Managing Churn to Maximize Profits

Aurélie Lemmens

Rotterdam School of Management, Erasmus University Rotterdam

Sunil Gupta

Harvard Business School

Forthcoming at *Marketing Science*

Aurélie Lemmens (corresponding author) is Associate Professor of Marketing at the Rotterdam School of Management (lemmens@rsm.nl). 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. under a VENI (451-09-005) and VIDI (452-12-011) grants. Part of the work has been carried out while she was visiting Harvard Business School. The authors are deeply indebted to Eva Ascarza (Harvard), Christophe Croux (K.U. Leuven), Hannes Datta (Tilburg University), Bram Foubert (Maastricht) and Jason Roos (Rotterdam School of Management), for their considerate support.

Managing Churn to Maximize Profits

Abstract

Customer defection threatens many industries, prompting companies to deploy targeted, proactive customer retention programs and offers. A conventional approach has been to target customers either based on their predicted churn probability, or their responsiveness to a retention offer. However, both approaches ignore that some customers contribute more to the profitability of retention campaigns than others. This study addresses this problem by defining a profit-based loss function to predict, for each customer, the financial impact of a retention intervention. This profit-based loss function aligns the objective of the estimation algorithm with the managerial goal of maximizing the campaign profit. It ensures (1) that customers are ranked based on the incremental impact of the intervention on churn and post-campaign cash flows, after accounting for the cost of the intervention and (2) that the model minimizes the cost of prediction errors by penalizing customers based on their expected profit lift. Finally, it provides a method to optimize the size of the retention campaign. Two field experiments affirm that our approach leads to significantly more profitable campaigns than competing models.

Keywords: Defection, Field Experiments, Loss Function, Machine Learning, Proactive Churn Management, Profit Lift, Stochastic Gradient Boosting.

1. Introduction

Customer defection is a global phenomenon, as exemplified by the estimated 20% annual churn rates for credit cards in the United States and 20%–38% annual churn rates 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. According to McKinsey & Co. reducing churn could increase earnings of a typical U.S. wireless carrier by nearly 10% (Braff, et al. 2003).

Not surprisingly, top executives cite customer retention as a top marketing priority, which they pursue with higher retention budgets (Forbes 2011) and more sophisticated, proactive churn management programs. These retention programs attempt to target potential churners with incentives (Ganesh, et al. 2000), such as special offers, discounts, personalized (e)mail, or gifts, all of which aim to boost targeted customers' behavioral loyalty (Winer 2001).

For years, marketing practice and research have mainly focused on churn prediction and proposed ways to target customers according to their estimated churn risk (e.g., Ascarza and Hardie 2013, Lemmens and Croux 2006, Neslin, et al. 2006, Risselada, et al. 2010, Schweidel, et al. 2011, for a review see, Ascarza, et al. 2018). Despite the popularity of this approach, recent studies have found that ranking customers on the basis of churn probability may lead to ineffective retention campaigns. Instead, Ascarza (2018) and Guelman, et al. (2012) propose to rank customers on the basis of their sensitivity to the intervention, regardless of their risk of churning. Using *uplift* random forests, they identify customers for whom the intervention will prompt the greatest lift in retention probability.

In both cases however, existing approaches fail to recognize the ultimate goal of companies to maximize the *profit* of their proactive retention campaigns. First, the rankings provided by both approaches are solely based on churn (risk or lift) and ignore the profit impacts of a retention intervention. The *profit lift* of a proactive retention incentive can be

estimated according to the intervention's (1) impact on the churn probability of the targeted customer, (2) impact on post-campaign cash flows, and (3) cost. A positive profit lift indicates that the intervention is likely to increase their retention probabilities and/or post-campaign revenues sufficiently to compensate for the intervention cost. In contrast, negative profit lifts signal cases in which targeting would lead to a loss for the company. By focusing on churn lift rather than on profit lift, prior approaches might end up targeting customers for whom the effect on retention is positive but the profit lift is negative.

Second, past approaches to obtain customer rankings ignore that prediction errors are more consequential in terms of campaign profit for some customers than others. Churn models attempt to minimize misclassifications of all customers' churn, regardless of their profit potential. Likewise, uplift models aim at estimating each customer's conditional average treatment effect as accurately as possible. In reality however, not all customers are equally valuable, and incorrectly predicting the churn risk or lift would be costlier for some customers than others. For example, failing to predict the response of a high profit-lift customer will have a larger financial impact on the campaign profit than failing to predict the response of a customer who is insensitive to the intervention.

We propose a new approach that addresses these two limitations. Our approach defines a profit-based loss function that estimates, for every customer, the expected profit lift taking into account the customer-specific cost of a prediction error. In contrast to existing approaches, the profit-based loss function fully aligns with the managerial objective of the retention campaign. The profit-based loss function weights more heavily those customers who have the greatest (positive or negative) impact on the campaign profit. This weighted estimation offers more accurate predictions for high profit-lift customers, and thus boosts the profitability of the campaign. We empirically demonstrate the superiority of our approach for two retention campaigns: for a European interactive television subscription firm (Datta, et al. 2015), and for

a special interest membership organization in North America.

The remainder of this article is organized as follows: In Section 2, we review existing approaches to customer retention and explain how they differ from a profit-based estimation approach. In Section 3, as a key building block for our approach, we define the profit lift of a retention program. In Section 4, we construct the profit-based loss function. Section 5 describes the various steps of our approach. Section 6 outlines the alternative methods we use as benchmarks. In Section 7, we present our empirical applications and results. Finally, we conclude in Section 8 with some limitations and potential extensions.

2. Existing Approaches for Proactive Retention Management

Most customer retention research focuses on predicting churn. In a modeling tournament organized by the Teradata Center at Duke University, 44 academics and practitioners proposed models (for overviews, see Blattberg, et al. 2008, Neslin, et al. 2006), including logistic regression (Knox and Van Oest 2014, Risselada, et al. 2010), probit models (Datta, et al. 2015), decision trees or CART (Huang, et al. 2012), neural nets (Huang, et al. 2012), random forests (Larivière and Van den Poel 2005), bagging and boosting (Lemmens and Croux 2006), hazard models (Bolton 1998, Braun and Schweidel 2011, Donkers, et al. 2007, Schweidel, et al. 2008a), hidden Markov models (Ascarza and Hardie 2013, Schweidel, et al. 2011, Schweidel and Knox 2013), simultaneous equation models (Reinartz, et al. 2005), probability models (Borle, et al. 2008, Fader and Hardie 2010, Fader, et al. 2010, Singh, et al. 2009), and heuristics (Wübben and Von Wangenheim 2008).

Most of these articles focus on predicting churn, instead of estimating the *impact* of marketing interventions on churn. This perspective contrasts with other areas of marketing that focus on the incremental "effect of a marketing action to inform targeting decisions" (Ascarza 2018, p.82, see e.g., Khan, et al. 2009, Kumar, et al. 2008, Lewis 2005b, Neslin, et al. 2009, Rossi, et al. 1996, Venkatesan and Kumar 2004, Venkatesan, et al. 2007). Some notable

exceptions include estimating the impacts of promotional activities (Schweidel, et al. 2008b), retention dollars (Reinartz, et al. 2005), or targeting the social network of an individual (Godinho de Matos, et al. 2018) on churn and profit, but the effects are estimated at an *aggregate* or *segment*-specific level. Guelman, et al. (2012), Datta, et al. (2015) and Ascarza (2018) propose targeting customers according to their *individual* churn lift or change in churn due to marketing intervention, by estimating the heterogeneous treatment effect of the marketing action on churn.

Despite their differences (focus on churn risk or churn lift), all these approaches seek to minimize prediction errors for all customers regardless of their contribution to the campaign profit. Churn models aim to minimize the percentage of churners classified as non-churners or vice versa; uplift models seek to reduce prediction errors in churn lift for all customers. The same limitation applies to studies that model customer churn and usage jointly to account for the dependence between these processes (Ascarza and Hardie 2013, Borle, et al. 2008, Datta, et al. 2015, Donkers, et al. 2007), which is conceptually and mathematically different from penalizing the prediction errors of a model based on profit lift.

Empirical research thus tends to ignore the risk of using a loss function that is not aligned with managerial objectives. Yet the loss function is integral to the model specification. It implicitly defines the model under consideration (Engle 1993) and should reflect the focal business problem (Christoffersen and Jacobs 2004, Granger 1969). When different loss functions apply to in-sample estimations and out-of-sample evaluations, the mismatch leads to suboptimal model selection and predictions (Engle 1993, Granger 1993). We note some exceptions: Using profit-based loss functions, Blattberg and George (1992) model customers' price sensitivity, Bult (1993) and Bult and Wittink (1996) estimate responses to mail, and Glady, et al. (2009) model temporal changes in usage. With conjoint analyses, Toubia and Hauser (2007) and Gilbride, et al. (2008) identify managerially relevant loss functions.

Bayesian statistics also highlight the importance of selecting managerially relevant priors (Montgomery and Rossi 1999). Data science and machine learning advances reiterate this importance (Chintagunta, et al. 2016), such that firms such as Amazon.com seek to include managerially relevant loss functions in their data acquisition strategies (Saar-Tsechansky and Provost 2007). Surprisingly, this focus has been missing in the retention literature. Our proposal to define a profit-based loss function to estimate the profit lift of a retention intervention addresses this gap. While conceptually straightforward, this approach requires significant changes in the model and estimation as indicated below.

3. Defining the Profit of Proactive Retention Actions

Imagine a proactive retention campaign that targets customers with a predefined retention incentive.[1] The firm's decision to target customer *i* is denoted T_i, so $T_i = 1$ indicates targeting, and $T_i = 0$ indicates no targeting. For every targeted customer *i*, the firm generates a *profit lift* π_i that represents the net impact of the intervention for this customer. In the potential outcome framework for causal inference (Rubin 2005), it corresponds to the conditional average treatment effect (CATE) of the retention program.[2] Following Neslin, et al. (2006), the profit of the campaign is the sum of the profit lift of all targeted customers,

$$\Pi = \sum_i^N \pi_i \, T_i, \tag{1}$$

where *N* is the total number of customers. In practice, we do not observe CATE because we do not know what the behavior of a targeted customer would have been if she was not targeted, nor the behavior of a non-targeted customer if she was targeted. Instead, we only observe one of both outcomes. One solution to estimate CATE is to run a randomized control trial, in which the intervention is randomized across a representative sample of customers. By observing the

[1] We assume a constant, exogenously determined retention offer and optimize the customer target for a specific intervention. In the last section, we briefly discuss how to generalize our approach for multiple offers.
[2] Some companies call this construct Delta CLV, suggesting a comparison of the customer lifetime value (CLV) of a customer if targeted versus not targeted. We prefer the term profit lift, to acknowledge that we incorporate the cost of the intervention.

behavior of customers in both treatment and control groups, we can estimate the causal impact of the campaign at the customer level (Rosenbaum 2017).[3]

The expected profit lift of a retention action given the intervention cost δ is

$$E(\pi_i|\delta) = E(CLV_i - \delta|X_i, T_i = 1) - E(CLV_i|X_i, T_i = 0), \quad (2)$$

where $E(CLV_i - \delta|X_i, T_i = 1)$ is the net residual[4] customer lifetime value (CLV) of customer *i* if targeted with an offer that costs δ, and $E(CLV_i|X_i, T_i = 0)$ is the (net) residual CLV if customer *i* is not targeted (Provost and Fawcett 2013). If a customer is targeted, her net residual CLV is the discounted value of the cash flows after the campaign minus the per customer cost of the retention intervention (Fader and Hardie 2010). When a customer is not targeted, the cost of the intervention is not incurred.

We consider the residual CLV for periods subsequent to the intervention (taking place at the beginning of period *t* = 1) in the general case where future retention probabilities and cash flows vary over time and given an infinite time horizon. Let $r_{it}^{(1)}$ and $r_{it}^{(0)}$ denote the retention probabilities of customer *i* in the period *t* following the intervention, depending on whether this customer is targeted (1) or not (0).[5] Likewise, let $m_{it}^{(1)}$ and $m_{it}^{(0)}$ denote the cash flows generated by customer *i* in the period *t* following the intervention if targeted or not, conditional on customer *i* being alive. In addition, *d* is the discount rate for post-campaign cash flows. We distinguish between two types of retention incentives. Unconditional incentives, such as thank you presents, can be sent to customers without their prior consent and without any conditions. Alternatively, conditional incentives (e.g., discounts, gifts) can be provided to customers only if they agree to purchase or renew their subscription. In many contractual

[3] We take the viewpoint of the firm and define the treatment as the firm sending a retention incentive to a customer (as proposed by Ascarza 2018). Thus, the treatment (TE) and the intent-to-treat (ITT) effects coincide.
[4] By using the residual CLV, we ignore transactions and costs (including acquisition cost) that precede the campaign as they are irrelevant to the current campaign.
[5] The retention probability $r_{it} = \prod_{k=1}^{t} \tilde{r}_{ik}$ with $\tilde{r}_{ik}$ the retention probability going from period *k*-1 to period *k*. For instance, the retention probability two periods after intervention equals to the product of the retention probability in period 1 (right after intervention) and the retention probability in period 2.

settings, customers who are up for renewal receive a discount if they renew their contract. Depending on whether the retention incentive is conditional or not, we rewrite Equation (2) as

$$E(\pi_i|\delta) = \left(\sum_{t=1}^{\infty} \frac{r_{it}^{(1)} m_{it}^{(1)} - r_{it}^{(0)} m_{it}^{(0)}}{(1+d)^t}\right) - \frac{r_{i1}^{(1)} \delta}{(1+d)} \quad (3)$$

for unconditional incentives, and

$$E(\pi_i|\delta) = \left(\sum_{t=1}^{\infty} \frac{r_{it}^{(1)} m_{it}^{(1)} - r_{it}^{(0)} m_{it}^{(0)}}{(1+d)^t}\right) - \frac{\delta}{(1+d)} \quad (4)$$

for conditional ones.[6] The difference between Equations (3) and (4) reflects that only customers who accept the offer prompt the cost of the conditional incentive. Also note that the overhead (fixed) costs of the retention campaign are not taken into account in the profit lift since they do not affect the customer ranking. Finally, note that the churn lift definition provided by Ascarza (2018), given by $r_{i1}^{(1)} - r_{i1}^{(0)}$ is a special case of Equations (3) and (4) when the intervention cost and cash flows are ignored.

Theoretically, the net residual CLV should be estimated over an infinite time horizon, but, for practical purposes, most companies and academics focus on a specific time period and use a truncated CLV (Glady, et al. 2015). In our empirical application, we estimate the impact of the intervention on the next period, as detailed in Section 7 (and further drop the time subscript in what follows). Indeed, estimating the *causal* effect of a retention intervention over an infinite, or even long, period of time is impractical because the company would need to ensure that no confounder influences the outcome of interest during this period. In practice, it is unlikely that the unconfoundedness assumption required for causal inference would not be violated (Rosenbaum and Rubin 1984).[7]

[6] We assume that the cost of the incentive is incurred in the same period as the first post-intervention cash flow is received, and thus discount its cost by one period.

[7] Our discussions with several customer retention managers affirmed that the main barrier to using A/B testing is that they do not want to isolate groups of customers for a long time. Practical constraints make it impossible to exclude the risk of contamination by post-treatment marketing interventions or external factors that have nothing to do with the retention campaign but that endanger the comparability of the treatment and control groups. A multi-period horizon thus would be practically intractable, as is also the case for the data sets in our empirical application.

The expected profit lift can take any positive to negative value. It will be positive if the residual CLV, conditional on targeting, is larger than the combination of residual CLV in the absence of targeting and the cost of the retention intervention. For example, customers might intend to churn but change their mind after receiving the retention incentive or those who did not intend to churn might increase their spending in response to the incentive, because of "delight" (Blattberg, et al. 2008). The expected profit lift instead is negative if the combination of the residual CLV in the absence of targeting and the cost of the intervention is greater than the residual CLV, conditional on targeting. Such counterproductive outcomes may occur if retention offers wake the "sleeping dogs" by reminding them of their dissatisfaction with the firm's service, thereby increasing their probability to churn (Ascarza, et al. 2016).

4. Developing a Profit-Based Loss Function

In this section, we describe the classic loss function used in the domain of retention management. Based on the definition of profit lift proposed in the previous section, we then propose a new profit-based loss function. The profit-based loss function can be used with any estimation technique, including logistic regression (via likelihoods) and more advanced machine learning methods. For this study, we chose to rely on stochastic gradient boosting (SGB), a greedy algorithm based on gradient descent (Friedman 2001, 2002) because it supports flexible specifications of the loss function and provides powerful optimization.

4.1. Classic 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 binary churn outcome, and x be a set of covariates for *N* customers. Let *F* be the function that maps *x* to *y*. The SGB estimation method we describe in the next section (or another binary prediction model, such as logistic regression) provides fitted values of $F(x_i)$ for every customer *i* based on the values of the x variables. When the fitted values are between 0 and 1, as with a logistic regression, they are called (churn) probabilities. When they are not restricted to this

interval, as in the SGB method, they are called scores. Scores can be mapped to probabilities using various transformations[8] (Greene 2003). For proactive retention programs, these estimated probabilities (or scores) represent the main input to rank-order customers, so companies can target the customers with the highest scores.

In a logistic regression, the estimation of probabilities relies on maximum likelihood, which aims to maximize the sum over the individual (weighted) log-likelihoods:

$$logL_i = w_i \left(\tilde{y}_i \log p(x_i) + (1 - \tilde{y}_i) \log \left(1 - p(x_i)\right) \right), \quad (5)$$

where $\tilde{y}_i = 1$ when customer *i* is a churner or 0 when she is a non-churner (Hastie, et al. 2009). Most churn models assume a constant weight ($w_i = 1$) for all customers, so the cost of misclassification is the same for all of them. Some models add a customer-specific weight $w_i \neq 1$ (depending on the model used, weights must sum to one or not), which leads to a weighted estimator (Cosslett 1993, Manski and Lerman 1977). Weighted estimators can impose different costs on type I (false positives) and type II (false negatives) errors. This option is also available for imbalanced data (Lemmens and Croux 2006), such that different weights would be assigned to churners and non-churners to account for the skewness of the *y* distribution.

Instead of maximizing a likelihood function, machine learning algorithms minimize a loss function. Most likelihoods have exact loss counterparts. Estimating a model with the log-likelihood in Equation (5) is the same as minimizing the binomial loss function,

$$\Psi_i = w_i log\left(1 + e^{-2y_i F(x_i)}\right), \quad (6)$$

where $y_i = 1$ for a churner and -1 for a non-churner, so $\tilde{y}_i = (y_i + 1)/2$ (for the proof, see Web Appendix A; Hastie, et al. 2009). A loss function is defined by three components: its margin (here, $y_i F(x_i)$), its functional form, and, possibly, the weighing structure w_i. First, the margin defines the variable to predict (here, *y*) and qualifies the accuracy of a prediction of the

[8] The estimated scores $\widehat{F}_i$ between $]-\infty, +\infty[$ can be transformed into defection probability estimates $\widehat{p}_i$ between [0,1] (e.g., when computing CLV) using the logistic (inverted-logit) formula, $\hat{p}_i = \frac{1}{1+\exp\left(-2\hat{F}(x_i)\right)}$.

outcome of interest. In our preceding example, the margin $y_i F(x_i)$ captures the extent to which $F(x_i)$ is a good predictor of y_i. The more negative the margin becomes (i.e., *y* and *F* of opposite signs), the larger the prediction error is. The goal is to estimate a positive score *F* for $y_i = 1$ and a negative score *F* for $y_i = -1$.

Second, the functional form defines the loss assigned to a given observation, according to the estimation/prediction error associated with this observation. It indicates the predictions that need improvement. In our example, the loss is a monotone decreasing function of the margin $y_i F(x_i)$. The loss associated with a negative margin (i.e., higher error) is greater than that associated with a positive margin (smaller error). It only depends on *y* via the margin (i.e., in combination with *F*), so this loss cannot distinguish false positives ($y_i = -1$ and $F(x_i) > 0$) and false negatives ($y_i = 1$ and $F(x_i) < 0$) and instead penalizes both equally.

Third, the weight w_i determines an additional penalty assigned to a prediction error, specific to a given individual, similar to the weighted estimator we described previously. It can depend on *y* (penalize type I and type II errors differently) or other variables (e.g., cash flows generated by a customer). In most applications, each individual is weighted equally.

Although statistically sound, the likelihood and the loss function in Equations (5) and (6) do not align with the objectives of retention programs to maximize profits. They depend solely on whether a customer is well-classified as a churner or not, rather than on her profit lift.

4.2. Profit-Based Loss Function

In contrast with a classic loss function, a profit-based loss function seeks to ensure that the firm targets customers with a positive profit lift and does not target customers with a negative profit lift. To achieve these goals, we adapt the loss function in several ways.

First, we adapt the margin by replacing y_i with $E(\pi_i)$, to represent $E(\pi_i|\delta)$, as defined in Equation (2), which reflects the new outcome of interest. The new margin ensures that customers with a higher profit lift earn a higher score $F(x_i)$ than customers with a lower profit

lift. The customer ranking based on these scores depends on the profit that a decision to target each of them would generate. The profit-based loss function thus becomes

$$\Psi_i = w_i log\left(1 + e^{-2E(\pi_i)\,F(x_i)}\right). \quad (7)$$

Second, we weight the loss attached to each customer as a function of their expected profit lift to specify which prediction errors have the largest (positive or negative) impact on profit and thus should be penalized more. We empirically test three weighting schemes: (i) *symmetric weighting*, where $w_i = |E(\pi_i)|$ for all customers; (ii) *right weighting*, such that $w_i = |E(\pi_i)|$ for $E(\pi_i) \geq 1$ and $w_i = 1$ otherwise; and (iii) *left weighting*, where $w_i = |E(\pi_i)|$ for $E(\pi_i) \leq 1$ and $w_i = 1$ otherwise. Symmetric weighting ensures that predictions of the profit lift will be the most accurate for customers with the most extreme (positive or negative) profit lift values. Both ignoring a customer who would have contributed greatly to campaign success (i.e., $E(\pi_i)$ is much greater than 0) and mistakenly targeting a customer who reacts very negatively to the campaign (i.e., $E(\pi_i)$ is much smaller than 0) would have detrimental impacts in Equation (1), so symmetric weighting penalizes both equally. In contrast, right weighting focuses exclusively on customers with the most positive expected profit lifts, while left weighting focuses on customers with the greatest losses only. These latter weighting schemes mimic the notion of penalizing type I versus type II errors in classification settings.[9] We treat the choice of the weighting scheme as an empirical question. The next subsection provides useful insight into the relative performance of these weighting schemes depending on the data characteristics.

4.3. Monte Carlo Simulations and Statistical Properties

The profit-based loss function belongs to the category of weighted estimators with endogenous weights (Solon, et al. 2015). These estimators are known to have different statistical properties than unweighted estimators. In Web Appendix B, we report the results of

[9] We thank the review team for this suggestion.

two Monte Carlo simulations that study the statistical properties of the profit-based loss estimator. The simulations show why and when a weighted estimator outperforms an unweighted one. We find that an estimator that uses the profit-based loss function has the same statistical properties as the weighted estimator for endogenously stratified samples (Cosslett 1993, Donkers, et al. 2003, King and Zeng 2001a, 2001b, Manski and Lerman 1977). On average, these estimators are less efficient than estimators that use an unweighted loss, because weighing observations dilutes the information by assigning low weights to some observations.

However, the simulations reveal the mechanism by which a weighted estimator can offer more profitable campaigns than an unweighted one: The weighted estimator offers less bias and is more efficient than the unweighted estimator at the *individual* level for observations that receive a greater weight. In fact, only the predictions for the observations that receive a small weight exhibit a greater bias and lower efficiency. This cross-customer reallocation mechanism ensures that the weighted estimator fits the behavior of customers who have the greatest impact on the profit of the retention campaign better than the unweighted estimator does. We also find that this mechanism is stronger when the accuracy of statistical models is poorer (*smaller signal-to-noise ratio*), for *smaller sample sizes* and in presence of *less extreme weights*. Noting that retention models reputably have low predictive power (see the recent review by Ascarza, et al. 2018) and that field experiments are usually based on small treatment (and control) groups, we expect a substantial effect of weighting on the profitability of retention campaigns. Finally, the simulations show that the relative performance of the various weighting schemes (symmetric, right and left weighting) likely depends on the expected profit lift distribution. In general, it is more beneficial to put greater weight on the under-represented part of the distribution (e.g., use right weighting if the share of positive profit lifts is small).

5. Integrated Profit-Based Approach

We integrate the profit-based loss function into the overall design of retention programs

in three stages, as depicted in Figure 1: (1) Estimate the expected profit lift of a retention intervention, (2) optimize the targets of the retention campaign, and (3) evaluate the targeting decisions. Central to our approach, each stage uses a different sample of customers, which we refer to as the calibration sample, validation sample, and test sample, respectively. The calibration sample is used only for model estimation. We use the validation sample to determine the target size. The number of customers to target is chosen to maximize a holdout profit measure. As we further explain below, determining the target size based on a holdout profit measure allows us to account for the fact that our model might over- or under-estimate the profit lift of customers on the calibration sample. Finally, the test sample contains customers who have not been used for estimation or for determining target size, so that we obtain a true holdout evaluation of campaign performance. To generate these three samples, we randomly split the data into three equal sets. To ensure the results are generalizable and not driven by any specific random split, we generated 100 different splits (Ascarza 2018), for both empirical applications. With this bootstrapping procedure, we also can test whether holdout performance is statistically superior to that of benchmark approaches.

Insert Figure 1 about here

5.1. Estimation Stage

We first estimate the heterogeneous treatment effect of the intervention on churn probabilities and cash flows separately, allowing for the possibility that the campaign can affect both processes differently. This step reflects the most recent benchmark in the literature (Ascarza 2018, Guelman, et al. 2012, Hitsch and Misra 2018). Once we obtain an estimate of the components of the expected profit lift, we plug them into Equations (3) and (4) and we use SGB with the profit-based loss function defined in Equation (7). This step allows us to penalize customers according to their respective impact on campaign profitability.

5.1.1. Lift in Churn and Lift in Cash Flows. To estimate the heterogeneous treatment effect of an intervention, we use data from randomized experiments, in which some share of customers gets assigned to the treatment condition and targeted with a retention offer, while others are assigned to the control condition and do not receive a retention offer (in practice, this would mean running a pilot on a small sample of randomly chosen individuals). This approach is central to the potential outcome framework (Rubin 1974). Customers randomly allocated to two conditions should be, on average, similar in both observed and unobserved covariates across conditions, so we use these data to estimate the impact of the intervention. We observe churn in the period after intervention denoted by $y_i^{(1)}$ if the customer is in the treatment group and $y_i^{(0)}$ if the customer is in the control group. In addition, we observe customers' cash flows in the period following the company intervention: $m_i^{(1)}$ if the customer is in the treatment group and $m_i^{(0)}$ if in the control group. Finally, we observe customer-specific covariates and the cost of the retention offer (see Section 7). With this information, we estimate the customer-specific effect of the intervention on retention probabilities and cash flows.

Although estimating an average treatment effect of an intervention is straightforward (it only requires comparing the average outcome in the treatment and control groups), the estimation of the heterogeneous treatment effects is more complex as it requires comparing the outcomes for matched individuals. Machine learning for causal inference, and uplift models in particular, offers a solution to this problem by matching pairs of customers in the treatment and control groups on the basis of their available covariates, and then comparing their respective churn and cash flow outcomes (Athey and Imbens 2016). Various uplift models can be used, depending on the nature of the dependent variable. For the binary retention model, we follow Ascarza (2018) and estimate the lift in retention probabilities $\hat{r}_i^{(0)}$ and $\hat{r}_i^{(1)}$ in the period following intervention using uplift random forests (Guelman, et al. 2015). To estimate the lift

in cash flows $\hat{m}_i^{(0)}$ and $\hat{m}_i^{(1)}$ we use the uplift *k*-nearest neighbors (*k*NN) for continuous outcomes (Alemi, et al. 2009, Su, et al. 2012).[10]

5.1.2. Profit Lift Estimation by SGB with a Profit-Based Loss Function. We combine all estimates obtained in the previous step to calculate the expected profit lift according to Equations (3) or (4). Then, we estimate the profit lift scores with SGB using the profit-based loss function defined in Equation (7). This weighted loss function allows us to penalize customers according to their respective impact on campaign profitability. Any (machine learning) estimation method could be used with the profit-based loss function, but we choose SGB because of its superior predictive performance for churn prediction (it won the Teradata Churn modeling tournament; Lemmens and Croux 2006, Neslin, et al. 2006) and other analyses (Hastie, et al. 2009). Moreover, it uses a flexible optimization algorithm based on gradient descent, so it can be used with any loss function. As a greedy numerical optimization algorithm (Friedman, et al. 2000, Friedman 2002), SGB sequentially combines predictions by simple models, typically regression trees (Breiman, et al. 1983), then makes initial guesses about each customer's outcome. It tries to predict residual errors by fitting a tree. At each iteration, a new tree is estimated to fit the residuals of the previous iteration. The estimation runs until no improvement occurs. We provide a description of regression trees in Web Appendix C.

Before the estimation, a loss function Ψ is chosen and used at each iteration b to compute the difference or error between the fitted scores $F_b(x_i)$, and the actual values to be predicted (in our case, the expected profit lifts). Once the loss function is defined, the estimation starts by setting each observation to an initial value, denoted by $\hat{F}_0(x_i)$, which can take any value in $(-\infty, \infty)$. From this initial guess, we compute the error (i.e., difference between the

[10] Uplift *k*NN computes the Euclidean distance between every pair of observations using all observed characteristics. Next, it selects *k* (here, *k* = 1) observation(s) in the calibration sample that is/are the closest to each observation in the validation (and test) sample and that received the treatment; $\hat{m}_i^{(1)}$ is the (average) cash flow for this/these *k* nearest neighbor(s). The same thing is done for the control sample. Both uplift random forests and k-nearest neighbors are implemented in the uplift R package (Guelman 2014).

fitted values $\hat{F}_0(x_i)$ and actual values). The next step fits a tree model $T(x_i, \Theta_0)$ of the *errors* against the predictors x and computes the fitted values of these errors.[11] The number of terminal nodes is relatively small (maximum 8 nodes) to avoid overfitting. These fitted errors are then combined with the predicted values $\hat{F}_0(x_i)$. The combination produces "boosted" fitted values (i.e., the original guess is *boosted* by the fitted errors), denoted $\hat{F}_1(x_i)$. This process repeats to compute the error from the boosted fitted values (difference between the fitted values $\hat{F}_1(x_i)$ and actual values), fit a tree model of the new errors, and combine the fitted values of these new errors to $\hat{F}_1(x_i)$. We repeat these steps *B* times until the model converges. Web Appendix D provides the estimation details.

5.2. Optimization Stage: Target Size Determination

The second stage determines how many and which customers to target to maximize the profit of the retention campaign. We use the validation sample for this purpose. The first step is to rank customers. Using the model estimated on the calibration sample, we predict the (holdout) profit lift scores for customers in the validation sample knowing their covariates values. We then rank them in order of decreasing scores, such that $\hat{F}(x_1) \geq \cdots \geq \hat{F}(x_i) \geq \cdots \geq \hat{F}(x_N)$. The final step is to determine the campaign size *S*, that is, the number of customers to target, starting at the top of the ranking. Because customers can have a negative profit lift, the optimal campaign size is usually smaller than 100%.

Two common approaches to determine target size include selecting the top decile (Lemmens and Croux 2006, Schweidel and Knox 2013), or applying a budget constraint (Datta, et al. 2015), which we present in Section 6.2. Instead, we optimize the campaign size using full enumeration search, combined with offline evaluation. In particular, we calculate the holdout profit of a campaign of size *S* going from 1 to *N* (total number of customers in the validation

[11] Fitting the errors gradually forces the model to predict the residual variance in the dependent variable that was unexplained in the prior iteration. Thus the estimation progressively concentrates on customers whose behaviour is difficult to predict.

sample) and identify the target size that maximizes the holdout campaign profit.

Calculating the holdout profit of a campaign of any target size *S* for a given customer ranking (i.e., the predictions of a model) is not straightforward, because we do not observe the actual profit lift of customers (i.e., we cannot observe the same unit at the same time in both the treatment and control groups). Offline policy evaluation provides a solution to this problem (Li, et al. 2012). This evaluation strategy is common with randomized experiments (Ascarza 2018, Hitsch and Misra 2018). It is "offline" in the sense that it is not necessary to effectively target the customers identified by a given policy. Instead, analysts can leverage the random treatment allocation to test the performance of any policy. For each target size *S*, we estimate the impact of the campaign in the period following the intervention, according to the per customer profit lift π_S it generates. That is,

$$\pi_S = \frac{1}{N_t}\sum_{i \in \text{Treatment}} \left(m_i^{(1)}\, \mathrm{I}\left(y_i^{(1)}=-1\right)-\delta\right) - \frac{1}{N_c}\sum_{j \in \text{Control}} m_j^{(0)}\, \mathrm{I}\left(y_j^{(0)}=-1\right), \text{ or} \qquad (8a)$$

$$\pi_S = \frac{1}{N_t}\sum_{i \in \text{Treatment}} m_i^{(1)}\, \mathrm{I}\left(y_i^{(1)}=-1\right) - \frac{1}{N_c}\sum_{j \in \text{Control}} m_j^{(0)}\, \mathrm{I}\left(y_j^{(0)}=-1\right) - \delta, \qquad (8b)$$

depending on whether the offer is conditional (8a) or unconditional (8b), where N_t stands for the number of customers in the top *S* who actually received the retention incentive during the randomized experiment, and N_c is the number of customers in the top *S* who did not receive it. The first part indicates average post-campaign net cash flows (less the action cost) of customers in a target of size *S* that were effectively treated. The second part denotes the average post-campaign cash flows of customers in a target of size *S* that were not treated. The difference captures the *actual* per customer impact of the intervention on customers who belong to a target of size *S*. Note that π_S is an unbiased estimate of the actual profit of the targeting decisions conditional on *S* (Hitsch and Misra 2018). To obtain the total (holdout) profit of the designed retention campaign of size *S*, denoted Π_S, we multiply π_S by the number of customers targeted,

$$\Pi_S = S\pi_S. \qquad (9)$$

Once we know Π_S for every target size *S*, we select the target size *S** that yields the highest

holdout profit on the validation sample.[12]

5.3. Evaluation Phase: Holdout Profit of the Retention Campaign

Recall that *S** is determined with the validation sample, so strictly-speaking the value of Π_{S*} on the validation sample is an in-sample measure of the campaign profit. We therefore use a third sample (test sample) to evaluate the holdout profit for a campaign of size *S**. The holdout profit of a campaign of target size *S** equals Π_{S*}, calculated using Equation (9) for *S* = *S** on the *test* sample.

6. Benchmark Models

We compare our approach against several benchmarks, including alternative estimation methods to rank-order customers and alternative approaches to determining target size.

6.1. Benchmark Estimation Methods to Rank Customers

6.1.1. Classic Loss. We estimate a churn model with SGB and the loss function defined in Equation (6) and rank customers on the basis of their estimated churn risk (Lemmens and Croux 2006).

6.1.2. Reordered Classic Loss. We reorder the classic loss scores (obtained from 6.1.1.) by accounting for the profit that each customer is expected to generate if targeted with a retention action. Therefore, we predict the retention probabilities $\hat{r}_i^{(0)}$ and $\hat{r}_i^{(1)}$ using SGB with a classic loss function by setting the treatment dummy to 0 or 1, then integrating the estimates of post-campaign cash flows $\hat{m}_i^{(0)}$ and $\hat{m}_i^{(1)}$ (estimated with *k*-nearest neighbors; see Section 5.1.1), and finally plugging them into the profit lift formulas in Equations (3) or (4). This

[12] An alternative to using offline policy evaluation would be to let *S** equal the number of customers in the validation sample whose profit lifts are predicted positive by the estimation algorithm (here, SGB). However, such an approach would be sensitive to the scale of the predicted scores. It would overestimate (underestimate) the target size if the model overestimates (underestimates) the number of customers with a positive profit lift. Instead, offline evaluation is insensitive to the scale of the predicted scores. The scores only serve to determine the ranking of customers, whereas the optimal target size is determined based on the *holdout* campaign profit. This approach is particularly useful when the predicted scores are not scaled (units have no meaning) as is the case with most machine learning methods.

method is a so-called indirect (two-step) estimation approach (Hitsch and Misra 2018). This approach, or some version of it, is typically used by a vast majority of practitioners and scholars who are aware that ranking solely based on churn is flawed and that cash flows should also be taken into account.

6.1.3. Uplift Models. We estimate the lift in churn (Ascarza 2018) and the lift in cash flows as described in Section 5.1.1., then combine them using Equations (3) or (4). This approach does not incorporate a profit-based loss function and corresponds to the first step of our estimation procedure.

6.2. Benchmark Methods to Determine the Optimal Target Size

6.2.1. Fixed Target Size. Companies often select target sizes by relying on managerial judgments or actual churn rates in their industry (i.e., higher churn rate prompts a larger target size). For this comparison, we define the target size according to the *churn rate* in our validation sample. Alternatively, we could determine the target size based on the available *budget* (Datta, et al. 2015). Given the action cost, we calculate the number of customers that can be targeted with a specific budget. For illustration, we use a budget of 1,000 Euros or dollars.

6.2.2. Optimized Target Size Using Aggregate Metrics. Verbeke, et al. (2012) calculate the optimal target size by combining information about the proportion of churners in the target, together with the average profit of targeting a customer. In their study, the probability of response to incentives and customer value used to calculate profit are hypothetical and assumed to be constant across all individuals. We extend this targeting rule by calculating the average treatment effect from the randomized controlled trial. It offers a crude approximation of the optimal target size selection we propose.

6.2.3. Buffer after Optimization. Finally, we consider the possibility that our target size optimization might be too restrictive and use an alternative target size that reflects a 10% buffer,

such that it is larger than the optimized target size.[13]

7. Empirical Applications

We test our approach on two different customer databases from two different industries. The first data corresponds to an interactive television subscription service, provided by a firm located in continental Europe and used by Datta, et al. (2015).[14] The second data refers to a subscription-based membership organization located in North America.

7.1. Interactive Television Subscription (Europe)

A major digital television provider in continental Europe offers access to local and international digital channels and video-on-demand (VOD) services. Customers pay for subscription that includes unlimited usage of the basic iTV service (prices vary from approx. €20 to €100 per month depending on the type of service). In addition, customers can buy various additional packages (e.g., sports), for which they pay a higher monthly fee that varies across packages. Finally, customers can use the VOD rental service, for which they are charged on a pay-per-use basis, with an average price of €3 per VOD rental. To increase market penetration, the company offers new customers a free trial period of three months. About 40% of customers who use this service during the free trial period do not renew the service.

To decrease this high "churn rate", the company used an intervention between August 2006 and July 2007. The retention offer was conditional on renewal, and its cost was about €12 per targeted customer, in line with practices in other industries.[15] Not every subscriber ends the free trial period at the same time, so the intervention spanned nine waves, and during each wave, some proportion of free trial customers were targeted before the end of their trial period (treatment group), while others were not targeted (control group). We cannot identify the

[13] We thank a reviewer for this suggestion.
[14] We are extremely grateful to the authors of this paper for sharing their data with us.
[15] Discussions with managers responsible for proactive retention programs confirm similar numbers. For example, a North European telecom firm cited an average cost of approximately 7 euros per customer.

decision rules used to split the customers, so we used propensity score matching on the samples and performed a randomization check before and after the matching to ensure the final treatment and control groups are comparable[16] (see Web Appendix E for details).

For each customer in the matched sample (2,595 treated and 2,595 not treated), we observe the month in which the retention offer was sent, whether the customer renewed the subscription after the intervention,[17] the cash flows before and after the intervention, and other demographic and usage data (e.g., number of months the individual is a customer of the company, customer gender, age, language, household size, income based on zip code, installation method).

7.2. Special Interest Membership Organization (North America)

This special interest membership organization offers an annual membership for the right to use its services and receive discounts to attend events. The annual fee is approximately $180. Each year, the organization sends out renewal letters to customers one month before their membership expires. The company ran a field experiment for five consecutive months that tested whether adding a thank you gift to the renewal communication increased renewal rates. Each month, the company identified customers who were up for renewal and split them (randomly and evenly) between a treatment group that received a gift with the letter and a control group that received only the renewal letter. The per customer cost of the retention gift was about $12. In total, 2,100 customers were involved in the experiment, and 1,044 of them were targeted. A randomization check confirms that randomization was done properly (Web Appendix E). This data set includes information on the month in which the renewal letter was sent, whether the customer renewed for the next year, and demographic and usage

[16] The management team has changed since 2007, so we cannot specify the decision rules used previously. However, considering the vast customer data and the flexibility of our matching algorithm, we believe matching can capture them relatively accurately. The post-matching randomization check confirms that the matched treatment and control groups do not differ.

[17] We use two operationalizations to measure churn and cash flow after intervention: (1) one month after the intervention, and (2) three months after the intervention. In the empirical section, we present the results for one month but using three months does not affect our conclusions.

characteristics such as tenure (years), location (state where the member lives), whether the subscriber attended any organized or special interest event, and whether the subscriber had logged in to the organization's website.

7.3. Results

In the following sections, we compare the performance of a retention campaign with a profit-based loss function against the benchmarks from Section 6. We also explore the mechanisms that lead to the improved performance of our approach.

7.3.1. Financial Impact of Retention Campaign. In each bootstrap iteration, we apply the integrated profit-based approach described in Section 5, and calculate the corresponding holdout profit using Equations (8) and (9). Table 1 contains the average holdout profits over all bootstrapped samples for the classic loss function, reordered classic loss function, uplift model, and profit-based loss function (our approach). For our approach, we report the results for the weighting scheme that gives the highest performance. For Study 1, all three weighting schemes give similar results (€4,967 for symmetric weighting, €4,945 for right weighting and €5,026 for left weighting). For Study 2, right weighting ($1,328) significantly outperforms symmetric weighting ($413) and left weighting (minus $55).[18] We also report the bootstrapped differences between our approach and alternative ones, as well as the *p*-values computed from the bootstrapped standard errors.

Insert Table 1 about here

First, our approach leads to a more profitable retention campaign than all benchmark models. In both applications, the differences are highly significant. Note that this is the case for all three weighting schemes. In all three cases and both studies, our approach outperforms all other benchmarks. For example, in Study 1, the campaign profit for our approach is 168% higher than for the classic loss, 300% higher than for the reordered classic loss, and 23% better

[18] See Section 7.3.4 for more details on the relative performance of the various weighting schemes.

than for the uplift method (which includes both uplift in churn and cash flow but does not use a profit-based loss function). In Study 2, our approach is the only one that provides a positive campaign profit; all others lead to losses.

As expected, the classic loss function produces the lowest profit, because it is the only estimation method that optimizes a non-profit–related criterion. As we show subsequently (Table 3), this method performs better when the assessment criterion reflects the optimization criterion (i.e., predicting churn rather than profit lift). By itself, this result confirms our main premise: Firms need to align their estimation method, and in particular their loss function, with their evaluation criterion, and both should fit their managerial objectives.

Second, reordering the scores to take the profit lift into account does not improve performance. Results for the classic loss and reordered classic loss are very similar to each other (as further illustrated by Figure 2 in which both curves follow very similar patterns). In general, the two-step reordering approach that minimizes the churn misclassification rate across all customers in the first step, and then reorders customers according to their profit lift, performs significantly worse than our approach that incorporates customers' profit in the first step. This result corroborates findings by Hitsch and Misra (2018, page 2), who note that "methods that are trained to directly predict the incremental effect of targeting yield larger profits than conventional methods that indirectly predict the incremental effect based on the conditional expectation function that is trained on the outcome level." For example, predicting churn (*outcome level*) for treated and control customers and then calculating the expected lift (*incremental effect*) is an indirect approach. Indirect approaches underperform direct approaches because the estimation uses the wrong metric. Reordering the ranking does not compensate for the loss function's goal of minimizing errors in churn instead of profit lift.

Third, our approach significantly outperforms the uplift model that does not have profit-based loss function in estimation. The advantage of our approach over the uplift approach is

that it *directly* estimates the profit lift (the profit lift is the dependent variable) and penalizes prediction errors for customers with the largest impact on campaign profits. As such, it aligns with firms' managerial objectives of maximizing campaign profit.

Fourth, we note that in Study 2, all approaches except our approach yield a negative total impact. The retention intervention was very ineffective in the first place and had a negative net impact on profits earned from many customers. Despite this condition, the profit-based loss function can still identify a target size for which the total profit is positive.

7.3.2. Impact of Campaign Profits on Firm Revenue and Profit. Based on Table 1, we can assess the impact of our approach on the increase in firm profits from a proactive retention campaign. In Study 1, our approach generates a per customer profit of €4.63 (€5,026/1,085 targeted customers). Given that the average annual revenue per customer in this data set is about €588, the profit earned from a proactive retention campaign would contribute about 1% to the firm's annual revenue. This is substantial, considering that this profit results from a single campaign, captures its effect over a single period ahead, and its impact on firm profit (not revenue) would be even higher (e.g., in 2018, operating profit for Comcast in the U.S. was about 25% of revenue, which implies that a single retention campaign for the firm in our data set could increase its profits by about 4%). In Study 2, our approach generates a per customer profit of $4.99 ($1,328/266 targeted customers). Given the annual subscription fee of $180, this represents an increase of about 3% in annual revenue from a single campaign. Note, none of the other approaches were able to achieve positive campaign profits. In summary, our approach has the potential to enhance a firm's future profits.

7.3.3. Profit as a Function of Campaign Size. Before we compare various target size optimization strategies, we explore how the profit of a campaign varies with its size by calculating a holdout cumulative profit for target sizes from 1% to 100% for each of the bootstrapped samples. Figure 2 reveals the average profit over all bootstrap samples for the

four estimation methods.[19] The profit with a 100% target size, when all customers receive a retention offer (i.e., it is the same across methods), indicates a positive impact in Study 1 but a negative one in Study 2. In the latter case, targeting all customers is not profitable.

Insert Figure 2 about here

This analysis confirms the superior performance of our approach and also reveals, at least in part, why it works well. For both applications, customers who generate positive profits are ranked first. For Study 1, the cumulative profit curve keeps increasing until it reaches its peak, at around 60% of the sample. Its slope is positive and larger than the slope of other methods, so our approach keeps adding profitable customers to the target, whereas other methods add less profitable or non-profitable ones. In Study 2, the overall poor impact of the intervention leads to a slightly different figure but similar conclusions. Our approach is the only one to achieve a positive campaign profit with small target sizes, because customers who contribute to the profit of the campaign are included first. The uplift model is the second best alternative, but it fails to rank high-profit customers first and thus requires a much larger target size—including a fair share of negative profit lift customers—before it reaches its maximum value (which is close to zero and far inferior to our approach). Finally, the profit curve for the classic loss function provides a good visualization of the problem of focusing on churn. Most customers who contribute to campaign profit are ranked low. For Study 1, the slope of the classic loss is largest from 90% to 100%; for Study 2, it is only positive from 70% to 80%. A different way of looking at it consists of decomposing, for each decile, the average treatment effect into the actual profit of a customer in the treatment group and the actual profit of a customer in the control group (see Web Appendix F)

[19] Note that the profit figures in Table 1 do not directly correspond to Figure 2. The former figures are obtained by determining the optimal target size per bootstrap iteration such that we obtain 100 holdout campaign profit measures. This approach allowed us to test whether two approaches perform significantly differently from each other. In contrast, Figure 2 does not fix the target size but averages the profit curves obtained at each iteration. The latter is useful to see how the profit evolves with the campaign size.

7.3.4. The Role of Weighting. To complement these results and gain further insights into the role of the weighting scheme on the performance of our approach, we explore the relative performance of symmetric, right and left weighting for both applications. Figure 3 reveals their average profit over all bootstrap samples as a function of campaign size. Results are in line with those from the Monte Carlo simulation (Web Appendix B). The relative performance of the three weighting schemes depends on the distribution of the expected profit lift, and confirms that it is more beneficial to put greater weight on the under-represented part of the distribution. For Study 1, the share of customers with a negative expected profit lift (as inputted in the profit-based loss function) is slightly inferior to the proportion of customers with positive expected profit lifts (68% positive, 32% negative). As a result, left weighting slightly outperforms the other schemes (see Section 7.3.1). In contrast, customers with a positive expected profit lift (as inputted in the profit-based loss function) in Study 2 are largely under-represented (17%), so right weighting is far more profitable than the other schemes. Figure 3 illustrates the function of the weighting scheme with regard to the performance of the prediction algorithm.

Insert Figure 3 about here

7.3.5. Determining Target Size. In Table 2, we compare the holdout profit of our target size optimization approach (Section 5.2) to the benchmarks (Section 6.2). For more details on the determination of the target size, we refer readers to Web Appendix G.

Insert Table 2 about here

As expected, fixed target sizes determined prior to the estimation lead to significantly lower profits than our optimization strategy. Targeting as many customers as the number of expected churners or using a fixed budget is not a good strategy. Fixing the target size based on the churn rate offers 17% and 77% less profit than the optimized target size in Study 1 and Study 2. The fixed budget constraint leads to 92% and 77% less profit than the optimized target size in the two studies. In addition, our optimization strategy is superior to Verbeke et al.'s

(2012) optimization approach, which determines the target size on the basis of aggregate metrics that do not reflect customer heterogeneity in profit lift. Their approach yields 93% and 84% less profit than the optimized target size. Finally, in contrast with common practice, it is not preferable to add a buffer of customers to the retention campaign to ensure the target includes "good" ones. Doing so actually decreases the profit of the campaign by 3% and 47% for Study 1 and Study 2 respectively, because it adds non-profitable customers.

7.3.6. Overlap of Customer Rankings. To understand the differences among various estimation methods, we investigate the extent to which the target identified by our approach overlaps with the targets of other estimation methods (see Ascarza 2018 for a similar approach). We rank-order customers according to the scores obtained by the various methods and split the four rankings into 10 deciles (the first decile corresponds to priority customers for targeting). For each decile, we then calculate the percentage overlap in customers targeted across methods. Figure 4 reveals the percentage overlap between our approach and all three alternatives. For instance, it indicates that almost 30% of the customers in the first decile provided by our (profit-based loss) approach also belong to the first decile provided by the uplift model (line with +). A value of 100% would mean that both groups perfectly overlap (i.e., the two approaches are identical in identifying profitable customers), whereas the 45-degree line represents a situation where the overlap between groups is purely due to chance. Figure 4 shows that the level of customer overlap between our approach (profit loss) and the classic loss or reordered classic loss functions is close to random. In other words, our approach ranks customers very differently than these alternative rankings because they rely on different criteria. The greatest customer overlap is between our approach and the uplift model, but even this overlap is limited.

Insert Figure 4 about here

The higher profits obtained by our approach reflects contributions from customers who do not overlap, because this method identifies more profitable customers and excludes

customers who have detrimental effects on total profits.

7.3.7. Drivers of churn vs. drivers of profit lift. The differences between the profit-based and the classic loss functions translate into discrepancies in the factors explaining churn vs. those that explain profit lift. We compute the relative variables' importance for both approaches (Friedman 2001). In Study 1, household income contributes to almost 50% of the "performance" of the profit loss solution compared to only 11% for the classic loss solution, after the treatment dummy (19%) and one of the sport package dummies (17%). In Study 2, customers' geographical location (48% for the profit loss vs. 25% for the classic loss) and tenure (28% for the profit loss vs. 58% for the classic loss) are the main drivers of profit lift and churn, but in different proportions.

7.3.8. Model aligned with managerial objective. The additional profits earned with the profit-based loss function do not imply that it should be used in all circumstances. Table 3 reports the holdout gini coefficient and top decile lifts for both studies, averaged across all bootstrap samples. The top decile lift measures the accuracy of the model in predicting churn among the top 10% riskiest customers. In turn, the gini coefficient provides a measure of model accuracy in predicting churn for the entire customer base. The higher the top decile lift and gini coefficient, the better the model predicts churn (see Lemmens and Croux (2006) for detailed definitions of these metrics).

Insert Table 3 about here

The (reordered) classic loss function provides the most accurate predictions of customer churn behavior, because it is the only approach that seeks to minimize errors in churn prediction without considering other dimensions of customer profit. This shows that, if the goal is to accurately predict churn, the classic loss function is the best, but if the goal is to maximize retention campaign profit, our approach is more suitable.

7.3.9. Replication for other estimation methods. Finally, we replicate our results for a

different estimation approach. Namely, we compare the performance of a logistic regression (classic loss) and a weighted regression (profit loss) with the weights defined in equations (3)-(4). The profit loss offers significantly ($p < .01$) larger holdout profits than the classic loss in both studies. In Study 1, €3,331 for the classic loss and €4,907 for the profit loss; in Study 2, minus $1,112 for the classic loss and $416 for the profit loss. This replication confirms that the loss function's choice drives the improvement in profits regardless of the estimation approach.

8. Conclusion, Limitations, and Further Research

We propose a method to optimize the profit of proactive retention campaigns. Our approach defines the profit lift of a retention intervention according to the potential outcome framework for causal inference (Rubin 2005). We demonstrate the benefits of using a profit-based loss function in estimating the financial impact of a targeted marketing intervention. Our findings highlight the need for marketing academics and practitioners to pay attention to the choice of loss function, a feature that is often neglected in model estimation processes. In particular, this choice should match managers' objectives.

Our approach potentially fits many contexts, within and outside marketing, where organizations seek to target a set of individuals with a specific intervention (e.g., catalog, mail, charitable giving, and personalized promotions). Estimating heterogeneous treatment effects is an exciting topic, featured in studies across economics and econometrics (Imbens and Rubin 2015), management (Godinho de Matos, et al. 2017), and computer science (Pearl and Mackenzie 2018). For each application, it is critical to carefully determine the appropriate loss function. When building their own "goal-oriented" loss functions, decision makers should (1) ensure that the margin specifies the true outcome of interest (i.e., goal of the intervention) and (2) use a weighting scheme that prioritizes customers who have the largest impact on the success of the intervention. This is relevant even in non-profit contexts, such as for predicting patient compliance with medical treatments. In this case, the loss function could incorporate

patient-specific health risks and benefits associated with complying with the medical treatment.

Our results also show that the optimization of the target size of a retention campaign has a significant impact on profits. Retention literature is surprisingly silent on this topic; it mostly focuses on ranking customers. We find that selecting a target size that maximizes the campaign profit leads to significantly more profitable campaigns than using the common rules. We thus hope managers attend to not only the estimation method used to rank customers but also the number of customers to target.

Several limitations of this paper offer fruitful research opportunities. First, our approach can rank-order customers according to the profit lift they produce, in response to a specific retention campaign. Both field experiments reflect a single, specific retention incentive, in line with recent attempts to estimate heterogeneous treatment effect models for customer relationship management (Ascarza 2018, Hitsch and Misra 2018, Provost and Fawcett 2013). An interesting further challenge would be to explore variations in customer responses depending on the type and depth of retention interventions, then determine the costs at which each response is maximized (Venkatesan, et al. 2007). Firms might estimate consumers' profit lift distributions for various costs by testing various retention incentives, then use these estimates to determine the optimal intervention per customer.

Second, our approach does not consider the long-run impact of retention interventions. Assumptions of unconfoundness make it difficult and impractical to estimate the profit lift of a single intervention over a long period of time. Managers are often unwilling to isolate a group of customers from any marketing intervention for a long time period. Concerns about legal customer privacy protections, which mandate that companies may only keep customer data for the shortest amount of time possible, also complicate experiments that run for long periods.[20]

[20] https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/how-long-can-data-be-kept-and-it-necessary-update-it_en

Third, we do not model whether a customer's sensitivity to an intervention depends on expectations of retention offers in the future (Lewis 2005a). In digital, connected economies, customers are more aware of the attractive discounts that others receive when they indicate an intention to churn. This phenomenon of strategic churning is an interesting area for research and could be captured using dynamic structural models (Khan, et al. 2009).

We hope our work will foster more research in the area of predicting the individual treatment effects and remind the reader of the importance of aligning the loss function used for model estimation with the managerial objectives of the campaign.

9. References

Alemi F, Erdman H, Griva I, Evans CH (2009) Improved statistical methods are needed to advance personalized medicine. *Open Translational Medicine Journal*. 1:16-20.

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Science*. 32(4):570-590.

Ascarza E, Iyengar R, Schleicher M (2016) The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *J Marketing Res*. 53(1):46-60.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *J Marketing Res*. 55(1):80-98.

Ascarza E, Neslin SA, Netzer O, Anderson Z, Fader PS, Gupta S, Hardie BGS, Lemmens A, Libai B, Neal D, Provost F, Schrift R (2018) In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*. 5(1-2):65-81.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*. 113(27):7353-7360.

Blattberg RC, George EI (1992) Estimation under profit-driven loss functions. *J Bus Econ Stat*. 10(4):437-444.

Blattberg RC, Kim B-D, Neslin SA (2008) *Database marketing: Analyzing and managing customers* (Springer, New York).

Bobbier T (2013) Keeping the Customer Satisfied: The Dynamics of Customer Defection, and the Changing Role of the Loss Adjuster. *CILA Report*.

Bolton RN (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.

Borle S, Singh SS, Jain DC (2008) Customer lifetime value measurement. *Management science*. 54(1):100-112.

Braff A, Passmore WJ, Simpson M (2003) Going the distance with telecom customers. *McKinsey Quarterly*. (4):82-93.

Braun M, Schweidel DA (2011) Modeling customer lifetimes with multiple causes of churn. *Marketing Science*. 30(5):881-902.

Breiman L, Friedman J, Olshen R, Stone C (1983) *Classification and regression trees* (Wadsworth Publishing).

Bult JR (1993) Semiparametric versus parametric classification models: An application to direct marketing. *J Marketing Res*. 30(3):380-390.
Bult JR, Wittink DR (1996) Estimating and validating asymmetric heterogeneous loss functions applied to health care fund raising. *International Journal of Research in Marketing*. 13(3):215-226.
Chintagunta P, Hanssens DM, Hauser JR (2016) Editorial— Marketing science and big data. *Marketing Science*. 35(3):341-342.
Christoffersen P, Jacobs K (2004) The importance of the loss function in option valuation. *J Financ Econ*. 72(2):291-318.
Cosslett SR (1993) Estimation from Endogenously Stratified Samples. Maddala GS, Rao CR and Vinod HD eds. *Handbook of Statistics* (Elsevier, North Holland), 1-43.
Datta H, Foubert B, Van Heerde HJ (2015) The challenge of retaining customers acquired with free trials. *J Marketing Res*. 52(2):217-234.
Donkers B, Franses PH, Verhoef PC (2003) Selective sampling for binary choice models. *J Marketing Res*. 40(4):492-497.
Donkers B, Verhoef PC, de Jong MG (2007) Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics*. 5(2):163-190.
Engle RF (1993) On the limitations of comṕaring mean square forecast errors: Comment. *J Forecasting*. 12(8):642-644.
Fader PS, Hardie BGS (2010) Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. *Marketing Science*. 29(1):85-93.
Fader PS, Hardie BGS, Shang J (2010) Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*. 29(6):1086-1108.
Forbes (2011) Bringing 20/20 foresight to marketing: CMOs seek a clearer picture of the customer. *Forbes Insights*.1-13.
Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics*. 28(2):337-407.
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 29(5):1189-1232.
Friedman JH (2002) Stochastic gradient boosting. *Comput Stat Data An*. 38(4):367-378.
Ganesh J, Arnold MJ, Reynolds KE (2000) Understanding the customer base of service providers: an examination of the differences between switchers and stayers. *J Marketing*. 64(3):65-87.
Gilbride TJ, Lenk PJ, Brazell JD (2008) Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Science*. 27(6):995-1011.
Glady N, Baesens B, Croux C (2009) Modeling churn using customer lifetime value. *Eur J Oper Res*. 197(1):402-411.
Glady N, Lemmens A, Croux C (2015) Unveiling the relationship between the transaction timing, spending and dropout behavior of customers. *International Journal of Research in Marketing*. 32(1):78-93.
Godinho de Matos M, Ferreira P, Smith MD (2017) The effect of subscription video-on-demand on piracy: Evidence from a household-level randomized experiment. *Management Science*. 64(12):5610-5630.
Godinho de Matos M, Ferreira P, Belo R (2018) Target the ego or target the group: Evidence from a randomized experiment in proactive churn management. *Marketing Science*. 37(5):793-811.
Granger CWJ (1969) Prediction with a generalized cost of error function. *Journal of the Operational Research Society*. 20(2):199-207.
Granger CWJ (1993) On the limitations of comparing mean square forecast errors: Comment. *J Forecasting*. 12(8):651-652.

Greene WH (2003) *Econometric analysis*, 6 ed. (Prentice Hall).
Guelman L (2014) *uplift: Uplift Modeling*. R package version 0.3.5.
Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. *Modeling and simulation in engineering, economics, and management* (Springer), 123-133.
Guelman L, Guillén M, Pérez-Marín AM (2015) Uplift random forests. *Cybernet Syst*. 46(3-
Hastie T, Tibshirani R, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2 ed. (Springer, New York).
Hitsch GJ, Misra S (2018) Heterogeneous treatment effects and optimal targeting policy evaluation, SSRN.
Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. *Expert Syst Appl*. 39(1):1414-1425.
Imbens GW, Rubin DB (2015) *Causal inference for statistics, social, and biomedical sciences: An introduction* (Cambridge University Press).
Khan R, Lewis M, Singh V (2009) Dynamic customer management and the value of one-to-one marketing. *Marketing Science*. 28(6):1063-1079.
King G, Zeng L (2001a) Explaining rare events in international relations. *Int Organ*. 55(3):693-715.
King G, Zeng L (2001b) Logistic regression in rare events data. *Political analysis*. 9(2):137-163.
Knox G, Van Oest R (2014) Customer complaints and recovery effectiveness: A customer base approach. *J Marketing*. 78(5):42-57.
Kumar V, Venkatesan R, Bohling T, Beckmann D (2008) Practice Prize Report—The power of CLV: Managing customer lifetime value at IBM. *Marketing science*. 27(4):585-599.
Larivière B, Van den Poel D (2005) Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst Appl*. 29(2):472-484.
Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. *J Marketing Res*. 43(2):276-286.
Lewis M (2005a) Incorporating strategic consumer behavior into customer valuation. *J Marketing*. 69(4):230-238.
Lewis M (2005b) Research note: A dynamic programming approach to customer relationship pricing. *Management science*. 51(6):986-994.
Li L, Chu W, Langford J, Moon T, Wang X (2012) An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, 19-36.
Manski CF, Lerman SR (1977) The estimation of choice probabilities from choice based samples. *Econometrica*.1977-1988.
Montgomery AL, Rossi PE (1999) Estimating price elasticities with theory-based priors. *J Marketing Res*. 36(4):413-423.
Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *J Marketing Res*. 43(2):204-211.
Neslin SA, Novak TP, Baker KR, Hoffman DL (2009) An optimal contact model for maximizing online panel response rates. *Management Science*. 55(5):727-737.
Pearl J, Mackenzie D (2018) *The book of why: the new science of cause and effect* (Basic Books, New York).
Provost F, Fawcett T (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking* (O'Reilly Media, Inc.).

Reinartz W, Thomas JS, Kumar V (2005) Balancing acquisition and retention resources to maximize customer profitability. *J Marketing*. 69(1):63-79.
Risselada H, Verhoef PC, Bijmolt THA (2010) Staying power of churn prediction models. *Journal of Interactive Marketing*. 24(3):198-208.
Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*. 79(387):516-524.
Rosenbaum PR (2017) *Observation and experiment: an introduction to causal inference* (Harvard University Press).
Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Science*. 15(4):321-340.
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*. 66(5):688-701.
Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*. 100(469):322-331.
Saar-Tsechansky M, Provost F (2007) Decision-centric active learning of binary-outcome models. *Inform Syst Res*. 18(1):4-22.
Schweidel DA, Fader PS, Bradlow ET (2008a) A bivariate timing model of customer acquisition and retention. *Marketing Science*. 27(5):829-843.
Schweidel DA, Fader PS, Bradlow ET (2008b) Understanding service retention within and across cohorts using limited information. *J Marketing*. 72(1):82-94.
Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. *Management Science*. 57(3):471-486.
Schweidel DA, Knox G (2013) Incorporating direct marketing activity into latent attrition models. *Marketing Science*. 32(3):471-487.
Singh SS, Borle S, Jain DC (2009) A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. *Quantitative Marketing and Economics*. 7(2):181-205.
Solon G, Haider SJ, Wooldridge JM (2015) What are we weighting for? *Journal of Human resources*. 50(2):301-316.
Su X, Kang J, Fan J, Levine RA, Yan X (2012) Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*. 13(10):2955-2994.
Toubia O, Hauser JR (2007) Research note—On managerially efficient experimental designs. *Marketing Science*. 26(6):851-858.
Venkatesan R, Kumar V (2004) A customer lifetime value framework for customer selection and resource allocation strategy. *J Marketing*. 68(4):106-125.
Venkatesan R, Kumar V, Bohling T (2007) Optimal customer relationship management using Bayesian decision theory: An application for customer selection. *J Marketing Res*. 44(4):579-594.
Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *Eur J Oper Res*. 218(1):211-229.
Winer RS (2001) A framework for customer relationship management. *Calif Manage Rev*. 43(4):89-105.
Wübben M, Von Wangenheim F (2008) Instant customer base analysis: Managerial heuristics often “get it right”. *J Marketing*. 72(3):82-93.

Table 1. Average campaign holdout profit for different estimation methods (and bootstrapped differences with profit-based loss)

	Study 1: Interactive Television Subscription		Study 2: Special Interest Membership Organization	
Estimation Method	**Holdout Profit (in Euro)**	**Difference (*p-value*)**	**Holdout Profit (in US$)**	**Difference (*p-value*)**
Classic Loss	€ 1,872.23	3,154 *(.000)*	$ (1,669.37)	2,997 *(.000)*
Reordered Classic Loss	€ 1,253.74	3,773 *(.000)*	$ (1,709.16)	3,037 *(.000)*
Uplift Model	€ 4,092.97	933 *(.000)*	$ (1,305.07)	2,633 *(.000)*
Our Approach	**€ 5,026.36**		**$ 1,327.76**	

Notes: The last row "our approach" refers to the results provided by the profit-based loss function. The "holdout profit" column reports the average holdout profit obtained across all bootstrapped iterations. The "difference" column reports the bootstrapped difference between the holdout profit given by our approach and each alternative approach, together with the *p*-values (in parentheses) obtained using the bootstrapped standard errors. All reported differences are significant at the 1% probability level.

Table 2. Average holdout campaign profit for different target size determination methods (and bootstrapped differences with profit-based loss)

	Study 1: Interactive Television Subscription		Study 2: Special Interest Membership Organization	
Target Size Determination	**Holdout Profit (in Euro)**	**Difference (*p-value*)**	**Holdout Profit (in US$)**	**Difference (*p-value*)**
Fixed Target Size:				
Based on Churn Rate	€ 4,164.06	862 *(.000)*	$ 308.56	1,019 *(.000)*
Based on Budget	€ 398.76	4,628 *(.000)*	$ 304.98	1,023 *(.000)*
Optimized Target Size:				
Based on Verbeke	€ 345.46	4,681 *(.000)*	$ 207.94	1,120 *(.000)*
10% Buffer	€ 4,882.36	144 *(.001)*	$ 700.28	627 *(.010)*
Our Approach	**€ 5,026.36**		**$ 1,327.76**	

Notes: The table provides the results of the profit-based loss function using the fixed and optimized target size selection approaches described in Section 6.2. The last row "our approach" refers to the proposed optimized target size selection using offline evaluation (Section 5.2.). The "holdout profit" column reports the average holdout profit obtained across all bootstrapped iterations. The "difference" column reports the bootstrapped difference between the holdout profit given by our target size optimization approach and each alternative approach, together with *p*-values (in parentheses) obtained using the bootstrapped standard errors. All reported differences are significant at the 1% probability level.

Table 3. Average holdout churn predictive accuracy for different estimation methods

	Study 1: Interactive Television Subscription		Study 2: Special Interest Membership Organization	
Estimation Method	**Gini Coefficient**	**Top Decile Lift**	**Gini Coefficient**	**Top Decile Lift**
Classic Loss	.277	2.014	.102	1.183
Reordered Classic Loss	.218	1.820	.110	1.215
Uplift Model	.150	1.435	-.027	.011
Our Approach	.142	1.291	-.016	.084

Notes: The table reports the average gini coefficients and top decile lifts obtained across all bootstrapped iterations. The last row "our approach" refers to the results provided by the profit-based loss function.

Figure 1: Profit-based analysis step by step

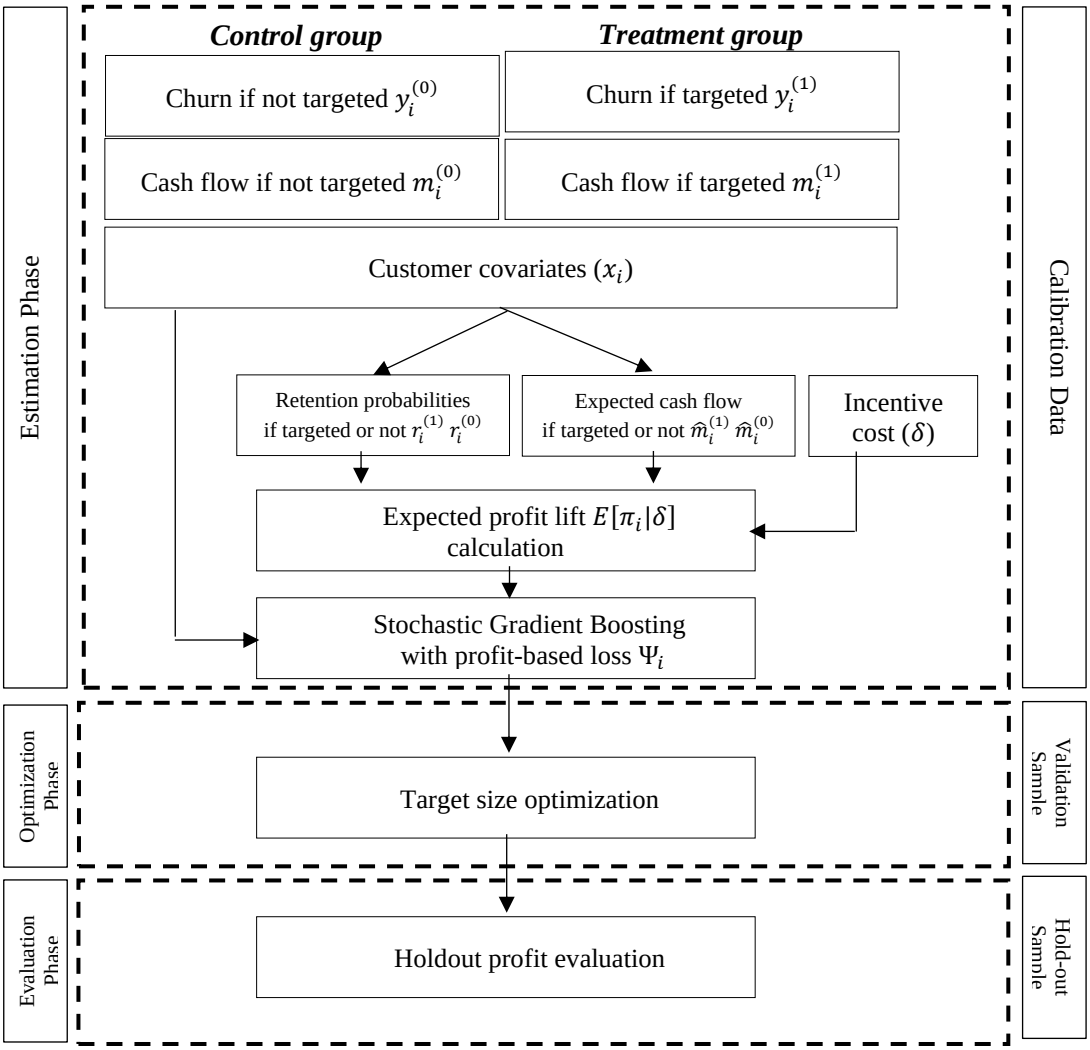
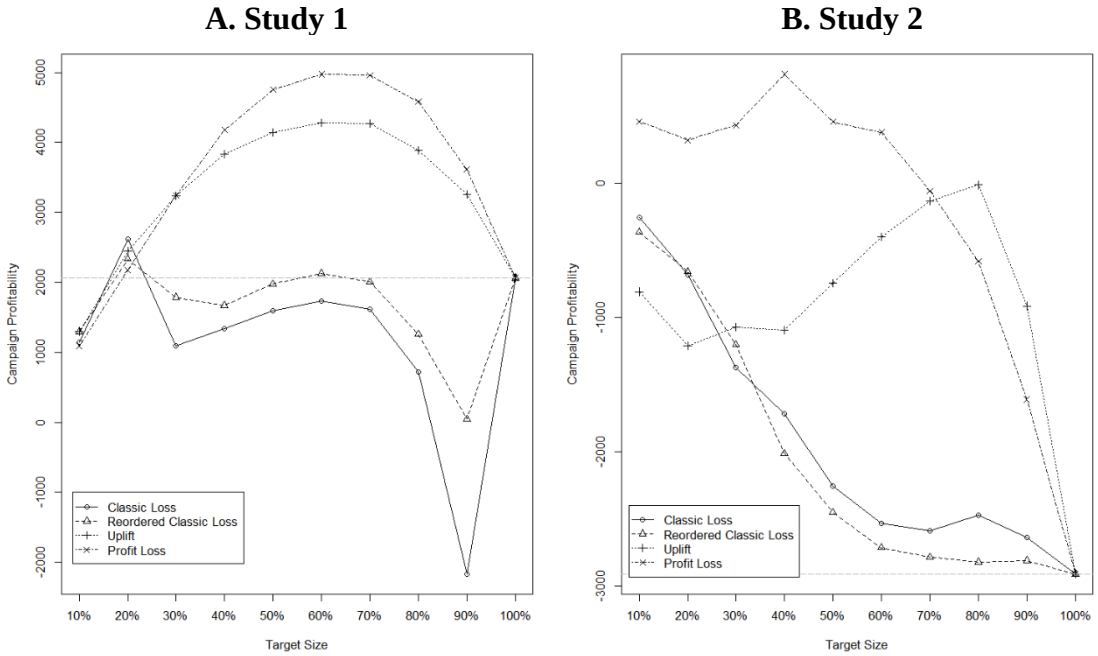
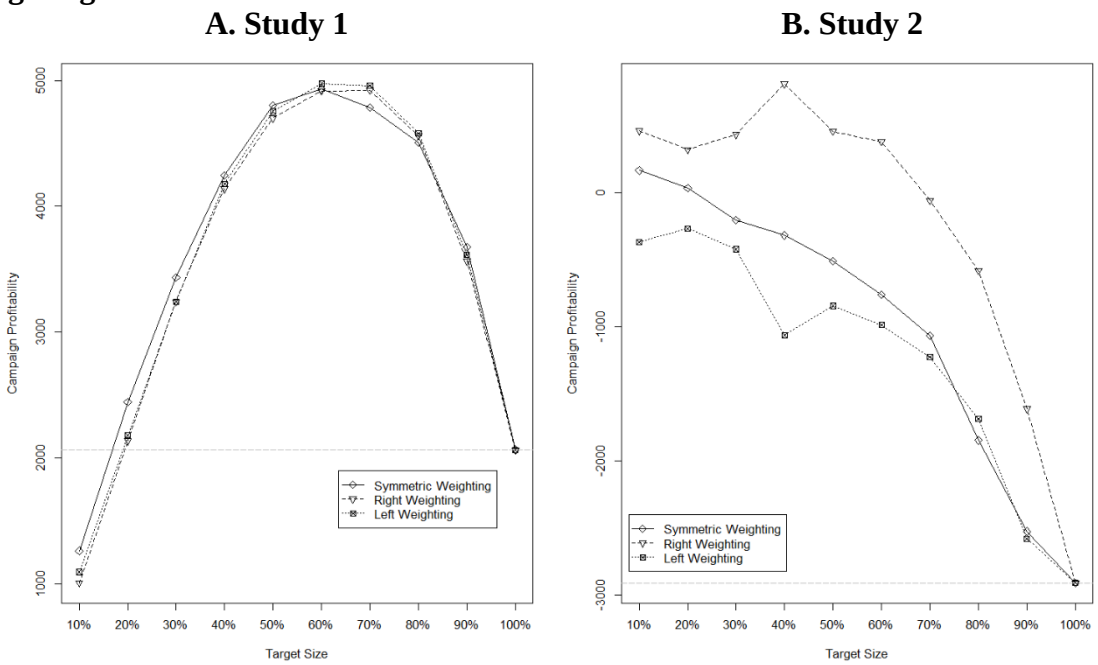


Figure 2. Average holdout campaign profit as a function of target size for different estimation methods



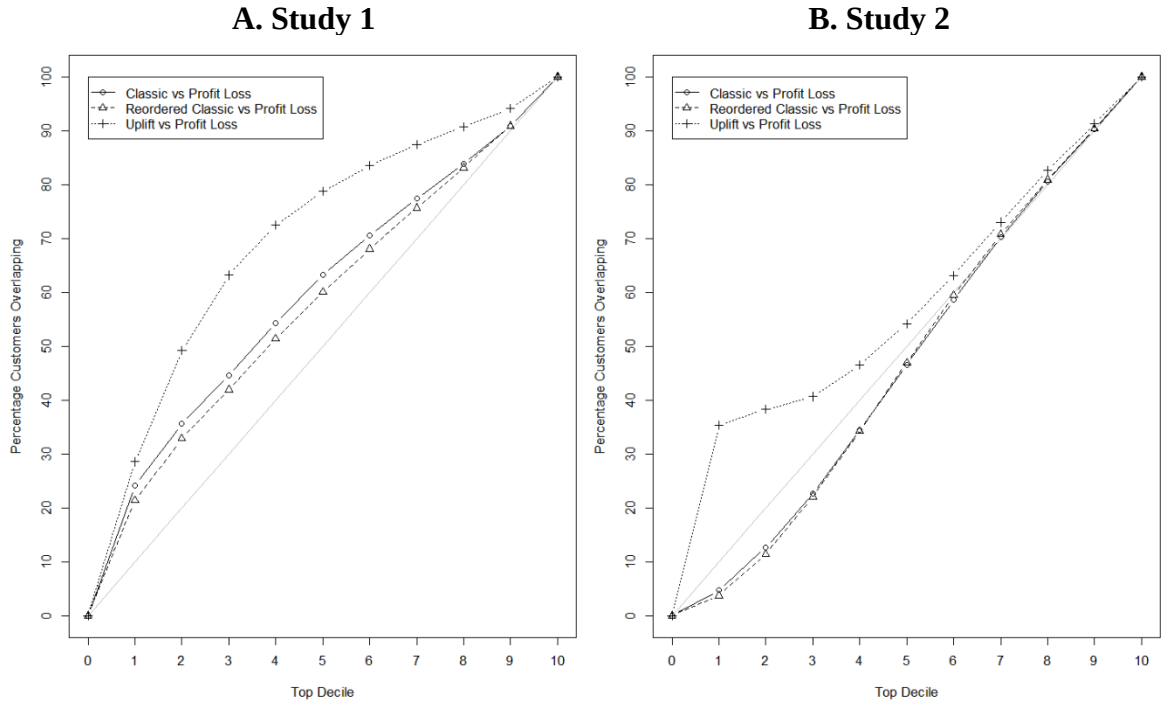
Notes: The curves represent the holdout profits of the campaign averaged over all bootstrap iterations. The horizontal grey dashed line represents a campaign targeting everyone.

Figure 3. Average holdout campaign profit as a function of target size for different weighting schemes



Notes: The curves represent the holdout profits of the campaign averaged over all bootstrap iterations. The horizontal grey dashed line represents a campaign targeting everyone.

Figure 4. Average percentage customers overlapping for different estimation methods



Notes: The 45% line corresponds to the level of overlap between two random rankings.

Web Appendix A. Equivalence of Log-Likelihood and Loss Function

Let $\tilde{y}_i = 1$ for a churner and $\tilde{y}_i = 0$ for a non-churner. The log-likelihood over all customers i=1,.., n can be written as

$$logL = \sum_{i=1}^{N} \{\tilde{y}_i \log p(x_i) + (1 - \tilde{y}_i) \log \big(1 - p(x_i)\big)\}. \quad \text{(A1)}$$

When converting the dependent variable $\tilde{y}_i$ to $y_i = 1$ for a churner and $y_i = -1$ for a non-churner using the transformation $\tilde{y}_i = (y_i + 1)/2$, we can rewrite the log-likelihood as

$$logL = -\sum_{i=1}^{N} \log\Big(1 + \exp\big(-2y_i F(x_i)\big)\Big). \quad \text{(A2)}$$

proving the relationship between equation (7) and equation (8) in the paper. The proof is as follows. Using y_i instead of $\tilde{y}_i$, replacing the probabilities $p(x_i)$ by the scores $F(x_i)$ and using the logistic (i.e. inverted-logit) formula in footnote 5, equation (A1) becomes

$$logL = \sum_{i=1}^{N} \left\{ \frac{(y_i + 1)}{2} \log\left(\frac{1}{1 + \exp\big(-2F(x_i)\big)} \right) + \frac{(1 - y_i)}{2} \log\left(1 - \left(\frac{1}{1 + \exp\big(-2F(x_i)\big)} \right) \right) \right\}. \quad \text{(A3)}$$

Using the proprieties of the log function, we can rewrite (A3) into

$$logL = \sum_{i=1}^{N} \left\{ -\frac{(y_i + 1)}{2} \log\big(1 + \exp\big(-2F(x_i)\big)\big) + \frac{(1 - y_i)}{2} \log\big(\exp\big(-2F(x_i)\big)\big) - \frac{(1 - y_i)}{2} \log\big(1 + \exp\big(-2F(x_i)\big)\big) \right\}. \quad \text{(A4)}$$

After factoring out the factor common to the first and third terms and simplifying the second term, we obtain

$$logL = \sum_{i=1}^{N} \{-\log\big(1 + \exp\big(-2F(x_i)\big)\big) + (y_i - 1)F(x_i)\} \quad \text{(A5)}$$

$$= -\sum_{i=1}^{N} \left\{ \log\big(1 + \exp\big(-2F(x_i)\big)\big) - \log\Big(\exp\big((y_i - 1)F(x_i)\big)\Big) \right\}. \quad \text{(A6)}$$

Using again the proprieties of the log and exponent functions, we can rewrite (A6) into

$$logL = -\sum_{i=1}^{N} \log\left(\frac{1 + \exp\big(-2F(x_i)\big)}{\exp\big((y_i - 1)F(x_i)\big)} \right) = -\sum_{i=1}^{N} \log\big[\big(1 + \exp\big(-2F(x_i)\big)\big)\exp\big(-(y_i - 1)F(x_i)\big)\big]. \quad \text{(A7)}$$

After distributing $\exp\big(-(y_i-1)F(x_i)\big)$ across the sum $\big(1+\exp\big(-2F(x_i)\big)\big)$, we obtain

$$logL = -\sum_{i=1}^{N} \log\big[\exp\big(-(y_i-1)F(x_i)\big) + \exp\big(-2F(x_i)-(y_i-1)F(x_i)\big)\big]$$
$$= -\sum_{i=1}^{N} \log\Big(\exp\big(-y_iF(x_i)+F(x_i)\big) + \exp\big(-F(x_i)-y_iF(x_i)\big)\Big). \quad \text{(A8)}$$

Using again the properties of the exponents function, we get

$$logL = -\sum_{i=1}^{N} \log\Big(\exp\big(-y_iF(x_i)\big)\exp\big(F(x_i)\big) + \exp\big(-F(x_i)\big)\exp\big(-y_iF(x_i)\big)\Big). \quad \text{(A9)}$$

Factoring out $\exp\big(-y_iF(x_i)\big)$, (A9) becomes

$$logL = -\sum_{i=1}^{N} \log\Big(\exp\big(-y_iF(x_i)\big)\Big(\exp\big(F(x_i)\big) + \exp\big(-F(x_i)\big)\Big)\Big). \quad \text{(A10)}$$

Bringing in the negative sign into the logarithm,

$$logL = \sum_{i=1}^{N} \log\left(\frac{\exp\big(y_iF(x_i)\big)}{\exp\big(F(x_i)\big)+\exp\big(-F(x_i)\big)}\right). \quad \text{(A11)}$$

Given that $y_i = 1$ or $y_i = -1$, $\exp\big(F(x_i)\big) + \exp\big(-F(x_i)\big)$ is equivalent to $\exp\big(y_iF(x_i)\big) + \exp\big(-y_iF(x_i)\big)$

$$logL = \sum_{i=1}^{N} \log\left(\frac{\exp\big(y_iF(x_i)\big)}{\exp\big(y_iF(x_i)\big)+\exp\big(-y_iF(x_i)\big)}\right). \quad \text{(A12)}$$

Given the corollary (Hastie, Tibshirani, and Friedman 2009, p. 346),

$$\frac{\exp(A)}{\exp(A)+\exp(-A)} = \frac{1}{1+\exp(-2A)} \quad \text{(A13)}$$

A12 is equivalent to

$$logL = \sum_{i=1}^{N} log\left(\frac{1}{1+exp\big(-2y_iF(x_i)\big)}\right)$$
$$= -\sum_{i=1}^{N} \log\Big(1+\exp\big(-2y_iF(x_i)\big)\Big). \quad \text{(A14)}$$

Web Appendix B. Monte Carlo Simulation: Weighted vs. Unweighted SGB

In this Monte Carlo simulation study, we study the relative prediction bias and efficiency of SGB when using a weighted loss function vs. an unweighted loss function. Let z_i be the dependent variable of interest (e.g. profit lift) and $F(x_i)$ be the estimated scores given a set of covariates x_i for customer *i*. The (un)weighted loss function is defined as

$$\Psi_i = w_i log\left(1 + e^{-2z_i F(x_i)}\right) \tag{B1}$$

with $w_i = 1$ for all customers *i* for the unweighted estimator. As explained in Section 4.2, we consider three weighting schemes for the weighted estimator:

(1) Symmetric weighting: $w_i = |z_i|$ for all customers *i*,
(2) Right weighting: $w_i = |z_i|$ for $z_i \geq 1$ and $w_i = 1$ otherwise,
(3) Left weighting: $w_i = |z_i|$ for $z_i \leq 1$ and $w_i = 1$ otherwise.

Following Carsey and Harden (2013), we simulate the data using the following data generating process:

$$z_i = \beta_{0i} + \beta_{1i} x_i + \varepsilon_i \tag{B2}$$

with x_i drawn from a random uniform $U(-1,1)$. We assume heterogeneous parameters $\beta_{0i} = \beta_0 + u_{0i}$ and $\beta_{1i} = \beta_1 + u_{1i}$ with $\beta_0 = 2$ and $\beta_1 = 5$, and u_{0i}, u_{1i} respectively drawn from a random uniform $U(-\beta_0/4\,, \beta_0/4)$ and $U(-\beta_1/4\,, \beta_1/4)$. Finally, we add random noise $\varepsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = \theta^2 \text{var}(\beta_{0i} + \beta_{1i} x_i)$. We control the signal-to-noise ratio by varying θ (see below).

Simulation Study 1. Relative Prediction Bias

In the first simulation study, we set the sample size *n* = 5,000 observations and the signal-to-noise ratio to 2:1, i.e. $\theta = .5$. We generate 1,000 data sets using Equation (B2). For each weighting scheme $k = 1, 2 \text{ and } 3$ defined above, we calculate the relative prediction bias (RPB) of the weighted estimator compared to the unweighted estimator *u* for each observation *i* as

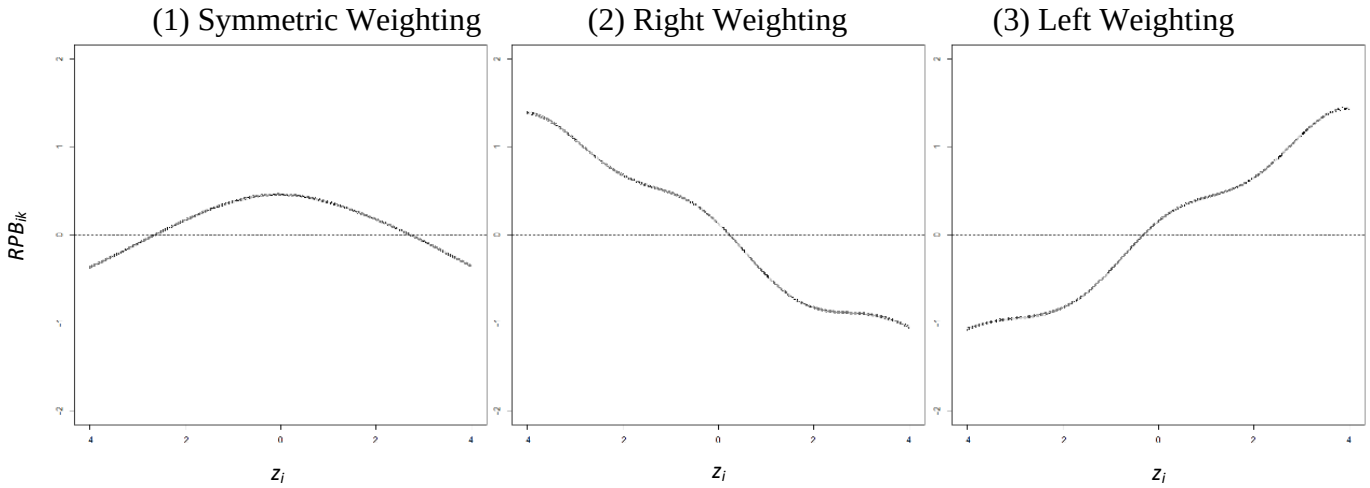
$$RPB_{ik} = \log \left| \frac{\hat{z}_{ik} - z_i}{\hat{z}_{iu} - z_i} \right|. \tag{B3}$$

Tofallis (2015) recommends the logarithm to limit the impact of extremely large values, which tend to occur when *z* is close to zero. A positive (negative) value indicates that the prediction for observation *i* obtained using weighing scheme *k* underperforms (outperforms) the prediction of the unweighted estimator. We investigate how this relative individual bias varies as a function of z_i by fitting a generalized additive model with smoothing splines. Figure B1 shows the fitted curves, together with the two standard error confidence intervals (not visible here because they are narrow) for all three weighting schemes.

The Monte Carlo simulation shows that the weighted estimators provide a smaller prediction bias than the unweighted estimator for the observations that get the highest weights. In contrast, observations with the lowest weights are predicted less accurately using a weighted estimator than when using an unweighted estimator. In particular, the symmetric weighting scheme offers

more accurate predictions on both extremes of the dependent variable distribution (*RPB* < 0). Instead, the right and left weighting offer better predictions only on one side of the dependent variable distribution, with the right weighting scheme performing best for the most positive values of *z* and the left weighting scheme for the most negative values of *z*. These figures illustrates the "bias reallocation mechanism" induced by weighting some observations more than others. On average, the weighted estimator does not predict better than the unweighted estimator. However, it does so locally for the most heavily weighted observations. Thus, introducing weights into the loss function provides a way to control where the observations should be predicted with the highest accuracy. This feature is useful in settings such as retention management campaigns where some customers have a largest impact than others on the performance of the campaign.

Figure B1. Relative Prediction Bias of the Weighted vs. Unweighted SGB Estimators



Simulation Study 2. Relative Efficiency

In the second simulation study, we investigate the relative efficiency of the weighted estimators vs. the unweighted estimator using SGB as in Study 1. In particular, we focus on how the relative efficiency is affected by the sample size, the share of positive vs. negative values of the dependent variable, the amount of noise in the data generating process as well as the concentration of the weights. To do so, we use the following design. We vary the sample size *n* from 1,000, 5,000, 10,000, 50,000 and 100,000 observations, the proportion of positive values of the dependent variable (*pp*) from 25%, 50% to 75%, and the signal-to-noise ratio from 2:1, 3:2, 1:1 and 1:2 (that is, $\theta = 1/2, 2/3, 1, 2$). Finally, we vary the presence of extreme values of *z* using different distributions of the error term. In addition to the Normal distribution, we use a Student's *t* distribution with one degree of freedom (*fatter tails, i.e. more extremes*) and a Truncated Normal distribution with truncation at the 1st and 3rd quartile of the equivalent Normal distribution (*shorter tails, i.e. less extremes*). These distributions differ in the share of observations that contribute to 80% of the sum of all positive *z* values: about 13% for the Student's *t*, 26% for the Normal, and 29% for the Truncated Normal,[21] while preserving an equal proportion of positives vs negatives (*pp* = 50%). We generate 1,000 data sets for each cell of the design.

[21] In our application, the number of customers that contribute to 80% of the sum of all positive profit lifts.

Given the results from the first simulation study, we can compare the efficiency of two biased estimators by looking at the ratio of their mean squared errors. The latter captures the trade-off between the squared bias of the estimator and its variance (Hastie, Tibshirani, and Friedman 2009). For each weighting scheme $k = 1, 2$ and 3 defined above, we calculate its mean squared prediction error over the *S* iterations,

$$MSE_k = \frac{1}{Sn}\sum_{s=1}^{S}\sum_{i=1}^{n}\left[\left(\hat{z}_{is}^{k} - z_{is}\right)^2\right], \tag{B4}$$

and define the relative efficiency of estimator *k* with respect to the unweighted estimator *u* as

$$RE_k = \frac{MSE_u}{MSE_k}. \tag{B5}$$

A value of RE_k smaller than one indicates that the unweighted estimator *u* is relatively more efficient that the weighted estimator *k*. In contrast, a value RE_k larger than one indicates that the weighted estimator is relatively more efficient. In addition, we also evaluate the efficiency of the estimators over two subsets of the data: (i) the positive values of *z* only, (ii) the negative values of *z* only. Given the results of the first simulation study, the right and left weighing schemes are expected to behave differently on each of the subsamples. Formally, we define

$$MSE_k^{+} = \frac{1}{Sn}\sum_{s=1}^{S}\sum_{i=1}^{n}\left[\left(\hat{z}_{is}^{k} - z_{is}\right)^2\right] \text{ for all } z_{is} \geq 0 \tag{B6}$$

$$MSE_k^{-} = \frac{1}{Sn}\sum_{s=1}^{S}\sum_{i=1}^{n}\left[\left(\hat{z}_{is}^{k} - z_{is}\right)^2\right] \text{ for all } z_{is} < 0, \tag{B7}$$

$$\text{and } RE_k^{+} = \frac{MSE_u^{+}}{MSE_k^{+}} \;\; and \;\; RE_k^{-} = \frac{MSE_u^{-}}{MSE_k^{-}}. \tag{B8}$$

Results are reported in Table B1. The left panel reports RE_k, RE_k^{+} and RE_k^{-} for varying sample sizes *n* for an equal proportion of positive and negative values of *z* (*pp* = 50%) and a signal-to-noise ratio of 2:1 ($\theta = .5$). The middle panel reports the same metrics for varying proportions of positive values of the dependent variable (*pp*) given a sample size *n* = 5,000 and a signal-to-noise ratio of 2:1 ($\theta = .5$). Finally, the right panel reports the same metrics for varying signal-to-noise ratios θ for a sample size *n* = 5,000 and an equal proportion of positive and negative values of *z* (*pp* = 50%). In bold, we depict the cases where the weighted estimator is more efficient than the unweighted estimator (*RE* larger than one).

Simulation results show that, on average, weighted estimators are less efficient than the unweighted estimator. Weighing dilutes the information leading to efficiency loss. The efficiency loss is most pronounced for small sample sizes. In contrast, increasing the sample size slightly compensates for the efficiency loss. The larger the sample size, the less impact any weighing scheme has on the estimation (Chambers 1996). Results are consistent for all three weighing schemes.

Positive vs. Negative Observations

More interesting results come from comparing the efficiencies for the positive and negative values of *z*. Table B1 shows that, even though weighted estimators are relatively inefficient on average, they actually are relatively more efficient than the unweighted estimator for the observations that receive the largest weights. For *n* = 1,000, the estimator based on right weighing is 25% more efficient than the unweighted estimator for the positive observations. Likewise, the estimator based on left weighing is 24% more efficient than the unweighted

estimator for the negative observations. This increase in efficiency comes at a cost of a lower efficiency for the observations that are weighted the least. Note that, for the symmetric weights, we do not find differences between the positive and negative values of *z* for the simple reason that it weights both equally. Instead, the efficiency for the largest values of *z* (both negative and positive) is larger than the efficiency for the values of *z* close to zero (detailed results are available upon request).

Sample Size
Importantly, the respective benefits of right and left weighing vary with the sample size. Table B1 shows that the respective benefits of right and left weighing are most prominent for small sample sizes. As mentioned above, the impact of weighing become smaller for large sample sizes. In practice, uplift models commonly rely on small samples as they require randomized control trials. We therefore expect a substantial impact of weighting in such contexts.

Share of Positive vs. Negative Observations
Weighting is also most beneficial when the share of the heavily weighted observations is small. In particular, the relative efficiency of the estimator based on right weighing (respectively, left weighing) is over 50% superior to the unweighted estimator for 25% positive (respectively, negative) observations for a sample size n = 5,000. Weighting acts in the same fashion as oversampling does, and works best when "balancing" the various parts of the expected profit lift distribution (Donkers et al. 2003, Solon et al. 2005). The smaller the set of customers reacting positively to a retention campaign, the more important to weight (i.e. as a way to resample) them more. Intuitively, the expected profitability of the campaign depends more critically on the share of customers with a positive profit lift when this share of customers is proportionally small.

Signal-to-Noise Ratio
The relative efficiency of the weighted estimator also depends on the amount of noise in the data. Our simulations show that the noisier the data, the larger the relative efficiency of weighing for the observations that are most heavily weighted. In real-life applications, the signal-to-noise ratios of retention and uplift models tend to be very small (Ascarza et al. 2018), which would imply large effects of weighting.

Presence of Extremes
The advantage of weighting gets smaller in presence of more extreme *z* values. Under the Student's *t* distribution, very few (about 13%) observations contribute to 80% of the sum of all positive *z* values (the same holds for the negative values). In such cases, few cases will receive a larger weight than the rest and have a disproportionally large influence on the estimator. Instead, under the Truncated Normal distribution, more (about 29%) observations generate 80% of the sum of all positive *z* values (the same holds for the negative values). Therefore, the estimator will be influenced by a larger set of cases. As a result, right (resp. left) weighting gives more efficient estimators across the range of *all* positive (resp. negative) values of *z* (not just the most extreme ones).

Conclusions
The simulations indicate that weighing leads to efficiency loss at the aggregate level, but at the same time, offers substantial efficiency gains for the most heavily weighted observations. The benefits of weighing are particularly large when the most heavily weighted observations are under-represented. The proportion of customers for whom one expects a positive vs. negative profit lift thus provides an indication as to which of the weighting schemes is the most

appropriate for a given application. Our empirical applications confirm this result (see Section 7.3). Finally, differences between weighted and unweighted estimators tend to disappear for very large sample sizes and/or very large signal-to-noise ratios.

Table B1. Relative Efficiency of the Weighted Estimators w.r.t the Unweighted Estimator, as a Function of the Sample Size *n*, Percentage of Positive Observations *pp* and Signal-to-Noise Ratio θ.

	Sample Size *n* (*pp* = 50% and $\theta = .5$)					% Positives *pp* (*n* = 5,000 and $\theta = .5$)			Noise-to-Signal Ratio θ (*n* = 5,000 and *pp* = 50%)				Presence of Extremes (*n* = 5,000, *pp* = 50% and $\theta = .5$)		
	1,000	5,000*	10,000	50,000	100,000	25%	50%*	75%	1/2*	2/3	1	2	Student's t Distribution	Normal Distribution*	Truncated Normal
All Observations															
Symmetric Weighing	0.840	0.836	0.841	0.846	0.847	0.858	0.836	0.857	0.836	0.818	0.816	0.877	0.508	0.836	0.904
Right Weighing	0.194	0.202	0.201	0.211	0.213	0.093	0.202	0.452	0.202	0.257	0.320	0.377	0.387	0.202	0.235
Left Weighing	0.194	0.203	0.207	0.210	0.213	0.455	0.203	0.095	0.203	0.257	0.321	0.375	0.963	0.203	0.234
Positive Observations															
Symmetric Weighing	0.829	0.832	0.842	0.846	0.847	0.821	0.832	0.871	0.832	0.818	0.816	0.882	0.533	0.832	0.903
Right Weighing	**1.256**	**1.137**	**1.108**	**1.098**	**1.092**	**1.549**	**1.137**	**1.046**	**1.137**	**1.225**	**1.411**	**1.705**	0.419	**1.137**	**1.803**
Left Weighing	0.106	0.111	0.114	0.116	0.118	0.179	0.111	0.070	0.111	0.144	0.180	0.210	0.961	0.111	0.125
Negative Observations															
Symmetric Weighing	0.851	0.839	0.840	0.846	0.847	0.872	0.839	0.821	0.839	0.819	0.815	0.873	0.166	0.839	0.905
Right Weighing	0.104	0.111	0.111	0.117	0.118	0.070	0.111	0.178	0.111	0.143	0.182	0.212	0.089	0.111	0.125
Left Weighing	**1.235**	**1.133**	**1.123**	**1.094**	**1.090**	**1.051**	**1.133**	**1.539**	**1.133**	**1.230**	**1.417**	**1.702**	**1.048**	**1.133**	**1.822**

Notes: In bold, we depict the cases where the weighted estimator is more efficient than the unweighted estimator (RE larger than one). The columns with a * correspond to the same condition of the simulation design (i.e., *n* = 5,000; *pp* = 50%; θ = .5; Normal distribution).

Web Appendix C. Regression Trees

Trees (called CART; Breiman et al. 1983) have been very popular among marketing practitioners (Verhoef et al. 2003). These nonparametric models are graphically insightful. However, they are somewhat less known among marketing academics (see Risselada, Verhoef, and Bijmolt 2010 or Schwartz, Bradlow, and Fader 2014, for exceptions). Let $T(x, \Theta)$ be a tree model that fits a dependent variable to the covariates x. It can be written as a piecewise regression function,

$$T(x, \Theta) = \sum_{l=1}^{L} \omega_l \mathrm{I}(R(x) = R_l) \,, \tag{C1}$$

where $\Theta = \{R_1, \ldots, R_L, \omega_1, \ldots, \omega_L\}$, are the tree parameters. The tree has L terminal nodes with R_l the l^{th} terminal node. Based on the value of its x variables, each customer is classified into one of the L terminal nodes, as indicated by the indicator function $\mathrm{I}(R(x) = R_l)$. A customer classified into the l^{th} terminal node receives fitted value ω_l (in our case, a churn score). One can think about the classification of customers in terminal nodes as the repartition of customers in segments in latent-class analysis, except that a customer belongs to one segment exclusively.

Trees are estimated using a greedy algorithm that finds at each step the split that maximizes the reduction in impurity (Breiman 1983). Having found the best split, the data are partitioned into the two resulting regions and the operation is repeated on each of the two regions. The number of terminal nodes is determined by first fitting a tree with a large number of nodes and subsequently pruning it. The splitting process stops when some minimum node size (i.e. number of observations per node) is reached (in our case, we fix the minimum node size to 10 observations per node, see Ripley 1996). Next, pruning is done by removing the least important nodes using a cost-complexity criterion described in Hastie et al. (2009, p. 308). This criterion is conceptually similar to the information criteria used for segmentation (e.g. BIC or AIC) and ensures a trade-off between the goodness of fit of the tree to the data and the tree size (i.e. model complexity), thus avoiding overfitting. We set a maximum of 8 terminal nodes for the pruned tree. Blattberg et al. (2008, pp. 423-441) provide an extensive overview of classification and regression trees with an example.

Web Appendix D. Estimation Details of the SGB Algorithm

Let z_i denote the dependent variable and x_i a set of independent variables. A key feature of SGB is the way the fitted values $\hat{F}_{b-1}$ obtained at iteration $b-1$ are combined with the tree fitted values $T(x_i, \Theta_b)$ obtained at iteration b. Given the process described above, the updated estimate of parameters $\widehat{\Theta}$ is given by

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi\left(z_i, \hat{F}_{b-1}(x_i) + T(x_i, \Theta_b)\right). \tag{B1}$$

The optimal combination can be found by computing the gradient of the loss function. The gradient is the partial derivative of the loss function w.r.t. $F_{b-1}(x_i)$,

$$grad_{ib} = \left[\frac{\partial \Psi(z_i, F_{b-1}(x_i))}{\partial F_{b-1}(x_i)}\right]_{F_{b-1}(x_i)=\hat{F}_{b-1}(x_i)}, \tag{B2}$$

where $grad_{ib}$ is the gradient for the i^{th} customer at the b^{th} iteration. Intuitively, a large gradient indicates a large difference in the loss between $\hat{F}_b(x_i)$ and $\hat{F}_{b-1}(x_i)$. 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. Steepest descent chooses $h_b = -\rho_b grad_b$ with ρ_b called the "step distance." It minimizes the loss function between the fitted values of iterations b and $b-1$. In simple terms, $\hat{\rho}_b$ provides the optimal way to combine the fitted values from one iteration to the next,

$$\hat{F}_b(x_i) = \hat{F}_{b-1}(x_i) - \hat{\rho}_b grad_b. \tag{B3}$$

In order to diminish the risk of overfitting, Friedman (2002) proposes two modifications. First, randomization is added to the algorithm. At each iteration $b = 1, \dots, B$, a randomly selected subsample (without replacement) of N' customers is drawn from the calibration data with $N' \leq N$. Following Hastie et al. (2009), we choose $N' = 3000$ customers, i.e. 30% of the original calibration sample. For both applications, we experimented with various sizes and found out that 30% was the best choice. This random subsample is used for estimation during this particular iteration. Second, Friedman (2002) proposes that the "model should not learn too quickly from the data." Therefore, he suggests multiplying $\hat{\rho}_b$ in equation (B3) by a learning parameter ν, with $0 < \nu \leq 1$. Taken small enough, it ensures that the fitted values are slowly converging over the iterations. In our applications, we selected the learning rates that led to the best performance, i.e. .001 for the first application and .0005 for the second one. They prove to lead to the best holdout results for both the misclassification loss and for the profit-based loss functions. The algorithm runs over B iterations until it converges such that the difference between the loss at $b-1$ and the loss at b is less than 1.0^{-6}. For every decision made, we use the exact same settings for both loss functions to ensure a fair comparison. Table D1 summarizes the SGB estimation process (the R code is available upon request).

Table D1: Pseudo-code of the SGB algorithm

1. ***Initiatize*** $\hat{F}_0(x_i)$.
2. ***For b in*** $\mathbf{1}, \dots, \boldsymbol{B}$**:**
   a. For $i = 1, \dots, N$, compute the negative gradient as
   $$-grad_{ib} = -\left[\frac{\partial \Psi\left(z_i, F_{b-1}(x_i)\right)}{\partial F_{b-1}(x_i)}\right]_{F_{b-1}(x_i)=\hat{F}_{b-1}(x_i)}.$$
   b. Take a random sample without replacement of N' observations from the data, with $N' < N$.
   c. Fit a regression tree with L terminal regions $R_{b1}, \dots, R_{bL}$ on the random sample with the negative gradient as dependent variable and all the customer covariates as independent variables.
   d. Compute the optimal terminal node predictions $\rho_{b1}, \dots, \rho_{bL}$
   $$\hat{\rho}_{bl} = \underset{\rho_{bl}}{\operatorname{argmin}} \sum_{x_i \in R_{bl}} \Psi\left(z_i, \hat{F}_{b-1}(x_i) + \rho_{bl}\right).$$
   e. Update $\hat{F}_b(x_i) \leftarrow \hat{F}_{b-1}(x_i) + \nu \hat{\rho}_{bl(x)}$ with $bl(x)$ the index of the terminal node into which an observation falls into at iteration b given the values of its x, and ν is the learning rate parameter $0 < \nu \leq 1$.

Web Appendix E. Matching and Randomization Checks

Study 1.

We could not guarantee that the targeting was made at random. However, the data base was large enough and contained enough information on each customer that we were able to using matching in order to create matched samples that do not suffer from sampling selection bias. We use propensity score matching with nearest neighbor, as described in Ho, Imai, King, and Stuart (2007). Nearest neighbor matches a treated unit to a control unit that is closest in terms of a distance such as a logit. Propensity score matching greatly reduces the dependence of causal inferences on hard-to-justify, but commonly made, statistical modeling assumptions. The matched data provide inferences with substantially more robustness and less sensitivity to modeling assumptions. The approach was also used by Datta et al. (2015). The matched samples contain 2,595 customers in the treatment group and 2,595 customers in the control group. We performed randomization checks before and after matching in order to ensure that the treatment and control groups are comparable. We compare the distributions of the continuous variables in both samples using the Welch two sample t-test (H0: true difference in means is equal to zero) and using the asymptotic Pearson chi-squared test (H0: independence) for the categorical variables. Table E1 confirms that while the samples were not random before matching ($p<.05$ for most variables), the distributions are not different from each other after matching (all p-values > .10).

Table E1. Randomization Check Before and After Matching (Study 1)

	Before matching			**After matching**		
	Mean Control Group	Mean Treatment Group	*p*	Mean Control Group	Mean Treatment Group	*P*
Continuous variables						
Customer tenure (in	134.89	142.69	.00	155.36	154.39	.72
Customer age (in years)	46.59	47.04	.12	47.30	47.16	.69
Household size	3.08	3.11	.37	3.09	3.14	.23
Household income (in	24,230	24,306	.53	24,512	24,503	.95
Categorical variables						
Installation (DIY/Full)			.00			.73
Language (A/B/C/D)			.00			.91
Gender (M/F/U)			.01			1.00
Sport package 1 (yes/no)			.00			.91
Sport package 2 (yes/no)			.50			1.00

[22] This variable measures the number of months that an individual is a customer of the company. It can be larger than 3 months when the customer had access to other services (e.g. Internet, phone) from the provider before subscribing to the interactive television subscription.

Study 2.

In order to confirm that the randomization was done properly, we compared the distribution of individuals across the treatment and control groups, using the Welch two sample t-test (H0: true difference in means is equal to zero) for the continuous variable, and the asymptotic Pearson chi-squared test (H0: independence) for the categorical variables. Customer tenure was first standardized for confidentiality reasons. The results available in Table E2 confirm that the randomization was made properly (all p-values > .10).

Table E2. Randomization Check (Study 2)

	Mean Control Group	**Mean Treatment Group**	***P***
Continuous variables			
Customer tenure (in years)	.01	-.01	.57
Categorical variables			
Attendance of any event (yes/no)			.56
Online logging (yes/no)			.62
Download activity (yes/no)			.20
Special interest attendance (yes/no)			.52

Web Appendix F. Empirical Comparison of the Classic and Profit-based Loss Function

As simulations showed (Web Appendix B), the profit-based loss function works by minimizing the bias and maximizing efficiency toward observations that receive the highest weight, such that it provides more accurate predictions for the most valuable customers. To illustrate this, we first rank customers according to the scores estimated by SGB, using either the classic loss or the profit-based loss function. For both customer rankings, we group customers per decile, D1 to D10 (Ascarza 2018), and calculate for each experimental condition the average actual profit earned from a customer. Using Equation (8), we calculate the average actual profit of a customer in the treatment group for a given decile with

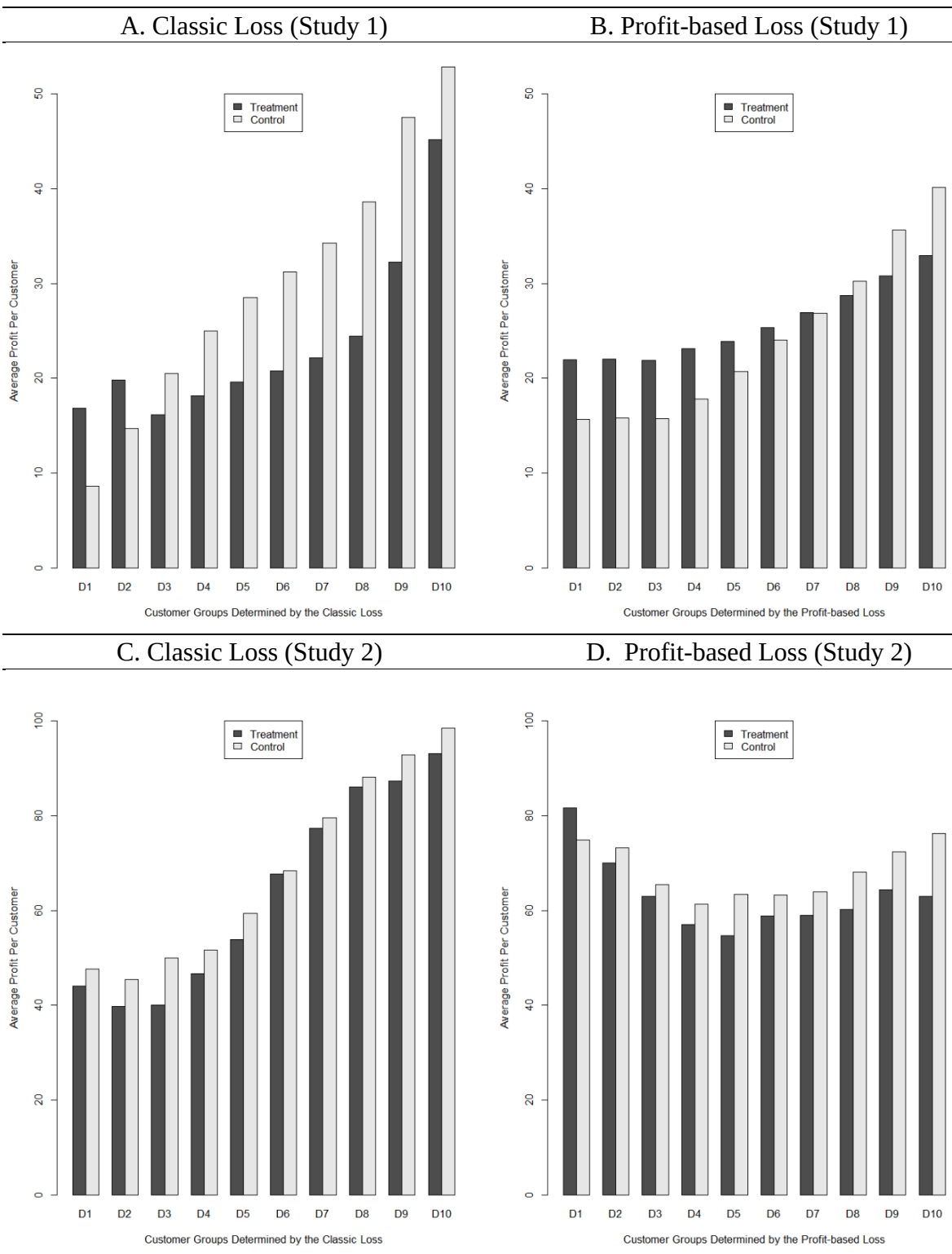
$$\pi_{Treatment} = \frac{1}{N_t}\sum_{i \in \text{Treatment}} \left(m_i^{(1)}\, \mathrm{I}\left(y_i^{(1)}=-1\right)-\delta\right), \quad \text{(F1)}$$

where N_t is the number of treated customers in the decile. The average profit of a customer in the control group for a given decile is

$$\pi_{Control} = \frac{1}{N_c}\sum_{j \in \text{Control}} m_j^{(0)}\, \mathrm{I}\left(y_j^{(0)}=-1\right). \quad \text{(F2)}$$

Comparing $\pi_{Treatment}$ with $\pi_{Control}$ provides an estimate of the average treatment effect in a specific decile. In other words, it shows for which decile the intervention is the most beneficial. Figure F1 contains the results from both empirical applications.

Figure F1. Average profit per customer across experimental conditions for different group deciles, based on SGB scores, using classic or profit-based loss functions

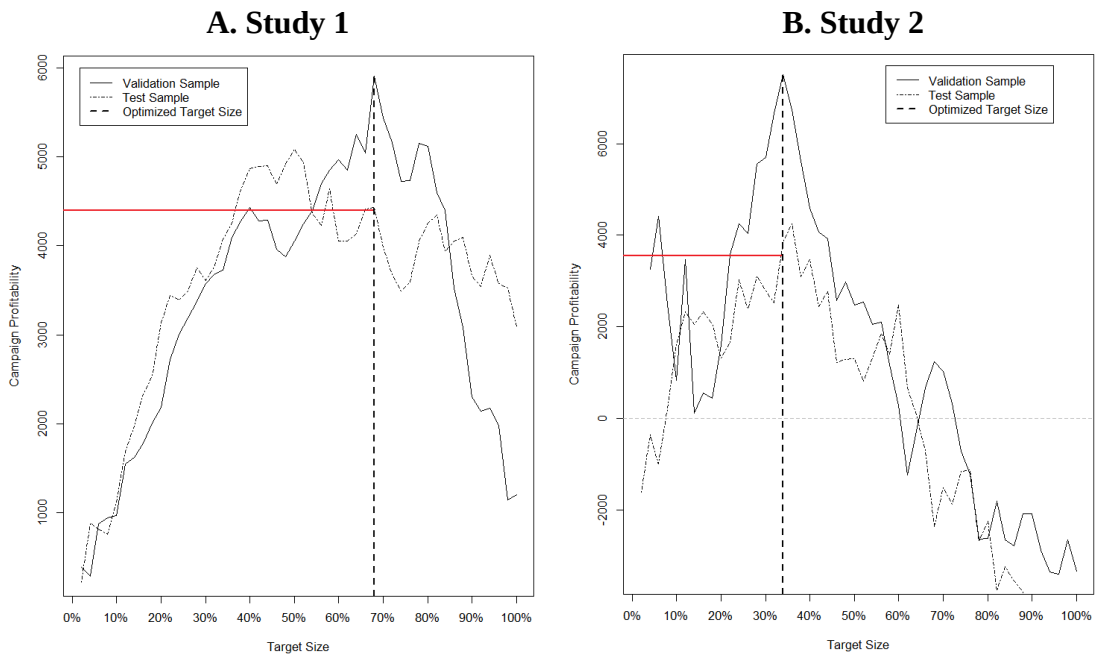


For Study 1, the classic loss function in Panel A reveals a different pattern than the profit-based loss function in Panel B. In Panel B, the intervention has the strongest impact (difference between treatment and control groups) in the top decile, which then slowly decreases across deciles. The treatment effect is positive for the first 60% of the ranking (D1–D6), reflecting how the profit-based loss function ranks high profit lift customers higher than lower profit lift customers. In contrast, Panel A indicates a negative impact of the classic loss function for most deciles except D1 and D2. Therefore, the classic loss function missed out on many high profit lift customers who were assigned lower rankings or else intervened with low profit lift customers who were placed too high in the ranking. By focusing on the "wrong" criterion (accurately predicting churn), this function performs significantly worse than the profit-based loss function. The results for Study 2 similarly confirm that the profit-based loss function (Panel D) establishes a larger treatment effect for the first decile than the classic loss function (Panel C). However, the differences are less pronounced, likely because the total impact of the campaign was small in the first place (see Figure 2). However, for customers in the first decile, the profit-based loss function exerts a positive impact (average profit in the treatment group is larger than average profit in the control group), whereas the classic loss function has a negative impact for all deciles. Thus, the improved performance of the profit-based loss function emerges because it assigns higher rankings to high profit lift customers.

Web Appendix G. Illustration of the Target Size Optimization Procedure

Figure G1 illustrates the procedure described in Section 5.2. It shows the holdout profit curve at a random bootstrap iteration as a function of the target size, using both validation data (solid line) and the holdout test sample (dashed line). The optimal target size is the maximum reached on the validation sample (solid line). For this particular bootstrap sample, the method recommends a target size of 68% for Study 1 and 34% for Study 2 (vertical dashed line). To evaluate holdout performance, we calculate the campaign profit for this target size with the third test sample (dashed line). It corresponds to the intersection of the horizontal line with the vertical dashed line. The performance is slightly inferior (confirming that it is a true holdout evaluation) but close to the maximum performance attained if the optimal target size for the test sample were known.

Figure G1. Holdout campaign profit on the validation and test (holdout) sample and optimized target size for a random bootstrap iteration



Notes: The curves represent the holdout profits of the campaign for a random bootstrap iteration. The vertical dashed line represents the optimized target size and the horizontal line is the corresponding holdout campaign profit on the test sample.

References

Carsey, Thomas M., and Jeffrey J. Harden (2013). *Monte Carlo Simulation and Resampling Methods for Social Science*. Sage Publications.

Chambers Ray L. (1996), *Weighting and Calibration in Sample Survey Estimation*, pp.125-148, in Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini Birth.

Ho Daniel E., Imai Kosuke, King Garry, and Elizabeth A. Stuart (2007), Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15 (3), 199–236.

Ripley, Brian D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

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.

Schwartz, Eric M., Eric T. Bradlow and Peter S. Fader (2014), Model Selection Using Database Characteristics: Developing a Classification Tree for Longitudinal Incidence Data, *Marketing Science*, 33(2), 188-205.

Tofallis Chris (2015). A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation, *Journal of the Operational Research Society*, 66(8),1352-1362

Verhoef, Peter C., Penny N. Spring, Janny C. Hoekstra, and Peter S. H. Leeflang (2003), The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing in the Netherlands, *Decision Support Systems*, 34 (4), 471-481.