Retention futility:
Targeting high risk customers might be ineffective

Eva Ascarza[†]

August 2017

Forthcoming at the *Journal of Marketing Research*

[†]Eva Ascarza is the Daniel W. Stanton Associate Professor of Business at Columbia Business School (email: ascarza@columbia.edu). The author benefited from comments by Bruce Hardie, Kamel Jedidi, Oded Netzer, the participants of the Choice Symposium session on Customer Retention, and the audience at the marketing division brown bag at Columbia Business School, the seminars at Harvard Business School, University of Oxford, University of Maryland, University of Notre Dame, Cornell University, University of Michigan, and University of Chile, and the 2016 Marketing Analytics and Big Data Conference at Chicago Booth and 2017 Marketing Science Conference at USC. The author is grateful to Matt Danielson for his help collecting data and to Yanyan Li for excellent research assistantship.

Abstract

Retention futility: Targeting high risk customers might be ineffective

Companies in a variety of sectors have increasingly started managing customer churn proactively, generally by detecting customers at the highest risk of churning and targeting retention efforts towards them. While there is a vast literature on developing churn prediction models that identify customers at the highest risk of churning, no work has investigated whether it is indeed optimal to target those individuals. Combining two field experiments with machine learning techniques, we demonstrate that customers identified as having the highest risk of churning are not necessarily the best targets for proactive churn programs. This finding is not only contrary to common wisdom, but also suggests that retention programs are sometimes futile not because firms offer the wrong incentives, but because they do not apply the right targeting rules. We propose an approach for proactive churn management that, through experimentation, identifies the observed heterogeneity in response to the intervention and targets customers based on their sensitivity to the intervention, regardless of their risk of churning. We empirically demonstrate that the proposed approach is significantly more effective than the standard practice of targeting customers with the highest risk of churning. More broadly, we encourage firms/researchers using randomized trials (or A/B tests) to look beyond the average effect of interventions and leverage the observed heterogeneity in customers' response to select customer targets.

Keywords: Churn, retention, proactive churn management, field experiments, heterogeneous treatment effect, random forest, customer relationship management.

"[...]the client wanted to *identify* customers with *high risk of defection* and implement ways to retain them. Proven results helping companies reduce churn was a key factor in the client's choice of Accenture"

Accenture: Client case study

"More sophisticated predictive analytics software use *churn prediction models* that predict customer churn by assessing their propensity of *risk to churn*. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn"

Wikipedia: Customer Attrition

1 Introduction

Churn management is a top priority for most businesses (Forbes 2011) as it directly ties to firm profitability and value. Churn prediction plays a central role in churn management programs. By predicting churn before it happens, marketers can proactively target activities to customers who are at risk of churning in order to persuade them to stay (Neslin et al. 2006; Blattberg, Kim, and Neslin 2008). In practice (e.g., Accenture Analytics 2014; Wikipedia 2017), as the above quotations highlight, targeting is generally achieved by assigning a churn propensity to each customer, selecting those that are at the highest risk of churning, and contacting them with a retention program that is aimed to retain them (e.g., by providing incentives to stay). Because churn prediction plays a crucial role in the design of proactive churn management programs, researchers in the areas of marketing, statistics and computer science have developed a variety of methods to accurately predict which customers are at *highest risk of churning.*

However, while a significant amount of work has tested the accuracy of such methods, no work has investigated whether proactive churn management programs should be targeted to individuals with the highest risk of churning. The main goal of this paper is to fill that gap. We empirically examine whether firms should target their retention efforts to customers with the highest risk of churning. More specifically, we challenge the most common practice for proactive churn management and claim that, when the main purpose of churn prediction is to *select* customers for proactive/preventive retention efforts, identifying customers at high risk of churning does not suffice to drive the firm's targeting decisions. We argue that, because customers respond differently to retention interventions, firms should not target those with the highest risk of churning but rather those with the highest sensitivity to the intervention. Researchers and practitioners might have (implicitly) assumed that these two groups of

customers are the same. We demonstrate that this is not the case — customers' risk of churning does not necessarily relate to their sensitivity to the retention incentive. Therefore, failure to account for customer differences in the response to the retention intervention often results in less effective, and even futile, proactive churn management programs.

While understanding customer heterogeneity in the sensitivity to retention actions might have been difficult decades ago, this task is much easier nowadays. Advances in technology, developments in data analysis, and the increased popularity and ease of implementation of field experimentation have enhanced firms' ability to gain insights about customers. Through experimentation, firms can better understand customer heterogeneity in the response to marketing interventions. We encourage firms to broaden the use of randomized experiments and use them to *identify customers to target*, as doing so would increase the effectiveness of their actions. Consequently, we propose an approach for proactive churn management that (1) leverages the firm's capabilities by running a retention pilot, (2) identifies the observed heterogeneity in the response to the intervention, and (3) selects target customers based on their sensitivity to the intervention, ensuring that the retention efforts are not futile.

We empirically validate our proposed approach by analyzing customer behavior in two field experiments conducted in different markets (South America and Asia) and covering two different sectors (telecommunications and professional memberships). Combining these field experiments with machine learning techniques, we demonstrate that our approach is more effective than targeting customers based on their risk of churning. We find that, across the two studies, the same retention campaign would result in a *further reduction* of 4.1 and 8.7 percentage points in churn rate, respectively, if each focal firm followed the proposed approach instead of the industry standard of targeting customers at the highest risk of churning. We consistently find that customers identified as being at the highest risk of churning are not necessarily the best targets for proactive churn programs. In particular, we find that the overlap between the group of customers with the highest sensitivity to the retention efforts and those with the highest risk of churn is roughly 50%; hence, the relationship between these two variables does not differ from independence (or random overlap). Finally, it is

important to highlight that this result is not driven by our modeling assumptions as we estimate the sensitivity to retention efforts in a nonparametric way.

Our approach also allows us to identify which customer characteristics (among the observed variables) best predict sensitivity to the retention intervention. In both applications, we identify several variables that highly correlate with being "at risk" but have no relationship with the sensitivity to the intervention, implying that selecting customer targets based on those variables would likely result in futile retention efforts. Furthermore, we find a set of characteristics that *positively* correlate with churn — or being "at risk" — but *negatively* correlate with the sensitivity to the intervention (or vice versa). In such cases, if the firm were to target based on these variables, they would be directing the resources to customers for whom the intervention is most harmful and would likely increase churn.

Finally, unlike churn scoring models, the proposed approach not only ranks customers by *whom should be targeted* first, but also identifies *the level of marketing intensity* at which the retention campaign becomes ineffective or futile. This insight is crucial for companies when making decisions such as how many customers to target or how much resources to allocate to a retention campaign. Across the two applications investigated in this research, we find that half of the retention money is wasted. Most importantly, estimating (observed) customer heterogeneity in the response to the campaign allows us to identify *which* half.

Our findings are generalizable to a large variety of business settings beyond the ones investigated in this work (e.g., credit card, software providers, online and offline subscriptions, leisure memberships) in which customer level data are available and where managing customer churn is a concern. Compared to the standard practice (i.e., targeting those "at risk"), the proposed method requires a market test as an additional step. While this step might be seen as challenging for some firms, implementing such a test is well within the capabilities of any firm already running proactive churn management programs. Furthermore, the method is generalizable (can be applied to a wide range of business contexts), easily scalable (can handle very large datasets), and is estimated using existing packages in R (freely available software), facilitating its use by practitioners.

A history of proactive churn management

The issue of customer retention/churn gained traction in the late 1990s and early 2000s, when the marketing field started devoting attention to customer relationship management (CRM). The earliest work on customer retention focused on identifying the drivers for such behavior, highlighting service quality, satisfaction and commitment as important constructs determining the lifetime of customers (e.g., Bolton 1998; Bolton and Lemon 1999; Ganesh, Arnold and Reynolds 2000; Gruen, Summers and Acito 2000; Lemon, White and Winer 2002). These findings become extremely relevant when Gupta, Lehman and Stuart (2004), among others, quantified the potential impact of retaining customers on the long-term profitability of the firm. Not surprisingly, firms across various sectors (from telecom and pay TV to credit cards among others) increasingly started to *proactively* manage churn by detecting those customers at the highest risk of churning and targeting their retention efforts towards them.[1] The rationale behind such a practice is straightforward: Targeting customers with the highest propensity to churn enables firms to focus their efforts on customers who are truly at risk of churning and to potentially save money that would be wasted in providing incentives to customers who would have stayed regardless (Neslin et al. 2006).

Because churn prediction played such a crucial role in determining which customers should be targeted/contacted in proactive churn management programs, marketing researchers started proposing a variety of methods to predict which customers are at the highest risk of churning. Traditionally, methods such as logistic regression and classification trees had been widely used in practice (see Neslin et al. (2006) for a review of methods and their performance). More recently, longitudinal methods such as hidden Markov models and new machine learning tools — including random forests, support vector machines and bagging and boosting algorithms — have been proposed to predict customers' propensity to churn (e.g., Lemmens and Croux 2006; Risselada, Verhoef, and Bijmolt 2010; Schweidel, Bradlow and Fader 2011; Ascarza and Hardie 2013).

[1] This practice differs from *untargeted* approaches to reduce churn, which aim at increasing satisfaction and switching costs across all customers, or from *reactive* churn management programs, in which firms wait for the customer to churn (or attempt to churn) before offering her incentives to stay (Blattberg et al. 2008).

Two streams of work have investigated approaches that go beyond targeting those at the highest risk of churning. The first approach has recognized that the cost of misclassifying customers largely depends on the profitability of each customer (Verbeke et al. 2012). Accordingly, Lemmens and Gupta (2017) incorporate a profit-based loss function in the model estimation, thus reducing prediction errors for customers with higher expected profitability.

The second approach, mainly driven by practitioners, has recognized the need to examine the *incremental* effect of the firm's actions rather than merely the behavior incurred by the customer (e.g., why contact a customer who would have bought anyway?). Differential response modeling or uplift models, originally developed in the domain of direct marketing, have be proposed for churn management (Siegel 2013; Provost and Faucett 2013; and Guelman, Guillén and Pérez-Marín 2012; 2015). However, to the best of our knowledge, no work has investigated the effectiveness of these approaches against the practice of targeting customers at the highest risk of churning, neither have proposed guidelines to firms as to how to collect the information needed to estimate the incremental impact of retention efforts.

The problem of modeling churn expanded outside the marketing literature both in scope and volume. As firms started integrating proactive churn management tools into their information systems, researchers in the areas of statistics, information systems, computer science, engineering and operations, developed methods to identify customers' at the highest risk of churning (see Hadden et al. (2007) for a review on early methods and Ngai et al. (2009); Huang, Kechadi, and Buckley (2012) for more recent developments). Such papers sought to identify which methods are best suited to accurately estimate customers' propensity to churn and how to incorporate such (risk scoring) algorithms in firms' information systems.

Looking back, the vast majority of the academic work on customer retention/churn in the past two decades has focused on developing methods to predict, based on historical data, which customers are at the highest risk of churning.[2] Firms then use these methods in their proactive retention programs to select customer targets. However, while the literature has

[2]For a review of the broader literature on customer retention (beyond proactive churn management), refer to Ascarza, Fader and Hardie (2017) and Ascarza, Neslin et al. (2017).

provided a thorough investigation into the accuracy of these methods in predicting *which customers are more likely to churn*, no work has investigated *whether it is optimal for firms to target those individuals*. In other words, are customers with the highest risk of churning those for whom proactive churn management programs are most effective?

While there is a long tradition in marketing to estimate the heterogeneous effect of marketing actions to inform targeting decisions (e.g., Rossi et al. 1996; Ansari and Mela 2003), such a view has not resonated in the context of proactive churn management, partly because the field has implicitly assumed that customers with the highest risk of churning also have highest sensitivity to retention interventions, partly because firms did not have enough variation in their databases to estimate the heterogenous effect of retention actions.[3] In the following section we propose an approach for proactive churn management that overcomes this limitation. We then use this approach to empirically test the relationship between customers risk of churning and the effectiveness of proactive churn management programs.

Proposed targeting method for proactive churn management

The firm's targeting problem

The firm is faced with the problem of deciding which customers should be targeted in the next retention campaign, the primary goal of which is to increase long-term profitability by reducing churn among current customers. The most common approach in practice, and what previous literature has suggested, is to target the customers who are at the highest risk of churning. In this paper we argue that such a targeting rule is not necessarily optimal.

Let us consider a customer i, with observed characteristics X_i (e.g., past behavior, demographics). The practice followed by most firms is to calculate the customer's probability to churn given her observed characteristics, $P[Y_i|X_i]$, and decide whether to target her based on this metric. The main limitation of this approach is that the decision variable — whether to target or not — is not incorporated in the problem specification.

[3]Unlike purchasing, customers can only churn once, limiting the number of observations per customer. Also, most firms do not vary their proactive retention strategies as often as marketing variables change in other contexts (e.g., price, display).

Alternatively, let T_i denote whether customer i is targeted; this variable takes value 1 if the customer is targeted, 0 otherwise, and let us define

$$LIFT_i = P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1] \text{ and} \quad (1)$$

$$RISK_i = P[Y_i|X_i, T_i = 0], \quad (2)$$

where $P[Y_i|X_i, T_i = 1]$ is the probability that the customer will churn *if she is targeted* and $P[Y_i|X_i, T_i = 0]$ is the probability that she will churn *if she is not targeted*.[4] We argue that firms should target their retention efforts to customers with highest $LIFT_i$, for whom the *impact* of the intervention is *highest*, regardless of their intrinsic propensity to churn.[5]

Furthermore, contrary to conventional wisdom, customers sometimes react negatively to the firm's intervention. Blattberg et al. (2008) noted that one of the potential concerns of proactive churn management is that, in some cases, a retention campaign might even encourage "not-would-be churners" to churn. For example, the intervention could make customers realize their (latent) need to churn (Berson, Smith and Thearling 2000) or could break the inertia that prevented them from churning (Ascarza, Iyengar and Schleicher 2016). If firms target on the basis of their risk of churning, these specific customers would likely be selected for the campaign. On the contrary, targeting customers based on *LIFT* minimizes the likelihood that such customers will be targeted because their *LIFT* will be negative.

Estimating the incremental effect of the campaign

While estimating $RISK_i$ is straightforward as it only requires data readily available in the firms' database, estimating $LIFT_i$ requires the comparison of two outcomes that cannot be both observed — a customer is either targeted or not. As a consequence, additional

[4]Note that in the case of proactive churn management $P[Y_i|X_i] = P[Y_i|X_i, T_i = 0]$ as firms compute the risk of the customer churning before she receives any targeted incentive.

[5]Bodapati (2008) makes a similar claim in the context of product recommendations, arguing that recommendation systems should maximize the likelihood of "modifying customers' buying behaviors relative to what the customers would do without such a recommendation intervention." Conceptually, the difference between that context and ours is that in recommendation systems, the norm had been to target customers whose probability of incurring on the behavior of interest was already high (i.e., highest probability to buy), whereas in the case of proactive retention campaigns the general norm is to target customers with lowest probability of incurring on the behavior of interest (i.e., lowest probability to renew). Methodologically, the two papers are distinct. While Bodapati (2008) assumes a two-step model for purchase probability — relying on parametric assumptions about the impact of the firm's recommendation on product awareness and satisfaction — ,we estimate the difference in churn probability directly, and nonparametrically.

variation in the data and assumptions about how such data are generated will be needed. Assuming there is no prior information about how customers respond to specific retention interventions, we encourage the firm to run a (small-scale) pilot retention campaign in which the intervention is randomized across a representative sample of customers. This step will suffice to estimate the incremental effect of the campaign on the remaining customers.

At first glance, the need for a retention campaign pilot might seem cumbersome, costly, or difficult for the company to implement. However, firms are increasingly adopting the use of small- and large-scale experiments (e.g., A/B testing) as part of their regular business. In turn, every company that has the ability to individually target customers — more specifically, every company that is already implementing proactive churn management programs — is equipped to run randomized experiments. The pilot campaign only requires the firm to run the intended retention campaign, but instead of targeting specific customers based on some pre-specified rule, they should target (i.e., *treat*) a randomly selected group of customers.

Once the company has run the retention pilot, estimating the heterogeneous treatment effect is straightforward. More formally, following the potential outcomes framework for causal inference (e.g., Rubin 2011), we can assume the existence of potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$, corresponding to whether customer i would have churned with and without the treatment, respectively. Given this formulation, the firm would estimate the Conditional Average Treatment Effect (CATE), defined as $\mathbb{E}[Y_i^{(0)} - Y_i^{(1)}|X_i]$, which corresponds to the treatment effect conditional on a given set of covariates X_i.[6] Given that the outcome variable is binary (i.e., $Y_i = 1$ if customer churns, 0 otherwise), we can express CATE as

$$\mathbb{E}[Y_i^{(0)} - Y_i^{(1)}|X_i] = P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1],$$

which corresponds to $LIFT_i$, as defined in (1). As for covariates, we use information readily available in the firm's database (e.g., previous purchases) such that predicting the $LIFT_j$

[6]Because the retention pilot is randomized across customers, it is reasonable to assume unconfoundedness (Rosebaum and Rubin 1983), that is, that the treatment is independent on the potential outcomes, given the set of covariates X_i. As a consequence, one can use the experimental data to consistently estimate the heterogeneous treatment effect. Note that we define 'treatment' as the action of targeting a customer (i.e., the firm sending a retention incentive to a particular customer) hence our experimental design allows us to consistently estimate the treatment effect. Alternatively, if one defined 'treatment' as a customer receiving and responding to the offer, we would be estimating the intention-to-treat effect. We chose to define treatment as the action of targeting because that is the decision variable for the firm.

for any remaining customer j (who was not part of the pilot) is straightforward, as it only involves the evaluation of the estimated model on a new set of observed covariates X_j.

A variety of methods have been proposed with regards to how to estimate CATE. In one stream of work, researchers in the areas of statistics, economics/econometrics, biostatistics, and political science have explored methods to consistently estimate heterogeneous treatment effects. Generalized linear models or generalized additive models have been traditionally used for such purposes (see Feller and Holmes (2009) offer an overview). In a second stream of work, focusing more on predictions than on inference, marketing practitioners and researchers from the areas of data mining and computer science have developed so-called "uplift" models.[7] The main goal those models is to predict which individuals would respond more favorably to an intervention, without focusing on the asymptotic characteristics of the estimates or their interpretation. Most recently, Guelman et al. (2015) build on the latter stream and propose a method to estimate uplift using random forests, combining approaches previously used for uplift modeling (Rzepakowski and Jaroszewicz 2012) with machine learning methods (Breiman 2001), achieving accuracy and stability on their predictions.

In parallel, researchers in the areas of statistics and economics also start recommending the use of tree-based methods for conducting causal inference (Athey and Imbens 2016, Wager and Athey 2017). At its core, both approaches share the goal of finding partitions of the data (based on observed characteristics) that differ in the impact of the intervention (i.e., the magnitude of the treatment effect). The main difference is that the methods proposed for causal inference (as developed by Athey and Imbens 2016 and Wager and Athey 2017) employ an "honest" estimation whereby one sample is used to construct the trees/partitions and another to estimate the treatment effect. Such an approach not only identifies individuals with larger/smaller treatment effects, but also enables the researcher to obtain consistent estimates and valid confidence intervals for the treatment effects. In this research we combine

[7]Uplift modeling is also known as incremental response, true-lift, or net modeling. See Radcliffe and Surry (1999) for early work on the topic, and Sołtys, Jaroszewicz and Rzepakowski (2015); and Jaroszewicz (2016) for reviews of the various methods and applications. Moreover, some of the software packages that include uplift modeling are Incremental Response Modeling using SAS, Spectrum Uplift^{TM} and the R package Uplift.

the algorithm proposed by Guelman et al. (2015) with recursive data splits (in the spirit of Athey and Imbens (2016)) to compute treatment effects and confidence intervals of the treatment effect in different groups of customers.[8]

Finally, once the heterogeneous treatment effect model is estimated on the pilot sample, we use the model to predict the (out-of-sample) $LIFT_j$ for all remaining customers j who were not part of the retention pilot and will be part of the actual campaign.

Selecting customer targets

How should the firm select which and how many customers to target? First, as suggested earlier, the firm should prioritize the retention efforts towards customers with highest $LIFT_j$, as doing so will increase the effectiveness of the campaign. Second, the value of $LIFT_j$ should be used not only as a "ranking" metric to better allocate resources but also to determine which (or how many) customers should be targeted — that is, to decide how much resources should be put in place. Recall that $LIFT_j$ is the expected *effect of the treatment* (or campaign). Hence the firm should target only those customers for whom $LIFT_j > z$, where z represents the minimum effect (i.e., increase in average churn probability) that the firm wants to achieve with the retention campaign.[9]

Bringing it all together

Figure 1 outlines our proposed approach for proactive churn management. To summarize, firms should leverage the knowledge that customers respond differently to marketing actions by relating such customer heterogeneity to the variables they already observe (e.g., past behavior, characteristics). In order to obtain those insights, we recommend firms run small-scale randomized tests and use the results to prioritize the retention efforts on customers whose sensitivity (i.e., incremental effect) is expected to be the highest, conditional on their

[8] We choose this algorithm because of its accuracy — Guelman et al. (2015) demonstrate its superiority over other non tree-based methods — and its availability — R package `uplift` freely available at `https://cran.r-project.org` — as it facilitates its use among analysts and practitioners. Details about the algorithm and implementation are presented in Section 1.

[9] This number varies across firms and across campaigns as it depends on multiple factors such as the cost of running the campaign, and the specific risk (or loss) the company is willing to take (balancing the gains of saving customers vs. the costs of losing others). Because we do not have such information for our field studies, we will assume $z = 0$ in our validation analyses.

observed characteristics. While there is a general consensus among both academics and practitioners that measurement of the impact of marketing actions requires a focus on incremental outcomes (obtained via randomized interventions), there has been less widespread recognition of the focus on incrementality when selecting target customers. Hence, with our approach we encourage firms to employ A/B testing (or small-scale pilots) not only to evaluate marketing actions but also to *identify* customer targets.

Insert Figure 1 here

Finally, the main difference between our proposed approach and common industry standard is that we focus on the *incremental effect* of the campaign rather than on the propensity to churn. However, it is worth noting that estimating $RISK_i (= P[Y_i|X_i, T_i = 0])$, as traditional churn scoring models do, and estimating $LIFT_i (= P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1])$, as our approach does, both rely on the *same observed* set of variables (X_i). Therefore, in cases where it is optimal to target those customers with highest risk of churning, our proposed approach "nests" to the already-in-use proactive retention programs. If customers with the highest propensity to churn are indeed those for whom the retention campaign is more effective, the proposed method would recommend targeting the *same customers* as traditional churn scoring models would recommend. The proposed approach therefore not only generalizes existing practices for proactive churn management but also allows firms to test the optimality of their current retention efforts.

Validation

We assess the performance of the proposed approach by analyzing two field experiments. Both studies involve a firm running a retention campaign in which a marketing intervention (treatment) is *randomized* across customers. While the type of intervention varies across companies and contexts, in both cases the treatment involved an incentive (e.g., a reward) that was expected to increase retention among customers. In each of the studies we observe *churn* behavior for both treated and non-treated customers. We also observe individual-level information such as multiple forms of past behavior and other customer characteristics,

variables that already exist in the firms' database, which are normally used to assess customers' risk of churning. These variables are collected *prior* to the experiment and thus are independent to the intervention. Hereafter we refer to these variables as *'covariates.'*

First, we introduce a methodology to empirically compare the proposed approach with that of targeting customers at the highest risk of churning. We then describe the details specific to each study and present the results obtained in each case. We conclude with a general discussion of the findings.

Methodology

Our main goal is to measure the effectiveness of using our proposed method to select customer targets and compare it with that of the standard practice for proactive churn management of targeting customers at the highest risk of churning. We leverage the experimental set-up of the field studies to simulate what the impact of these retention campaigns would be had the focal firms implemented our approach instead of the standard practice. We replicate the validation exercise for each of the field studies.

Broadly, our validation strategy is as follows. For each study, we will split the data into two samples. One sample will be used to resemble the pilot study from which the firm estimates the heterogeneous treatment effect. The other sample will be used to predict what the outcome of the retention campaign would be under two different scenarios: (1) if the firm targeted customers based on their risk of churning, and (2) if the firm used our approach to target customers. We then compare the outcomes across scenarios and quantify the benefits of following our approach over the standard practice. We proceed as follows:

Step 1: Data split

For each of the datasets, we randomly allocate 50% of the customers to the calibration sample and the remaining 50% to the validation sample. Note that both treated and non-treated customers are included in each sample. That way, the calibration sample will resemble the outcome of the retention pilot, and the validation sample will be used to evaluate the effectiveness of marketing campaigns under different targeting practices (scenarios).

Step 2: Estimate a model for incremental churn (i.e., *LIFT* model)

Using the observed data from customers in the calibration sample (including *'treatment'* condition, post-experiment *'churn'*, and pre-experiment *'covariates'*), we estimate a heterogeneous treatment effect model using churn as a dependent variable. This model will be used to predict the customers' sensitivity to the marketing intervention. As discussed earlier, we employ the algorithm proposed by Guelman et al. (2015) to estimate the *LIFT* model.[10]

Step 3: Estimate a ("traditional") model for risk of churning

Using the calibration data, we also estimate a "traditional" churn model that will be later used to predict customers' propensity to churn. Among the customers in the calibration sample, we select those who belong to the control group (i.e., who did not receive any incentive) and model churn as a function of the customers' observed characteristics.[11] This step mirrors the standard churn scoring models used in practice. Because we are most interested in predicting risk of customers outside this sample, we employ the method that maximized cross-validation accuracy, in our case a LASSO binary regression.[12]

Step 4: Predict churn metrics in the validation sample

Using the models estimated in steps 2 and 3, we predict two variables of interest among customers in the validation sample:

1. $[RISK]$ Using the risk scoring model estimated in Step 3, we predict the risk of churning for each customer in the validation sample. Specifically, we define

$$RISK_j = P(Y_j = 1 | X_j = x_j), \tag{3}$$

[10] See Web Appendix A1.1 for details about the estimated model. The R code used for the empirical application is made available as a supplemental file.

[11] Note that we can only use control observations (i.e., customers who did not receive any incentive) to calibrate the RISK model, whereas both control and treatment customers are used to calibrate the LIFT model. We corroborate that this difference in sample size is not driving, not even partially, the results we obtain (Web Appendix A2.1).

[12] We tried multiple approaches to estimate the churn scoring model, including GLM, random forests, and SVMs. We choose the LASSO approach combined with a GLM model as it provided the most accurate forecasts in our applications. See Web Appendix A1.2 for details about the estimated models. Furthermore, we check the robustness of the results when using random forest to estimate the *RISK* model (to be consistent with the *LIFT*) modeling approach. Our results remain largely unchanged when using such an approach (See Web Appendix A2.2 for results).

where j denotes a customer in the validation sample. This step corresponds to the firm's practice of assessing customers' risk of churning before selecting targets for a retention campaign.

2. [$LIFT$] Using the incremental churn model estimated in Step 2, we predict, for each customer in the validation sample, the following quantities:
   - The probability of churn *if not targeted*, defined as $\mathrm{P}(Y_j = 1|T_j = 0, X_j = x_j)$
   - The probability of churn *if targeted*, defined as $\mathrm{P}(Y_j = 1|T_j = 1, X_j = x_j)$

   and then define $LIFT_j$ as the expected incremental effect of the campaign, given X_i:

$$LIFT_j = P(Y_j = 1|T_j = 0, X_j = x_j) - P(Y_j = 1|T_j = 1, X_j = x_j). \quad (4)$$

   We subtract *targeted* from *not targeted* such that positive values of $LIFT_j$ mean that the campaign *reduces* churn (i.e., it would be beneficial for the firm).

   Note that $LIFT$ represents the customer's sensitivity to the intervention. One might have assumed that $LIFT < 0$ is not a possible outcome as it implies that the retention campaign increases, rather than decreases, the customers' likelihood of churning. However, we want to account for this possibility because it is possible for retention campaigns to increase churn (Berson et al. 2000; Blattberg et al. 2008, Ascarza et al. 2016). Moreover, even if a retention campaign is overall positive (i.e., it reduces churn at the aggregate level), it is also possible that it had a negative effect on *some* customers. By allowing $LIFT$ to be negative, we account for such a possibility.[13]

Step 5: Measure customer heterogeneity in treatment effect

We leverage the richness of the experimental design to evaluate the effect of the intervention in different groups of customers, depending on their level of *RISK* and *LIFT*. More specifically, we model customer heterogeneity by measuring the treatment effect among subpopulations of customers, defined by the deciles of each variable of interest (either *RISK* or

[13] We also created a discrete metric for sensitivity to the retention intervention corresponding to whether the customer would have changed her behavior as a consequence of the intervention (in the spirit of the "switchable" customer, Gensch 1984). The analyses with such metric were almost equivalent to those obtained with the $LIFT$ metric and are available from the authors.

LIFT). By doing so we can compare the treatment effect among customers in the top decile of *RISK* with that of those in the second decile, third decile, and so forth. Similarly, we can also compare the magnitude of the treatment effect on customers in the top *RISK* decile with that of customers in the top *LIFT* decile. We choose to model heterogeneity in this fashion not only for its flexibility — we do not impose any parametric relationship between the treatment effect and the level of *RISK* or *LIFT*, hence allowing for linear, U-shape relationships, and so forth — but also because decile split is a segmentation method commonly employed by firms (e.g., Bauer, C.L. 1988; Bayer 2010). Moreover, a metric commonly used to assess the performance of a churn model is the 'top-decile lift', which implies that firms would target the top 10% of the customers in terms of risk of churn.

We perform this analysis as follows. For each of the two metrics, we split the validation sample into ten groups of customers of equal size (based on the deciles for each metric) and calculate the average treatment effect within each group. Because the treatment/control allocation in the field experiment was fully random, all of these subgroups contain both customers who received the retention incentive (treatment group) and those who did not (control group). This aspect of the data is important because we use such variation (between treatment and control conditions) to calculate the treatment effect in each of the subgroups. We proceed as follows:

- We calculate the *RISK* deciles $(r_1, r_2, ..., r_9)$ and split the validation sample into ten equally-sized groups of customers such that group R_1 includes all customers whose $RISK_j$ is lower than r_1 (i.e., $R_1 = \{j\}_{RISK_j < r_1}$), R_2 includes customers with $r_1 < RISK_j <= r_2$, and R_{10} includes those with $RISK_j > r_9$. We then compute, for each group R_d, the difference in the actual churn rate — i.e., proportion of customers who churn — across experimental condition. More formally, we calculate each *treatment effect* (TE) as follows:

$$TE_{R_d} = \underbrace{\frac{1}{N_l} \sum_{l \in \text{Control}} \mathbb{1}\big[(Y_l = 1)\big]}_{\text{Churn rate in control}} - \underbrace{\frac{1}{N_{l'}} \sum_{l' \in \text{Treatment}} \mathbb{1}\big[(Y_{l'} = 1)\big]}_{\text{Churn rate in treatment}} \qquad \text{for } d = 1, ..., 10, \quad (5)$$

where N_l is the number of customers belonging to the R_d group who *did not* receive the retention incentive (i.e., control). The same holds for$N_{l'}$, representing the number of customers l' in the R_d group who *did* get the retention incentive (i.e., treatment). We subtract treatment from control such that positive TE means that the treatment is beneficial for the firm (i.e., it reduces churn among that group of customers).

- Second, we split the validation sample on the basis of predicted *LIFT*, creating ten equally-sized groups L_1, L_2, ... , L_{10} with respect to each customer's $LIFT_j$. Similarly, we calculate the treatment effect (TE_{L_d}) in each of the groups L_d, with $d = 1, ..., 10$.

We compute these quantities for two main reasons. First, it helps us understand the extent to which each of the churn metrics relates to the customer's sensitivity to retention actions. For example, are customers with the highest risk of churning the most sensitive to the retention campaign? If that were the case, then we should find that $TE_{R_{10}} \geqslant TE_{R_9} \geqslant ... \geqslant TE_{R_1}$. Second, measuring TE by group also helps identify which groups of customers should and should not be targeted by the firm. For example, if $TE_{R_5} < 0$, then targeting all customers in the R_5 group would likely increase churn.

Step 6: Evaluate the impact of the campaign under different targeting rules

Finally, we evaluate the impact of the retention campaign under two scenarios: (1) if the company based their targeting decisions on the basis of customers' propensity to churn (i.e., *RISK*), as is commonly used in practice and suggested by most papers in the literature. And (2) if the firm selected customer targets based on the incremental effect of the campaign (measured by *LIFT*), as we propose in this research. For example, let us consider the case in which the company were to target 30% of its customers.[14] The goal of this (final) step is to compare what the impact of the same campaign would be if the firm targeted the 30% of customers with the highest *RISK* versus if it targeted the 30% of customers with highest *LIFT*. We proceed as follows:

[14]The decision to target 30% could come from budget limitations, as companies usually operate, or from having calculated the proportion of customers who are expected to respond positively to the campaign, approach we recommend in this research, for which companies have to estimate the heterogeneous treatment effect (*LIFT*).

1. We rank customers on the basis of their $RISK_j$ (descending order), and:

   - For each value of $P = 10\%, 20\%, 30\%, \ldots, 100\%$, we select the top P of customers, which we denote as 'target subgroup'. Note that as P increases, the number of customers in each group increases, with $P = 100\%$ corresponding to the firm targeted the whole customer base.

   - We then estimate the impact of the campaign for each 'target subgroup' by comparing the churn rates across experimental conditions. Specifically, we compute the impact of the campaign as

   $$IC_{R_P} = \frac{1}{M_k} \sum_{k \in \text{Control}} \mathbb{1}\big[(Y_k = 1)\big] - \frac{1}{M_{k'}} \sum_{k' \in \text{Treatment}} \mathbb{1}\big[(Y_{k'} = 1)\big], \tag{6}$$

   where M_k is the number of customers in the top P with respect to RISK who did not receive the retention incentive. Similarly, $M_{k'}$ refers to customers k' in the top P-risk group who did get the retention incentive. As we did with the treatment effect, we subtract treatment from control such that positive IC indicates an effective campaign.

   Because we sorted customers on the basis of $RISK_j$, IC_{R_P} corresponds to the cumulative average of the treatment effects (TE_{R_d}). For example, targeting the customers with the 30% highest $RISK$ corresponds to target those in the top three deciles R_{10}, R_9, and R_8 as defined in step 5. In other words, $IC_{R_{30\%}} = (TE_{R_{10}} + TE_{R_9} + TE_{R_8})/3$.

2. Similarly, we rank the customers on the basis of their $LIFT_j$ (in descending order) and compute IC_{L_P} for all percentiles P.

Note that the impact of the campaign for the $RISK$ and $LIFT$ approaches should be identical for $P = 100\%$, as this case corresponds to the firm targeting *all* customers in the sample.

Because we are analyzing field experiments in which the treatment/control allocation is random, selecting groups of customers based on their predicted $LIFT$ (or $RISK$) does not

suffer from endogenous self-selection because these metrics were obtained solely from customers' pre-campaign behavior. Furthermore, the observed churn behavior of the customers in the validation sample was neither used to estimate the models nor to allocate customers into target subgroups. Therefore, differences between IC_{R_P} and IC_{L_P} are purely based on the actual post-campaign behavior and do not directly rely on any modeling assumptions. In some sense, these differences are "model free."

"Bootstrap" cross-validation — Replicate Steps 1–6

To make sure that the results are replicable and not driven by the specific (random) split in Step 1, we run steps 1 though 6 multiple times. In particular, we generate 1,000 different splits between calibration and validation samples and then summarize the results, reporting the average and standard deviation of the quantities in Equations 5 and 6 across all iterations.

Study 1: Wireless service (Middle East)

The first application corresponds to a wireless provider located in the Middle East. The country where the focal provider (the one we collected the data from) is located is a well-connected market, with more than 4 million subscribers. There are two main players in this market, with our focal provider owning the biggest share of the market.

Intervention: The firm conducted an experiment to test whether giving customers free credit (bonuses) when recharging their amounts affected their likelihood to remain active. Customers in this experiment belong to prepaid plans in which a pre-paid credit is added to the account and gives users the right to make calls, send texts, and download data. In order to keep the account active, customers need to refill their balances within a certain period of time, which depends on the plan they belong to; otherwise, the account is deactivated.

The firm selected customers who have refilled their accounts sometime between one and four weeks prior to the experiment (all customers were active at the moment of the intervention) and had not initiated a call in the week prior to the experiment. Among the 12,137 customers fitting these criteria, the company randomly assigned treatment and control groups. Treated customers (68% of the sample) received a text offering additional credit if

they recharged a specific amount within the 3 days following the intervention. The company then tracked whether the customers were active (or inactive) 30 days after the experiment.

Data and randomization: The company tracks multiple measures of activity such as texts, calls, data uploads/downloads and recharges, as well as the type of (prepaid) plan the customer belongs to. We obtain customers' information for the month prior to the experiment, along with the tenure of each customer (i.e., how long ago she opened the account).

To test whether the randomization succeeded, we compare customers across the two experimental conditions along a set of observed variables. Due to privacy concerns of the focal provider, all the variables are standardized at the population level, then summarized per condition (see Table 1). With the exception of one variable (voice volume), all other variables are not statistically different across conditions, suggesting that the randomization was executed appropriately.[15] In addition to the variables described in Table 1, we also observe several dummy variables the company stores in its database, including location (country area), type of plan the customer belongs to, and other internal segmentation variables, which we include in all our analyses. For this application we have a total of 37 variables.

Study 2: Special interest membership organization (North America)

The second application corresponds to a (subscription-based) membership organization located in North America. This organization offers an annual subscription/membership that gives members the right to use its services (both online and offline services), and also offers them a discount (sometimes as high as 100% discount) to attend events.[16] Each year, one month before a customer's membership is close to expiring, the organization sends out a renewal letter. If the membership is not renewed, the benefits can no longer be received.

[15]Further conversations with the firm managers indicated that the difference in voice volume was a mere coincidence because such variable was never used to select targets and all other variables were equally distributed across experimental groups. The company tracks the usage variables (e.g., data, SMS and voice volume) at different time frames (e.g., within the last week, last two weeks, last 4 weeks, etc.) For brevity sake, here we only report those from the two weeks prior to the experiment while we use all the variables for our analysis.

[16]The organization prefers to remained unidentified. The reader can think about any cultural, professional, or special interest organization that offers annual memberships.

Intervention: The focal organization ran a field experiment that tested whether adding a gift to the renewal communication would increase renewal rates. Because not every subscriber ends the membership at the same time, then experiment was run during five consecutive months. Each month, the company identified the customers who were up for renewal and split them (randomly and evenly) between a treatment group that received a "thank you" gift with the letter and a control group that received only the renewal latter. The intervention was not targeted to any specific type of customer. Rather, every customer who was up for renewal was part of the experiment. At the end of the experiment, we obtained all the information from a random sample of the customers involved in this experiment (N$= 2,100$).

Data and randomization: For each customer in the sample, we observe the month in which the renewal letter was sent, whether the customer renewed her subscription or no, and other demographic and usage characteristics such as tenure in the organization (in number of years), location (which state the member lives in), whether the subscriber attended any organized event (0/1), and whether the subscriber had logged in into the organization website. The company did not target on any of these, or other, characteristics.

As with the first study, we confirm that the randomization was appropriate by comparing the distribution of customers across the two experimental conditions. Table 1 describes the observed variables by group. In addition to the variables presented in Table 1, we create a categorical (dummy) variable capturing whether it was the first renewal occasion for the customer. Historically the organization has observed that first-time renewal rates are systematically lower than those from customers who had been subscribers for more than one year. We also include the interaction between the first year dummy and the four usage variables as well as several dummy variables indicating geographical location of the customer. We observe a total of 50 variables.

Validation results

We start by analyzing customer heterogeneity in the treatment effect by comparing churn rates across experimental conditions for customers with different levels of *RISK* and *LIFT*

(recall Step 5). We first discuss the results for the first Study 1 (wireless) and then compare them with those obtained in Study 2 (membership).

Figure 2a shows, for different levels of *RISK*, the churn rate of customers in each of the experimental conditions of Study 1. The first column on the left (R_{10}) corresponds to the customers identified as being at the highest risk of churning whereas the last column on the right (R_1) corresponds to those at the lowest risk of churning. Comparing churn rates across conditions, we can easily observe the extent to which treatment reduced churn in each of the *RISK*-groups. For instance, the intervention slightly reduced churn among customers in highest risk group, R_{10}, where churn rate is 95.3% for control and 93.7% for treatment (i.e., 1.6 percentage points *reduction* in churn rate). The group R_7 is the group for which the intervention had a greatest impact, reducing churn from 71.2% (control) to 68.3% (treatment), corresponding to a 2.9 percentage points reduction. We also observe that the intervention was not beneficial (i.e., increase churn) among some groups of customers. For instance, the treatment increases churn in groups R_5, R_3, and R_1, with R_3 being the most harmful (churn rate is 9.1% for control and to allow 11.9% for treatment, corresponding to a 2.8 percentage points *increase* in churn rate).

Figure 2b summarizes churn rates when customers are grouped on the basis of their *LIFT*. We highlight two insights: First, the intervention clearly reduced churn among customers with highest *LIFT* (in particular, among L_{10}, L_9, L_8 and L_7 groups), with the differences in churn rates being substantially larger than those observed in the "best" *RISK*-groups. For instance, churn reduced 9.2 percentage points (from 57.0% and 47.8%) among customers in $LIFT_{10}$. Second, the intrinsic churn rate (57.0% and 47.8%) for customers with highest *LIFT* (those in L_{10}) is not necessarily the highest, implying that customers who are more sensitive to the retention efforts are not necessarily at the highest risk of churning.

Insert Figure 2 here

Similarly, we examine churn rates for different levels of *RISK* and *LIFT* in Study 2 (Figures 2c and 2d, respectively). In this case the focal company should not have targeted customers at high risk of churning. In turn, these are the customers for whom the intervention

was most harmful. For example, among the R_{10} group (those whose *RISK* is in the highest decile), the churn rate was 79.4% in the control compared to 82.7% in the treatment (i.e., the intervention *increased* churn by 3.3 percentage points among that group of customers). A similar effect was found for customers in the R_9, R_8 and R_7 groups. On the contrary, churn was reduced among customers who had a lower risk of churning (those in groups R_5, R_4, and R_1). This finding contradicts the conventional wisdom that retention programs should target high-risk customers. Also note that this pattern — non monotonic relationship between the treatment effect and *RISK* — differs from that Study 1, suggesting that the relationship between levels of *RISK* and the response to the intervention is not easily predictable.

On the contrary, when we examine the heterogeneity in treatment effect with respect to customers' *LIFT* (Figure 2d), we find an identical pattern to the previous application. Customers with the highest levels of *LIFT* (L_{10}–L_6) respond positively to the treatment — churn rates are about 5 percentage points lower for treated customers than for control customers — whereas the treatment increases churn among those with lowest levels of *LIFT* (L_4–L_1). To better visualize the differences in churn rates across conditions, and for an easier comparison across studies, Figure 3 shows the magnitude of the treatment effects, TE_{R_d} and TE_{L_d} (i.e., churn rate in the control minus churn rate in the treatment groups) for different levels of *RISK* and *LIFT* for each of the empirical applications. The squares represent the average (across all iterations) of the treatment effects for different levels of *RISK* while the circles correspond to levels of *LIFT*. The dotted line marks the average effect of the campaign, which corresponds to the expected effect of the campaign should the firm target customers randomly. Comparing the results across both studies, customer *LIFT* is a strong discriminatory variable for targeting marketing efforts whereas the patter for *RISK* is rather unclear.[17]

Insert Figure 3 here

[17]We corroborate that the proposed approach not only sorts customers from greater to lower treatment effect, but also provides a consistent estimate of the size of the effect. Please refer to Web Appendix A3.1 for more details.

Impact of the retention campaign if targeting based on RISK or LIFT

We now compare what the impact of the campaign would be if the firm targeted the *same proportion* of customers, selecting them based on either their *RISK* or their *LIFT*. Figure 4 depicts the impact of the campaign under each of the scenarios, assuming the firm targets 10% of customers, 20% of customers, etc. The squares/circles represent the average across iterations, and the bars represent the standard deviation around the mean.[18] As discussed earlier (Step 6), this analysis is equivalent to "accumulate" the treatment effect across deciles. For example, the impact of targeting the top 10% *LIFT* customers equals the treatment effect for the 10$^{\text{th}}$ *LIFT*-decile. The impact of targeting the top 20% *LIFT* customers corresponds to the average of the treatment effects of the 10$^{\text{th}}$ and 9$^{\text{th}}$ *LIFT*-deciles.[19] The straight dotted line corresponds to the impact of the campaign if the company targets customers at random, which is the average treatment effect of this campaign.

Insert Figure 4 here

There are several patterns to note. First, the impact of targeting customers based on *LIFT* decreases as the percentage of customers being targeted increases (i.e., $IC_{L_{10\%}} > IC_{L_{20\%}} > \cdots > IC_{L_{100\%}}$). This pattern should be expected because the *LIFT* approach selects the "best" (i.e., more sensitive) customers first. Therefore, as more customers are targeted, the effectiveness of the campaign should decrease. Second, and most importantly, the proposed approach is substantially more effective than the "at risk" approach. In other words, both firms would have saved more customers if targeted their retention efforts based on customers' *LIFT* than based on customers' *RISK*. For example, with reference to Study 1, if the firm had targeted the 40% customers with highest *RISK*, the retention campaign would have reduced churn by 1.9 percentage points ($IC_{R_{40\%}} = 0.019$). However, is the same proportion of customers had been targeted but selected based on their *LIFT*, the *same campaign* would have caused a 6.0 percentage points churn reduction ($IC_{L_{40\%}} = 0.060$). The

[18] As top deciles increase, more customers are included in each group, hence the error bars become narrower. Web Appendix A3.2 shows the results for one single iteration.

[19] For example, comparing Figures 3 and 4, we have that $IC_{L_{10\%}} = 0.091$ in Figure 4a equals $TE_{L_{10}} = 0.091$ in Figure 3a, $IC_{L_{20\%}} = 0.080$ in Figure 4a corresponds to $(TE_{L_{10}} + TE_{L_9})/2 = (0.091 + 0.069)/2$ in Figure 4a, and so forth.

equivalent result is even more pronounced for Study 2, in which the company would have *increased* churn by 4.4 percentage points if targeting the 40% of customers with highest *RISK* whereas it would have *reduced* churn by 4.3 points if targeting the 40% of customers with highest *LIFT*. In this case the difference is churn reduction between targeting based on *RISK* and based on *LIFT* is high as 8.7 percentage points. Finally, and not surprisingly, both methods would give similar effectiveness if the company decided to target most customers.

Differences between customers' *RISK* and *LIFT*

We further leverage our data and explore the differences between customers' *RISK* and *LIFT*. Specifically, we quantify the level of overlap between customers with highest risk of churning (top *RISK*) and those that are most sensitive to the retention intervention (top *LIFT*). We then investigate which observed characteristics (e.g., metrics of past behavior) are better predictors for each of the two metrics. This analysis holds managerial relevance for several reasons. First, doing so helps identify the most important variables that predict customers' sensitivity to retention interventions. Second, albeit correlational, it adds to the understanding of why certain interventions work better (or worse) on some types of customers. Finally, investigating differences between *RISK* and *LIFT* predictors will inform firms about what types of "at risk" customers should be left alone.

We turn to quantify the level of overlap between the *RISK* and the *LIFT* metrics. The results thus far suggest that we should not expect high levels of overlap between the groups of customers of high (low) *RISK* and those with high (low) *LIFT* — otherwise, the lines for the *RISK* and *LIFT* approaches in Figures 3 and 4 would be similar. We corroborate this pattern by leveraging the results from Step 6 to quantify the level of overlap among the *RISK* and *LIFT* groups. Figure 5 shows, for each size of subgroup (e.g., 10% of sample, 20% of sample), the proportion of customers who overlap between the top *RISK* and the top *LIFT* groups, for each of the studies. The squares represent the percentage of customers in each top P% *RISK* percentile who also belong to the top P% *LIFT* percentile. So, a value of 100% would denote perfect overlap between the groups. That is, the customers identified as

having highest levels of *RISK* also have the highest levels of *LIFT*. On the other hand, the (dotted) 45° line represents the level of overlap if there were no relationship between the two groups (in other words, if the chance of overlap between *RISK* and *LIFT* were random).

Insert Figure 5 here

As Figure 5 illustrates, the relationship between these two metrics is rather weak. In Study 1, among the 10% of customers with highest *RISK*, only 16% of them also belong to the top 10% *LIFT* group. If we look at the 50% top *RISK* customers, only 52% of them belong to the highest *LIFT* group, suggesting that, if this company were to target on the basis of highest *RISK*, more than half of the resources would be allocated to customers who are not very sensitive to the campaign — or, indeed, even to customers who might increase churn as a consequence of the campaign. Regarding Study 2, consistent with the finding that the highest risk customers should not be targeted (Figure 2c), the level of overlap between *RISK* and *LIFT* is not only weak but slightly negative. Only 6% of customers in the top 10% *RISK* belong to the top 10% *LIFT* group. If we look at that metric for the 50% percentile, the level of overlap is a mere 40%,[20] suggesting the inefficiency of targeting customers on the basis on highest risk for this retention campaign.

Finally, we explore which customer characteristics are predictive of customers' *RISK* and *LIFT*. We compute, for each decile ($R_{10}, R_9, ..., R_1$ and $L_{10}, ..., L_1$), the average value of the observed characteristics. In the interest of brevity, we only report the variables that are most relevant for each study (Figure 6a for Study 1 and Figure 6b for Study 2); the full set of results is reported in the Web Appendix A3.3.

Insert Figure 6 here

We start analyzing Study 1. As expected, the patterns between each of the variables and the *RISK* deciles are different from the patterns between those same variables and *LIFT*. For example, consider the variable 'Days no recharge', which represents how long it has been since the customer put money in her account. Customers with longer times are, not surprisingly, more likely to churn in the following month. However, this variable is

[20] This level of overlap is lower than the case where customers are picked at random, in which case the overlap should be 50%.

not predictive of sensitive to the incentive (as captured by the almost flat line between the variable and the *LIFT* deciles). The variable 'Data volume' reveals an interesting pattern. Customers who used low levels of data in the previous week exhibit very a high risk of churn (as represented by the upward relationship between such variable and the *RISK* deciles). Conversely, these customers have negative *LIFT*, implying that if the company decided to send the retention incentive to customers with low data consumption (as they belong to a "high churn" segment), such campaign would likely increase churn. Among the variables selected, only 'Tenure' and 'Last recharge' have similar relationship patterns with *RISK* and *LIFT*. On the contrary, all other usage-related metrics (e.g., revenue in the last week, number of days without consumption, number of days since last recharge) are not predictive of the extend to which the campaign altered behavior. This finding suggests that the intervention employed in this campaign (i.e., sending a text) did not reach customers who were at risk of leaving due to inactivity. And, if the firm wanted to prevent churn among this type of customers, a different type of intervention should be employed and tested. While designing campaigns incentives is beyond the scope of this research, these findings are informative to the firm as to what type of interventions work for which type of customers.

We now investigate the relationship between the observed characteristics and the *RISK* and *LIFT* metrics of Study 2. Consistent with the results from Figure 2c, the majority of the variables show the opposite pattern when predicting *RISK* than when predicting *LIFT*. For example, while customers in their first year of membership (top middle figure) are at the highest risk of churning, they are precisely the ones whose reaction to the intervention was the most harmful for the firm. That is, contrary to the firm's intentions, the intervention encouraged "newer" customers to cancel their subscription. This finding suggests that the intervention not only did not resonate with "newer" members but was perceived negatively. A finding that deserves further investigation is that related to off-line engagement (captured by the variables 'Attendance' and 'Special events'). Whereas these two variables are predictive of customer *RISK*, they are not correlated with the extent to which the intervention affected behavior (i.e., the *LIFT* lines in Figure 6b are flat). It would be interesting to investigate

if such a relationship (or lack of) would differ if the firm ran a retention campaign with an intervention that, for example, highlighted future events.

Summary of results

Combining the results across the two field experiments, we have demonstrated that targeting based on *LIFT* is more effective at reducing customer churn than targeting on the basis of *RISK*. In particular, we find that the same retention campaign would result in a *further reduction* of 4.1 and 8.7 (studies 1 and 2, respectively) percentage points in churn rate, if each focal firm followed our proposed approach instead of the industry standard of targeting customers at the highest risk of churning. This result is consistent across both studies representing two different business settings (wireless/telecom and special-interest organization) and located in two different markets (Asia and North America). Hence, our proposed approach of selecting targets based on their *LIFT* may be generalizable to a variety of proactive churn management programs.

With respect to the question of heterogeneity in treatment effects (Figure 3), we have demonstrated that customers with higher propensity to churn (operationalized by *RISK*) are not necessarily those who are more sensitive to the retention efforts. In turn, we do not find a consistent pattern between customers' risk of churning and their response to the retention action, suggesting that this pattern is likely to be campaign- and context-specific. On the contrary, the pattern between customers predicted *LIFT* and response to the treatment is strong and consistent across both applications.

Finally, our method allowed us to quantify the level of overlap among customers "at risk" and those with higher sensitivity to the marketing intervention as well as to identify which customer characteristics are most relevant to predict each of these metrics. Overall, we find that the overlap between the propensity to churn (i.e., *RISK*) and the sensitivity to the retention campaign (i.e., *LIFT*) is not different from independence (or random overlap). It is important to highlight that this (lack of) relationship between *RISK* and *LIFT* is not driven by our choice of the modeling approach. Unlike parametric methods for binary data (e.g., logistic regression), the random forest estimates the differential impact of the retention

campaign in a nonparametric way. As a result, the magnitude of the impact of the campaign on customer churn does not depend on where in the probability space each customer is located (as it would be if one used a logistic regression, for example). Furthermore, this (lack of) relationship between *RISK* and *LIFT* is not due to the selection of customers eligible for the experiments. In our second application, *all* customers that were up for renewal participated in the experiment. As such, our data covers the full range of *RISK* levels among customers. We acknowledge than in the first study we might not cover the entire range of *RISK* levels because customers with certain characteristics — in theory, at a higher risk of churning — were selected for the experiment. Nevertheless, the strong consistency across the two studies gives us some confidence about the generalizability of this result.[21]

Regarding which variables are "driving" churn (i.e., *RISK*) versus sensitivity to the intervention (*LIFT*), both applications presented evidence of different drivers behind each of the metrics.[22] While the drivers for *RISK* are expected to be more generalizable across contexts, the drivers for *LIFT* are campaign-specific. That is, if the interventions investigated were of a different nature (e.g., giving a new handset versus money incentive, or a price discount versus a thank you gift), we would expect to see different variables being related to the impact of the campaign. Nevertheless, across both applications we found a distinct lack of consistency between the patterns for *RISK* and *LIFT*. Future research should investigate these relationships in the interest of better designing incentives for retention campaigns.

All in all, we have demonstrated that, contrary to the conventional wisdom, proactively targeting high-risk customers might not be an effective strategy to reduce churn because by doing so, firms are wasting resources on customers who are not responsive (or even respond negatively) to the campaign. In other words, our analyses not only demonstrate that half of the retention money is wasted but also identifies *which half*. Consequently, firms should first

[21] We perform a simulation study in which we simulate churn behavior in the context of a randomized intervention. We manipulate the correlation between *RISK* and *LIFT* and summarize the expected outcomes in each scenario. See all details in Web Appendix A3.4. The simulation analyses suggest that the patterns in the first application are consistent with a correlation of 0.2 between *RISK* and *LIFT* constructs while the second application is consistent with the scenario of a correlation of -0.2.

[22] We use "driving" acknowledging that such patterns are only correlational and might not imply causation.

explore customer heterogeneity on the sensitivity to their retention efforts, and then target customers whose sensitivity is the highest, regardless of their intrinsic propensity to churn.

Proactive churn management in a broader context

Contractual and noncontractual settings

The two applications considered in this research were *contractual* settings in which there was a clear metric to capture customer churn. Following Schmittlein, Morrison and Colombo (1987), these are generally called *contractual* settings, term used when the loss of the customer is observed by the firm. On the other hand, the term *noncontractual* is used for settings where the loss of the customer is not observed. While proactive churn management programs, in practice, have been mainly applied to contractual settings (e.g., telecommunications, financial services, utilities, memberships), firms in noncontractual settings (e.g., online games, retailers) can also leverage our proposed approach to select targets in their proactive campaigns. As part of their marketing activities, many of these firms constantly run targeted interventions aimed at increasing activity of "dormant" customers. While these interventions are not called 'proactive churn management,' they are proactive at managing churn in the sense that their goal is to "retain" customers by, for example, encouraging them to make another transaction with the firm.

Extending our approach to these noncontractual settings is straightforward. Building on the notation introduced earlier, noncontractual firms first need to decide how to operationalize the dependent variable (Y_i). Recall that in our case Y_i was defined as 'whether the customer churns'. For example, a noncontractual firm could operationalize Y_i = as whether customer i makes a transaction in the month following the intervention. Then, defining $LIFT_i = P[Y_i|X_i, T_i = 1] - P[Y_i|X_i, T_i = 0]$,[23] the approach proposed in this pa-

[23]Note that when Y_i corresponds to churn, lift was defined as $LIFT_i = P[Y_i|X_i, T_i = 0] - P[Y_i|X_i, T_i = 1$ (i.e., control minus treatment). We define the difference such that positive $LIFT$ represents what is beneficial to the firm, that being more transactions or less churn.

per would identify the customers who are more likely to make a transaction *because* of the intervention.[24]

From *LIFT* to *Value-LIFT*

It is also important to note that churn (or customer retention) is only one measure of interest in the customer relationship. In many business contexts, other behaviors (e.g., consumption) are also important determinants of the value of a customer (Ascarza and Hardie 2013; Lemmens and Gupta 2017). For example, in our second application — the special interest organization — every customer pays the same annual fee, implying that churn is the main differentiator for customer value. On the other hand, there exist settings in which customer revenue directly depends on consumption, implying that some customers will be more valuable than others even if they all had the same churn propensity. Examples of this kind include telecommunications (like our first empirical application), financial services, energy utilities, health care, or online games (with in-app purchases).[25] In the latter case, proactive retention campaigns should not only focus on retaining customers but also retaining high value customers, and when possible, increasing the value of their current customers.

More generally, the ultimate goal of any marketing intervention should be to increase the expected value of customers — i.e., not only considering the revenues in the next period, but accounting for future periods as well. That is, campaigns should be targeted to maximize "profit lift" (Lemmens and Gupta 2017), defined as the increase in expected customer lifetime value (CLV) depending on whether the customer is targeted or not. In other words, and with reference to the notation introduced earlier, the goal of the campaign would be to maximize

$$\textit{Value-LIFT}_i = \mathbb{E}[CLV_i|X_i, T_i = 1] - \mathbb{E}[CLV_i|X_i, T_i = 0], \tag{7}$$

where CLV_i is now a continuous metric representing the discounted value of the (post-campaign) customer profitability. While causal random forests can be easily applied to the

[24] Previous work in marketing — Gönül, Kim and Shi (2000) in the context of catalog mailing, Bodapati (2008) in the context of product recommendations — has already highlighted the importance of targeting based on the incremental effect of targeted marketing interventions. Unlike previous work, our experimental approach does not need any distributional assumption about the propensity to incur in the behavior of interest (e.g., make a purchase, churn) and does not require multiple observations per customer in order identify which customers (based on their observed characteristics) should be targeted.

[25] Another dimension in which some customers can be more valuable than others is by their level of influence on their connections (Nitzan and Libai 2010; Ascarza, Ebbes, Netzer and Danielson 2017). While our approach could, potentially, incorporate customer influence, we do not consider that case in our application as such data are not available to us.

case of a continuous dependent variable,[26] the real challenge of estimating *Value-LIFT*$_i$ is that one needs a very long time horizon to estimate the impact of the marketing intervention in consumer behavior, or alternatively, strong assumptions about the impact of the marketing campaign need to be made.

For example, in the case of a contractual relationship we can express CLV_i as

$$CLV_i = \sum_{t=1}^{\infty} \frac{\lambda_{it} \prod_{\tau=1}^{t} r_{i\tau}}{(1+d)^t}, \tag{8}$$

where λ_{it} is the profit per customer in each period, r_{it} is the probability that a customer renews in each period, and d is the discount rate. In order to simplify the expression of CLV, most past work in marketing has assumed constant margins and retention probabilities.[27] However, the main purpose of a retention campaign is to *alter* the probability that a customer will renew, making assumption about constant retention rates problematic. Furthermore, the retention campaign might also affect consumption or expenditure. For example, a retention campaign might cause a "delight" factor (Blattberg, Kim and Neslin 2008), increasing customer's profitability. As a result, while one could simplify the above formula to make the estimation of *Value-LIFT*$_i$ tractable, one needs to be careful about the validity of the assumptions being made.[28]

This is not to say that estimating *Value-LIFT* is impossible; if one had exogenous variation in retention activities and observed individual retention and expenditure for several periods after the campaign, it would be possible to model the impact of the campaign in

[26]The `uplift` R package allows for continuous Y_i if one uses $k-$nearest neighbors instead of random forests. Note, however, that the accuracy of $k-$nearest neighbors approach is likely to decrease as many covariates are incorporated in to the model. Alternatively, one could easily adapt the causal random forest algorithm to accommodate a continuous dependent variable.

[27]Assuming constant margins ($\lambda_{it} = \lambda_i$) and constant retention probabilities ($r_{it} = r_i$), we obtain that $CLV_i = \sum_{t=1}^{\infty} \frac{\lambda_i r_i^t}{(1+d)^t}$, which can be further simplified to $CLV_i = \frac{\lambda_i(1+d)}{(1+d-r_i)}$ (see Fader and Hardie (2012) for derivations and for a detailed discussion about different CLV formulae). Then, we can simplify the expression in (7), if we further assume that that the retention campaign affects customers' propensity to churn in the current period, but not their future expenditure and retention propensities. In that case, the marketing intervention should target customers for whom *Value-LIFT*$_i = \frac{\lambda_i(1+d)}{(1+d-(1-\mathbb{P}[Y_i|X_i,T_i=0]))}\big(\mathbb{P}[Y_i|X_i,T_i=0] - \mathbb{P}[Y_i|X_i,T_i=1]\big)$ is the highest. Note that the second term of the formula corresponds to the *LIFT*$_i$ metric as defined in (1) and estimated as described in Section 1.

[28]In Web Appendix A3.5 we show the results of targeting a retention campaign on the basis of a restricted, very conservative, version of *Value-LIFT* which only takes into account the period after the campaign.

future behavior, which could be then incorporated in a CLV framework. Ideally, one should model such an impact in an integrated model for consumption and retention (e.g., Ascarza and Hardie 2013) to capture not only the impact of the intervention in each behavior but also the possible interdependencies between those two processes. This approach would be similar to the one suggested by Braun, Schweidel and Stein (2015) for non-contractual businesses, where the authors estimate the differences in discounted expected residual transactions (DERT) depending on the customer requested level of service.

Conclusion

In this paper we propose a different approach for proactive churn management programs. Contrary to what past research and marketing practice have suggested, we claim that proactive churn management programs should not necessarily be targeted to customers who are at the highest risk of churning. Rather, they should conduct pilot field experiments to model customer heterogeneity in the response to the retention incentive and target only customers whose propensity to churn will decrease *in response to* of the intervention.

Combining data from two field experiments with machine learning techniques, we have empirically demonstrated the superiority of our proposed approach compared to the current practice of targeting customers with the highest risk of churning. We show that firms could further reduce customer churn by focusing their retention efforts on the customers identified as having highest sensitivity to the marketing intervention. In particular, we find that the same campaign would reduce churn by an *additional* 4.1 and 8.7 percentage points (Studies 1 and 2, respectively) relative to the standard practice of targeting customers at highest risk.

In addition to its effectiveness in reducing churn, our proposed method has other desirable characteristics that facilitate its use among practitioners. First, the method is scalable to large customer populations and large set of covariates. Second, the method is estimated using existing R packages that are freely available. Third and most importantly, the method can be applied to a wide variety of business context where retaining customers is a concern. In particular, only two conditions are needed for a business setting to leverage the insights

from our research: (1) the company observes customer behavior at the individual level, and (2) the firm is capable to interact with customers in a one-on-one basis (i.e., they can run individually-targeted campaigns). Examples of these business contexts include credit card companies, software providers, online and offline subscriptions (e.g., The Economist, Amazon Prime), leisure memberships (e.g., museums, aquariums, ski resorts), and virtually, any context in which the firm keeps track of individual behavior. Compared to current practice, our approach requires an additional step, a randomized market test. Implementing such a step is within the capabilities of firms that are already running proactive churn management programs as they already have the capacity to target and track behavior individually.

Furthermore, the current research highlights the importance of understanding customer heterogeneity in the response to marketing actions, and in particular, the use of "pilots" or A/B tests to better understand such heterogeneity. We encourage firms to broaden the use of randomized experiments and leverage those data to better understand the heterogeneity in response to the marketing actions. Put differently, we recommend marketers/analysts/researchers to look beyond the average effect of campaigns, and leverage the observed heterogeneity in customers' responses to those campaigns to inform future decisions. More broadly, this research adds to the growing body of literature on the use of big data and supervised machine learning methods to move beyond prediction and inform decisions/policy (Athey 2017).

We acknowledge that the proposed method does not explicitly incorporate competitors' actions. As highlighted by Subramanian, Jagmohan and Zhang (2013), the firm's most valuable customers might easily be the ones that competitors seek to poach, making those customers most responsive to retention efforts (provided the offered incentive compensates the competitive alternatives), while those with lower value to the firm might not be as attractive to competitors, hence might be the most insensitive to retention actions. On the other hand, as documented by Du, Kamakura and Mela (2007), customers who have low levels of expenditure with the focal firm might be spending most of their share of wallet with competing firms, potentially making these customers more sensitive to incentives from the focal firm. Because our approach measures the sensitivity to the retention incentive, we

are (implicitly) capturing these effects; however, we are not isolating the heterogeneity in sensitivity that is driven by the competitors' actions. Understanding those drivers would be interesting for firms as they would be able to focus on customers that are sensitive to their actions and not necessarily sensitive to competitors' offers (Musalem and Joshi 2009).

Even though we were able to collect experimental data from two different contexts — adding to the generalizability of our findings — our empirical approach imposes some limitations. First, the churn rates observed in our two contexts were similar in magnitude. While we anticipate/speculate that the proposed approach is beneficial regardless of the churn rate observed in the market, we acknowledge that the expected benefit of using the proposed approach might be smaller in settings where churn rates are very low.[29] Second, we apply our approach to a single retention campaign (per application), whereas firms typically implement multiple campaigns as part of their retention efforts. For example, wireless providers as well as firms in the financial services (e.g., insurance companies, banks) continuously implement proactive campaigns, targeting customers whose contracts are close to expire. Other companies (e.g., arts, sports, and special interest memberships) generally run campaigns in a more ad-hoc way, e.g., if they observe that their retention rates have recently decreased. It would be interesting to analyze multiple campaigns from the same company (or a similar pool of customers) so we can learn more about how to leverage insights obtained from previous campaigns. An ideal scenario would be to analyze the case of the same company testing different incentives. Applying our approach to such setting would identify *which types of incentives* should be sent to *what types of customers*.

Another aspect that deserves more attention is the length of the assessment period. In both applications we used one month to measure the impact of the retention intervention because that was the timing each of the collaborating companies used. Longer assessment periods would allow the researcher to measure long-term effects of retention incentives and potentially identify the best targeting rules for optimizing both short- and long-term outcomes.

[29] See Web Appendix A3.6 for an exploration of this issue using simulations.

Finally, several methodological aspects of our approach merit further investigation. For example, what is the optimal size for a retention pilot? In our validation analyses we used half of the available data as a calibration sample (replicating what the pilot would be) for convenience and for consistency across the two studies. However, a smaller sample size might have been enough to rank new customers in an effective way, implying that more churn would be avoided if the "remaining" customer sample is larger. Similarly, how stable (over time) is the heterogeneity in sensitivity to the retention action? In practice, a company would implement the pilot, run the analysis and then run the real campaign. That is, there will be a 1- or 2-month gap between the calibration and validation data. While there are not obvious reasons why the relationship between the covariates and the sensitivity to the intervention would change over time, it would useful to empirically investigate this question. We hope that future research will address these and other related issues.

References

Accenture Analytics (2014), Nordic Telco: Analytics Help Reduce Churn and Improve Marketing Campaigns (accessed July 13, 2017), https://www.accenture.com/us-en/success-nordic-telco-analytics-marketing-campaigns.

Ansari, Asim and Carl Mela (2003), E-customization. *Journal of Marketing Research* 40(2), 131–145.

Ascarza, Eva, Peter Ebbes, Oded Netzer and Matt Danielson (2017), Beyond the Target Customer: Social Effects of CRM Campaigns. Forthcoming at the *Journal of Marketing Research*.

Ascarza, Eva, Peter S. Fader, and Bruce G.S. Hardie (2017), Marketing Models for the Customer-Centric Firm. *Handbook of Marketing Decision Models*, edited by Berend Wierenga and Ralf van der Lans, Springer.

Ascarza, Eva and Bruce G.S. Hardie (2013), A joint model of usage and churn in contractual settings. *Marketing Science* 32(4), 570–590.

Ascarza, Eva, Raghuram Iyengar, and Martin Schleicher (2016), The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research* 53(1), 46–60.

Ascarza, Eva, Scott A. Neslin, Oded Netzer, Zachery Anderson, Peter S. Fader, Sunil Gupta, Bruce G.S. Hardie, Aurélie Lemmens, Barak Libai, David Neal, Foster Provost, Rom Schrift (2017), In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. Available at SSRN:2903548.

Athey, Susan and Guido W. Imbens (2016), Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27), 7353–7360.

Athey, Susan (2017), Beyond Prediction: Using Big Data for Policy Problems. *Science*, 355(6324), 483–485.

Bauer, Connie L. (1988), A direct mail customer purchase model. *Journal of Direct Marketing* 2(3), 16–24.

Bayer, Judy (2010), Customer segmentation in the telecommunications industry. *Journal of Database Marketing & Customer Strategy Management* 17(3-4) 247–256.

Berson, Alex, Stephen Smith and Kurt Thearling (2000), *Building Data Mining Applications for CRM.* McGraw-Hill.

Blattberg, Robert C., Byung-Do Kim, Scott A. Neslin (2008), *Database Marketing: Analyzing and Managing Customers.* Springer, New York, NY.

Bodapati, Anand V. (2008), Recommendation systems with purchase data. *Journal of Marketing Research* 45(1), 77–93.

Bolton, Ruth N. (1998), A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science* 17(1), 45–65.

Bolton, Ruth N. and Katherine N. Lemon (1999), A dynamic model of customers' usage of services: Usage as an antecedent and consequence of satisfaction. *Journal Marketing Research* 36(2), pp.171–186.

Braun, Michael, David A. Schweidel, and Eli M. Stein (2015), Transaction attributes and customer valuation. *Journal of Marketing Research*, 52(December), 848–864.

Breiman, L. (2001), Random Forests. *Machine Learning* 45, 5–32.

Cameron N. (2013), How predictive analytics is tackling customer attrition at American Express, *CMO*, Accessed 13 July 2017.

Du, Rex Yuxing, Wagner A. Kamakura, and Carl F. Mela (2007), Size and share of customer wallet. *Journal of Marketing* 71(2), 94–113.

Fader, Peter S. and Bruce G.S. Hardie (2012), Reconciling and Clarifying CLV Formulas. Available at `http://brucehardie.com/notes/024`. Last accessed April 24, 2017.

Feller, Avi and Chris C. Holmes (2009), Beyond toplines: Heterogeneous treatment effects in randomized experiments. Unpublished manuscript, Oxford University.

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

Ganesh, Jaishanker, Mark J. Arnold, and Kristy E. Reynolds (2000), Understanding the customer base of service providers: an examination of the differences between switchers and stayers. *Journal of Marketing* 64(3), 65–87.

Gensch, Dennis H. (1984), Targeting the switchable industrial customer. *Marketing Science* 3(1), 41–54.

Gönül, Füsun F., Byung-Do Kim, and Mengze Shi (2000), Mailing smarter to catalog customers. *Journal of Interactive Marketing* 14(2), 2–16.

Gruen, Thomas W, John O. Summers and Frank Acito (2000), Relationship marketing activities, commitment, and membership behaviors in professional associations. *Journal of Marketing* 64(3), 34–49.

Guelman, Leo, Montserrat Guillén and Ana M. Pérez-Marín (2012), Random forests for uplift modeling: an insurance customer retention case. *In Modeling and Simulation in Engineering, Economics and Management* 123–133. Springer Berlin Heidelberg

Guelman, Leo, Montserrat Guillén and Ana M. Pérez-Marín (2015), Uplift random forests. *Cybernetics and Systems* 46(3-4), 230–248.

Gupta, Sunil, Donald R. Lehmann, and Jennifer Ames Stuart (2004), Valuing customers. *Journal of Marketing Research* 41(1), 7–18.

Hadden, John, Ashutosh Tiwari, Roy Roy and Dymitr Ruta (2007), Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research* 34(10), 2902–2917.

Huang, Bingquan, Mohand T. Kechadi and Brian Buckley (2012), Customer churn prediction in telecommunications. *Expert Systems with Applications* 39(1), 1414–1425.

Jaroszewicz, Szymon (2016), Uplift Modeling. *Encyclopedia of Machine Learning and Data Mining.* Springer.

Lemmens, Aurélie and Christophe Croux (2006), Bagging and boosting classification trees to predict churn. *Journal of Marketing Research* 43(2), 276–286.

Lemmens, Aurélie and Sunil Gupta (2017), Managing Churn to Maximize Profits. Available at SSRN:2964906.

Lemon, Katherine N., Tiffany B. White and Russell S. Winer (2002), Dynamic customer relationship management: Incorporating future considerations into the service retention decision. *Journal of Marketing* 66(1), 1–14.

Linoff, Gordon S. and Michael J.A. Berry (2011). *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons.

Musalem, Andrés and Yogesh V. Joshi (2009), Research note-How much should you invest in each customer relationship? A competitive strategic approach. *Marketing Science* 28(3), 555–565.

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu and Charlotte H. Mason (2006), Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* 43(2), 204–211.

Ngai, Eric W., Li Xiu and Dorothy C. Chau (2009), Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications* 36(2), 2592–2602.

Nitzan, Irit and Barak Libai (2011), Social effects on customer retention. *Journal of Marketing* 75(6), 24–38.

Provost, Foster and Tom Fawcett (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

Radcliffe, Nicholas J. and Patrick D. Surry (1999), Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI.* Edinburgh, Scotland.

Risselada, Hans, Peter C. Verhoef and Tammo H. Bijmolt (2010), Staying power of churn prediction models. *Journal of Interactive Marketing* 24(3), 198–208.

Rossi, Peter, Robert E. McCulloch and Greg M. Allenby (1996), The value of purchase history data in target marketing. *Marketing Science* 15(4), 321–340.

Rubin, Donald B. (2011), Causal inference using potential outcomes. *Journal of the American Statistical Association* 100(469), 322–331.

Rzepakowski, Piotr, and Szymon Jaroszewicz (2012), Decision trees for uplift modeling. *Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE.*

Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), Counting Your Customers: Who-Are They and What Will They Do Next?. *Management science* 33(1), 1–24.

Schweidel, David A., Eric T. Bradlow, and Peter S. Fader (2011), Portfolio dynamics for customers of a multiservice provider. *Management Science* 57(3), 471–486.

Siegel, Eric (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die.* John Wiley & Sons.

Sołtys, Michal, Ssymom Jaroszewicz and Piotr Rzepakowski (2015), Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery* 29(6), 1–29.

Subramanian, Upender, Jagmohan S. Raju, and Z. John Zhang (2013), The strategic value of high-cost customers. *Management Science* 60(2) 494–507.

Verbeke, Wouter, Karel Dejaeger, David Martens, John Hur, and Bart Baesens (2012), New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229.

Wager, Stefan and Susan Athey (2017), Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Forthcoming at the *Journal of the American Statistical Association.*

Wikipedia, The Free Encyclopedia, s.v. "Customer Attrition," (accessed May 12, 2017), https://en.wikipedia.org/wiki/Customer_attrition.

TABLES and FIGURES

Study 1: Wireless provider	Control (N = 3,857)	Treatment (N = 8,280)	Difference p-value
Tenure	0.002	−0.001	0.881
Consecutive days with no recharge	0.015	−0.007	0.256
Days since last recharge	−0.013	0.006	0.317
Revenue from last recharge	−0.003	0.001	0.816
Consecutive days with no outbound usage	0.018	−0.008	0.186
Data volume last two weeks (in logs)	−0.007	0.003	0.625
SMS volume last two weeks (in logs)	−0.017	0.008	0.200
Voice volume last two weeks (in logs)	0.043	−0.020	0.001

Study 2: Membership organization	Control (N = 1,056)	Treatment) (N = 1,044)	Difference p-value
Tenure	0.013	−0.013	0.565
Attendance (binary)	0.311	0.298	0.527
Online activity (binary)	0.413	0.401	0.591
Download activity (binary)	0.164	0.143	0.180
Special interest attendance (binary)	0.050	0.058	0.460

Table 1: Randomization check for the data from the two studies. All continuous variables were first standardized then summarized across conditions.

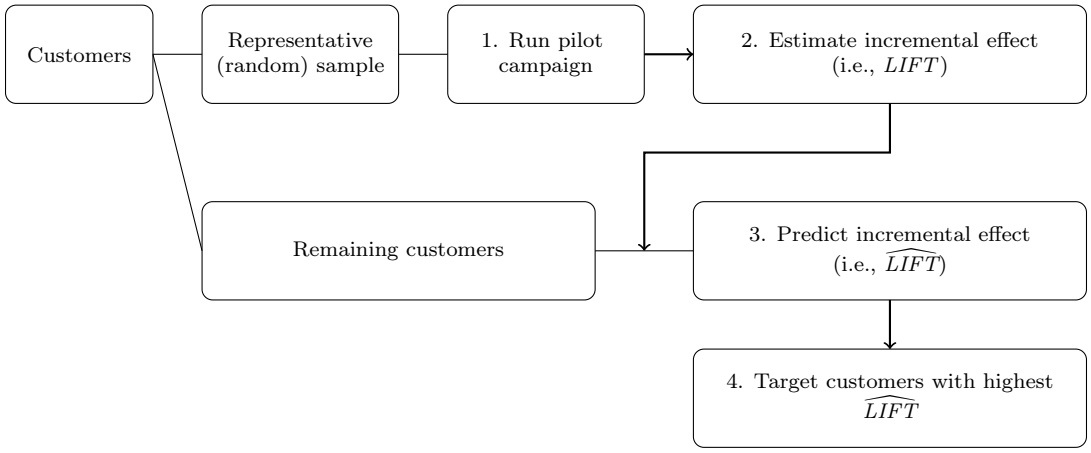
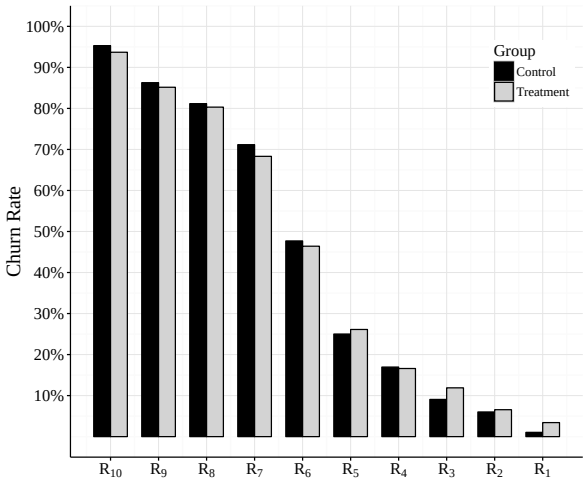
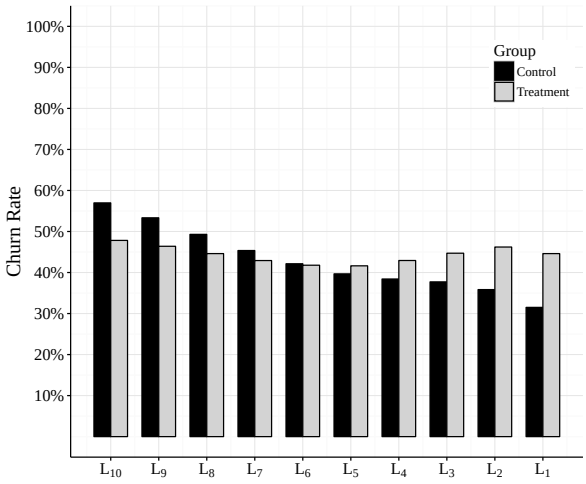


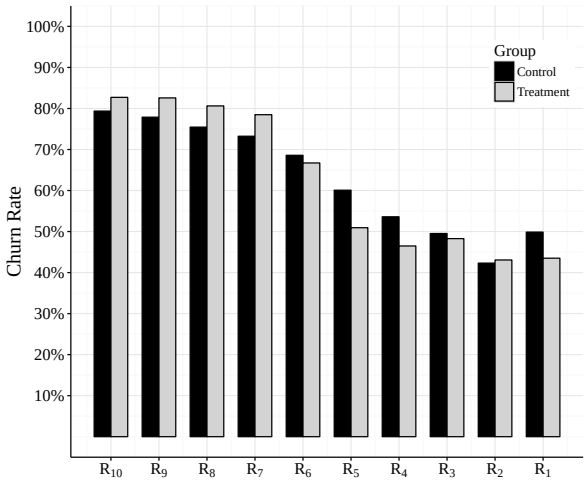
Figure 1: Proposed approach to select customer targets in proactive churn management programs.



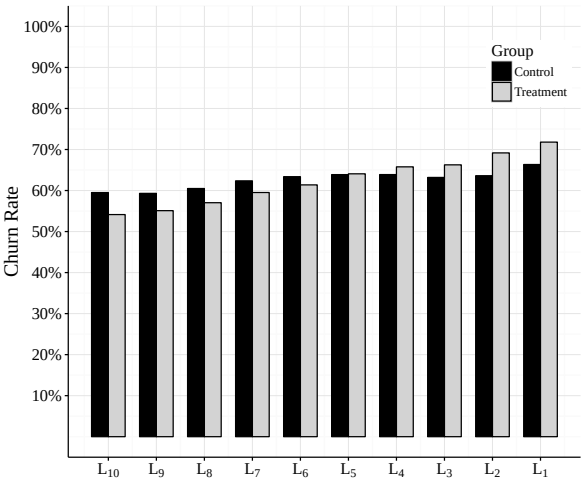
(a) [Study 1] By levels of *RISK*



(b) [Study 1] By levels of *LIFT*

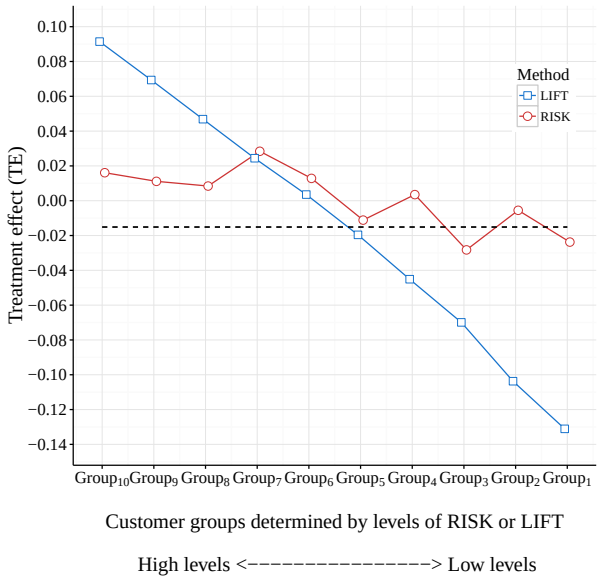


(c) [Study 2] By levels of *RISK*

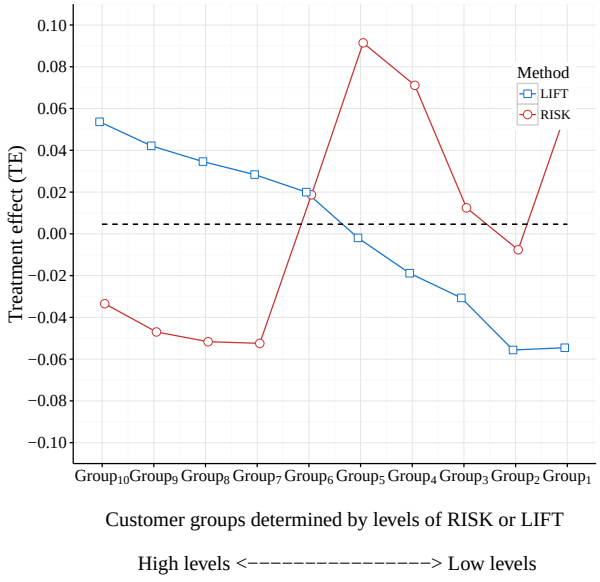


(d) [Study 2] By levels of *LIFT*

Figure 2: Heterogeneity in treatment effects. We compare differences in churn rates across conditions, when targeting customers with different levels of churn propensity (i.e., *RISK*) vs. targeting customers with different levels of sensitivity to the retention intervention (i.e., *LIFT*)

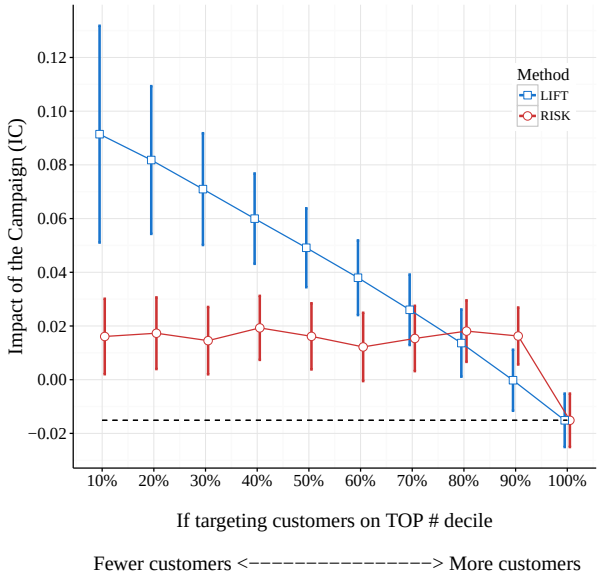


(a) [Study 1] Treatment effect (TE) by deciles

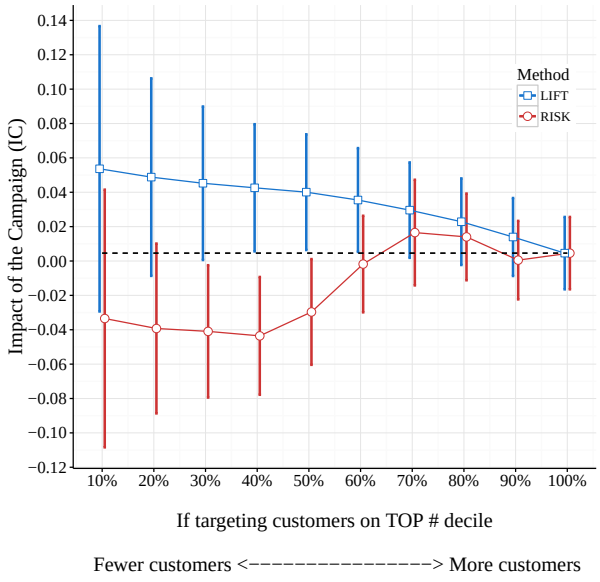


(b) [Study 2] Treatment effect (TE) by deciles

Figure 3: Treatment effect (TE) for different group deciles, depending on whether customers are grouped by levels of *RISK* (represented by the squares) or *LIFT* (represented by the circles). The dotted (straight) line corresponds to the average effect of the campaign if the firm targeted randomly.

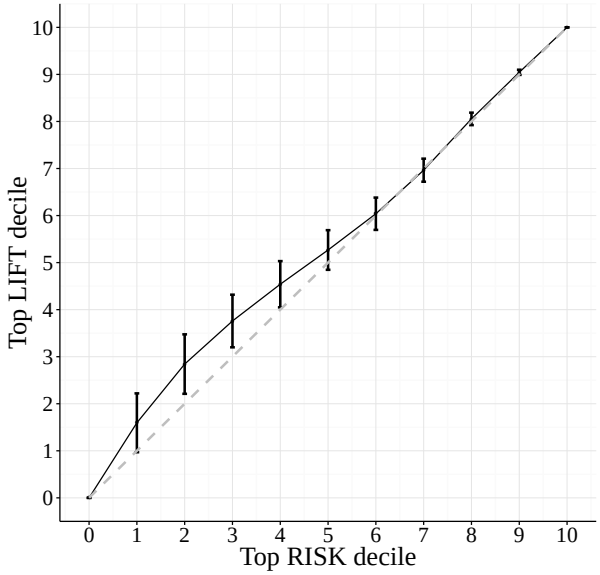


(a) [Study 1] Impact of the campaign (IC) by top deciles

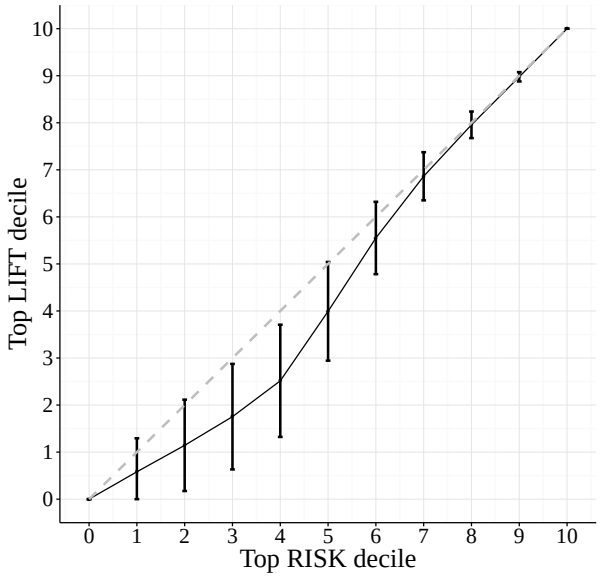


(b) [Study 2] Impact of the campaign (IC) by top deciles

Figure 4: Impact of the campaign (IC) under different scenarios. *RISK* assumes the company targets customers with higher levels of risk of churning. *LIFT* assumes that the company targets customers with high levels of sensitivity to the retention campaign. The dotted (straight) line corresponds to the impact of the campaign if all customers were targeted.

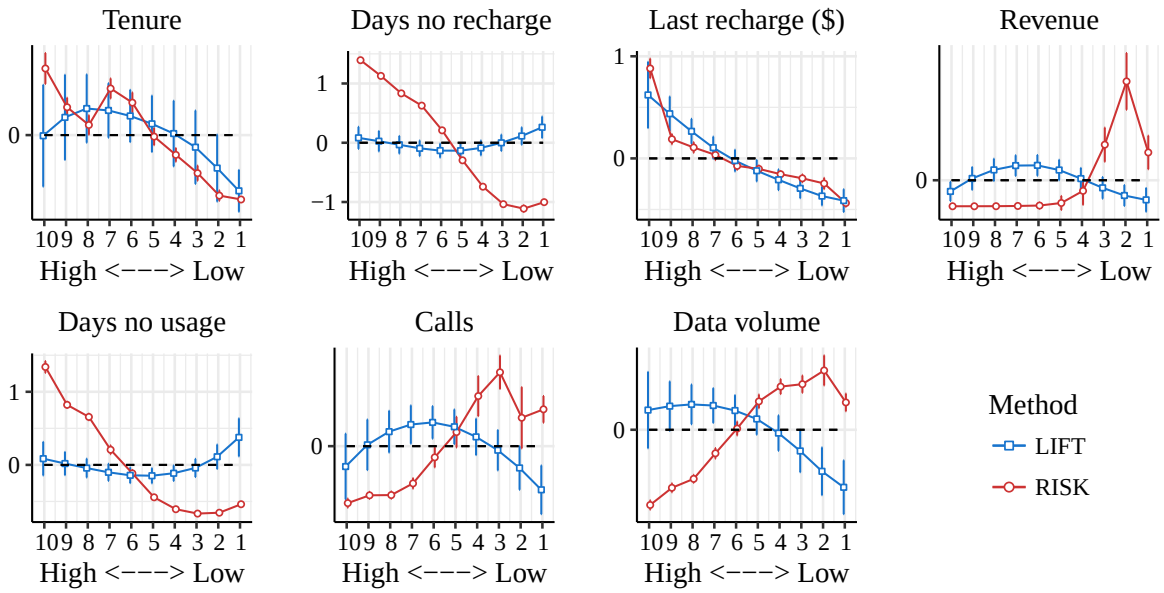


(a) [Study 1] Overlap between *RISK* and *LIFT*

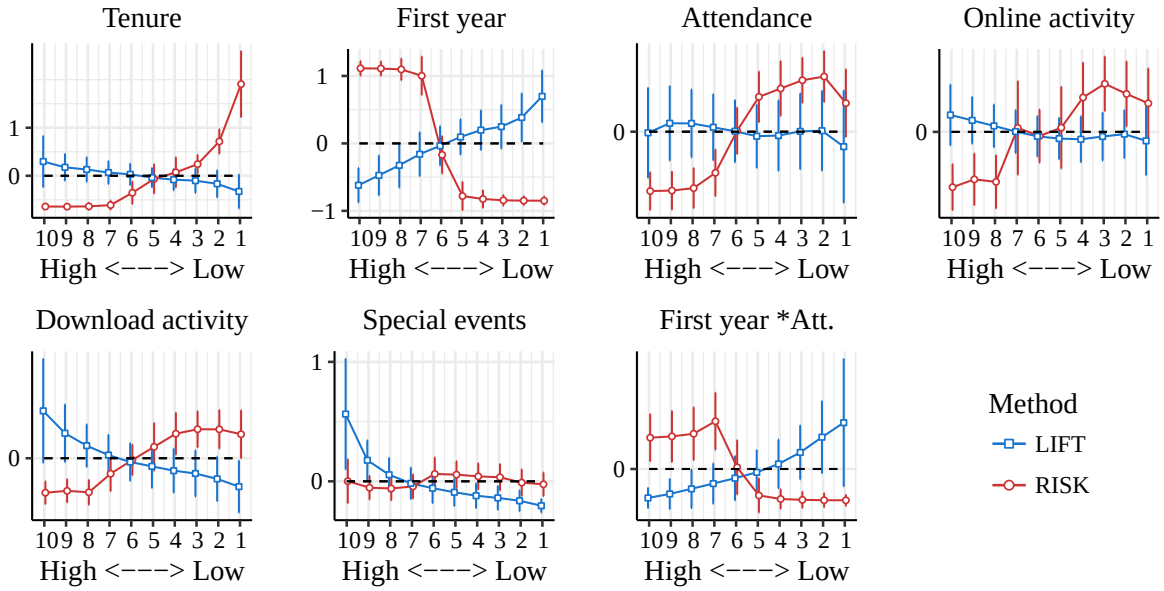


(b) [Study 2] Overlap between *RISK* and *LIFT*

Figure 5: Level of overlap across groups defined by top *RISK* deciles vs. top *LIFT* deciles. The (dotted) 45° line represents the level of overlap if there was no relationship between the two groups.



(a) [Study 1] Average levels of each observed variable by levels of *LIFT* and *RISK*



(b) [Study 2] Average levels of each observed variable by levels of *LIFT* and *RISK*

Figure 6: Average value of each of the observed characteristics for each decile ($R_{10}, R_9, ..., R_1$ and $L_{10}, ..., L_1$). Squares/circles represent the average across iterations and the bars represent the standard deviation.

Retention futility: Targeting high risk customers might be ineffective

Eva Ascarza

Web Appendix

In this appendix we present a set of additional results that were not incorporated in the main manuscript due to space limitations.

A1 Details about *LIFT* and *RISK* models

A1.1 Uplift (i.e., *LIFT*) model

We estimate the *LIFT* model using the uplift random forest algorithm proposed by Guelman et al. (2015), which combines approaches previously used for tree-based uplift modeling (Rzepakowski and Jaroszewicz 2012) with machine learning ensemble methods (Breiman 2001). Like traditional random forests, this algorithm grows an ensemble of trees, each of them built on a (random) fraction of the data. Each tree is grown by randomly selecting a number of variables (among all the available covariates) for splitting criteria.

The trees grow as follows: First, the split rule is chosen to maximize a measure of distributional divergence on the treatment effect (Rzepakowski and Jaroszewicz 2012). In other words, each split (or partition of the data) maximizes the difference between the differences in churn probabilities between treatment and control individuals in each of the two resulting subtrees. Second, each tree will keep growing until the average divergence among the (resulting) subtrees is smaller than the divergence of the parent node. More specifically, let CR_Ω be the churn rate (i.e., proportion of customers churning) in a partition of the population Ω. We denote Ω^t and Ω^c the group of treated and control individuals in that partition, and Ω_1 and Ω_2 the subtrees resulting from splitting Ω. For ease of illustration, let as consider Euclidian distance as divergence measure. For each tree of the ensemble, the algorithm works as follows:

- First, the algorithm picks the split such that $(CR_{\Omega_1^t} - CR_{\Omega_1^c})^2 - (CR_{\Omega_2^t} - CR_{\Omega_2^c})^2$ is maximized.

- Second, the tree stops growing when $\sum_{i=1,2}(CR_{\Omega_i^t} - CR_{\Omega_i^c})^2)/2 < (CR_{\Omega^t} - CR_{\Omega^c})^2$.

Once all the trees are grown, the predicted treatment effect is obtained by averaging the 'uplift' predictions across all trees of the ensemble, which corresponds to the expected treatment effect given the observed covariates.

Regarding divergent criteria, we tested (1) the Kullback-Leibler (KL) distance or Relative Entropy, (2) the L1-norm divergence, and (3) the Euclidean distance.[1] While they all provided similar performance, L1 did marginally better for the first application and KL did slightly better for the second application. The results reported in the manuscript use the best fitting criteria for each application. (We replicated the full analysis using the other metrics and obtained very robust results.) Regarding the number of trees, we vary the number of trees from 10 to 200 in intervals of 10. The (out-of-sample) model fit notably increased as the number of trees increased, with a marginal improvement after having reached 80–100 trees. Hence, we chose 100 trees for both applications. Finally, to avoid having very few observations in a final node (which could result in unstable results due to outliers), we set the minimum criteria to split to 20 observations.

The R code used for the empirical application is made available as a supplemental file.

A1.2 Churn (i.e., *RISK*) model

We tested multiple approaches to estimate the churn scoring model, including GLM, random forests, and SVMs. To select the best *RISK* model we perform a 10-fold cross-validation in which the calibration data is randomly partitioned into 10 equal sized subsamples such that 9 subsamples are used as training data and the remaining subsample is used for testing the model. The cross-validation process is repeated 10 times such that each subsample is used once as the testing data. Importantly, note that we do not use the validation sample

[1] See Rzepakowski and Jaroszewicz (2012) for a description of such metrics.

(i.e., the 50% of customers selected in Step 1) to evaluate the model performance or as any source for model selection. As metric for accuracy we use the area under the curve (AUC) of the receiver operating characteristics (ROC). The best performing method was the LASSO approach combined with a GLM model, which provides an AUC of 0.907 for the first empirical application and and AUC of 0.658 for the second application. Following Tibshirani (1997), we standardized all variables before estimating the model.

A2 Robustness of the results with different specifications of the *RISK* model

A2.1 Increasing the number of observations for the *RISK* model

The impact of targeting based on *RISK* or *LIFT* ultimately depends on the accuracy of the models used to predict customer RISK or LIFT. For example, even if customers should be targeted based on *LIFT*, it could be possible that our *LIFT* model is not good enough to accurately predict customers' *LIFT*, making it impossible for us to show such a relationship in the data. (Ditto for the *RISK* approach.) Therefore, given that we calibrate RISK and LIFT models (Step 3) using different sample sizes, it could be possible that the *LIFT* approach dominates the *RISK* approach because the latter model is calibrated on a smaller sample, making it, potentially, less accurate. This is unlikely given the great accuracy of the *RISK* model (as reported in Section A1.2, the AUC for the first and second applications were 0.907 and 0.658, respectively).

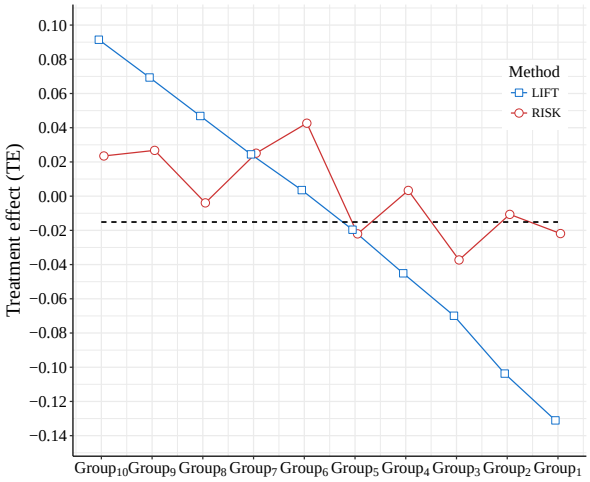
Nevertheless, we corroborate empirically that the superior performance of the LIFT approach is not driven by the size of the data used to calibrate the RISK model. In particular, we replicate the main analysis (Section 4.3) increasing the size of the data used in the RISK estimation (i.e., altering Step 3 in Figure 3). More specifically, we do the following:

1. We calibrate the RISK model (Step 3) using all control observations (3,587 observations for the first application and 1,056 for the second application). The AUC using the full sample was 0.916 for the first application and 0.681 for the second application.

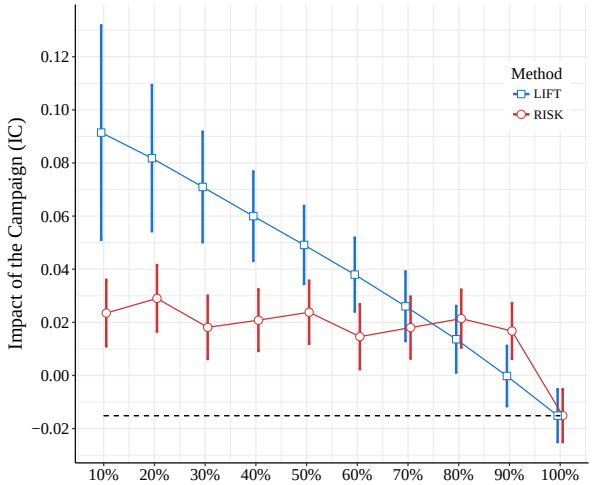
2. Using that model, we predict RISK for the observations in the validation sample (Step 4). Note that we are using some of the observations twice, once to calibrate the model and then to predict RISK, thus increasing the accuracy of the RISK model.

3. We evaluate the effect of the retention campaign by deciles of RISK (Step 5).

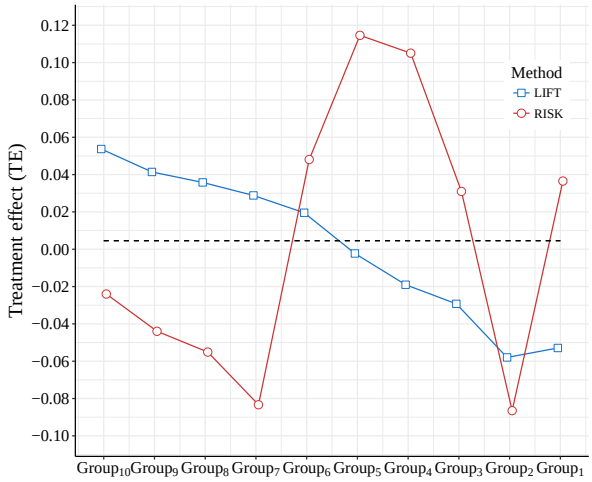
We then compare the effect of the retention campaign for the RISK (overly-accurate approach) with the LIFT (as obtained in the main manuscript). Below we recreate the figures appearing in the main manuscript corresponding to the heterogeneity in treatment effect (Figures 3a and 3b), the impact of the campaign (Figures 4a and 4b), and the level of overlap between the two metrics (Figures 5a and 5b)). As the figures show, the results remain unchanged, verifying that the superiority of the LIFT approach is not driven by the difference in sample size when calibrating each of the models.



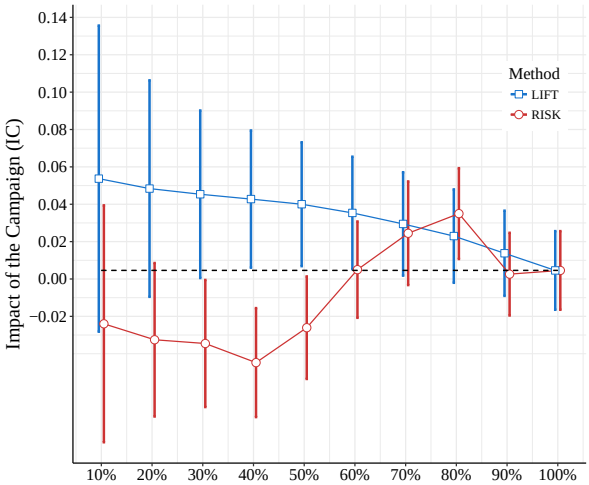
(a) [Study 1] Treatment effect (TE) for different group deciles



(b) [Study 1] Impact of the campaign under different scenarios



(c) [Study 2] Treatment effect (TE) for different group deciles

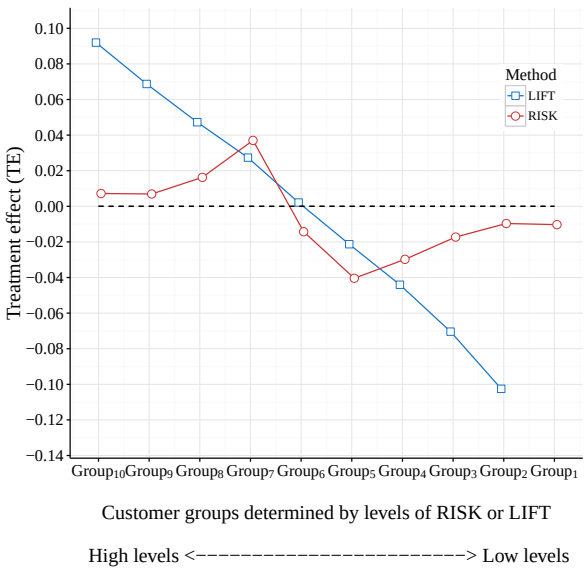


(d) [Study 2] Impact of the campaign under different scenarios

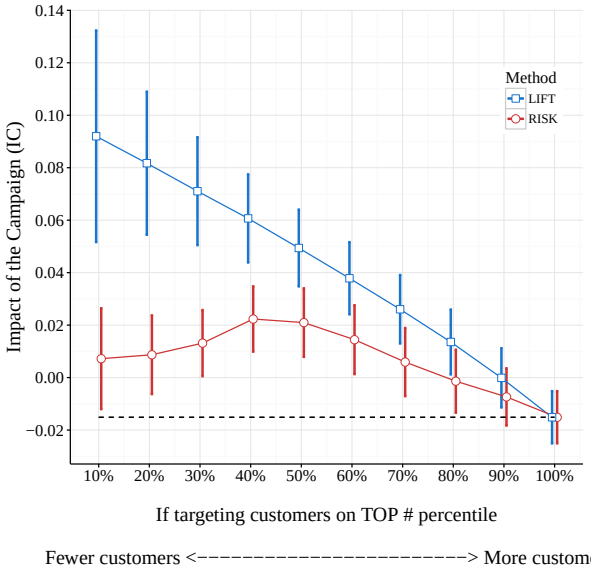
Figure A1: Replication of treatment effect (TE) and Impact of the campaign (IC) results using the full sample to calibrate the *RISK* model

A2.2 Using the same model approach (i.e., random forest) to estimate $RISK$ and $LIFT$

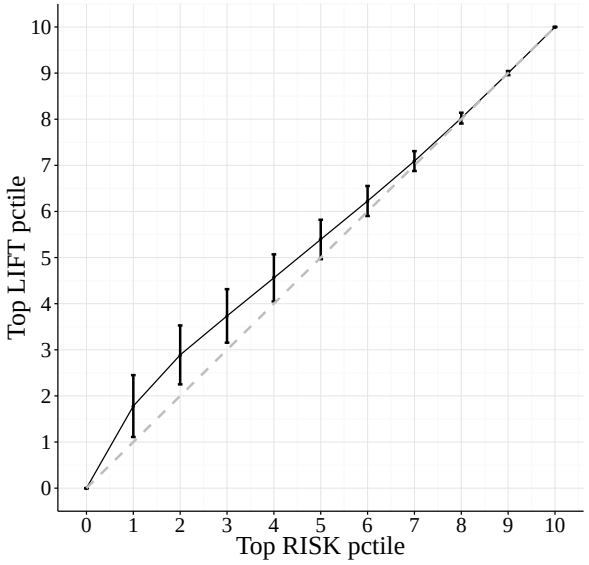
In addition to run the full analysis with the best performing method (as presented in the main manuscript), we also replicated the analysis by using the *RISK* estimates from the best performing random forest. The rationale behind this analysis was to estimate both *RISK* and *LIFT* using the same modeling approach. Below we recreate the figures appearing in the main manuscript corresponding to the heterogeneity in treatment effect (Figures 3a and 3b), the impact of the campaign (Figures 4a and 4b), and the level of overlap between the two metrics (Figures 5a and 5b).



(a) Treatment effect (TE) for different group deciles

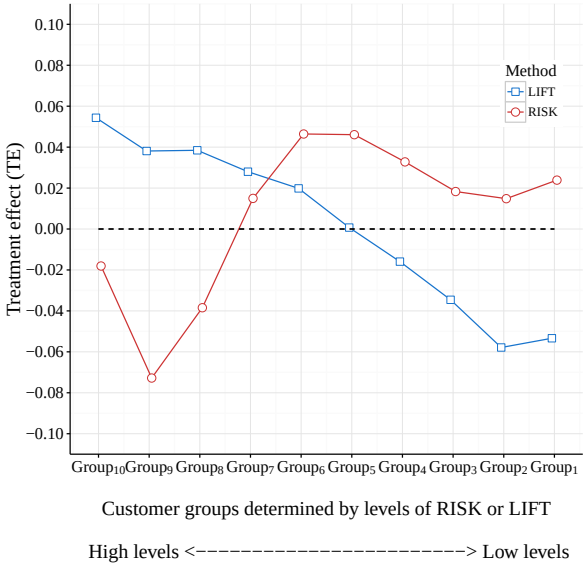


(b) Impact of the campaign under different scenarios

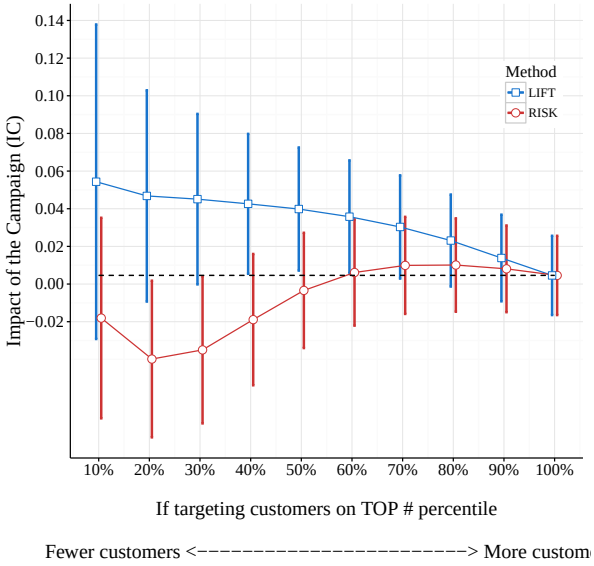


(c) Level of overlap across groups defined by top *RISK* deciles vs. top *LIFT* deciles

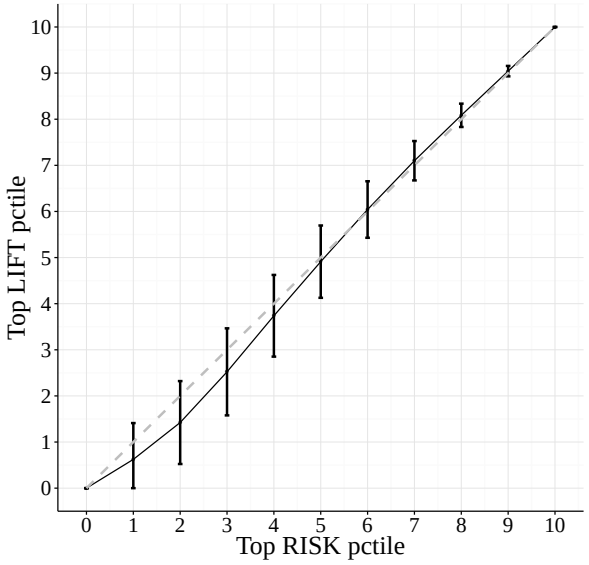
Figure A2: [Study 1] Replication of treatment effect (TE), impact of the campaign (IC), and overlap results using random forest to estimate both *RISK* and *LIFT*.



(a) Treatment effect (TE) for different group deciles



(b) Impact of the campaign under different scenarios



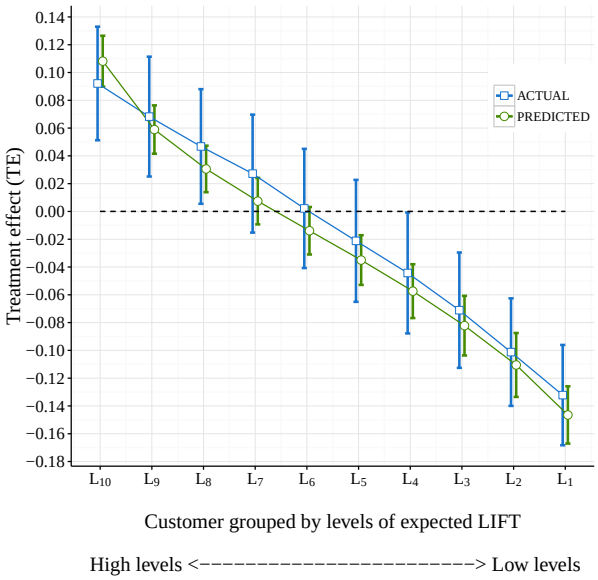
(c) Level of overlap across groups defined by top *RISK* deciles vs. top *LIFT* deciles

Figure A3: [Study 2] Replication of treatment effect (TE), impact of the campaign (IC), and overlap results using random forest to estimate both *RISK* and *LIFT*.

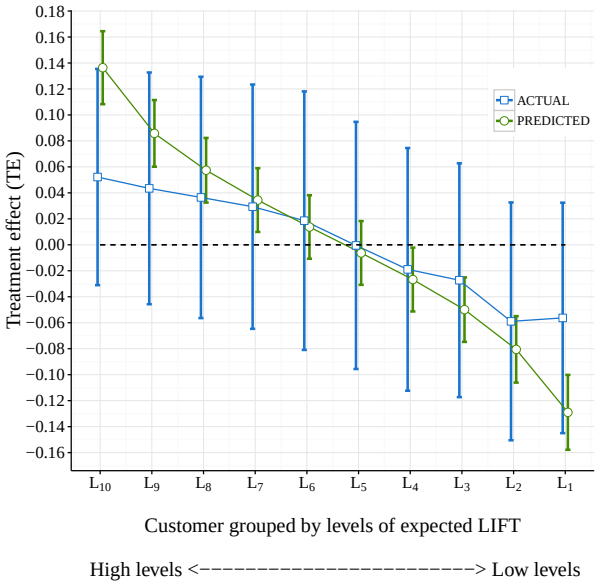
A3 Additional analyses/results

A3.1 Predicted vs. acual $LIFT$

In this appendix we compare predicted and actual $LIFT$ by comparing, by decile, the average $LIFT$ — as predicted by the causal uplift model — with the magnitude of the treatment effect — computed as the difference in observed churn rates between control and treated observations. With reference to Figure A4, we observe that predicted $LIFT$ (green circles) accurately estimates the magnitude to actual $LIFT$ (blue squares). Not surprisingly, the intervals around those estimates are wider for the actual data than for the estimates.



(a) Study 1

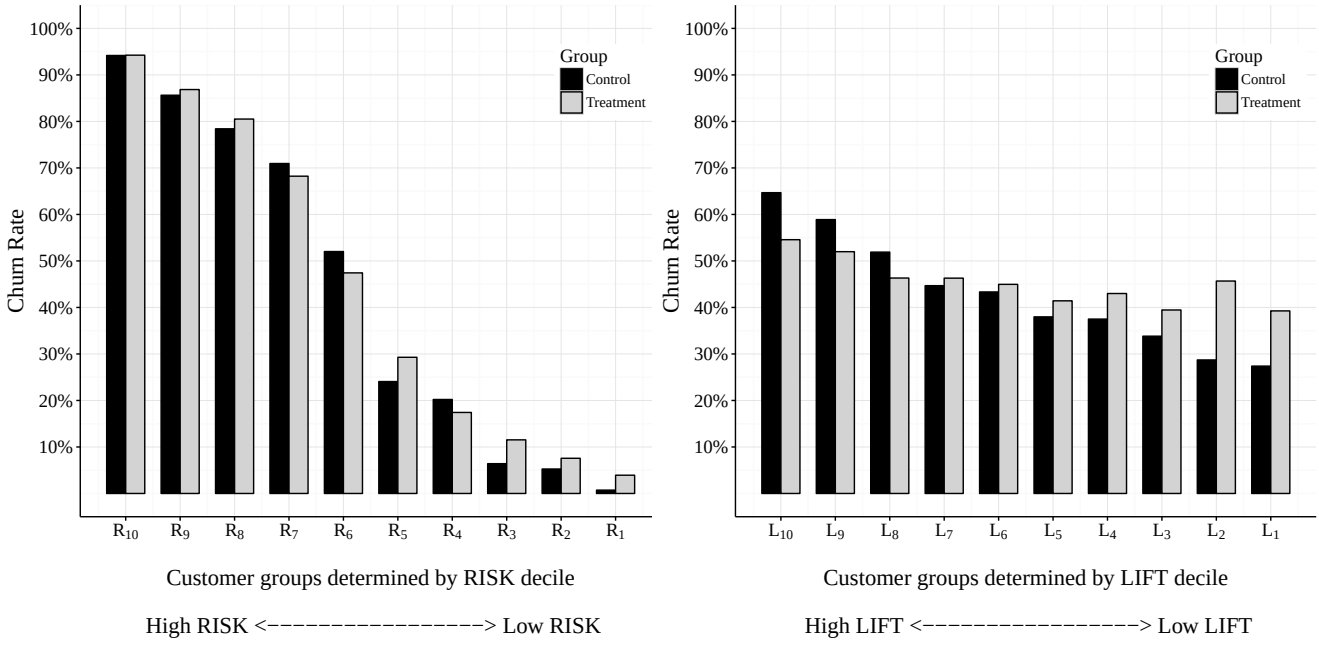


(b) Study 2

Figure A4: Predicted vs. actual *LIFT*. Green (circles) represent the average predicted *LIFT*, representing the expected treatment effect in each decile. Blue (square) represtent the (actual) average treatment effect in each decile.

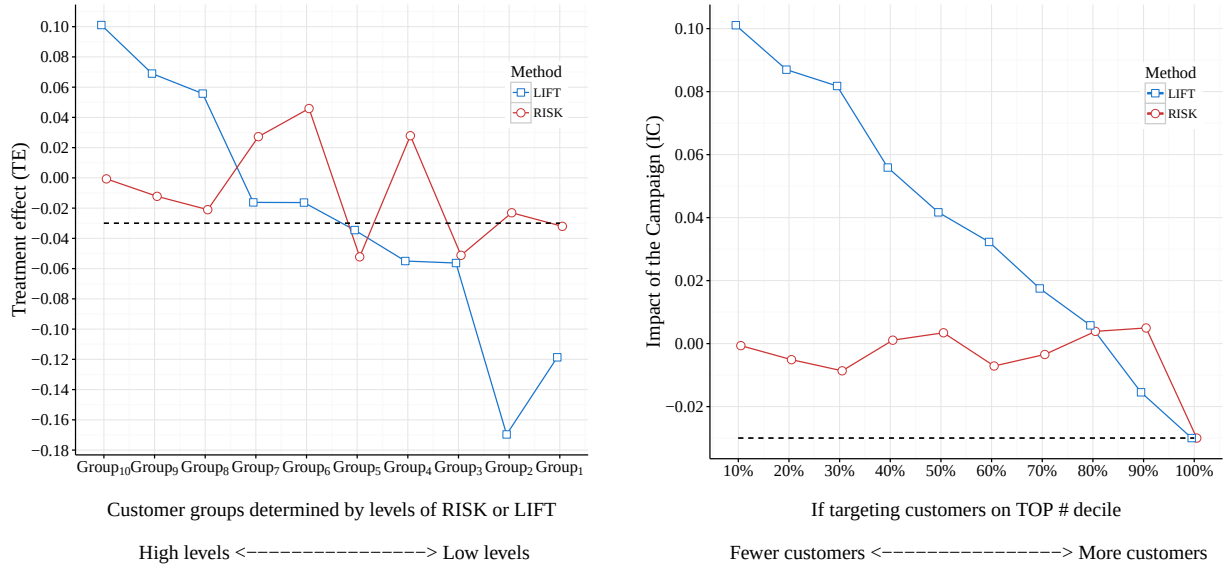
A3.2 Results for one iteration

In this appendix we show the results for one single iteration. The iteration was randomly chosen in R. We draw from an Uniform(0,1), multiply that number by 1000 (as the number of iterations in our analysis), and took the integer number closes to that figure. We performed this procedure just once. While the figures are less smooth (not surprisingly, due to the aggregation), we observe that all patterns of the results are very similar to those obtain when aggregating across iterations. In particular, Figure A5 corresponds to Figures 2a, 2b, 3a and 4a from the main manuscript.



(a) Customers grouped by levels of *RISK*

(b) Customers grouped by levels of *LIFT*



(c) [Study 1] Treatment effect (TE) for different group deciles

(d) [Study 1] Impact of the campaign (IC) under different scenarios

Figure A5: [Study 2] Analysis of churn rates, treatment effect (TE), and impact of the campaign (IC), for one iteration.

A3.3 Differences between customers' $RISK$ and $LIFT$ (results for all variables)

In the main manuscript we only discussed the most relevant variables for each application. In this appendix we present the result for all variables used in the estimation. For the first application we have 37 variables (consisting on the ones described in the main manuscript and multiple dummy variables indicating whether the customer was participating in some specific plans the the focal company offers) and the second application has 50 variables (consisting on the variables described in the main manuscript, interactions between them, and dummy variables indicating the region in which the customer was registered).

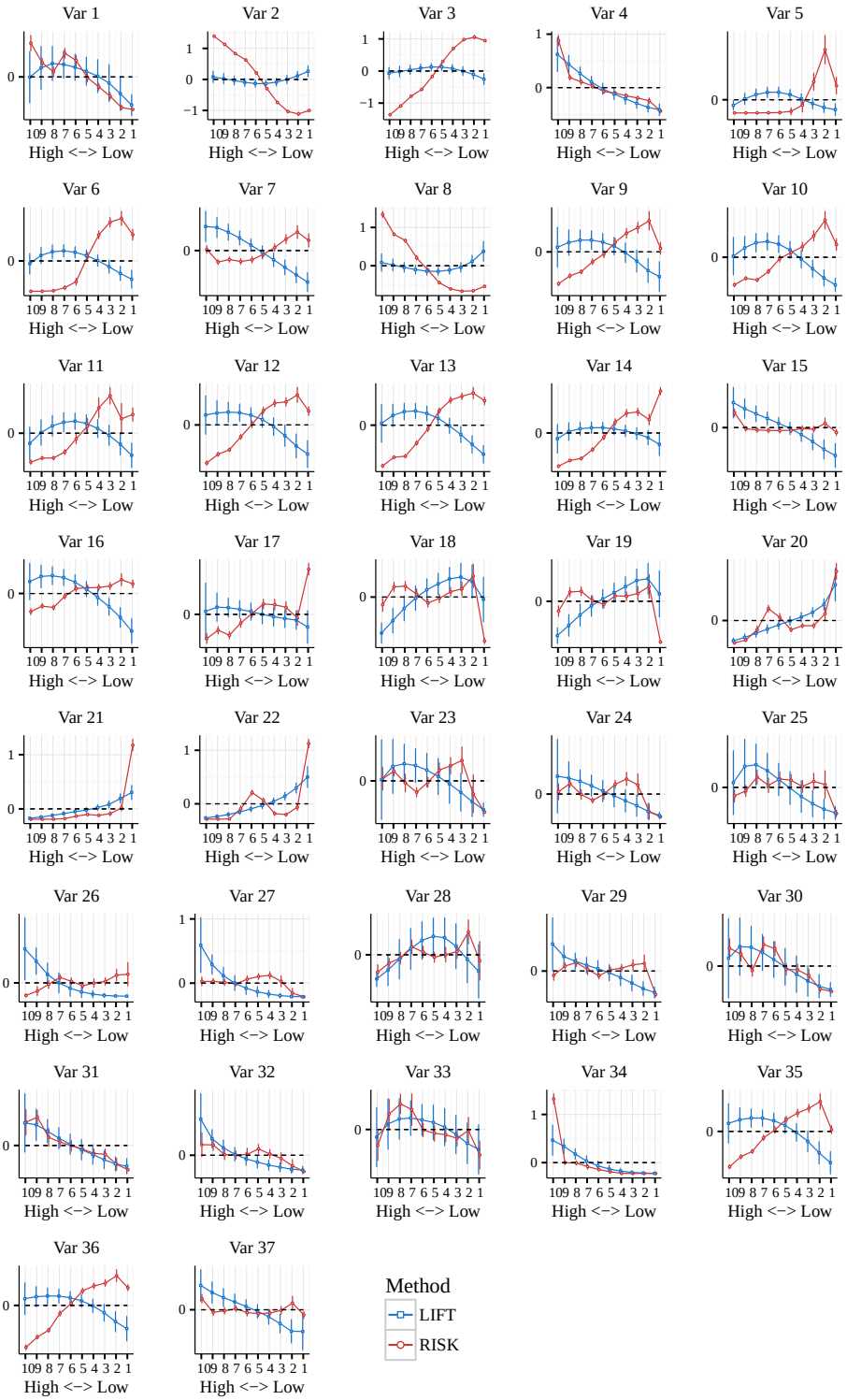


Figure A6: [Study 1] Observed characteristics as a function of *LIFT* and *RISK* deciles

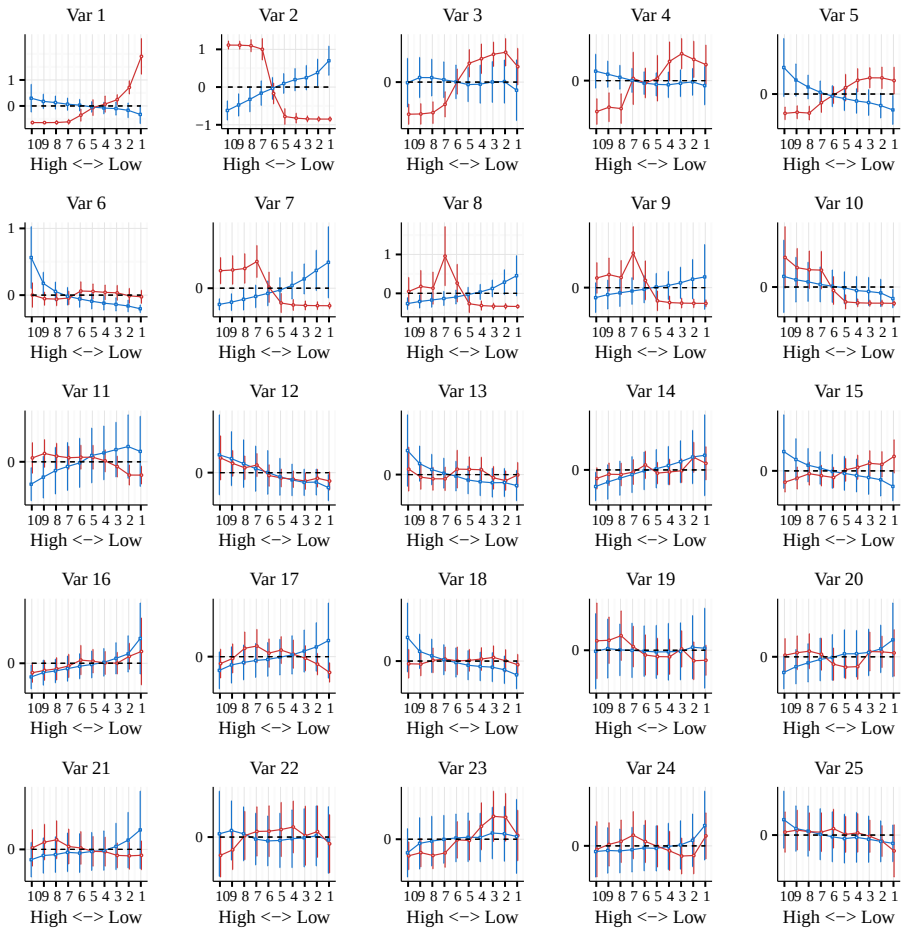


Figure A7: [Study 2] Observed characteristics (variables 1–25) as a function of *LIFT* and *RISK* deciles

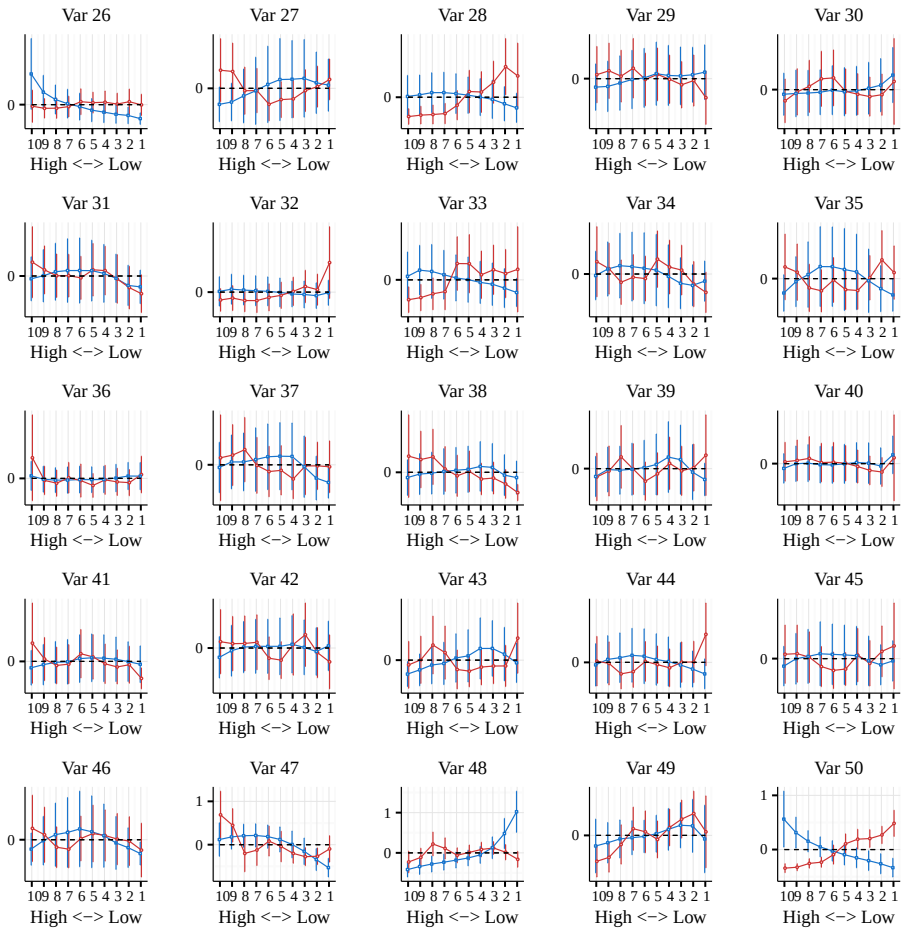


Figure A8: [Study 2] Observed characteristics (variables 25–50) as a function of *LIFT* and *RISK* deciles

A3.4 Simulation study

We conduct a simulation analysis to explore how different levels of correlation between RISK and LIFT correspond to the level of overlap, as reported in the main manuscript. We assume a market context with $N = 5,000$ customers and firm that is trying to prevent churn among them. Customers have an intrinsic propensity to churn (i.e., *RISK*) that is heterogenous across the population. The probability that a customer will churn in the next renewal occasion can be altered if the person receives an incentive. Customers are also heterogeneous in the way the respond to the incentive. In particular, we simulate each customer propensity to churn as follows

$$\text{Churn}_i = \begin{cases} 1 \text{ if ChurnPropensity}_i >= 0 \\ 0 \text{ if ChurnPropensity}_i < 0 \end{cases}$$

where

$$\text{ChurnPropensity}_i = X_i - Z_i \text{Mktg}_i + \varepsilon_i.$$

The term X_i represents the intrinsic (or baseline) propensity to churn (i.e., RISK), Mktg_i is a dummy variable that takes value 1 if customer i gets a retention incentive, 0 otherwise, the term Z_i captures the individual sensitivity to the treatment (i.e., LIFT), and ε_i is assumed normally distributed with mean 0 and variance 1.

We vary the values of X_i and Z_i to cover a variety of business contexts—firms with high/low levels of churn as well as effective/ineffective marketing interventions. For example, a customer with very high X_i is likely to churn; but such churn can be prevented by a marketing action ($\text{mktg}_i = 1$) if Z_i is very low (or "very" negative). Finally, we allow for different levels of correlation between *RISK* and *LIFT* by jointly drawing the individual

quantities X_i and Z_i as follows

$$\begin{pmatrix} X_i \\ Z_i \end{pmatrix} \sim \text{MultivariateNormal} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{A1}$$

The term ρ captures correlation between X_i and Z_i, which we vary from -1 to 1, in intervals of 0.2. Figure A9 shows the level of overlap for all levels of ρ. Comparing these figures with those obtained in the empirical applications, it seems that in the first context (telecommunications) the correlation between$RISK$ and $LIFT$ is close to 0.2. Similarly, the resulting treatment effects (Figure A10) are very similar to those obtained using the real data. Comparing the results from the second application (special interest membership), the correlation between $RISK$ and $LIFT$ is clearly negative, possibly around -0.2.

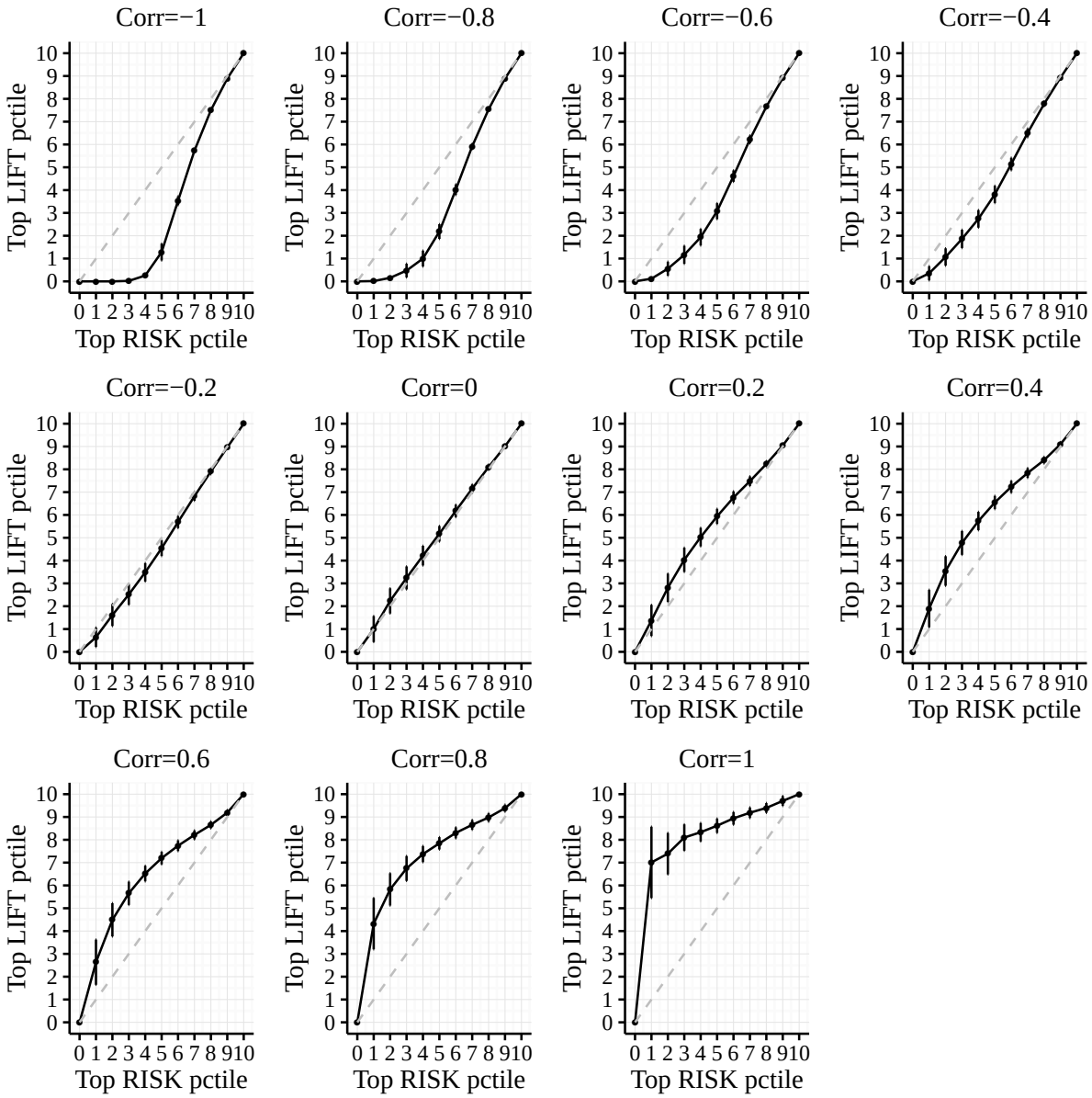


Figure A9: Level of overlap across groups defined by top *RISK* deciles vs. top *LIFT* deciles.The (dotted) 45° line represents the level of overlap if there was no relationship between the two groups

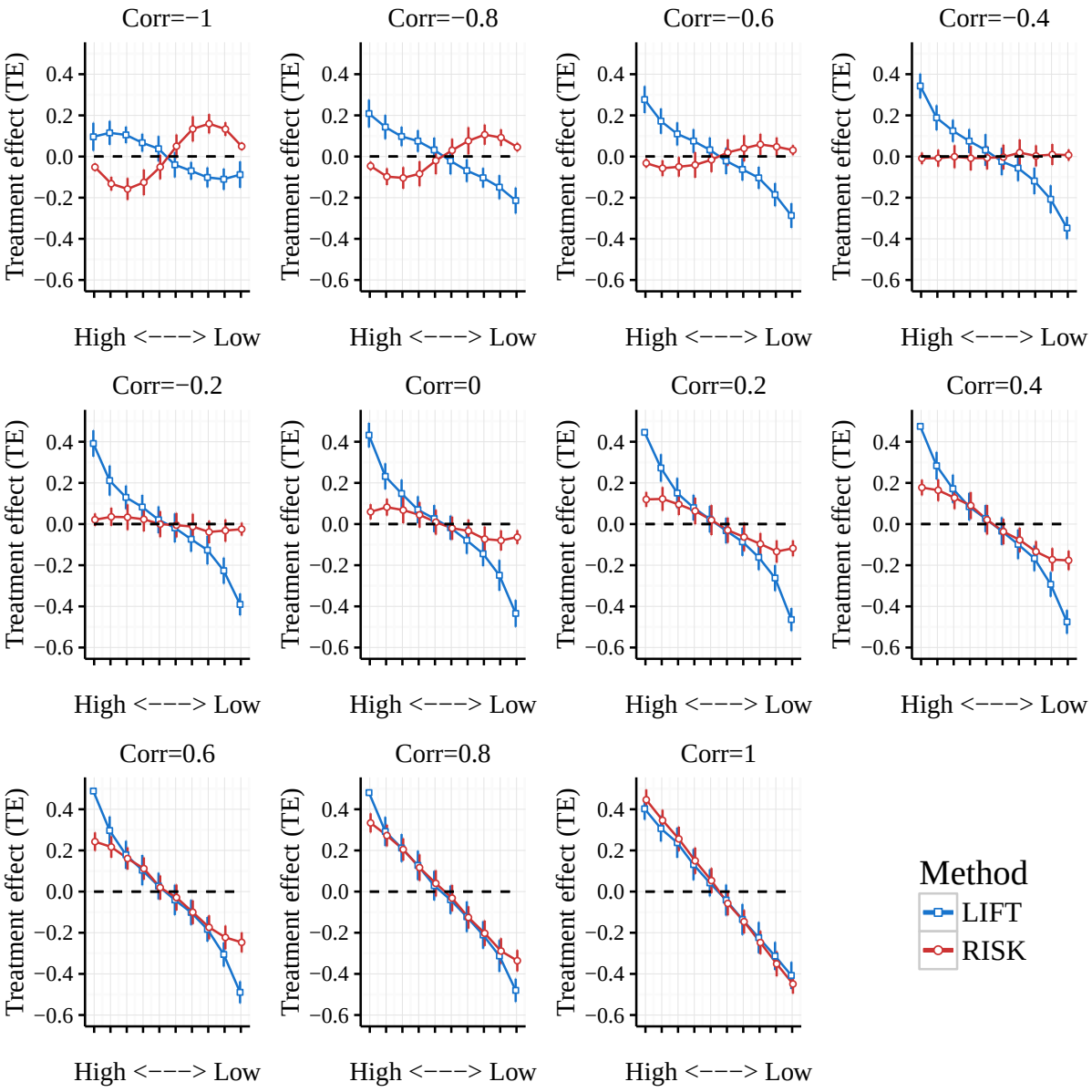


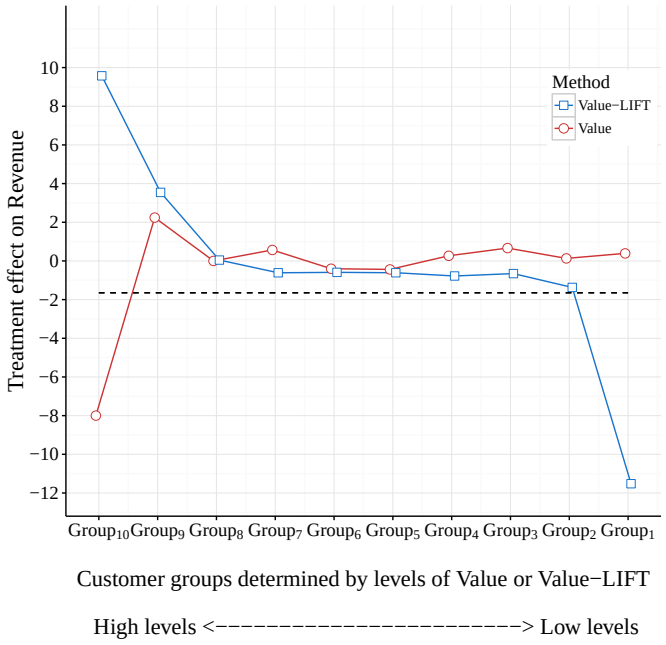
Figure A10: Treatment effect (TE) for different group deciles, depending on whether customers are grouped by levels of *RISK* (represented by the squares) or *LIFT* (represented by the circles). The dotted (straight) line corresponds to the average effect of the campaign if the firm targeted randomly

A3.5 *Value-LIFT*

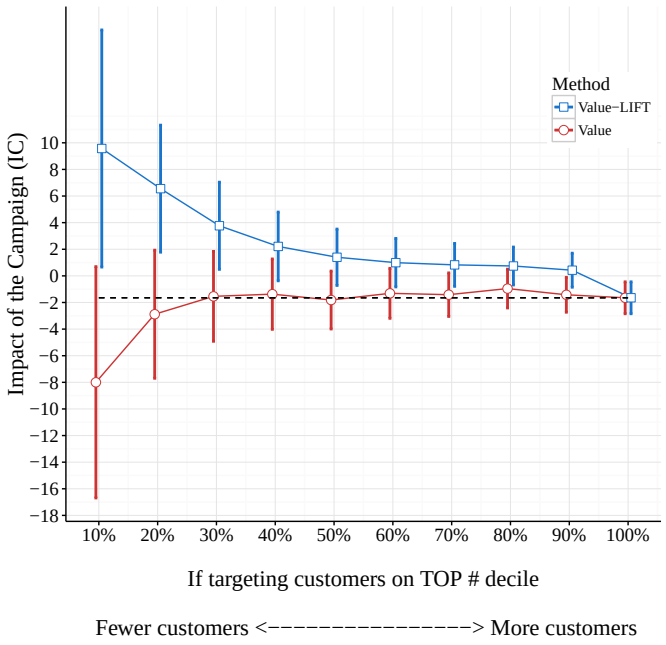
We leverage our first application to show the results of targeting a retention campaign on the basis of a restricted version of *Value-LIFT*. In our case, because the focal company was mainly interested in retention, we could not obtain data on customer profitability after the campaign, hence we cannot observe the change, if any, in customer expenditure. Furthermore, we only obtained retention behavior for the period right after the intervention, preventing us from estimating the impact of the campaign beyond the period of study. Therefore, for the purpose of this exercise, we define *Value-LIFT*$= \lambda_i LIFT_i$, where λ_i is the level of expenditure during the month prior to the campaign.[2] We perform the analysis as described in Section 1 with the main difference that now, when we measure the effect of the campaign, we do not only sum the number of customers that were retained (in each decile) but we also sum all their expenditures. As a comparison, we also compute the effect of the intervention if the company were to target by levels of current expenditure. This comparison is prompted by the common practice of targeting "high value customers" without considering the impact of the intervention in their future value.[3] Figure A11 shows the effect of the campaign, measured as the difference in revenues between control and treated customers, by levels of *Value* versus *Value-LIFT*. From the figure, we can observe that the customers with highest *Value-LIFT* (and not those with highest *Value*) are those for whom the intervention will be most beneficial to the firm. We also quantify what the overall impact of the campaign would be if the company targeted top 10% value customers, top 20% value customers, and so forth. The results (bottom figure) corroborate the claim that companies would notably improve the impact of their campaigns by targeting customers with highest *Value-LIFT* rather than those with high current value.

[2]Note that because we only consider the period after the campaign, *Value-LIFT* is a linear function of *LIFT*. This is likely not to be the case when one incorporates behavior from future periods.

[3]In this application we abstract from computing the discounted value of all future transactions (i.e., computing actual customer lifetime value), as applying such metric would require making several assumptions that are not easily tested in our setting and are not critical for the purpose of this research.



(a) Heterogeneity in the effect of the intervention on post-campaign revenue for different group deciles



(b) Impact of the campaign for different top deciles

Figure A11: [First empirical application] Customers are grouped by levels of *Value* (represented by the squares) or *Value-LIFT* (represented by the circles).The dotted (straight) line corresponds to the impact of the campaign if all customers were targeted)

A3.6 Simulation results for different churn rates

The churn rates for Studies 1 and 2 are 44% and 62%, respectively. Other industries generally face lower churn rates, (e.g., 12% annual churn rate for post-paid customers in telecommunications, 10% pay-TV or streaming services), implying that, by its own nature, the treatment effect of any intervention cannot be very large. In that case, the potential gain of using a better method for targeting will likely be lower than what we find in both studies, where there was a lot more "room for improvement." Nevertheless, that does not mean that the *LIFT* approach will not help companies with lower churn rates than the ones reported here. To the extent that a firm's intervention can have an effect reducing churn, and to the extent that there will be heterogeneity in the customer base, the *LIFT* approach identifies those customers that the firm should give priority.

We corroborate this intuition via simulations. Using the same approach as in Web Appendix A3.4, we simulated four environments which have churn rates of 50%, 25%, 10% and 5%. In all cases we assumed the firm runs a randomized intervention of exact same characteristics. We obtained the expected result: As the churn rate decreases (from 50% down to 5%), the benefit of using *LIFT* approach vs *RISK* decreases, on average. However, regardless of the churn rate, using *LIFT* is always superior as it identifies customers who will be more sensitive to the treatment. These results are shown in Figure A12.

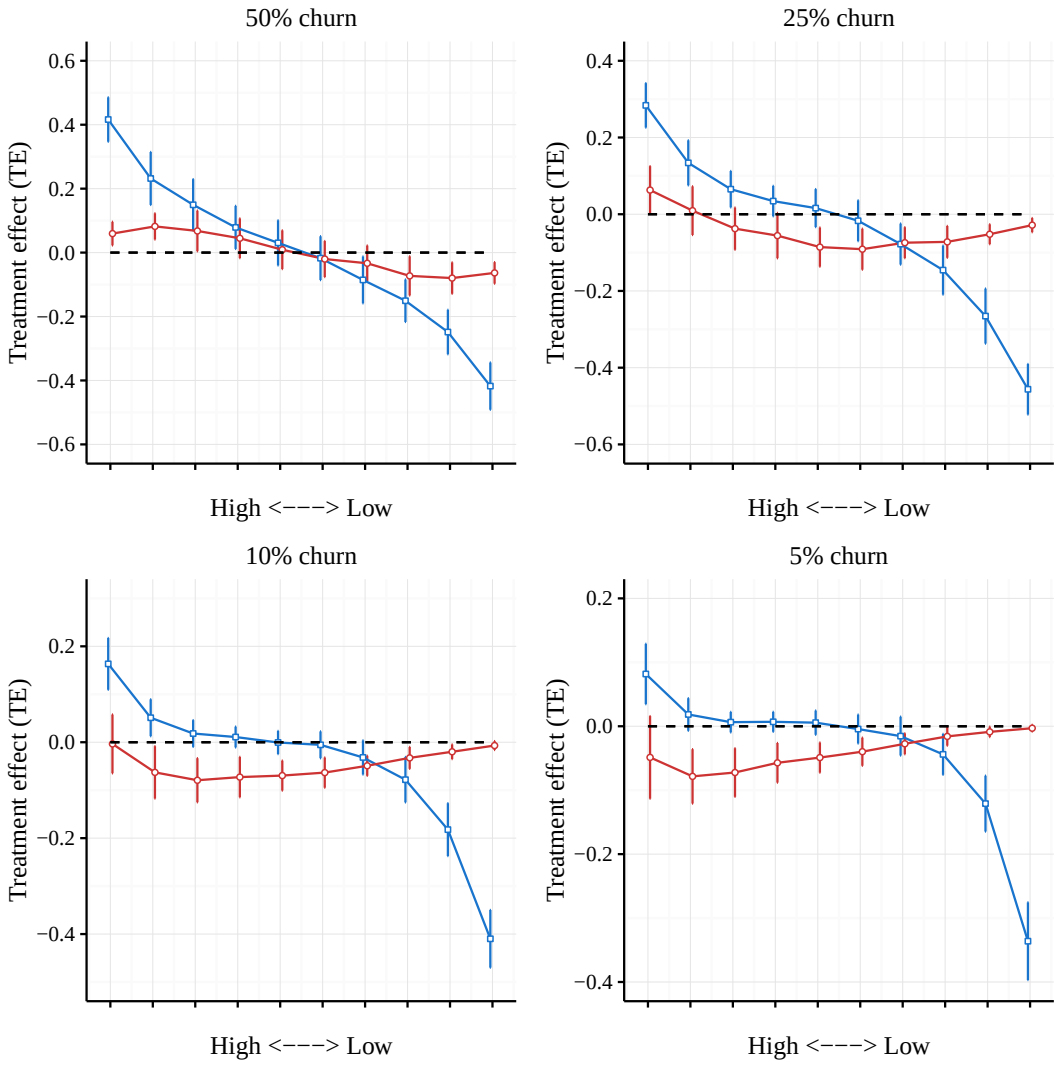


Figure A12: Treatment effect (TE) for simulated data. Varing churn rate from 50% (Top left) to 5% (Bottom left)

References

Breiman, L. (2001), Random Forests. *Machine Learning* 45, 5–32.

Guelman, Leo, Montserrat Guillén and Ana M. Péez-Marín (2015), Uplift random forests. *Cybernetics and Systems* 46(3-4), 230–248.

Rzepakowski, Piotr, and Szymon Jaroszewicz (2012), Decision trees for uplift modeling. *Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE.*

Tibshirani, Robert (1997), The lasso method for variable selection in the Cox model. *Statistics in medicine.* 16(4), 385–395.