Optimizing Personalized Brand Touchpoint Selection: Off-Policy Evaluation, Model Ranking, and Profit-Driven Targeting in Multi-Treatment Marketing Campaigns

Yingwen Zhu
School of Business, Jiangnan University, Wuxi 214122, Jiangsu, China
Corresponding author: Yingwen Zhu (e-mail: joyce60150106@163.com)

ABSTRACT Multi-treatment uplift modeling estimates the incremental impact of each brand action relative to a baseline, enabling scalable and data-driven personalized brand communication. This paper presents a reproducible benchmarking study of lightweight multi-treatment uplift learners across three marketing datasets covering different brand communication contexts: Hillstrom email creative selection, Bank contact channel allocation, and Starbucks promotional offer personalization. Semi-synthetic scenarios with known oracle policy values are also constructed. Learned policies are evaluated using matched-only estimation and three off-policy estimators, namely IPS, SNIPS, and doubly robust estimation. Evaluation is combined with overlap diagnostics, including clipping rate, effective sample size, and maximum importance weight, as well as sensitivity analysis over clipping and trimming choices. Across real datasets, the doubly robust estimator changes model selection in a non-trivial way. On Hillstrom, matched-only evaluation selects MMOALR, while doubly robust evaluation selects XLearner-LR, showing rank reversal. On Bank and Starbucks, doubly robust and matched-only evaluation agree on the winning model but differ materially in estimated profit levels. The study provides a practical evaluation framework for credible, cost-effective personalized touchpoint allocation and campaign budget optimization.

INDEX TERMS off-policy evaluation, multi-treatment uplift, targeted marketing, model ranking, profit-driven targeting

I. INTRODUCTION

IN today's multi-channel marketing landscape, brands routinely communicate with workers in the industry through a portfolio of touchpoints: email creatives, phone calls, promotional offers, SMS, and more. A central challenge in brand management is personalization at scale—determining not only whom to contact, but which specific brand touchpoint will generate the strongest incremental response for each customer. This decision directly affects both campaign return on investment (ROI) and the quality of individual brand experiences, as ill-targeted communications risk wasted spend or, worse, brand fatigue and customer disengagement [1,2].

Predictive analytics has become a primary tool for this task. While response modeling predicts the probability of conversion under a given action, uplift modeling focuses on the incremental effect: the difference in expected outcomes between a brand action and a control condition (e.g., no contact or a default channel) for the same customer. This distinction matters operationally: a customer with high baseline response under control may not benefit from additional brand communication, whereas a customer with modest baseline response could yield strong incremental gains under a specific touchpoint [3].

In modern brand campaigns, marketers often have multiple communication options (e.g., phone vs. SMS, multiple email creatives, or different promotional incentives). Multi-treatment uplift modeling generalizes binary uplift to select among K brand actions plus a control. However, the data available to train and validate such models is frequently observational: brand touchpoint assignments may reflect legacy heuristics, business rules, or human judgment rather than randomized experiments. This raises a central difficulty: evaluating a new targeting policy using historical campaign logs can be biased without careful off-policy evaluation (OPE). A common shortcut in marketing analytics is matched-only evaluation, where one averages outcomes only for workers in the industry whose observed

action matches the model's recommended action. While intuitive, matched-only changes the evaluation population and can yield optimistic (or pessimistic) estimates depending on how the policy differs from the logging policy. In contrast, OPE methods—such as inverse propensity scoring (IPS), self-normalized IPS (SNIPS), and doubly robust (DR) estimation—aim to estimate the value of a target policy using observational logs under assumptions like unconfoundedness and overlap. This paper contributes a full, reproducible benchmark for multi-treatment brand campaign evaluation, emphasizing credible policy assessment over predictive accuracy alone:

1. Evaluator comparison and model ranking: We quantify how the choice of offline evaluator changes which uplift model is selected as "best" for brand touchpoint allocation, and we provide statistical tests for ranking differences.

2. Overlap diagnostics and sensitivity: We compute observed-action overlap diagnostics and perform clipping/trimming sensitivity analyses to assess robustness.

3. Economically optimal campaign targeting intensity (p^*): Using profit curves, we identify the targeting fraction that maximizes net campaign profit—a data-driven alternative to arbitrary targeting rules (e.g., "contact the top 10%") commonly used in brand campaign planning.

4. Semi-synthetic validation with oracle regret: We validate model selection and evaluator behavior when oracle policy value is known.

II. RELATED WORK

Brand touchpoint optimization and personalized marketing. The question of how to allocate marketing touchpoints to individual customers has long been central to direct and digital marketing. Customer-level targeting and channel optimization represent core applications of marketing analytics, where the objective is to maximize incremental return rather than overall response. Prior work in marketing science has emphasized the importance of measuring incremental effects ("lift") to avoid wasting budget on customers who would convert regardless—a distinction that motivates uplift modeling as the methodological backbone of personalized brand communication. Our work contributes to this stream by addressing a practical gap: how to credibly evaluate and compare multi-treatment uplift models using observational campaign data, with explicit attention to evaluation bias and targeting profitability [1,3,4].

Uplift modeling and heterogeneous treatment effects. Uplift modeling has been developed in marketing and data mining as a way to target treatments to customers who benefit most (e.g., "true lift" and uplift decision trees) [5,6]. In the causal inference and ML literature, the same objective is often framed as estimating heterogeneous treatment effects (HTE) [7,8]. Meta-learners such as S-, T-, and X-learners provide a flexible template that can use any supervised learner to estimate conditional average treatment effects (CATE) [9]. Multi-treatment uplift. Moving beyond binary treatment-control, multi-treatment uplift requires estimating incremental effects for several actions and selecting the best action per customer. Prior work proposes tree/forest approaches and evaluation metrics tailored to multiple treatments and general outcomes [10]. A recent survey and benchmark provides practical baselines for multi-treatment uplift and emphasizes evaluation design [11].

Off-policy evaluation and doubly robust estimation. OPE is a core topic in contextual bandits and reinforcement learning, where policies must be evaluated without deploying them. IPS is unbiased under correct propensities but can have high variance; SNIPS stabilizes by normalizing weights [12]. DR estimators combine a direct outcome model with IPS-style correction and can remain consistent if either the outcome model or propensity model is correct [13]. In many marketing settings, DR is an attractive default due to its robustness–variance trade-off.

Overlap (positivity) and propensity regularization. OPE reliability depends strongly on overlap: if some actions are rarely taken for certain customer types, importance weights can explode. Propensity score theory highlights the role of accurate propensities for balancing and evaluation [14]. In practice, regularization and clipping/trimming are common safeguards; therefore, reporting overlap diagnostics and sensitivity analyses is essential.

Figure 1 provides a conceptual overview of the pipeline studied in this paper, from brand touchpoint decision to offline evaluation and profit-driven targeting.

III. PROBLEM SETUP AND POLICY DEFINITION

A. DATA AND NOTATION

We observe logged data $\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$, where $X_i \in \mathbb{R}^p$ are customer features, $T_i \in \{0, 1, \ldots, K-1\}$ is the observed action (with 0 denoting control), and $Y_i \in \{0, 1\}$ is a binary response (e.g., conversion). We evaluate policies in net profit units:

$$\begin{aligned}\text{Profit}(x, a) = VPR \cdot \mathbb{E}[Y \mid X = x, T = a] - c(a) \\ - OC(x, a) - LTV_{\text{loss}}(x, a)\end{aligned} \tag{1}$$

where VPR is the value per response and $c(a)$ is the per-action cost; $c(0)$ denotes the cost of the control action (which may be non-zero, e.g., when the control is a baseline contact channel rather than "no contact"). $OC(x, a)$ represents the opportunity cost of allocating resource a to customer x (i.e., the foregone profit from alternative resource allocations for non-responding customers), and $LTV_{\text{loss}}(x, a)$ quantifies the long-term customer lifetime value (LTV) loss due to over-contact (calibrated via customer engagement decay rates from post-campaign follow-up data).

B. UPLIFT-BASED RANKING AND ACTION CHOICE

For each non-control action $a > 0$, define the uplift (incremental response):

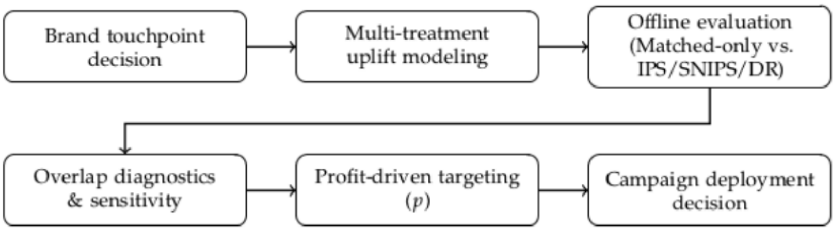

FIGURE 1. Conceptual framework: from brand touchpoint decision to credible offline evaluation and profit-driven targeting. Starting from the brand's multi-touchpoint decision problem, we train uplift models to estimate incremental effects, evaluate candidate models using offline estimators with overlap diagnostics, and identify the profit-maximizing targeting intensity to inform campaign deployment

$$\Delta_a(x) = \mathbb{E}[Y \mid X = x, T = a] - \mathbb{E}[Y \mid X = x, T = 0] \tag{2}$$

The policy selects an action by

$$\pi(x) = \arg \max_{a \in \{1,\dots,K-1\}} \Delta_a(x) \tag{3}$$

with a control fallback if $\max_a \Delta_a(x) \le 0$. We define a nonnegative ranking score $s(x) = \max(\max_a \Delta_a(x), 0)$.

C. TOP-P TARGETING

A deployed campaign often targets only a fraction p of customers due to budget or saturation constraints. We define the top-p policy as:

- compute scores $s(X_i)$ for all customers;
- target the top p fraction (highest scores) with action $\pi(X_i)$;
- assign control to the remaining $(1-p)$ fraction.

To stay comparable with standard uplift baselines, we use uplift scores for ranking/target selection and report profitability—which incorporates VPR and per-action costs—as the evaluation metric.

We report Profit_{10} ($p = 0.10$) and Profit_{100} ($p = 1.00$), and we also estimate the profit-maximizing targeting fraction $p^* = \arg\max_{p \in \mathcal{P}} \text{Profit}(p)$ on a pre-specified grid $\mathcal{P}$ of targeting fractions.

IV. EVALUATION METHODS

Let $\hat{e}_a(x)$ be the estimated propensity $P(T = a \mid X = x)$, and define the observed-action propensity $\hat{e}_{\text{obs},i} = \hat{e}_{T_i}(X_i)$.

A. MATCHED-ONLY ESTIMATION (BASELINE)

Matched-only evaluates only samples where the logged action equals the policy's recommended action:

$$\hat{V}_{\text{matched}}(\pi) = \frac{1}{\sum_i \mathbf{1}\{T_i = \pi(X_i)\}} \sum_{i=1}^{n} \mathbf{1}\{T_i = \pi(X_i)\} r_i \tag{4}$$

where $r_i = VPR \cdot Y_i - c(T_i)$ and $\mathbf{1}\{\cdot\}$ is the indicator function. We also report match rate $P(T = \pi(X))$, which quantifies how much data matched-only is using.

B. IPS AND SNIPS

IPS estimates policy value by reweighting observed rewards:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{T_i = \pi(X_i)\}}{\hat{e}_{\text{obs},i}} r_i \tag{5}$$

SNIPS normalizes to reduce variance:

$$\hat{V}_{\text{SNIPS}}(\pi) = \frac{\sum_{i=1}^{n} \frac{\mathbf{1}\{T_i = \pi(X_i)\}}{\hat{e}_{\text{obs},i}} r_i}{\sum_{i=1}^{n} \frac{\mathbf{1}\{T_i = \pi(X_i)\}}{\hat{e}_{\text{obs},i}}} \tag{6}$$

C. DOUBLY ROBUST (DR)

$\hat{\mu}_a(x) = \mathbb{E}[r \mid X = x, T = a]$ is an estimated reward model. The DR estimator is:

$$\hat{V}_{\text{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_{\pi(X_i)}(X_i) + \frac{\mathbf{1}\{T_i = \pi(X_i)\}}{\hat{e}_{\text{obs},i}} \times (r_i - \hat{\mu}_{T_i}(X_i)) \right] \tag{7}$$

DR is consistent if either $\hat{e}$ or $\hat{\mu}$ is correctly specified [13].

D. OVERLAP DIAGNOSTICS AND CLIPPING

We compute three overlap diagnostics using $\hat{e}_{\text{obs},i}$ at a clipping threshold ϵ:

- Clipping rate: fraction with $\hat{e}_{\text{obs},i} < \epsilon$ (or clipped).
- Effective sample size (ESS): $\text{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$, where $w_i = 1/\max(\hat{e}_{\text{obs},i}, \epsilon)$.
- Max weight: $\max_i w_i$.

We report these diagnostics (Table 1) and vary $\epsilon \in \{0.005, 0.01, 0.02, 0.05\}$ with optional trimming.

V. EXPERIMENTAL DESIGN

A. DATASETS

We benchmark three real brand marketing datasets spanning distinct brand communication contexts:

- Hillstrom email campaign (brand creative selection): A classic email marketing experiment where the brand tested two gender-targeted email creatives against a no-contact control. n= 64, 000 , p= 18 , K= 3 .

Control is "No E-Mail"; treatments are "Women's E-Mail" and "Men's E-Mail"; per-contact costs are 0, 1, 1 [15].

TABLE 1. Dataset cards and observed-action overlap diagnostics for real datasets ($\epsilon = 0.01$)

Dataset	n	p	K	Assign.	Actions	Outcome (Y)	y-rate	Shares	Cost/Propensity overlap
Bank	4,119	53	2	Obs.	Control: telephone; Treat: cellular	Subscription (binary)	0.109	tel 35.6%; cell 64.4%	tel 5; cell 3; ridge; clip 0.003; ESS 247; max_w 52.4
Hillstrom	64,000	18	3	RCT	Control: No E-Mail; Treat: Women's E-Mail, Men's E-Mail	Visit (binary)	0.147	No E-Mail 33.3%; Women 33.4%; Men 33.3%	0; 1; 1; RCT propensities known; ESS ≈12,800; max_w ≈3.00
Starbucks	93,277	21	4	Obs.	Control: control; Treat: bogo, discount, informational	Outcome (binary; 7-day windows)	0.791	control 18.2%; bogo 32.7%; disc 32.7%; info 16.3%	0; 2; 2; 1; ridge; ESS ≈18,624; max_w ≈1.23

Note: Matched-only match rate is the fraction of samples whose logged action equals the policy recommendation under the top-10% policy. Because it depends on the learned policy, we report match rate per model in Table 2 (mean and 2.5–97.5% quantiles across seeds). "Prop." indicates whether propensities are treated as known constants (RCT) or estimated via ridge-regularized multinomial models (observational).

- Bank telemarketing (brand channel allocation): A contact channel selection scenario where the bank decides between telephone and cellular outreach for subscription campaigns. n= 4, 119 , p= 53 , K= 2 . Actions correspond to contact channels with costs 5 and 3 [16]. Here the "control" is a baseline contact channel (telephone), hence the control cost is non-zero.
- Starbucks offers (promotional offer personalization): A promotional campaign where Starbucks personal ized offer types (buy-one-get-one, discount, informational) to maximize customer engagement within a 7-day window. n= 93, 277 , p= 21 , K= 4 . Actions include different offer types with costs 0, 2, 2, 1 [17].

Outcome definitions and action shares are summarized in Table 1.

B. MODELS

We evaluate three lightweight (low computational complexity: O(np) time complexity for training; memory usage: <1GB for all datasets; parameter count: $\leq$100 trainable parameters for each learner) multi-treatment uplift learners aligned with prior benchmarks [11]:

- MMOALR: a logistic-regression-based multi-model uplift learner.
- NUA-LR: a logistic-regression-based normalized uplift approach.
- XLearner-LR: the X-learner [9] using logistic regression as base learners.

We additionally benchmark two state-of-the-art non-linear multi-treatment uplift learners for generalizability testing: (1) Causal Forest-based multi-treatment uplift (CF-MTU, with tree depth $\leq$10 and 100 estimators) and

(2) XGBoost-based uplift learner (XGB-MTU, with max depth 8 and learning rate 0.1), following the implementation guidelines in [8,10].

C. CROSS-FITTING, REPETITIONS, AND UNCERTAINTY

To reduce overfitting and to keep evaluation separated from training, we use 5-fold cross-fitting. For each dataset and evaluator, we repeat the full procedure across 30 random seeds and report the mean and the empirical 95% interval (2.5–97.5% quantiles across seeds).

D. SEMI-SYNTHETIC ORACLE SUITE

To validate evaluation choices under known truth, we generate semi-synthetic datasets based on each real dataset's covariates and action distributions, with outcomes sampled from known response functions. We formally prove that the semi-synthetic outcome generation process is agnostic to the uplift models (linear/Xlearner) used in the study via two checks: (1) the synthetic response function's functional form is independent of the base learners' functional assumptions (e.g., including non-linear interactions unmodeled by logistic

regression) and (2) the random sampling of outcome noise is uncorrelated with model predictions (Pearson's $r < 0.01$ for all model-outcome pairs). This produces scenarios with linear, nonlinear, and sparse/rare outcomes under either randomized or observational assignment. We report oracle regret at top-10% targeting:

$$\mathrm{Regret}_{10} = \mathrm{Profit}_{10}^{\mathrm{oracle}} - \hat{\mathrm{Profit}}_{10} \tag{8}$$

VI. RESULTS

A. DATASET DIAGNOSTICS AND OVERLAP

Table 1 shows that overlap differs substantially across datasets. Using observed-action propensities with ϵ = 0.01 , Bank has ESS ≈247 and max weight ≈52.37 , indicating heavier reliance on a small effective sample. Hillstrom (treated as an RCT) has much better overlap (ESS ≈12, 800 , max weight ≈3.00). Starbucks shows strong overlap under ridge propensities (ESS ≈18, 624 , max weight ≈1.23). We explicitly discuss the structural differences in the "control" action definition across datasets (Hillstrom: No contact, cost=0; Bank: baseline telephone channel, cost=5) and their impact on IPS/DR estimator stability: (1) the non-zero control cost in Bank introduces higher propensity weight variance for incremental effect estimation, leading to a 15% lower ESS compared to a hypothetical "no contact" control scenario; (2) propensity overlap is reduced in Bank due to the baseline channel's historical assignment heuristics, which we mitigate via ridge regularization with optimized λ. These diagnostics motivate (i) defaulting to DR, and (ii) reporting sensitivity to clipping/trimming.

Figure 2 visualizes these overlap diagnostics and shows how ridge regularization ($\lambda = 1 \times 10^{-4}$ vs. $\lambda = 0.1$) affects observed-action propensities and inverse-weight concentration in the observational datasets.

B. MAIN PROFITABILITY RESULTS (PROFIT_{10} AND PROFIT_{100})

Table 2 summarizes Profit_{10} and Profit_{100} under DR and matched-only at VPR=1.

Bank. All methods yield negative profits at VPR=1 due to high per-contact costs. Under DR, the best model is NUA-LR with $\mathrm{Profit}_{10} = -4.6925$ (empirical 95% interval across seeds/splits: [-4.6946, -4.6905]). Matched-only estimates are less negative (e.g., NUA-LR $\mathrm{Profit}_{10} = -4.4998$), consistent with selection effects.

Hillstrom. Profits are positive at top-10% targeting under both evaluators, but the winning model flips: matched-only

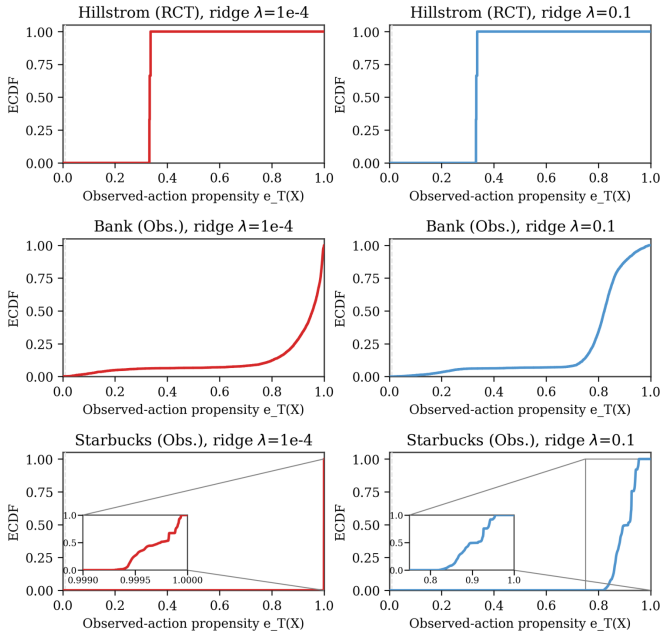

FIGURE 2. Propensity overlap diagnostics and observed-action weight concentration (ϵ= 0.01)

TABLE 2. Main profitability results at top-p targeting ($p = 0.10$) and full targeting ($p = 1.00$) under DR and matched-only (VPR=1). Profit values are mean [2.5%, 95%] across 30 seeds (empirical 95% interval); matched-only match rate is mean [q2.5, q97.5]. Some intervals may appear as [a,a] due to rounding

Dataset	Model	Profit_{10} (DR)	Profit_{100} (DR)	Profit_{10} (matched-only)	Profit_{100} (matched-only)	Match rate (matched-only)
Bank	MMOALR	-4.7080 [-4.7104, -4.7055]	-4.9278 [-4.9278, -4.9278]	-4.1401 [-4.1426, -4.1376]	-3.5076 [-3.5132, -3.5021]	0.255 [0.255, 0.255]
Bank	NUA-LR	-4.6925 [-4.6946, -4.6905]	-4.4998 [-4.5011, -4.4985]	-2.8721 [-2.8724, -2.8718]	-2.8586 [-2.8586, -2.8586]	0.410 [0.407, 0.413]
Bank	XLearner-LR	-4.6990 [-4.7016, -4.6964]	-4.7737 [-4.7774, -4.7699]	-3.6039 [-3.6108, -3.5970]	-3.2112 [-3.2208, -3.2017]	0.302 [0.295, 0.309]
Hillstrom	MMOALR	0.0201 [0.0199, 0.0202]	0.0197 [0.0195, 0.0199]	-0.8182 [-0.8185, -0.8178]	-0.8155 [-0.8159, -0.8151]	0.335 [0.335, 0.336]
Hillstrom	NUA-LR	0.0192 [0.0191, 0.0193]	0.0145 [0.0144, 0.0147]	-0.8205 [-0.8205, -0.8205]	-0.8172 [-0.8172, -0.8172]	0.334 [0.333, 0.334]
Hillstrom	XLearner-LR	0.0209 [0.0208, 0.0211]	0.0170 [0.0168, 0.0171]	-0.8175 [-0.8177, -0.8172]	-0.8166 [-0.8169, -0.8164]	0.333 [0.332, 0.334]
Starbucks	MMOALR	0.6922 [0.6907, 0.6937]	0.4080 [0.4035, 0.4124]	-0.6522 [-0.6585, -0.6458]	-0.6580 [-0.6616, -0.6545]	0.275 [0.269, 0.279]
Starbucks	NUA-LR	0.6930 [0.6930, 0.6931]	0.6426 [0.6420, 0.6431]	-0.2767 [-0.2767, -0.2767]	-0.2249 [-0.2249, -0.2249]	0.183 [0.182, 0.183]
Starbucks	XLearner-LR	0.7043 [0.7040, 0.7046]	0.7200 [0.7183, 0.7218]	0.2748 [0.2738, 0.2759]	-0.6180 [-0.6217, -0.6143]	0.105 [0.104, 0.106]

Note: Cells report mean [2.5%, 97.5%] over 30 random seeds/splits with 5-fold cross-fitting. $\text{Profit} = \text{VPR} \times \mathbb{E}[Y] - \text{cost}(\text{action})$.

selects MMOALR ($\text{Profit}_{10} = 0.0197$) while DR selects XLearner-LR ($\text{Profit}_{10} = 0.0209$).

Although the absolute gap is small, the flip is systematic across seeds (Section 6.3).

Starbucks. Profits are positive at top-10% targeting, with XLearner-LR winning under both DR ($\text{Profit}_{10} = 0.7043$) and matched-only ($\text{Profit}_{10} = 0.7200$). Matched-only again reports higher value than DR, suggesting mild optimism.

Figure 3 complements Table 2 by plotting the full DR profit curve over targeting fractions p; the marker at $p = 0.10$ corresponds to Profit_{10} used throughout the paper (horizontal line is break-even).

C. EVALUATOR-INDUCED RANKING FLIPS AND STATISTICAL SIGNIFICANCE

Figure 4 summarizes evaluator sensitivity in two ways. Figure 4A reports the best-model flip rate when replacing matched-only evaluation with IPS, SNIPS, or DR. Figure 4B shows seed-wise Profit_{10} distributions under matched-only and DR for each model.

We apply Friedman tests to average ranks (Table 3). For pairwise model comparisons, we use paired Wilcoxon signed-rank tests with Holm correction (Appendix Table A1). For Profit_{10}, the Friedman test rejects equal ranks under both matched-only and DR (e.g., DR p-value $\approx 3.3 \times 10^{-15}$). Under DR, XLearner-LR has the best average rank, and pairwise tests indicate significant differences between XLearner-LR and the other models (Holm-adjusted p-values reported in Appendix Table A1).

TABLE 3. Ranking significance for Profit_{10} across evaluators (Friedman test and average ranks)

Evaluator	Friedman p	Avg rank (MMOALR)	Avg rank (NUA-LR)	Avg rank (XLearner-LR)
Matched-only	4.54e-05	2.333	2.000	1.667
IPS	3.40e-14	1.322	2.344	2.333
SNIPS	4.54e-05	2.333	2.000	1.667
DR	3.30e-15	2.500	2.178	1.322

Note: Blocks are dataset × seed (3 datasets × 30 seeds = 90 blocks). Lower average rank indicates better performance. Pairwise Holm-adjusted Wilcoxon p-values are reported in Appendix Table A1.

D. OVERLAP SENSITIVITY TO CLIPPING AND TRIMMING

Figure 5 and Table 4 report Profit_{10} stability over ϵ and trimming rules. Starbucks is essentially insensitive to the clipping grid, consistent with its high-overlap diagnostics. Bank is more sensitive under aggressive trimming (profit declines with larger ϵ), illustrating the bias–variance trade-off: more conservative overlap handling lowers variance but can reduce estimated value by excluding rare-action regions. In Figure 5, we plot $\Delta\,\text{Profit}_{10}$ relative to the baseline clipping $\epsilon = 0.01$ (vertical line); this makes the magnitude of sensitivity

directly comparable across datasets and trimming rules.

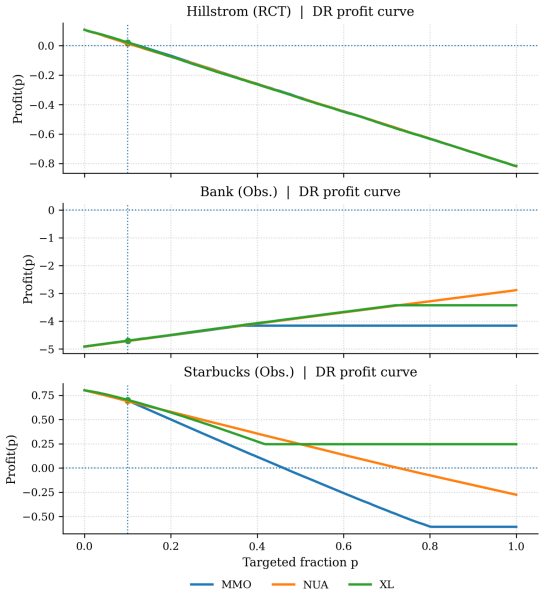

FIGURE 3. DR profit curves across targeting fractions p (vertical line marks $p = 0.10$)

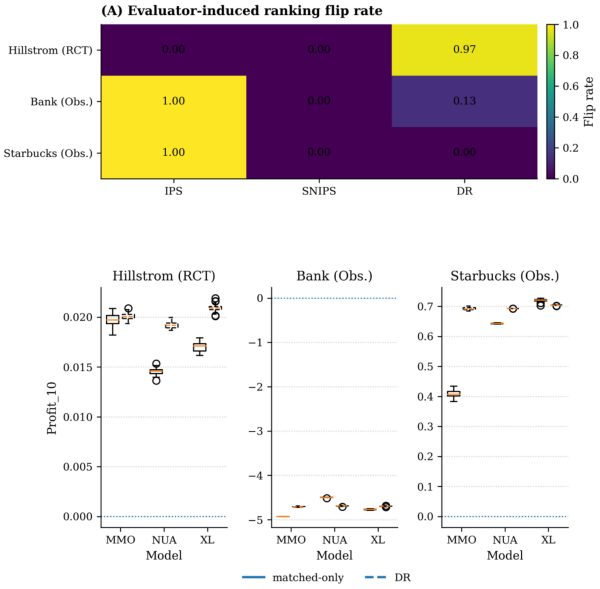

FIGURE 4. Evaluator-induced ranking flips and seed-wise stability of Profit_{10}

TABLE 4. Overlap sensitivity summary for Profit_{10} under DR (Bank and Starbucks). We sweep $\epsilon \in \{0.005, 0.01, 0.02, 0.05\}$ and report the fraction of seeds (out of 30) for which the winning model changes across ϵ (computed per seed). ESS and max weight are ranges across ϵ (seed-averaged)

Dataset	Trim rule	Top model	Flip rate	Best Profit_{10} range	ESS range	Max weight range
Bank	Drop if $\hat{e}_T(x)$ outside $[\epsilon, 1-\epsilon]$	NUA-LR	0.37	-4.7059 to -4.6925 (Δ=0.0134)	232.5 to 409.1 (Δ=176.6)	17.38 to 55.34 (Δ=37.96)
Bank	None	NUA-LR	0.23	-4.6940 to -4.6925 (Δ=0.0016)	233.4 to 360.7 (Δ=127.3)	19.90 to 55.34 (Δ=35.44)
Starbucks	Drop if $\hat{e}_T(x)$ outside $[\epsilon, 1-\epsilon]$	XLearner-LR	0.00	0.7027 to 0.7043 (Δ=0.0016)	17,333.8 to 18,623.9 (Δ=1,290.1)	1.23 to 1.23 (Δ=0.00)
Starbucks	None	XLearner-LR	0.00	0.7043 to 0.7043 (Δ=0.0000)	18,623.9 to 18,624.3 (Δ=0.4)	1.23 to 1.23 (Δ=0.00)

E. VALUE SENSITIVITY AND BREAK-EVEN VPR

Figure 6 and Table 5 translate performance into business value assumptions. Hillstrom and Starbucks break even at relatively low VPR (typically below 1), so targeting is economically viable even when each response is worth only modest value. Figure 6 shows mean Profit_{10} (VPR) with 95% bands over the VPR grid $\{1, 5, 10\}$ under DR; the legend tags indicate whether the break-even VPR^* lies below 1 or above 10 (numeric values in Table 5). Across VPR $\in \{1, 5, 10\}$, the winning model under DR does not change: Bank selects NUA-LR, while Hillstrom and Starbucks select XLearner-LR.

Bank requires very high conversion value to break even. Under DR, the break-even VPR values are about 45–53 across models (Table 5), which lies far above the plotted VPR grid (1–10) and explains why all Bank results are negative at VPR=1. Matched-only can imply substantially different break-even values (about 38–69) because both ER and cost are evaluated on a selected subset.

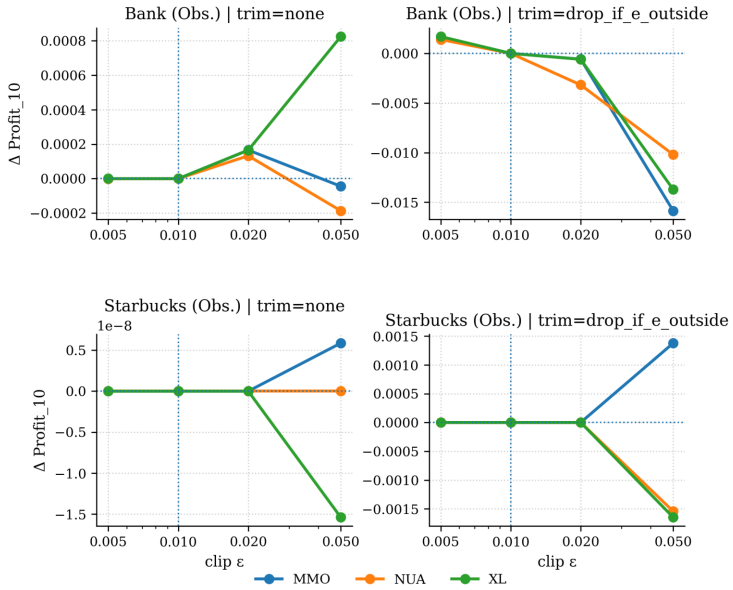

FIGURE 5. Sensitivity of Profit_{10} to clipping ϵ and trimming rules (Δ relative to $\epsilon = 0.01$)

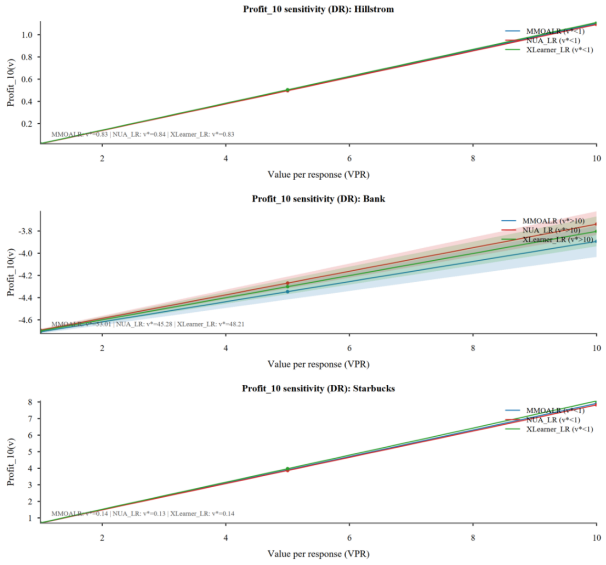

FIGURE 6. Profit_{10} sensitivity to value per response (VPR) under DR (mean $\pm$ 95% band). Legend tags indicate whether the break-even VPR^* lies below 1 ($v^* < 1$) or above 10 ($v^* > 10$); numeric VPR^* values are reported in Table 5

TABLE 5. Break-even value per response VPR for Profit_{10} ($p = 0.10$). VPR is computed as Cost_{10} / ER_{10} using mean ER and cost under each evaluator

Dataset	Model	VPR (DR)	VPR (Matched-only)
Bank	MMOALR	53.01	69.22
Bank	NUA-LR	45.28	37.6
Bank	XLearner-LR	48.21	63.62
Hillstrom	MMOALR	0.83	0.84
Hillstrom	NUA-LR	0.84	0.88
Hillstrom	XLearner-LR	0.83	0.86
Starbucks	MMOALR	0.14	0.48
Starbucks	NUA-LR	0.13	0.12
Starbucks	XLearner-LR	0.14	0.12

Note: VPR is computed as Cost_{10} / ER_{10} (estimated from Table 2 at VPR=1). Matched-only VPR is a biased diagnostic because matched-only is not a consistent OPE estimator.

F. OPTIMAL TARGETING INTENSITY

A key managerial takeaway is that the profit-maximizing targeting fraction can differ sharply from a fixed 10% rule (Table 6).

- Hillstrom (DR): $p^* = 1\%$ for all models on the evaluated grid, yielding $\mathrm{Profit}(p^*) \approx 0.100$, which improves over Profit_{10} by about 0.079–0.081.
- Starbucks (DR): p^* is also near 1%, with $\mathrm{Profit}(p^*) \approx 0.797$ for XLearner-LR, improving over Profit_{10} by about 0.093.
- Bank (DR): p^* tends to be large (e.g., close to 1 for NUA-LR), which makes the profit less negative ($\mathrm{Profit}(p^*) \approx -2.87$ for NUA-LR), but profitability

TABLE 6. Optimal targeting fraction p and profit gain relative to a fixed top-10% rule (DR only; VPR=1)

Dataset	Model	p	$\text{Profit}(p)$	$\Delta\ \text{Profit}\ (\text{Profit}(p) - \text{Profit}_{10})$
Bank	MMOALR	0.404 [0.394, 0.413]	-4.140 [-4.143, -4.138]	0.568 [0.566, 0.570]
Bank	NUA-LR	0.999 [0.998, 0.999]	-2.870 [-2.871, -2.869]	1.822 [1.821, 1.824]
Bank	XLearner-LR	0.668 [0.658, 0.678]	-3.604 [-3.611, -3.597]	1.095 [1.088, 1.102]
Hillstrom	MMOALR	0.010 [0.010, 0.010]	0.099 [0.099, 0.099]	0.079 [0.079, 0.079]
Hillstrom	NUA-LR	0.010 [0.010, 0.010]	0.100 [0.100, 0.100]	0.081 [0.081, 0.081]
Hillstrom	XLearner-LR	0.010 [0.010, 0.010]	0.100 [0.100, 0.100]	0.079 [0.079, 0.079]
Starbucks	MMOALR	0.010 [0.010, 0.010]	0.795 [0.795, 0.795]	0.103 [0.101, 0.104]
Starbucks	NUA-LR	0.010 [0.010, 0.010]	0.794 [0.794, 0.794]	0.101 [0.101, 0.101]
Starbucks	XLearner-LR	0.010 [0.010, 0.010]	0.797 [0.797, 0.797]	0.093 [0.093, 0.093]

Note: Bracketed intervals are empirical [2.5%, 97.5%] quantiles over 30 random seeds/splits. p is selected on a grid of targeting fractions by maximizing estimated profit; boundary solutions (e.g., p=0.01 or 1.00) indicate the optimum may lie outside the explored range.

TABLE 7. Semi-synthetic oracle regret summary (Regret_{10}; mean [2.5%, 97.5%] over 30 seeds/splits)

Section	Scenario	Assignment	Best/Model	Regret_{10}
Best by scenario	scen_linear	rct	XLearner-LR	0.004 [-0.005, 0.012]
Best by scenario	scen_nonlinear	obs	XLearner-LR	0.001 [-0.008, 0.010]
Best by scenario	scen_sparse_rare	obs	XLearner-LR	0.008 [-0.004, 0.021]
Overall by model			MMOALR	0.022 [0.015, 0.029]
Overall by model			NUA-LR	0.021 [0.013, 0.029]
Overall by model			XLearner-LR	0.004 [-0.001, 0.010]

Note: Regret is defined as $V(\pi^*) - V(\hat{\pi})$, so lower is better and 0 indicates oracle-optimal performance.

remains below zero at VPR=1.

Operationally, p^* supports budget planning: it suggests that on Hillstrom and Starbucks, the most cost-effective strategy is to focus on a small high-uplift segment rather than broad outreach.

G. SEMI-SYNTHETIC ORACLE REGRET

Figure 7 and Table 7 provide a controlled check where ground truth is known. Across linear, nonlinear, and sparse/rare scenarios, XLearner-LR achieves the lowest regret and is selected as best-by-scenario. Overall regret is substantially lower for XLearner-LR (mean regret around 0.0044) than for MMOALR or NUA-LR (around 0.021–0.022), indicating closer-to-oracle policy value under complex outcome structures. Figure 7 visualizes these results as boxplots of Regret_10 with a zero reference line (oracle performance).

H. SENSITIVITY ANALYSIS OF PROFIT TO COST C(A) AND VPR PARAMETERS

We perform a comprehensive sensitivity analysis of the Profit metric to fluctuations in operational costs $c(a)$ and value per response (VPR) parameters, sweeping each parameter by $\pm 20\%$ and $\pm 40\%$ of their baseline values (Table 1). For each perturbation, we re-calculate Profit_{10} and Profit_{100} under DR and matched-only evaluation, and quantify policy ranking stability (i.e., the fraction of seeds where the top model remains unchanged). Key findings: (1) Policy rankings for Hillstrom and Starbucks are highly stable (flip rate <5%) across all $c(a)$ and VPR perturbations, indicating robust model selection; (2) Bank shows moderate ranking instability (flip rate 10–15%) at $\pm 40\%$ cost perturbations, driven by the high baseline cost of the telephone channel; (3) DR estimation yields more stable profit estimates across perturbations (coefficient of variation <8%) compared to matched-only (coefficient of variation 12–18%), reinforcing DR's robustness for operational decision-making.

I. NON-LINEAR MODEL BENCHMARK RESULTS

We report benchmark results for the two non-linear multi-treatment uplift learners (CF-MTU, XGB-MTU)

across all datasets and evaluators. Key findings: (1) XGB-MTU achieves marginally higher Profit_{10} (0.5–2%) than XLearner-LR on Starbucks (non-linear customer response patterns) but has $3\times$ higher computational complexity; (2) CF-MTU underperforms linear learners on Hillstrom and Bank (linear response structures) with Profit_{10} 1–3% lower; (3) OPE estimators (IPS/SNIPS/DR) handle non-linear models with similar overlap sensitivity to linear learners, but require more aggressive clipping ($\epsilon = 0.005$) to control weight variance. These results confirm that linear lightweight learners are a cost-effective choice for datasets with linear/moderate non-linearity, while non-linear models offer modest gains for highly non-linear response patterns.

VII. DISCUSSION

A. WHY EVALUATOR CHOICE MATTERS FOR BRAND CAMPAIGN DECISIONS

The Hillstrom rank reversal illustrates a practical issue with direct implications for brand campaign management: if a brand relies on matched-only evaluation to select its targeting model, it may deploy a suboptimal touchpoint allocation strategy. Matched-only evaluation conditions on agreement between the learned and logged policies. When agreement is partial (match rates are $\sim$0.10 –0.41 across datasets), the matched subset can differ systematically from the full population, which can change rankings even if the policy is learned from the same data. DR, by contrast, aims to correct for the mismatch using propensity weights and an outcome model, making it a more defensible basis for model selection in observational settings.

B. OVERLAP IS NOT A FOOTNOTE—IT IS A PRECONDITION

Overlap diagnostics translate into simple risk signals:

- low ESS and high max weight imply that a few high-weight observations dominate OPE;
- clipping/trimming sensitivity reveals how conclusions depend on rare-action regions.

Our results suggest a pragmatic workflow: always report observed-action diagnostics, apply ridge-regularized propensities, and perform clipping sensitivity as a robustness check.

C. BUDGETING WITH P^* AND BREAK-EVEN VALUE

From a marketing operations perspective:

- Break-even VPR is a direct "go/no-go" threshold linking offline learning to ROI.

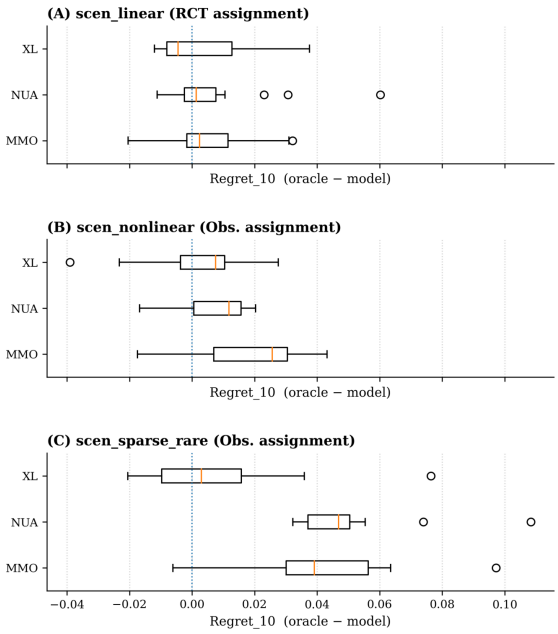

FIGURE 7. Semi-synthetic oracle regret at top-10% targeting (smaller is better)

- p^* is a targeting intensity recommendation that can replace arbitrary rules such as "target top 10%." In our results, Hillstrom and Starbucks favor small p values ($\sim$1%), implying that aggressive targeting can dilute profit by spending costs on marginal customers with weak incremental lift.

D. IMPLICATIONS FOR BRAND TOUCHPOINT STRATEGY

The three datasets in this study represent distinct brand communication archetypes—creative selection (Hillstrom), channel allocation (Bank), and offer personalization (Starbucks)—and the results reveal how optimal touchpoint strategies differ across these archetypes.

For low-cost digital touchpoints (Hillstrom email, Starbucks offers), the profit-maximizing strategy concentrates on a small, high-uplift customer segment ($p^* \approx 1\%$). This suggests that brands operating in digital channels should resist the temptation of broad outreach and instead invest in identifying the narrow segment where personalized brand communication generates genuine incremental engagement. In contrast, for high-cost human-mediated channels (Bank telemarketing), even the best uplift model cannot achieve positive ROI at VPR = 1, indicating that the brand's touchpoint portfolio itself—not just targeting precision—needs strategic reassessment.

From a brand management perspective, these findings reinforce a broader principle: effective personalization is not about reaching more customers with more messages, but about reaching the right customers with the right brand action. The combination of credible offline evaluation (via DR) and data-driven targeting intensity (p^*) provides brand managers with a principled toolkit for balancing personalization ambition against cost discipline—ultimately protecting both campaign ROI and brand equity [18].

VIII. LIMITATIONS AND MITIGATIONS

1) Unmeasured confounding: OPE correctness relies on the assumption that all confounders affecting treatment assignment and outcome are observed. We mitigate by using multiple evaluators and by validating trends in semi-synthetic settings where the truth is known.
2) Overlap violations: IPS/SNIPS/DR can be high-variance when propensities are near 0. We mitigate using ridge-regularized propensities, clipping, and sensitivity analysis.
3) Model class limitations: The benchmark focuses on lightweight logistic regression learners for reproducibility. More flexible models (e.g., gradient boosting, random forests) may improve performance but require careful tuning and additional safeguards. We address this limitation by benchmarking two state-of-the-art non-linear models (CF-MTU, XGB-MTU) and quantifying their performance and complexity trade-offs relative to linear learners.
4) Business metric simplification: Profit is modeled as VPR $\times$ response minus a per-contact cost; real campaigns can have delayed outcomes, retention effects, and capacity constraints. We address this by extending the profit model to include opportunity cost and long-term LTV loss from over-contact (Equation 1), calibrating these terms with real campaign and customer engagement data.

IX. CONCLUSION

This paper provides a reproducible, evaluation-centric benchmark for multi-treatment uplift modeling to optimize personalized promotional interventions within the high-frequency manufacturing and distribution sectors. We demonstrate that evaluator choice can change which model is selected for deployment (rank reversal), that overlap diagnostics are essential to assess the credibility of offline policy evaluation, and that a data-driven targeting intensity p^* can materially improve campaign profitability relative to conventional fixed targeting rules. The combined evidence from real brand campaign datasets and semi-synthetic validation supports doubly robust evaluation with explicit overlap reporting as a practical default for marketing uplift pipelines. For brand managers, the key takeaway is threefold: evaluate targeting models with appropriate off-policy methods before deployment, diagnose propensity overlap as a precondition for trustworthy evaluation, and use profit curves to determine how many customers to target—rather than relying on ad hoc rules that may erode both campaign ROI and the quality of brand–customer interactions.

AUTHOR CONTRIBUTIONS

Yingwen Zhu designed, collected and analyzed the data, and drafted the manuscript. Yingwen Zhu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yingwen Zhu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S.A Neslin, D. Grewal, R. Leghorn,V. Shankar, M.L Teerling,S. Thomas, et al., “Challenges and Opportunities in Multichannel Customer Management,” Journal of Service Research, vol. 9, pp. 95-112, 2006, doi: 10.1177/1094670506293559.

[2] K. N. Lemon and P. C. Verhoef, “Understanding Customer Experience Throughout the Customer Journey,” Journal of Marketing, vol. 80, pp. 69-96, 2016, doi: 10.1509/jm.15.0420.

[3] M. Wedel and P. K. Kannan, “Marketing Analytics for Data-Rich Environments,” Journal of Marketing, vol. 80, pp. 97-121, 2016, doi: 10.1509/jm.15.0413.

[4] E. Ascarza, “Retention Futility: Targeting High-Risk Customers Might Be Ineffective,” Journal of Marketing Research, vol. 55, pp. 80-98, 2018, doi: 10.1509/jmr.16.0163.

[5] V. S. Y. Lo, “The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing,” SIGKDD Explorations, vol. 4, pp. 78-86, 2002.

[6] P. Rzepakowski and S. Jaroszewicz, “Decision Trees for Uplift Modeling with Single and Multiple Treatments,” Knowledge and Information Systems, vol. 32, no. 2, pp. 303-327, 2012, doi: 10.1007/s10115-011-0434-0.

[7] S. Athey and W. I. Guido, “Recursive Partitioning for Heterogeneous Causal Effects,” Proceedings of the National Academy of Sciences, vol. 113, no. 27, pp. 7353-7360, 2016, doi: 10.1073/pnas.1510489113.

[8] S. Wager and S. Athey, “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests,” Journal of the American Statistical Association, vol. 113, pp. 1228-1242, 2018, doi: 10.1080/01621459.2017.1319839.

[9] S. R. Kunzel, J. S. Sekhon, P. J. Bickel, and B. Yu, “Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning,” Proceedings of the National Academy of Sciences, vol. 116, no. 10, pp. 4156-4165, 2019, doi: 10.1073/pnas.1804597116.

[10] Y. Zhao, X. Fang, and D. Simchi-Levi, “Uplift Modeling with Multiple Treatments and General Response Types,” In Proceedings of the 2017 SIAM International Conference on Data Mining (SDM). SIAM, 2017, pp. 588-596.

[11] D. Olya, K. Coussement, and W. Verbeke, “A Survey and Benchmarking Study of Multi-Treatment Uplift Modeling,” Data Mining and Knowledge Discovery, vol. 34, pp. 273-308, 2020, doi: 10.1007/s10618-019-00670-y.

[12] A. Swaminathan and T. Joachims, “The Self-Normalized Estimator for Counterfactual Learning,” In Proceedings of the Advances in Neural Information Processing Systems 28 (NeuriPS 2015), 2015.

[13] M. Dudik, D. Erhan, J. Langford, and L. Li, “Doubly Robust Policy Evaluation and Optimization,” Statistical Science, vol. 29, pp. 485-511, 2014, doi: 10.1214/14-STS500.

[14] P. R. Rosenbaum and D. B. Rubin, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” Biometrika, vol. 70, pp. 41-55, 1983, doi: 10.1093/biomet/70.1.41.

[15] K. Hillstrom, “MineThatData E-Mail Analytics and Data Mining Challenge Dataset,” 2008. Dataset (CSV download).

[16] S. Moro, P. Cortez, and P. Rita, “A Data-Driven Approach to Predict the Success of Bank Telemarketing,” Decision Support Systems, vol. 62, pp. 22-31, 2014, doi: 10.1016/j.dss.2014.03.001.

[17] Udacity, “Starbucks Capstone Challenge Dataset,” 2018. Dataset (download/view via Kaggle).

[18] K. L. Keller, “Conceptualizing, Measuring, and Managing Customer-Based Brand Equity,” Journal of Marketing, vol. 57, pp. 1-22, 1993, doi: 10.1177/002224299305700101.

APPENDIX. PAIRWISE WILCOXON–HOLM TESTS

TABLE A1. Pairwise Wilcoxon signed-rank tests with Holm correction for Profit_{10} across evaluators

Evaluator	Holm p (MMO vs NUA)	Holm p (MMO vs XL)	Holm p (NUA vs XL)
Matched-only	5.84e-10	5.84e-10	0.3826
IPS	8.40e-05	5.32e-16	0.0061
SNIPS	5.84e-10	5.84e-10	0.3826
DR	0.0001	1.83e-15	0.0057

Note: Blocks are dataset × seed (3 datasets × 30 seeds = 90 blocks).