EVA ASCARZA*

Companies in a variety of sectors are increasingly 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 research has investigated whether it is indeed optimal to target those individuals. Combining two field experiments with machine learning techniques, the author demonstrates 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. Accordingly, firms should focus their modeling efforts on identifying the observed heterogeneity in response to the intervention and to target customers on the basis of their sensitivity to the intervention, regardless of their risk of churning. This approach is empirically demonstrated to be significantly more effective than the standard practice of targeting customers with the highest risk of churning. More broadly, the author encourages firms and 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, machine learning

Online Supplement: http://dx.doi.org/10.1509/jmr.16.0163

Retention Futility: Targeting High-Risk Customers Might Be Ineffective

> 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.
>
> —Client case study, Accenture Analytics (2014)

> Marketers can also use big data to identify which customers are at highest risk of churn—and re-engage them before they defect.
>
> —AIMIA Institute (Rogers 2013)

> The challenge, of course, is to identify customers who are at the highest risk of churn before they switch to another carrier.
>
> —*Analytics Magazine* (2016)

> 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.
>
> —"Customer Attrition," Wikipedia

Churn management is a top priority for most businesses (Forbes Insights 2011) because it is directly tied to firm

*Eva Ascarza is the Daniel W. Stanton Associate Professor of Business, Columbia Business School, Columbia University (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; 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, University of Chile, and London Business School; and the 2016 Marketing Analytics and Big Data Conference at the University of Chicago Booth School of Business and the 2017 Marketing Science Conference at University of Southern California. The author is grateful to Matt Danielson for his help collecting data and to Yanyan Li for excellent research assistantship. Coeditor: Rajdeep Grewal; Associate Editor: Fred Feinberg.

ISSN: 0022-2437 (print)
1547-7193 (electronic)

Journal of Marketing Research
Vol. LV (February 2018), 80–98
DOI: 10.1509/jmr.16.0163

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 to persuade them to stay (Blattberg, Kim, and Neslin 2008; Neslin et al. 2006). In practice (e.g., Accenture Analytics 2014), as the opening quotations highlight, targeting is generally achieved by assigning a churn propensity to each customer, selecting those who 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 the highest risk of churning.

Although a significant amount of work has tested the accuracy of such methods, no research has investigated whether proactive churn management programs should be targeted to individuals with the highest risk of churning. The main goal of this article is to fill that gap. I empirically examine whether firms should target their retention efforts to customers with the highest risk of churning. More specifically, I 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. I 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. I 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 today. Advances in technology, developments in data analysis, and the increased popularity and ease of implementation of field experimentation have enhanced firms' abilities to gain insights about customers. Through experimentation, firms can better understand customer heterogeneity in the response to marketing interventions. I encourage firms to broaden the use of randomized experiments and use them to identify customers to target, because doing so would increase the effectiveness of their actions. Consequently, building on the uplift modeling literature (e.g., Guelman, Guillén, and Pérez-Marín 2015; Radcliffe and Surry 1999), I recommend 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 on the basis of their sensitivity to the intervention, ensuring that the retention efforts are not futile.

I empirically validate this approach by analyzing customer behavior in two field experiments conducted in different markets (North America and the Middle East) and covering two different sectors (telecommunications and professional memberships). Combining these field experiments with machine learning techniques, I demonstrate that this approach is more effective than targeting customers on the basis of their risk of churning. I find that, across the two studies, the same retention campaign would result in an additional reduction of 4.1 and 8.7 percentage points in churn rate, respectively, if each focal firm followed the recommended approach instead of the industry standard of targeting customers at the highest risk of churning. I 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, I 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 approximately 50%; thus, 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 any modeling assumptions because the sensitivity to retention efforts is estimated in a nonparametric way.

This approach can be further leveraged to identify which customer characteristics (among the observed variables) best predict sensitivity to the retention intervention. In both applications, I 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 on the basis of those variables would likely result in futile retention efforts. Furthermore, I 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 on the basis of 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 recommended 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 deciding how many customers to target or how many resources to allocate to a retention campaign. Across the two applications investigated in this research, I find that half of the retention money is wasted. Most importantly, estimating (observed) customer heterogeneity in the response to the campaign allows the marketer/analyst to identify *which* half.

The findings are generalizable to a large variety of business settings beyond those 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 with the standard practice (i.e., targeting those "at risk"), targeting customers by their sensitivity to the intervention requires a market test as an additional step. Although this step might be viewed 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 data sets), and is estimated using existing (freely available) packages in R (Guelman, Guillén, and Pérez-Marín 2015), 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 to devote 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, Lehmann, 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 (e.g., telecom, pay TV, credit cards) increasingly started to proactively manage churn by detecting those customers at the highest risk of churning and targeting their retention efforts toward 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 have been widely used in practice (for a review of methods and their performance, see Neslin et al. [2006]). 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., Ascarza and Hardie 2013; Lemmens and Croux 2006; Risselada, Verhoef, and Bijmolt 2010; Schweidel, Bradlow, and Fader 2011).

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 been proposed for churn management (Guelman, Guillén, and Pérez-Marín 2012, 2015; Provost and Fawcett 2013; Siegel 2013). However, to the best of my knowledge, no work has investigated the effectiveness of these approaches against the practice of targeting customers at the highest risk of churning, nor has it 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 (for a review on early methods, see Hadden et al. [2007]; for more recent developments, see Ngai, Xiu, and Chau [2009] and Huang, Kechadi, and Buckley [2012]). This research 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 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?

Although there is a long tradition in marketing of estimating the heterogeneous effect of marketing actions to inform targeting decisions (e.g., Ansari and Mela 2003; Rossi, McCulloch, and Allenby 1996), 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 the highest sensitivity to retention interventions, partly because firms did not have enough variation in their databases to estimate the heterogeneous effect of retention actions.[3] The following section describes an approach for proactive churn management that overcomes this limitation. I then use this approach to empirically test the relationship between customers' risk of churning and the effectiveness of proactive churn management programs.

LIFT-BASED TARGETING 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 as suggested by previous literature, is to target the customers who are at the highest risk of churning. In this article, I argue that such a targeting rule is not necessarily optimal.

[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 customers to churn (or attempt to churn) before offering them incentives to stay (Blattberg, Kim, and Neslin 2008).

[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).

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

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 his or her observed characteristics, $P[Y_i|X_i]$, and decide whether to target him or her on the basis of 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.

Alternatively, let T_i denote whether customer i is targeted; this variable takes a value of 1 if the customer is targeted, 0 otherwise. Define

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

$$\text{(2)} \qquad \text{RISK}_i = P[Y_i|X_i, T_i = 0],$$

where $P[Y_i|X_i, T_i = 1]$ is the probability that the customer will churn if (s)he is targeted and $P[Y_i|X_i, T_i = 0]$ is the probability that (s)he will churn if (s)he is not targeted.[4] I 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, Kim, and Neslin (2008) note 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 customers on the basis of their risk of churning, these specific customers would likely be selected for the campaign. In contrast, targeting customers on the basis of LIFT minimizes the likelihood that such customers will be targeted because their LIFT will be negative.

Estimating the Incremental Effect of the Campaign

Whereas estimating RISK_i is straightforward as it requires only data that are readily available in the firms' database, estimating LIFT_i requires the comparison of two outcomes that cannot both be observed—a customer is either targeted or not. Consequently, additional variation in the data and assumptions about how such data are generated will be needed. Assuming that there is no prior information about how customers respond to specific retention interventions, I 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 on the basis of some prespecified 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), one 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, and 0 otherwise), the CATE can be expressed as

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

which corresponds to LIFT_i, as defined in Equation 1. As for covariates, information readily available in the firm's database (e.g., previous purchases) should be used, such that predicting the LIFT_j for any remaining customer j (who was not part of the pilot) is straightforward as it merely involves the evaluation of the estimated model on a new set of observed covariates X_j.

A variety of methods have been proposed with regard to estimating 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 traditionally been used for such purposes (for an overview, see Feller and Holmes [2009]). 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 of

[4]Note that in the case of proactive churn management, $P[Y_i|X_i] = P[Y_i|X_i, T_i = 0]$ because firms compute the risk of the customer churning before (s)he 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 the one explored here is that in recommendation systems, the norm is 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 articles are distinct. Whereas Bodapati 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, I estimate the difference in churn probability directly and nonparametrically.

[6]Because the retention pilot is randomized across customers, it is reasonable to assume unconfoundedness (Rosenbaum and Rubin 1983)—that is, that the treatment is independent on the potential outcomes, given the set of covariates X_i. Therefore, the experimental data can be used to consistently estimate the heterogeneous treatment effect. Note that I define "treatment" as the action of targeting a customer (i.e., the firm sending a retention incentive to a particular customer); thus, my experimental design allows me to consistently estimate the treatment effect. Alternatively, if "treatment" were defined as a customer receiving and responding to the offer, I would be estimating the intention-to-treat effect. I chose to define treatment as the action of targeting because that is the decision variable for the firm.

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

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, Guillén, and Pérez-Marín (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 have also begun to recommend 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. (For a unified review of these types of methods, see Gutierrez and Gérardy [2016].) In this research, I use the algorithm proposed by Guelman, Guillén, and Pérez-Marín (2015) to compute customers' LIFT and combine it with data splits (in the spirit of Athey and Imbens [2016]) to obtain estimates for the treatment effect 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, the firm should 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 previously, the firm should prioritize the retention efforts toward 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 many resources should be put in place. Recall that $LIFT_j$ is the expected effect of the treatment (or campaign). Thus, the firm should target only those customers for whom $LIFT_j > z$, where z represents the minimum effect (i.e., change in average churn probability) that the firm wants to achieve with the retention campaign.[9]

[8] I chose this algorithm because of its accuracy (Guelman, Guillén, and Pérez-Marín [2015] demonstrate its superiority over other non-tree-based methods) and its availability (the R package Uplift is available at https://cran.r-project.org), which facilitate its use among analysts and practitioners. Details about the algorithm and implementation are presented in the "Methodology" section.

[9] For simplicity, and because I do not have firm-level information for the current field studies, I use z = 0 in the validation analyses. However, this number will vary 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). Firms should also incorporate their risk preferences when computing the confidence intervals for the estimated LIFT.

Bringing It All Together

Figure 1 outlines the LIFT-based 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). To obtain those insights, I recommend that 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 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. Thus, with this approach, I encourage firms to employ A/B testing (or small-scale pilots) not only to evaluate marketing actions but also to identify customer targets.

Finally, the main difference between this approach and the common industry standard is that the former 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 the current approach does, rely on the same observed set of variables (X_i). Therefore, when it is indeed optimal to target customers with highest risk of churning, the current approach "nests" to the already-in-use proactive retention programs. In other words, if customers with the highest propensity to churn are indeed those for whom the retention campaign is more effective, the LIFT-based method would recommend targeting the same customers as traditional churn-scoring models would recommend. This approach therefore not only generalizes existing practices for proactive churn management but also allows firms to test the optimality of their current retention efforts.

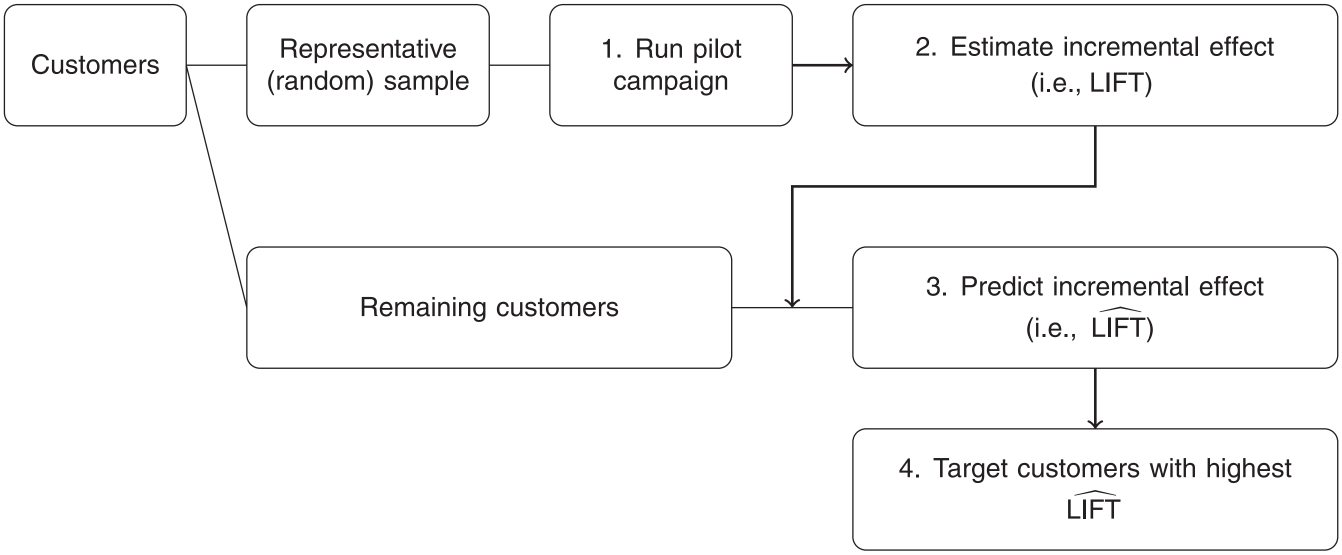
Validation

I compare the performance of this approach with that of targeting "high-risk" customers 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, churn is observed for both treated and nontreated customers. I 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 before the experiment and thus are independent to the intervention. Hereinafter, these variables are referred to as "covariates."

First, I introduce a methodology to empirically assess the effectiveness of each approach (i.e., targeting customers on the basis of RISK vs. targeting customers on the basis of LIFT). I then describe the details specific to each study and present the results obtained in each case. I conclude with a general discussion of the findings.

Figure 1

PROACTIVE CHURN MANAGEMENT PROGRAMS: HOW TO SELECT CUSTOMER TARGETS ON THE BASIS OF THEIR SENSITIVITY TO THE INTERVENTION



Methodology

The main goal is to measure the effectiveness of the standard practice for proactive churn management of targeting customers at the highest risk of churning (i.e., RISK) and compare it with that of selecting customer targets on the basis of their sensitivity to the intervention (i.e., LIFT). I leverage the experimental setup of the field studies to simulate what the impact of these retention campaigns would be had the focal firms implemented each of the two approaches. I replicate the validation exercise for each of the field studies.

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

Step 1: Data split. For each of the data sets, I randomly allocate 50% of the customers to the calibration sample and the remaining 50% to the validation sample. Note that both treated and nontreated 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 the treatment condition, post-experiment churn, and preexperiment covariates), I 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 previously, I employ the algorithm proposed by Guelman, Guillén, and Pérez-Marín (2015) to estimate the LIFT model.[10]

Step 3: Estimate a ("traditional") model for risk of churning. Using the calibration data, I also estimate a "traditional" churn model that will be used subsequently to predict customers' propensity to churn. Among the customers in the calibration sample, I 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. This step mirrors the standard churn-scoring models used in practice. Because I am most interested in predicting the risk of customers outside this sample, I employ the method that maximizes cross-validation accuracy—in this case, a LASSO binary regression.[11,12]

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

[11]I tried multiple approaches to estimate the churn-scoring model, including generalized linear models, random forests, and SVMs. I chose the LASSO approach combined with a generalized linear model because it provided the most accurate forecasts in my applications (for details about the estimated models, see Web Appendix A1.2). Furthermore, I checked the robustness of the results when using random forests to estimate the RISK model (to be consistent with the LIFT modeling approach). The results remain largely unchanged when using such an approach (for results, see Web Appendix A2.1).

[12]Note that only control observations (i.e., customers who did not receive any incentive) can be used to calibrate the RISK model, whereas both control and treatment customers are used to calibrate the LIFT model. I corroborate that this difference in sample size is not driving, not even partially, the results obtained (Web Appendix A2.2).

Step 4: Predict churn metrics in the validation sample. Using the models estimated in steps 2 and 3, I predict two variables of interest among customers in the validation sample:

1. RISK: Using the risk-scoring model estimated in Step 3, I predict the risk of churning for each customer in the validation sample. Specifically, I define

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

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, I predict, for each customer in the validation sample, the following quantities:
   - The probability of churn, *if not targeted*, defined as $P(Y_j = 1 | T_j = 0, X_j = x_j)$.
   - The probability of churn, *if targeted*, defined as $P(Y_j = 1 | T_j = 1, X_j = x_j)$,

   and then LIFT_j is computed as the expected incremental effect of the campaign, given X_i:

$$\text{LIFT}_j = P(Y_j = 1 | T_j = 0, X_j = x_j) - P(Y_j = 1 | T_j = 1, X_j = x_j). \tag{4}$$

I 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 because it implies that the retention campaign increases, rather than decreases, the customers' likelihood of churning. However, I want to account for this possibility because it is possible for retention campaigns to increase churn (Ascarza, Iyengar, and Schleicher 2016; Berson, Smith, and Thearling 2000; Blattberg, Kim, and Neslin 2008). 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. Allowing LIFT to be negative accounts for such a possibility.[13]

Step 5: Measure customer heterogeneity in treatment effect. I 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, I model customer heterogeneity by measuring the treatment effect among subpopulations of customers, defined by the deciles of each variable of interest (either RISK or LIFT). By doing so, the treatment effect among customers in the top decile of RISK can be compared with that of those in the second decile, third decile, and so forth. Similarly, the magnitude of the treatment effect on customers in the top RISK decile can also be compared with that of customers in the top LIFT decile. I choose to model heterogeneity in this fashion not only for its flexibility (I do not impose any parametric relationship between the treatment effect and the level of RISK or LIFT, thus allowing for linear, U-shaped relationships, etc.) but also because decile split is a segmentation method commonly employed by firms (e.g., Bauer 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.

I perform this analysis as follows. For each of the two metrics, I 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 such variation (between treatment and control conditions) is used to calculate the treatment effect in each of the subgroups. I proceed as follows:

- I 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., $R_1 = \{j\}_{\text{RISK}_j < r_1}$), R_2 includes customers with $r_1 < \text{RISK}_j < = r_2$, and R_{10} includes those with $\text{RISK}_j > r_9$. I then compute, for each group R_d, the difference in the actual churn rate (i.e., proportion of customers who churn) across two experimental conditions. More formally, I calculate the treatment effect (TE) in each RISK decile as follows:

$$\text{TE}_{R_d} = \underbrace{\frac{1}{N_s} \sum_{s \in \text{Control}} \mathbb{1}[(Y_s = 1)]}_{\text{Churn rate in control}} - \underbrace{\frac{1}{N_{s'}} \sum_{s' \in \text{Treatment}} \mathbb{1}[(Y_{s'} = 1)]}_{\text{Churn rate in treatment}} \quad \text{for } d = 1, ..., 10, \tag{5}$$

- where N_s 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 $N_{s'}$, representing the number of customers s' in the R_d group who did get the retention incentive (i.e., treatment). Treatment is subtracted 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).
- Next, the validation sample is split 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, the treatment effect (TE_{L_d}) is calculated in each of the groups L_d, with $d = 1, ..., 10$.

I compute these quantities for two main reasons. First, doing so helps shed light on 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 $\text{TE}_{R_{10}} \geq \text{TE}_{R_9} \geq ... \geq \text{TE}_{R_1}$. Second, measuring TE by group also helps identify which groups of customers should and should not be targeted by the firm. For example, if $\text{TE}_{R_5} < 0$, 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, I evaluate the impact of the retention campaign under two scenarios: (1) if the company made its targeting decisions on the basis of customers' propensity to churn (i.e., RISK), as is commonly used in practice and suggested by most of the literature, and (2) if the firm selected customer targets on the basis of the incremental effect of the campaign (measured by LIFT), as

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

recommended in this research. For example, 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. I proceed as follows:

1. I rank customers on the basis of their $RISK_j$ (descending order), and:
   - For each value of P = 10%, 20%, 30%, ..., 100%, I select the top P of customers, denoted as the "target subgroup." Note that as P increases, the number of customers in each group increases, with P = 100% corresponding to the firm targeting the whole customer base.
   - I then estimate the impact of the campaign for each "target subgroup" by comparing the churn rates across experimental conditions. Specifically, the impact of the campaign is computed as

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

   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 receive the retention incentive. As with the treatment effect, I subtract treatment from control, such that positive IC indicates an effective campaign. Because I 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, I 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 I am analyzing field experiments in which the treatment/control allocation is random, choosing groups of customers on the basis of their predicted LIFT (or RISK) does not suffer from endogenous self-selection because these metrics were obtained solely from customers' precampaign behavior. Furthermore, the observed churn behavior of the customers in the validation sample was used neither 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 postcampaign 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 ensure that the results are replicable and not driven by the specific (random) split in step 1, I run steps 1–6 multiple times. In particular, I 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.

[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. The latter approach, for which companies have to estimate the heterogeneous treatment effect (LIFT), is recommended in this research.

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 (from which the data were collected) is located is a well-connected market, with more than 4 million subscribers. There are two main players in this market, with the 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 prepaid credit is added to the account and gives users the right to make calls, send texts, and download data. To keep the account active, customers need to refill their balances within a certain period of time, which depends on the plan to which they belong; 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 who fit 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 three 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 to which the customer belongs. I obtain customers' information for the month prior to the experiment, along with the tenure of each customer (i.e., how long ago (s)he opened the account).

To test whether the randomization succeeded, I 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, several dummy variables the company stores in its database are also observed, including location (country area), type of plan to which the customer belongs, and other internal segmentation variables, which are included in all analyses. For this application, there is 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

[15]Conversations with the firm managers indicated that the difference in voice volume was a mere coincidence because this 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, short message service, voice volume) at different time frames (e.g., within the last week, last two weeks, last four weeks). For brevity, only those from the two weeks prior to the experiment are reported, though all the variables are used in the analysis.

Table 1

RANDOMIZATION CHECK FOR THE DATA FROM THE TWO STUDIES

	Control (N = 3,857)	*Treatment (N = 8,280)*	*Difference* p-*Value*
A: Study 1: Wireless Provider			
Tenure	.002	−.001	.881
Consecutive days with no recharge	.015	−.007	.256
Days since last recharge	−.013	.006	.317
Revenue from last recharge	−.003	.001	.816
Consecutive days with no outbound usage	.018	−.008	.186
Data volume last two weeks (in logs)	−.007	.003	.625
SMS volume last two weeks (in logs)	−.017	.008	.200
Voice volume last two weeks (in logs)	.043	−.020	.001
B: Study 2: Membership Organization			
	Control (N = 1,057)	*Treatment (N = 1,044)*	*Difference* p-*Value*
Tenure	.013	−.013	.565
Attendance (binary)	.311	.298	.527
Online activity (binary)	.413	.401	.591
Download activity (binary)	.164	.143	.180
Special interest attendance (binary)	.050	.058	.460

Notes: All continuous variables were first standardized then summarized across conditions. SMS = short message service.

high as 100%) 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.

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, the 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 letter. 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, I 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, I observe the month in which the renewal letter was sent, whether the customer renewed his or her subscription, and other demographic and usage characteristics such as tenure in the organization (in number of years), location (U.S. state of residence), whether the subscriber attended any organized event (0/1), and whether the subscriber had logged in to the organization website. The company did not target on any of these, or other, characteristics.

As with the first study, I 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, I 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. I also include the interaction between the first-year dummy and the four usage variables as well as several dummy variables indicating the customer's geographical location. Fifty variables are observed in total.

Validation results. I begin 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). I first discuss the results for Study 1 (wireless) and then compare them with those obtained in Study 2 (membership).

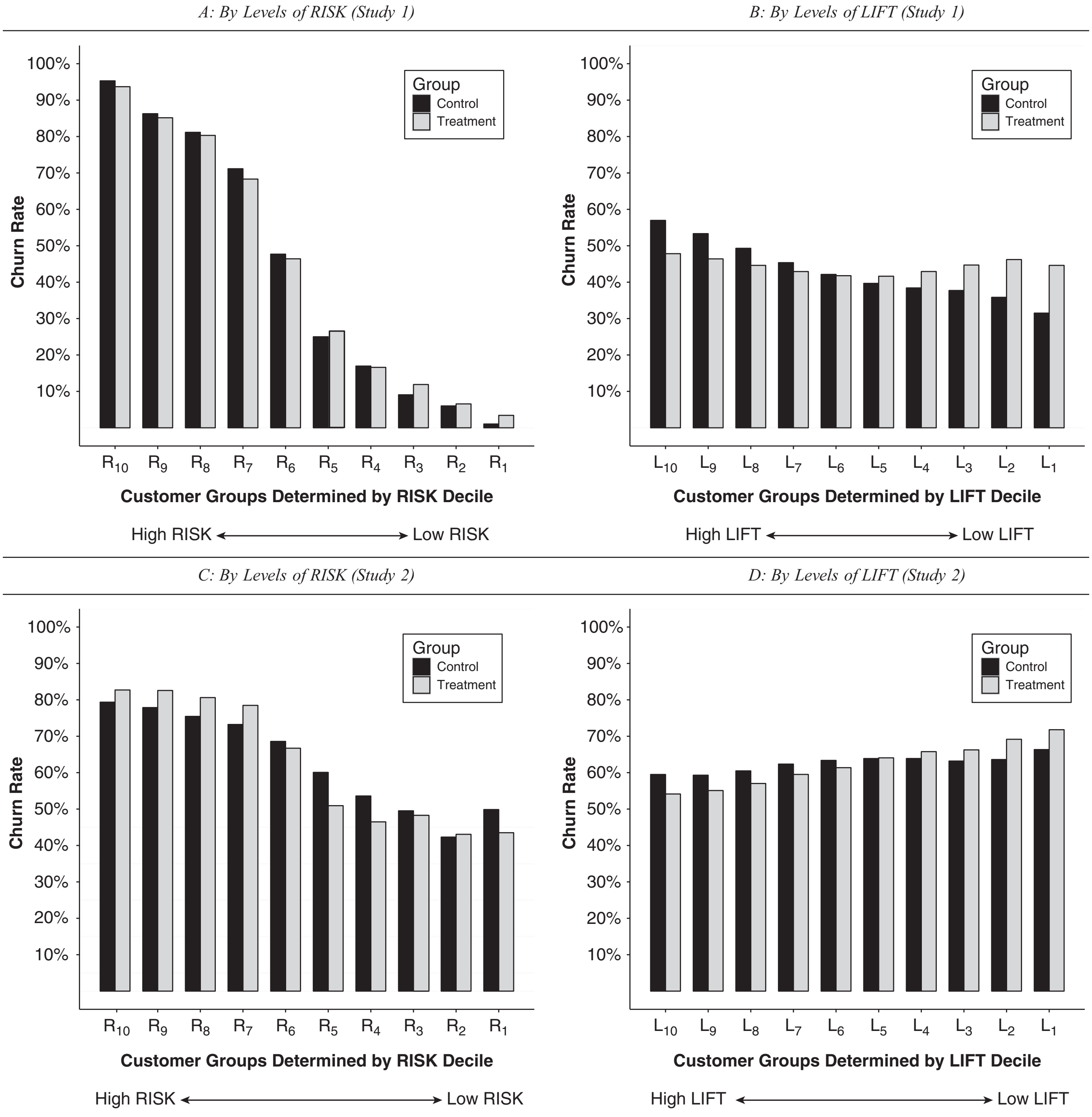
Figure 2, Panel A, shows, for different levels of RISK, the churn rate of customers in each of the experimental conditions in Study 1. R_{10} corresponds to the customers identified as being at the highest risk of churning, whereas R_1 corresponds to those at the lowest risk of churning. Comparing churn rates across conditions, the extent to which treatment reduced churn in each of the RISK groups is evident. 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). 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. The intervention was not beneficial 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 11.9% for treatment, corresponding to a 2.8 percentage points increase in churn rate).

Figure 2, Panel B, summarizes churn rates when customers are grouped on the basis of their LIFT. Two insights can be highlighted: 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

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

Figure 2

CHURN RATES ACROSS EXPERIMENTAL CONDITIONS FOR DIFFERENT GROUP DECILES, DEPENDING ON WHETHER CUSTOMERS ARE GROUPED BY LEVELS OF RISK OR LIFT



Notes: Churn rates are estimated for each experimental condition, when targeting customers with different levels of churn propensity (i.e., RISK) versus targeting customers with different levels of sensitivity to the retention intervention (i.e., LIFT).

57.0% and 47.8%) among customers in L_{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.

Similarly, Figure 2, Panels C and D, illustrate churn rates for customers in Study 2, by different levels of RISK and LIFT, 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 with 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. In contrast, 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. Note also that this pattern—the nonmonotonic relationship between the treatment effect and RISK—differs from that in Study 1, suggesting that the relationship between levels of RISK and the response to the intervention is not easily predictable.

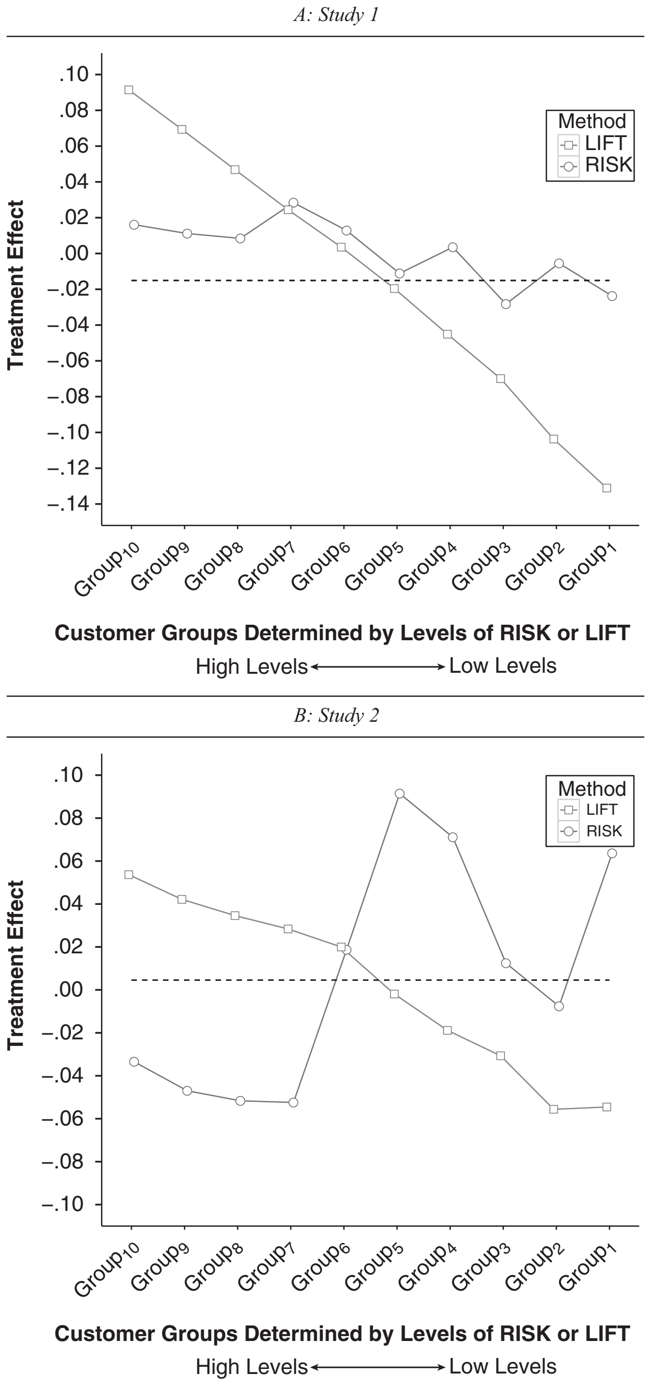
In contrast, when the heterogeneity in treatment effect is examined with respect to customers' LIFT (Figure 2, Panel D), I find an identical pattern to the previous application. Customers with the highest levels of LIFT (L_{10} through L_6) respond positively to the treatment—churn rates are approximately five percentage points lower for treated customers than for control customers—whereas the treatment increases churn among those with lowest levels of LIFT (L_4 through 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 circles represent the average (across all iterations) of the treatment effects for different levels of RISK while the squares correspond to levels of LIFT. The dotted line marks the average effect of the campaign, which corresponds to the expected effect of the campaign if the firm targeted customers randomly. Comparing the results across both studies, customer LIFT is a strong discriminatory variable for targeting marketing efforts, whereas the pattern for RISK is rather unclear.[17]

Impact of the retention campaign if targeting is based on RISK or LIFT. Next, I compare what the impact of the campaign would be if the firm targeted the same proportion of customers, selecting them on the basis of 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, and so on. The squares/circles represent the average across iterations, and the bars represent the standard deviation around the mean.[18] As discussed in 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 tenth LIFT decile. The impact of targeting the top 20% LIFT customers corresponds to the average of the treatment effects of the tenth and ninth LIFT deciles.[19] The straight dotted line corresponds to the impact of the campaign if the company targeted customers at random, which is the average treatment effect of this campaign.

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., $IC_{L_{10\%}} > IC_{L_{20\%}} > ... > IC_{L_{100\%}}$). This pattern should be expected because the LIFT approach selects the "best" (i.e., most sensitive) customers first. Therefore, as more customers are targeted, the effectiveness of the campaign should decrease. Second, and most importantly, the LIFT-based approach is substantially more effective than the "at risk" approach. In other words, both firms would have saved more customers if they had targeted their retention efforts on the basis

Figure 3

TREATMENT EFFECTS FOR DIFFERENT GROUP DECILES, DEPENDING ON WHETHER CUSTOMERS ARE GROUPED BY LEVELS OF RISK OR LIFT



Notes: The dotted line corresponds to the average effect of the campaign if the firm targeted randomly.

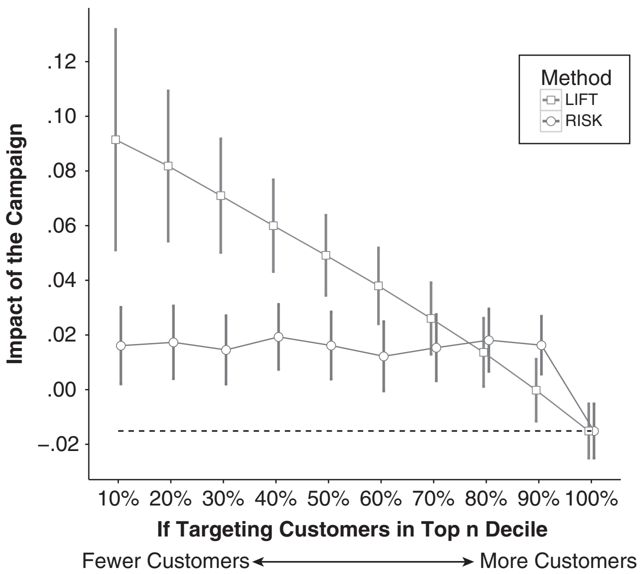
[17]I corroborate that the recommended approach not only sorts customers from greater to lower treatment effect but also provides an estimate of the size of the effect. For more details, see Web Appendix A3.1.

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

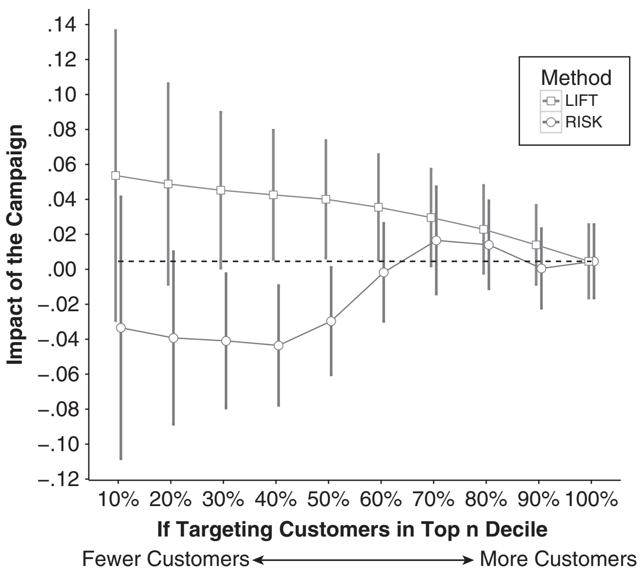
[19]For example, comparing Figures 3 and 4, $IC_{L_{10\%}} = .091$ in Figure 4, Panel A, equals $TE_{L_{10}} = .091$ in Figure 3, Panel A; $IC_{L_{20\%}} = .080$ in Figure 4, Panel A, corresponds to $1/2\ (TE_{L_{10}} + TE_{L_9})$ in Figure 3, Panel A, and so forth.

Figure 4
IMPACT OF THE CAMPAIGN UNDER DIFFERENT SCENARIOS

A: Study 1



B: Study 2



Notes: 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 line corresponds to the impact of the campaign if all customers were targeted. Bars represent standard deviations. The dotted line corresponds to the impact of the campaign if all customers were targeted.

of customers' LIFT rather than on the basis of 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 ($IC_{R_{40\%}}$ = .019). However, if the same proportion of customers had been targeted but selected on the basis of their LIFT, the same campaign would have caused a 6.0 percentage points churn reduction ($IC_{L_{40\%}}$ = .060). The 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 in churn reduction between targeting on RISK and LIFT is as high as 8.7 percentage points. Finally, and not surprisingly, both methods would provide similar effectiveness if the company decided to target most customers.

Differences between customers' RISK and LIFT. I further leverage the data and explore the differences between customers' RISK and LIFT. Specifically, I 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). I 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, although the findings are correlational, they add 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.

Next, I quantify the level of overlap between the RISK and the LIFT metrics. The results thus far suggest that the level of overlap between the groups of customers with high (low) RISK and those with high (low) LIFT should not be high; otherwise, the lines for the RISK and LIFT approaches in Figures 3 and 4 would be similar. I 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 solid line represents 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. In contrast, 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).

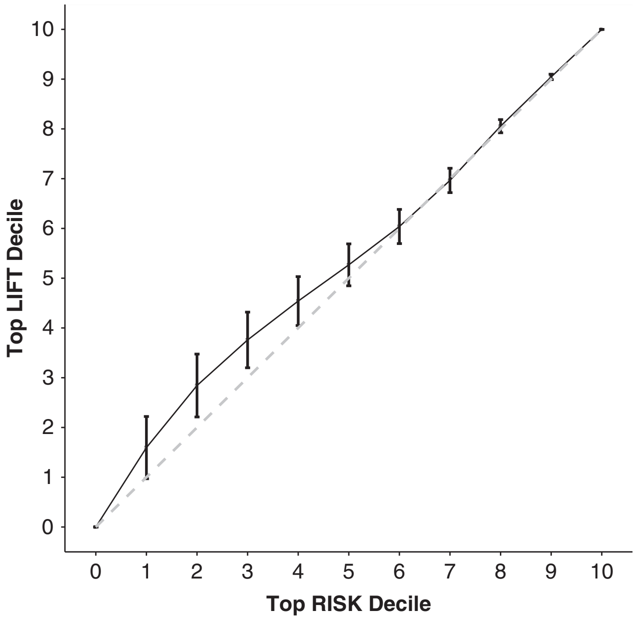
As Figure 5 illustrates, the relationship between these two metrics is rather weak. In Study 1, among the 10% of customers with highest RISK, only 16% of them also belong to the top 10% LIFT group. Of the top 50% of RISK customers, only 52% of them belong to the highest LIFT group, suggesting that if 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 2, Panel C), 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. For the 50th 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.

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

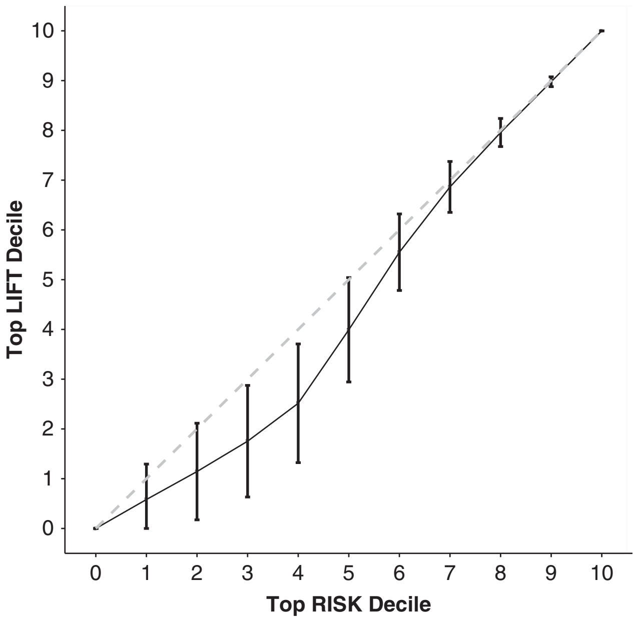
Figure 5

LEVEL OF OVERLAP ACROSS GROUPS DEFINED BY TOP RISK DECILES VERSUS TOP LIFT DECILES

A: Study 1



B: Study 2



Notes: The dotted 45° line represents the level of overlap if there were no relationship between the two groups. Bars represent standard deviations.

Finally, I explore which customer characteristics are predictive of customers' RISK and LIFT. I 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, I report only the variables that are most relevant for each study (Figure 6, Panel A, for Study 1 and Figure 6, Panel B, for Study 2); the full set of results is reported in Web Appendix A3.3.

In 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 (see Figure 6, Panel A). For example, consider the variable "days no recharge," which represents how long it has been since the customer put money in his or her account. Customers with longer times are, not surprisingly, more likely to churn in the following month. However, this variable is not predictive of the sensitivity 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 a very high risk of churn (as represented by the upward relationship between this 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 (because they belong to a "high churn" segment), such a campaign would likely increase churn. Among the variables selected, only "tenure" and "last recharge" have similar relationship patterns with RISK and LIFT. In contrast, 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 extent 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 these types 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.

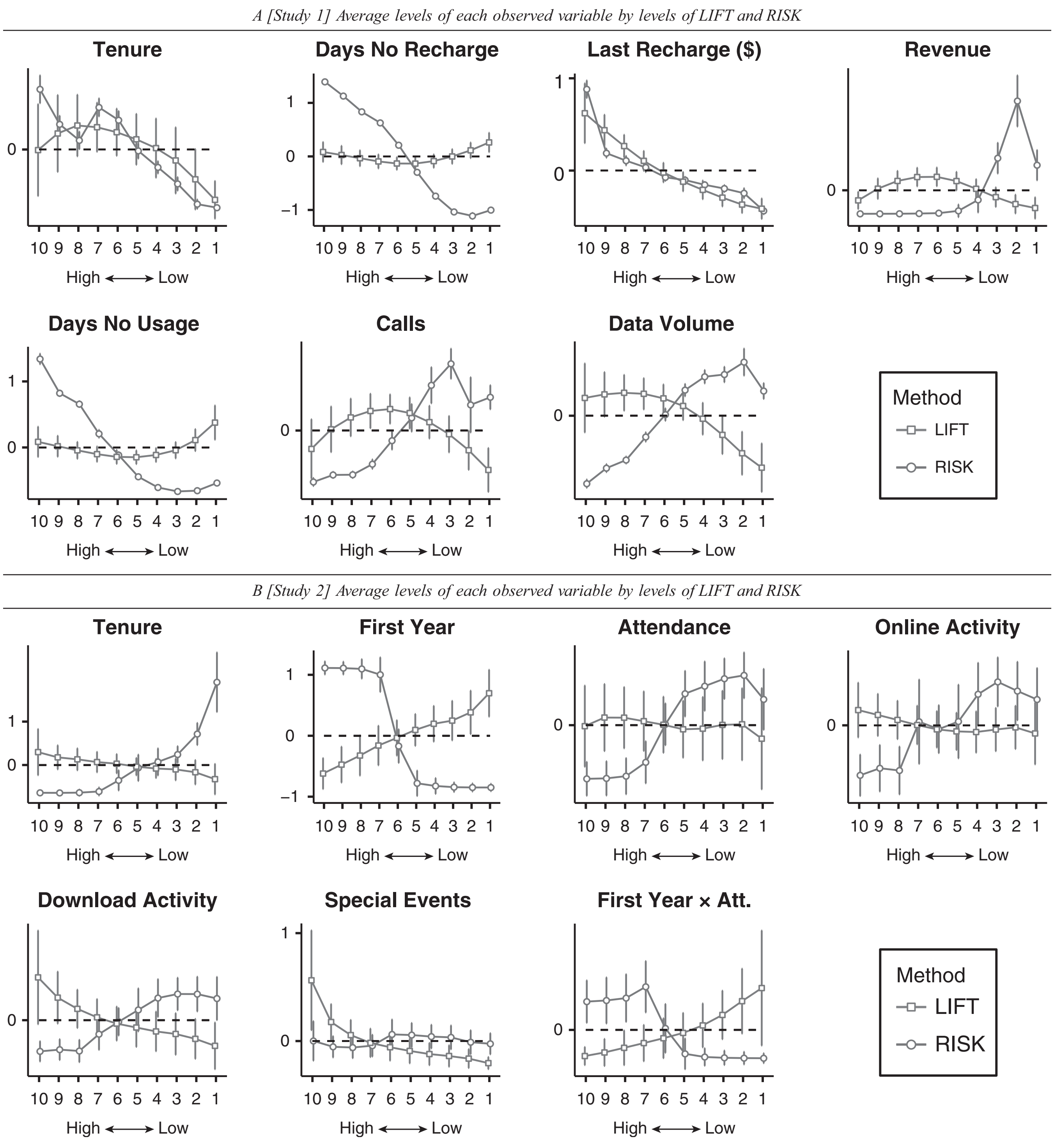
Figure 6, Panel B, displays the relationship between the observed characteristics and the RISK and LIFT metrics in Study 2. Consistent with the results from Figure 2, Panel C, the majority of the variables show the opposite pattern when predicting RISK than when predicting LIFT. For example, while customers in their first year of membership 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 related to offline 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 6, Panel B, are flat). It would be worthwhile to investigate whether such a relationship (or lack thereof) 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, I have demonstrated that targeting on the basis of LIFT is more effective at reducing customer churn than targeting on the basis of RISK. In particular, I 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

Figure 6

AVERAGE VALUE OF EACH OF THE OBSERVED CHARACTERISTICS FOR EACH DECILE ($R_{10}, R_9, \ldots, R_1$ AND $L_{10}, \ldots, L_1$)



Notes: Squares/circles represent the average across iterations and the bars represent standard deviations.

selected customers by their sensitivity to the intervention instead of following 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) in two different markets (Middle East and North America); thus, they may be generalizable to a variety of proactive churn management programs.

With respect to the question of heterogeneity in treatment effects (Figure 3), I 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, I 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. In contrast, the pattern between customers' predicted LIFT and their response to the treatment is strong and consistent across both applications.

Finally, this validation approach allowed me 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, I 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 my 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 the second application, *all* customers who were up for renewal participated in the experiment. As such, the data cover the full range of RISK levels among customers. I acknowledge that, in the first study, the data used in the experiment might not cover the entire range of RISK levels because customers with certain characteristics (who were, in theory, at a higher risk of churning) were selected for the study. Nevertheless, the strong consistency across the two studies provides 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] Whereas 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 vs. money incentive, giving a price discount vs. a thank-you gift), different variables would be expected to be related to the impact of the campaign. Nevertheless, across both applications, there was 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, I have demonstrated that, contrary to 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 who may even respond negatively) to the campaign. In other words, the present analyses not only demonstrate that half of the retention money is wasted but also identify *which half*. Consequently, firms should first 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 settings in which there was a clear metric to capture customer churn. Following Schmittlein, Morrison, and Colombo (1987), these are generally called "contractual settings," a term used when the loss of the customer is observed by the firm. Conversely, 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 the recommended 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. Although 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 the present approach to these noncontractual settings is straightforward. Building on the notation introduced previously, noncontractual firms first need to decide how to operationalize the dependent variable (Y_i). Recall that, in this 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 described in this article 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 the second application (the special interest organization), every customer pays the same annual fee, implying that churn is the main differentiator for customer value. However, there exist many settings in which customer revenue

[21]I perform a simulation study in which churn behavior is simulated in the context of a randomized intervention. I 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 .2 between RISK and LIFT constructs while the second application is consistent with the scenario of a correlation of −.2.

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

[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). The difference is defined such that positive LIFT represents what is beneficial to the firm (more transactions or less churn).

[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 on the basis of the incremental effect of targeted marketing interventions. Unlike previous work, the present experimental approach does not need any distributional assumption about the propensity to engage in the behavior of interest (e.g., make a purchase, churn) and does not require multiple observations per customer to identify which customers (based on their observed characteristics) should be targeted.

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 the first empirical application), financial services, energy utilities, health care, or online games (with in-app purchases).[25] Arguably, proactive retention campaigns should not only focus on retaining customers but, more specifically, focus on retaining high-value customers and, when possible, increasing the value of their current customers.

In general, the 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 CLV depending on whether the customer is targeted. In other words, and with reference to the notation introduced previously, the goal of the campaign would be to maximize

$$\text{(7)} \qquad \text{Value-LIFT}_i = \mathbb{E}[\text{CLV}_i | X_i, T_i = 1] - \mathbb{E}[\text{CLV}_i | X_i, T_i = 0],$$

where CLV_i is now a continuous metric representing the discounted value of the (postcampaign) customer profitability. While causal random forests can be easily applied to the case of a continuous dependent variable,[26] the real challenge of estimating Value-LIFT_i is that a very long time horizon is needed to estimate the impact of the marketing intervention in consumer behavior; alternatively, strong assumptions about the impact of the marketing campaign need to be made.

For example, in the case of a contractual relationship, CLV_i can be expressed as

$$\text{(8)} \qquad \text{CLV}_i = \sum_{t=1}^{\infty} \frac{\lambda_{it} \prod_{\tau=1}^{t} r_{i\tau}}{(1+d)^t},$$

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. To simplify the expression of CLV, most previous 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 the 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 per period. As a result, although Equation 8 could be simplified 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 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 noncontractual businesses, wherein the authors estimate the differences in discounted expected residual transactions depending on the customer requested level of service.

CONCLUSION

Contrary to prior research and marketing practice, I demonstrate that proactive churn management programs should not necessarily be targeted to customers who are at the highest risk of churning. Rather, firms 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* the intervention.

Combining data from two field experiments with machine learning techniques, I empirically find that customers' risk of churning is not necessarily related to their sensitivity to the retention campaign. In turn, across both studies, I consistently find no overlap between the groups of customers who are at the highest risk of churning and those who are most responsive to the retention incentives (and thus should be targeted). I 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, 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 of churning.

In addition to its effectiveness in reducing churn, this approach has other desirable characteristics that facilitate its use among practitioners. First, the method is scalable to large customer populations and large sets 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 contexts in which retaining customers is a concern. In particular, only two conditions are needed for a business setting to leverage the insights from this research: the company can (1) observe customer behavior at the individual level and (2) interact with customers on a one-on-one basis (i.e., it can run individually

[25]Another dimension in which some customers can be more valuable than others is by their level of influence on their connections (Ascarza, Ebbes, et al. 2017; Nitzan and Libai 2011). While my approach could potentially incorporate customer influence, I do not consider that case in my applications because such data are not available to me.

[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 the 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 have that $\text{CLV}_i = \sum_{t=1}^{\infty} \frac{\lambda_i r_i^t}{(1+d)^t}$, which can be further simplified to $\text{CLV}_i = [\lambda_i(1+d)]/(1+d-r_i)$ (for derivations and a detailed discussion about different CLV formulae, see Fader and Hardie [2012]). Then, one can further simplify the expression in Equation 7 if assuming 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 $\text{Value-LIFT}_i = ([\lambda_i(1+d)]/\{1+d-[1-\mathbb{P}(Y_i|X_i, T_i = 0)]\})(\mathbb{P}[Y_i|X_i, T_i = 0] - \mathbb{P}[Y_i|X_i, T_i = 1])$ is the highest. Note that the second term of the formula corresponds to the LIFT_i metric as defined in Equation 1 and estimated as described in the methods section.

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

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 tracks individual behavior. Compared with current practice, this 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 because 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. Firms are encouraged to broaden the use of randomized experiments and leverage those data to better understand the heterogeneity in response to the marketing actions. Put differently, marketers, analysts, and researchers are recommended 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).

I acknowledge that the proposed method does not explicitly incorporate competitors' actions. As highlighted by Subramanian, Raju, and Zhang (2013), the firm's most valuable customers might easily be the ones who competitors aim to poach, making those customers most responsive to retention efforts (provided that the offered incentive compensates the competitive alternatives), whereas those with lower value to the firm might not be as attractive to competitors and, thus, might be the most insensitive to retention actions. However, 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 the current approach measures the sensitivity to the retention incentive, these effects are (implicitly) captured; however, the approach does not isolate the heterogeneity in sensitivity that is driven by the competitors' actions. Understanding those drivers would be valuable for firms as they would be able to focus on customers who are sensitive to their actions and not necessarily sensitive to competitors' offers (Musalem and Joshi 2009).

Even though I was able to collect experimental data from two different contexts, adding to the generalizability of my findings, my empirical approach imposes some limitations. First, the churn rates observed in the two contexts were similar in magnitude. While I anticipate/speculate that targeting retention efforts based on customers' sensitivity to the intervention is beneficial regardless of the churn rate observed in the market, I acknowledge that the expected benefit of using this approach might be smaller in settings in which churn rates are very low.[29] Second, this approach is applied 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 financial services (e.g., insurance companies, banks) continuously implement proactive campaigns, targeting customers whose contracts are close to expiring. Companies in other industries (e.g., arts, sports, 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 worthwhile to analyze multiple campaigns from the same company (or a similar pool of customers) to 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 the current approach to such a setting would identify which types of incentives should be sent to which types of customers.

Another aspect that deserves more attention is the length of the assessment period. In both applications, one month was used 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.

Finally, several methodological aspects of the current approach merit further investigation. For example, what is the optimal size for a retention pilot? In the validation analyses, half of the available data were used 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 would be a one- or two-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 be useful to empirically investigate this question. It is my hope that future research will address these and other related issues.

[29]For an exploration of this issue using simulations, see Web Appendix A3.6.

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.

Analytics Magazine (2016), "IBM Business Analytics Helps Improve Customer Retention," (accessed March 25, 2016), http://analytics-magazine.org/ibm-business-analytics-helps-improve-customer-retention/.

Ansari, Asim, and Carl Mela (2003), "E-Customization," *Journal of Marketing Research*, 40 (2), 131–45.

Ascarza, Eva, Peter Ebbes, Oded Netzer, and Matt Danielson (2017), "Beyond the Target Customer: Social Effects of Customer Relationship Management Campaigns," *Journal of Marketing Research*, 54 (3), 347–63.

Ascarza, Eva, Peter S. Fader, and Bruce G.S. Hardie (2017), "Marketing Models for the Customer-Centric Firm," in *Handbook of Marketing Decision Models*, 2nd ed., Berend Wierenga and Ralf van der Lans, eds. Cham, Switzerland: Springer.

Ascarza, Eva, and Bruce G.S. Hardie (2013), "A Joint Model of Usage and Churn in Contractual Settings," *Marketing Science*, 32 (4), 570–90.

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, et al. (2017), "In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions," *Customer Needs and Solutions*, available at https://doi.org/10.1007/s40547-017-0080-0.

Athey, Susan (2017), "Beyond Prediction: Using Big Data for Policy Problems," *Science*, 355 (6324), 483–85.

Athey, Susan, and Guido W. Imbens (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 113 (27), 7353–60.

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), 247–56.

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

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

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 of Marketing Research*, 36 (2), 171–86.

Braun, Michael, David A. Schweidel, and Eli M. Stein (2015), "Transaction Attributes and Customer Valuation," *Journal of Marketing Research*, 52 (6), 848–64.

Breiman, Leo (2001), "Random Forests," *Machine Learning*, 45 (1), 5–32.

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, (March), http://brucehardie.com/notes/024/reconciling_clv_formulas.pdf.

Feller, Avi, and Chris C. Holmes (2009), "Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments," unpublished manuscript, Oxford University.

Forbes Insights (2011), "Bringing 20/20 Foresight to Marketing: CMOs Seek a Clearer Picture of the Customer, Coremetrics," ftp://public.dhe.ibm.com/software/hk/pdf/product_tab_03_wp-forbes-bringing-foresight-to-marketing_4.pdf.

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*, Kurt J. Engemann, Anna M. Gil-Lafuente, and José M. Merigó-Lindahl, eds. Berlin: Springer, 123–33.

Guelman, Leo, Montserrat Guillén, and Ana M. Pérez-Marín (2015), "Uplift Random Forests," *Cybernetics and Systems*, 46 (3/4), 230–48.

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

Gutierrez, Pierre, and Jean-Yves Gérardy (2017), "Causal Inference and Uplift Modelling: A Review of the Literature," in *Proceedings of the 3rd International Conference on Predictive Applications and APIs, PMLR*, Vol. 67, 1–13, http://proceedings.mlr.press/v67/gutierrez17a.html.

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), 2902–17.

Huang, Bingquan, Mohand T. Kechadi, and Brian Buckley (2012), "Customer Churn Prediction in Telecommunications," *Expert Systems with Applications*, 39 (1), 1414–25.

Jaroszewicz, Szymon (2016), "Uplift Modeling," in *Encyclopedia of Machine Learning and Data Mining*, Claude Sammut and Geoffrey I. Webb, eds. New York: Springer, 1–6.

Lemmens, Aurélie, and Christophe Croux (2006), "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43 (2), 276–86.

Lemmens, Aurélie, and Sunil Gupta (2017), "Managing Churn to Maximize Profits," working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=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*. New York: 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), 555–65.

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), 204–11.

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), 2592–602.

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*. Sebastopol, CA: 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," in *Proceedings of Credit Scoring and Credit Control VI*. Edinburgh, Scotland, http://www.maths.ed.ac.uk/~mthdat25/uplift/cscc99-1.

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

Rogers, David (2013), "Big Data, Big Answers," AIMIA Institute (June 18), http://web.archive.org/web/20150719081543/https://aimiainstitute.aimia.com/archive/article/big-data,-big-answers/a81/.

Rosenbaum, Paul R., and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55.

Rossi, Peter, Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–40.

Rubin, Donald B. (2011), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100 (469), 322–31.

Rzepakowski, Piotr, and Szymon Jaroszewicz (2012), "Decision Trees for Uplift Modeling," in Data Mining (ICDM), 2010 IEEE 10th International Conference on IEEE, http://ieeexplore.ieee.org/document/5693998/?reload=true.

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), 471–86.

Siegel, Eric (2013), *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: John Wiley & Sons.

Sołtys, Michal, Szymon 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), 211–29.

Wager, Stefan, and Susan Athey (2017), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, (published electronically April 21), DOI: 10.1080/01621459.2017.1319839.

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