HARVARD | BUSINESS | SCHOOL

Managing Churn to Maximize Profits

Aurélie Lemmens
Sunil Gupta

Working Paper

14-020

September 4, 2013

Managing Churn to Maximize Profits

Aurélie Lemmens

Tilburg School of Economics and Management

Sunil Gupta

Harvard Business School

This Version: September 2013

Aurélie Lemmens (corresponding author) is Associate Professor of Marketing at Tilburg University. Address: P.O. Box 90153, 5000 LE Tilburg,The Netherlands, E-mail: a.lemmens@uvt.nl, Phone: +31 13-466 2057. Sunil Gupta is the Edward W. Carter Professor of Business Administration at Harvard Business School. The first author received financial support from the N.W.O. in the framework of a VENI grant. Part of the project has been made while she was visiting Harvard Business School. The manuscript benefited from the many comments received from the marketing department members at Tuck Business School, Dartmouth College. Finally, we are deeply indebted to Prof. Dr. Christophe Croux (K.U. Leuven) for his valuable contribution on an earlier version of this project.

Managing Churn to Maximize Profits

Abstract

Customer defection or churn is a widespread phenomenon that threatens firms across a variety of industries with dramatic financial consequences. To tackle this problem, companies are developing sophisticated churn management strategies. These strategies typically involve two steps – ranking customers based on their estimated propensity to churn, and then offering retention incentives to a subset of customers at the top of the churn ranking. The implicit assumption is that this process would maximize firm's profits by targeting customers who are most likely to churn.

However, current marketing research and practice aims at maximizing the correct classification of churners and non-churners. Profit from targeting a customer depends on not only a customer's propensity to churn, but also on her spend or value, her probability of responding to retention offers, as well as the cost of these offers. Overall profit of the firm also depends on the number of customers the firm decides to target for its retention campaign.

We propose a predictive model that accounts for all these elements. Our optimization algorithm uses stochastic gradient boosting, a state-of-the-art numerical algorithm based on stage-wise gradient descent. It also determines the optimal number of customers to target. The resulting optimal customer ranking and target size selection leads to, on average, a 115% improvement in profit compared to current methods. Remarkably, the improvement in profit comes along with more prediction errors in terms of which customers will churn. However, the new loss function leads to better predictions where it matters the most for the company's profits. For a company like Verizon Wireless, this translates into a profit increase of at least $28 million from a single retention campaign, without any additional implementation cost.

Keywords: Churn Management, Defection Prediction, Loss Function, Stochastic Gradient Boosting

INTRODUCTION

Customer churn or defection is a widespread phenomenon across a variety of industries. A recent report estimated 20% annual churn rates for credit cards in the US, and 20%-38% annual churn rate for mobile phone carriers in Europe (Bobbier 2013). As customer acquisition costs continue to rise, managing customer churn has become critically important for the profitability of companies. A report by the consulting company McKinsey estimated that reducing churn could increase earnings of a typical US wireless carrier by as much as 9.9% (Braff, Passmore, and Simpson 2003).

Not surprisingly top executives both in the US and UK indicate customer retention as their number one marketing priority and report higher retention budgets (Forbes 2011). Companies are beginning to use sophisticated churn management or retention campaigns. Such retention programs consist of targeting the customers identified as potential churners with special retention incentives (Bolton, Kannan, and Bramlett 2000; Ganesh, Arnold, and Reynolds 2000; Shaffer and Zhang 2002). These incentives take multiple forms, such as special offers, discounts, personalized (e-) mailing etc., with the common objective of increasing the targeted customers' behavioral loyalty.

Thus far, the most common approach among marketing academics and practitioners to design targeted retention programs has consisted of the following two steps. First, a classification method estimates the propensity of each customer to defect, such that a ranking of customers can be established based on their estimated churn probability (Neslin 2002). Subsequently, a fraction of customers, starting from the top of the ranking, is selected as target group for the retention incentive. All customers in the target group are offered a retention action. While very straightforward and easy to implement, this approach to design retention campaigns suffers from an obvious but nonetheless extremely critical shortcoming: neither of

the two steps are constructed to maximize the financial profits of the retention campaign and therefore lead to poor retention investments.

Our goal in this paper is to rethink the approach to design targeted retention programs in a way that maximizes the financial profits of the retention investments made by companies. To do so, we develop a binary classification method that uses a gain/loss matrix, which incorporates the gain of targeting and retaining the most valuable churners and the cost of incentives to the targeted customers. Using this gain/loss matrix, we construct a customer-heterogeneous profit-based loss function. We then use the stochastic gradient boosting optimization algorithm to minimize the losses. This greedy algorithm is based on stage-wise steepest or gradient descent. Finally, we determine the optimal target size for maximizing the financial profits, rather than selecting an arbitrary number of customers. The optimal size of the retention program takes into account the compromise that the firm has to make between increasing the target size to reduce the loss from defection, and decreasing the target size to reduce the cost of retention actions.

Our results show that our approach leads to far more profitable retention campaigns than the traditional churn modeling approaches. In addition, the additional profits come at no cost for companies. The implementation of the retention campaign is unchanged, only the composition and size of the target group changes compared to traditional approaches.

The remainder of the paper is organized as follows. We start by briefly outlining the current approaches and the two steps used in these approaches to model churn. Next, as a key building block for our approach we define the profit of a retention campaign. We then build a profit-based loss function based on the profit formula proposed above. In the next two sections, we explain the stochastic gradient boosting algorithm, show how we integrate the new loss function in the algorithm and describe the process for optimizing the target size. This is followed by the description of the data we use, and the results of our approach compared to

existing methods. We conclude by discussing the limitations and potential extensions of our approach.

A BRIEF OVERVIEW OF THE CHURN MODELING APPROACHES

As indicated earlier, traditional churn management approaches consist of two steps: a model to predict the propensity of churn for each customer, followed by selecting the top few percent of likely churners who are offered the retention incentives.

In carrying out the first step, various prediction methods are used as highlighted by the churn modeling tournament organized by the Teradata Center at Duke University, where 44 respondents participated, divided roughly 50-50 between academics and practitioners (Neslin et al. 2006, also see Blattberg, Kim, and Neslin 2008, pp. 256-259 for an overview). In presence of cross-sectional data, academics have used logistic regression (Lemon, White, and Winer 2002), discriminant analysis, finite mixture (Andrews, Ainslie, and Currim 2002), hierarchical Bayes (Yang and Allenby 2003), decision trees or CART (Risselada, Verhoef, and Bijmolt 2010), neural nets (Thieme, Song, and Calantone 2000; West, Brockett, and Golden 1997), random forests (Lariviere and Van den Poel 2005), and bagging and stochastic gradient boosting (Lemmens and Croux 2006). Other methods, when longitudinal data are available, include hazard models (Bhattacharya 1998; Bolton 1998; Schweidel, Fader, and Bradlow 2008) as well as hidden Markov models (Ascarza and Hardie 2013).

While these methods vary in multiple dimensions (parametric vs. nonparametric, homogenous vs. heterogeneous parameters, time-invariant vs. time-varying effects, etc.), they all implicitly aim – via their loss function – at minimizing the percentage of misclassified customers (*misclassification rate*), i.e. the percentage churners classified as non-churners and non-churners classified as churners, or equivalently, at maximizing the number of churners predicted churners as well as the number of non-churners predicted as non-churners. They

ignore that each customer is not equally important to the firm. In particular, the benefits (costs) associated with an (in)correct evaluation of a customer's churn propensity substantially varies across a firm's customer base and depends on the potential profit leveraged by the decision of targeting or not a given customer. Therefore, minimizing the overall misclassification rate of a prediction model is not equivalent to maximizing the profits of a retention campaign. Actually, a method can potentially do worse in terms of classification error and still yield higher profits, as we will demonstrate in this paper.

In many academic fields, including marketing, empirical researchers have long ignored the perils of using a loss function that is not aligned with the managerial objectives of the company. Most of the time, a different loss function is used for in-sample estimation and out-of-sample evaluation. Such mismatch leads to suboptimal model selection and predictions (Engle 1993; Granger 1993). A few notable exceptions exist. Blattberg and George (1992) optimize manufacturers' prices by estimating price sensitivity using a profit-based loss function. Bult (1993), as well as Bult and Wittink (1996) propose a loss function that account for the asymmetry in the cost of targeting mailing to the wrong customers. Glady, Baesens, and Croux (2009) model the probability of a net increase in customer lifetime value following a marketing action using an asymmetric misclassification cost. Bayesian statistics, and Bayesian decision analysis in particular, is perhaps one of the only modeling approach for which the importance of selecting a relevant loss function is more salient (Rossi and Allenby 2003). The loss function quantifies the loss to the decision maker of taking a given action given the state of nature (Gilbride, Lenk, and Brazell 2008). In the next section we develop a profit-based loss function for our approach.

In the second step of target size selection, traditional approaches also ignore the companies' goal of maximizing profits. Current practice chooses an arbitrary target size, e.g. the *top-decile* of the ranking (Lemmens and Croux 2006). A somewhat better approach, recently

suggested by Verbeke et al. (2012), consists of computing the expected profit of targeting an average customer, given the average value of a customer to the firm and an average positive response rate to a retention action, and therefore evaluating the target size that would lead to the highest profits. However, this approach ignores the heterogeneity in individual targeting opportunity of customers. Later, we show how we construct an optimal target size that maximizes profits and does significantly better than the arbitrary target size selection.

DEFINING THE PROFIT OF TARGETED RETENTION ACTIONS

Using the conceptual framework proposed by Neslin et al. (2006), we postulate that the profit of a retention action is heterogeneous across the customer base. At the customer level, the profit of targeting a given individual depends on four elements: (i) the customer's future churn behavior in absence of a retention action, (ii) the value of the customer to the firm, (iii) the probability that the customer, if targeted, will respond positively to the retention action and therefore not defect, and (iv) the cost of the retention action. At the company level, the profit of the overall retention campaign also depends on the target size of the retention campaign.

First, the financial profit of a retention action depends on the churn propensity of each customer in the absence of a retention action. This differs across the customer base, and hence the opportunity of targeting a given customer does as well. Targeting a future churner is financially more profitable than targeting a customer who has no intention to leave. The action cost that the company would incur in the latter case is unnecessary (Neslin 2002).

Second, the financial profit of a retention action depends on the value that each customer generates for the firm. In other words, the targeting opportunity of a customer does not only depend on her churn propensity, but also on her value. Not all customers spend the same

money with the firm and, as a consequence, losing a high-value customer is financially more damaging to the firm than losing a low-value customer.

Third, the financial profit of a retention campaign depends on the customer response to the retention action. Not all targeted customers will successfully be retained, even after being correctly identified by the firm as future churners and targeted with a retention incentive. Some will defect anyway; hence their targeting opportunity is lower than if they would respond positively to a retention action.

Fourth, the financial profit of a retention campaign depends on the campaign cost. Offering retention incentives to customers is not costless. The costs incurred force companies to make a trade-off in terms of the number of customers to target. On the one hand, as the firm targets more customers, it reaches a larger set of churners. However, targeting a very large group of customers to ensure that many churners are reached could be very expensive and suboptimal.

Given a customer base of *N* customers, the profit Π of the overall retention campaign across all customers i being targeted can be written as

$$\Pi = \sum_{i \in \text{target}}^{N} \pi_i, \tag{1}$$

with π_i the profit of targeting a given customer *i* depending on whether customer *i* is a churner or not. Let $y_i = +1$ if customer *i* would churn in the next period if no action is taken, and $y_i = -1$ if customer *i* would not churn. We can write the individual targeting profit of customer *i* as

$$\pi_{y_i=+1} = \gamma_i(V_i - \delta) \text{ and } \pi_{y_i=-1} = -\varphi_i \delta. \tag{2}$$

where, γ_i is the probability that a targeted customer who intends to churn would accept the retention action (*positive response* probability) and hence stay with the firm. Likewise, φ_i is

the probability that a targeted customer who does not intend to churn would accept the retention action. In practice, γ_i and φ_i are likely to be different from each other (Neslin et al. 2006).

In equation (2), V_i represents the lifetime value of customer i (Gupta, Lehmann, and Stuart 2004). We assume, for simplification, that customers who accept the retention offer reach a defection probability equal to the population churn rate after accepting the offer.[1] Conceptually, we argue that a retention offer that a customer accepts effectively reduces her intrinsic motivation to churn, by (temporarily) overcoming her reasons to leave the firm.[2] We believe that this assumption is especially realistic in cases where the retention actions are associated with a contractual obligation of the customer to extend her contract with the company (which is common practice in the telecommunication industry). For instance, customers may receive an email in which they can "Redeem your free 100MB data pack" with the condition (sometimes hidden in a footnote) that their contract will be automatically extended for another 2 years.

In equation (2), δ represents the cost of the retention action targeted at a customer. For simplification purpose, we assume that the cost of the retention action is homogenous across customers and we assume that it includes the cost of contacting customers.[3]

In non-mathematical terms, the profit of targeting a customer who would churn in absence of a retention action equals a fraction of the net value of the customer to the firm (i.e. revenues minus costs), where the fraction is the probability of the customer responding positively to the action. As to the non-churners, the profit of targeting a customer who is not intending to churn is negative and equals to a fraction of the cost of the action, where the frac-

[1] Future research could extend our approach by making this new churn rate consumer-specific, which makes the model and its estimation significantly more complex.

[2] A similar argument is used in the Cell2Cell case, following upon the churn modeling tournament organized by the Teradata Center at Duke University (Neslin 2002). More details on the operationalization of the customer value variable are provided in the results section.

[3] The cost of contacting the customer is typically negligible compared to the cost of the retention offer and the average customer value (see Neslin et al. 2006).

tion is the propensity of this non-churner to accept the retention action. We assume that part of the non-churners will not take the offer even if they have no intention to leave the company. For instance, some customers value their freedom to be able to leave at any time more than the offer made to them.

The customer heterogeneity in the targeting profit in equation (1) and (2) implies that it is more crucial to make accurate targeting decisions (i.e. accurately predict the churn propensity) towards some customers than towards others. In particular, the higher the value of a customer with a high churn probability but a high response probability to a retention action, the larger the benefits of targeting her or the loss of not targeting her. In other words, predictions that underestimate the churn propensity of high-value and highly-responsive future churners should be favored less than predictions that underestimate the churn propensity of low-value and non-responsive future churners. Likewise, predictions that overestimate the churn propensity of non-churners, whatever their value to the firm, should also be avoided, especially when the action cost is large and the response rate of these non-churners is high.

As we explain in the next section, existing approaches to model customer defection ignore customer heterogeneity in the opportunity cost of mis-predicting churn. In mathematical terms, all customers are given the same penalty (or loss) for mis-prediction and the method intends to minimize the total prediction error over the customer base. In other words, it tries to correctly predict all customers' churn propensity. However, given that a prediction method can by nature not perfectly predict all customers' churn propensity, this consists of a big missed opportunity for marketing practitioners, who could substantially increase the profits of their retention campaigns if they focus their prediction efforts towards the customers that are likely to generate the highest profits. In the next section, we develop a new loss function based on the profit of retention actions.

DEVELOPING A PROFIT-BASED LOSS FUNCTION

Let $(y_1, X_1), \ldots, (y_i, X_i), \ldots, (y_N, X_N)$ be a (calibration) sample of known values of y, the churn binary outcome, and X a set of customer characteristics for N customers. Let $y_i = 1$ if customer i is a churner and $y_i = -1$ if customer i is a non-churner. Let F be the function that maps X to y. This function represents the score attached to each customer, which characterizes, in our application, the *targeting opportunity score* of a customer. The higher the score $F(X_i)$, the greater the opportunity to target individual i. In general, the score $F(X_i)$ can take any value, either positive or negative. As we describe below, the score has traditionally been modeled as the churn or defection propensity of a customer, hence ignoring the fact that the other elements (customer value, response probability, action cost and target size) influence the opportunity to target this individual. In this section, we will explain how to incorporate these elements such that the score would now represent a more realistic view of the targeting opportunity of each customer.

For all customers $i = 1, \ldots, N$, the goal is to obtain a function $F^*(X_i)$ that minimizes the expected value of a specific loss function $\Psi(y_i, F(X_i))$ over the joint distribution of all(y_i, X_i), $i = 1, \ldots, N$,

$$F^*(X_i) = arg\ min_{F(X_i)} E_{y,X}\left[\Psi\left(y_i, F(X_i)\right)\right], \tag{3}$$

with $F^*(X_i)$ representing the value of $F(X_i)$ where the expected loss is minimized. For notation simplification, we further denote $F_i = F(X_i)$ the predicted score for customer i in the remainder of the paper.

Traditional Approach: Misclassification Loss Function

In a classification context, the most commonly-used loss functions penalize *classification errors* in predictions. For such misclassification loss functions, the customer is predicted to be a churner when $F_i \geq 0$, while the customer is predicted to be a non-churner when $F_i < 0$. Two types of classification error can be made: (a) cases where a customer is predicted as a churner while she is a non-churner, and (b) cases where a customer is predicted as a non-churner when she is a churner. As a result, classification errors are all cases where the product $y_i F_i$, subsequently called the *margin*, is negative, either because $y_i = +1$ and $F_i < 0$, or because $y_i = -1$ and $F_i > 0$. In contrast, well-classified customers have a positive margin $y_i F_i$, either $y_i = +1$ and $F_i > 0$; or $y_i = -1$ and $F_i < 0$. As such, the margin can be used to characterize whether predictions are correct or incorrect.

Using the margin, the loss function can be represented by a 2 x 2 matrix, depending on whether the customer is a churner or not and whether she is predicted as churner or not. The diagonal elements equal 0 (i.e. correct predictions) and the non-diagonal elements are different from 0 (i.e. incorrect predictions). In table 1, we show the losses for the zero-one loss function, where all misclassifications are charged a single unit (Hastie, Tibshirani, and Friedman 2009, p. 20). This is a common loss function used, for instance, in the case of the Bayes optimal classifier, 1-nearest neighbor classification and some versions of the classification trees.

Table 1: Zero-one misclassification loss matrix

Loss Matrix	Predicted churner $\boldsymbol{F_i \geq 0}$	Predicted non-churner $\boldsymbol{F_i < 0}$
Actual churner $\boldsymbol{y_i = +1}$	Well-classified customer $\Psi = 0$	Misclassified customer $\Psi = 1$
Actual non-churner $\boldsymbol{y_i = -1}$	Misclassified customer $\Psi = 1$	Well-classified customer $\Psi = 0$

Mathematically, one can write such a loss function $\Psi_{\text{one-zero}}(y_i, F_i)$ using the margin $y_i F_i$

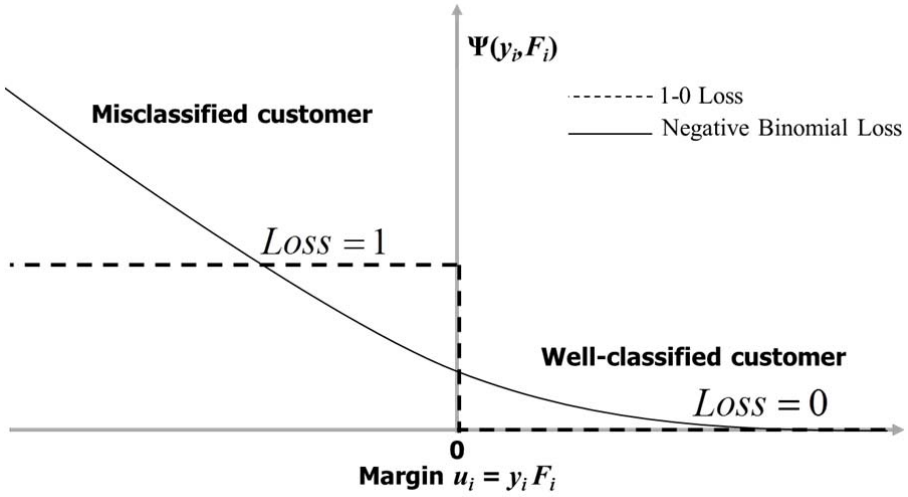
$$\Psi_{\text{one-zero}}(y_i, F_i) = I(y_i F_i < 0) \tag{4}$$

As drawbacks, the one-zero loss function is non-differentiable and it does not distinguish between extreme misclassification (e.g. $y_i F_i = -0.9$) and mild misclassification (e.g. $y_i F_i = -0.001$). A more sophisticate loss function to address these problems is the negative binomial log-likelihood or deviance (also known as cross-entropy, Hastie, Tibshirani, and Friedman 2009, p. 346), which is a monotone continuous approximation of the one-zero loss function:

$$\Psi_{\text{negative-binomial}}(y_i, F_i) = log(1 + e^{-2y_i F_i}). \tag{5}$$

With this continuous misclassification loss function, the more negative the margin, the worse the classification accuracy and the higher the loss. As represented in Figure 1, correctly classified cases receive a very small loss, while misclassified cases receive a positive loss that exponentially increases with the value of the margin. The higher the margin, the worse the classification accuracy and the higher the associated loss becomes. In addition, the negative binomial log-likelihood is a differentiable and continuous function. The negative binomial log-likelihood is a popular loss function for classification methods, as for instance in the original gradient boosting algorithm (Friedman 2001).

Figure 1: Misclassification loss functions: one-zero and negative binomial log-likelihood



We claim that, while statistically relevant, this misclassification loss function is not appropriate in the current churn prediction context as it does not take into account the heterogeneity in the targeting profit of the customers. The latter determines the loss related to different prediction errors and should therefore be accounted for in order to maximize the profit of the retention campaign.

New Approach: Profit-Based Loss Function

Based on the previous sections, we move away from a score representing the churn of defection propensity of customer *i* to a score that now represents the targeting opportunity score of customer *i*. This score captures the expected profit that the firm would make by targeting this individual. As described in equations (1)-(2), this targeting profit depends not only on whether the customer is a churner or not, but also on a number of other customer-specific parameters that influence her targeting profitability. Therefore, the margin should become $u_i = sign(\pi_i)F_i$ replacing the classical margin definition $u_i = y_i F_i$. There could indeed be cases where a customer is a churner ($y_i = 1$) but is not associated with a positive targeting profit (i.e. her net value ($V_i - \delta$) is negative). The margin is positive (negative) when a cus-

tomer i for whom the targeting profit is positive (i.e. $\pi_i \geq 0$) receives a score F_i greater (smaller) than 0; or when a customer i for which the targeting profit is negative (i.e. $\pi_i < 0$) receives a score F_i smaller (greater) than 0.

In Table 2, we construct a new 2 x 2 loss matrix based on this new margin. The cases where the margin is positive correspond to optimal targeting decisions and therefore receive a zero loss.[4] They are represented by the diagonal elements of the profit-based loss matrix in Table 2. In contrast, the off-diagonal elements have a negative margin and correspond to non-optimal targeting decisions. Following the targeting profit equation (2), the loss associated with not targeting a profitable customer – such a customer has to be by definition a churner – equals $\pi_i = \gamma_i(V_i - \delta)$. The loss associated with targeting a non-profitable customer equals $|\pi_i| = |-\varphi_i\delta| = \varphi_i\delta$ when she is a non-churner, and $|\pi_i| = \gamma_i(\delta - V_i)$ when she is a churner. As a result, the loss associated with a negative margin can be written as

$$\Psi_{\text{one-zero}}(\pi_i, F_i) = |\pi_i|. \tag{6}$$

Table 2: Profit-based loss matrix

Loss Matrix	Targeted customer $\boldsymbol{F_i \geq 0}$	Non-targeted customer $\boldsymbol{F_i < 0}$
Positive targeting profit $\boldsymbol{\pi_i \geq 0}$	Targeted churner $\Psi = 0$	Non-targeted profitable churner $\Psi = \pi_i$
Negative targeting profit $\boldsymbol{\pi_i < 0}$	Targeted non-churner or Targeted unprofitable churner $\Psi = \lvert\pi_i\rvert$	Non-targeted non-churner $\Psi = 0$

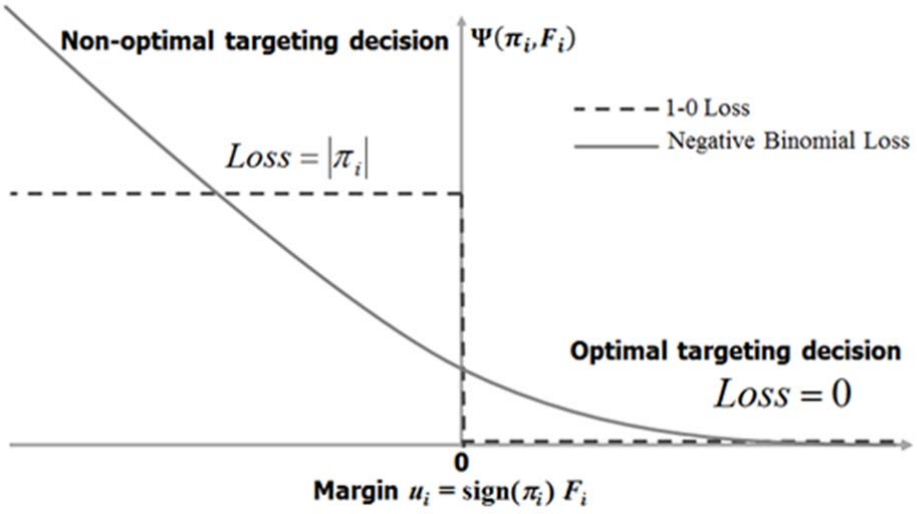
[4] In order to compute losses, we evaluate targeting with respect to the optimal targeting decisions. Other references can also be used, e.g. evaluating the loss with respect to targeting no customer at all. In such cases, targeting a profitable customer leads to a negative loss, i.e. a positive gain, while not targeting a profitable customer leads to a zero loss (compared to no retention action). However, the difference between both cases is unchanged, as also for the estimated scores.

The differentiable and continuous approximations of the profit-based loss functions can be obtained using the same negative binomial transformation as for the classical loss function above,

$$\Psi_{\text{negative-binomial}}(\pi_i, F_i) = |\pi_i| log(1 + e^{-2u_i}) \quad (7)$$

In Figure 2, we report the discontinuous version as well as their continuous and differentiable approximation of the profit-based loss function.

Figure 2: Profit-based loss function



In the next section, we explain how the stochastic gradient boosting algorithm works and how we integrate the profit-based loss function in the algorithm.

OPTIMIZATION WITH STOCHASTIC GRADIENT BOOSTING

Stochastic gradient boosting (hereafter, S.G.B.) is a greedy numerical optimization algorithm, originating from the machine learning literature and invented at Stanford University by Friedman and colleagues (Friedman 2002; Friedman, Hastie, and Tibshirani 2000). S.G.B.

sequentially combines the predictions of an ensemble of prediction models, typically regression trees (Breiman et al. 1983). Like other aggregation or ensemble methods such as bagging, Bayesian model averaging or random forests, the idea is that the predictive power of multiple trees can be combined. S.G.B. linearly combines *M* regression trees $tree_0(X), \dots, tree_M(X)$ to each other with weights $\beta_0, \dots, \beta_M$. The aggregated predicted scores *F* assigned to every customer can therefore written as (for ease of exposition we are dropping the consumer subscript i in this section)

$$F(X; \beta_0, \dots, \beta_M) = \sum_{m=0}^{M} \beta_m tree_m(X). \quad (8)$$

While other aggregation methods estimate the trees in parallel across iterations and combine all of them with equal weight, S.G.B. gradually refines its predictions over the iterations in such a way that the prediction errors made at the prior iteration are given more attention and minimized at the next iteration. Intuitively, the idea is the algorithm progressively concentrates on the customers that are difficult to classify until no more improvement can be made (Lemmens and Croux 2006).

Roughly, the iterative procedure works in two steps. First, it fits a regression tree with *L* terminal nodes on the pseudo-residuals obtained at the prior iteration. The regression tree is estimated using the well-established CART algorithm (Breiman et al. 1983), which consists growing trees by recursively splitting the population of customers in nodes using the entropy as splitting criterion. As suggested by Friedman (2001), the number of nodes is kept relatively small (2 to 5) in order to avoid overfitting. These pseudo-residuals represent the prediction errors made by the algorithm at the prior iteration, as specified by the loss function chosen by the researcher. The regression tree allocates each customer to one of the *L* terminal nodes. Customers allocated to the same node are predicted to have the same pseudo-residuals. As a consequence, their scores require the same update. Second, scores are updated per segment

(or node) in such a way that the respective loss in each segment is minimized. Once all scores are updated, a new iteration starts until the resulting loss does not change anymore.

Gradient Descent Optimization

The minimization per segment (or terminal node) is done using the gradient descent optimization. Gradient descent optimization works by finding a local minimum of a differentiable (loss) function. As the negative gradient of a function points out to the direction of "the steepest descent" of this function, the optimization takes steps proportional to the negative of the gradient (i.e. first-derivative) of the function at the current point until no more improvement is found. In our case, the goal is to minimize the profit-based customer-specific loss function defined in the previous sections. Hence, at each iteration, the algorithm takes a step in the direction of the negative gradient of the profit-based loss function. Note that, in practice, any differentiable loss function can be used depending on the specific needs and goals of the researcher. For details on the S.G.B. algorithm, we refer to the Appendix and to Friedman (2001, 2002).

We apply the S.G.B. algorithm to a calibration sample of customers (see data section below) and estimate the M regression trees and β_m weighting parameters on the calibration data. Using these estimates, we can compute (using equation (8)) the predicted score F for any customer (in or outside the calibration sample) for which we know the values of the X variables. This score represents the predicted targeting opportunity of the customer. The higher the score of this customer compared to the other customers' score, the higher the expected relative profit of targeting him rather than another customer. Consequently, targeting the customers in the top of the ranking with a retention action is predicted to yield the highest profit for the company.

OPTIMAL TARGET SIZE SELECTION

Using S.G.B., we can rank a set of customers in terms of the predicted targeting opportunity. In order to maximize the profit of a retention campaign, an important question remains: how many customers, starting from the top of the ranking, should the firm target with a retention action? In order to determine the optimal target size, we compute the cumulative targeting profit for various target sizes on a separate validation sample (see data section below). We use a different sample for two reasons. First, it reduces the overfitting that would be induced by ranking and selecting customers on the same data set. Second, we want to determine the optimal number of customers to target on a proportion sample (in which the proportion of churners is similar to the actual proportion of churners). The data set used for estimation instead contains an oversampled number of churners in order to facilitate the characterization of churners (Lemmens and Croux 2006).

Knowing the X values of the customers in the validation sample, we calculate their F scores and rank them accordingly. Next, we calculate the actual cumulative targeting profit for various target sizes using the profit equations (1) and (2). Finally, we select the target size that yield the highest total profit over the customer base. We assess the performance of our approach and the current approaches on a separate test or hold-out sample (see data section), which has been used neither for model estimation, nor for target size selection.

DATA

The study is performed on the database provided by the Teradata Center at Duke University in the context of the 2002 Churn Modeling Tournament. This database contains three datasets of mature subscribers (i.e. customers who were with the company for at least six

months) of a major U.S. wireless carrier.[5] As in most telecommunication companies in the US, customers are in a contractual relationship with the firm. The actual average monthly churn rate in the data is around 1.8%. The company wants to know if a certain customer should receive a retention offer 31-60 days after the sampling date. A delay of one month is justified as the implementation of proactive customer retention incentives requires some time.

The first dataset is a *balanced* sample of 10,000 customers, which contains an over-sampled number of churners. In this sample, the number of churners is perfectly balanced by the number of non-churners (i.e. 50 % churners and 50 % non-churners). It is used as calibration sample to fit the S.G.B. model. The reason for oversampling churners in the calibration sample is to avoid the possibility that the vast majority of non-churners may dominate the statistical analysis and hinder the detection of churn drivers, thus eventually decreasing the predictive accuracy of the model. A discussion on such a procedure, its advantages and drawbacks can be found in Donkers, Franses, and Verhoef (2003), King and Zeng (2001a,b) and Lemmens and Croux (2006).

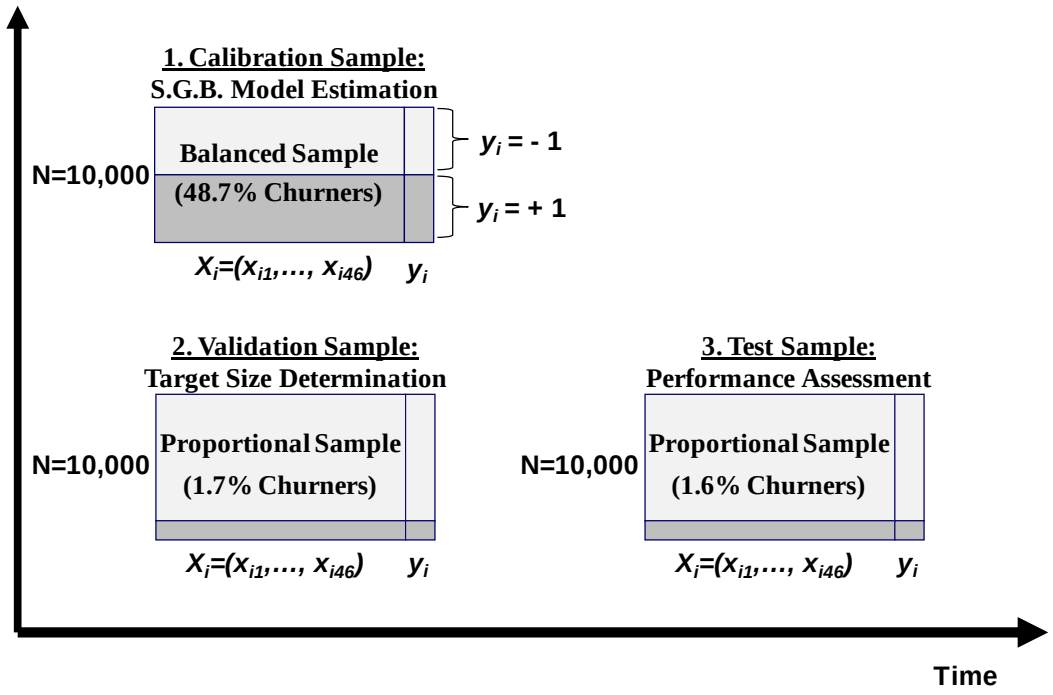
The second dataset is a *proportional* sample, in which the proportion of churners is representative of the proportion of churners in the customer base (i.e. 1.68% churners). Similar to the calibration sample, this sample also contains 10,000 customers. As detailed in the previous section, this second dataset is used as validation sample to determine the optimal target size of the retention campaign. To do so, we use the model estimates from the calibration sample on the validation sample data and compute the profit per customer for different target sizes. We then select the target size that yields the highest profits.

The third dataset is a *proportional* sample that contains 10,000 customers, of whom 1.57% are churners. It is used as a hold-out sample in order to assess the predictive performance of the model estimated on the calibration sample, given the optimal target size deter-

[5] For this study, we randomly select 10,000 customers from each of these datasets.

mined from the validation sample. This third dataset contains customer data during a period of time subsequent to the period recorded in the first two datasets, in order to match a real-life setting. Note that all three samples contain a different set of customers. Figure 4 illustrates graphically the various data sets used.

Figure 4: Calibration, Validation and Test Samples and Procedure



To predict the churn potential of customers, U.S. wireless operators usually take into account between 50 and 300 subscriber variables (Hawley 2003). From the high number of variables contained in the initial database (171 variables), we retain the same 46 variables as in Lemmens and Croux (2006). In that study, these variables were selected by excluding all the variables that contained more than 30% missing values, as well as the variables that were the least relevant following the results of a principal components analysis. Variables include 31 continuous and 15 categorical variables. Retained predictors include behavioral (e.g. the

average monthly minutes of use over the previous three months), company interaction (e.g. mean unrounded minutes of customer care calls), and customer demographics (e.g. the number of adults in the household, or the education level of the customer) variables (see Lemmens and Croux (2006) for an overview).

BENCHMARKS

In order to show the managerial relevance and impact of our approach, we test it against a number of benchmarks inspired from the existing business and academic practice. The comparison of various approaches can be characterized along the following dimensions: (i) how scores are estimated in step 1, (ii) how customers are ranked (hereafter denoted step 1b), and (iii) how the target size is determined in step 2. In Table 3, we provide a summary of all our benchmark approaches.

For the score estimation in step 1 various binary classification methods are used in practice. According to Verhoef (2003), the most popular among academics and managers include logistic regression and classification trees and, more recently S.G.B. (with the classical misclassification loss function). Prior research on the same data sets found that that S.G.B. substantially outperforms the other methods (Lemmens and Croux 2006). We therefore use S.G.B. and apply different loss functions to it. This choice also allows to evaluate how the specification of a profit-based loss function improves the performance of a given prediction method (here, S.G.B). In the following application, we therefore test the performance of S.G.B. with the profit-based loss function (IIIa and IIIb in Table 3) against the S.G.B. algorithm with its original misclassification loss function (Ia and Ib in Table 3).

For the customer ranking in step 1b, the classical approach consists of using the customer scores directly and rank customers from the highest to the lowest predicted score F_i (I and III in Table 3). As such, customers are ranked according to their expected churn probabil-

ity when using the misclassification loss function (I) or according to their expected targeting profit when using the profit-based loss function (III).

A potentially compelling variant that we investigate is to use the original misclassification loss but to subsequently re-order customers such that factors other than churn that determine their profitability are taken into account as well (II in Table 3). This reordering of customers is done by transforming their expected churn probabilities into expected targeting profits. We achieve this by transforming the estimated scores $\widehat{F}_i$ into defection probability estimates $\widehat{p}_i$ using the inverted log-odds formula,[6] taking into account the high skewness distribution of the churn outcome variable. To do so, we use the intercept-correction method suggested by Lemmens and Croux (2006). Based on the estimated churn probabilities, we then calculate the expected profit of targeting customer *i* as

$$\hat{\pi}_i = \widehat{p}_i[\gamma_i(V_i - \delta)] + (1 - \widehat{p}_i)[-\varphi_i\delta], \tag{9}$$

the equivalent of equation (2) when churn is a probability variable rather than a binary outcome. We can now rank customers based on their expected profit $\hat{\pi}_i$, rather than based on the score $\widehat{F}_i$. By doing so, this modified customer ranking now takes into account the expected heterogeneity in profit. This ranking is likely to yield higher profits than the unmodified ranking and can potentially be a good competitor to the profit-based loss function. The question remains opens as to whether this modified ranking compensates for the fact that the estimation of the scores aims at minimizing the misclassification rate rather than maximizing the profit of the retention campaign. By comparing both versions, we can assess the impact of directly incorporating the targeting profit in the loss function rather than modifying the customer ranking in the second step.

[6] Note that the ranking is not affected by the log-odds transformation of the scores into probabilities, only the scale of the variable changes.

For determining the target size in step 2, we propose two alternatives: either select an arbitrary target size, as often done in business practice (a in Table 3); or use the optimal target size selection we propose (b in Table 3). To ensure a fair comparison with the optimal target size selection, we use information from the validation sample to decide on the arbitrary target size. We therefore use as target size the percentage of churners in the validation sample. i.e. 1.68%. We assume that the firm wants to target as many customers as there are potential churners in the data. When comparing approaches Ia and IIIb, we can assess the total impact of our approach, relative to a traditional approach that does not take profit into account in the design of the retention campaign. In turn, when comparing IIb and IIIb, we can assess the impact of changing the loss function into the S.G.B. algorithm.

Table 3: Competing approaches

Competing approach	**Step 1: Score Estimation**	**Step 1b: Customer Ranking**	**Step 2: Target Size Selection**
(Ia)	S.G.B. with misclassification loss	Churn propensity	Fixed size
(Ib)	S.G.B. with misclassification loss	Churn propensity	Optimized size
(IIa)	S.G.B. with misclassification loss	Expected profit	Fixed size
(IIb)	S.G.B. with misclassification loss	Expected profit	Optimized size
(IIIa)	S.G.B. with profit loss	Expected profit	Fixed size
(IIIb)	**S.G.B. with profit loss**	**Expected profit**	**Optimized size**

RESULTS

We apply our proposed profit-oriented approach to the Teradata database described above. We estimate S.G.B. with the profit-based loss function on the calibration data, select the optimal target size on the validation data and evaluate the performance of the approach on a hold-out test sample. In addition, we also apply all the benchmarks described in the previous section.

We examine the performance of our approach against the benchmark approaches across a wide range of action costs (δ), \$30 to \$70; and consumer response probabilities to the retention actions (γ_i), 10% to 50%. We also assume that the response probability of non-churners is always higher than the response probability of the churners, here 70%. Finally, we measure the lifetime value of a customer to the firm V_i using the monthly base cost of her calling plan during the observation period. We assume that the more expensive the calling plan of the customer, the higher the revenues the firm would lose if the customer churns. To compute the value of a customer over her lifetime, we use the population churn rate as discounting factor. Implicitly, we make the assumption that the probability of churn for a customer retained with a retention action goes back to the *average* population churn rate.

In Table 4, we report the profit for our hold-out sample of 10,000 customers (in US \$) for each of the six competing approaches summarized in Table 3. In addition, we also report the chosen target size for the cases where the target size is optimized (Ib, IIb and IIIb). First, we find that the profit loss function is consistently more profitable than the classical misclassification loss function. In the case of a response rate = 30% and an action cost = \$40, the optimal target size for the profit loss (IIIb) is 2.21% of the customer base and the hold-out profit is approximately \$3,433. In contrast, using the misclassification loss function and optimal target size (Ib) produces a profit of \$836. In other words, our profit-based approach improved the profit in this scenario by \$2,597 or 310%. For a company such as Verizon Wireless, the largest wireless services provider in the United States with 111.3 million subscribers, such difference in predictive performance would yield a difference of more than \$28 million. In general, we find that our approach yields better profits (\$3,700) than the classical loss function (\$1,718) across all scenarios, as the last row of Table 4 testifies. The improvement in profit is on average equal to 115%. Such improvement is substantial, especially if one thinks it only came by just specifying a different loss function, and without any changes

in how these retention programs are implemented. Knowing that the S.G.B. with the classical loss was already substantially outperforming other models that companies use in practice, such as logit or CART, such a big profit difference is significant and managerially relevant.

Regarding the optimization of the target size, we find that the hold-out profits generally improve when the target size is optimized rather than a priori determined (IIIb compared to IIIa). For a response rate = 30% and an action cost = $40, the hold-out profit in our approach is $3,433 when target size is optimized (2.21%), versus $2,985 when the target size is determined a-priori (1.68%), an increase of 15%. Across scenarios, optimizing the target size (for the profit loss) yields an increase of 84% in profit ($3,700 compared to $2,014). Note, however that, in some cases, using a target size equal to the number of expected churners in the hold-out sample (i.e. using as proxy the proportion of churners in the validation sample) works very well. The happens because the optimization is done on the validation sample but the profits are measured on a different hold-out sample, which, by nature, can lead to small discrepancies in the size of the optimal target size. In general, we find that using a profit loss function has a stronger effect on the resulting profits than optimizing the target size.

In addition, we also find that the optimal target size increases as the positive response rate goes up and as the action costs go down. These relationships are intuitive. The higher the positive response rate and the lower the action costs, the less expensive it becomes to target customers and the higher the number of churners retained if targeted.

In Table 4, we also see that, in general, reordering the scores obtained from the misclassification loss in order to take profit into account (II) does not improve the performance of the targeting decisions. On average, the hold-out profits equal $1,452 across scenarios (which means $2,248 less compared to the profit loss). The only exception occurs when the retention costs are very low and the positive response rate is very high. This makes targeting customers relatively inexpensive and therefore targeting more customers is better.

The two-step reordering approach (II) is not working well since in its first step it minimizes the churn misclassification rate across all customers, regardless of their individual targeting profits. As a consequence, rearranging the customer ranking (using the targeting profit formula) does not allow the method to recover from the misclassification errors made in the first step. In contrast, our approach minimizes the misclassifications towards the most profitable customers right from the start.

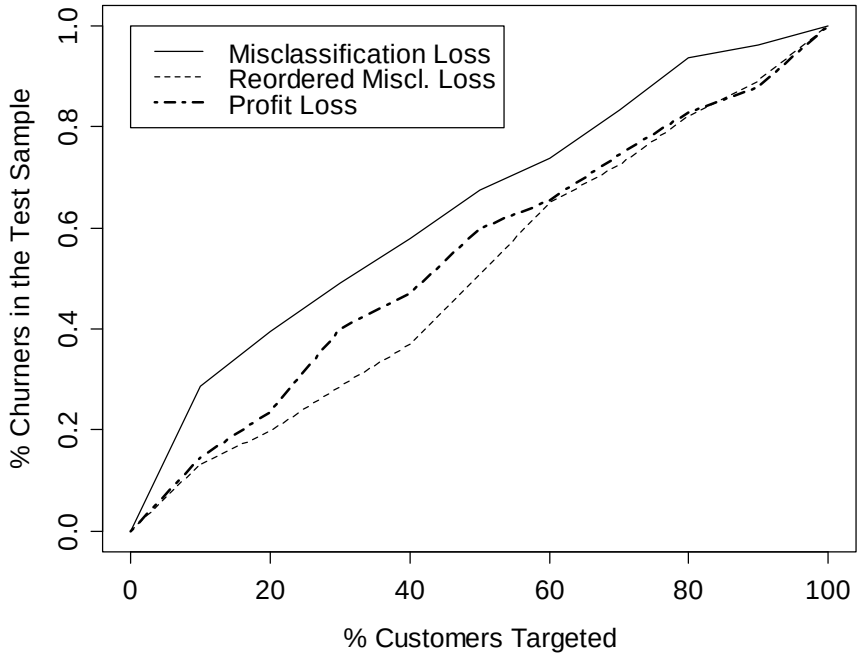
Table 4: Hold-out profit comparison across the competing approaches (US $)

		I. Misclassification loss			**II. Misclassification Loss Reordered on Expected Profit**			**III. Profit Loss**		
		a. Fixed Size (1.68%)	b. Optimized Size		a. Fixed Size (1.68%)	b. Optimized Size		a. Fixed Size (1.68%)	b. Optimized Size	
Positive Response Rate γ_i	Retention Action Cost δ	Hold-out Profit ($)	Target Size (% cust.)	Hold-out Profit ($)	Hold-out profit ($)	Target Size (% cust.)	Hold-out Profit ($)	Hold-out profit ($)	Target Size (% cust.)	Hold-out Profit ($)
10%	$70	(6,339)	0.00%	-	(6,102)	0.00%	-	(5,573)	0.00%	-
20%	$70	(5,035)	0.00%	-	(4,217)	0.00%	-	(3,725)	0.00%	-
30%	$70	(3,730)	0.41%	(487)	(2,333)	0.00%	-	(1,039)	0.61%	**924**
40%	$70	(2,425)	0.81%	124	(448)	0.00%	-	3,266	1.81%	**3,956**
50%	$70	(1,121)	0.81%	1,036	1,437	1.61%	134	5,044	1.81%	**7,027**
10%	$60	(5,235)	0.00%	-	(4,956)	0.00%	-	(4,673)	0.00%	-
20%	$60	(3,919)	0.00%	-	(3,066)	0.00%	-	(2,170)	0.61%	(514)
30%	$60	(2,602)	0.41%	(220)	(1,177)	0.00%	-	105	0.61%	**1,335**
40%	$60	(1,285)	0.81%	664	713	1.61%	(316)	3,166	1.81%	**6,513**
50%	$60	31	1.01%	1,717	2,603	1.61%	1,253	7,174	2.01%	**7,432**
10%	$50	(4,131)	0.00%	-	(3,810)	0.00%	-	(3,317)	0.00%	-
20%	$50	(2,803)	0.00%	-	(1,915)	0.00%	-	(1,443)	0.61%	**499**
30%	$50	(1,474)	0.81%	273	(21)	1.61%	(775)	**2,786**	1.61%	1,494
40%	$50	(145)	1.01%	1,287	1,874	1.61%	799	3,532	2.01%	**7,044**
50%	$50	1,183	6.61%	3,612	3,769	1.61%	2,372	**8,330**	2.01%	7,175
10%	$40	(3,027)	0.00%	-	(2,664)	0.00%	-	(2,405)	0.00%	-
20%	$40	(1,687)	0.41%	(127)	(764)	0.00%	-	102	0.61%	**906**
30%	$40	(346)	1.01%	836	1,135	1.61%	336	2,985	2.21%	**3,433**
40%	$40	995	6.61%	3,026	3,035	6.41%	2,812	6,704	2.01%	**8,415**
50%	$40	2,335	6.61%	8,171	4,935	8.01%	7,474	**9,486**	2.41%	9,082
10%	$30	(1,923)	0.00%	-	(1,518)	0.00%	-	(1,271)	0.61%	**58**
20%	$30	(571)	0.81%	386	387	1.61%	(134)	1,637	2.21%	**1,821**
30%	$30	782	6.61%	2,371	2,291	6.41%	2,154	4,125	2.01%	**5,356**
40%	$30	2,135	6.61%	7,551	4,196	8.01%	7,148	7,850	2.21%	**9,366**
50%	$30	3,487	6.61%	12,730	6,101	8.01%	**13,045**	9,676	2.21%	11,167
Average profit over 10,000 customers		(1,474)		1,718	(21)		1,452	2,014		**3,700**

It is important to realize that the improvement in financial profits is not associated with an improvement in the number of churners targeted. On the contrary, the profit loss actually identifies fewer churners in the top of its ranking than the misclassification loss, but the churners correctly identified are more profitable. A comparison of the lift curves in Figure 5

confirms these results. The chart reports the percentage of churners in the test sample contained in a given target, as a function of the target size. The chart shows that the traditional predictive performance criteria (e.g. lift, top-decile, …) can be extremely misleading for managers since they focus almost exclusively on the predictive accuracy of churn rather than on financially-relevant predictive performance criteria.

Figure 5: Lift Charts for the Various Loss Functions for Different Target Sizes (%)



Finally, we investigate whether the identification of the most important churn triggers is sensitive to the choice of the loss function. In other words, we assess whether a profit-based ranking yields a different set of customer characteristics than a ranking based on the classical loss function. In Figures 6 and 7, we report the 15 most important churn triggers, ordered by importance, according to the classical misclassification loss (Figure 6) and as per the profit loss (Figure 7).

Figure 6: Most Important Churn Triggers according to the Misclassification Loss

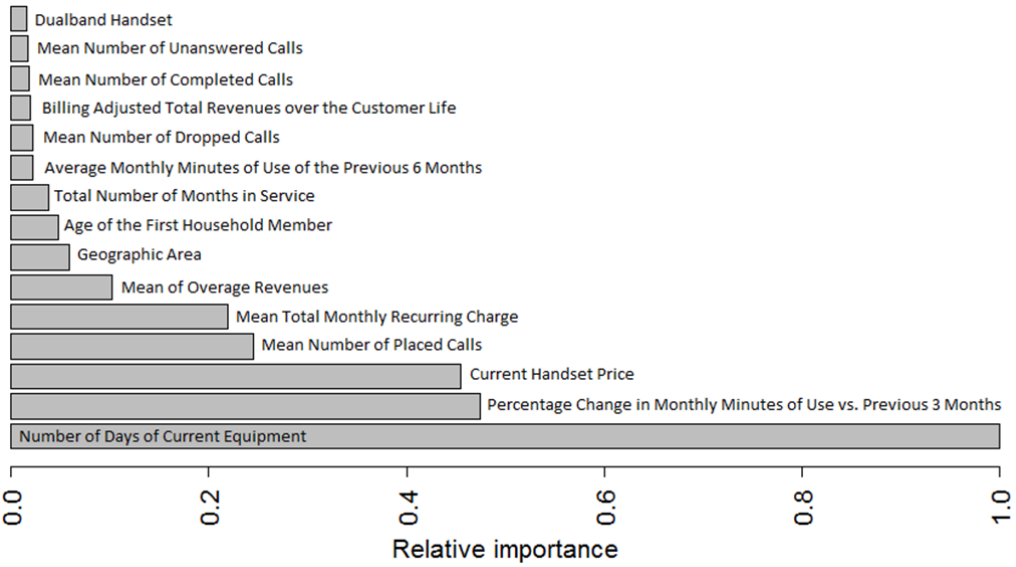
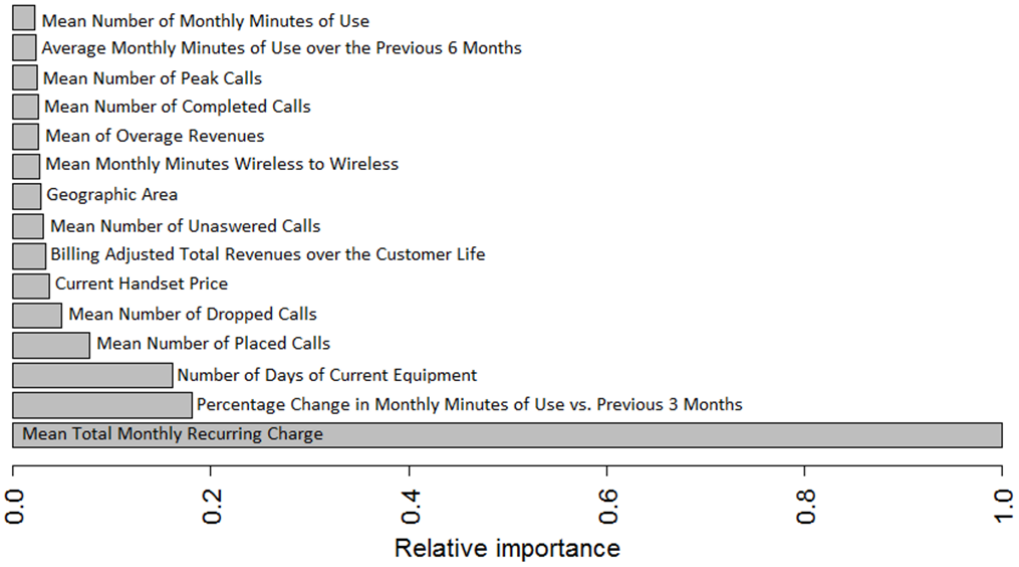


Figure 7: Most Important Churn Triggers according to the Profit Loss



The variables' relative importance is computed following Friedman (2001) using the relative empirical improvement in fit resulting for each node split (Breiman et al. 1983), averaged across all iterations of the boosting algorithm. While 12 of the 15 triggers identified as most important by the classical approach are also relevant when adopting a profit-based approach, the order of importance does vary between both approaches. Interestingly, the profit-based approach recommends focusing on the monthly customer value (i.e. mean total monthly recurring charge), while the classical approach identifies the age of the equipment as the most important churn trigger.

CONCLUSIONS, LIMITATIONS AND FUTURE RESEARCH

In this paper, we provide a novel method for determining which customers to target in order to maximize the profit of a retention campaign. We explain the need for companies to pay more attention to the choice of the loss function. The latter should match their marketing objectives. Aligning the loss function to the objectives of their marketing actions rather than blindly applying any binary prediction method at hand has a substantial effect on the impact of their actions. We develop a profit-based loss function and use S.G.B. to optimize this loss function. In addition, we also optimize the size of the target. We show that our method leads to substantial improvements for companies with no additional implementation cost. Interestingly, we find that the profit-based loss function leads to more errors in terms of forecasting which customers are likely to leave the firm. However, these errors are less costly in terms of firm profitability than the errors made by the traditional loss function.

Our contribution complements the expanding literature on customer lifetime value in two substantive ways. First, it differs from a stream of literature that uses CLV to guide targeting decisions, but ignores heterogeneity in customer churn. For instance, the work by Kumar et al. (2008), Ventakesan and Kumar (2004), as well as Venkatesan, Kumar, and Bohling

(2007) determines how much marketing resources should be spent on every customer based on their heterogeneity in future spending. Such targeting decisions are useful for customer development (rather than retention) strategies. However, in contrast to our approach, they do not focus on the heterogeneity in the defection probability of customers, which is assumed to be homogenous across the customer base and independent of the marketing actions. Second, our approach also differs from recent developments on probability models of customer behavior to predict future CLV (e.g. Fader, Hardie, and Lee 2005) that account for heterogeneity in defection probability across the customer base, but do not focus on how customer characteristics and firm actions can influence it. As a result managers cannot use these models for managing specific retention campaigns.

Our paper also contains a number of limitations that offer fruitful avenues for future research. First of all, we do not observe the actual response rate of customers to retention actions and we are therefore unable to estimate the response probability of each customer to a given action. However, we show by making a number of realistic assumptions, that using a managerially-relevant loss function – even if partially inaccurate, has a huge impact of targeting decisions and has the potential to greatly improve profits. Using more precise estimates of the customers' response probability to retention actions has the potential to leverage the benefits of our approach even more. When data are available, it would be interesting to model the response probability of customers, as done by Neslin et al. (2009). They use an optimization method to maximize the response rates, which could be incorporated into our retention modeling framework. This approach can potentially be used in sequence or in parallel with our optimization method. Based on an estimate of the customers' response probability, we can update our loss function and our customer ranking. Along those lines, it would also be interesting to test in a field experiment the effect of different action costs of the customer response

probabilities as well as the possibility of using different actions towards different customer segments.

Second, our approach assumes that, once a customer who intended to churn is effectively retained after being targeted with a retention action, her churn propensity returns to the population churn rate and her spending pattern is unchanged. Some customers could potentially experience a so-called delight effect (see Blattberg, Kim, and Neslin 2008, p. 630). Using longitudinal data on customer spending, it would be possible to relax this assumption by jointly modeling future spending stream of customers (see e.g. Rust, Kumar, and Venkatesan 2011), and their defection behavior (see e.g. Fader, Hardie, and Lee 2005; Fader and Hardie 2010).

Finally, the approach presented in this paper can be potentially used in other contexts where other non-profit related loss functions can be the main priority of the decision maker. For instance, patient compliance to a medical treatment in the health sector is an important variable to predict. In such a context, the loss function can incorporate the risk for a given patient health of not complying (i.e. churning) to the medical treatment.

REFERENCES

Andrews, Rick L., Andrew Ainslie, and Imram S. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39, 479-487.

Ascarza, Eva And Bruce G.S. Hardie (2013), "A Joint Model of Usage and Churn in Contractual Settings," *Marketing Science*, Forthcoming.

Bhattacharya, C.B. (1998), "When Customers are Members: Customer Retention in Paid Membership Contexts," *Journal of the Academy of Marketing Science*, 26, 31-44.

Blattberg, Robert C. and Edward I. George (1992), "Estimation Under Profit-Driven Loss Functions," *Journal of Business and Economic Statistics*, 10 (4), 437-444.

Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin (2008), *Database Marketing: Analyzing and Managing Customers*, International Series in Quantitative Marketing. Springer.

Bobbier, Tony (2013), "Keeping the Customer Satisfied: The Dynamics of Customer Defection, and the Changing Role of the Loss Adjuster," CILA Report, http://www.cila.co.uk/publication/articles/keeping-customer-satisfied-tony-boobier, accessed July 2013.

Bolton, Ruth N. (1998) "A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction," *Marketing Science*, 17 (1), 45-65.

Bolton, Ruth N., P.K. Kannan and Matthew D. Bramlett (2000), "Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value," *Journal of the Academy of Marketing Science*, 28, 95-108.

Braff, Adam, William J. Passmore, and Michael Simpson (2003), "Going the Distance with Telecom Customers," *McKinsey Quarterly,* November, 1-16.

Breiman, L., Jerome H. Friedman, Richard Olshen, and Charles Stone (1983), *Classification and Regression Trees*, Wadsworth.

Bult, Jan Roelf (1993), "Semiparametric versus Parametric Classification Models: An Application to Direct Marketing," *Journal of Marketing Research*, 30, August, 380-390.

Bult, Jan Roelf., Dick R. Wittink, (1996), "Estimating and Validating Asymmetric Heterogeneous Loss function applied to Health Care Fund Raising," *International Journal of Research in Marketing*, 13, 215-226.

Donkers, Bas, Philip H.B.F. Franses, and Peter Verhoef (2003), "Selective Sampling for Binary Choice Models," *Journal of Marketing Research*, 40, 492-497.

Engle, Robert F. (1993), "A Comment on Hendry and Clements on the Limitations of Comparing Mean Squared Forecast Errors," *Journal of Forecasting*, 12, 642-644.

Fader, Peter S. and Bruce G.S. Hardie (2010), "Customer-Base Valuation in a Contractual Setting: the Perils of Ignoring Heterogeneity," *Marketing Science*, 29 (1), 85-93.

Fader, Peter S., B. Hardie, K.L. Lee (2005), "RFM and CLV: Using Iso-value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42 (November), 415–430.

Forbes (2011), "Bringing 20/20 Foresight to Marketing: CMOs Seek a Clearer Picture of the Customer," *Forbes Insights*, 1-13.

Friedman, Jerome H. (2001), "Greedy Function Approximation: a Gradient Boosting Machine," *The Annals of Statistics*, 29 (5), 1189-1232.

Friedman, Jerome H. (2002), "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis*, 38, 367-378.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2000), "Additive Logistic Regression: a Statistical View of Boosting," *The Annals of Statistics*, 28 (2), 337-374.

Ganesh, Jaishankar, Mark J. Arnold, and Kristy E. Reynolds (2000), "Understanding the Customer Base of Service Providers: An Examination of the Differences between Switchers and Stayers," *Journal of Marketing*, 65, 65-87.

Gilbride, Thimothy J., Peter J. Lenk, and Jeff D. Brazell (2008), "Market Share Constraints and the Loss Function in Choice-Based Conjoint Analysis," *Marketing Science*, 27 (6), 995-1011.

Glady, Nicolas, Bart Baesens, and Christophe Croux (2009), "Modeling Churn using Customer Lifetime Value," *European Journal of Operational Research*, 197, 402-411.

Granger, Clive W.J. (1993), "On the Limitations of Comparing Mean Square Forecast Errors: Comment," *Journal of Forecasting*, 12, 651-652.

Gupta, Sunil, Don R. Lehmann, and Jennifer A. Stuart (2004), "Valuing Customers," *Journal of Marketing Research*, 41 (February), 7-18.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning (*second edition), Springer: New York.

Hawley, David (2003), "International Wireless Churn Management Research and Recommendations," *Yankee Group Report*, June.

King, Gary and Langsche Zeng (2001a), "Explaining Rare Events in International Relations," *International Organization*, 55, 693-715.

King, Gary and Langsche Zeng (2001b), "Logistic Regression in Rare Events Data," *Political Analysis*, 9, 137-163.

Kumar, V., Rajkumar Venkatesan, and Timothy Bohling (2008), "Practice Prize Report: The Power of CLV: Managing Customer Lifetime Value at IBM," Marketing Science, 27 (4), 585–599.

Lariviere, Bart and Dirk Van den Poel, (2005), "Predicting Customer Retention and Profitability by Using Random Forest and Regression Forest Techniques," *Expert Systems with Applications*, 29 (2), 472-484.

Lemmens, Aurélie and Christophe Croux (2006), Bagging and Boosting Classification Trees to Predict Churn, *Journal of Marketing Research*, 43 (May), 276-286.

Lemon, Katherine N., Tiffany B. White, and Russell S. Winer (2002), "Dynamic Customer Relationship Management: Incorporating Future Considerations into the Service Retention Decision," *Journal of Marketing*, 66 (January), 1-14.

Neslin, Scott A. (2002), "Cell2Cell Case Notes," Hanover, NH: Tuck School of Business, Dartmouth College.

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason (2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43 (2), 204-211.

Neslin, Scott A., Thomas P. Novak, Kenneth R. Baker, and Donna L. Hoffman (2009), "An Optimal Contact Model for Maximizing Online Panel Response Rates," *Management Science*, 55 (5), 727–737.

Risselada, Hans, Peter C. Verhoef, and Tammo H.A. Bijmolt (2010), "Staying Power of Churn Prediction Models," *Journal of Interactive Marketing*, 24 (3), 198-208.

Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22 (3), 304-328.

Rust, Roland, V. Kumar and Rajkumar Venkatesan (2011), "Will the Frog Change into a Prince?: Predicting Future Customer Profitability," *International Journal of Research in Marketing*, 28 (4), 281-294.

Schweidel, David, Peter Fader, Eric Bradlow (2008), "A Bivariate Timing Model of Customer Acquisition and Retention," *Marketing Science*, 27, 829-843.

Shaffer, Greg and John Z. Zhang (2002), "Competitive One-to-One Promotions," *Management Science*, 48, 1143-1160.

Thieme, R. Jeffrey, Michael Song, and Roger J. Calantone (2000), "Artificial Neural Network Decision Support Systems for New Product Development Project Selection," *Journal of Marketing Research*, 37, 499-507.

Venkatesan, Rajkumar, and V. Kumar (2004), "A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy," *Journal of Marketing*, 68 (October), 105-125.

Venkatesan, Rajkumar, V. Kumar, and T. Bohling (2007), 'Optimal CRM Using Bayesian Decision Theory; An Application to Customer Selection," *Journal of Marketing Research*, 44 (4), 579-594.

Verbeke, Wouter, Karel Dejarger, David Martens, Joon Hur, and Bart Baesens (2012), New Insights into Churn Prediction in the Telecommunication Sector: a Profit Driven Data Mining Approach, *European Journal of Operational Research*, 218, 211-229.

Verhoef, Peter C. (2003), "Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development*," Journal of Marketing*, 67 (4), 30-45.

West, Patricia M., Patrick L. Brockett, and Linda L. Golden (1997), "A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice," *Marketing Science*, 16, 370-391.

Yang, Sha and Greg M. Allenby (2003), "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research*, 40, 282-294.

Appendix: Stochastic Gradient Boosting Algorithm

The boosting algorithm starts with an initial guess $F_0(X)$ of all customers' scores,

$$F_0(X) = \frac{1}{2} log\left(\frac{p_0}{1-p_0}\right)$$

with p_0 equals to the proportion of churners in the sample. At each iteration $m = 1, \dots, M$, we randomly select a subsample (without replacement) of N' customers from the calibration data with $N' \leq N$. The randomization is a feature added by Friedman in his 2002 paper to ensure that the method does not suffer from overfitting.

We then compute the negative gradient values $grad_{im}$ of all customers given the customers' scores at $m-1$. The latter represent the pseudo-residuals from the previous iteration. Subsequently, a regression tree with *L* terminal nodes is then fitted to estimate the relationship between the negative gradient and the set of customer characteristics X. The tree allocates customers into *L* different segments based on their *X* values (depending on the terminal node they fall in) and predicts a different constant to each segment *l*. Formally,

$$tree_m(X) = \sum_{l=1}^{L} \bar{y}_{lm}\, 1(i \in l\,),$$

where $\bar{y}_{lm} = mean_{i \epsilon l}(grad_{im})$ is the mean of the negative gradient values

$$grad_{im} = -\left[\frac{\partial \Psi(y_i, F(X_i))}{\partial F(X_i)}\right]_{F(X)=F_{m-1}(X)}$$

in each region l.

As the tree predicts a constant value $\bar{y}_{lm}$ within each region l, we obtain, for the m^{th} tree,

$$\gamma_{lm} = argmin_{\gamma} \sum_{i \epsilon\, l} \Psi(y_i, F_{m-1}(X_i) + \gamma).$$

The current approximation $\mathrm{F}_{\mathrm{m}-1}(X)$ is then separately updated in each region

$$F_m(X) = F_{m-1}(X) + \nu \gamma_{lm} 1(i \epsilon\, l),$$

where ν is the learning rate or s parameter $0 < \nu \leq 1$, or

$$F_m(X) = F_{m-1}(X) + \beta_{lm}\, 1(i \epsilon\, l),$$

with $\beta_{lm} = \nu\, \gamma_{lm}$. The procedure runs over *M* iterations until no substantial improvement anymore is found in the resulting loss.