Retention futility:
Targeting high-risk customers might be ineffective

Eva Ascarza

Web Appendix

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

A1 Details about *LIFT* and *RISK* models

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

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

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

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

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

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

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

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

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

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

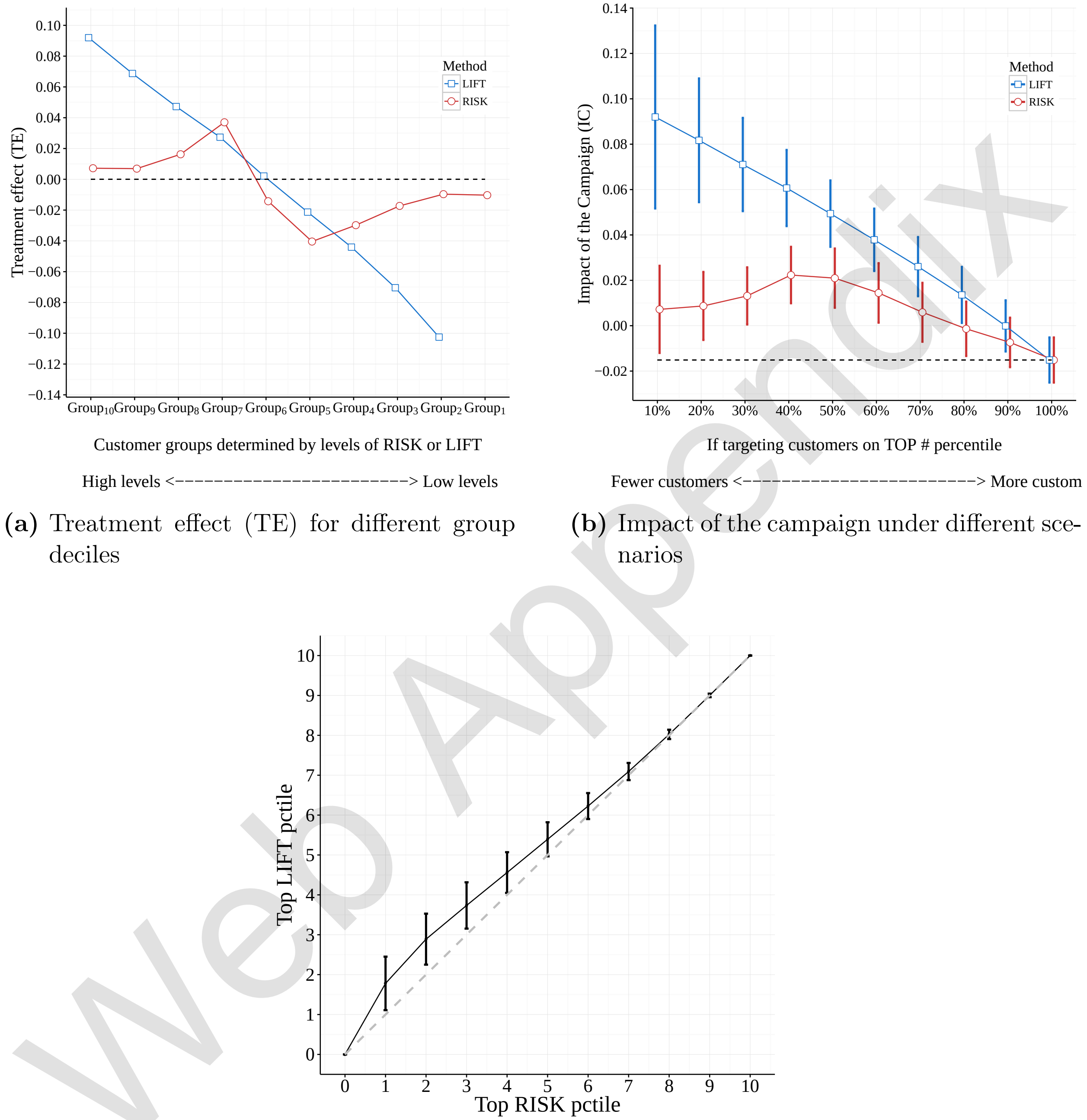
[1] See Rzepakowski and Jaroszewicz (2012) for a description of such metrics.

the 50% of customers selected in Step 1) to evaluate the model performance or as any source for model selection. As metric for accuracy I use the area under the curve (AUC) of the receiver operating characteristics (ROC). The best performing method was the LASSO approach combined with a GLM model, which provides an AUC of .907 for the first empirical application and an AUC of .658 for the second application. Following Tibshirani (1997), I standardized all variables before estimating the model.

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

A2.1 Using the same model approach (i.e., random forest) to estimate *RISK* and *LIFT*

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

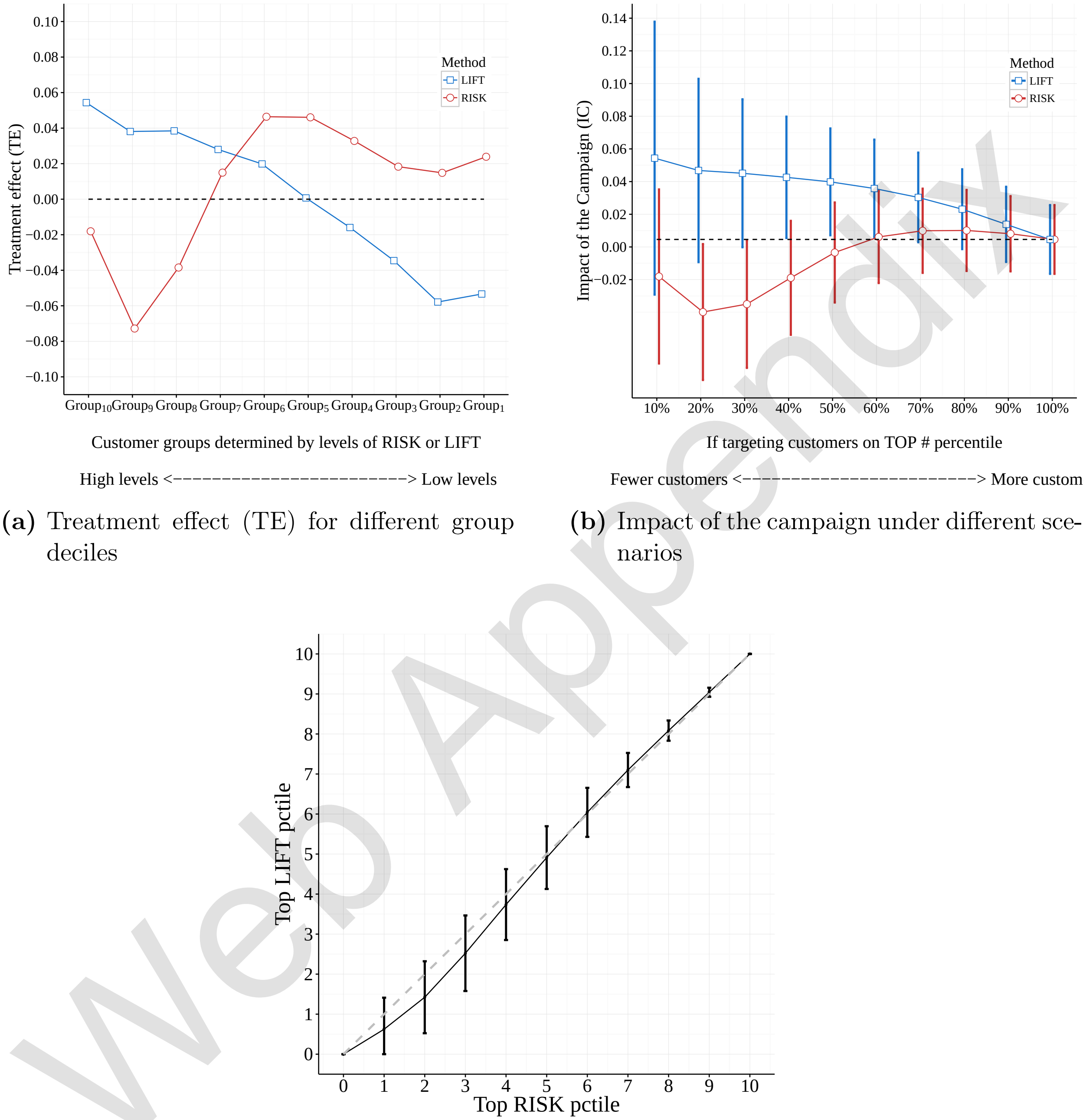


(a) Treatment effect (TE) for different group deciles

(b) Impact of the campaign under different scenarios

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

Figure A1: [Study 1] Replication of treatment effect (TE), impact of the campaign (IC), and overlap results using random forest to estimate both *RISK* and *LIFT*.



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

(b) Impact of the campaign under different scenarios

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

Figure A2: [Study 2] Replication of treatment effect (TE), impact of the campaign (IC), and overlap results using random forest to estimate both *RISK* and *LIFT*.

A2.2 Increasing the number of observations for the *RISK* model

The impact of targeting based on *RISK* or *LIFT* ultimately depends on the accuracy of the models used to predict customer RISK or LIFT. For example, even if customers should be targeted based on *LIFT*, it could be possible that the *LIFT* model is not good enough to accurately predict customers' *LIFT*, making it impossible for me to show such a relationship in the data. (Ditto for the *RISK* approach.) Therefore, given that I calibrate *RISK* and *LIFT* models (Step 3) using different sample sizes, it could be possible that the *LIFT* approach dominates the *RISK* approach because the latter model is calibrated on a smaller sample, making it, potentially, less accurate. This is unlikely given the great accuracy of the *RISK* model (as reported in Section A1.2, the AUC for the first and second applications were .907 and .658, respectively).

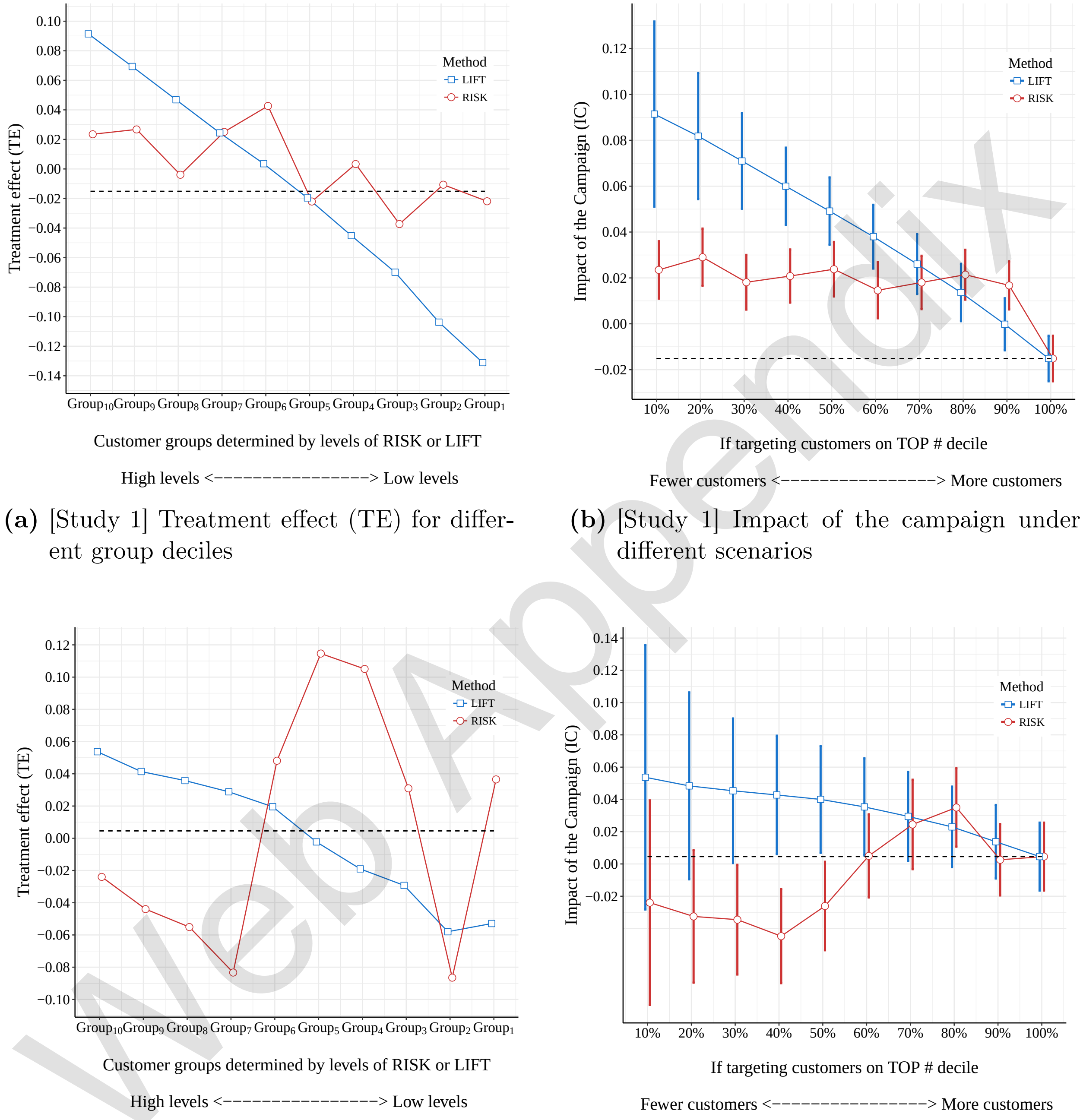
Nevertheless, I corroborate empirically that the superior performance of the *LIFT* approach is not driven by the size of the data used to calibrate the RISK model. In particular, I replicate the main analysis (Section 4.3) increasing the size of the data used in the *RISK* estimation (i.e., altering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

1. I calibrate the RISK model (Step 3) using all control observations (3,587 observations for the first application and 1,056 for the second application). The AUC using the full sample was .916 for the first application and .681 for the second application.

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

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

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



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

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

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

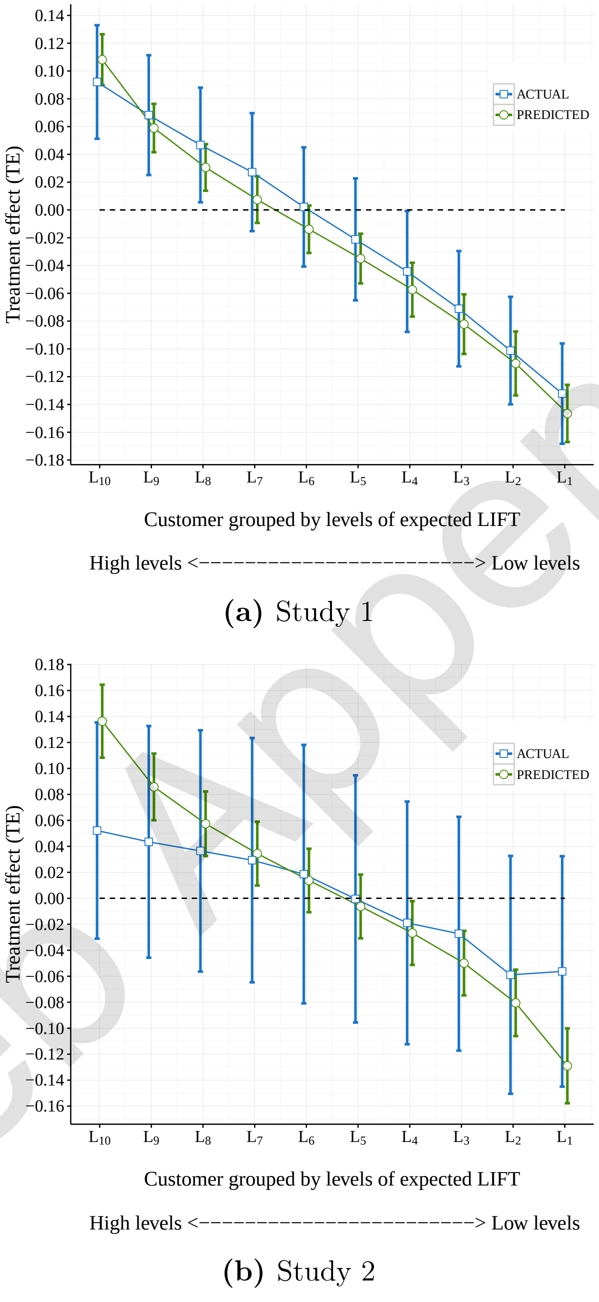
(d) [Study 2] Impact of the campaign under different scenarios

Figure A3: Replication of treatment effect (TE) and Impact of the campaign (IC) results using the full sample to calibrate the *RISK* model

A3 Additional analyses/results

A3.1 Predicted vs. acual $LIFT$

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



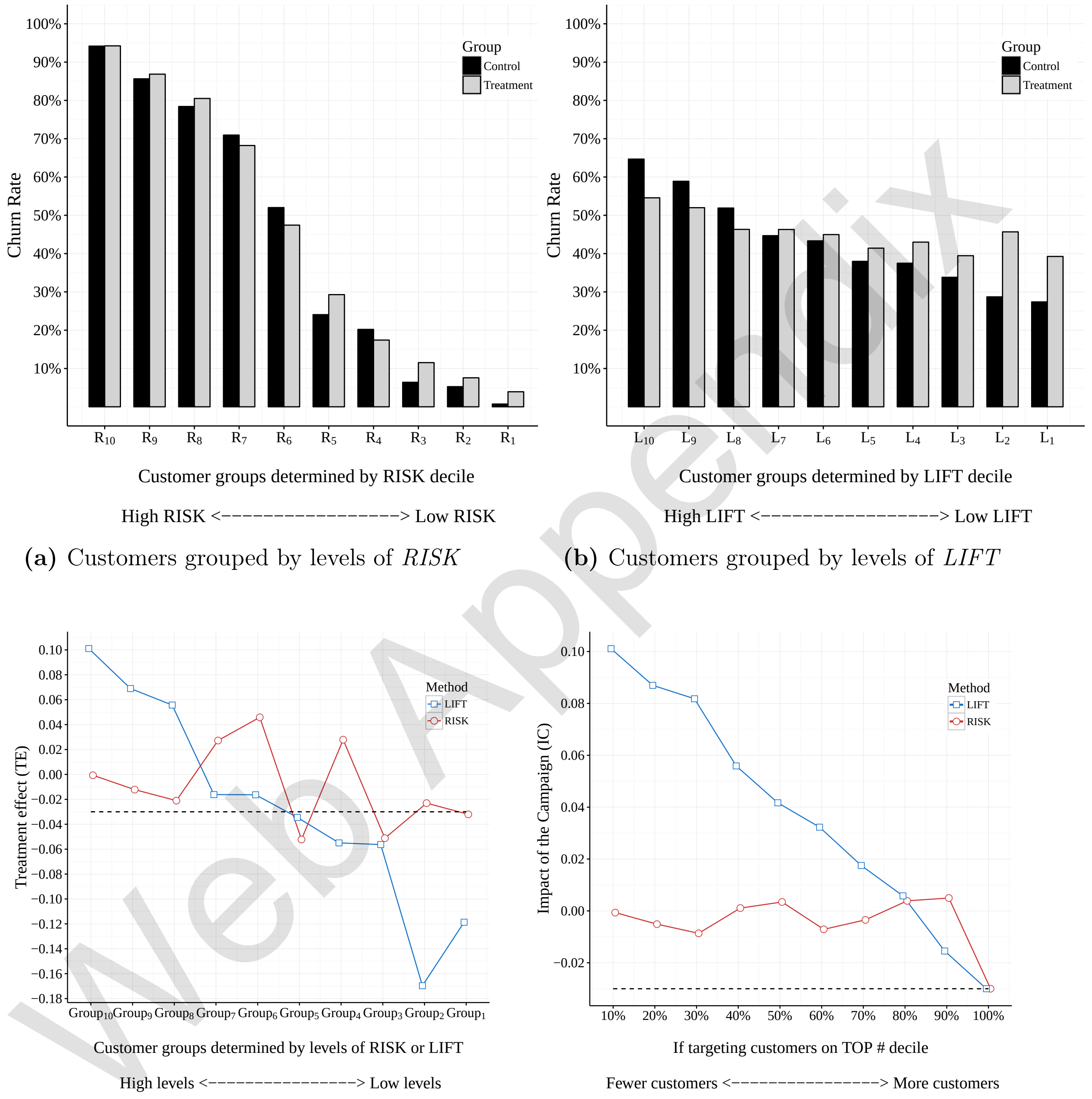
(a) Study 1

(b) Study 2

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

A3.2 Results for one iteration

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



(a) Customers grouped by levels of *RISK*

(b) Customers grouped by levels of *LIFT*

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

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

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

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

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

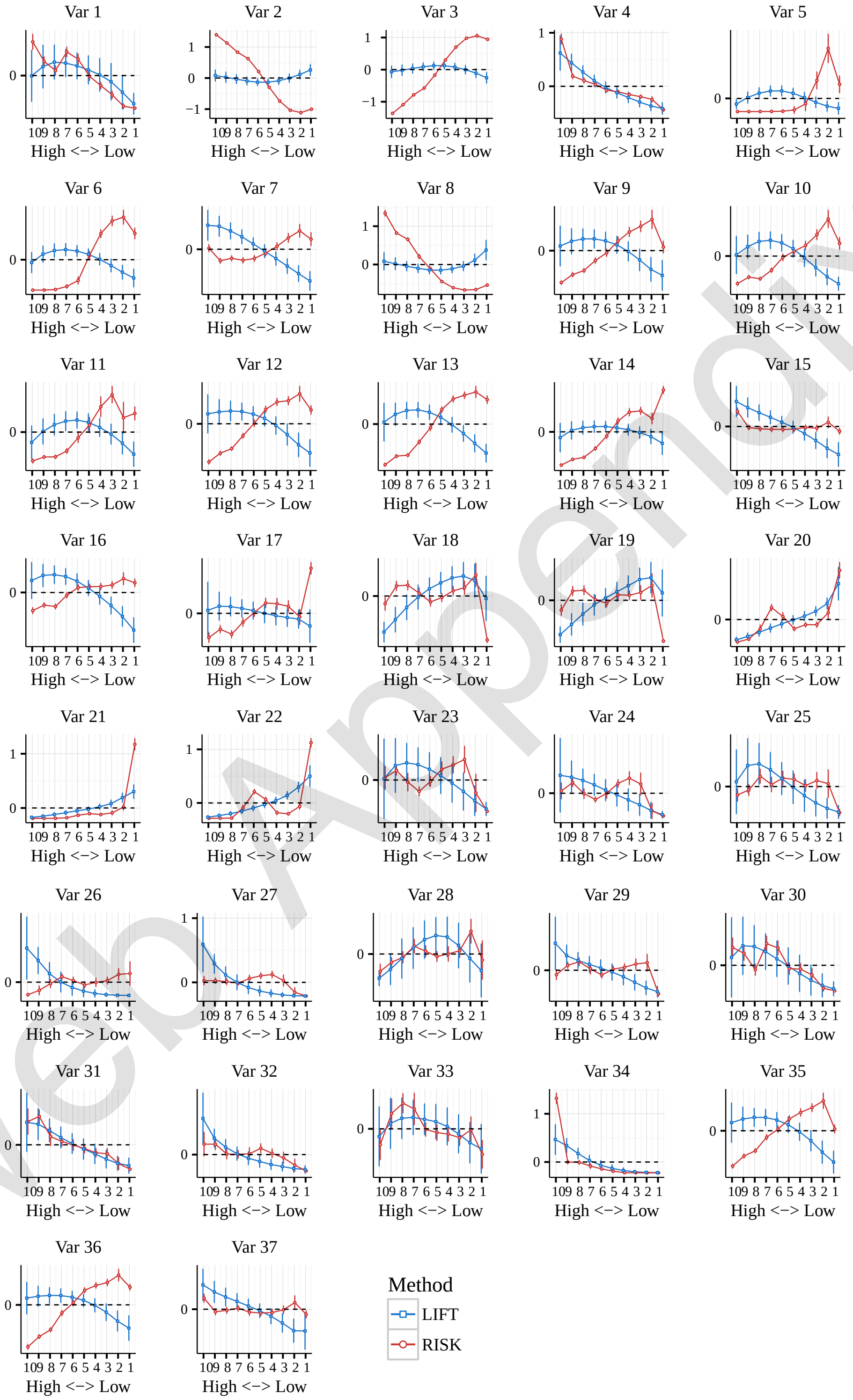


Figure A6: [Study 1] Observed characteristics as a function of *LIFT* and *RISK* deciles

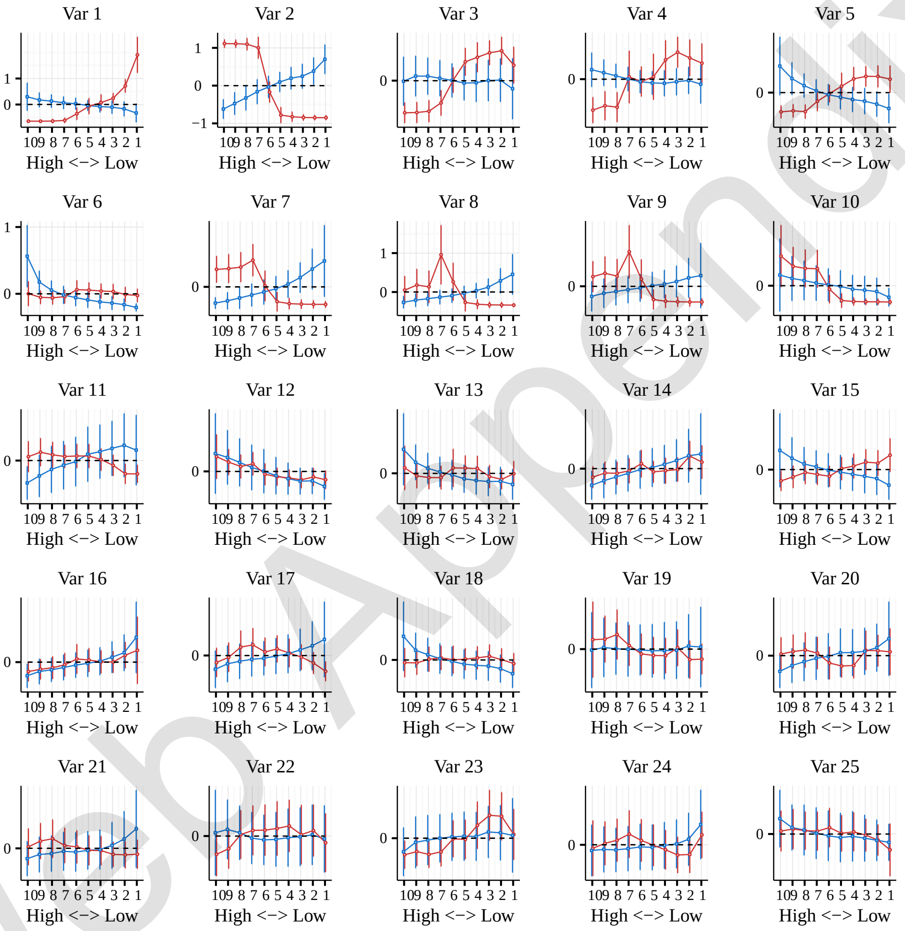


Figure A7: [Study 2] Observed characteristics (variables 1–25) as a function of *LIFT* and *RISK* deciles

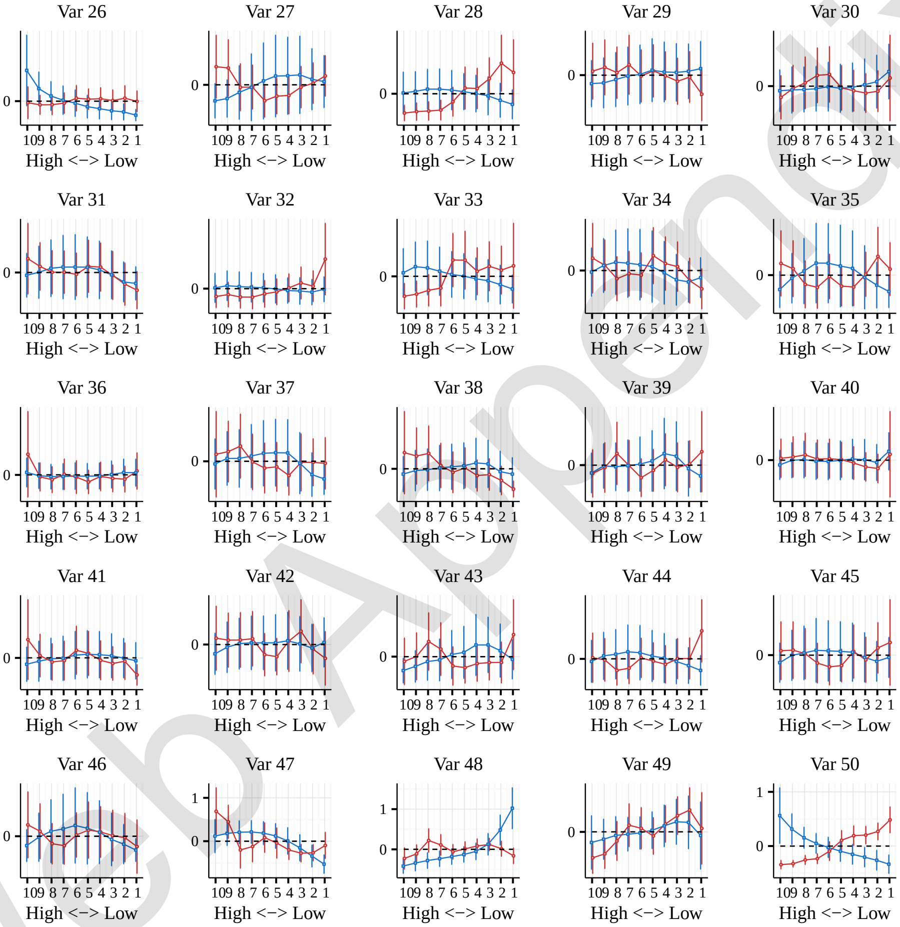


Figure A8: [Study 2] Observed characteristics (variables 25–50) as a function of *LIFT* and *RISK* deciles

A3.4 Simulation study

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

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

where

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

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

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

Z_i as follows

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

The term ρ captures correlation between X_i and Z_i, which I vary from -1 to 1, in intervals of .2. Figure A9 shows the level of overlap for all levels of ρ. Comparing these figures with those obtained in the empirical applications, it seems that in the first context (telecommunications) the correlation between*RISK* and *LIFT* is close to .2. Similarly, the resulting treatment effects (Figure A10) are very similar to those obtained using the real data. Comparing the results from the second application (special interest membership), the correlation between *RISK* and *LIFT* is clearly negative, possibly around $-.2$.

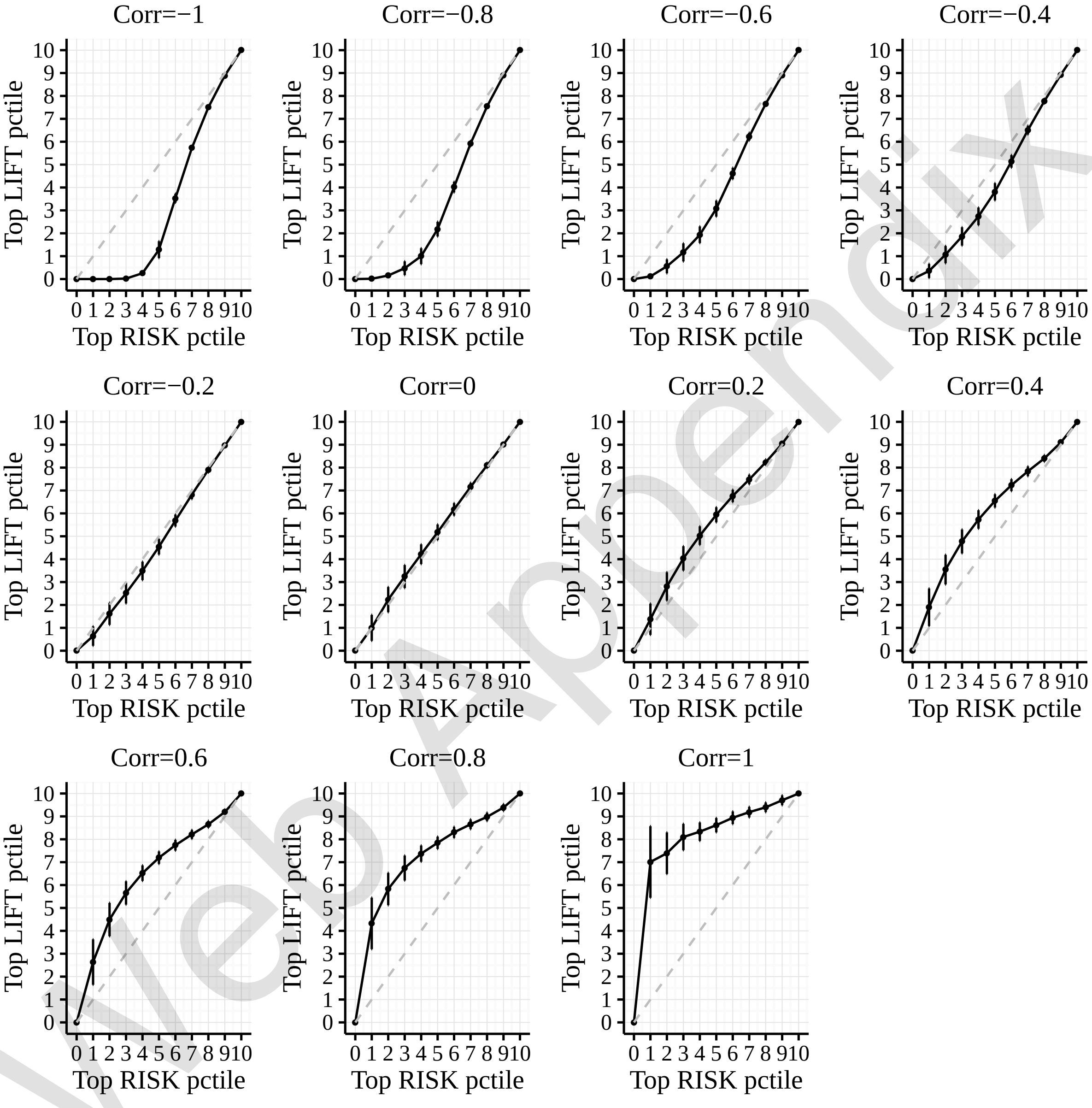


Figure A9: Level of overlap across groups defined by top *RISK* deciles vs. top *LIFT* deciles.The (dotted) 45° line represents the level of overlap if there was no relationship between the two groups

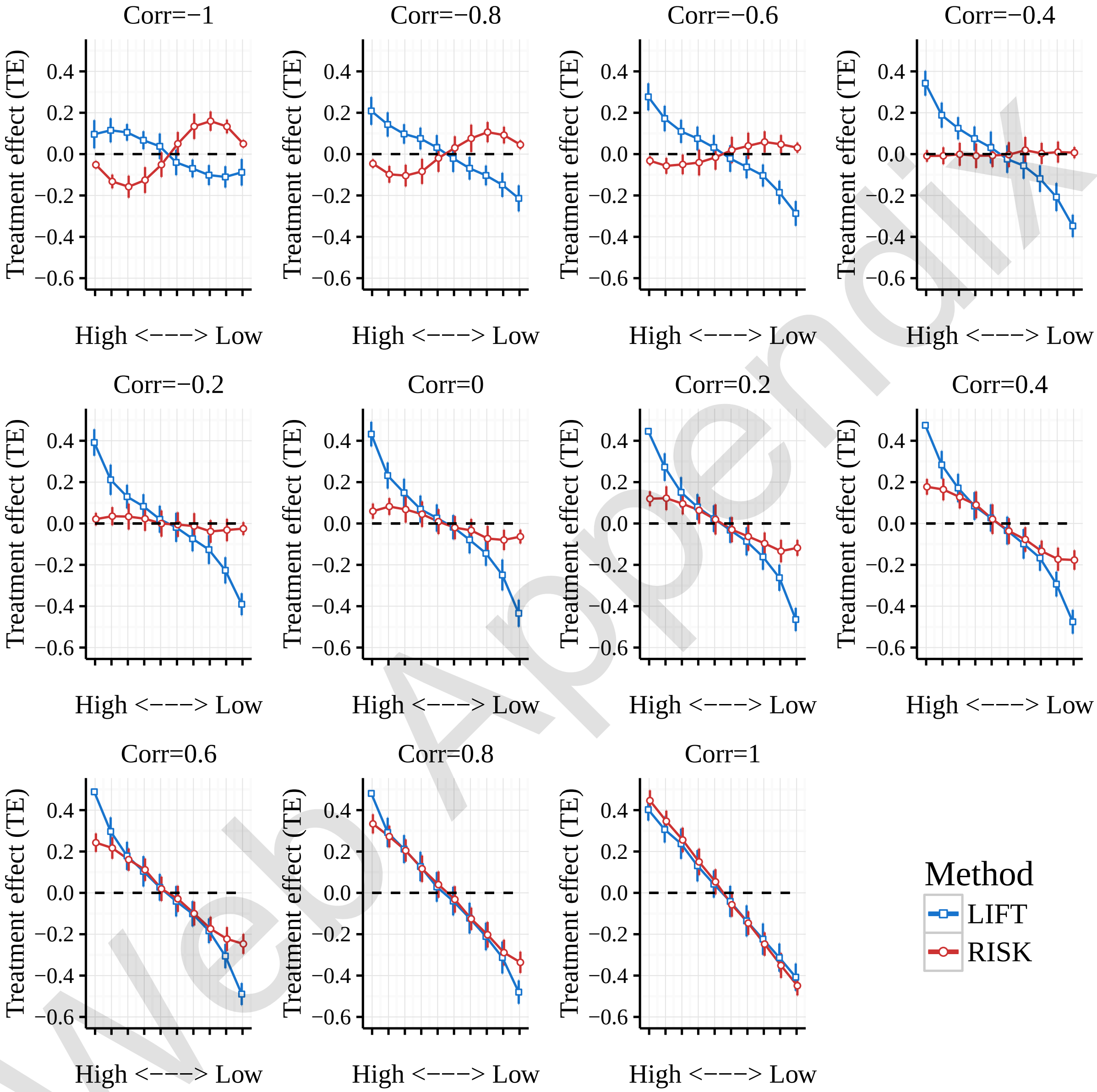


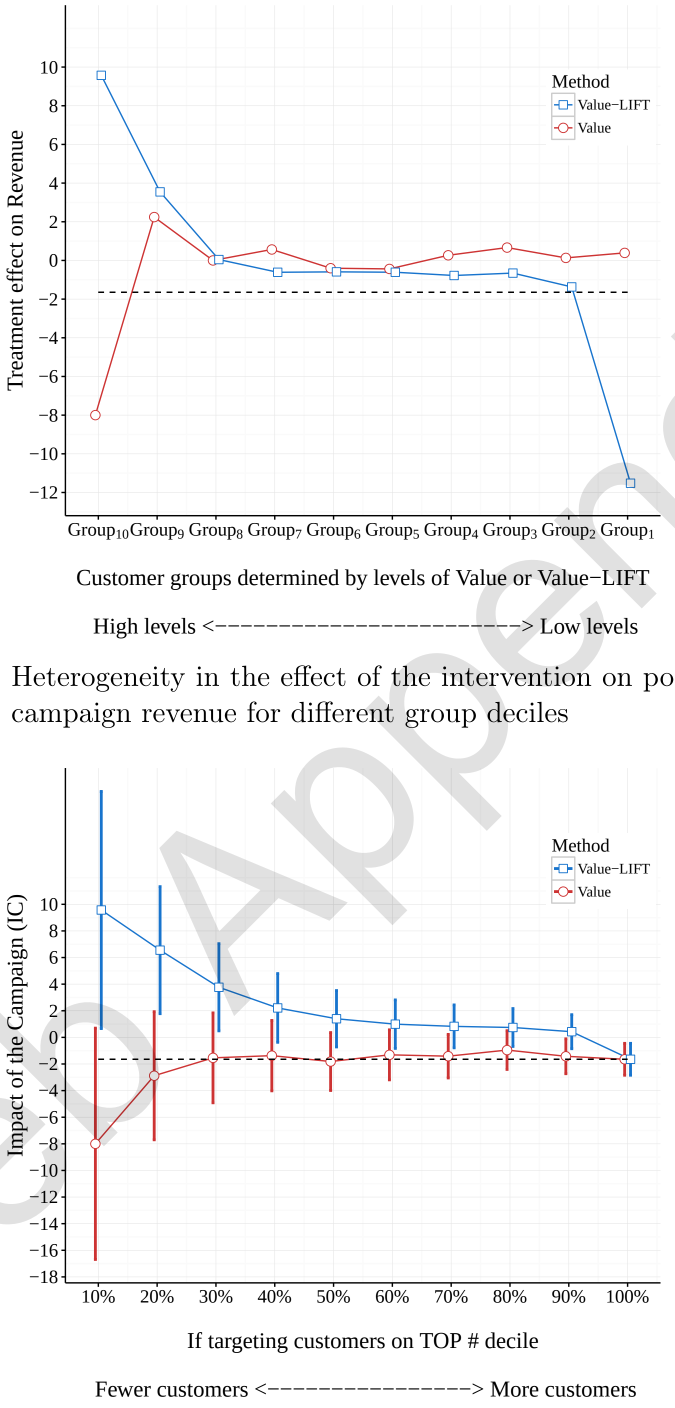
Figure A10: Treatment effect (TE) for different group deciles, depending on whether customers are grouped by levels of *RISK* (represented by the squares) or *LIFT* (represented by the circles). The dotted (straight) line corresponds to the average effect of the campaign if the firm targeted randomly

A3.5 *Value-LIFT*

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

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

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



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

(b) Impact of the campaign for different top deciles

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

A3.6 Simulation results for different churn rates

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

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

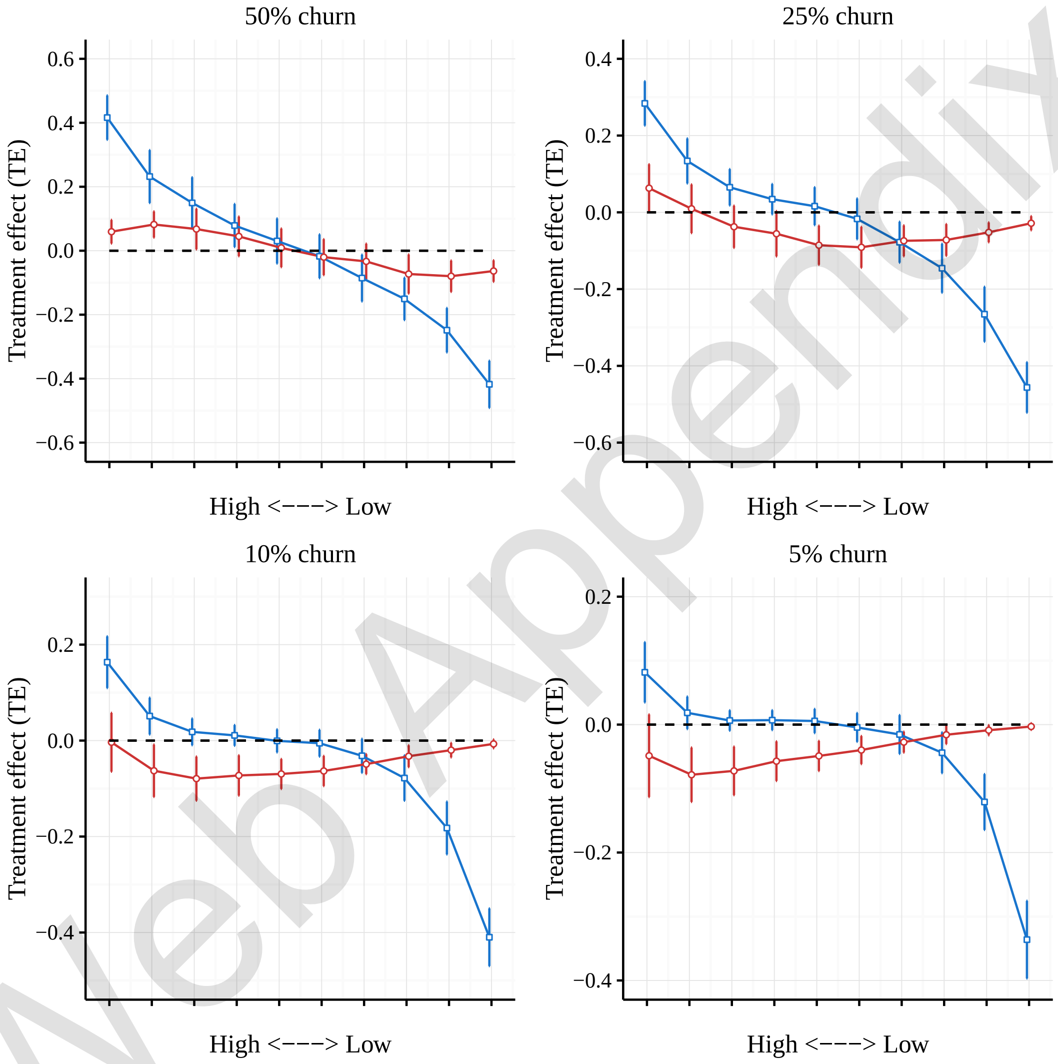


Figure A12: Treatment effect (TE) for simulated data. Varing churn rate from 50% (Top left) to 5% (Bottom left)

References

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

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

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

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