to participate in the written exam and thus to successfully complete the module.

# Task Description and Background

The business problem stems from the marketing domain and is related to managing customer churn (e.g., Reichheld & Sasser, 1990). Customer churn is a severe problem in many service industries including telecommunications, banking and insurance, energy, and many more. Taking a company perspective, acquiring new customers (e.g., to compensate for leaving customers) is substantially more costly than retaining existing customers (e.g., Colgate & Danaher, 2000). Moreover, long-term customers generate higher profits, are less sensitive to competitive actions, and may act as promoters through positive word of mouth (e.g., Ganesh, et al., 2000; Reichheld, 1996; Zeithaml, et al., 1996). Although relationship management instruments such as loyalty programs often reduce churn, customer attrition remains a major threat to the financial health of many companies (e.g., Schweidel, et al., 2008; Thomas, et al., 2004). For example, T-Mobile USA lost half a million of its most lucrative customers in the first quarter of 2012 (Bensinger & Tibken, 2012). It is known that contract churn rates for many types of communication services are in the three-percent-per month range (Kim, 2010). This means that a provider needs to refresh nearly 100% of its customer base about every three years.

The purpose of a churn model is to predict the likelihood that some customer ends her/his business relationship in the near future. Equipped with such information, the company can initiate a marketing campaign to proactively prevent customer attrition. In the mobile telecommunication industry, for example, the company could contact customers with high churn risks and offer them a new phone if they decide to renew their contract. In the SWT, students are given a real-world data set from a large mobile telecommunication company in the US, and are asked to develop a prediction model to estimate customers’ churn risk as accurately as possible.

# Data

The customer data set includes 100,000 observations (customers) each of which is described by 171 attributes. Roughly speaking, the attributes capture demographic, sociographic and micro-geographic characteristics of the customers, information about their interactions with the company, and their service usage behavior in particular (e.g., how many calls, average length per call, etc.). A description of the attributes is available in the file BADS\_SWT\_DataDescription.xls.

**Disclaimer:** Note that the description of the data was provided by the anonymous telco company. In that sense, it represents a real-world data description. This implies that the information provided in the data description is not necessarily correct. For example, the descriptive statistics related to individual variables, which are given in BADS\_SWT\_DataDescription.xls might suffer errors. It might be a good idea to check and maybe re-compute descriptives using your own codes.

The data set includes two special attributes *Customer\_ID*, and *churn*. The former is a unique identifier of the customer. The latter is a binary target (dependent) variable. The two states of this variable capture whether a customer did churn (*churn*=1) or not (*churn*=0), after showing some ‘behavior’, which is represented by the remaining variables. For example, when gathering the data in the first place, the company did first collect the values of the customer attributes, say in one month t, and then observed whether the customer churned in a subsequent month, e.g., t+1.

# Task Specification

Let **x**i denote a vector that contains the values of the attributes of customer i. The SWT is then formally given as estimating the conditional probability of yi=1 given **x**i; p(yi=1|**x**i), where the scalar yi represents the value of the target variable of customer i; that is, whether customer i churned (yi=1) or not (yi=0).

The data set has been randomly split in two parts of equal size, a training and a test set. The information whether a customer churned is available in the training set only. The task of the assignment is to build the churn model using the data of the training set. Subsequently, the model is to be applied to the test set to predict churn probabilities for the customers therein. Next, the results are to be submitted in a CSV file with the following format:

Customer\_ID; EstimatedChurnProbability

1, 0.45

2, 0.32

3, 0.87

…

Hence, the submission file includes two columns (using the comma as separator) that give the value of the customer ID and the churn probability as estimated by the previously developed churn model. **Note that it is key to apply the churn model to the test set. Only predictions for test set customers will count toward your prediction score.**

# Performance Evaluation

Churn models and more generally scoring models for binary events can be assessed in a number of ways. The SWT focusses on predictive accuracy. The predicted churn probabilities (of test set customers!) will be compared to the actual churn outcomes, which are known to the lecturer only. The specific indicator used to summarize this comparison and calculate predictive accuracy is the lift-measure. This measure is commonly used in direct marketing to assess targeting models (e.g., Ling & Li, 1998). The lift measure grounds on a list of customers ordered according to their model-estimated scores (from highest to lowest risk in churn prediction). For some decile *d* of the ordered customer list, the lift measure *Ld* is defined as:

|  |  |
| --- | --- |
|  | ( 1 ) |

where and denote the fraction of actual churners among all customers and those ranked in the top-*d* decile, respectively. If a retention program were targeting customers at random, the fraction of actual churners reached with the campaign would equal  (that fraction that can be expected by change). The lift measure quantifies how much a churn model improves over such a random targeting.

An important feature of the lift measure is that it captures, under some assumptions, the profit of a marketing campaign (e.g., the profit of a retention program; Neslin, et al., 2006). Although the lift measure can be defined for any decile *d* of the ranked customer list, it is common practice to set *d* to some small value. In the assignment this value is 0.1, which means that the accuracy of students’ churn predictions will be assessed on the 10% of test set customers with highest estimated churn probabilities.

# Written Report

In addition to submitting test set probability predictions, the SWT also requires students to prepare a written report, which documents the individual steps they took in the modeling process. The report should clarify which modeling approaches have been employed. It should also **elaborate** and **motivate** relevant decisions taken in the modeling process. Even approaches that have eventually been discarded and are not reflected in the final churn model might deserve a mention. Basically, the written report is meant to be a comprehensive documentation of everything that has been done to solve the SWT.

# Expectation and Grading

Predictive accuracy is an important feature of any forecasting model; this is no different for churn models. However, with respect to the SWT, accuracy is not the only – in fact not even the main – factor that determines students’ grade. Predicting future customer behavior involves randomness and uncertainty. A carefully and adeptly devised churn model may, in some unfortunate circumstances, predict less accurately than a simplistic model that has been created without care and skill. Therefore, the grading of the SWT depends mainly on the way in which students approach the modeling task, which methods they use and why, how much effort they put into model building, and their overall level of competency. This is why the written report is very important and the main determinant of students’ grade.

The core task of the assignment is to build a prediction model for a binary target variable. It is well-possible to “solve” this task with maybe ten lines of R code. Obviously, the resulting solution would not be very good. On the other hand, there are an indefinite number of approaches and strategies one could explore in the course of trying to build the very best prediction model possible. This way, one can easily invest hundreds of working hours into the task. Essentially, students decide for themselves how much effort to devote to the modeling task and the SWT, respectively. However, it is sensible to use the KDD process (discussed in the lecture) as a guideline. Every step (and maybe sub-step) of the KDD process should be considered in the SWT. For example, the churn data set contains many missing values. Thus, it is expected that students think about an approach to remedy this issue. Similarly, the number of variables is fairly large (e.g., about 170). Hence, a solution that misses out on variable selection would probably not achieve a good grade.

# Fair Use Agreement

The data set used in the assignment has been used in a public churn modeling competition organized by the center for customer relationship management at Duke University (Neslin, et al., 2006). There may be a way to acquire the full data set, which would then facilitate a re-identification of the customers in the test set and their actual churn values in particular. Such approach is strictly forbidden, and also easy to detect. It is expected that students **DO NOT** engage in, participate, or support such fraudulent approach in any way. Any violations of this rule will have serious consequences and lead to an exclusion from the SWT. Accordingly, students would not be allowed to participate in the written exam and could not complete the module in this semester.

# Organization and Deadlines

Students work on the SWT in groups. The recommended group size is five students. However, smaller groups are also possible. In exceptional cases, groups with more than five participants might be feasible as well; students are asked to contact the course organizer in such exceptional cases to explain their motivation for larger groups.

To register a group, please send an email to Anna-Lena Bujarek ([bujarek@wiwi.hu-berlin.de](mailto:bujarek@wiwi.hu-berlin.de)). Your email should give, for each group member, first and last name as well as student ID (Matrikelnummer).

Deadlines:

* Submission of final predictions 29th of Jan. 2016
* Submission of written report 05th of Feb. 2016

# References

Bensinger, G., & Tibken, S. (2012). T-Mobile struggles to stem customer losses. In *The Wall Street Journal*.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of Marketing Science, 28*, 375-387.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers *Journal of Marketing, 64*, 65-87.

Kim, G. (2010). AT&T Churn Rate Offers Lesson. Last accessed: November, 2014 [<http://technews.tmcnet.com/voice-quality/topics/phone-service/articles/93062-att-churn-rate-offers-lesson.htm>]

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. In R. Agrawal, P. E. Stolorz & G. Piatetsky-Shapiro (Eds.), *Proc. of the 4th Intern. Conf. on Knowledge Discovery and Data Mining* (pp. 73-79). New York City, NY, USA: AAAI Press.

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

Reichheld, F. F. (1996). Learning from customer defections. *Havard Business Review, 74*, 56-69.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to service. *Havard Business Review, 68*, 105-111.

Schweidel, D. A., Fader, P. S., & Bradlow, E. T. (2008). Understanding service retention within and across cohorts using limited information. *Journal of Marketing, 72*, 82-94.

Thomas, J. S., Blattberg, R., & Fox, E. (2004). Recapturing lost customers. *Journal of Marketing Research, 41*, 31-56.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing, 60*, 31-46.